Methodologies in Predictive Visual Analytics

by

Yafeng Lu

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2017 by the
Graduate Supervisory Committee:

Ross Maciejewski, Chair
Nancy Cooke
Huan Liu
Jingrui He

ARIZONA STATE UNIVERSITY

May 2017

ABSTRACT

Predictive analytics embraces an extensive area of techniques from statistical modeling to machine learning to data mining and is applied in business intelligence, public health, disaster management and response, and many other fields. To date, visualization has been broadly used to support tasks in the predictive analytics pipeline under the underlying assumption that a human-in-the-loop can aid the analysis by integrating domain knowledge that might not be broadly captured by the system. Primary uses of visualization in the predictive analytics pipeline have focused on data cleaning, exploratory analysis, and diagnostics. More recently, numerous visual analytics systems for feature selection, incremental learning, and various prediction tasks have been proposed to support the growing use of complex models, agent-specific optimization, and comprehensive model comparison and result exploration. Such work is being driven by advances in interactive machine learning and the desire of end-users to understand and engage with the modeling process. However, despite the numerous and promising applications of visual analytics to predictive analytics tasks, work to assess the effectiveness of predictive visual analytics is lacking.

This thesis studies the current methodologies in predictive visual analytics. It first defines the scope of predictive analytics and presents a predictive visual analytics (PVA) pipeline. Following the proposed pipeline, a predictive visual analytics framework is developed to be used to explore under what circumstances a human-in-the-loop prediction process is most effective. This framework combines sentiment analysis, feature selection mechanisms, similarity comparisons and model cross-validation through a variety of interactive visualizations to support analysts in model building and prediction. To test the proposed framework, an instantiation for movie box-office prediction is developed and evaluated. Results from small-scale user studies are pre-

sented and discussed, and a generalized user study is carried out to assess the role of predictive visual analytics under a movie box-office prediction scenario.

To my father and in memory of my grandma

ACKNOWLEDGMENTS

I would first like to express my sincere gratitude to my advisor, Dr. Ross Maciejewski, for his continuous support throughout my PhD study and related research, for his patience, motivation, and immense knowledge. He gave me unreserved help on research guidance and for the achievement of personal goals. He helped me to have a joyful journey in research. I would also like to thank my committee members, Dr. Huan Liu, Dr.Nancy Cooke and Dr. Jingrui He, who have provided me insightful comments and supported me further in my endeavors.

Furthermore, I would like to mention that some of the material presented here was supported by the NSF under Grant No. 1350573 and in part by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001. I also would like to thank my colleagues at the Visual Analytics and Data Exploration Research (VADER) Lab for their constructive criticism and helpful suggestions regarding this work as well as for their support. Last but not least, I would like to thank my family and friends who have provided me with their love and affection and have believed in my abilities. They been a pillar of strength behind me through the years allowing me to focus and achieve my goals.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

Predictive analytics is the practice of identifying patterns within data to predict future outcomes and trends. Such work is relevant across all scientific disciplines as data is constantly being collected to explain phenomena. In the Big Data era, increasing amounts of data have accelerated the need for predictive analytics methods as the ability to collect data has outstripped the pace of analysis. This has led to the rapid development of novel predictive analytics algorithms, predominantly black-box methods, where data is given as an input and prediction results are returned as output. As predictive analytics methods have been refined, prediction models have achieved high levels of accuracy in many applications. One success story is a partnership between The Weather Company and IBM to predict the impact of weather on business performance. They hypothesize that retailers can use such prediction models to improve supply chain management and demand forecasting in the energy sector or adjust staffing in retail sectors. Other examples include Kroger's grocery stores, which have seen revenue growth coming from behavioral models of individuals, using big data to move from coarse demographic targeting to individualized coupons and customized loyalty rewards [13]. However, as the models become more complex and the size of the data grows, new challenges in predictive analytics have arisen.

One well-documented example of the complexities in predictive analytics is Google Flu Trends [14]. Launched in 2008, Google had developed a linear model that touted a 97% accuracy rate for predicting cases of influenza-like-illness based on word search frequencies. However, in February 2013, Nature reported that Google Flu Trends was predicting more than double the cases of influenza-like-illness than reported by the

Centers for Disease Control [15]. Another classic example includes training a neural network to detect camouflaged enemy tanks [16]. In this example, researchers used pictures of forests with and without camouflaged tanks. However, the pictures with tanks were taken on cloudy days and those without tanks were taken on sunny days leading the classifier to identify cloudy versus non-cloudy, as opposed to identifying tanks. Other such examples exist in domains that require a high-degree of expertise for prediction, such as cyber security [17]. Such failures in predictive analytics applications have led to a demand for methods that can incorporate human-in-the-loop intelligence as part of the predictive analytics pipeline [17]. Users report issues in data reliability, a desire to understand the inner workings of a model, and a need to update portions of the model to apply domain-specific content [18]. To facilitate this, visualization has been used as a method to overview the data, illustrate the model rationale, present the prediction result, and validate the model response. In addition, visual analytics, as a means of supporting human-in-the-loop processes, has been developed to integrate human knowledge and machine learning for predictions.

Given the highly specialized nature of many predictive analytics tasks, research in the visual analytics community has focused on developing systems for explicit predictive analytics methods including regression (e.g. [19]), classification (e.g. [2]), clustering (e.g. [20]), and decision making (e.g. [9]). One major goal of such systems is to improve model comprehension [21] and improving comprehensibility of various phases of the predictive modeling process can lead to desirable outcomes including optimized predictions. Visual analytics methods in this domain are referred to as predictive visual analytics (PVA) in this thesis. Here, a PVA method is seen as a complement to the traditional predictive analytics pipeline where PVA focuses on utilizing visualization techniques to improve users' performance, perception, and externalization of insights.

This thesis first reviews extant literature in PVA and proposes a PVA pipeline. Following this pipeline, a PVA framework with novel visual analytics tools is presented and evaluated on revenue prediction for movie box-office using social media data and movie's meta data. Illustrated through the movie box-office prediction task, this thesis shows that in many prediction analyses, experts' domain knowledge is very useful because they are sensitive to the data, such as noticing grouping of features, abnormal signals, and multi-correlation. In this thesis, multiple machine learning models are embedded to provide guidance for the analysts, and the raw data and features not used in those models are all exposed to the end user. A parallel coordinate plot (PCP) is used to show feature correlations, and a similarity widget is used to explore the local performance of the embedded models. In the case studies of the proposed visual analytics framework, an exploration and analysis procedure is provided to guide the analysts through the exploration but still leaves interactive modeling and decision making steps to the analyst. Different data sources can provide complementary information, and different models can also contribute to the same analytics task from multiple perspectives. Visual analytics tools display those multi-facet data and data sources while integrating different models to provide a comprehensive analysis toolkit for the analysts. In the movie box-office prediction problem, three data sources (Twitter, IMDB, and YouTube) are integrated together for the prediction of upcoming movies' opening weekend gross. For the prediction of each single movie's revenue, this framework supports the creation of three different types of models: Support Vector Machine (SVM) [22], Linear Regression (LIN) [23] and Multilayer Perceptron (MLP) [24]. The results given by these models configured by the user can be displayed together with the selected dataset and a 95% confidence range. Considering the context of the movie industry, a temporal model is also provided to estimate the total revenue available for the weekend for all new released

movies. Those models, none of which are accurate enough to solve the problem alone, can contribute to the prediction through the visual analytics toolkit.

This thesis discusses the proposed visual analytics framework and view designs, as well as experimental studies for box office prediction. It shows that visual analytics can help effectively analyze problematic datasets, integrating multiple data sources, multiple learning models and experts' domain knowledge into general and specific analytics tasks. Inspired from the successful case studies on predicting movie revenue with visual analytics on social media data, this thesis also analyzes and discusses the visual characteristics of network properties in social media data. In addition, a user's role in predictive visual analytics will be discussed through a user study by testing the prediction performance under conditions of different models and objects.

Chapter 2

PROBLEM STATEMENT

This thesis studies the methodologies in predictive visual analytics through the problem of using social media data to predict movie box-office revenue. Although much research has been done on social media data analysis and movie revenue prediction, many challenges still exist in this field and the evaluation of predictive visual analytics. This section will discuss the problems in social media data analysis and the role of human-in-the-loop in prediction, and develop a list of research questions studied in this thesis.

2.1 Problems in Social Media Data Analysis

The noisy and unstructured nature of social media data presents opportunities and challenges in knowledge mining and retrieval. Previous works have shown the power of such data in many aspects, such as anomaly detection, human behavior analysis, and sentiment analysis of election and product reviews. Many works from the visualization community have emerged to explore social media data and analyze it to study network structures [25, 26], spatiotemporal distributions [27, 28], and social problems [29, 30]. However, the problems of using social media data for predictive analytics caused by its content, format, volume and value are not yet solved. In order to use social media data to develop prediction models, features are usually extracted and selected to encode information from raw data that allows machine learning algorithms to classify an unknown object or estimate an unknown value. Features are important to predictive models; the quality and quantity of features have great influence on whether the model is good or not. Feature engineering by

itself is hard; however, problems with feature engineering are further exacerbated in social media data due to the size and dimensionality of the data. Some of these problems are discussed below.

Data Collection: In the scientific community, Twitter is a social media giant famous not only for its popularity but also for sharing its data. Twitter provides a glance into its millions of users and billions of tweets through a Streaming API which provides a sample of all tweets matching some parameters preset by the API user. The Twitter Streaming API [31] allows anyone to retrieve at most a 1% sample of all the data by providing some parameters. According to the documentation, the sample will return at most 1% of all the Tweets produced on Twitter at a given time. Once the number of tweets matching the given parameters passes 1% of all the tweets on Twitter, Twitter will begin to sample the data returned to the user. The representative and bias of this sampled data in different applications has been discussed [32, 33], and their results show that for some prediction tasks, this sampling does affect the modeling process and model's performance. Other than the sampling in Twitter's Streaming API, Twitter only allows users to crawl most recent tweets posted within a time window (which is 7 days for now) using keyword search. It is possible that the data collection process can be unexpectedly interrupted, which may cause corruption and missing elements in the data.

Tweets Volume: The volume of the social media data, for instance the number of tweets posted per day, has been used as a significant feature in prediction models. Though this number can be understood as how popular a topic is, trust issues exist, and whether the volume should be used after filtering out suspicious posts or not is a question. Methods for detecting those suspicious posts and users is another research question. Thus, based on different filtering strategies, the volume of tweets may change, as well as the role of the volume in the modeling process.

Sentiment Analysis: Opinion mining on social media data has been studied widely in recent years. Sentiment classification has demonstrated its usefulness on election prediction [34] and customer review analysis for product retooling [34]. However, the performance on social media data is still unsatisfactory due to the distinct data characteristics [35, 36]. First, social media posts are mostly short and unstructured. For example, Twitter allows no more than 140 characters and uses many informal words such as "cooool" and "OMG". The short texts can hardly provide sufficient statistical information for learning based models. Second, it is laborious and time consuming to obtain ground truth for training data, which is needed to build an effective supervised learning model.

2.2 Problems in Human-in-the-Loop Process

Human-in-the-loop is a means of mitigating model complexity and specializing on domains and problems. Human-in-the-loop solutions take advantage of human knowledge and intelligence in the analytic process so that features that can hardly been captured by automatic models are able to be used through human interactions. Predictive visual analytics uses human-in-the-loop methods to solve prediction problems. Providing human knowledge in the predictive model construction process can be beneficial. However, opening the black-box of predictive models for human intervention is not without issues. By giving users the option to integrate their domain knowledge, it has also allowed them to inject bias into the model. What's the point of using technology to learn something new when you are bending it to fit your preexisting notions? More seriously, how can the knowledge integration be regulated or constrained so that it gets the benefits of domain knowledge, social and emotional intuition, and minimizes the costs of introducing bias? How much human-in-the-loop is the right amount? A recent study [18] ran experiments on incentivized forecasting

tasks where participants could choose to use forecasting outputs from an algorithm or provide their own inputs. The study found that letting people adjust an imperfect algorithm's forecasts would increase both their chances of using the algorithm and their satisfaction with the results. However, the authors also found that participants in the study often worsened the algorithm's forecasts when given the ability to adjust them. This further brings into question how much interaction should be provided in the PVA pipeline. Results from the forecasting study also indicated that people were insensitive to the amount that they could adjust the forecasts, which may indicate that interaction as a placebo could be an option. Given these problems and studies, it is clear that more research is needed to provide clear guidelines for predictive visual analytics methodologies.

2.3 Research Problems

Considering the aforementioned problems in social media data analysis and human-in-the-loop processes for prediction, this thesis focuses on developing visual analytics frameworks for the following research questions:

- Propose a general predictive visual analytics pipeline.
- Develop visual analytics tools to analyze social media data with a goal of making predictions.
- Analyze visual characteristics and their effects on predictive analytics.
- Evaluate the role of human in predictive visual analytics.

To study the role of human in predictive visual analytics, an evaluation environment is also developed.

Chapter 3

RELATED WORK

Predictive analytics is a core research topic with roots in statistics, machine learning, data mining, and artificial intelligence. Definitions of predictive analytics ranged from broad—every machine learning technique is predictive analytics—to narrow—making empirical predictions for the future [37]. A discussion on predictive analytics and predictive visual analytics will be first presented as related work to this thesis. Following the concepts in predictive analytics and the scope of predictive visual analytics, representative predictive visual analytics techniques are categorized and discussed as they are related to the development of the predictive visual analytics framework in the thesis. Recent visual analytics systems for social media analysis include Whisper [38], which focused on information propagation in Twitter, SensePlace2 [39], which focused on the analysis of geographically weighted Tweets, and TweetXplorer [40] which combined geographical visualization of Tweets along with their social networks. Other applications have explored the use of social media analytics for improving situational awareness in emergency response. Thom et al. [41] and Chae et al. [27] developed spatiotemporal visual analytics systems that integrated various social media data sources for anomaly event detection and disaster management. The proposed framework takes cues from this previous work and is developed to integrate data from multiple sources and provide an environment for user-in-the-loop predictive analytics. Relevant to the evaluations of predictive visual analytics in the thesis, extant research work on evaluating predictive visual analytics is discussed.

3.1 Predictive Analytics

Predictive analytics covers the practice of identifying patterns within data to predict future outcomes and trends. With respect to analytics, three common terms are descriptive, prescriptive, and predictive analytics. Descriptive analytics focuses on illustrating what has happened and what the current status of a system is. Prescriptive analytics uses data to populate decision models that produce optimal (or near-optimal) decisions of what should be done, and predictive analytics applies algorithms to extrapolate and model the future based on available data. In this sense, one can think of descriptive as a passive, historical analysis, prescriptive as active analysis suggesting how to exploit new opportunities, and predictive as an intersecting medium where historical data is used to produce knowledge of what is likely to happen as a means of driving decisions. Predictive analytics uses predictive modeling which should be empirical (based on observations) rather than theoretical (based on hypothesis). Arguably, the main tasks in predictive analytics are relevant to numerical predictions (where the most common predictive analytics methods are regressions), and categorical predictions (where the most common methods focus on classification and clustering) [42]. As an introduction to predictive analytics techniques, a brief definition of regression, classification, and clustering is provided.

Regression analysis is a statistical technique for modeling the relationships between variables [23]. Specifically, regression analysis focuses on understanding how a dependent variable changes when a predictor (or independent variable) is changed. Linear regression is perhaps one of the most common predictive analytics techniques available to analysts with implementations in Excel, SAS, JMP, and many other common software packages. Much of its power comes from the interpretability of the model, where relationships tend to be readily explorable by end users. For different

relationships between variables, regression models can be linear and non-linear and to explore local patterns, segmented or piecewise regression models can be used. Challenges exist in data transformation, feature selection, and model comparison, and widely used techniques to address these challenges include stepwise feature screening and comparing models through performance measures, such as the p-value and R^2.

Classification broadly covers the problem of identifying which category a new observation belongs to based on information from a training data set in which observations have known category memberships. Classifiers learn patterns using the data attribute features from the training set and these patterns can be applied to unknown instances to predict their categories. Well-known classification methods include Bayesian classification, logistic regression, decision trees, support vector machines (SVM), and artificial neural networks (ANN). Challenges with classification exist in learning large and/or streaming data (e.g., real-time security classification [43]), defining proper cost functions in model optimization for domain specific tasks where the error cost varies on instances, obtaining enough labeled data samples, and understanding what characteristics the models have learned.

Similar to classification, *clustering* also attempts to categorize a new observation into a class membership. However, clustering is an unsupervised method that discovers the natural groupings of a data set with unknown class labels. Clustering has been widely used in pattern recognition, information retrieval, and bioinformatics, and popular applications include gene sequence analysis, image segmentation, document summarization, and recommender systems. Challenges with clustering exist in feature extraction due to the high dimensionality and unequal length of feature vectors, metric learning, and clustering evaluation due to unknown ground truth. This thesis specifically considers clustering as a prediction task given the current use of

clustering for prediction [44, 45] along with a variety of work in visualization focused on clustering analysis.

In the context of the research presented here, predictive analytics is considered to be the method of analysis in the process of prediction modeling which consists of building and assessing a model aimed at making empirical predictions [37]. Predictive analytics overlaps with the process of knowledge discovery, but the emphasis is on predictions, specifically forecasts of the future, unknown, and 'what if' scenarios [46, 47]. The goal of prediction modeling is to make statements about an unknown or uncertain event, which can be numerical (prediction), categorical (classification), or ordinal (ranking) [37]. In this context, a paper falls into the scope of predictive analytics if it satisfies the following conditions:

1. The analysis process has a clear prediction goal. While open-ended explorations and exploratory data analysis play a role in predictive analysis, the task must be to ultimately make a statement about an unknown event.

2. The analysis process uses quantitative algorithms, such as statistical methods, machine learning models, and data mining techniques to make grounded and reasonable predictions. This means the modeling process should be data-driven as opposed to purely theory-driven.

3. The predictions or the prediction models themselves have a means of being evaluated.

Finally, if the model developed only extracts or explains features, patterns, correlations, and causalities but does not make reference to future predictions or 'what if' scenarios, it is not considered to fall under the scope of predictive analytics. The reason for the chosen scope is that in order to make a prediction, the model needs to

be applied to unknowns. This scope is useful in categorizing and reviewing predictive visual analytics papers and clarifying the gap in this field.

3.2 Scope of Predictive Visual Analytics

Predictive analytics methods primarily rely on a black-box approach consisting of a four-step process of data cleaning, feature selection, modeling, and validation [48, 49]. This thesis broadly considers predictive visual analytics to cover the domain of visualization methods and techniques that have been used to support the predictive analytics process. In the context of the research presented here, a paper falls into the scope of predictive visual analytics if the paper satisfies the following conditions:

1. The predictive visual analytics method is specific to prediction problems, not only confirmatory or explanatory problems. This means the task of the predictive visual analytics system, method, or technique, is to support analysts in making predictions.

2. The predictive visual analytics method enables the user to interact with at least one step in the predictive analytics process through exploratory visualization (as opposed to traditional interactions in user interfaces such as save and undo).

3. The predictive visual analytics method supports both prediction and visual explanation, which allows analysts either to improve model performance with respect to the general accuracy or to improve an analyst's understanding of the modeling process and output.

Moreover, predictive visual analytics methods should share the same goal as predictive analytics methods, which is to make accurate predictions. In addition, predictive visual analytics could also focus on improving users' satisfaction and confidence in the resulting predictions. While decision-making systems overlap with this definition,

this thesis does not specifically survey such tools as this falls more in the realm of prescriptive analytics. If a system supports decision making, this decision has to be made directly from predictive algorithms to be categorized in this survey.

To further clarify the scope of predictive visual analytics in this thesis, it is important to note that there are visual analytics papers that are related to predictive analytics but are considered to be out of the scope. Specifically, visual analytics works that use predictive analysis methods for guiding the design of a visualization (e.g., placing a flow field [50]) are not part of the definition as the goal of such papers is not to make a prediction but to use prediction to help improve the rendering process. Another example of a related, but excluded, work is the work by Tzeng and Ma [51] which proposed a visualization method to explain the behavior of artificial neural networks (ANNs). This paper focused on the design of a visualization but provided no interactive analytics for the classification using the ANN. Such methods that do not provide interactivity are also considered to be outside of the scope of predictive visual analytics definition of predictive visual analytics. While not considered in the above defined scope of predictive visual analytics, these works are still related to this research as they contribute to the development of predictive visual analytics for broadening capabilities and improving presentations. Similarly, uncertainty visualization [52], risk visualization, and ensemble visualization are also excluded.

3.3 Predictive Visual Analytics Techniques

Numerous solutions have been proposed to address predictive analytics for both novice and expert users (e.g., R [53], SAS [54], Weka [55], JMP [56], Excel). These software packages and tools provide a variety of machine learning algorithms that can be used for predictive analytics tasks, such as feature selection, parameter optimization and result validation. Many of these systems offer basic visualizations including

residual plots, scatterplots and linecharts. However, most visualizations are only used to display the final results and statistical evaluation report but do not provide interactive means for manipulation, feature selection or model refinement; instead, these systems often opt to show baseline models or simple statistical measures for result validation, working as more of a black-box system. To support flexible modeling procedure and exploration of the data and the model, researchers in the visual analytics community have been developing methods for improving model building and predictive analytics.

Besides improving prediction accuracy, predictive visual analytics also works to supporting comprehensibility, which is required for domain experts to understand the prediction model and integrate their knowledge in cases falling within their domain of expertise and to make decisions with confidence. In terms of supporting different analyses, predictive visual analytics techniques can be generally categorized into two types. One is developed for a particular application scenario and therefore categorized as an agent-based predictive visual analytics technique. The other is developed for a type of prediction task, such as clustering or classification, which is categorized as method-based predictive visual analytics technique.

Application-based techniques are usually proposed with specific needs and applications, such as bioinformatics analysis (e.g, the work by Barlow et al. [57]) and criminal analysis (e.g, the work by Malik et al. [58]). To support agent-based particular requirements, these techniques and systems often integrate specialized prediction approaches, visualization views, and interactions for a special domain.

There are systems developed for special expertise. For example, in public health, Afzal et al. [9] presents a decision history tree view for analyzing epidemic simulation results and strategy decision making. This work has a specific application scenario in epidemic predictions with large simulation data. It uses branching time paths to

show prediction models' results and allows users to add/remove mitigation measures to explore different epidemic prediction trends over time. In bioinformatic analytics, Barlow et al. [57] use prediction model and visual analytics to support the analysis of protein flexibility neighborhoods. Model results are visualized on a color-coded 2D flexibility plot, where a slider can be adjusted by the user to scan any region. Their system allows experts to understand what causes proteins to change shape under varying empirical parameters. For law enforcement, Malik et al. [58] developed a proactive and predictive environment for domain experts in decision making and prediction problems about community policing and law enforcement. To support the analysis, they proposed a spatiotemporal prediction method utilizing a Seasonal Trend Decomposition based on Loess (STL) smoothing and a kernel density estimation (KDE) approach. For video content filtering, Höferlin et al. [59] present an interactive learning approach integrating active learning, cascading classifiers, and visual analytics for video frame classifier training. It visualizes the class distribution and classification performance of each cascade node and enables the user to select video instances for labeling with a 2D projection and a video context view to investigate the video content. Taking new annotated data, active learning updates the classifier's predictions. For traffic control, Buchmüller et al. [60] develop a visual analytics system for understanding airplane movements (trajectories) and predicting flights' departure, arrival, and behaviors. Users are allowed to include/exclude flights based on the trajectories and their domain knowledge. The prediction view visualizes the predicted flight density over gridded areas, and it allows the user to manually change feature values, such as weather conditions, to explore possible outcomes. In these agent-based predictive visual analytics works, the user is often a domain expert with some knowledge of how to improve the model performance.

Other than these systems, some agent-based systems focus on problems requiring more general domain knowledge. For example, the text document retrieval model proposed by Heimerl et al. [1] is applicable to users with some experience in using a document search engine. This is a knowledge many people have, or can be acquired in a short time. This visual analytics approach is derived from active learning in which the user participate in the loop of classifying documents into relevant and irrelevant groups. In each iteration of the active learning classification, current model performance is visualized. The system also visualizes some selective instances and their distance to the classification boundary for the user to select and annotate the data. Users can decide which data to include for the next model iteration and manually assign the class label, and the system will highlight the data points whose prediction will change based on this update for users to preview the result before running the model. Another example lies in business intelligence. Lu et al. [61, 62] present a visual analytics framework for box office prediction. This work integrated a linear regression model, a time series model, and different visual analytics views for investigating social media data and model performance. The results have indicated that users being familiar with movies or having experience in developing prediction models could make comparable predictions to the experts from box office websites.

Method-based techniques are usually designed for a particular prediction model or task regarding the modeling procedure, such as feature selection. For example, decision tree construction is one of the most frequent problems that visual analytics has tackled. Ankerst et al. [63] present a visual classification approach on decision tree construction. They used the circle segments visualization to present data attributes, and users can manually select features, split nodes, and change data labels while constructing the tree, and backtrack previous interactions. More recent work about decision trees, BaobabView [8], also enables the model developer to grow the

tree, prune branches, split and merge nodes, and tune parameters as part of the tree construction process. Baobabview presents an ad-hoc tree-look visualization and interactions for decision tree construction.

In addition to decision trees, predictive visual analytics has also been used in broader works for classification, clustering, and regression. For example, iVisClassifier [2] proposed a classification system based on linear discriminant analysis (LDA). This system links parallel coordinates plots, heat maps, and reconstructed images for the user to explore the data structure with reduced dimensions. Users can manually label unknown data and trigger a new round of LDA by removing/adding labeled instances. iVisClustering [64] implements a document clustering system based on Latent Dirichlet Allocation topic modeling (distinct from the LDA used in iVisClassifier), and visualizes the output of LDA topic model by displaying cluster relationships based on keyword similarity in a node-link cluster tree view. IVisClassifier also supports the control of the model parameters through a Term-Weight view to update the LDA model. Scatter/gather clustering [3] supports interactive clustering by enabling users to set soft constraints on the clustering method and compare clustering results. Users can set the number of clusters, and the system will update the model by applying scatter (changing to more clusters) or gather (changing to less clusters) operations. Dis-Function [65] allows the user to interact directly with a multidimensional scaling (MDS) scatter plot to update the underlying clustering distance function. Mühlbacher et al. [6] propose a visual analytics system for segmented linear regression. It allows users to analyze the prediction performance by setting different regression targets with regressors being single features or pairwise interactions.

There are also visual analytics techniques focusing on a particular task in predictive analytics. For example, INFUSE [5] is a visual analytics system focusing on feature selection for classification. It proposes a visual feature glyph which displays

the feature performance by different measurements in cross-validation, according to which users can select a feature subset for classification modeling. Squares [12] is a visual analytics technique for comparing and selecting classification models by visualizing the results of multiclass classification on both instance level and class level performance. Small multiples are used to support the analysis of the probability distribution of each class label from one model, and Squares uses squares to visualize the prediction result and error type for each instance. In these ways, users can discover detail differences between classifiers which can hardly be compared by accuracy. For more complex models, Rauber et al. [11] present a visual analytics work on understanding the prediction results and the hidden layer states of a neuron network. It uses dimension reduction techniques to project data instances and neurons in multilayer perceptrons and convolutional neural networks to present both the classification results and the relationships between artificial neurons.

Among these predictive visual analytics techniques, the benefit of using them lies in improving prediction accuracy and model comprehension by enabling human-in-the-loop approaches. To evaluate the effectiveness of predictive visual analytics in terms of accuracy and domain knowledge integration, this thesis uses box office prediction, which is a numerical prediction task which can be evaluated by different measures.

3.4 Human Factors in Predictive Analytics

Researchers have also studied the effect of human adjustments to statistical forecasts in management science, such as sales forecasting [66], time series forecasting [67], and hiring [68]. Here, controversial results emerge; some studies indicate that human judgments when performing forecasting result in lower accuracy [69, 70], whereas others indicate a significant benefit to forecasting accuracy [66, 71].

The field of human factors has long been interested in the relationship of humans and technology, and the effect of biases on human decision-making has been studied in human factors throughout multiple different contexts. In particular, confirmation bias- seeking out information to confirm decisions [72, 73], overconfidence bias- being too confident in abilities which leads to taking risks [74], and anchoring- over-reliance on first piece of information found [75] are specific human biases that are known to affect decision-making. In regard to predictive analytics, each of these biases may impact how humans utilize different analytical tools and methods. A human may be more or less prone to utilize a predictive model depending on the bias that is in play. The concept of trust can also effect biases, and thus the usage of predictive analytics. A great deal of human factors work has focused on how humans trust machinery, specifically automation and autonomy [76]. Recent work by Hoff and Bashir [77] identified that human's trust in automation is highly dynamic and dependent on a multitude of factors. Those factors can be distilled into three main areas of trust: dispositional trust (dependent on culture, age, gender, personality), situational trust (type of system, system complexity, task, etc.), and learned trust (past or current experience with systems). These aspects of trust in automation are likely relevant and valuable to the context of predictive analytics.

With these uncertain human factors, researchers have also studied the reasons why people prefer to use human predictions over automatic models in many scenarios [78, 79]. One reason is that humans seem to have an inherent distrust of algorithmic models, and examples of this distrust are found in various fields including organizational planning [80], hiring [68], and clinical predictions [81]. Here, people rely on their judgment and intuition much more than prediction algorithms, although investigations show that the prediction performance could be improved if they followed some principles and computational models. Other studies have also indicated

that lack of trust stems from the limitation of automatic techniques and challenges of model explainability [21, 82].

This algorithm aversion phenomenon is further discussed by Simmons and Massey [83] where studies indicate that people are less likely to use forecasts from an algorithm after seeing it perform and learning that it is imperfect, even if they also see that it outperforms the human forecaster who serves as the alternative forecasting method. A related study [18] found that people are much more willing to use forecasts from an imperfect algorithm when they can retain a slight amount of control over the algorithms forecasts. This study found that letting people adjust an imperfect algorithm's forecasts would increase both their chances of using the algorithm and their satisfaction with the results. However, the authors also found that participants in the study often worsened the algorithm's forecasts when given the ability to adjust them. Dietvorst [84] also studied the decision process that leads people to rely on human predictions instead of algorithmic predictions. In this study, he identified which prediction method to use depends on (1) the status quo prediction method which is a default choice, and (2) whether an alternative method can meet people's counter-normative reference points.

Given these results, it is clear that more studies are needed to provide guidelines for methodologies and designs in predictive visual analytics. By giving users the option to integrate their domain knowledge, we have also allowed them to inject bias into the model. What's the point of using technology to learn something new when you are bending it to fit your pre-existing notions? More seriously, how can we regulate or constrain knowledge integration so that we get the benefits of user knowledge and social and emotional intuition while minimizing the costs of introducing bias? How much human-in-the-loop is the right amount?

Chapter 4

PREDICTIVE VISUAL ANALYTICS PIPELINE

4.1 PVA Pipeline

This thesis considers predictive visual analytics to fall squarely under the umbrella of human-machine integration and a key aspect of predictive visual analytics should be supporting model comprehensibility [21]. This thesis defines predictive visual analytics as visualization techniques that are directly coupled (through user interaction) to the predictive analytics process. The four steps of the predictive analytics pipeline (data preprocessing, feature engineering, model building, and model selection and validation) serve as a basis for defining the PVA pipeline (Figure 4.1). The definition of the PVA pipeline is further informed by the knowledge discovery process of Pirolli

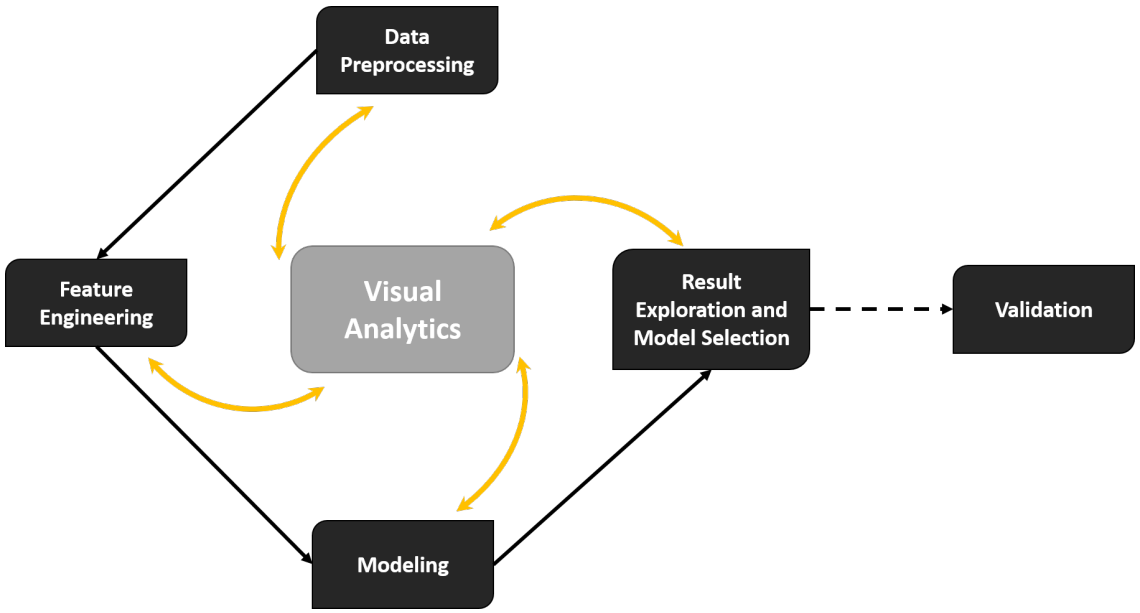


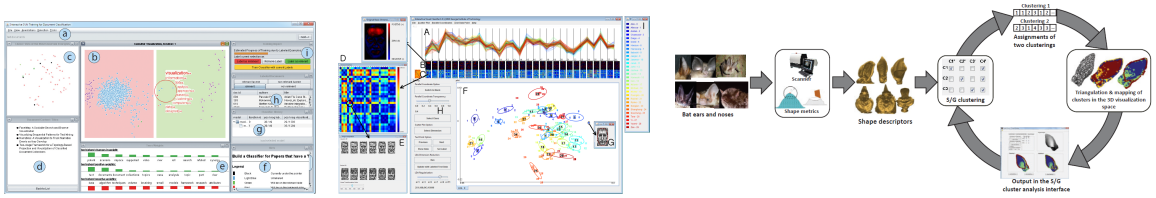
Figure 4.1: The Predictive Visual Analytics Pipeline

Table 4.1: The Goal of Each Step in the PVA Pipeline and the General Predictive Analytics Procedure

	PA	PVA Exclusive
Overall Goal	Make Prediction	Support Explanation
Data Preprocessing	Clean and format data	Summarize and overview the training data
Feature Selection and Generation	Optimize prediction accuracy	Support reasoning and domain knowledge integration
Modeling	Optimize prediction accuracy	Support reasoning and domain knowledge integration
Result Exploration and Model Selection	Model quality analysis	Get insights; Select the proper model; Feedback for model updates
Validation	Test for overfitting	Get insights from other datasets

and Card [49] and a variety of recent surveys on topics ranging from visual analytics pipelines and frameworks [85–87] to human-centered machine learning [88–90] to knowledge discovery.

As a starting point for defining the PVA pipeline, a four step pipeline of predictive analytics and general data mining [48] consisting of *Data Preprocessing*, *Feature Engineering*, *Model Building*, and *Model Selection and Validation* is developed. Chen et al. [91] extended this pipeline by adding an *Adjustment Loop* and a *Visualization* step allowing for the application of different visual analytics methods within the general data mining framework. Similar to previously proposed frameworks, the proposed PVA pipeline here is also built on top of the typical process of knowledge discov-



(a) (b) (c)

Figure 4.2: Examples of PVA Systems Representing the Entire PVA pipeline: (a)A Document Classifier Training System [1], (b)iVisClassifier [2], and (c) Scatter/Gather Clustering [3]

ery. In this thesis, Chen et al.'s pipeline is extended, splitting the process into five steps: *Data Preprocessing*, *Feature Engineering*, *Modeling*, *Result Exploration and Model Selection*, and *Validation*, as shown in Figure 4.1. The step *Model Selection and Validation* is separated into two distinct steps, *Result Exploration and Model Selection* and *Validation*, and the interactive analytics loop is represented by bidirectionally connecting the first four steps with *Visual Analytics*. The proposed pipeline highlights two specific aspects of PVA systems:

- Visual Analytics can be integrated into any of the first four steps iteratively so that these steps need not proceed in a specific order in every iteration.
- In the validation step, model testing can be applied. Users are able to go back to the first four steps after validation, but the integration level must be shallow to prevent overfitting and conflation of testing and training data.

Given the tight coupling of interaction in the pipeline, a detailed categorization of interactions found in the predictive visual analytics literature, building off of Yi et al.'s interaction taxonomy [92] is also provided. To illustrate the difference between

PVA and the general predictive modeling process, the goals of predictive analytics are summarized and compared to the goals of predictive visual analytics in Table 4.1.

To date, few visual analytics systems have been developed to support the entire PVA pipeline. Some examples of the most comprehensive PVA systems are presented in Figure 4.2. Heimerl et al. [1] discussed three text document classification methods covering the first four steps in the PVA pipeline and used statistical model validation on the test dataset to demonstrate the effectiveness of the interactive visual analytics method (Figure 4.2a). iVisClassifier [2] (Figure 4.2b) proposed a classification system based on linear discriminant analysis (LDA). This system emphasizes the data preprocessing step by providing parallel coordinates plots, heat maps, and reconstructed images for the user to explore the data structure with reduced dimensions. iVisClassifier's feature engineering and modeling steps are embedded in the classification and involve significant manual work where users need to label the unknown data and trigger a new round of LDA by removing/adding labeled instances. However, iVisClassifier has only been demonstrated using case studies without a well-established testing and validation step. Scatter/gather clustering [3] (Figure 4.2c) has black-box data preprocessing and feature extraction steps, but the system supports interactive clustering as part of the modeling phase where users can set soft constraints on the clustering method and compare clustering results.

Other representative examples that cover the complete PVA process include a visual analytics framework for box office prediction [62] using iterative data integration, feature selection, and modeling with results exploration and model validation measure; a predictive policing visual analytics framework [58]; Peak-Preserving [93] time series prediction; and iVisClustering [64] which implements a document clustering system based on latent Dirichlet allocation topic modeling (distinct from the LDA used in iVisClassifier). iVisClustering also supports relation analysis between

clusters and documents by using multiple views to visualize the clustering results and a Term-Weight view to control the parameters that update the clustering model.

The analysis of papers indicates that visual analytics developers tend to focus only on portions of the PVA pipeline. Even in cases when all of the steps of the pipeline are found within a single system, several steps will often lack a direct connection to any visualization. Instead, many steps are often left as black-boxes to the user in order to focus on a subset of steps within the PVA pipeline. The most commonly neglected step tends to be the data preprocessing step and the formal validation step that utilizes testing of the prediction model. The formal validation step is quite rare in visual analytics papers, though simple performance measures are often reported. Given the lack of full pipeline support, this thesis categorizes the surveyed papers according to which of the PVA pipeline steps were supported.

4.2 Data Preprocessing

Data Preprocessing has two objectives. The first objective is to understand the data, and the second objective is to prepare the data for analysis. Typical preparation approaches include data cleaning, encoding, and transformation. Examples of systems where data preprocessing is firmly integrated into the predictive analysis loop include the work by Krause et al. [4] which presents a visual analytics system, COQUITO, focusing on cohort construction by iteratively updating queries through a visual interface (Figure 4.3). The usability of this system has been demonstrated in diabetes diagnosis and social media pattern analysis. Other examples include the Peak-Preserving time series predictions by Hao et al. [93] which supports noise removal prior to building prediction models for seasonal time series data. Lu et al. [61] propose a system for predicting box-office revenue from social media data. Their system allows users to refine features by deleting noisy Twitter data which then updates

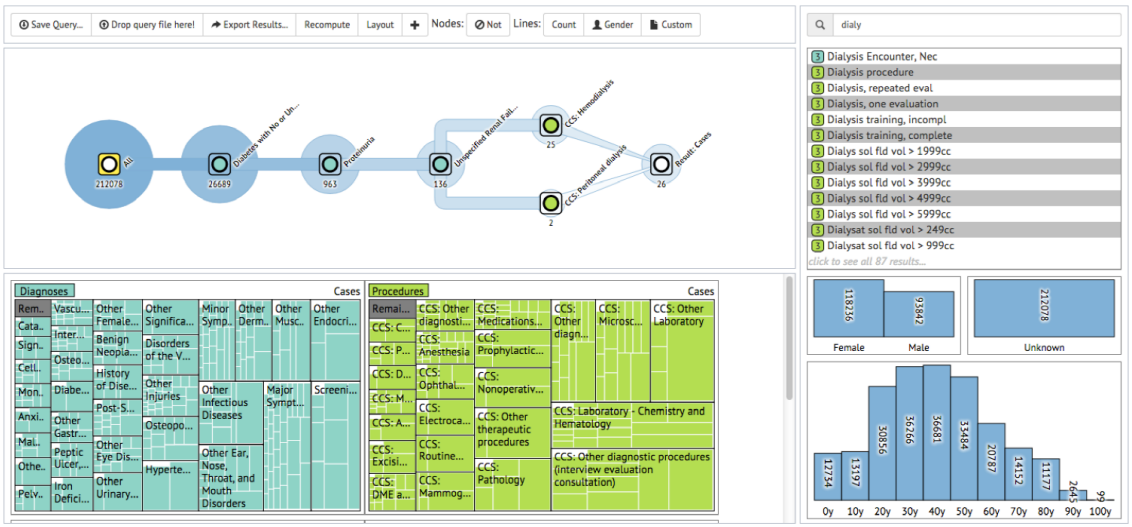


Figure 4.3: An Example of PVA work on Data Preprocessing (COQUITO [4])

the feature values. Other systems, such as iVisClustering [64] and iVisClassifier [2], mention the data encoding process (i.e., given a text document, iVisClustering encodes the document set as a term-document matrix using a bag-of-words model with stemming and stop words removal) but offer no visual analytics support in preprocessing.

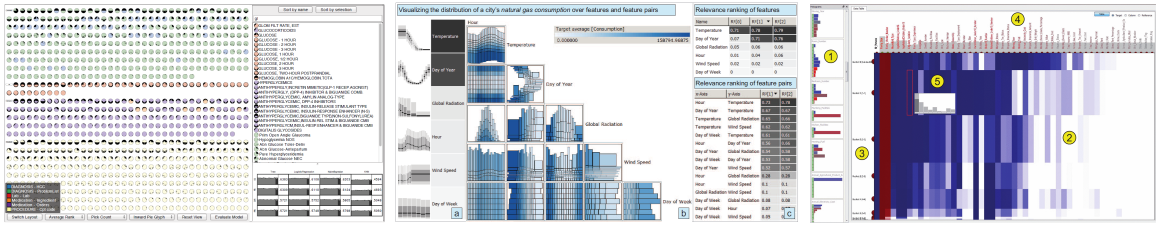
What is found is that in predictive visual analytics, the data preprocessing step is commonly removed from the main analytic workflow. This is likely due to the time-consuming nature of data cleaning, and the fact that specific visualizations and interactions may be used for data preprocessing but are unlikely to be returned to during analysis. As such, visual analytics systems that focus solely on supporting the preprocessing step (e.g., Wrangler [94]) are often preferred to implement data preprocessing, and researchers should consider how to better integrate these tools to support a full predictive analytics pipeline.

4.3 Feature Engineering

Once data is ready for analysis, the second phase of the PVA pipeline is feature engineering. Feature engineering covers both feature generation and feature selection techniques, and has become a key focus in many visual analytics systems (e.g., DimStiller [95], rank-by-feature framework [96]) due to the complexity of feature engineering in large, high-dimensional datasets. A recent survey by Sacha et al. [97] further documents the role of visualization in feature engineering, specifically dimensionality reduction.

In PVA systems, feature selection has been supported by parallel coordinates [62, 98], scatter plots [99], and matrix views [100]. For example, INFUSE [5] (Figure 4.4a) supports feature selection by comparing different measures in classification. INFUSE proposes a visual feature glyph which displays the performance of the feature during cross-validation, and users can select a feature subset for classification modeling. Mühlbacher et al. [6] proposed a visual analytics system for segmented linear regression which supports feature selection and segmentation on single features as well as pairwise feature interactions (Figure 4.4b). SmartStripes [7] helps experts identify the most useful subset of features by enabling the investigation of dependencies and interdependencies between different feature and entity subsets (Figure 4.4c).

Other relevant works have focused on feature space exploration coupled with visual interfaces to adjust feature metrics. For example, Guo et al. [101] developed a visual analytics system to support the exploration of local linear relationships among features in multivariate datasets. Dis-Function [65] allows the user to interact directly with a visual representation of the data to define an appropriate distance function. Dis-Function projects data points into a 2D scatter plot. The user may drag points to the region of the scatterplot that they consider more appropriate, and the system



(a) (b) (c)

Figure 4.4: Examples of Feature Engineering in the PVA pipeline: (a) INFUSE [5], (b)Segmented Linear Regression [6], and (c) Work by May et al. [7]

adjusts the distance function accordingly. The weights of the learned distance function provide a human-interpretable measure of feature importance. Krause et al. [102] proposed an interactive partial dependence diagnostics for users to understand how features affect the classification and understand the local effect with detail inspections. "What if" questions are also supported and users can change feature values to explore possible outcomes and this is useful for design medical treatment.

In surveying PVA papers, what is noted is that data transformation is often treated as a second class citizen even though many predictive analytics algorithms require data inputs to have certain statistical distribution properties [103]. Instead, the majority of techniques focus on dimension reduction, reconstruction (e.g., [104]), and feature space exploration. Currently, few systems provide support for feature generation. Examples of systems that support feature generation include FeatureInsight [105], which supports building new dictionary features for a binary text classification problem through visualizing summaries of errors and sets of errors, and Prospect [106], which uses a scatterplot and confusion matrix to visualize model performance and the agreement of multiple models so that the user can remove label noise and select models as well as generate new features to differentiate samples.

4.4 Modeling

Once features are selected by the analyst, the analyst enters the modeling stage of the PVA pipeline. In this stage, machine learning and statistical models are typically applied to the data. The underlying goal is to fit a representation onto known data to predict unknown data. It is observed that PVA methods often focus on applying a specific type of modeling process to the data (e.g., decision trees, support vector machines, hierarchical clustering, linear regression). From this survey analysis, PVA tools use three primary model types, *regression*, *classification*, and *clustering*. What this survey reveals is that model building is not usually separated from feature selection and result exploration, and interactions are often designed to support the iterative refinement of the model while exploring data space, feature space, and the results.

In predictive visual analytics, regression modeling has been used for box office prediction [62], epidemic diffusion analysis [9], and ocean forecasts [107]. In these systems, visual analytics methods have focused on data subspace exploration, training set modification, outlier removal, model parameter tuning, and modeling with different targets and different optimization functions. For example, Guo et al. [101] present a visual analytics system that helps analysts discover linear patterns and extract subsets of data following the patterns. They integrate automatic linear trend discovery and the interactive exploration of the multidimensional attribute space to support model refinement and data subset selection. Other work includes Mühlbacher et al. [6] which developed a system to support segmented linear regression model building. This system supports feature selection and local model exploration.

Clustering is a common prediction task in many applications where labeled data is unavailable. Clustering challenges include choosing an appropriate similarity metric

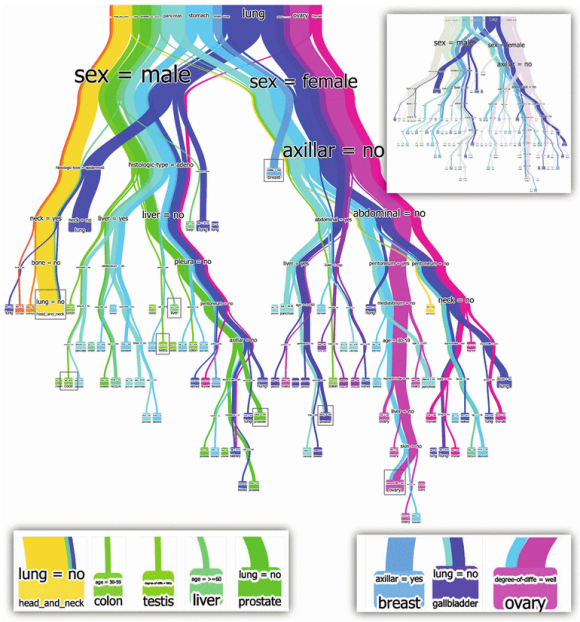


Figure 4.5: An Example of Modeling with Visual Analytics (BaobabView [8])

and validation due to the fact that models generated by clustering may not generalize. In clustering, visual analytics has been used for clustering manipulation, exploration, and evaluation. For example, ClusterSculptor [108] utilizes a fine-grained bottom-up pre-clustering to reduce the data size and allows the user to apply a top-down clustering strategy with visual regroup. Rules can be learned from the clustering built by the user and used on more data points. Following a similar strategy, Andrienko et al. [109] propose a visual analytics clustering method which starts by clustering a small set of data and then assigning new data to existing clusters. Clustering results are presented to the users, and users are able to interactively define new clusters and revise results to boost performance. Scatter/Gather Clustering [3] allows the users to set indirect constraints on the number of clusters, and the system will perform scatter (changing from N clusters to $N+1$ clusters) or gather (changing from N clusters to $N-1$ clusters) iterations to update the clustering result. Dis-Function [65] allows

the user to move data points on a 2D projected view and the system will learn and update its underlying distance function in the clustering method.

For classification, visual analytics has been extensively used to support active and incremental learning models where users interactively label a subset of the data to train the model. Heimerl et al. [1] presented a user-driven method that incorporates active learning for document classification. Visual cues are overlaid on unlabeled text documents representing their distance to the classification boundary and users can decide which data to include for the next model iteration. Users can also manually assign the class label as part of the incremental modeling approach, and the system will highlight the data points whose prediction will change based on this update. Paiva et al. [110] used a Neighbor Joining tree and a similarity layout view for interpreting misclassified instances to support the labeling and training set selection as part of an incremental learning procedure. This thesis finds that this concept of interactively labeling data as part of a model learning process is also supported by other PVA works [59, 104, 111].

Another focus of classification modeling in predictive visual analytics is decision tree construction. Work here has been demonstrated to improve both model accuracy and model comprehensibility. For example, Ankerst et al. [63] presented a visual classification approach on decision tree construction where circle segments are used to visualize the data attributes and clustering results. Users can manually select features, split nodes, and change data labels while constructing the tree. Backtracking in the tree construction phase is also supported. More recent work includes BaobabView [8] (Figure 4.5) which supports manual decision tree construction through visual analytics. BaobabView enables the model developer to grow the tree, prune branches, split and merge nodes, and tune parameters as part of the tree construction process. Additionally, neural networks and support vector machines have also been incorpo-

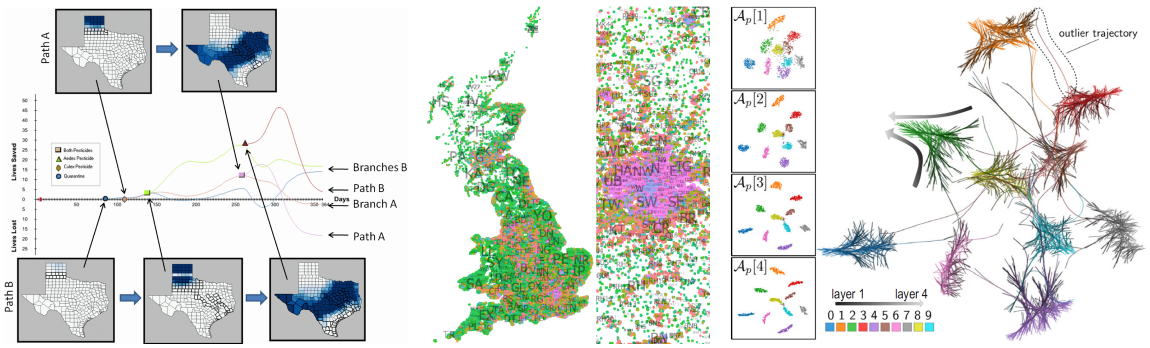
rated into predictive visual analytics works where the focus is on enabling users to understand the black-box modeling process of these algorithms [51, 112, 113].

What is found in the modeling stage of the PVA pipeline is that a major focus is on both model configuration and model comprehensibility. Currently, some of the most popular classification algorithms are inherently black-box in nature, which has led to researchers asking questions about how and why certain algorithms come to their conclusions. Challenges here include how much of the model should be open and configurable to the user and what the best sets of views and interactions are for supporting modeling. Again, one can observe a relatively tight coupling of this stage in the PVA pipeline with the feature engineering stage. This is likely due to the iterative nature of the knowledge foraging process [49].

4.5 Result Exploration and Model Selection

Once a model is generated, the next step in the PVA pipeline is to explore the results and compare the performance among several model candidates (if more than one model is generated). In this step, scatterplots, line charts, and other diagnostic statistical graphics are often the primary means of visualization, and many variations of these statistical graphics have been proposed, e.g., the line chart with confidence ranges and future projections [114], node-link layout for hierarchical clustering results [115], etc. In this phase, systems tend to support connect interactions to highlight and link relationships to explore and compare the outputs of the modeling process under different feature inputs.

Examples of result exploration in PVA include Afzal et al. [9] which presents a decision history tree view to analyze disease mitigation measures. Users can analyze the future course of epidemic outbreaks and evaluate potential mitigation strategies by flexibly exploring the simulation results and analyzing the local effects in the map



(a) (b) (c)

Figure 4.6: Examples of Result Exploration for Predictive Analytics: (a) A Decision History Tree View by Afzal et al. [9], (b) Uncertainty Visualization in Area Classification Results [10], and (c)Similarities Between Artificial Neurons [11]

view (Figure 4.6a). Different paths can be displayed revealing prediction outcomes under different settings by deploying selected strategies. In this way, the user can explore model results and decide which strategy to use while comparing multiple cases. Slingsby et al. [10] present geodemographic classification results using a map view, parallel coordinates, and a hierarchical rectangular cartogram. The parallel coordinates view is used to drive the Output Area Classification model and compare classification results given different parameterizations (Figure 4.6b). Rauber et al. [11] use dimension reduction techniques to project data instances and neurons in multilayer perceptrons and convolutional neural networks to present both the classification results and the relationships between artificial neurons (Figure 4.6c), and Dendrogramix [116] interactively visualizes clustering results and data patterns from accumulated hierarchical clustering (AHC) by combining a dendrogram and similarity matrix. iVisClustering [64] visualizes the output of Latent Dirichlet allocation by displaying cluster relationships based on keyword similarity in a node-link cluster tree view. Users can explore the model results, and interactions support model refinement.

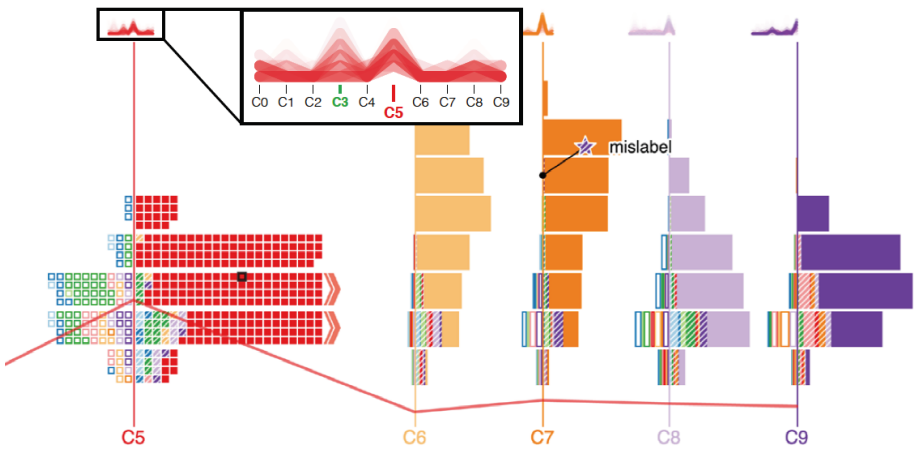


Figure 4.7: An Example of Model Selection (Squares [12])

Alsallakh et al. propose the confusion wheel [117] and visualize true positive, false positive, false negative, and false positive classification results.

Along with result exploration, model selection methods have been employed to compare prediction results and model quality under different parameterizations. For example, Squares [12] is a visual analytics technique to visualize the results of multiclass classification on both instance level and class level performance (Figure 4.7). Squares uses small multiples to support the analysis of the probability distribution of each class label in a classification, and it uses squares to visualize the prediction result and error type for each instance. Pilhöfer et al. [118] use Bertin's Classification Criterion to optimize the display order of clustering results from different models so that the relationship between results can be explored. Other techniques have explored methods for visually comparing clustering results under different parameters [58, 119, 120] in geographical displays.

From this survey, one can observe that there are many PVA works supporting result exploration and model selection. However, one undersupported topic is model

comparison, i.e., comparing the results of two different classifiers such as a decision tree and a support vector machine. Furthermore, it is noted that many systems also have a distinct lack of provenance and history support. Result exploration often gets tied into the feature engineering process as many systems have been developed for feature steering and selection. As features are modified, results from the model are presented. Without the ability to save results, however, comparison can be difficult even within a model.

4.6 Validation

Finally, once a model is generated and the results are explored, model validation is performed to test its quality. After training, data that has not been used in the first four stages of the PVA pipeline can be used to evaluate the performance of the model. This step is critical to verify that the model created can be applied to future or unknown data without a significant drop in accuracy. Statistical measures such as accuracy, precision, recall, mean square error (MSE), and $R^2_{predicted}$ are commonly used to evaluate model performance. Currently, the user's enjoyment measurement [105] is not considered as a part of the validation step in the PVA pipeline, but, arguably, such measures should be an integral part of a PVA system. Similarly, efficiency and scalability are also not considered. As such, validation in PVA has two forms. First, visual analytics can be employed as part of the statistical model validation. Second, validation with respect to the use of throughput and efficiency of the PVA system should also be considered for validating the proposed PVA approach.

Common visualizations used in machine learning and data mining for model validation include residual plots, the receiver operating characteristic (ROC) curve, and the auto-correlation function (ACF) plot. Hao et al. [93] present a visual analytics approach for peak-preserving predictions where they visualize the certainty of the

model on future data using a line chart with a certainty band. This work explores model accuracy on the training time series data with color codes, and then discusses the model accuracy as an offline comparison with the true value of the test data. K-fold cross-validation is involved in INFUSE [5] for feature selection, and Andrienko et al. [109] apply the classifier to a new set of large-scale trajectories and calculate the mean distance of the class members to the prototype for validation after developing a classifier using their PVA system.

While validation in the context of predictive analytics would refer to questions of the model's accuracy, predictive visual analytics must also be concerned with validating the visualizations and interactions proposed. This is often done through case studies. An example of this was the 2013 VAST Challenge on box office predictions [121, 122] where participating teams submitted predictions of future ticket receipts and ratings of upcoming movies using their visual analytics system (over the course of 23 weeks). The performance of these tools has been reported in follow-up papers [61, 123] and provides insights into the current design space of PVA. Other works have validated their PVA systems by including statistical tests for the models generated using their PVA systems compared to other approaches. For example, BaobabView [8] compares its classification accuracy to the automatic implementation of C4.5 [124] on an evaluation set. Heimer et al. [1] separate training and test data and provide a detailed performance comparison of the three models they discussed to illustrate that the user-driven classification model outperforms others. Similar examples can be found in works from Ankerst et al. [63], Kapoor et al. [125], and Seifert and Granitzer [126].

What is observed in the survey is that validation is perhaps the most underserved stage in the PVA pipeline, both from the statistical and user point of view. In many PVA systems, the user is allowed to interact until the model outputs match their

expectation; however, such a process may be dangerous as it allows the user to inject their own biases into the process. More research should be done exploring to what extent humans should be involved in the predictive analytics loop. This requires validation on the user side and methods for measuring a user's model comprehension. Insight generation should also be considered alongside measures of the predictive accuracy of the model.

Chapter 5

A PREDICTIVE VISUAL ANALYTICS FRAMEWORK

Following the proposed predictive visual analytics pipeline, a predictive visual analytics framework consisting of several visual analytics tools is proposed in this thesis and one particular real world prediction problem, movie box-office prediction, is used to show the effectiveness of the framework. Movie box-office prediction, as one predictive analytics problem, is selected in this thesis for several reasons: First, the data is public and easy to collect. In this thesis, movies released from 2013 to 2016 are tracked and both IMDB movie data and Twitter data are collected. The true value of movie revenue is also public and can be used to evaluate the prediction results. Second, it is easy to find users with adequate knowledge of movies so that they can be considered experts in the analysis. Experts have domain knowledge in making predictions and predictive visual analytics can assist them to use such knowledge to adjust the model's prediction. Recruiting users familiar with movies, it is able to support the integration of domain knowledge and machine learning and evaluate human participation in such a predictive analytics process. To focus on the research problems proposed in this thesis, some visual analytics tools in this framework are specifically developed to support visual analytics on social media data.

5.1 Predictive Analytics for Movie Revenue

In the United States, social media platforms are being used by two out of three adults and trends increasingly [127]. People use social media to read news, post events, express opinions and share experiences. Because of the nature of such use, social media data conveys information for business analysis and has caught attention

in many marketing research topics, such as word-of-mouth [128], relational marketing [129], and products adoption and diffusion [130].

As a representative business analysis problem, movie revenue prediction has been studied widely and many social media features have been explored for opening weekend's box office prediction before release. An early study by Simonoff et al. [131] leveraged classical numerical and categorical movie features (e.g., time of year, genre, MPAA rating, budget) and proposed a logged response regression model. Although this work demonstrated that the first weekend gross is an effective predictor for the total revenue of a movie, the prediction accuracy using data before release was lower than 45%. To improve the prediction accuracy, researchers tried many social media features and the power of social media data in box office prediction emerged quickly. Zhang et al. [132] demonstrated that pre-release prediction of movie gross can be enhanced by utilizing metrics extracted from news sources in both regression models and k-nearest neighbor models. Focusing more on text features, Joshi et al. [133] explored the relationship between film critic reviews and box office performance. Turning from news and reviews to less structured and more enriched social media posts, further work by Asur et al. [134] found that the rate of Tweets per day could explain nearly 80% of the variance in movie revenue prediction. The predictive power of search volume in different websites has been demonstrated [135], and a white-paper report from Google [136] claimed a 94% prediction accuracy in box office prediction by utilizing the volume of internet trailer searches for a given movie title. While many factors have been analyzed, the integration of both classical factors and social factors has been suggested for better predictions [137].

In this thesis, similar to previous works, classical movie features and social media features are both used in the movie box-office prediction models, but different to previous works, the emphasis here is on integrating experts' domain knowledge and

machine learning models through a predictive visual analytics framework to support the predictive analytics.

5.2 Predictive Visual Analytics Toolkit

In prediction tasks, the visual analytics tools provide more comprehensive information and allow users to interactively steer their exploration and bring in their own domain knowledge. For a specific movie box-office prediction problem, the design contains the following parts: single movie prediction, prediction for the total revenue of the movies released at the weekend, feature exploration and selection, similar movie's performance comparison, sentiment analysis, and an interactive modeling. The development of this toolkit started as part of the 2013 VAST Challenge, and the environment of using some of the tools and models is also related to the contest.

5.2.1 Linear Regression for Movie Predictions

Traditional variables used in box office prediction models include structured variables (e.g., MPAA rating,movie budget) and derived measures (e.g., popularity of the movie stars, popular sentiment regarding the movie). Based on an initial literature search, multiple linear regression is chosen to produce an initial prediction range for the opening weekend box-office revenue. As in a contest where traditional variables used by other researchers were not always available (for example, theater count is not provided for every movie in IMDB), a variety of different variables that could be mined from the social media data and IMDB were explored, see Table 5.1. After initial model fitting and evaluation using R [138], the best fit of this linear regression model is found to be of the form:

$$OW = \beta_0 + \beta_1 TBD + \beta_2 Budget + \varepsilon \tag{5.1}$$

Table 5.1: Variable Description

Variable	Description
OW	3-day Opening Weekend Gross
Budget	Approximate movie budget from IMDB. (unit is "million" of dollars)
Genre(category)	The movie's genre(s) according to IMDB
TUser	Number of unique users who tweeted about a movie
TBD	The average daily number of Tweets over the 2 weeks prior to release
TSS	Tweet Sentiment Score - A summation of each individual word's sentiment polarity as calculated via SentiWordNet [139]
MSS	Movie Sentiment Score - A derivation of the overall sentiment of a movie
MSP	Movie Star Power - A summation of the Twitter followers of the three highest billed movie stars (as listed by IMDB)

This model is used to predict upcoming movies' opening weekend gross and the model is updated weekly as new movies are entered into the dataset. Parameters are fit using movie data beginning in January, 2013. The first prediction was for the May 17th weekend and used data from 39 movies for training. This weekly model reported an $R\text{-}adj^2 \approx 0.60$ with $p < .05$. As more movies came in, this model became more stable. The final parameters were $\beta_0 \approx 4.9 \times 10^3$, $\beta_1 \approx 4462$, and $\beta_2 \approx 2.3 \times 10^5$.

The drawback of this model is that it does not fit the data overly well and its predictions have a large variance. The hypothesis was that a visual analytics toolkit

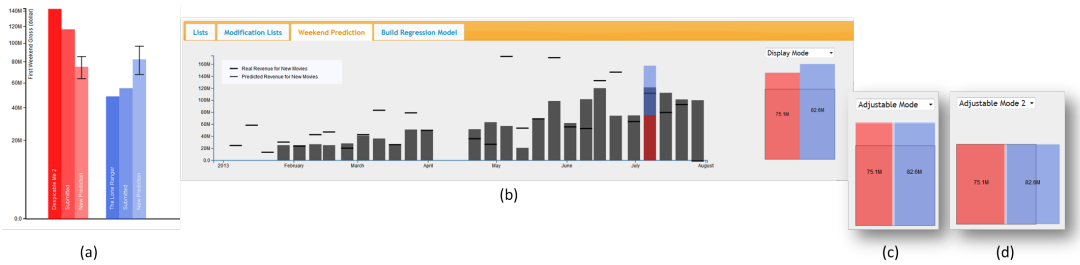


Figure 5.1: The Weekend Prediction View for Newly Released Movies and the Prediction Adjustment Widget

could partially enable analysts to overcome poor data (partially due to the noise in social media data and partially due to the closed world nature of the contest). In order to facilitate better model prediction, a simple bar graph view (Figure 5.1(a)) was created which, for historical movies, showed the model prediction and its 95% confidence interval error range, the submitted prediction, and the actual box office gross. For newly released movies, only the model prediction and user submission was shown. This view was critical in the analysis process, and the primary view into the data consists of an overview of the tweets per day and the model predictions of the movies under analysis as shown in Figure 5.4(a).

5.2.2 Temporal Modeling for Weekend Prediction

While the regression model is able to provide one point for analysis, the goal was to also provide a big picture overview. For any given weekend, there is likely a maximum amount of money available in the market. In order to approximate this value, a simple moving average model was employed. This model approximated subsequent weekend grosses for movies under the assumption that movies would run for three weeks following their opening weekend, and each weekend their box office

take would be reduced by 50%. Thus, for any given weekend, it approximated the gross as:

$$\textit{Weekend Gross}(t) = \sum_{\forall_i} \mathrm{OW}_i(t) + \sum_{\forall_i, j=1}^{j=3} .5^j \mathrm{OW}_i(t-j),$$

where t is the current weekend and i is the index to a movie that exists at time t. Then, the weekend gross prediction uses a moving average:

$$\textit{Weekend Gross}(t+1) = \tfrac{1}{3} \sum_{j=0}^{j=2} \textit{Weekend Gross}(t-j).$$

Finally, it approximates the available revenue for new movies as:

$$\textit{New Movie Gross}(t+1) = \textit{Weekend Gross}(t+1) - \sum_{\forall_i, j=1}^{j=3} .5^j \mathrm{OW}_i(t+1-j).$$

While this prediction is crude, it provided the analysts with a valuable bound in which to explore the revenue predictions.

Results from the temporal weekend prediction and the linear regression models were then visualized in two different views as shown in Figure 5.1. The first view consists of a linked bar graph combined with stacked bars as shown in Figure 5.1 (b). The primary portion of the bar graph consists of light gray bars indicating the predicted total weekend market for the new movies and the dark gray short line indicates the actual weekend market for each calendar week whose date is shown on the x-axis. The stacked color bar graph is visualized only for the weekend under analysis, and the color design is the same as the movie's color in the prediction bar graph.

The second view, Figure 5.1 (c) and (d), is used to enable users to interactively adjust predictions while also visualizing the bounds of the total weekend prediction. In this view, a gray square is drawn, the area of which is scaled linearly to the total weekend prediction. Colored rectangles are superimposed onto the gray square, where the area of each colored rectangle represents the linear regression prediction for each

movie being released on that weekend. If the sum of the individual predictions is equal to the total prediction, the colored rectangles will fit exactly into the gray square in both Figure 5.1 (c) and Figure 5.1 (d). The color design is the same as those of the bar graph, and modifying the size of a bar in any view will modify the size across all views.

The system was designed to allow for three types of prediction adjustments.

1) Users are allowed to change the amount of the total gross prediction but the ratio between the movies will remain consistent.

2) Users are allowed to change the amount of an individual prediction but the total weekend prediction is kept consistent.

3) Users are allowed to arbitrarily change each movie's prediction and ignore the weekend gross.

The first two adjustment functions are directly implemented in view (b). The white translucent lines on the boundary and the top of the colored rectangles are control bars used to modify the gross. The vertical control bar can be dragged from left or right to change the predicted gross of adjacent movies without affecting the total prediction. The top horizontal control bar can be moved up and down to change the total prediction while keeping the original ratio of each movie's take. In this view, every colored rectangle has the same height, and their width represents the ratio of their gross. View (c) provides a method for analysts to compare what proportion one movie takes of the predicted total weekend gross and adjust each movie individually without affecting others. In view (c), the height of each movie is equivalent to the height of gray square, and the width of the bars is allowed to extend out of the square. When one movie is aligned to the left in the gray square, the analysts can quickly observe the proportion of the movie to the total predicted weekend gross.

By implementing and integrating multiple comparison methods, the model analysis process found to be able to quickly bound user's analysis. While flexible, these bounds provided one with an early estimate of the total expected weekend gross in which to compare the predictions of the linear regression models. This multiple model comparison was a critical step for our overall box office prediction and was regularly used for all movie analyses.

While the results of our temporal predictions were of low quality, the combination of predictions and bounding of the problem space provided critical information for comparison and analysis. This thesis will further discuss, in the case study section, how the combination of both models was critical for successful predictions. Overall, the addition of multiple models predicting similar information can help guide analysts to a better ground truth. Similar to principles employed in the Delphi method [140], where predictions are solicited from multiple experts and used to come to a common conclusion, this system allows users to solicit predictions from multiple models to aid in their analysis. This bounded adjustment widget can be used in other hierarchical predictions which have both individual and total predictions, such as sub-topic trend prediction in a time period.

5.2.3 Feature Analysis and Selection

Feature values of movies can give insights and hints about their box office success. Moreover, they can be used as predictors for a movie's opening weekend revenue. Using Twitter, Youtube and IMDB data sources, four groups of features are extracted for model building with 119 features listed in the Feature Selection Table (Figure 5.2). Given the large number of features, it is necessary to provide the users with a suitable starting point for analysis. As such, known predictive features for movie analysis from previous work [131] (e.g., budget, number of screens the movie opens on, etc.) are

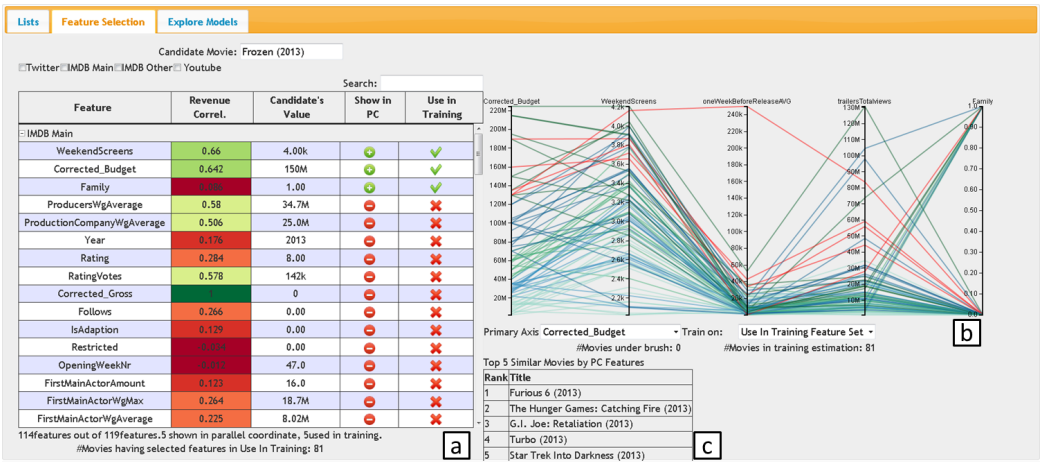


Figure 5.2: Feature Selection Page with Frozen as an Example. View (a) is the Feature Selection Table, View (b) is a Parallel Coordinate Plot, and View (c) is the Most Similar Movies List.

utilized. Thus, when the users begin their exploration process, they are presented with a baseline model to compare against. Other options would include integrating automatic feature selection as an entry point for analysis (e.g., [141, 142]).

The goal was to augment model building by adding tools for a user to modify and explore various features. In order to quickly enable this exploration, the Feature Selection Table (Figure 5.2(a)) utilizes a variety of interactions and visual overlays. First, for the candidate movie being predicted (in this case Frozen), features which are not available are grayed out. Second, each of the columns in the feature selection table provides the details of a movie. The first three columns include information on the feature's name, the correlation to the revenue, and the candidate movie's value. These columns can be automatically sorted from high to low or low to high simply by clicking on the column header. The Revenue Correlation column is also color coded to directly highlight correlated features. A myriad of work has been done in feature selection [7, 143, 144] and correlation is traditionally used as one of the major

factors in feature selection. A high correlation of a feature to the response variable (in this case the movie revenue) indicates that this feature could greatly impact the model. A green to red divergent color scale [145] is used to represent the correlation value where green represents a high absolute value of correlation and red represents a low value of correlation, with .5 being the midpoint value. Although correlation here is univariate (meaning it does not show correlations between multiple features) and non-linear dependencies are not taken into account, it still provides important information to users for feature detection and analysis.

The final two columns in the Feature Selection Table are associated with the parallel coordinate plot and the model training data selection. The "Show in PC" column, when selected, will add that feature as an axis of the parallel coordinate plot. The "Use in Training" column, when selected, will add all data elements that contain all of the features selected into the training set. To quickly see what features have been selected, the analyst can sort the features by clicking the column header. When features are selected, the footer information about the Feature Selection Table will update and tell the user how many features have been added to the training set, as well as the amount of movies that exist having all of these features. In this manner, the analyst can determine how many data elements can be used to train a model and they can quickly make decisions about the tradeoff between the use of more features or more training samples. For example, if a user chooses to select a Twitter feature, only 112 movies in this data set have associated Twitter data. Thus, the number of elements in the training set decreases. However, Twitter data may have a high correlation to the opening weekend gross. As such, the analyst can actually build multiple models with multiple features for training and analysis.

Another way to select the training data is through the interaction with the parallel coordinate plot view. Consider the case in which a user has sorted the features

by correlation to revenue, selected some features with higher correlation to the gross, and selected features that he/she suspects are important. These selected features can now be further explored in the PCP view (Figure 5.2(b)) by simply activating the "Show in PC" cell in the corresponding table row. Referring to the candidate movie's value, shown in the fourth column, the user can further filter out movies far away from this value in the PCP view. Figure 5.2(b) shows features of the movie "Frozen" with highly correlated features in different group and the movie's genre, "Family". Pairwise correlations between features are explored in the PCP view. For example, the WeekendScreens (the number of screens in which a movie was released during its opening weekend) and the oneWeekBeforeReleeaseAVG (the daily average number of Tweets that are related to a movie one week before its opening) variables are correlated. These axes can be dragged and dropped to explore more pairwise dimension correlations so that an analyst can choose features with low multi-correlation in order to improve the model performance. Users can then interactively select ranges by brushing on each axis to filter the data and can select an option to train the model using only the selected data.

The PCP view can also be used to generate insight into the data. For example, by brushing and selecting only Family movies using the Boolean genre feature "Family," one can define the training set to be only those movies that are considered to be "Family" movies. Moreover, the PCP view allows the analyst to select a primary axis, this selection defines the feature on which it bases the PCP line color scheme. For example, if one colors the lines based on the genre axis "Family" it can be seen that family movies rarely obtain a very high gross. From there, the user could train the model for only Family movies or could look for genre crossover movies such as Family and Animation.

The final item in the Feature Analysis and Selection widget is the "Top 5 Similar Movies by PCP Features" view, Figure 5.2(c). Given the feature vector corresponding to the features selected in the parallel coordinate plot, the system automatically calculates a Euclidean distance metric between the candidate movie and all other movies that appear in the PCP view. The five movies with the smallest Euclidean distance are then summarized in a tabular view.

5.2.4 Similarity Widget

While the Feature Analysis and Selection Tools show the top 5 most similar movies, a series of tools have also been developed for enabling users to explore temporal and sentiment similarities with regards to social media trends and specific feature similarities such as genre and ratings. Figure 5.3 shows the similarity widget page. Items in this similarity view focus primarily on similarity across social media (as opposed to the previous widget which used a Euclidean distance metric across many features, this view is a pairwise feature similarity). The left side of Figure 5.3 shows the various similarity options provided while the center view displays line charts or wordles depending on the selection. It has ten predefined metrics (eight of them are described in Table 5.2 and one "Make Your Own Similarity" option. The rightmost area shows the model predictions and the actual weekend gross for similar movies via a bar graph.

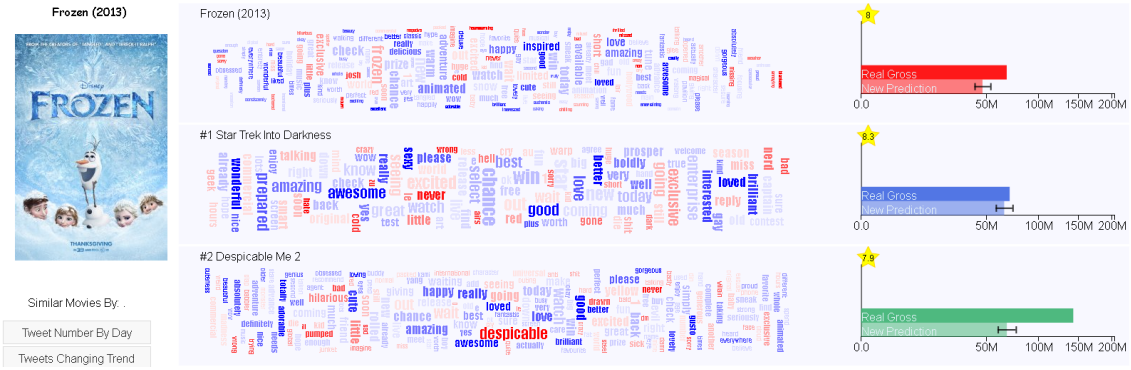
This widget enables analysts to quickly find and compare the accuracy of predictions based on various criteria of similarity, and to perceive if the given prediction model typically underestimates, overestimates or is relatively accurate with regards to movies that the analyst deems to be similar. In this manner, a user can further refine their final prediction value. Ten similarity criteria are defined with distance calculation methods focusing on matching temporal trends through sequential nor-

Table 5.2: Calculations of Similarity Criteria

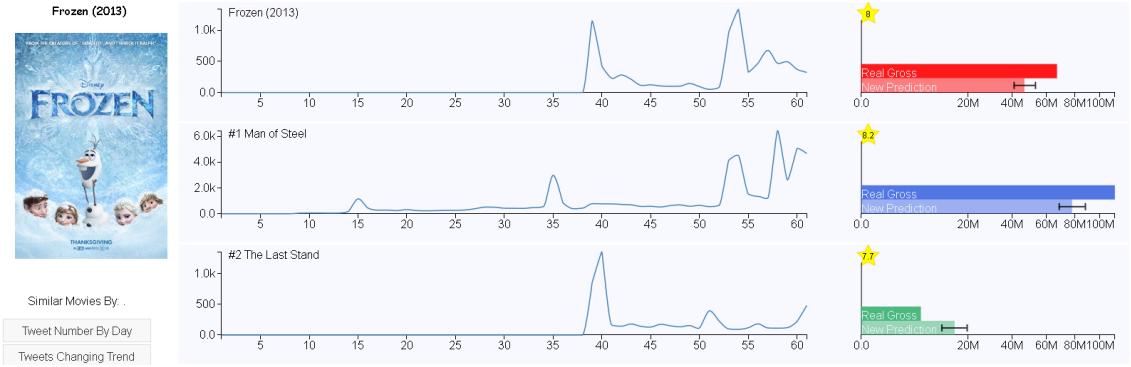
Similarity Criteria	Distance Measurement
Tweet Number by Day	$Dis(v,s) = \sum_{i=1}^{14} \lvert TBD_i(v) - TBD_i(s) \rvert$
Tweet Changing Trend	$Dis(v,s) = \sum_{i=1}^{14} \left\lvert \frac{TBD\_i(v)}{Max(TBD\_j(v),j=1,2,...14)} - \frac{TBD\_i(s)}{Max(TBD\_j(s),j=1,2,...14)} \right\rvert$
Sentiment River	$Dis(v,s) = \sum_{i=1}^{14} \left\lvert \frac{MSS\_i(v)}{Max(MSS\_j(v),j=1,2,...14)} - \frac{MSS\_i(s)}{Max(MSS\_j(s),j=1,2,...14)} \right\rvert$
MSS	$Dis(v,s) = \lvert MSS(v) - MSS(s) \rvert$
MPAA	same MPAA rating and close release date
Genre	$Dis(v,s) = 1 - \frac{card(Genre(v) \cap Genre(s)) \times 2}{card(Genre(v)) + card(Genre(s))}$
MSP	$Dis(v,s) = \lvert MSP(v) - MSP(s) \rvert$
Sentiment Wordle	$Dis(v,s) = 1 - \frac{card(SWordle(v) \cap SWordle(s))}{card(SWordle(v))}$

malization or Euclidean distance metrics for magnitude comparisons. In all similarity matches, this view shows the top five most similar movies. These views allow users to directly compare Tweet trends and sentiment words between movies deemed to be similar in a category. Figure 5.3 contains snapshots from Frozen's similarity page cropped to the top two most similar movies by Sentiment Wordle and Youtube Trailer Comments.

Though similarity metrics used in this page are not directly transformed into modeling features, by providing an analyst with insight into these secondary variables, coupled with the model performance with similar movies included in the training set, further refinement of the prediction is made possible. For example, an analyst may compare the absolute difference between Tweets/Youtube comments of two movies, or they can inspect the trend of the Tweets through line chart comparison using



(a) Similarity by Sentiment Wordle



(b) Similarity by Youtube Trailer Trend

Figure 5.3: Similarity Widget View with Frozen as an Example.

the Tweets Changing Trend similarity metric. This tool also allows users to quickly compare the current movies under analysis to recently released movies with the same MPAA rating and genre. When the user builds a model involving Twitter features, the top 5 most similar movies listed in the Feature Selection and the Explore Models page can be compared in the similarity page.

5.2.5 Tweet Sentiment Visual Analytics

Sentiment embedded in social media has great value but is also hard to perceive effectively. In sentiment visual anlayitcs, the goal is to visually represent the sentiment

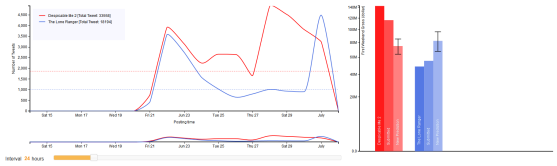
conveyed in social media data and support interactive analysis in the context of sentiment classification problems on social media text.

While structured data is relatively straightforward to extract, unstructured data requires a large amount of pre-processing and manipulation. In a box-office prediction scenario, tweets were collected for the two-week period prior to the release date based off the hashtag provided by a movie's official Twitter account. In order to approximate the popular sentiment of a movie, each tweet is processed using a dictionary based classifier, SentiWordNet [139]. This process assigns each word in the tweet with a score from -1 to 1 with -1 being the highest negative sentiment score and 1 being the highest positive sentiment score. Next, each tweet is assigned a sentiment score by summing the sentiment score of all words in the tweet and scaling the range from $-.5$ to .5 (TSS in Table 1). Finally, the movie sentiment score (MSS in Table 1) is calculated as

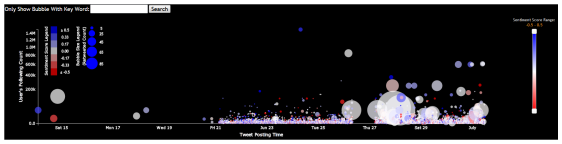
$$MSS = \frac{Positive\ Score}{Positive\ Score + Negative\ Score} \tag{5.2}$$

where *Positive Score* is the sum of all tweets for a given movie with a TSS greater than zero and *Negative Score* is the absolute value of the sum of all tweets for a given movie with a TSS less than zero.

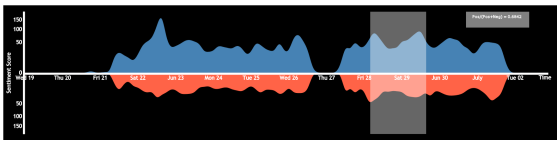
Once the sentiment scores for tweets were extracted, these values were then visualized to the end user. Figure 5.4 (b-d) shows the bubble plot view, the sentiment river view, and the sentiment wordle view. In the sentiment wordle view (Figure 5.4 (d)), the 200 most frequently mentioned words are extracted and visualized. Both the bubble plot and the wordle plot enabled interactive searching and filtering by keywords and users. Users posting irrelevant messages could be removed from the tweet count and mismatched sentiment could be modified by the end user. One example usage is that the user can first mouse over a bubble to see the user name and the text content of this tweet and then a left click opens a pop-up dialog for sentiment



(a) Tweet Line Chart



(b) Bubble Plot



(c) Sentiment River

(d) Sentiment Wordle

Figure 5.4: Tweet Trend and Sentiment Views for Despicable Me 2

modification while a right click allows tweet removal. The primary use found for the views in Figure 5.4 were for data cleaning. The primary lesson learned was that visualization tools are a necessity for data cleaning due to the noisiness of social media data and the problems inherent in sentiment matching using a sentiment dictionary (e.g., phrases such as "I want to see this movie so bad" are marked as negative due to the word "bad", and words such as "Despicable" give negative sentiment when they are merely references to a movie title). While the wordle view provided a quick way to assess the sentiment of popular words, it was necessary to hover over the bubble plot or open a tweet list view through the search bar in order to fully explore the context of a tweet. The bubble plot and wordle plot helped users to deal with the challenge of sentiment analysis and cleaning of noise from social media data.

5.2.6 Interactive Model Building

Based on recent literature and the general use of prediction models, this framework supports the creation of three different types of models: Support Vector Machine (SVM) [22], Linear Regression (LIN) [23] and Multilayer Perceptron (MLP) [24]. A

Figure 5.5: Front Page of the Frozen Weekend with View (a) the Tweet and Youtube Comments Line Graph, View (b) the Opening Weekend Gross Bar Graph, and View (c) the List of Tweets and Users

linear regression model is used as a baseline model that the system provides users with its prediction result together with a 95% confidence interval for each movie. The baseline model results are shown in both the front page (see Figure 5.5(b)) and the similarity page's right-hand bar graphs (Figure 5.3).

Besides exploring the baseline model, the user can build more complex models, bringing in domain knowledge and analytic insights. For instance, the user is allowed to interactively set up parameters and build models with different feature sets, training instances (movies) and model types. Several error measures are used to give the analyst feedback about the quality of fit and the prediction stability. By using the interactive Feature Selection and Explore Models pages, the user can iteratively change the features, training sets and model types to improve a model's quality. The model's accuracy is measured using the adjusted R^2, denoted R^2_{adj}.

Base Line Model

The model proposed in section 5.2.1 is used as the base line model, which is described as follows:

$$OW = \beta_0 + \beta_1 TBD + \beta_2 Budget + \varepsilon \tag{5.3}$$

It uses the budget and the average number of daily Tweet (TBD)s for a movie as regressors and the opening weekend gross as response. With all 110 movies in the training set, the estimation of parameters in Equation 5.3 are $OW = 6.878 \times 10^6 + 1303 \times TBD + 0.26 \times Budget$ with $R^2_{adj} \approx 0.6$ and $P \ll 0.05$.

Advanced Models

As most of the attributes are proportional to the box office success (e.g. the more budget, the higher weekend gross potential) one can even achieve good results using linear regression model. More advanced models can be built using a Support Vector Machine (SVM) or a Neural Network, i.e. Multilayer Perceptron (MLP). To achieve good results, these algorithms have to be finely configured by setting input parameters based on the input data. A grid search (parameter optimization method) is run to find out the best parameter settings. The SVM here uses a linear kernel and a nu-parameter of 0.4, which constrains the influence of a single instance (movie) to the model. Considering the relatively small number of movies when compared to the large feature space we also tested an RBF kernel. However this did not achieve better R^2_{adj} results than with the linear kernel. The MLP here uses the backpropagation learning rule and use a learning rate of 0.3, 200 training epochs and a momentum rate of 0.85 to achieve good results.

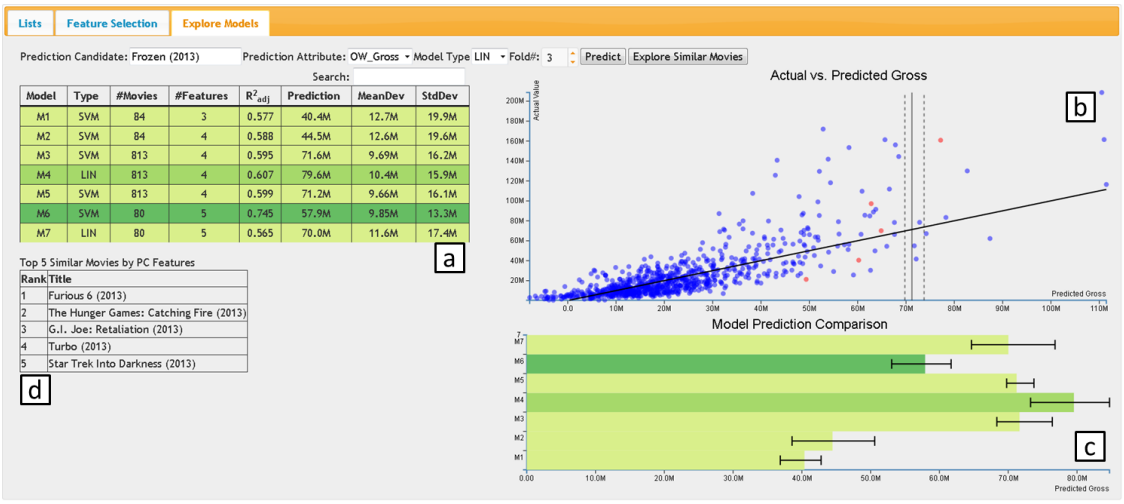


Figure 5.6: Multiple Method Modeling with Frozen as the Candidate Movie with View (a) a Model History Table, View (b) a Scatterplot, and View (c) a Model Prediction Comparison Plot

Multiple Methods Modeling

Predictive models help to reveal relationships between the predictors and the response variable, but no matter how good the prediction is, no cause-effect relationship can be implied. Also, the accuracy of one prediction can hardly be generalized to all other predictions. In statistical analysis, experts usually explore residual distributions, outliers, influential points, and model stability. In this system, besides using statistical methods, visual analysis methods are applied for exploring the residual distribution.

In the page "Explore Models" the user can select which algorithm to use, set the number of folds for the stability test, train models to predict the movie's revenue, and compare between models. The Explore Models view is shown in Figure 5.6. For model building, the feature and training set configurations from the Feature Selection page are applied. After the prediction is executed, the analyst can use the Actual vs. Predicted Gross view (Figure 5.6(b)) to obtain an overview of the residuals, as

was presented in [146]. A diagonal line indicating "the perfect prediction" is also drawn. This means, the closer the data points lie to the reference line, the better the overall fit of the model. The top 5 most similar movies are highlighted in red to quickly guide comparison and analysis. The user can change these similar movies based on adding/removing features in the parallel coordinates plot view (Feature Selection page). To submit a good prediction for a particular movie, it may be more important that the model fits for similar movies than fits the overall training set. In other words, if the model predicts well for similar movies this may be an indicator that it also gives good results for the prediction candidate.

These tools also enable the exploration of influential points. An influential point is an outlier in both the predictor and the response domain, and these points are known to have a noticeable impact on the model coefficients [23]. If an influential point is removed from the training set, the fit of the model will change by a relatively large degree and usually fit other points better. This fact can be used to improve prediction results. Instead of using statistic diagnostics, such as Cook's D and DFFITS [147], the user is allowed to directly remove such instances and only train on selected movies. In this way, influential points can be implicitly removed via exploring differences between different models.

Finally, the Model History Table (Figure 5.6(a)) enables the comparison of multiple models so that the analyst can review the predictions by re-investigating their scatterplots. In combination with the Model Comparison view (Figure 5.6(c)), the user can also get an overview of the prediction deviations, review the increase or decrease of prediction precision and select his/her final prediction. The goal is to build a model which can help the analyst to better predict the upcoming movie's opening weekend gross, not to build an adequate model that fits all the training data very well.

To estimate the performance and to test the model's stability, an n-fold cross-validation [148, 149] is provided. For the cross-validation we partition the data into n folds. Each fold includes num_{movies}/n instances. The movies of each fold are predicted once, using the other folds for training. This way, the method ensures that the model generalizes and is not overfit to the training data. For the prediction candidate, every fold is used once to predict the outcome. Thus, for each prediction the candidate movie gets n results. The dashed vertical line in the scatterplot shows the range of these results. A smaller range indicates that the model is stable. This range is also shown in the bar graph below the scatter plot, where all predictions can be compared.

Auxiliary Analysis

Instead of depending totally on an automatic model, most industry predictions also utilize an expert's domain knowledge. For example, if a movie is released next to an expected blockbuster, its performance could be also impacted. With the system, analysts can query any movie by its title to investigate features. Users can also go to previous weekends to see how much money those movies made. A user can also investigate the Twitter and Youtube data to explore the advertising campaign and public sentiment. Usually a successful movie has either an effective advertisement campaign, positive public reactions, or both. From the bubble plot shown in Figure 5.4(a), large bubbles usually are Tweets from the movie production company and the bubble size indicates the spread power. If the large bubbles separate along the time line, it is likely that the company has continued advertising its movie.

Chapter 6

CASE STUDIES

The proposed predictive visual analytics framework consists of several views and interactive analytics tools. To show its effectiveness in movie box-office prediction, some tools have been used in the 2013 VAST Box-office Challenge to predict movie's opening weekend gross weekly and the results were compared with peer teams. Adding other features in the framework, a predictive analytics procedure has been proposed after the contest and a small-scale user study was run to evaluate the framework with a larger dataset. This section will discuss these two case studies on applying the developed visual analytic framework on movie box-office predictions. The results of these studies inspired a further study on the visual properties of the social media data, which has been demonstrated to be useful in the case study. Therefore, the network structure of the movie related Tweets is explored in this thesis as the third case study to analyze its role in predicting movie box-office.

6.1 VAST Box Office Challenge: Predicting Despicable Me 2 and the Lone Ranger

A version of this framework (including tweets sentiment views shown in Figure 5.4, temporal modeling shown in Figure 5.1, similarity widget shown in Figure 5.3 and a linear regression baseline model) was used to predict 23 movies over the course of 3 months in the VAST 2013 Box Office Challenge. The VAST Box Office Challenge was a closed world contest in which contestants were provided with a set of Twitter indices, bitly links, and access to the Internet Movie Database. A web-enabled visual analytics toolkit was developed to enable analysts to quickly extract, visualize and clean information from social media sources (specifically bitly and Twitter). These

tools were then combined with linear regression and temporal modeling for movie box office prediction and sentiment analysis for movie review rating prediction. This section will discuss the box office prediction process using these various tools developed as well as lessons learned from the contest.

An example prediction process focuses on the July 4th holiday in the United States when Despicable Me 2 and The Lone Ranger were released. This weekend was challenging for two reasons. First, the data stream from the contest was broken, providing only 6 days worth of Tweets, and, second, the predictions were for a five-day weekend as opposed to the typical three-day weekend. Using the available data, a rough estimate was obtained for the Despicable Me 2 box office value in the range of \$76M +/- \$13M and \$85M +/- \$13M for The Lone Ranger. Next, the expected three-day weekend total was explored and the time series model approximates that \$124M is available for the two movies for the three-day weekend. A quick look at Figure 5.1 shows that the regression predictions are well outside the bounds of the time series model prediction.

Given the misalignment between the two models, the similarity views were explored to determine which movies The Lone Ranger and Despicable Me 2 are most similar to based on the predicted review score as well as various other metrics. Despicable Me 2 is compared to a variety of animated movies and one can see that the predicted \$73M is actually low when compared to animated movies such as Monsters University. Next, various similarity views are explored for The Lone Ranger and one can see that it is likely similar to World War Z, which had a weekend gross of \$66M.

After looking at the available information, it's determined that Despicable Me 2 should perform similarly to Monsters University, and a three-day gross was predicted to be \$85M. Based on the temporal prediction, this left only \$39M for The Lone Ranger; however, given the other evidence, it seemed likely that The Lone Ranger

Table 6.1: Comparison With Peer Teams Predictions

Team	Gross Prediction			
	Entry	Average Error	STD	MRAE
VADER(Interactive)	23	11.213	9.416	0.467
Team Prolix	23	16.466	15.195	0.424
Uni Konstanz Boxoffice	14	17.056	15.743	3.929
CinemAviz	21	17.219	17.677	1.970
Team Turboknopf	8	21.9	15.606	0.685
elvertoncf - UFMG	3	12.677	9.806	3.009
Philipp Omentisch	5	30.657	38.028	0.678
CDE IIIT	2	60.6	62.084	0.537

would underperform. Finally, the three-day prediction values were linearly scaled to be a five day prediction, resulting in a final five day prediction of $116.5M for Despicable Me 2 and $55.45M for The Lone Ranger. The actual three-day gross for Despicable Me 2 was $83.5M and $29M for The Lone Ranger. The actual five-day gross for Despicable Me 2 was $143M and $48.7M for The Lone Ranger.

6.1.1 Comparison With Peer Teams

Eight teams (Team VADER represents the predictions in this thesis) from various research institutes participated in the VAST Box Office Challenge. Data was also collected from 4 professional movie prediction websites. The prediction performance is compared with respect to peer teams from the VAST challenge and professional predictions.

Table 6.1 provides summary statistics of the performance of each team that participated in the VAST Box Office Challenge. For the gross prediction the measures reported are the average error (in terms of millions of dollars), the standard deviation (STD) of the average error term and the mean relative absolute error (MRAE), which is the percentage of bias deviating from the real value.

$$MRAE = \frac{1}{N}\sum_{i=1}^{N}\frac{|Prediction_i - RealValue_i|}{RealValue_i} \tag{6.1}$$

These statistics can be interpreted by their magnitude, where smaller values indicate more accurate predictions. Data collected in Table 6.1 was provided to all challenge participants after the contest was closed.

In terms of average error and standard deviation, team VADER reported the lowest values in gross prediction across all teams. With respect to the MRAE for gross prediction and viewer rating, VADER's results are slightly worse than Team Prolix (MRAE of .424 for Prolix compared with VADER's .467), and similar in range to Philipp Omentisch, CDE IIIT and Team Turboknopf. While Team Prolix was able to achieve a smaller MRAE over the contest than VADER team, comparatively, they have a much larger average error and standard deviation indicating more inconsistency in their predictions.

6.1.2 Comparison With Professional Predictions

In order to explore the hypothesis that the use of visual analytics will enable end-users to develop better box-office predictions, corresponding movie predictions have also been collected from four professional prediction websites for comparison. For the comparison to the professional prediction websites, the results of the VAST Box Office challenge are explored again. Given that these results were collected and verified by the contest organizers, this should be considered as an adequate means of justifying

Table 6.2: Comparison With Professional Predictions.

Prediction Source	Entry	Average Error	STD	Average MRAE
VADER(interactive)	21	12.729	9.425	0.285
VADER (No interaction)	21	23.051	22.011	0.501
boxoffice.com	21	8.538	7.466	0.191
filmgo.net	6	12.75	7.409	0.297
hsx	20	9.06	7.397	0.205
boxofficemojo	14	9.864	7.527	0.224

their validity. For the comparison in Table 6.2, only 21 movies are shown in the chart as two movies, The Bling Ring and The To Do List, were limited release movies which opened in only 5 and 591 theaters respectively and most expert prediction sites do not provide predictions for limited release movies.

Results in terms of the MRAE are given in Figures 6.1 for the opening weekend gross. It provides a comparison of the MRAE with that of several expert prediction websites. From Figure 6.1, it is clear that VADER's prediction outperformed the experts in the case of three movies (Epic, Hangover 3 and Fast and Furious 6), and in the case where VADER had the largest error (After Earth) the team relied heavily on the analytical component with no interaction.

Table 6.2 gives the average error, standard deviation and MRAE for the predicted movies. What the results show is that for the model used, the predictions of the team utilizing an interactive tool were a dramatic improvement over just the model itself (see Table 6.2 VADER (Interactive) versus VADER (No Interaction)). This provides a strong indication that the hypothesis that the use of visual analytics will enable end-users to develop better box-office predictions when compared to a purely

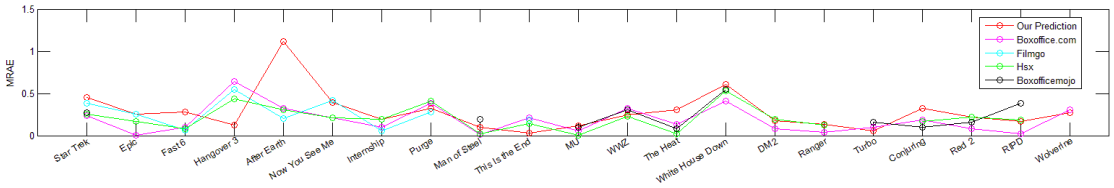


Figure 6.1: The Relative Absolute Error of Box Office Weekend Gross Predictions, Where the X-Axis is the Predicted Movies.

statistical solution is valid. However, this thesis does not wish to overstate its claims. This contest provides only a single data point for exploring how one group of analysts in a closed world setting were able to utilize a visual analytics toolkit for improved prediction. What this demonstrates is the need for further controlled studies in which a group of analysts perform similar model predictions and results are compared between analysts using a visual analytics platform and analysts using only results from a given regression model. However, results from the contest indicate that a visual analytics toolkit can enhance business intelligence.

Further analysis of the data also indicates that these tools enabled the team of novice box office analysts to quickly close the gap between the experts. Table 6.2 shows the average error and standard deviation for the team's predictions and compares them to four well known professional prediction websites. What one can see is that both the average error and average MRAE are slightly lower than filmgo.net indicating that the proposed methodology enabled the group of novice analysts to be competitive when compared to expert analysts. The significance of this relies on three major assumptions:

1. The professional prediction websites have more experience in box office prediction than the team.

2. The professional prediction websites have access to more data than the team was allowed in the closed world contest.

3. Access to more data can enable better predictive models as evidenced by [132, 134, 136, 150]

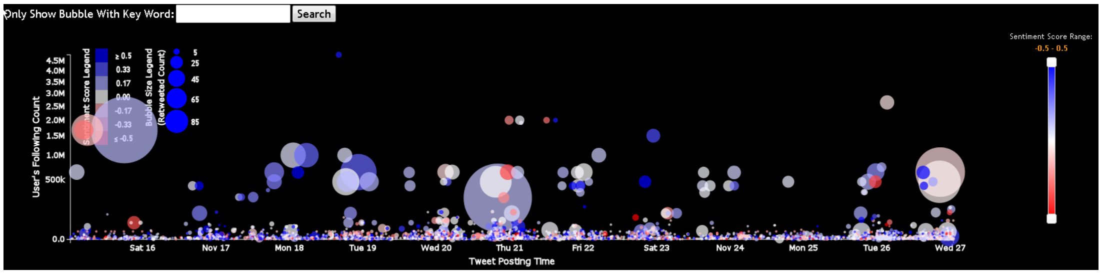
First, it seems reasonable that a professional prediction website would have much more experience than a computer science team who has never previously attempted to predict box office sales. Second, it is clear that utilizing data sources (specifically the number of theaters a movie is released in) will result in a better prediction model (a larger R^2). From these assumptions, it becomes clear that (in this instance) the application of a visual analytics toolkit can enable individuals that are knowledgeable with respect to data analysis to quickly understand information being presented to them in new domains and make predictions that are in line with expert predictions. Overall, the prediction error (.285) was slightly lower than that of filmgo (.297), but approximately 50% worse than boxoffice.com (.191). However, if the After Earth and Now You See Me weekend (during which the team relied heavily on the model and very little on the interactive visuals) is removed, the MRAE drops to .239 which puts the team's performance near the prediction range of boxofficemojo. Other sources of error can be accounted for in disrupted Twitter and bitly data feeds. These interruptions were pronounced for The Heat, White House Down, Monsters University and World War Z. However, even with those interruptions, the predictive analysis process was still quite robust with only The Heat being a significantly worse prediction than the professional sites.

6.2 Using Multiple Models: Predicting Disney's Frozen

This section demonstrates how an analyst would use the system (focusing on the feature exploration and model building views shown in Figure 5.2 and Figure 5.6) to predict Frozen's opening weekend gross. This process consists of multiple steps, which can be iteratively traversed in different ways. However, this thesis suggests the following procedure. First, the user gets an overview of the Twitter and Youtube comments using the dual-y-axis line chart (Figure 5.5 view (a)) to compare movies released together. Second, details can be investigated using the detail pages of the candidate movie (the sentiment bubble plot and sentiment wordle shown in Figure 6.2). Third, the user can explore similar movies and compare their gross (in Figure 5.3, as well as how well the baseline model performed for them. After having a general impression of the expected revenue, the user can navigate to the Feature Selection tab to explore and select features or filter movies to create a model (Figure 5.2. Finally the user can build and explore different models and their prediction ranges in the Explore Models view (Figure 5.6). Step 4 and step 5 can be iteratively applied until the user feels they can make a confident prediction.

To illustrate these 5 steps, Frozen is taken as an example. Starting on the overview page, the line chart in Figure 5.5 (a) indicates that there are 4 movies released on the same weekend (Frozen, Black Nativity, Homefront, and Oldboy). One can quickly see that online chatter (Tweet and YouTube comment volume) about Frozen is not dominating the other weekend movies, in fact it is trending similarly to the movie Black Nativity. This phenomenon indicates that it is unlikely that Frozen will obtain an anomalously large gross as the market will be shared by competitors.

In the second step, using the detailed view of Frozen (see Figure 6.2) the Tweet sentiment is analyzed. One can see frequent Tweet keywords and the sentiment



(a) Bubble Plot

(b) Sentiment Wordle

Figure 6.2: A Tweet Bubble Plot in (a) and a Sentiment Wordle in (b) for Movie Frozen

polarity. Also, the retweet volumes provides information about users' interest in the movie and the advertisement campaign. For example in Figure 6.2 one can see that Frozen does not have a large Tweet and retweet volume compared to other blockbusters; however it does have a very positive sentiment (blueish dots). The movie sentiment score for Frozen is approximately 0.8 which is very high among all 112 movies having Twitter data.

In the third step the similarity widget is explored (see Figure 5.3). This reveals that movies similar to Frozen were under-predicted with the baseline model, which predicts about $44M for Frozen. The fourth step focuses on the analysis and selection of the movies features (see Figure 5.2). There are two main views for feature selection: the correlation view showing relationships between a feature and the revenue, and; the

Table 6.3: Results for Frozen and Hunger Games. The Opening Weekend Gross for Frozen is $67M and for the Hunger Games it is $158M.

subject	user1	user2	user3	user4	user5	user6	user7	BoxOffice.com	BoxofficeMojo
Prediction (Frozen)	55.9	59	50	60	57.7	62.5	58	47	44.7
Abs Error	11.1	8	17	7	9.3	4.5	9	20	22.3
Prediction (Hunger Games)	71.1	135	NA	100	95.9	86	75	166	167
Abs Error	86.9	23	NA	58	62.1	72	83	8	9

relationship among features depicted in the PCP view. From the baseline model one selects the number of opening screens, the budget and the weekly average of Tweet counts as an initial feature selection. This gives us a model with $R^2_{adj} \approx 0.58$ (M1 in Figure 5.6). To further improve the model, one adds another feature, view counts of the movie's YouTube trailers, and built both an SVM and LIN model. R^2_{adj} improved to approximately 0.6 while the prediction deviations from the different folds decreased. Next, using the background knowledge, the user explores the genre of this movie (in this case the genre is "Family"). While adding the Family feature to the Parallel Coordinates, one finds that the gross distribution for Family movies is significantly different to most non-Family genres. Thus, for the last prediction iteration, one adds the family feature to the model. The model obtained an R^2_{adj} score of 0.745. Finally, the user reviews the Model Prediction Comparison graph and decided to finalize the prediction between $60M to $70M based on the best performing models.

Table 6.4: Results for Divergent and Muppets. The opening weekend gross for Divergent is $56M and for the Muppets it is $16.5M.

subject	user1	user2	user3	user4	user5	user6	user7	BoxOffice.com	BoxofficeMojo
Prediction (Divergent)	54.1	53	40	50	30.1	47.5	48	66	51
Abs Error	1.9	3	16	6	25.9	8.5	8	10	5
Prediction (Muppets)	50.6	21.5	28	15	35	21.4	20	25	22
Abs Error	34.1	5	11.5	1.5	18.5	4.9	3.5	8.5	5.5

6.2.1 User Study

In order to evaluate the effectiveness of this framework for predictive analytics, this thesis performed a user study. On March 20th, 2014 seven graduate students from China, India, the United States and Germany were enlisted and asked to predict the results of four different movies. The first two movies predicted were to provide them with baseline training, the next two movies were to be released on March 21st, thus having them do an actual future prediction. The movies they were predicting included Disney's Frozen (2013) and The Hunger Games: Catching Fire (2013) (which were used for training) and Divergent (March 21, 2014) and Muppets Most Wanted (March 21, 2014) (which were the movies to be predicted). For Frozen and the Hunger Games, their weekend box office data was removed for the training exercise in order to simulate the prediction process.

Of the seven participants, six were male, one was female and all were PhD students. Prior to participation, they were surveyed about their cinema affinity and

data visualization knowledge on a scale from 1-5 (with 1 being the lowest). From the seven participants four claimed to be visualization experts. Five subjects rated their movie affinity as low (1-2), and two rated medium (3-4). Their machine learning knowledge was mostly low, with only two participants claiming a basic knowledge of machine learning and prediction related tasks (these students had all taken regression analysis and/or data mining courses, as such it feels that they can be considered to have a relatively high level of expertise in the modeling and analysis process). The two subjects that rated their movie affinity as low were those that rated their machine learning and predictive analytics knowledge as high. Thus, there are three subjects that were casual users with limited domain knowledge and limited analytics experience, two subjects that had some domain knowledge and limited analytics experience, and two subjects that had expertise in data mining and predictive analytics but limited domain knowledge.

To introduce the system, an example analysis of the movie After Earth was walked through and the proposed analytics process (similar to the case study) was explained. Subjects were then asked to predict Frozen and The Hunger Games. During the analysis and prediction process of these two movies, they were open to ask any questions, such as the meaning of a feature, how to use a special function of the system, and what information could help to choose proper features and improve the model performance. After they submitted their final prediction about a movie, they were told the real gross so that they could make a comparison and adjust their strategy for the next movie. After practicing with these two movies, they used the system (unaided) to predict the new movies Divergent and Muppets Most Wanted.

To get a deeper understanding of the users analysis processes this study was carried out as talk-aloud [151] session. The users were asked to speak their thoughts out loud explaining their actions. The voice and system interaction were recorded by

video. After the study the key results were summarized and classified into System Usability, Social Media Exploration, Feature Selection and Model Comparison.

6.2.2 System Usability

Key findings here indicated more details on system design. All subjects reported ease of use and interaction with the system. Furthermore, the length of the user study demonstrated the subjects' engagement. No instructions were given on the time needed to make a prediction; however, subjects spent over 1 hour on average tuning system parameters and exploring the data. Subjects also were excited to compare their results Monday and indicated they wanted to try this again. Design issues they faced were that they wanted even more transparency in the data. As no subject was a self-rated expert in cinema (most indicating they had seen less than two movies in the past 6 months) many of the subjects wanted more information about the movie features. They suggested direct links to the IMDB pages for the movies to allow even greater detail views. Overall, the most used views were the similarity page and the feature selection page.

Subjects all started their analysis on the overview page, exploring time series trends and comparing how they felt the movies on the weekend would fare when compared to others. They typically looked at the Twitter and YouTube volumes and sentiment data. At the beginning they found it difficult to interpret those visualizations as they were unfamiliar to a user; however, by the end of the study the users were requesting more features, wanting to create difference maps of the movies to look for keyword differences in the sentiment analysis and also to identify what was being discussed differently between YouTube and Twitter. As such, it is clear that more text analysis is needed for further insight generation. A clear example of gaining insight was shown during the analysis of the movie, Divergent. No subject had

heard of this movie; however, when inspecting the data they saw that The Hunger Games was often referred to in context with this movie. This grounding gave them the contextual clues which they needed in order to analyze Divergent.

Negative comments focused on the disconnect between the similar movies and the users' thought process. In the Feature Selection page, users are presented with the five most similar movies with respect to the selected PCP features. This is calculated as a Euclidean distance metric, and the calculation is a black-box to the user. As such, analysts were often wary of these movies and preferred to use the "create your own similarity" option on the similarity widget page. However, this again required more domain knowledge than some users had, with many again requesting details about what genre, rating, etc. a particular movie had. Future work should include better views for multi-dimensional similarity matches and more transparency in the similarity metrics. Yet, what the process highlights is that all subjects, even those with little self-proclaimed movie knowledge, are able to bring some background knowledge into the prediction process, which could be used to add value when compared to a purely automated prediction process.

6.2.3 Feature Selection

All users worked with the Feature Selection table to determine which data was available for a movie and remarked on how they felt the prediction was more reliable when they knew that the data existed. Again, this indicates that transparency in the model training can improve an analyst's confidence. During the feature selection process, most users started with the baseline settings, inspected the results and then iteratively chose more features with high correlations, reinspected and then iterated again. Other users again applied their domain expertise and chose features that seemed interesting to them. For example, the user that had seen 10 movies in the

theater in the past six months used his domain knowledge to select features which are not obviously highly correlated to the revenue, but these features considerably improved that subject's model.

Participants who decided to add Twitter related features typically based this choice on the genre of the movie, stating that Twitter users would be interested in Divergent but not in the Muppets. One user, with a basic background knowledge in prediction tasks commented on how the Parallel Coordinate view enabled her to choose features that were independent (i.e., not multi-correlated). Other users engaged the PCP view to filter out movies to create models based on genre or movie ratings. Overall, they spent a large amount of time exploring features and discussing what they felt these features meant. They also found it extremely helpful to see how the selection of different features impacted the amount of movies available for training.

Negative comments revolved around users' frustration in feature selection, noting that there should be a way to provide more details on what is likely to be a good feature. For the inspection of correlations, one user noted that it was hard to use the PCP view and had difficulty distinguishing the highlights. However, the users all liked the design of the framework, and commented on how it would be useful to change the domain to look at other specific problems of interest. It shows that a future work could be to explore how to improve the presentation of features. Obviously showing all features (in this case 116) is a huge amount of information overload; however, it is also desired to involve the user and allow him/her to use domain knowledge to guide the modeling and prediction process. It could also to explore several methods of automatic feature selection as a means of organizing information for visual presentation and exploration and performing user studies across various feature set visualizations in order to explore this area.

6.2.4 Model Comparison

As for the Feature Selection view, participants found the model comparison features extremely useful. Starting with some initial predictions, they tried to improve the model to reduce the errors. Users often focused on prequel movies (particularly during the Hunger Games prediction) and focused on developing a model that was a good fit for known prequels or known movies within a genre. One user repeated the feature inspection, selection and modeling until he was able to create a model that strongly fit to the prequel (in the case of the Hunger Games). Others tried to inspect all outliers and then made decisions based on their domain expertise regarding movie similarity. This would lead to an iterative model building and refinement loop. Users also inspected the scatterplot and would then access the similarity comparison tools to explore the impact of Twitter on the model prediction. Users noted that Twitter seemed to have an impact depending on the type of movies, and many came to the conclusion that Twitter was relevant when predicting Science Fiction movies (such as Divergent) but less relevant when predicting Family movies (such as the Muppets). Again, subjects indicated a desire for even further transparency of the inner workings of the model prediction.

6.2.5 Prediction Results

Table 1 and 2 show the results of this user study in both the training trial and the actual prediction trial. For the training results (Table 1), subjects were found to have a lower error than that of the experts for Frozen; however, for the Hunger Games, subjects found this very difficult to model. It is important to note that the user study went through the example of After Earth, Frozen and the Hunger Games for training in order to give subjects examples of a low outlier, a good fit, and a high

outlier respectively. In this way they can explore all possible scenarios prior to the actual prediction task.

For the actual results (Table 2), 5 of the 7 subjects were able to best BoxOffice.com predictions for Divergent and 2 of the 7 subjects were able to best both expert prediction websites. Only two subjects erred on the far low end of the spectrum for this movie (subjects 3 and 5). For the Muppets, 4 of the 7 subjects were able to best the experts, with one subject (subject 4) accurately predicting this would be a box office failure. Again, subject 5 was an outlier, and subject 1 predicted that the Muppets would be an outlier on the positive end of the spectrum.

Overall, the results of this study are quite positive. Given the subjects self-reported lack of movie knowledge, it is clear that the integration of social media and visual analytics for model building and prediction can quickly generate insight at a near professional prediction level. Subjects 2 and 7 had the highest self-reported domain knowledge and (as seen in Table 2) outperformed experts from BoxOffice.com (and Subject 2 outperformed the BoxofficeMojo results as well). The machine learning and regression experts were subjects 4 and 6 and they also outperformed the experts. The remaining subjects can all be considered more casual users and had a higher variability. In both future prediction cases, over half the subjects were able to best the experts over the course of a one hour training session. Furthermore, such work indicates that visual analytics can have a direct impact on the modeling and prediction process. As noted by Lazer et al. [14], there is a need for tools that can improve insight into large data analytics and an increased transparency can potentially lead to improved model efficacy. Future work will look at doing a more formal evaluation where a larger subject pool is recruited and more analysis between the three groups is performed.

6.3 Twitter Network Properties

Previous studies have indicated that social media data is useful for movie revenue prediction. To better predict movies' revenue, researchers have also explored more complex models. For example, purchase intention was mined from social media text and used to predict box office revenue through linear regression and support vector machines [152]. Zhang et al. [153] used neural networks and 16 variables to predict box office revenues and achieved a 75% accuracy one month before release. Social influence has been studied to explain box office distributions [154]. Multiple social media sources and many different metrics have been studied as predictors for the box office. However, the previously explored features are mostly descriptive analytics (e.g. number of tweets and distribution of different types of tweets) and content analytics (e.g. text mining and sentiment analysis) but little work on network analytics has been done. However, some work has shown that the network structure can be useful in business analysis. For example, Marc A. Smith et al. [155] proposed six conversational archetypes of Twitter social networks from the aspect of marketing. Seeing the promising contribution of social media to predictive analytics, this thesis further explored the Twitter network structure and properties with respect to movie revenue prediction.

A social media community, such as Twitter, Facebook, and blogs, can be considered a copy of a human community of people but organized in a digital world. There are different types of users, actions, messages, and behaviors. Questions asking for more than a single prediction number could be 'Are different users playing the same role in forecasting?'; 'What are some network community structures for different movies?'; and 'Does the different structures represent different marketing strategies and contribute differently to box office revenue?' Motivated by such questions, this

case study collected movie related tweets since 2013 and studied their Twitter network features.

In this analysis, Twitter networks for different movies are explored in three steps. First, the re-tweet network is extracted for each movie and the visual display of the network structure is explored. Similar to [155], users are clustered and the group-in-a-box layout [156] is generated for each movie. Exploring the network layouts, this thesis proposes a set of conversational archetypes based on the categorization proposed by Smith et al [155]. Two coders coded the re-tweet networks and further analysis about box office and movie types were conducted. Second, this study extracts a set of numeric network features (e.g., clustering coefficient, connected components, and density) and uses linear regression analysis to explore the predictive power of these network features. Third, ordinal logistic regression is used to see if network features (including both the conversational types as categorical features and the numerical network features) are useful to predict if a movie is a success or failure based on the ratio of its total domestic gross and its budget.

The results of this analysis reveal the predictive value of network analysis in box office prediction. From the conversational types, different social network organization patterns can be discovered and the distribution of successful movies are different in these types. Numerical network features have also been shown to be useful in opening weekend prediction via a linear regression model. Consistent with the results of the box office analysis across different conversational types, the ordinal logistic regression analysis indicates that movie network type could be useful to indicate whether a movie is successful or not when coupled with other movie meta data and network features. Although more analysis is required, this study indicates that network analysis on social media data could be useful for revenue prediction in short term and long term (the opening weekend prediction and the total revenue prediction).

6.3.1 Data Collection

Following the work of VAST Box Office Challenge 2013 [157], this thesis uses the same data collection strategy and collects tweets for 400 movies released between January 2013 and June 2016. These tweets are used in the network property analysis. In addition to Twitter data, movies' meta data from IMDB was also collected, parsed, and cleaned for movies under analysis. Particularly, Screen (the number of theaters a movie is released on at the opening weekend), Budget (the projected budget for a movie, in million dollars), OWG (opening weekend gross, the 3-day box office of the opening weekend in the U.S.) and Genre (the genre(s) of a movie) are used in this analysis. Movies released in fewer than 1000 screens are filtered out, as well as movies that had tweet collection interruptions.

This study uses 261 movies in total after filtering, and it uses tweets posted two weeks prior to each movie's release date, where 8,770,448 tweets and 4,666,243 distinct users are under analysis.

6.3.2 Conversational Archetypes of Movie Twitter Network

Tweeting on the social network is one important advertising approach of movies. Distribution companies, actors/actress, theaters and fans are all tweeting about their movies to boost sales. There are a couple of hypotheses on successful social network promotions [154]. For example, more discussions indicate more potential audiences, and fewer isolates means tighter connections and maybe better diffusion of the advertisement. This section will analyze the Twitter network structure of the movie related tweets and organize them into different archetypes for analysis. Inspired by the six twitter conversational archetypes in marketing proposed by Smith et al. [155],

this thesis provides a modified set of archetypes focusing on the movie marketing problem.

Twitter Network Visualization

For each movie, a network graph is generated according to the following steps. First, the retweet network is extracted where each user is a node and each retweet relationship is an edge pointing from the retweeting user to the original user. Second, the Clauset-Newman-Moore clustering algorithm [158] is used to cluster nodes into groups and the descriptive metrics for the clustering results are calculated. Third, the network structure is visualized in a node-link diagram and organized using the group-in-a-box layout [156], which assigns each group a region like treemap layout so that the area is proportional to the number of nodes. The node-link layout algorithm within one cluster is the Harel-Koren Fast Multiscale [159] or the Fruchteman-Reingold [160] if the Harel-Koren method reached the memory limitation. The last two steps are done using NodeXL [161]. Due to the machine's memory limitation, 42 out of 261 movies have no visualized network graphs for their entire two weeks retweet network (219, 84% of the movies have graphs). Figure 6.3e shows an example of the layout for the movie *Fast & Furious 6*, and one can see big clusters in dark blue, light blue, dark green, light green, red, etc. and many small isolated clusters on the right bottom of the graph. Edges are displayed as gray lines.

Movie Tweets Conversational Archetypes

In a previous analysis on marketing related tweets [155], crowd network structure has been organized into six types based on conversation styles. A brief summary of these six types is listed as follows.

- **Polarized Crowd** shows two big and dense groups that have few connections between them.
- **Tight Crowd** shows highly interconnected people with few isolated participants.
- **Brand Cluster** has many disconnected participants with medium or small sized communities low inter-connectivity.
- **Community Clusters** has many medium-sized communities that are somewhat connected, and a fair number of isolates.
- **Broadcast Network** has a hub which is the main media outlets and many people repeat its posts.
- **Support Network** shows a hub replying to many other people.

According to the development scenario of these six types, almost all movie related Twitter networks should belong to the **Broadcast Network** archetype because it describes the conversational state that some influential promoters (e.g. the movie company) are mainly repeated by others and there are few conversations between those none-influential users. However, when trying to cluster the movie twitter networks, it is found that this set of archetypes is not very proper to describe the observed properties and coders have difficulty agreeing with each other. Some problems are encountered. First, in the movie Twitter network, one can barely see communities where users are tightly connected to each other, but groups where most users are communicating with a center hub. This phenomenon indicates that most of the networks are showing the properties of a **Broadcast Network**. Second, even if a Broadcast Network is used to describe the Twitter networks for all of these movies, there are still some properties that cannot be captured. Some of them also have large amounts of

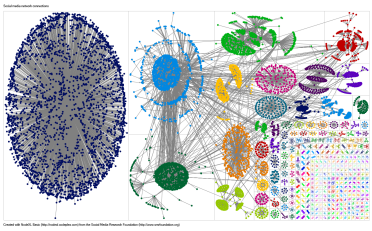
isolates, and some of them have multiple broadcasting centers. Third, in the context of movie promotion, the **Support Network** and **Tight Crowd** archetypes are not observed.

In order to better describe the conversational features represented in the network, this thesis proposes a set of modified archetypes as below.

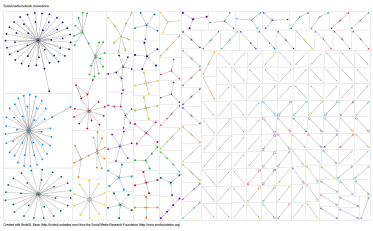
- **Broadcast:** The Broadcast conversational archetype describes the situation that in which information posted by the main sources diffuses very well through the whole network. It usually has one big group with a hub which is the information source. It can have more than one big groups, but the groups should be well connected. While there could be some middle-sized groups, the connections from the big group should cover the majority of the population, and there should not be many isolated small groups.

- **Community:** The Community conversational archetype describes the situation in which a couple of communities are obvious, and the information from the main sources in each community does not diffuse very well to other communities. This kind of archetype usually has some middle-sized groups, and there could be some connections between different groups but no dense connections covering the whole population. This archetype usually has many isolated small groups (e.g. with fewer than 10 users) due to the lack of good connectivity. There could be a hub within a group, but unlike Broadcast, the connections from a hub do not cover the majority of the population.

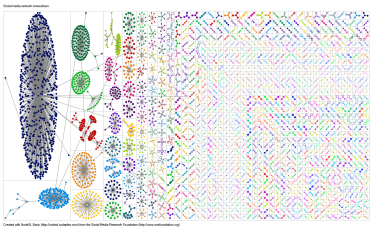
- **Broadcast with Community:** The Broadcast with Community conversational archetype describes the situation in which some main sources diffuse quite well and cover many users while the remaining users are forming poorly connected communities. This archetype usually has a big group with a hub and



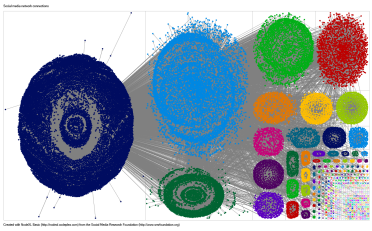
(a) Polarized (The Lone Ranger)



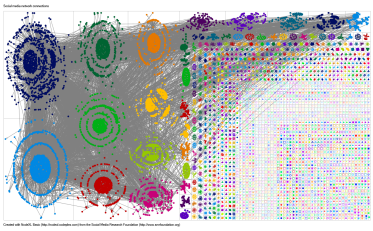
(b) Brand (Oz the Great and Powerful)



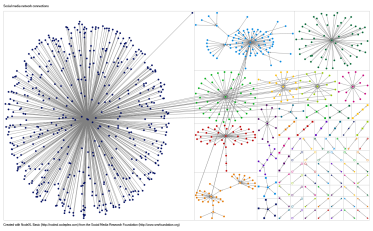
(c) Community (Iron Man 3)



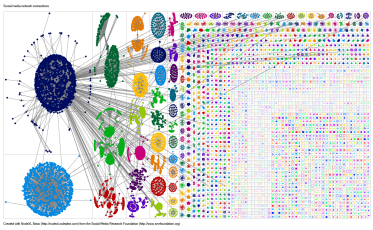
(d) Broadcast (Zootopia)



(e) Broadcast with Isolates (Fast & Furious 6)



(f) Broadcast with Community (Walking With Dinosaurs)



(g) Broadcast with Community and Isolates (Gangster Squad)

Figure 6.3: The Network Layout Examples for Different Conversational Archetypes

many middle-sized groups. Some of the middle-sized groups are retweeting the big group but other middle-sized groups are disconnected from the big group. For those disconnected groups, there could exist inter-cluster connections, but the center nodes are not connected directly to the hub of the main sources. This means that even they are not absolutely isolated, but they are not likely to spread the same information posted by the main source. To separate this archetype from Broadcast with Community and Isolates, isolated small groups in this type cover less than 1/3 of the population.

- **Broadcast with Isolates:** This is similar to Broadcast where it has a big group and some middle-sized or small groups and many connections between them. But this type also has many isolated small groups (more than 1/3 of the population) which are not connected to the hub of the big group. This means the information from the main sources diffuses well among the middle-sized clusters but does not reach those small groups.

- **Broadcast with Community and Isolates:** This is a hybrid of broadcast, community and isolates. It is similar to broadcast with community, but has many isolated small groups.

- **Brand:** This is similar to Community, but has even fewer connections between groups.

- **Polarized:** This type has two parties covering at least 2/3 population of the network. Each of them may have a hub and most nodes are well connected within the party. If many users of one party are retweeting the hub of the other one, it should be a Broadcast network.

Table 6.5: Conversational Archetype Distribution

Conversational Archetype	N Decision	Opening Weekend Gross			Total Domestic Gross		
		> $100M	> $80M	> 0.3 Budget	> $300M	> $200M	> 1.5 Budget
Polarized	18(4.1%)	0	0	10(55.6%)	0	0	7(38.9%)
Brand	8(1.8%)	0	0	4(50.0%)	0	2	2(25.0%)
Community	42(9.6%)	2	2	32(76.2%)	2	2	16(38.1%)
Broadcast	302(68.9%)	4	13	209(69.2%)	3	25	138(45.7%)
Broadcast with Isolates	15(3.4%)	0	2	11(73.3%)	0	2	7(46.7%)
Broadcast with Community	47(10.7%)	0	0	18(38.3%)	0	0	10(21.3%)
Broadcast with Community and Isolates	6(1.4%)	0	1	4(66.7%)	1	1	2(33.3%)
All	438(100%)	6	18	288(65.8%)	6	32	182(41.6%)

Figure 6.3 gives one example for each of the above seven archetypes. For the 219 movies that have the graph visualized for the whole data, two coders labeled each graph into one of the above seven types after discussing and understanding the description of these conversational types. Among the 219 cases with 438 decisions, the coders agreed on 196 graphs and disagreed on 23 graphs. The percent agreement is 89.5%, and the nominal Krippendorff's α is 0.791, which can be considered acceptable [162].

Among these archetypes, Broadcast still covers the majority of the movies, 302 decisions among 438 (69%). The distribution of decisions for each archetype is summarized in Table 6.5.

Movie Box Office Analysis by Network Archetypes

Different conversational archetypes represent different media information dissemination states, and social media platforms play an important role in advertisement and may contribute to the gross of a movie. In box-office analysis, there are usually two tasks. One is to look at the short-term gain and the other is to project the long-term gain. The first task is usually framed by the opening weekend gross and the second task is framed by the total gross of the movie. Both the actual value and the relative value to the budget are of interest since high revenue movies are usually very popular and also a successful movie should at least match its cost and make some profit.

To evaluate the value of different conversational archetype in box office prediction, movie revenue distribution in each archetypes is explored under both short-term and long-term analysis. For short-term analysis, the analyst first looks at the number of movies having a high absolute opening weekend gross (more than \$100M or more than \$80M) and then looks at the number of movies earning more than 30% ($\theta = 0.3$) of the budget on its opening weekend. Because the opening weekend of a movie's release typically accounts for 25% of the total domestic box-office gross [131], it is assumed that if a movie can make more than 30% of the budget on its first weekend it should match its cost and make some profit. For long-term analysis, these two aspects are also studied, the number of movies with a high domestic gross (more than \$200M) and the ratio of a movie's total revenue to its budget (greater than 1.5 times). The descriptive analysis result is organized in table 6.5.

Table 6.6: Movie Tweet Features

Feature	Description
Vertics	The number of vertices in the network
Edges_U	The number of unique edges in the network
Edges_D	The number of duplicated edges in the network
TBD7	The average daily number of tweets in one week(7 days) prior to the movie's release date
TBD14	The average daily number of tweets in two weeks(14 days) prior to the movie's release date
CC	The number of connected components
CC_S	The number of single vertex connected components
CC_MV	The maximum number of vertices in one connected component
CC_ME	The maximum number of edges in one connected component
Density	Density of the network
ClusteringCoeff	The global clustering coefficient of the network
Clusters	The number of clusters given by Clauset-Newman-Moore clustering algorithm
Cluster_M	The maximum size of the clusters
Modularity	The modularity of the clustering result
Cluster_10	The number of clusters smaller than size 10 (inclusive)
Cluster_5	The number of clusters smaller than size 5 (inclusive)
Cluster_2	The number of clusters smaller than size 2 (inclusive)

Looking at the first two columns (N Decision and Percentage), one can quickly observe that Broadcast is the most commonly coded conversational archetype (about 70%) and Brand and Broadcast with Community and Isolates are coded in only a few networks (they have 8 and 6 cases respectively). In the short-term analysis for the opening weekend gross, 65.8% of all movies making more than 30% of their budget. Two archetypes have much higher percentages than the average; Community has 76.2% and Broadcast with Isolates has 73.3%. This seems to indicate that a successful pre-release digital promotion needs more than one core information source, and the interest from the public users should be reflected as many isolated groups in the network. On the other side, Broadcast with Community has only a 38.3% short-term success rate, ranking the lowest among all types. Polarized and Brand are also much lower than the overall rate. This seems to be reasonable because such conversational archetypes do not indicate a healthy information diffusion state; where as both Polarized and Brand form disconnected groups where information is simply blocked and is not diffused well, and the Broadcast with Community type lacks public support (small isolates) and throughout connections like Broadcast.

In the long-term analysis for the total domestic gross, some similar patterns can also be observed that Broadcast with Community has the lowest success rate with only 21.3% of them making more than 1.5×Budget. Polarized (38.9%) and Brand (25.0%) both make less than the overall rate (41.6%), and the Broadcast with Isolates archetype still ranks highly with 46.7% of the movies making more than half of the budget in the opening weekend. Different from the short-term analysis, the success rate of Community drops below the overall rate in the analysis of the total domestic gross, and similarly the Broadcast with Community and Isolates also has a relatively lower long term revenue. These changes seem to indicate that community is not a positive feature in terms of a long-term projection because more intra-connections with

fewer inter-connections may be unhealthy for the growth of the network, although the influence on the short-term outcome is not obvious.

To analyze some specific high-revenue movies, it is found that the following three have made more than $300M dollars: *Iron Man Three*, *Frozen*, and *Despicable Me 2*. These three also made more than 1.5 times of their budget. However, among the movies that made more than $200M dollars, only 6 out of the 16 grossed more than 1.5 times their budget.

6.3.3 Network Factors in Predictive Modeling

In addition to looking at the layout of the network clustering result and analyzing the conversational archetypes, this study also explores a list of numeric network features (listed in Table 6.6). The first section contains five descriptive analytics features, the following two sections contain six global network analysis features and 6 clustering network analysis features.

Linear regression is used to predict movies' opening weekend gross, and ordinal logistic regression is used to predict the successful level defined by the ratio or a movie's total revenue to its budget. By fitting such models, the goal is to quantitatively analyze how useful network features are to the prediction of box office revenue.

In the network graph analysis, only 219 movies are evaluated due to computation limitations. Among the remaining 42 movies, some are quite popular, such as *Godzilla*, *X-Men: Days of Future Past*, and *The Maze Runner*. Considering the importance of these movies, the analysis also tries to include them in prediction modeling. The reason why they have no graphs is that 25 movies failed to complete the clustering algorithm, and another 17 failed to complete the layout algorithm after completing the clustering algorithm. Therefore, in the following modeling exploration, 261 movies will all be used in a full sample linear regression to predict the opening weekend gross

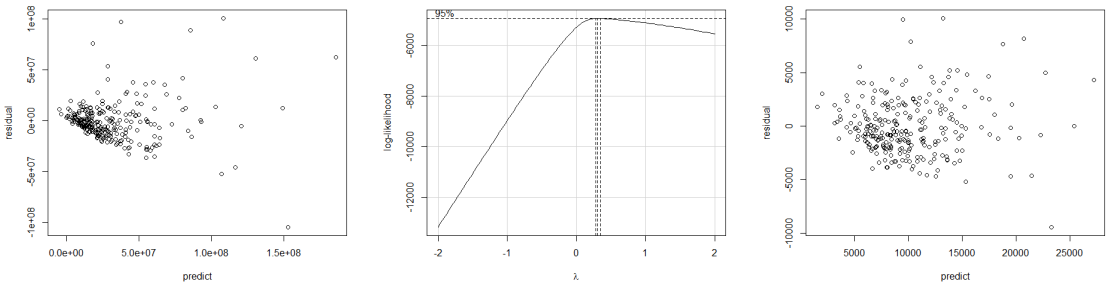
using only 13 features (*budget*, *screen*, 5 descriptive features, and 6 global network features). In addition, 236 movies (with 25 movies excluded for lacking clustering features) will be used in a full feature linear regression to predict the opening weekend gross using all 19 numerical features (*budget*, *screen*, 5 descriptive features, 6 global network features, and 6 clustering network features). Finally, both 19 numerical features and 7 categorical features (network conversational types) will be used in an ordinal logistic regression on 219 movies to predict the success level of the movie.

Network Features in Opening Weekend Prediction

Linear regression analysis is used to evaluate the prediction power of the numerical network features. For both the full sample linear regression and the full feature linear regression, the same analysis pipeline, data transformation, feature selection, and modeling and comparison [23], is followed. The analysis uses step-wise regression to find the best variable subset. Given this feature subset, all possible models are compared to explore the contribution of different features towards the box office prediction problem.

Analysis on All Movies This analysis runs on all 261 movies with 13 features through the following steps: data transformation, feature selection, and modeling and comparison [23]. Based on previous work, it is already known that the screen and budget are important variables and one can assume that when a movie has 0 screens and 0 budget, it makes 0 gross. Therefore, the full linear regression model is defined as no intercept.

Data Transformation: Based on previous work, logged regression models have been proposed with different predictors, such as screen, search volume, and tweet rate. To choose data transformation function, full regression models and residual analysis should be used. The residual plot for the initial full regression model without any



(a) Before Transformation (b) Box-Cox Parameter Plot (c) After Transformation

Figure 6.4: Residual Plot Before and After Square Root Transformation for the Full Model with 13 Features.

transformation is displayed in Figure 6.4a. This figure shows the residual distribution and has an obvious pattern which indicates that the response variable may need a non-linear transformation. A full regression model is run on all features and it shows the residual plot and the suggested λ in box-cox transformation. A Box-Cox parameter plot shown in Figure 6.4b suggests a square root transformation ($\lambda \approx 0.5$). Although $\lambda \approx 0.5$ is not the optimal value, it is very close to the optimal and square root transformation maintains more interpretability for the analysts. Taking this transformation, the updated residual becomes normal shown in Figure 6.4c, and the full model has $R^2_{adj} = 94.04\%$ and $p - value \ll 0.05$. The following analysis is carried out after this response transformation.

Variable Selection: After the response data transformation, variable selection is carried out in two steps, stepwise regression for feature subset selection and all possible regression models for model specific analysis. The stepwise regression is needed here because all 13 variables can easily create 2^{13} linear regression models even prior to adding interaction terms and this could be too much work for the All Possible Models analysis.

Table 6.7: Best Subsets Regression

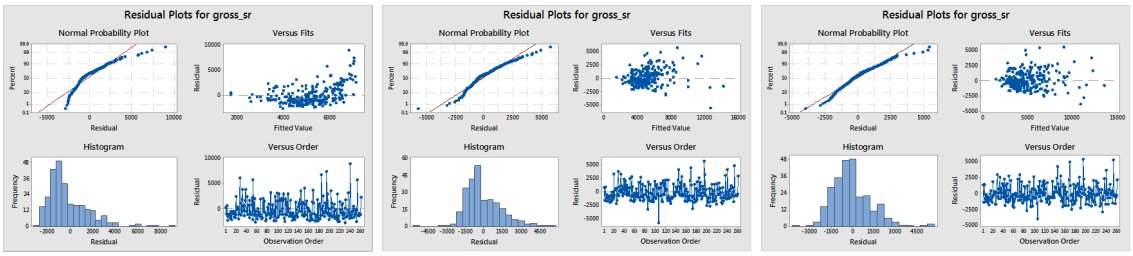
Var	Mallows's C_P	S	R^2_{adj}	Screen	Budget	CC_S	Density	TBD7	Edge_D
1	236.0	1870.0	88.65%	×					
1	875.5	2831.1	73.98%		×				
2	60.1	1499.6	92.70%	×		×			
2	68.4	1519.2	92.51%	×				×	
3	19.7	1398.5	93.65%	×	×	×			
3	25.0	1411.9	93.53%	×	×			×	
4	11.7	1375.5	93.86%	×	×		×	×	
4	12.2	1376.9	93.85%	×	×	×	×		
5	6.1	1358.4	94.01%	×	×	×	×	×	
5	7.9	1363.0	93.97%	×	×		×	×	×
6	6.0	1355.3	94.04%	×	×	×	×	×	×

Using forward-backward stepwise regression with $\alpha = 0.15$ for both to enter and to remove, the following variables are selected as a feature subset to minimize the prediction error of the linear regression model: *screen*, *budget*, *cc_singlevertexcc*, *edge_duplicate*, *TBD7*, and *density* (as used in Table 6.7). As it is no doubt that screen and budget are important variables, three are also some network features emerge in the model.

To find the best model subset, all possible regressions are performed. According to Mallows's C_p and the sum of squared error, the best models with different number of regressors are displayed in Table 6.7. The best subsets table shows that *screen* is a strong predictive indicator and as a simple linear regression model, it can achieve

R^2_{adj} = 88.65%. For the two-variable regression model, the best performing one takes *screen* and *CC_S* (the number of single vertex connected components) and its R^2_{adj} = 92.7%, while the second one with *TBD7* and *screen* has its R^2_{adj} = 92.51%. Among the three-variable models, the best one also uses *CC_S* and its R^2_{adj} = 93.65%. The goodness-of-fit (R^2_{adj} = 93.53%) drops slightly when *CC_S* is replaced by *TBD7*. The Mallows's C_p starts to become smaller when the fourth variable, *density*, is added into the model. With *screen*, *budget*, *density* and *TBD7*, R^2_{adj} = 93.86%, and with *screen*, *budget*, *density* and *CC_S*, R^2_{adj} = 93.85%. With five and six variables, the Mallows's C_p reaches good values (close to the degree of freedom). The model with *screen*, *budget*, *CC_S*, *density*, and *TBD7* has R^2_{adj} = 94.01%. The model with *screen*, *budget*, *density*, *TBD7* and *Edge_D* has R^2_{adj} = 93.97%. The model with all six variables has R^2_{adj} = 94.04%, but the variable *Edge_D* is not significant (it's p-value = 0.144, greater than 0.05) when the variable *CC_S* is used.

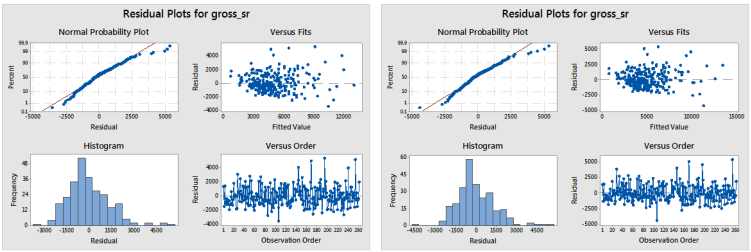
Model Comparison: Instead of finding the best proper model in the subset, the goal in the model comparison section is to explore how much accuracy Twitter features and Twitter network features can add to the model. First of all, the *CC_S* feature is the best candidate once *screen* is in the model. Comparing the best models with different number of variables, the statistical analysis on residuals for model evaluation are plotted in Figure 6.5. Clearly, the model with only *screen* doesn't fit very well since obvious patterns are shown and the distribution of residual is too far from being normal. When *CC_S*, and Budget are added to the model, the residual distribution becomes closer to normal, and this gets improved further when *density* and *TBD7* are added. Such analysis demonstrates that *screen* and *budget* are important features in box office prediction. However, social media features do help fit the model, and using multiple features from different aspects (e.g. *CC_S* and *density* are global network features while *TBD7* is a volume-based descriptive feature) benefit the prediction.



(a) Model with Screen

(b) Model with Screen and CC_S

(c) Model with Screen, Budget and CC_S



(d) Model with Screen, Budget, CC_S and Density

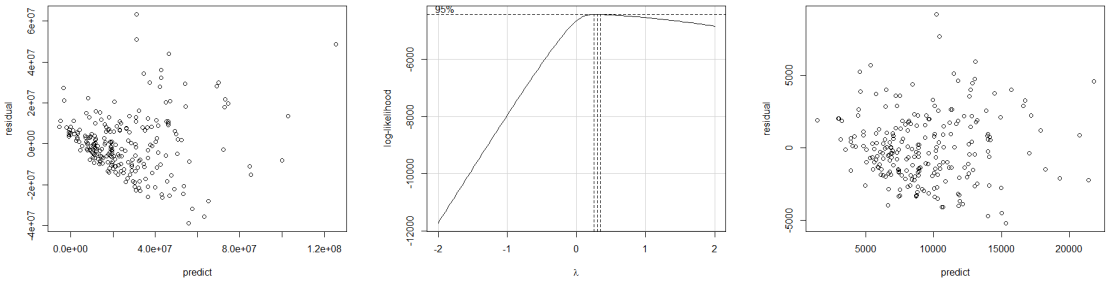
(e) Model with Screen, Budget, CC_S, Density and TBD7

Figure 6.5: Residual Analysis for Model Comparison with Different Features

Regression Analysis on All Features In order to evaluate the usability of network clustering features, a second set of regression analysis is run using all the features but only 236 movies having no missing data across the 19 numerical features.

Data Transformation: The residual plot of the full model is explored and the box-cox parameter plot decides that a square root transformation on the response should be proper. The residual plot before and after this transformation and the box-cox plot are shown in Figure 6.6. The full model has its $R^2_{adj} = 94.35\%$ and $p - value \ll 0.05$.

Variable Selection: There are 19 continuous features in the full model. Before running all possible models with feature subsets, the forward-backward stepwise



(a) Before Transformation (b) Box-Cox Parameter Plot (c) After Transformation

Figure 6.6: Residual Plot Before and After Square Root Transformation for the Full Model with 19 Features

regression is also used first to screen features. The feature screening process used $\alpha = 0.15$ to be the *F-to-enter* and *F-to-remove* significant level thresholds. This step selects eight features: *screen*, *budget*, *Edge_D*, *TBD7*, *CC*, *CC_S*, *Modularity*, and *Cluster_M*, while the variable *density* is also involved in the intermediate steps but eventually gets removed. The 2 best models for each different number of variables (only 1 when eight variables are all used) are listed in Table 6.8. Similar to the previous full sample regression and previous research works on box office prediction, *screen* is the most powerful predictor among all variables and can explain up to 90% variance of the data. However, this table also reveals that the network variable, *Modularity*, is another potentially strong predictive indicator ($R^2_{adj} = 80.25\%$) although *Modularity* does not fit as well as the model using *screen*. When more than one variable is used, all models keep the variable *screen* because of its significance. For models with two variables, *CC* and *CC_S* are used as the second variable alternatively. This finding supports the result in the previous analysis that connected component features are significantly contributing to the modeling of box office prediction.

Table 6.8: Best Subsets Regression for model with all features

Var	Mallows's C_P	S	R^2_{adj}	Screen	Budget	TBD7	CC	CC_S	Edge_D	Cluster_M	Modularity
1	154.8	1555.0	90.69%	×							
1	590.8	2264.8	80.25%								×
2	52.5	1333.0	93.16%	×				×			
2	52.9	1334.0	93.15%	×			×				
3	25.2	1265.1	93.84%	×	×		×				
3	26.4	1268.1	93.81%	×	×			×			
4	14.0	1234.6	94.13%	×	×		×			×	
4	15.2	1237.7	94.10%	×	×			×		×	
5	12.0	1227.1	94.20%	×	×		×			×	×
5	13.8	1231.7	94.16%	×	×			×		×	×
6	10.2	1219.8	94.27%	×	×		×		×	×	×
6	10.8	1221.3	94.26%	×	×		×	×		×	×
7	8.7	1213.2	94.33%	×	×	×	×		×	×	×
7	9.7	1216.0	94.31%	×	×		×	×	×	×	×
8	8.0	1208.9	94.37%	×	×	×	×	×	×	×	×

Table 6.9: Success Level Definition

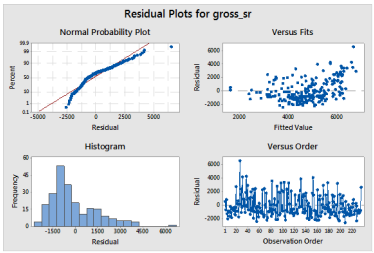
Success Level	Threshold
Failure	Revenue $<$ 0.75 Budget
Expectation	0.75 Budget $\leq$ Revenue $<$ 1.25 Budget
Making Profit	1.25 Budget $\leq$ Revenue $<$ 2 Budget
Success	Revenue $\geq$ 2 Budget

Model Comparison: Looking at the residual evaluation of the best models when adding more features (Figure 6.7b), one can observe that the model fits better when network features are used compared to the first model where only *screen* is used.

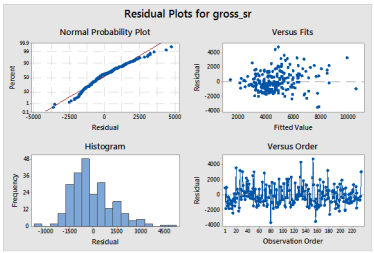
Network Features in Movie Success Level Prediction

Besides real gross value, the profit rate is another metric for movie success. There are many movies that cannot meet the cost even if their opening weekend gross is high. Therefore, it is desirable to explore whether network features are useful in predicting the success level of movies, which measures that the movie is considered a failure, expectation, making profit or a success by looking at the ratio of its total domestic revenue to its budget. To define these success levels, the thresholds in Table 6.9 are used, and the revenue indicates total domestic revenue.

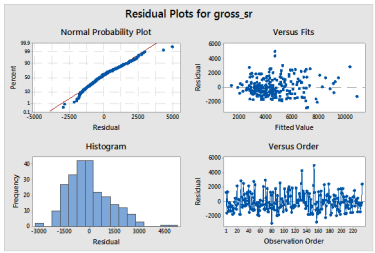
Success Level is an ordinal response which has four categories ordered from low (failure) to high (success). Ordinal logistic regression is chosen to use for Success Level prediction, and the network conversational archetypes are coded into seven binary variables. In order to find significant features in Success Level prediction, backward stepwise regression is used. The feature screening process starts with a full modeling using all features and removes one none-significant variable which has the



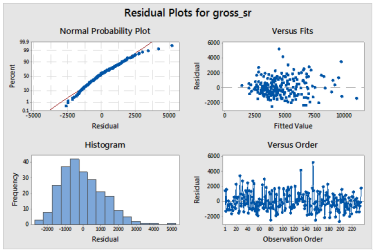
(a) Model with Screen



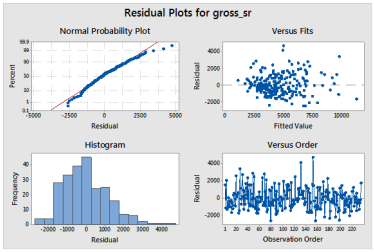
(b) Model with Screen and CC



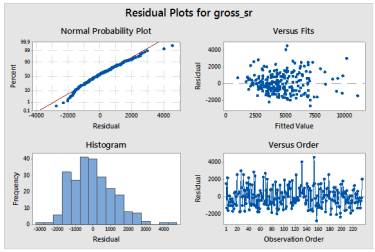
(c) Model with Screen, CC and Budget



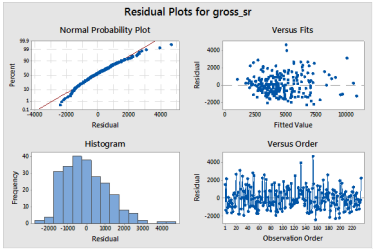
(d) Model with Screen, CC, Budget and Cluster_M



(e) Model with Screen, CC_S, Budget, Cluster_M and Modularity



(f) Model with Screen, CC_S, Budget, Cluster_M, Modularity and Edge_D



(g) Model with Screen, CC_S, Budget, Cluster_M, Modularity, Edge_D and TBD7

Figure 6.7: Residual Analysis for Model Comparison

largest p-value ($p \gg 0.05$) in each fitting until all remaining variables are significant. When the backward ordinal logistic regression stops, eight variables are left: *screen*, *budget*, *CC_M*, *Edge_D*, *Clusters*, *Cluster_10*, *Cluster_2*, and *Broadcast with Isolates*. The p-value of these variables are all smaller than 0.05 which means the they reach the significance level of 95%. The p-value of a Pearson χ^2 test equals 0.187, and the p-value of the Deviance χ^2 test is 1.0 which means the goodness-of-fit of this model is not non-significant (where not significant is the null hypothesis and get rejected). In the measures of association analysis between the response variable and the predictive probabilities, 79% data pairs are concordant and the statistic measures Somers' D = 0.58, Goodman-Kruskal Gamma = 0.59, and Kendall's Tau-a = 0.43 which means the model is adequate.

In this Success Level analysis, the value of *screen* and *budget* has also been supported. In addition, network features are also shown to be useful as significant variables. One point to state is that among the significant features, *Broadcast with Isolates* is also selected and this type of network contains a high percentage of successful movies (considering the OWG > 0.3 budget and the revenue > 1.5 budget) in the analysis in section 4.3. The estimated coefficient in the ordinal logistic regression for *Broadcast with Isolates* is positive (Coef = 2.255) which also indicates that this variable should positively contribute to the movie's success level.

Chapter 7

EVALUATING PREDICTIVE VISUAL ANALYTICS

Predictive visual analytics methods have been applied in a variety of domains ranging from healthcare, intelligence analysis, and emergency crisis management [9, 60, 61], and a great deal of research in the visual analytics community has focused on developing methods for explaining predictive models to users [11, 66, 102, 163] and enabling interactive model steering [3, 8, 64]. Ideally, by opening the black-box of data mining and machine learning algorithms, visual analytics can help users understand the reasoning behind prediction outcomes thereby improving model comprehensibility [21]. Along with opening the black-box, a variety of research methods advocate adding a human-in-the-loop component not only as a part of the analytics process to improve a user's understanding of the model, but also to enable user knowledge injection into the system [3, 5]. This intersection of human-machine analysis is seen as a critical stage in the visual analytics pipeline. Yet, this potential for adding user knowledge comes with increased risk. Inserting user knowledge into the modeling and prediction process may inherently bias the model itself [78, 164] as humans, while a wealth of contextual information, are biased in their own thought and knowledge [72, 75]. If human input is too closely tied to the model then the model may become biased in its assumptions and may, in general, become less accurate.

Such concerns are buoyed by research in the decision science field that has shown that in forecasting tasks, machine predictions consistently outperform human forecasters [69, 70, 165, 166]. In fact, work by Akes, Dawes, and Christensen [167] found that domain expertise diminished people's reliance on algorithmic forecasts which led to a worse performance. Studies have also shown that humans develop an algorithm

aversion in forecasting tasks [83, 168]. Specifically, humans quickly lose confidence in algorithmic forecasts after seeing algorithmic mistakes [169]. Given that the underlying goal of many predictive visual analytics methods is to inject domain knowledge into the analysis and point out potential algorithmic errors to the end user for updating and correction, these goals may be at odds with human behavior. As such, visual analytics could potentially contribute to algorithmic aversion during forecasting tasks and lead to reduced performance. Conversely, studies report that forecasters may desire to adjust algorithmic outputs to gain a sense of ownership of the forecasts due to a lack of trust in statistical models [18], and the goal of many predictive visual analtyics methods is to help a user develop trust in the model.

Given the conflicting demands of model accuracy, comprehensibility and trustworthiness, the question of how much human knowledge and interaction is needed or warranted in relation to the model becomes a critical question for predictive visual analytics. How much of the human is wanted in the loop of model prediction? Are humans able to accurately make predictions or outperform models with the aid of visual analytics? The goal of this study is to explore these questions. This research seeks to better understand prediction accuracy when using visual analytics support for human-in-the-loop predictions.

Inspired by work from the 2013 VAST Box Office Challenge [121–123], the goal of this study is to explore forecasting in a predictive visual analytics setting. Based on studies from management, economic sciences and psychology, it seems that algorithmic forecasts should outperform humans. However, the results in the VAST Challenge by team VADER [61] indicated that by leveraging visual analytics methods, users were able to improve the model, and a study exploring managerial intervention in sales forecasting did lead to an increase in the overall prediction accuracy [66].

As such, it is hypothesized that perhaps there is a range of algorithmic accuracy for which a human-in-the-loop forecasting process may be an optimal configuration.

To explore the hypothesis and provide an empirical evaluation of predictive visual analytics, a controlled user study was conducted to test the hypothesis that human-in-the-loop prediction, in the context of visual analytics, will outperform the computational solution given by a model at the middling level of accuracy (e.g., $.5 < R^2 < .7$). The visual analytics system used in this study is modified from a previous work [62, 170], and three prediction models varying at the goodness-of-fit were developed. Subjects were asked to predict the opening weekend gross of 9 movies in the same genre with similar levels of popularity using a box office predictive visual analytics system. The goal of this experiment is to study users' performance when making predictions with the aid of visual analytics tools under different levels of model accuracy. Critical relevant questions that this research seeks to better understand are as follows:

- Can a user (with visual analytics support) develop more accurate predictions than a middling level accuracy algorithmic model?
- How will a user respond to a predictive analytics task given black-box prediction models with different accuracy levels?
- Do users have preferences regarding the design and utilization of visual analytics tools for predictive analytics?
- Are there strategies that users implement during the analytics process that can be learned from to improve the predictive visual analytics design?

To answer these questions, this thesis designed an experiment in the context of box office prediction. Participants were required to use a web-based visual analytics system

to predict the box office gross of a movie during its opening weekend. Participants were given baseline computational models at different levels of accuracy and visual analytics tools on the movie's meta-data and Twitter data.

7.1 Hypothesis

This experiment examines the role of humans (compared to automatic models and/or pre-defined approaches) in predictive visual analytics. In this study, the hypotheses are focused on prediction performance, the relationship between a user's prediction and the model's prediction, and a user's analytic preference and patterns.

7.1.1 Prediction Performance

In many research scenarios, visual analytics use cases have shown that user's interactions can contribute to intelligence analysis, such as select features for modeling and adjust model results. With the help of users, real world prediction problems may be analyzed in a more comprehensive manner than simply relying purely on a model. Therefore, the first hypothesis is based on understanding the user's contribution to the predictive analytics process. The impact of the user should be reflected in the prediction performance when compared to using simple model's predictions.

Hypothesis 1: Participants can make better predictions than purely algorithmic models when using predictive visual analytics tools. The goal here is to further explore the conflicting results of user knowledge when applied to predictive analytics tasks. Whereas numerous studies indicate that users' predictions are generally worse than model predictions, other works have come to different conclusions [66]. Furthermore, some research has shown that users' confidence and satisfaction improve after being allowed to make changes based on a model's pre-

diction [18], but prediction outcome does not. As a result, users' contribution to prediction accuracy may be limited.

Hypothesis 2: The improvement of a user's prediction accuracy decreases when the model's prediction accuracy increases. The goal here is to explore if there seems to be a "sweet spot" for users to improve prediction accuracy. For example, if a model is 90% accurate, it is likely human intervention can only decrease the accuracy; however, if the model is 60% accurate, the integration of user knowledge could help bolster results.

7.1.2 Influence by Default Model's Prediction

In a previous research, which assessed the role of teamwork in predictive analytics [170], it was identified that participants tend to have prediction errors that are correlated with the given model. It is reasonable to expect that a participant could refer to the model's prediction as one important factor in making their own prediction.

Hypothesis 3: The model's prediction directly influences participants' decisions and this will result in users having a higher prediction accuracy when the underlying model has a higher prediction accuracy. It is expected that participants' performance will have a positive correlation with model accuracy which would indicate that the model has an influence on the participants' prediction (as opposed to seeing random predictions).

7.1.3 Participants' Behavior

Along with exploring prediction accuracy, this study can also be considered as an opportunity to analyze the role of a user in predictive visual analytics. By capturing their interaction logs, one can investigate the behavior patterns of participants to help him/her to understand what types of analytical tools users want to use during

their predictive analytics process. The following hypotheses are oriented to improve the understanding of how users behave during the predictive visual analytics process.

Hypothesis 4: Participants use easy-to-interpret visualization methods more than the complex methods.

Hypothesis 5: Participants develop a prediction strategy for repeating tasks.

7.2 Experiment Design

In order to test the above hypotheses, a variety of predictive models were developed and particularly chosen for box office prediction.

7.2.1 Dataset

The box office prediction task uses a movie's metadata and related tweets. The metadata was collected from the Internet Movie Database (IMDB). The study has used the movie's release date, genre, MPAA rating, and estimated budget. For social media data, following the data collection strategy of the 2013 VAST Box Office challenge [122], this study has collected movie related tweets from 280 movies that were widely released in the United States from January 2013 to February 2016. Tweets are crawled using the Twitter Streaming API [31] by searching the hashtag keywords extracted from each movie's official Twitter account. In this work, tweets posted two weeks prior to each movie's release date are used, totaling 13,415,382 tweets.

7.2.2 Default Models in the Experiment

Three models are established with varying degree of accuracy in order to evaluate participants' performance for varying model accuracies. The data is split into two chunks, according to their release date, for a time series validation. The estimation

data has 150 movies, and the validation data has 130 movies. All three models were fit using a square root data transformation on the response for a normal residual distribution. This results in the response for the model being the square root of the opening weekend gross.

The first model uses budget and the average tweet rate for 7 days (tbd7) as regressors.

$$\textbf{Model 1: } gross_{sr} = \beta_0 + \beta_1 budget + \beta_2 tbd7$$

The goodness of fit measures are that $R^2 = 60.73\%, R^2_{adj} = 60.20\%, R^2_{pred} = 55.33\%$ and the mean squared error on the estimation data is 2,172,086. Applying this model on the validation data, the mean squared error is 2,903,916.

The second model uses budget, tbd7, and the number of theaters a movie is released in (screen).

$$\textbf{Model 2: } gross_{sr} = \beta_0 + \beta_1 budget + \beta_2 tbd7 + \beta_3 screen.$$

The goodness of fit measures are that $R^2 = 69.28\%, R^2_{adj} = 68.65\%, R^2_{pred} = 64.89\%$ and the mean squared error on the estimation data is 1,710,968. Applying this model on the validation data, the mean squared error is 2,125,388.

The third model uses budget, tbd7, screen, and the interaction of budget and screen (budget×screen).

$$\textbf{Model 3: } gross_{sr} = \beta_0 + \beta_1 budget + \beta_2 tbd7 + \beta_3 screen + \beta_3 budgetscreen.$$

The goodness of fit measures are that $R^2 = 75.07\%, R^2_{adj} = 74.33\%, R^2_{pred} = 72.13\%$ and the mean squared error on the estimation data is 1,398,258. Applying this model on the validation data, the mean squared error is 2,083,217.

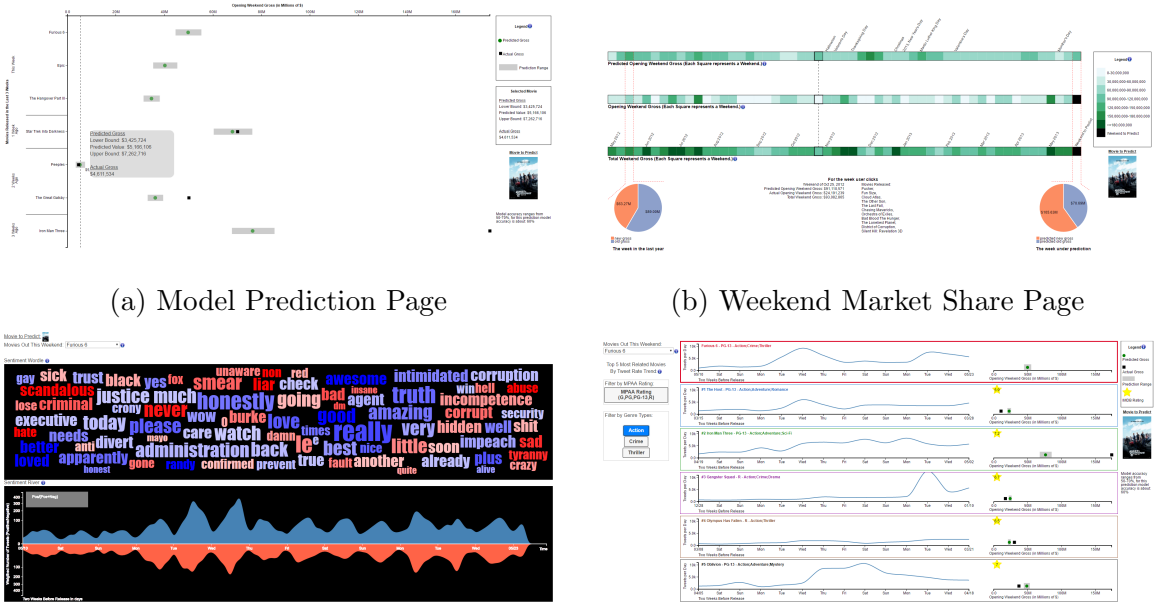
7.2.3 Other Factors

Beyond the model, there are many other factors that might impact participants' performance. Based on the experience from previous studies, this study controls specific factors to be held-constant and randomize other factors, while leaving some to vary.

User bias: In the work by Buchanan et al. [170], participants' collaboration and communication styles have been shown to be an important factor in prediction accuracy during teamwork, and personal bias has been studied in judgmental adjustments [78]. Although this study is only focused on individual participants, varying individual factors should not be ignored. To mitigate user bias, this study narrowed down the user portfolio to be full-time undergraduate and master students with small variance in age. In addition, this study increased the number of movies one participant predicts from 3 movies in the previous work [170] to 9 movies as additional data tends to average out randomness.

Movie bias: A movie, as the object of this prediction task, might also impact the performance and variance of different participants' predictions. In the previous study [170], a user's familiarity with the movie greatly impacted their perception and subsequent prediction procedures. The movie genre (the type of movie, such as action or romance) is another crucial factor. In a previous study, participants were unfamiliar with the movie *About Last Night*. In turn, this resulted in participants having a difficult time analyzing the movie. In order to mitigate such movie bias, well-known movies based on a high Tweet count were selected particularly. Additionally, the all 9 movies were selected from the same genre (Action).

Order bias: For repeating experiments, a concern is that the participant may becomes familiar with the system and the prediction task and perform differently



(a) Model Prediction Page

(b) Weekend Market Share Page

(c) Sentiment Analysis Page

(d) Movie Similarity Page

Figure 7.1: The Four Main Visual Components in the Experiment for Data Exploration and Predictive Analytics

according to the occurrence order of these movies. To avoid such bias, the order was randomized for the movies, as well as the models.

7.2.4 Interface Design

In the VAST box office challenge, different visual analytics systems were proposed from several teams [121, 122, 171]. Common among these systems were visual analytics tools for sentiment classification to analyze movie related tweets where detailed text was accessible through interactions. Prediction models (e.g, linear regression [122], SVM [121], and weighted average [171]) are also used to assist the analytics in the three systems. Major design differences occurred at the predictive model interaction level. Kruger et al. [121] focused more on developing a complex prediction model and relied on the model results greatly. Al-Masoudi et al. [171]

relied on user refinement of selecting historical movies to be the training dataset for a simple weighted average model. Lu et al. [122] combined two prediction models and the visual exploration of user-defined similar movies for the prediction. This evaluation experiment utilizes the previously developed visual analytics framework (where the visual designs were comparable to other teams participating in the contest) for box office analysis and prediction [170] that has been modified to: (1) facilitate the recording and implementation of the experiment, and; (2) provide a convenient way for analysts to explore and analyze the data.

Visual Interfaces

This visual analytics system consists of six visual components: Homepage, Model Prediction, Weekend Market Share, Sentiment Analysis, Movie Similarity, and Make Prediction. The system uses pre-processing to extract useful information from large-scale, noisy, and unstructured social media data. These data are then integrated into the visual analytics system so that the participants can easily use the abstracted information. The numerical and nominal features used in this study are shown in Table 7.1 and they are extracted from IMDB and Twitter.

The *Homepage* shows basic information for the weekend under prediction and a tutorial on using the system. It contains the date of the weekend under analysis and a brief introduction to the released movies on that weekend, so that the participant has a general context for the movie under prediction. A tutorial can be opened from the Homepage for a detailed introduction of each function in the system. The Make Prediction page allows the participant to submit his or her final prediction. This page also lists the prediction of the current default model and the weekend market share model. These two predictions can be referred to by the participant while analyzing the data and making a decision. The other four pages are the main exploration

Table 7.1: Variables Used in the Experimental Visual Interface

Variable	Description
Gross	3-day Opening Weekend Gross
Budget	Approximate movie budget from IMDB (unit is "million" of dollars)
Genre	The movie's genre(s) according to IMDB
MPAA	The movie's MPAA rating according to IMDB
TBD	The average daily number of Tweets over the 2 weeks prior to release
TSS	Tweet Sentiment Score calculated via SentiWordNet [139]
MSS	Movie Sentiment Score – A derivation of the overall sentiment of a movie

interfaces and the participants can move between these pages freely to interactively explore the data.

The *Model Prediction* page (Figure 7.1a) shows the single movie prediction model's results and performance for the previous three weeks and its prediction for the current week. These prediction values are given by one of the three default models randomly selected for the current prediction, and the basic performance of this model is shown on the bottom right corner of the page. This page uses a scatter plot with prediction ranges where the x-axis is the revenue value and the y-axis lists the movies ordered by their release date. Solid lines are used to separate movies by their release week and the top movies are those released in the current week under prediction. The green circle and a surrounding gray bar show the single movie model's prediction value and its 95% confidence range, whereas the black squares indicate the actual opening weekend gross for previous movies. Mousing over the scatter plot shows a dashed line referring to the opening weekend gross axis. To look at the exact values of a movie,

the participant can simply click on the green circle to open a pop-up context window and fill the content of the selected movie information on the right-hand side.

The *Weekend Market Share* page (Figure 7.1b) visualizes the result from the temporal model of weekend market prediction for the participants to identify seasonal patterns in the movie industry. This page has three horizontal bars where each one consists of 55 squares covering a whole year's weekends prior to the weekend under prediction. The revenue of each weekend is shown by the color of its square according to a sequential color scale where the light color means low revenue and the dark color means high revenue. The three horizontal bars correspond to the temporal model's prediction, the real value of the sum of all newly released movies, and the real total weekend gross of all currently playing movies. Mousing over these squares will line up the weekends from these three bars and clicking on the square can highlight this weekend and display the details of the released movies and the revenue of that weekend. To highlight special days, holidays are listed on top. The exact values of the temporal model's prediction are visualized as pie charts for the current weekend and the corresponding weekend in the last year for a convenient comparison.

The *Sentiment Analysis* page (Figure 7.1c) consists of a sentiment wordle and a sentiment river plot [61]. By exploring these plots, participants can determine the most frequent positive and negative words regarding a movie and perceive a general understanding of the public's review/expectation of a movie. Positive sentiment is shown as blue and negative sentiment is shown as red. Participants can switch between movies released on the same weekend to compare sentiment.

The visual analytics environment also supports the comparison of features between movies in the Movie Similarity page (Figure 7.1d). Similar movies can be filtered by selecting the movie's MPAA rating, and its genre(s). Once a metric is selected, the movies are first filtered by these metrics and the five most similar movies ordered by

tweet volume trend are displayed. The user can compare their social media trends as well as the regression model's predictions for each movie. The left side of this page lists the options of filtering for similar movies and the right side uses small multiple views to show the five most similar movies and the current movie under prediction (the top one). The view of each movie contains a line chart of the tweet volume trend and the single movie model's prediction.

System Utilities and Data Recording

In addition to supporting interactive analytics of the social media data, movie meta data, and the given prediction models, the visual analytics system is also designed to facilitate tutoring participants, guiding the experiment's process, recording prediction results, and saving interactions and durations for each user. The framework begins with a tutorial, which is followed by a practice prediction, and then the nine experimental predictions. Once started, a new record is created in the database to save the experimental information for the current participant. The system will automatically move to the next prediction after each submission and record the duration time for each prediction and all prediction results after the entire process is complete. In addition, the system records the analytics interactions, such as changing visual interface, clicking on a particular weekend, filtering for similar movies, etc.

7.3 User Study

This study recruited 20 participants (18 males and 2 females) for this experiment. Among the 20 participants, 3 are college undergraduate students and 17 are masters students. The average age is 23.25, ranging from 21 to 26. Each participant participated in a training session, a practice prediction session, and nine prediction sessions. Their workload was evaluated twice using the NASA Task Load Index (NASA TLX)

measure; once after the practice session and once after the last prediction was cast. Finally, the participants completed a demographic questionnaire about their background, knowledge of predictive analytics, computer usage, and familiarity with the movies presented.

Training

Each participant was trained at the start of the experiment. Training was presented using PowerPoint slides that covered the purpose of the study (to accurately predict movie's opening weekend gross), how to navigate through the visual analytics system, what information is available to them, and why this information is useful when predicting opening weekend gross. The Disney movie, Frozen, was selected to illustrate all points made during training. A short quiz was administered after the training (10 questions, 5 minutes). The quiz was immediately reviewed by the experimenter and participants were encouraged to ask questions about the tested material and the movie interface. At the same time, the researcher provided feedback on the recently tested material.

After taking the quiz, the researcher loaded the visual analytics interface, logged in each participant, and prepared the system for the prediction session. Participants were instructed to access the system "Tutorial" page in order to make sure they know where it is located. The training PowerPoint and the "calculator" were also opened for participants' reference. A reference sheet with a list of box office related concepts (e.g., MPAA rating, Opening Weekend Gross, etc.), scratch paper (for notes or performing calculations), and a prediction sheet to write down their final prediction for each movie.

Predictions

Exploration and predictions consisted of 10 chunks, a practice session and nine real predictions. The participants started with a 15-minute practice session. During the practice session participants were given the movie Fast & Furious 6 to explore. The default prediction model is randomly selected from the three models. An embedded timer is shown in the system and participants were encouraged to finalize their prediction when 10 minutes had elapsed. However, the system was designed so that participants could take longer to cast their prediction if they needed to. Participants were allowed to take longer because the primary focus of this study was to understand the human participation during the predictive visual analytics task. Participants were also encouraged to ask questions during and after the practice session to ensure comprehension of the task and interface. Once the participants completed the practice session they were given a 10-minute break.

After the break, the participants completed nine real predictions. During each session, participants were able to cast their prediction before the 15 minutes were up. However, participants were given a maximum time limit of 20 minutes to cast their prediction. The following movies were presented to the participants in a random order: 300: Rise of An Empire, The Amazing Spider-Man 2, Maleficent, Transformers: Age of Extinction, The Equalizer, Kingsman: The Secret Service, San Andreas, Terminator Genisys, and Mission: Impossible - Rogue Nation. Their written predictions are collected while the system also record their input values into the system and their actions and duration during the exploration for predictions. For each movie presented to the participant, the regression model being used in the visual analytics system was randomized, where each model was presented three times.

7.3.1 Questionnaires

Questionnaires

To evaluate the participants' mental workload, a NASA TLX was administered twice during the experiment, the first time after the practice session (the first prediction of movie Fast & Furious 6) and the second time after the last prediction. It was expected that participants' workload reduced over time due to familiarity with the task and system. A demographics questionnaire was also administered at the end of the experiment to evaluate the participants' background (age, gender, education), domain knowledge (movie familiarity, frequency of "going to the movies"), social media usage (frequency), and knowledge of predictive analytics (familiarity with mathematical models). Finally, participants were also asked a free-response question aimed to assess how they analyzed the data, as well as, whether they had used a particular strategy to analyze the data.

7.4 Experiment Result and Analysis

The practice prediction is excluded from the analysis because participants were able to ask questions about the prediction task during that time. As such, the analysis uses the predictions made for the 9 movies. A total of 180 predictions were analyzed. In addition, participants' actions (e.g., what interface features they accessed) and duration for each action taken (e.g. how long they viewed a particular data set) were recorded. The demographics questionnaire and the NASA TLX evaluations were used to gain a deeper understanding of how participants worked through the analysis tasks.

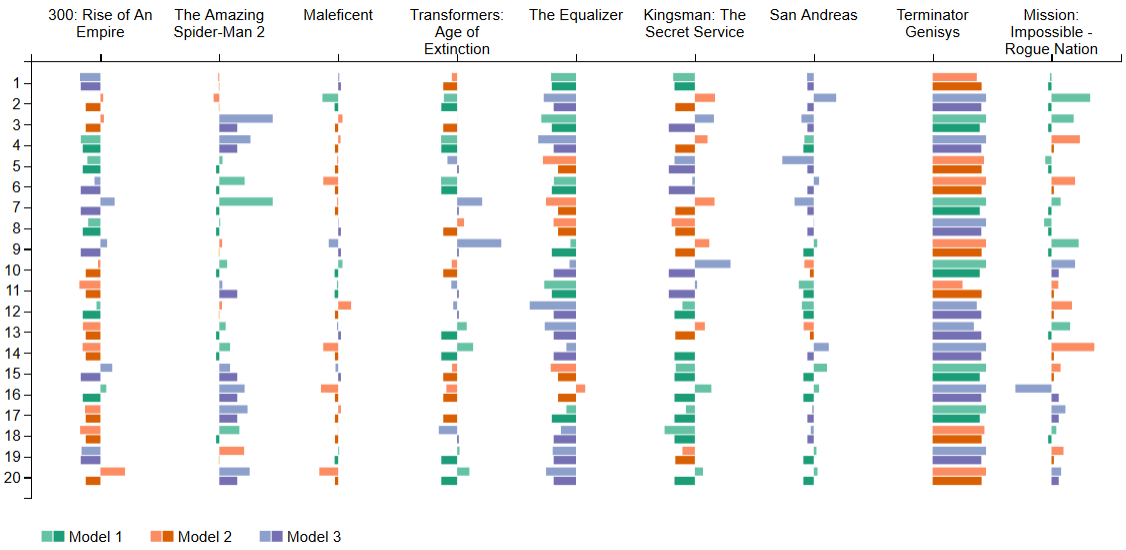


Figure 7.2: The RAE Comparison Display for Every Prediction

7.4.1 Prediction Performance

To test hypothesis 1, the participants' predictions were first compared to the model predictions. To test hypothesis 2, a two part process was used. First, the participants' predictions were grouped by the default model they used. Second, the participants' predictions among the different model groups were compared. The relative absolute error (RAE), which is the percentage bias deviating from the real value, is used to measure each prediction's accuracy.

$$RAE = \frac{|Prediction - RealValue|}{RealValue} \tag{7.1}$$

Figure 7.2 displays the RAE for each participant prediction and the model he/she was using. This display compares the RAE of each prediction between the participants and the default model. The x-axis lists the 9 real movie predictions. The y-axis lists the participant ids. Each cell has two bars displaying the error where overestimation is to the right-side and underestimation is to the left-side, and both bars are aligned in the middle. The top bar uses a lighter hue and represents the participant's prediction

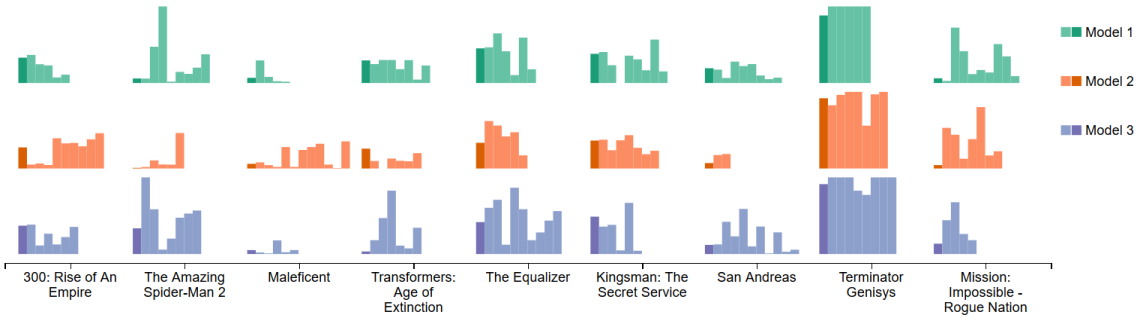


Figure 7.3: The RAE Display of Each Prediction Organized by Movie and Model

error, and the bottom bar uses a darker hue and represents the model's prediction error. The bar length is scaled by the error with a cutoff at 1.The shorter the bar, the more accurate the prediction. Among the 180 predictions, 72 (40%) have a lower RAE than the corresponding model while 108 (60%) have a higher RAE. To analyze the predictions in terms of different movies and different models, Figure 7.3 displays the RAE for each model's prediction and each participant prediction organized by movies and models (9 movies $\times$ 3 models). The plot indicates that participant predictions vary a great deal. For instance, some participants beat the model prediction whereas over 50% of the participants' error was larger than the model predictions. This was seen across all three models.

A one-way ANOVA performed on these data revealed that there was a significant difference between participants' prediction and overall model's prediction with a p-value = 0.009. Test results are given in Table 7.2. The standard deviation of the participants' RAE is 0.4639 whereas the model's RAE is 0.2589 indicating a larger error variance for participants' prediction outcomes. Next, a one-way ANOVA was performed on each set of predictions for Model 1, Model 2, and Model 3. The following rows in Table 7.2 show the result. Participants' predictions were not significantly different when using Model 1, Model 2, or Model 3 predictions with an $\alpha = 0.05$ (all

Table 7.2: Results of Four One-Way ANOVA Tests with Participant Prediction RAE and Model Prediction RAE as Responses and Equal Mean as the Null Hypothesis

	Model MRAE	Participant MRAE	P-Value	95% CI of Model	95% CI of Participant
All Three Models	0.2950	0.3994	0.009	(0.2399, 0.3501)	(0.3442, 0.4544)
Model 1	0.2793	0.3608	0.214	(0.1880, 0.3707)	(0.2694, 0.4521)
Model 2	0.2687	0.3801	0.131	(0.1661, 0.3713)	(0.2775, 0.4827)
Model 3	0.3370	0.4572	0.076	(0.2429, 0.4310)	(0.3631, 0.5513)

three p-values are greater than 0.05), and the result still clearly illustrates that participants do not outperform the models with respect to predictive accuracy. **Therefore, the first hypothesis that participants supported by a visual analytics interface can outperform algorithmic models in terms of accuracy, is not supported.** Instead, this experiment result showed the opposite, which was also shown in other studies [18, 71, 166]. The model used in this movie prediction task had a higher prediction accuracy than the average participant and was better than 60% of the participants' single predictions.

To test the second hypothesis, users' contribution decreases when model's accuracy increases, the results were further analyzed by comparing the difference between participants' predictions and the model predictions across different models. This analysis shows that, in Table 7.2, participants' performance was worse when using Model 3 (the most accurate model) and the best performance when they used Model

1 (the least accurate model). This difference is indicated by the MRAE values (mean RAE). ANOVA tests were performed with model RAE and Participant RAE being the response variable and model type as the factor. The results revealed that Model 3 has the largest RAE and Model 2 the smallest RAE. However, the p-value = 0.487 indicating no significant difference in performance when using the different models. Tests on the participants' RAE also shows that Model 3 has the largest RAE, followed by Model 1, and then Model 2. This difference is also not statistically significant with a p-value = 0.3. Comparing the MRAE values between the participants and the models, one can see that the participants' prediction accuracy increased from Model 3 to Model 1, but not with Model 2. These results indicate that the participants tended to perform better when the model performance improves from low accuracy to average accuracy, but not with the higher accuracy. This shows that using a more accurate reference prediction value may not necessarily improve participants' prediction. Though, this difference is not statistically significant and indicates only a slight trend. As such, **the hypothesis that a user's prediction accuracy will decrease as the model's prediction accuracy increases is also not supported**. However, the underlying trend that was discovered indicates further studies in this direction are needed to determine if increasing the levels of model accuracy tested will give insight into a threshold at which user interaction further degrades the prediction accuracy.

7.4.2 Influence by Default Model's Prediction

Each model has an overall accuracy; however, for any given movie the model may overperform or underperform, thus biasing the experiment. In this study, Model 3 had the worst performance for each of the 9 selected movies even though it was the most accurate one over all 280 movies. This occurred due to the constraints that were used to select the movies. Likewise, similar situations can arise in real world

applications. For instance, using a general model to predict a specific instance can lead stakeholders astray because the model is poorly fit.

The difference between the participant predictions and the corresponding model predictions is analyzed using one-way ANOVA test with model type being the factor. This difference indicates that participants tend to predict closer to the model's prediction when Model 2 was being used, although the difference is not statistically significant (p-value = 0.69). This is striking given the fact that participants' were told the accuracy of the model being used for each experiment and would have knowledge that Model 3 is the most accurate. What this indicates is that participants did not exhibit a higher degree of trust in a model based on reported accuracy. Furthermore, there is a decrease of the prediction difference from Model 1 to Model 2 as model accuracy improves. **Therefore hypothesis 3, participants prediction is influenced more by the model's prediction when the accuracy of the model increase, is not supported.** One possible reason that participants chose not to base their predictions on Model 3 may have been because they discovered that the model was actually inaccurate for this particular set of movies. Consequently, the participants may have been able to use the system, along with their own knowledge and intuition to assess this discrepancy. As such, further experiments focusing specifically on trustworthiness should be explored.

7.4.3 Participants' Behavior Analysis

This study has collected participants' answers from the demographics questionnaire and the NASA TLX. Additionally, the system recorded interface interactions using the system's log. These data was used to evaluate participants' behavior during their analysis process.

Demographics Questionnaire Analysis

In order to gain a deeper understanding into which interface features were used and why, a series of questions were asked regarding the ease/difficulty of use, frequency, and ranking of interface features. Among the 20 participants, 19 completed the entire questionnaire while 1 participant left some questions blank. Of the 19 participants that answered, 16 of them marked the most used interface feature as the easiest one to use. Overall, 17 participants marked the Sentiment Analysis Page as easiest to use/most understandable interface feature. This result is also reflected when assessing all 68 marked answers (some participants ranked more than 3 interfaces features). Again, the most frequently used interface features was also selected as the easiest to use/understand (49 marked responses, 72%). On the other hand, 9 participants marked the Weekend Market Share and Movie Similarity Page as the least usable and understandable interface. These two interface features were also the two least used ones. **These results support the fourth hypothesis which stated that participants use easy-to-interpret visual analytics tools more than complex tools.**

Finally, understandability and ease of use does not necessarily determine the usage of the computer-interface for all participants. For instance, out of the 9 participants that marked the Weekend Market Share Page as difficult to use, four ended up still using the Weekend Market Share data during their prediction tasks. Similarly, out of the 9 participants who marked the Movie Similarity Page as difficult, only one ended up not using the data presented on this page. As a result, it can be seen that difficulty is not the only determining factor of whether certain views are considered during predictive visual analytics tasks.

One interesting finding that emerged was that though no significant difference between participants' prediction performance was found, background knowledge seemed to play an important role in the accuracy of the prediction outcome. For example, participant No. 12, who had the lowest MRAE for all 9 predictions, described his strategy as follows: "I read a lot about movies and have heard of almost all of these movies. I recall descriptions about how people liked these movies and used the model as a starting point. I notice the model was better at predicting movies that were considered average and movies that are considered good to do better than the model." Again, anecdotal evidence from the participants seems to indicate that user knowledge can improve results; however, in this study, as well as others, the experimental results do not support such an anecdote. Furthermore, it has been tested if participants have significant differences among themselves in terms of their prediction performance and the influence of their performance by their familiarity to the movies (based on their own self-assessment measure). A t-test of the prediction errors between familiar movies and non familiar movies was performed and no significant difference was found (p-value = 0.753) indicating that movie familiarity (which could translate to user knowledge) does not improve prediction

System Log Analysis

To further explore how participants approached making a prediction, this system recorded the participants' interactions and the amount of time spent on each interface feature via timestamps. Out of the 20 participants, one encountered repeated system problems causing the experiment to be paused several times. Therefore, this participant's log was removed from the analysis. In the experiment, participants were allowed to move at their own pace. They were able to cast their prediction before the allotted time was up and after. However, participants were asked not to exceed

20 minutes per prediction task. For each prediction session, the average time-on-task was 7 minutes with a median time of 6 minutes. The shortest session lasted only 0.5 minutes and the longest one 16 minutes. Interestingly enough, the data indicates that there is no correlation between time-on-task and participants' prediction error (Pearson correlation = 0.155, P-value = 0.043). The ANOVA on these data also revealed that there is no significant difference between movie under prediction and time-on-task (P-value = 0.631). However, there is a significant difference between time-on-task across the 20 participants (p-value = 0.00). Hence, time-on-task seems to be a personal factor. Some participants may have needed more time because they were unfamiliar with the movies or did not understand the task or interface. Yet, other participants may have needed less time for similar reasons. Lastly, it is likely that additional factors contributed to the variation for time-on-task. For example, additional factors such as personal commitment and motivational states may have impacted participants' time commitment during their analysis and future studies should include such factors in the analysis.

The analysis of the log data revealed a significant (p-value = 0.000) difference of the mean time-usage among the six interface tabs. This significant (p-value = 0.004) difference remains even if time-usage is only compared for the four analytics tabs (Make Prediction, Weekend Market Share, Sentiment Analysis, and Movie Similarity). Specifically, log data analysis indicates that participants spend most time looking at the Model Prediction and Movie Similarity data. For example, average time spend on the Movie Similarity Page was 1.523 minutes and 1.496 minutes for the Model Prediction Page. In addition, participants spend an average of 1.196 minutes on the Sentiment Analysis Page and 1.107 minutes on the Weekend Market Share Page. These findings match the previously reported results. Specifically, participants reported that they found the Weekend Market Share Page as most difficult and least

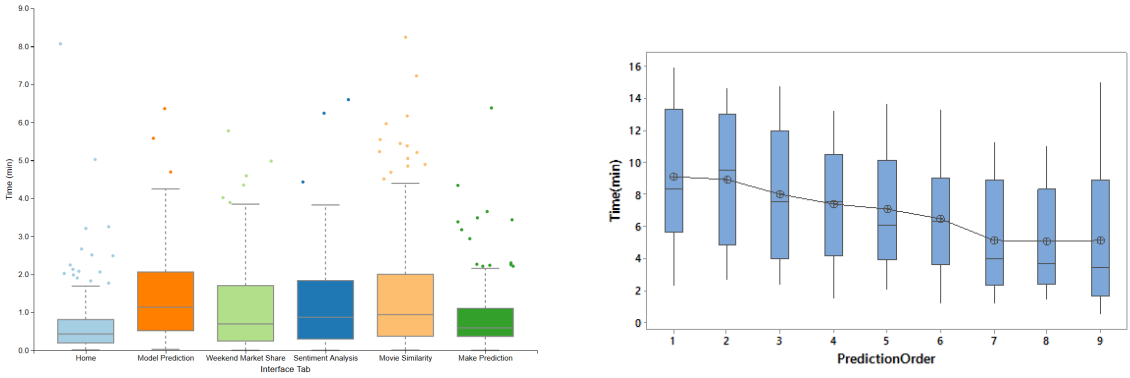
used interface feature. Yet, they still spend a considerable amount of time analyzing the data presented here. Consequently, participants were willing to invest time to use a relative difficult page; presumably this occurred because participants viewed the data in the Weekend Market Share Page as significant. Lastly, these results illustrate a key point – no matter how crucial certain analytical features are, users will not invest the same amount of time purely based on criticality, but will consider ease of use first.

In a next analysis whether time-on-task decreased as prediction session increased was assessed. Figure 7.4b depicts the box plot for time-on-task by sessions. The results show that there is a significant (p-value = 0.002) decrease on time-on-task as sessions increase. Yet, participants' prediction error did not increase with reduced time-on-task indicating that repeated predictive analytics tasks leads to familiarity of the interface and data presented.

This analysis also analyzed how many times participants switched between the different interfaces. For the 19 participants a total number of 171 prediction sessions were evaluated. The results indicate that the number of switches between interfaces varied greatly among participants, ranging from 4 to 77, with mean = 16.89 and median = 12. However, further analysis showed no significant relationship between the number of switches to the number of prediction errors.Considering the wide range of number of times participants switched between interface pages it seems that this tendency may be due to personal preference during data analysis.

NASA TLX Analysis

The analysis on the participants' NASA TLX responses revealed that there were no significant differences in difficulty and task demand ratings between the two time frames (after the practice session and after the last prediction). That is, participants



(a) The Box Plot of the Time Participants Used in Each Interface Tab

(b) The Box Plot of Time-on-Task per Prediction Task

Figure 7.4: Time-on-Task Analysis Plots

did not rate the prediction task as becoming easier over time. There were also no significant differences in the participants' emotional states between the two time frames (all p-values were well above p=.05). The NASA TLX results show that participants felt the same, as well as, viewed task difficulty the same throughout the entire experiment. However, ideally participants should have rated the task as easier as time went on, partly because they were becoming more familiar with it. Yet, participants' mental workload seemed to have remained high for the entire experimental session, indicating the complexity of the prediction task.

7.5 Discussion

Work in the visual analytics community has demonstrated that experts could benefit from tools that support the integration of domain knowledge with interactive visual exploration. In terms of predictive analytics, visual analytic techniques have also been used to help improve the comprehension of data, model, and prediction results, and visual analytics techniques that enable users to interact within

the predictive modeling process have also reported benefits. However, the social and management science communities have deliberated for decades as to whether or not humans can improve algorithmic prediction. Unfortunately, the literature does not provide a clear answer given that conflicting results (that humans improve the accuracy and that humans decrease the accuracy) have been reported. In recent years, researchers have also turned their focus to examining factors that may influence human prediction outcomes and which factors contribute to lower prediction outcomes. However, controlled experimental studies are relatively rare within the domain of visual analytics. Therefore, there is a need to systematically study how humans can improve predictive visual analytics outcomes. Specifically, in this study it examined how humans perform over a range of model accuracies, the goal was to explore if there was a model accuracy after which human returns diminished. Additionally, interactions with the predictive analytics system were also studied in order to understand what interface features were considered useful and which ones were not.

Aligned with the research from the Wharton School [18], the results proposed in this thesis also indicate that human predictions were significantly worse than model predictions. Although the hypotheses were not supported (i.e., a human-in-the-loop did not lead to better predictions), this research study remains valuable as it is the first (to the best of the author's knowledge) controlled user study for evaluating human participants during a predictive visual analytics task. Yet, numerous questions remain to be answered. For instance, how to better prepare the user for this task? Another question that would be worth exploring is, what type of knowledge is more beneficial to have, knowledge of predictive models or domain knowledge (e.g., knowledge of movies)? In this study, Participant 12 had an extensive knowledge of movies and, in turn, had the lowest error rate. Another avenue worth exploring is how can the usability of predictive tools themselves be improved (e.g., interfaces)? From the

results here it was clear that users preferred easy-to-use visual analytical interfaces. Furthermore, the participants may have performed worse because they shied away from the more difficult data sets/interface features.

It is important to note that this study is not without limitations. First, evaluating human participation in predictive visual analytics is breaking new ground. Hence, it is still being explored for effective predictive visual analytic system designs and methods of conducting controlled experiments where a complex cognitive process is involved. Second, it is possible that the models used in this study were not accurate enough for users to trust them. Third, different types of predictive visual analytics procedures need to be compared. For example, future experiments should allow participants to modify the models themselves. This interaction may foster a greater understanding and trust for the models. The added benefit of this approach is that one can investigate which predictive visual analytics approach leads to improved prediction results both in terms of accuracy and user satisfaction.

Chapter 8

CONCLUSION AND FUTURE WORK

In summary, this work has proposed and evaluated a framework of integrating social media and predictive analytics, and studied various predictive visual analytics methodologies. This works major contributions are as follows:

1. A predictive visual analytics pipeline.
2. An integrative multi-source predictive visual analytics implementation.
3. A predictive visual analytics approach linking sentiment analysis tools, similarity metrics, and regression models.
4. A predictive visual analytics framework derived from the Delphi method.
5. A visualization-centric categorization of social media network types.
6. An experiment to evaluate human participation in predictive visual analytics.

The effectiveness of the predictive visual analytics methodologies proposed in this thesis was demonstrated through a series of case studies and user studies. These studies focused on the practical problem of box-office prediction. Results indicate that predictive visual analytics technologies are effective for integrating multiple knowledge sources and making predictions: users with background knowledge could better predict box office revenue given a proper set of visual analytics tools and a suggested analytics approach. However, a follow-up user study that evaluated the human effect of predictive visual analytics revealed that users were liable to introduce bias into the prediction process. Users attitudes toward, and preferences regarding, the use

of visual analytics tools were analyzed to guide future design decisions. Given these finding, I will direct my future research efforts to target the following areas:

1. Visualization designs for explainable AI.
2. HCI aspects of predictive visual analytics.
3. Methodologies to broaden the scope of prediction.
4. Comprehensive evaluations of predictive visual analytics.

8.1 Explainable AI (XAI)

Data and computational resources are rapidly becoming more widely available. A central trend in modeling has been to employ increasingly large and sophisticated models that exploit these. This trend is typified by "deep learning" approaches that apply large scale network models to modeling problems such as prediction. Such approaches are often able to leverage large data to achieve impressive performance. However, this performance is not without a cost: such models are large, complex, and constructed automatically (especially in terms of feature engineering), making them difficult to interpret. While the predictive results may be accurate, if the generated model lacks interpretable meaning, then its predictive power is hampered [82]. Interpretability is an important concern whenever AI techniques are utilized, and this problem is exacerbated with the emergence of deep models. The challenges of interpreting complex models are often referred to as *explainable AI* (or XAI for short). Interpretable models can serve many goals for a variety of stakeholders [21].

An example in the requirement of explainable AI is the self-driving car. Google's self-driving car project utilizes machine learning in order to generate models that can accurately process and respond to input from its sensors [172]. The self-driving

cars have now logged over 2 million miles on public roads with only a couple dozen accidents, only one of which was caused by the autonomous vehicle [173]. This is an impressive safety record, but given the complexity of input and response the cars need to handle, it cannot be known if the cars will respond well in every situation. This is a prime example of an accurate predictive model that lacks interpretability in a domain where the interpretability of the model is of grave importance.

There has been a perceived trade-off between model interpretability and performance, however there may be other pathways to improving interpretability besides using simpler models with poorer performance. In terms of self-driving cars, it is conceivable that a better safety record would be traded for a simpler model that makes it easier to draft legislation and comply to regulations concerning autonomous vehicles [174], however it is preferable to have both safety and comprehensibility. Research in explainable AI has explored approaches including generating descriptions of complex models and for interpreting complex models through a series of simpler ones (e.g., LIME [175]).

8.2 Interactions with Predictive Visual Analytics System

In predictive visual analytics, interaction is the means by which experts explore data, steer the model, and integrate their knowledge into the prediction process. Understanding how experts interact with the system is critical for developing efficient and effective predictive visual analytics systems. Many advanced interaction techniques have been developed and used in visualization systems. Semantic interactions have been a hot topic recently in the visual analytics literature [176–178]. By allowing users to directly manipulate data in the visualization space, updates to the positions of data elements on the display can be tied back to weights in the ana-

lytic modules, which can then be translated to model updates. This is also called visual-to-parametric interaction.

Little work has been done to understand the reasoning process of the expert when they interact with the modeling process. Thus, I plan to analyze the participants' interactions through logging the timestamps and actions they perform on the system. Having analyzed such log data, one could learn which views have been used most and what are the procedures of using different views and tools; furthermore, one could learn whether the analysts were confirming their findings or exploring other evidence. The goal is to retrieve predictive analytics procedures and strategies of the analysts (if any) and further improve the predictive visual analytics toolkits. Knowing what kind of visualization tools have been used most and why participants use some tools more than others could help us improve the usability of visual analytics systems. Also, understanding the impact of different views on prediction performance could help us and develop more effective tools for predictive analytics.

8.3 Generalization on Prediction Tasks

In past research, the predictive visual analytics framework was employed and evaluated on box office prediction. However, the approach was not generalized to other prediction problems, such as stock market prediction and student performance prediction. The advantages of using box office prediction are twofold: 1) It is easy to collect data on real box office to evaluate performance, and 2) it is easy to find participants with some knowledge of movies who can quickly understand the relevant concepts in predictive visual analytics. However, research on predictive visual analytics should be applied to more applications, and similar to the user study in chapter 7 more controlled experiments should be conducted to evaluate the effects of predictive visual analytics. Thus, I plan to develop predictive visual analytics tools

for other applications, such as stock price prediction, housing and income prediction, and school performance prediction. In addition to real world problems, I also plan to generate synthetic datasets or use well-established datasets to evaluate the efficacy of predictive visual analytics so that fewer factors are uncontrolled.

8.4 Evaluations

Currently, only a preliminary evaluation has been done to investigate the role of humans in predictive visual analytics box office prediction being a proxy. However, many factors have not been tested yet and many other aspects and situations should be analyzed. In the future, I plan to extend the evaluation of predictive visual analytics to the following two aspects: First, more advanced and accurate models should be used as the default model to test participants trust and reliance on different models. Second, different interactions on developing the prediction should be considered and compared. For example, I want to conduct an experiment in which I allow some participants to modify the prediction value through data explorations and allow other participants to develop their own models to make predictions. In this way, in addition to investigating the contribution from humans to predictive analytics, one could also investigate which predictive visual analytics approach has better performance.

REFERENCES

[1] F. Heimerl, S. Koch, H. Bosch, and T. Ertl, "Visual classifier training for text document retrieval," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2839–2848, 2012.

[2] J. Choo, H. Lee, J. Kihm, and H. Park, "ivisclassifier: An interactive visual analytics system for classification based on supervised dimension reduction," in *IEEE Symposium on Visual Analytics Science and Technology*, Oct 2010, pp. 27–34.

[3] M. S. Hossain, P. K. R. Ojili, C. Grimm, R. Müller, L. T. Watson, and N. Ramakrishnan, "Scatter/gather clustering: Flexibly incorporating user feedback to steer clustering results," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2829–2838, Dec 2012.

[4] J. Krause, A. Perer, and H. Stavropoulos, "Supporting iterative cohort construction with visual temporal queries," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 91–100, 2016.

[5] J. Krause, A. Perer, and E. Bertini, "Infuse: interactive feature selection for predictive modeling of high dimensional data," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 1614–1623, 2014.

[6] T. Muhlbacher and H. Piringer, "A partition-based framework for building and validating regression models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 1962–1971, 2013.

[7] T. May, A. Bannach, J. Davey, T. Ruppert, and J. Kohlhammer, "Guiding feature subset selection with an interactive visualization," in *IEEE Conference on Visual Analytics Science and Technology*. IEEE, 2011, pp. 111–120.

[8] S. Van Den Elzen and J. J. van Wijk, "Baobabview: Interactive construction and analysis of decision trees," in *IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2011, pp. 151–160.

[9] S. Afzal, R. Maciejewski, and D. S. Ebert, "Visual analytics decision support environment for epidemic modeling and response evaluation," in *IEEE Conference on Visual Analytics Science and Technology*. IEEE, 2011, pp. 191–200.

[10] A. Slingsby, J. Dykes, and J. Wood, "Exploring uncertainty in geodemographics with interactive graphics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2545–2554, 2011.

[11] P. E. Rauber, S. Fadel, A. Falcao, and A. Telea, "Visualizing the hidden activity of artificial neural networks," *IEEE Transactions on Visualization and Computer Graphics*, vol. PP, no. 99, pp. 1–1, 2016.

[12] D. Ren, S. Amershi, B. Lee, J. Suh, and J. D. Williams, "Squares: Supporting interactive performance analysis for multiclass classifiers," *IEEE Transactions on Visualization and Computer Graphics*, no. 1, pp. 1–1, 2017.

[13] T. Groenfeldt, "Kroger knows your shopping patterns better than you do," October 2013, [Online; posted 27-October-2013]. [Online]. Available: http://www.forbes.com/sites/tomgroenfeldt/2013/10/28/kroger-knows-your-shopping-patterns-better-than-you-do/#5f330d93396d

[14] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The parable of google flu: Traps in big data analysis," *Science*, vol. 343, no. 6176, pp. 1203–1205, 2014.

[15] D. Butler, "When google got flu wrong." *Nature*, vol. 494, no. 7436, p. 155, 2013.

[16] H. L. Dreyfus and S. E. Dreyfus, "What artificial experts can and cannot do," *AI & society*, vol. 6, no. 1, pp. 18–26, 1992.

[17] L. F. Cranor, "A framework for reasoning about the human in the loop." *UPSEC*, vol. 8, pp. 1–15, 2008.

[18] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them," *Management Science*, 2016.

[19] H. Piringer, W. Berger, and J. Krasser, "Hypermoval: Interactive visual validation of regression models for real-time simulation," in *Computer Graphics Forum*, vol. 29, no. 3. Wiley Online Library, 2010, pp. 983–992.

[20] A. Tatu, F. Maaß, I. Färber, E. Bertini, T. Schreck, T. Seidl, and D. Keim, "Subspace search and visualization to make sense of alternative clusterings in high-dimensional data," in *IEEE Conference on Visual Analytics Science and Technology*. IEEE, 2012, pp. 63–72.

[21] M. Gleicher, "A framework for considering comprehensibility in modeling," *Big Data*, 2016.

[22] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[23] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*. Wiley, 2012.

[24] S. Haykin, *Neural networks: A comprehensive foundation*. Prentice Hall PTR, 1994.

[25] H. Bosch, D. Thom, F. Heimerl, E. Puttmann, S. Koch, R. Kruger, M. Worner, and T. Ertl, "Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2022–2031, 2013.

[26] M. A. Smith, B. Shneiderman, N. Milic-Frayling, E. Mendes Rodrigues, V. Barash, C. Dunne, T. Capone, A. Perer, and E. Gleave, "Analyzing (social media) networks with nodexl," in *Proceedings of the fourth international conference on Communities and technologies*. ACM, 2009, pp. 255–264.

[27] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert, and T. Ertl, "Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition," in *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2012, pp. 143–152.

[28] R. Maciejewski, S. Rudolph, R. Hafen, A. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. J. Grannis, and D. S. Ebert, "A visual analytics approach to understanding spatiotemporal hotspots," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 2, pp. 205–220, 2010.

[29] A. Malik, R. Maciejewski, Y. Jang, S. Oliveros, Y. Yang, B. Maule, M. White, and D. S. Ebert, "A visual analytics process for maritime response, resource allocation and risk assessment," *Information Visualization*, p. 1473871612460991, 2012.

[30] L. Yu, W. Wu, X. Li, G. Li, W. S. Ng, S.-K. Ng, Z. Huang, A. Arunan, and H. M. Watt, "iviztrans: Interactive visual learning for home and work place detection from massive public transportation data," p. ?, 2015.

[31] "Twitter streaming api," https://dev.twitter.com/streaming/overview, accessed: 2015-11-24.

[32] F. Morstatter, J. Pfeffer, and H. Liu, "When is it biased?: Assessing the representativeness of twitter's streaming api," in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW '14 Companion. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2014, pp. 555–556. [Online]. Available: http://dx.doi.org/10.1145/2567948.2576952

[33] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, "Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose," *arXiv preprint arXiv:1306.5204*, 2013.

[34] D. Oelke, M. Hao, C. Rohrdantz, D. Keim, U. Dayal, L.-E. Haug, H. Janetzko *et al.*, "Visual opinion analysis of customer feedback data," in *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*. IEEE, 2009, pp. 187–194.

[35] X. Hu, J. Tang, H. Gao, and H. Liu, "Unsupervised sentiment analysis with emotional signals," in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 607–618.

[36] X. Hu, L. Tang, J. Tang, and H. Liu, "Exploiting social relations for sentiment analysis in microblogging," in *Proceedings of the sixth ACM international conference on Web search and data mining*, 2013, pp. 537–546.

[37] G. Shmueli and O. Koppius, "Predictive analytics in information systems research," *Robert H. Smith School Research Paper No. RHS*, pp. 06–138, 2010.

[38] N. Cao, Y.-R. Lin, X. Sun, D. Lazer, S. Liu, and H. Qu, "Whisper: Tracing the spatiotemporal process of information diffusion in real time," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2649–2658, 2012.

[39] A. M. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford, "Senseplace2: Geotwitter analytics support for situational awareness," in *IEEE Conference on Visual Analytics Science and Technology*. IEEE, 2011, pp. 181–190.

[40] F. Morstatter, S. Kumar, H. Liu, and R. Maciejewski, "Understanding twitter data with tweetxplorer," in *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2013, pp. 1482–1485.

[41] D. Thom, H. Bosch, S. Koch, M. Worner, and T. Ertl, "Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages," in *IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, 2012, pp. 41–48.

[42] M. Kuhn and K. Johnson, *Applied predictive modeling*. Springer, 2013.

[43] S. Suthaharan, "Big data classification: Problems and challenges in network intrusion prediction with machine learning," *ACM SIGMETRICS Performance Evaluation Review*, vol. 41, no. 4, pp. 70–73, 2014.

[44] S. Trivedi, Z. A. Pardos, and N. T. Heffernan, "The utility of clustering in prediction tasks," *arXiv preprint arXiv:1509.06163*, 2015.

[45] A. D. King, N. Pržulj, and I. Jurisica, "Protein complex prediction via cost-based clustering," *Bioinformatics*, vol. 20, no. 17, pp. 3013–3020, 2004.

[46] W. W. Eckerson, "Predictive analytics," *Extending the Value of Your Data Warehousing Investment. TDWI Best Practices Report. Q*, vol. 1, p. 2007, 2007.

[47] E. Siegel, *Predictive analytics: The power to predict who will click, buy, lie, or die*. John Wiley & Sons, 2013.

[48] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[49] P. Pirolli and S. Card, "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," in *Proceedings of international conference on intelligence analysis*, vol. 5, 2005, pp. 2–4.

[50] C. Auer, J. Kasten, A. Kratz, E. Zhang, and I. Hotz, "Automatic, tensor-guided illustrative vector field visualization," in *IEEE Pacific Visualization Symposium*. IEEE, 2013, pp. 265–272.

[51] F.-Y. Tzeng and K.-L. Ma, "Opening the black box-data driven visualization of neural networks," in *IEEE Visualization*. IEEE, 2005, pp. 383–390.

[52] G.-P. Bonneau, H.-C. Hege, C. R. Johnson, M. M. Oliveira, K. Potter, P. Rheingans, and T. Schultz, "Overview and state-of-the-art of uncertainty visualization," in *Scientific Visualization*. Springer, 2014, pp. 3–27.

[53] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: http://www.R-project.org

[54] SAS Institute Inc., *SAS/STAT Software, Version 9.3*, Cary, NC, 2011. [Online]. Available: http://www.sas.com/

[55] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, Nov. 2009. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278

[56] S. Publishing *et al.*, *JMP 10 Modeling and Multivariate Methods*. SAS Institute, 2012.

[57] S. Barlowe, J. Yang, D. J. Jacobs, D. R. Livesay, J. Alsakran, Y. Zhao, D. Verma, and J. Mottonen, "A visual analytics approach to exploring protein flexibility subspaces," in *IEEE Pacific Visualization Symposium*. IEEE, 2013, pp. 193–200.

[58] A. Malik, R. Maciejewski, S. Towers, S. McCullough, and D. S. Ebert, "Proactive spatiotemporal resource allocation and predictive visual analytics for community policing and law enforcement," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 1863–1872, 2014.

[59] B. Höferlin, R. Netzel, M. Höferlin, D. Weiskopf, and G. Heidemann, "Interactive learning of ad-hoc classifiers for video visual analytics," in *IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2012, pp. 23–32.

[60] J. Buchmüller, H. Janetzko, G. Andrienko, N. Andrienko, G. Fuchs, and D. A. Keim, "Visual analytics for exploring local impact of air traffic," in *Computer Graphics Forum*, vol. 34, no. 3. Wiley Online Library, 2015, pp. 181–190.

[61] Y. Lu, F. Wang, and R. Maciejewski, "Business intelligence from social media: A study from the vast box office challenge," *IEEE Computer Graphics and Applications*, pp. 58–69, 2014.

[62] Y. Lu, R. Krüger, D. Thom, F. Wang, S. Koch, T. Ertl, and R. Maciejewski, "Integrating predictive analytics and social media," in *Proc. IEEE Conference on Visual Analytics Science and Technology*, 2014.

[63] M. Ankerst, C. Elsen, M. Ester, and H.-P. Kriegel, "Visual classification: an interactive approach to decision tree construction," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1999, pp. 392–396.

[64] H. Lee, J. Kihm, J. Choo, J. Stasko, and H. Park, "ivisclustering: An interactive visual document clustering via topic modeling," in *Computer Graphics Forum*, vol. 31, no. 3pt3. Wiley Online Library, 2012, pp. 1155–1164.

[65] E. T. Brown, J. Liu, C. E. Brodley, and R. Chang, "Dis-function: Learning distance functions interactively," in *IEEE Conference on Visual Analytics Science and Technology*. IEEE, 2012, pp. 83–92.

[66] B. P. Mathews and A. Diamantopoulos, "Managerial intervention in forecasting. an empirical investigation of forecast manipulation," *International Journal of Research in Marketing*, vol. 3, no. 1, pp. 3–10, 1986.

[67] R. Carbone, A. Andersen, Y. Corriveau, and P. P. Corson, "Comparing for different time series methods the value of technical expertise individualized analysis, and judgmental adjustment," *Management Science*, vol. 29, no. 5, pp. 559–566, 1983.

[68] S. Highhouse, "Stubborn reliance on intuition and subjectivity in employee selection," *Industrial and Organizational Psychology*, vol. 1, no. 3, pp. 333–342, 2008.

[69] R. M. Dawes, D. Faust, and P. E. Meehl, "Clinical versus actuarial judgment," *Science*, vol. 243, no. 4899, pp. 1668–1674, 1989.

[70] W. M. Grove, D. H. Zald, B. S. Lebow, B. E. Snitz, and C. Nelson, "Clinical versus mechanical prediction: a meta-analysis," *Psychological assessment*, vol. 12, no. 1, pp. 19–30, 2000.

[71] A. A. Syntetos, K. Nikolopoulos, and J. E. Boylan, "Judging the judges through accuracy-implication metrics: The case of inventory forecasting," *International Journal of Forecasting*, vol. 26, no. 1, pp. 134–143, 2010.

[72] M. B. Cook and H. S. Smallman, "Human factors of the confirmation bias in intelligence analysis: Decision support from graphical evidence landscapes," *Human Factors*, vol. 50, no. 5, pp. 745–754, 2008.

[73] R. S. Nickerson, "Confirmation bias: A ubiquitous phenomenon in many guises," *Review of general psychology*, vol. 2, no. 2, pp. 175–220, 1998.

[74] J. Klayman, J. B. Soll, C. González-Vallejo, and S. Barlas, "Overconfidence: It depends on how, what, and whom you ask," *Organizational behavior and human decision processes*, vol. 79, no. 3, pp. 216–247, 1999.

[75] A. Furnham and H. C. Boo, "A literature review of the anchoring effect," *The Journal of Socio-Economics*, vol. 40, no. 1, pp. 35–42, 2011.

[76] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, vol. 30, no. 3, pp. 286–297, 2000.

[77] K. A. Hoff and M. Bashir, "Trust in automation integrating empirical evidence on factors that influence trust," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 57, no. 3, pp. 407–434, 2015.

[78] C. Eroglu and K. L. Croxton, "Biases in judgmental adjustments of statistical forecasts: The role of individual differences," *International Journal of Forecasting*, vol. 26, no. 1, pp. 116–133, 2010.

[79] M. Lawrence, P. Goodwin, M. O'Connor, and D. Önkal, "Judgmental forecasting: A review of progress over the last 25years," *International Journal of Forecasting*, vol. 22, no. 3, pp. 493–518, 2006.

[80] R. Fildes and P. Goodwin, "Against your better judgment? how organizations can improve their use of management judgment in forecasting," *Interfaces*, vol. 37, no. 6, pp. 570–576, 2007.

[81] S. I. Vrieze and W. M. Grove, "Survey on the use of clinical and mechanical prediction methods in clinical psychology." *Professional Psychology: Research and Practice*, vol. 40, no. 5, p. 525, 2009.

[82] J. Pearl, "Comments on neuberg's review of causality," *Econometric Theory*, vol. 19, no. 04, jun 2003. [Online]. Available: https://doi.org/10.1017%2Fs0266466603004110

[83] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err." *Journal of Experimental Psychology: General*, vol. 144, no. 1, p. 114, 2015.

[84] B. J. Dietvorst, "People reject (superior) algorithms because they compare them to counter-normative reference points," 2016.

[85] K. A. Cook and J. J. Thomas, "Illuminating the path: The research and development agenda for visual analytics," Pacific Northwest National Laboratory (PNNL), Richland, WA (US), Tech. Rep., 2005.

[86] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler, "Visual analytics: Scope and challenges," in *Visual data mining*. Springer, 2008, pp. 76–90.

[87] X.-M. Wang, T.-Y. Zhang, Y.-X. Ma, J. Xia, and W. Chen, "A survey of visual analytic pipelines," *Journal of Computer Science and Technology*, vol. 31, no. 4, pp. 787–804, 2016.

[88] E. Bertini and D. Lalanne, "Surveying the complementary role of automatic data analysis and visualization in knowledge discovery," in *Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration.* ACM, 2009, pp. 12–20.

[89] D. Sacha, M. Sedlmair, L. Zhang, J. Lee, D. Weiskopf, S. North, and D. Keim, "Human-centered machine learning through interactive visualization: Review and open challenges," in *Proceedings of the 24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2016.

[90] G. K. Tam, V. Kothari, and M. Chen, "An analysis of machine-and human-analytics in classification," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 71–80, 2017.

[91] J. Lu, W. Chen, Y. Ma, J. Ke, Z. Li, F. Zhang, and R. Maciejewski, "Recent progress and trends in predictive visual analytics," *Frontiers of Computer Science*, 2016.

[92] J. S. Yi, Y. ah Kang, J. Stasko, and J. Jacko, "Toward a deeper understanding of the role of interaction in information visualization," *IEEE transactions on visualization and computer graphics*, vol. 13, no. 6, pp. 1224–1231, 2007.

[93] M. Hao, C. Rohrdantz, H. Janetzko, U. Dayal, D. A. Keim, L. Haug, and M.-C. Hsu, "Visual sentiment analysis on twitter data streams," in *IEEE Conference on Visual Analytics Science and Technology.* IEEE, 2011, pp. 277–278.

[94] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer, "Wrangler: Interactive visual specification of data transformation scripts," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, 2011, pp. 3363–3372.

[95] S. Ingram, T. Munzner, V. Irvine, M. Tory, S. Bergner, and T. Möller, "Dimstiller: Workflows for dimensional analysis and reduction," in *IEEE Symposium on Visual Analytics Science and Technology.* IEEE, 2010, pp. 3–10.

[96] J. Seo and B. Shneiderman, "A rank-by-feature framework for interactive exploration of multidimensional data," *Information visualization*, vol. 4, no. 2, pp. 96–113, 2005.

[97] D. Sacha, L. Zhang, M. Sedlmair, J. A. Lee, J. Peltonen, D. Weiskopf, S. North, and D. A. Keim, "Visual interaction with dimensionality reduction: a structured literature analysis," *IEEE Transactions on Visualization and Computer Graphics*, 2016.

[98] G. K. L. Tam, H. Fang, A. J. Aubrey, P. W. Grant, P. L. Rosin, D. Marshall, and M. Chen, "Visualization of time-series data in parameter space for understanding facial dynamics," *Computer Graphics Forum*, vol. 30, no. 3, pp. 901–910, 2011.

[99] S. Bremm, T. von Landesberger, J. Bernard, and T. Schreck, "Assisted descriptor selection based on visual comparative data analysis," in *Computer Graphics Forum*, vol. 30, no. 3. Wiley Online Library, 2011, pp. 891–900.

[100] A. Kapoor, B. Lee, D. Tan, and E. Horvitz, "Interactive optimization for steering machine classification," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2010, pp. 1343–1352.

[101] Z. Guo, M. O. Ward, and E. A. Rundensteiner, "Model space visualization for multivariate linear trend discovery," in *IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2009, pp. 75–82.

[102] J. Krause, A. Perer, and K. Ng, "Interacting with predictions: Visual inspection of black-box machine learning models," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2016, pp. 5686–5697.

[103] R. Maciejewski, A. Pattath, S. Ko, R. Hafen, W. S. Cleveland, and D. S. Ebert, "Automated box-cox transformations for improved visual encoding," *IEEE transactions on visualization and computer graphics*, vol. 19, no. 1, pp. 130–140, 2013.

[104] F. Zhou, J. Li, W. Huang, Y. Zhao, X. Yuan, X. Liang, and Y. Shi, "Dimension reconstruction for visual exploration of subspace clusters in high-dimensional data," in *IEEE Pacific Visualization Symposium*. IEEE, 2016, pp. 128–135.

[105] M. Brooks, S. Amershi, B. Lee, S. M. Drucker, A. Kapoor, and P. Simard, "Featureinsight: Visual support for error-driven feature ideation in text classification," in *IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2015, pp. 105–112.

[106] K. Patel, S. M. Drucker, J. Fogarty, A. Kapoor, and D. S. Tan, "Using multiple models to understand data," in *IJCAI Proceedings of International Joint Conference on Artificial Intelligence*, vol. 22, no. 1. Citeseer, 2011, p. 1723.

[107] T. Höllt, A. Magdy, G. Chen, G. Gopalakrishnan, I. Hoteit, C. D. Hansen, and M. Hadwiger, "Visual analysis of uncertainties in ocean forecasts for planning and operation of off-shore structures," in *IEEE Pacific Visualization Symposium*. IEEE, 2013, pp. 185–192.

[108] E. J. Nam, Y. Han, K. Mueller, A. Zelenyuk, and D. Imre, "Clustersculptor: A visual analytics tool for high-dimensional data," in *IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2007, pp. 75–82.

[109] G. Andrienko, N. Andrienko, S. Rinzivillo, M. Nanni, D. Pedreschi, and F. Giannotti, "Interactive visual clustering of large collections of trajectories," in *IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2009, pp. 3–10.

[110] J. G. S. Paiva, W. R. Schwartz, H. Pedrini, and R. Minghim, "An approach to supporting incremental visual data classification," *IEEE transactions on visualization and computer graphics*, vol. 21, no. 1, pp. 4–17, 2015.

[111] P. Bruneau and B. Otjacques, "An interactive, example-based, visual clustering system," 2013.

[112] M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, and S. Liu, "Towards better analysis of deep convolutional neural networks," *IEEE Transactions on Visualization and Computer Graphics*, vol. PP, no. 99, pp. 1–1, 2016.

[113] D. Caragea, D. Cook, H. Wickham, and V. Honavar, "Visual methods for examining svm classifiers," in *Visual Data Mining*. Springer, 2008, pp. 136–153.

[114] S. K. Badam, J. Zhao, S. Sen, N. Elmqvist, and D. Ebert, "Timefork: Interactive prediction of time series," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2016, pp. 5409–5420.

[115] S. Bremm, T. von Landesberger, M. Heß, T. Schreck, P. Weil, and K. Hamacherk, "Interactive visual comparison of multiple trees," in *IEEE Conference on Visual Analytics Science and Technology*. IEEE, 2011, pp. 31–40.

[116] R. Blanch, R. Dautriche, and G. Bisson, "Dendrogramix: A hybrid tree-matrix visualization technique to support interactive exploration of dendrograms," in *IEEE Pacific Visualization Symposium*. IEEE, 2015, pp. 31–38.

[117] B. Alsallakh, A. Hanbury, H. Hauser, S. Miksch, and A. Rauber, "Visual methods for analyzing probabilistic classification data," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 1703–1712, 2014.

[118] A. Pilhöfer, A. Gribov, and A. Unwin, "Comparing clusterings using bertin's idea," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2506–2515, Dec 2012.

[119] Y. Zhang, W. Luo, E. Mack, and R. Maciejewski, "Visualizing the impact of geographical variations on multivariate clustering," in *Computer Graphics Forum*, vol. 35, no. 3. Wiley Online Library, 2016, pp. 101–110.

[120] Y. Zhang and R. Maciejewski, "Quantifying the visual impact of classification boundaries in choropleth maps," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 371–380, 2017.

[121] R. Krüger, H. Bosch, D. Thom, E. Püttmann, Q. Han, S. Koch, F. Heimerl, and T. Ertl, "Prolix-visual prediction analysis for box office success," in *IEEE Conference on Visual Analytics Science and Technology*, 2013.

[122] Y. Lu, F. Wang, and R. Maciejewski, "Vast 2013 mini-challenge 1: Box office vast-team vader," in *IEEE Conference on Visual Analytics Science and Technology*, 2013.

[123] M. El-Assady, W. Jentner, M. Stein, F. Fischer, T. Schreck, and D. Keim, "Predictive visual analytics: Approaches for movie ratings and discussion of open research challenges," in *An IEEE VIS 2014 Workshop: Visualization for Predictive Analytics*, 2014.

[124] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[125] A. Kapoor, B. Lee, D. S. Tan, and E. Horvitz, "Performance and preferences: Interactive refinement of machine learning procedures." in *AAAI*. Citeseer, 2012.

[126] C. Seifert and M. Granitzer, "User-based active learning," in *IEEE International Conference on Data Mining Workshops*. IEEE, 2010, pp. 418–425.

[127] P. Foundation, "Social media usage: 2005-2015," 2015. [Online]. Available: http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/

[128] D. Godes and D. Mayzlin, "Firm-created word-of-mouth communication: Evidence from a field test," *Marketing Science*, vol. 28, no. 4, pp. 721–739, 2009.

[129] A. Rapp, L. S. Beitelspacher, D. Grewal, and D. E. Hughes, "Understanding social media effects across seller, retailer, and consumer interactions," *Journal of the Academy of Marketing Science*, vol. 41, no. 5, pp. 547–566, 2013.

[130] S. Tuarob and C. S. Tucker, "Fad or here to stay: Predicting product market adoption and longevity using large scale, social media data," in *ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers, 2013, pp. V02BT02A012–V02BT02A012.

[131] J. S. Simonoff and I. R. Sparrow, "Predicting movie grosses: Winners and losers, blockbusters and sleepers," *Chance*, vol. 13, no. 3, pp. 15–24, 2000.

[132] W. Zhang and S. Skiena, "Improving movie gross prediction through news analysis," in *Proceedings of the IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, 2009, pp. 301–304.

[133] M. Joshi, D. Das, K. Gimpel, and N. A. Smith, "Movie reviews and revenues: An experiment in text regression," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 293–296.

[134] S. Asur and B. A. Huberman, "Predicting the future with social media," in *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2010, pp. 492–499.

[135] S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts, "Predicting consumer behavior with web search," *Proceedings of the National academy of sciences*, vol. 107, no. 41, pp. 17 486–17 490, 2010.

[136] A. C. Reggie Panaligan, "Quantifying movie magic with google search," *Google Whitepaper — Industry Perspectives + User Insights*, 2013.

[137] A. Bhave, H. Kulkarni, V. Biramane, and P. Kosamkar, "Role of different factors in predicting movie success," in *2015 International Conference on Pervasive Computing (ICPC)*, Jan 2015, pp. 1–4.

[138] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, 2008. [Online]. Available: http://www.R-project.org

[139] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *Proceedings of the International Conference on Language Resources and Evaluation*, 2010.

[140] G. Rowe and G. Wright, "The delphi technique as a forecasting tool: Issues and analysis," *International journal of forecasting*, vol. 15, no. 4, pp. 353–375, 1999.

[141] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, April 2005.

[142] Z. Zhao, L. Wang, and H. Liu, "Efficient spectral feature selection with minimum redundancy," in *AAAI Conference on Artificial Intelligence*, 2010.

[143] H. Piringer, W. Berger, and H. Hauser, "Quantifying and comparing features in high-dimensional datasets," in *12th International Conference on Information Visualisation*. IEEE, 2008, pp. 240–245.

[144] J. Seo and B. Shneiderman, "A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections," in *IEEE Symposium on Information Visualization*. IEEE, 2004, pp. 65–72.

[145] M. Harrower and C. A. Brewer, "Colorbrewer. org: An online tool for selecting colour schemes for maps," *The Cartographic Journal*, vol. 40, no. 1, pp. 27–37, 2003.

[146] R. Krüger, H. Bosch, D. Thom, E. Püttmann, Q. Han, S. Koch, F. Heimerl, and T. Ertl, "Prolix - visual prediction analysis for box office success," in *IEEE Conference on Visual Analytics Science and Technology*, 2013.

[147] D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons, 2005, vol. 571.

[148] S. Geisser, *Predictive inference*. Chapman & Hall, New York, 1993.

[149] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, 1995, pp. 1137–1143.

[150] M. Joshi, D. Das, K. Gimpel, and N. A. Smith, "Movie reviews and revenues: An experiment in text regression," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 293–296.

[151] K. Ericsson and H. Simon, *Protocol Analysis: Verbal Reports as Data.* MIT Press, Cambridge, MA, 1993.

[152] T. Liu, X. Ding, Y. Chen, H. Chen, and M. Guo, "Predicting movie box-office revenues by exploiting large-scale social media content," *Multimedia Tools and Applications*, vol. 75, no. 3, pp. 1509–1528, 2016.

[153] Z. Zhang, B. Li, Z. Deng, J. Chai, Y. Wang, and M. An, "Research on movie box office forecasting based on internet data," in *2015 8th International Symposium on Computational Intelligence and Design (ISCID)*, vol. 2. IEEE, 2015, pp. 83–86.

[154] T. L. Broekhuizen, S. A. Delre, and A. Torres, "Simulating the cinema market: How cross-cultural differences in social influence explain box office distributions," *Journal of Product Innovation Management*, vol. 28, no. 2, pp. 204–217, 2011.

[155] M. A. Smith, L. Rainie, B. Shneiderman, and I. Himelboim, "Mapping twitter topic networks: From polarized crowds to community clusters," *Pew Research Center*, vol. 20, 2014.

[156] E. M. Rodrigues, N. Milic-Frayling, M. Smith, B. Shneiderman, and D. Hansen, "Group-in-a-box layout for multi-faceted analysis of communities," in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on.* IEEE, 2011, pp. 354–361.

[157] Y. Lu, F. Wang, and R. Maciejewski, "VAST 2013 Mini-Challenge 1: Box Office VAST-Team VADER," in *IEEE Conference on Visual Analytics Science and Technology*, 2013.

[158] A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," *Physical review E*, vol. 70, no. 6, p. 066111, 2004.

[159] D. Harel and Y. Koren, "A fast multi-scale method for drawing large graphs," in *International Symposium on Graph Drawing.* Springer, 2000, pp. 183–196.

[160] T. M. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *Software: Practice and experience*, vol. 21, no. 11, pp. 1129–1164, 1991.

[161] M. Smith, N. Milic-Frayling, B. Shneiderman, E. Mendes Rodrigues, J. Leskovec, and C. Dunne, "Nodexl: a free and open network overview, discovery and exploration add-in for excel 2007/2010," *Social Media Research Foundation*, 2010.

[162] K. Krippendorff, *Content analysis: An introduction to its methodology.* Sage, 2004.

[163] R. J. Crouser, L. Franklin, A. Endert, and K. Cook, "Toward theoretical techniques for measuring the use of human effort in visual analytic systems," *IEEE Transactions on Visualization and Computer Graphics*, vol. PP, no. 99, pp. 1–1, 2016.

[164] F. Petropoulos, R. Fildes, and P. Goodwin, "Do big losses in judgmental adjustments to statistical forecasts affect experts behaviour?" *European Journal of Operational Research*, vol. 249, no. 3, pp. 842–852, 2016.

[165] R. M. Dawes, "The robust beauty of improper linear models in decision making," *American psychologist*, vol. 34, no. 7, pp. 571–582, 1979.

[166] N. Silver, *The signal and the noise: Why so many predictions fail-but some don't*. Penguin, 2012.

[167] H. R. Arkes, R. M. Dawes, and C. Christensen, "Factors influencing the use of a decision rule in a probabilistic task," *Organizational Behavior and Human Decision Processes*, vol. 37, no. 1, pp. 93–110, 1986.

[168] F. Gul, "A theory of disappointment aversion," *Econometrica: Journal of the Econometric Society*, pp. 667–686, 1991.

[169] K. Yu, S. Berkovsky, R. Taib, D. Conway, J. Zhou, and F. Chen, "User trust dynamics: An investigation driven by differences in system performance," in *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. ACM, 2017, pp. 307–317.

[170] V. Buchanan, Y. Lu, N. McNeese, M. Steptoe, R. Maciejewski, and N. Cooke, "The role of teamwork in the analysis of big data — a study of visual analytics and box office prediction," *Big Data*, vol. 3, 2017.

[171] F. Al-Masoudi, D. Seebacher, M. Schreiner, M. Stein, C. Rohrdantz, F. Fischer, S. Simon, T. Schreck, and D. Keim, "Similarity-driven visual-interactive prediction of movie ratings and box office results," in *IEEE Conference on Visual Analytics Science and Technology*, 2013.

[172] E. Guizzo, "How googles self-driving car works," *IEEE Spectrum Online, October*, vol. 18, 2011.

[173] T. Higgins, "Google's self-driving car program odometer reaches 2 million miles," *The Wall Street Journal*, 2016. [Online]. Available: http://www.wsj.com/articles/googles-self-driving-car-program-odometer-reaches-2-million-miles-1475683321

[174] M. U. Scherer, "Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies," *SSRN Electronic Journal*, 2016. [Online]. Available: https://doi.org/10.2139%2Fssrn.2609777

[175] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *European Conference on Computer Vision*. Springer, 2016, pp. 3–19.

[176] A. Endert, L. Bradel, and C. North, "Beyond control panels: Direct manipulation for visual analytics," *IEEE Computer Graphics and Applications*, vol. 33, pp. 6–13, 2013.

[177] A. Endert, P. Fiaux, and C. North, "Semantic interaction for visual text analytics," in *SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 473–482.

[178] A. Endert, S. Fox, D. Maiti, and C. North, "The semantics of clustering: analysis of user-generated spatializations of text documents," in *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM, 2012, pp. 555–562.

APPENDIX A

DEMOGRAPHICS QUESTIONNAIRE

Please answer the following to the best of your ability. All answers will be kept confidential and will only be reported statistically (grouped with others' responses). Please feel free to leave a question blank if you feel uncomfortable answering it.

1. What is your age? _____
2. What is your gender? (circle):
   a Male
   b Female
3. What is your current level of education?
   a Less than High School
   b High School/GED
   c Some College
   d 2 year degree
   e 4 year degree
   f Master's
   g Doctoral
   h Professional (MD, JD, etc.)
4. If you have been or are enrolled in a post high school institution, what is your major?
   _____
5. Are you currently employed?
   a Yes
   b No
6. If yes to # 5, what is your job title?
   _____
7. Are you native English Speaker?
   a Yes
   b No
   c If No, then what is your native language?
   _____
8. How long have you lived in the United States?
   a Native (all my life)
   b Less than 1 year
   c 1 year
   d 2 years
   e 3 years
   f 4 years
   g Greater than 5 years
9. How long have you lived in the United States?
   a Native (all my life)
   b Less than 1 year
   c 1 year
   d 2 years
   e 3 years
   f 4 years
   g Greater than 5 years
10. Within a month: On average, how often do you watch movies (Theater, TV, Internet)?
    a 0
    b 1-2
    c 3-4
    d 5-6
    e 7-8
    f 9-10
    g 11 or more
11. Were you familiar with any of the movies presented today?
    a Yes
    b No
    c If, yes which one(s)? Please mark ALL that apply.

i Fast & Furious (2013, Practice)
ii The Amazing Spider-Man2 (2014)
iii The Equalizer (2014)
iv 300: Rise of an Empire (2014)
v Maleficent (2014)
vi San Andreas (2015)
vii Transformers: Age of Extinction (2014)
viii Mission Impossible-Rouge Nation (2014)
ix Terminator Genisys (2015)
x Kingsman: The Secret Service (2014)

12. If you answered Yes to # 10: Did you know how much money the movie(s) had made during the opening weekend?
    a Yes
    b No
    c If, yes which one(s)? Please mark ALL that apply.
        i Fast & Furious (2013, Practice)
        ii The Amazing Spider-Man2 (2014)
        iii The Equalizer (2014)
        iv 300: Rise of an Empire (2014)
        v Maleficent (2014)
        vi San Andreas (2015)
        vii Transformers: Age of Extinction (2014)
        viii Mission Impossible-Rouge Nation (2014)
        ix Terminator Genisys (2015)
        x Kingsman: The Secret Service (2014)

13. Were you familiar with how popular any of these movies were opening weekend?
    a Yes
    b No
    c If, yes which one(s)? Please mark ALL that apply.
        i Fast & Furious (2013, Practice)
        ii The Amazing Spider-Man2 (2014)
        iii The Equalizer (2014)
        iv 300: Rise of an Empire (2014)
        v Maleficent (2014)
        vi San Andreas (2015)
        vii Transformers: Age of Extinction (2014)
        viii Mission Impossible-Rouge Nation (2014)
        ix Terminator Genisys (2015)
        x Kingsman: The Secret Service (2014)

14. Do you follow new release movies on Social Media?
    a Yes
    b No

15. How often do you use Social Media?
    a Several times an hour
    b Hourly
    c Daily
    d Every couple days
    e Once a week
    f Every couple weeks
    g Less than once a month
    h Every couple of months
    i Once or twice a year

j Never

16. Were you familiar with predictive modelling prior to this experiment?

    a Yes
    b No
    c If yes, how? Please describe/list briefly (For example: learned in class, part of my degree program, part of my job, etc.)
    _____
    _____
    _____
    _____
    _____

17. How often have you used predictive modeling prior to this experiment?

    a Never
    b Rarely
    c Occasionally
    d Frequently
    e Very Frequently

18. How often do you use a computer?

    a Hourly
    b Daily
    c Every couple days
    d Once a week
    e Every couple weeks
    f Less than once a month
    g I do not use computers

19. Please rate the degree to which you agree with the following statement: I am proficient with computers.

    a Strongly Agree
    b Slightly Agree
    c Neutral
    d Slightly Disagree
    e Strongly Disagree

20. In what way do you use computers? (Mark **all** that apply)

    a I do not use computers
    b Internet
    c Email
    d Word processing
    e Spreadsheets
    f Computer Games
    g Other

21. This task was complicated.

    a Strongly Agree
    b Slightly Agree
    c Neutral
    d Slightly Disagree
    e Strongly Disagree

22. This task was boring.

    a Strongly Agree
    b Slightly Agree
    c Neutral
    d Slightly Disagree
    e Strongly Disagree

23. This task was easy.

    a Strongly Agree
    b Slightly Agree
    c Neutral
    d Slightly Disagree
    e Strongly Disagree

24. The user-computer interface was easy to use.

    a Strongly Agree
    b Slightly Agree
    c Neutral

d Slightly Disagree

e Strongly Disagree

25. What features of the movie interface were easy to use/understand? Please mark **all** that apply.

    a None

    b Home Page

    c Model Prediction Page

    d Weekend Market Share Page

    e Sentiment Analysis Page

    f Movie Similarity Page

    g Make Prediction Page

    Additional Comments:

    _____

26. What features of the movie interface were difficult to work with/understand? Please mark **all** that apply.

    a None

    b Home Page

    c Model Prediction Page

    d Weekend Market Share Page

    e Sentiment Analysis Page

    f Movie Similarity Page

    g Make Prediction Page

    Additional Comments:

    _____

    _____

    _____

    _____

27. Which features of the movie interface did you use to make your predictions? Please mark **all** that apply.

    a Home Page

    b Model Prediction Page

    c Weekend Market Share Page

    d Sentiment Analysis Page

    e Movie Similarity Page

    f Make Prediction Page

28. Which features of the movie interface did you use the most to make your predictions? Please rate the **3** most often used interface features (1=most often, 2=second most often, 3=third most often).

    a Home Page _____

    b Model Prediction Page _____

    c Weekend Market Share Page _____

    d Sentiment Analysis Page _____

    e Movie Similarity Page _____

    f Make Prediction Page _____

29. I would use this interface again.

    a Strongly Agree

    b Slightly Agree

    c Neutral

    d Slightly Disagree

    e Strongly Disagree

30. How did you go about analyzing the presented data? Did you look at all available data? Did you develop a specific strategy? Did you only consider certain pieces of data?

    Please list/describe briefly.

    _____

    _____

    _____

    _____

    _____

    _____

    _____

    _____

    _____

    _____

31. I enjoyed participating in this study.

a Strongly Agree

b Slightly Agree

c Neutral

d Slightly Disagree

e Strongly Disagree

32. Would you like to share anything else about this experiment with us? Please list/describe below.

_____

_____

_____

_____

APPENDIX B

TLX REPORT

Instructions:

Below you will be asked some questions about the task you just completed. Please read each question and think about the information being requested. Then, respond on each scale about how you felt or what you experienced within the task. Please consider each scale independent of the previous or following scales. If you have any questions, please ask the experimenter.

1. **How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)?**

The task was easy	**1 2 3 4 5 6 7 8 9 10**	The task was demanding
The task was simple	**1 2 3 4 5 6 7 8 9 10**	The task was complex
The task was forgiving	**1 2 3 4 5 6 7 8 9 10**	The task was exacting
The task was mentally effortless	**1 2 3 4 5 6 7 8 9 10**	The task was mentally difficult

2. **How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred?**

The task was slow	**1 2 3 4 5 6 7 8 9 10**	The task was rapid
The task was leisurely	**1 2 3 4 5 6 7 8 9 10**	The task was frantic

3. **How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)?**

Unsuccessful	**1 2 3 4 5 6 7 8 9 10**	Successful

4. **Please rate the following emotional dimensions felt during the task**

Insecure	**1 2 3 4 5 6 7 8 9 10**	Secure
Discouraged	**1 2 3 4 5 6 7 8 9 10**	Gratified
Irritated	**1 2 3 4 5 6 7 8 9 10**	Content
Stressed	**1 2 3 4 5 6 7 8 9 10**	Relaxed
Annoyed	**1 2 3 4 5 6 7 8 9 10**	Complacent

APPENDIX C

MOVIE INTERFACE QUIZ

1) I have access to the tutorial page throughout the entire study.

   a True
   b False

2) The goal of this task is to predict as accurately as possible how much the movies are going to make opening weekend.

   a True
   b False

3) I can get the actual gross amount for all movies displayed.

   a True
   b False

4) Each of the small green squares represent a week in a calendar year.

   a True
   b False

5) Why are some of the last squares in 'Weekend to Predict' colored black?

   a I can't remember.
   b They are colored black because that is the weekend to be predicted.
   c They are colored black because some of those values are not available.

6) The 'Sentiment Analysis' displays tweets about the given movie selected.

   a True
   b False

7) I can access the following data. Mark all that apply.

   a MPAA Rating
   b Genre/Category of movie
   c Accuracy of Model
   d Number of tweets per day
   e Release date
   f Prediction range

8) My prediction has to be within the given prediction range.

   a True
   b False

9) The 'Total Opening Weekend Gross' is:

a Predicted amount for the movie under investigation.

b Predicted amount for all movies to be released that weekend.

c Not sure.

10) My prediction amount can be given with commas, periods, asterisks, etc.

a True

b False