Predicting Demographic and Financial Attributes

in a Bank Marketing Dataset

by

Samira Ejaz

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved November 2015 by the
Graduate Supervisory Committee:

Hasan Davulcu, Chair
Janaka Balasooriya
Kasim Candan

ARIZONA STATE UNIVERSITY

May 2016

ABSTRACT

Bank institutions employ several marketing strategies to maximize new customer acquisition as well as current customer retention. Telemarketing is one such approach taken where individual customers are contacted by bank representatives with offers. These telemarketing strategies can be improved in combination with data mining techniques that allow predictability of customer information and interests. In this thesis, bank telemarketing data from a Portuguese banking institution were analyzed to determine predictability of several client demographic and financial attributes and find most contributing factors in each. Data were preprocessed to ensure quality, and then data mining models were generated for the attributes with logistic regression, support vector machine (SVM) and random forest using Orange as the data mining tool. Results were analyzed using precision, recall and F1 score.

DEDICATION

To my family for their love and support

ACKNOWLEDGMENTS

I would like to take this opportunity to thank all of those who have contributed to my thesis directly and indirectly. My sincere gratitude extends out to my committee chair Professor Hasan Davulcu for his continued guidance and support throughout this thesis journey. This work would not have been possible without his help and direction. In addition, I would also like to thank Professor Kasim Candan and Professor Janaka Balasooriya for their participation in the thesis committee. I appreciate the time and support from my entire committee.

I thank my family for their care and encouragement throughout my pursuit of this journey. My friends have also been a source a support. I am blessed to have so many people supporting my success.

TABLE OF CONTENTS

Page

LIST OF TABLES vii

LIST OF FIGURES viii

CHAPTER

1. INTRODUCTION 1

1.1 Purpose and Motivation 1

1.2 Scope 2

1.3 Outline 3

2. BACKGROUND 4

2.1 Introduction 4

2.2 Data 4

2.3 Algorithms 8

2.4 Tool 10

2.5 Evaluation Metrics 12

3. RELATED WORK 17

3.1 Introduction 17

3.2 Examination of Related Work 17

CHAPTER Page

4. IMPLEMENTATION 20

4.1 Introduction 20

4.2 Data Preprocessing 21

4.3 Imputation Analysis 29

4.4 Training and Testing 30

5. EVALUATION 38

5.1 Introduction 38

5.2 Format of Output from Orange 38

5.3 Experimental Results and Analysis 39

5.3.1 Age Group 40

5.3.2 Employment Status 42

5.3 3 Marital Status 44

5.3.4 Education Level 45

5.3.5 Housing Loan 47

5.3.6 Personal Loan 49

5.3.7 Term Deposit 51

6. CONCLUSION AND FUTURE WORK 53

Page

REFERENCES 56

APPENDIX

A. FINAL DATA MODELS 58

LIST OF TABLES

Table Page

1. Description of Attributes in the Original Dataset 7
2. All Unique Categories for Age Group with the Number of Instances for Each Category in Square Braces. A Total of 78 Unique Categories Exist and the Representation of Those Categories Range from 1 Instance to 1947 Instances. .. 23
3. Data Distribution and Categories of Age Group after Preprocessing.................. 24
4. All Unique Categories for Employment Status and the Number of Instances for Each Category. A Total of 11 Unique Categories Exist. 25
5. Data Distribution and Categories of Employment Status after Preprocessing 25
6. Data Distribution and Categories of Marital Status... 26
7. Data Distribution and Categories of Martial Status after Preprocessing 26
8. All Unique Categories for Education Level and the Number of Instances for Each Category. A Total of 7 Unique Categories Exist 27
9. Data Distribution and Categories of Education Level after Preprocessing 27
10. Data Distribution and Categories of Housing Loan... 28
11. Data Distribution and Categories of Personal Loan 28
12. Data Distribution and Categories of Term Deposit 29
13. Best F1 Score for Each Attribute on the Smallest Category Followed by the Precision, Recall and Algorithm Used. Attributes Are Sorted by the F1 Score. . 53

LIST OF FIGURES

Figure Page

1. All Dataset Attributes Categorized into Different Types with the Percentage of Each Type Shown as a Pie Chart. 5
2. Confusion Matrix Where the Columns Represent the Prediction and the Rows Represent the Actual Classification 13
3. Confusion Matrix of a Model with Some Predictive Power (Top) and a Confusion Matrix of a Model with Zero Predictive Power (Bottom) as Items Are Always Classified as Part of the Negative Class. 15
4. Precision and Recall 16
5. Results from Not Imputing the Attributes and Not Imputing the Class for Term Deposit. 30
6. Data Flow Model Created Using Orange to Depict the Process Flow of the Data from Input to Evaluation. 31
7. Left: Inputs to the File Widget for the Marital Status Test Class. Right: Settings Used for the Impute Widget. 32
8. Parameters Used for Logistic Regression, Random Forest and SVM Widgets 33
9. Code for Python Widget 1 to Print Model Information for Logistic Regression . 34
10. Code for Python Widget 2 to Print Model Information for SVM 34
11. Classification Tree Graph Widget to Depict the Data Model Generated by Random Forest 35

Figure Page

12. Sampling Settings for the Test Learners Widget and the Results for Logistic Regression, Random Forest and SVM on the 'No' Category of the Term Deposit Class 36

13. Confusion Matrix for Logistic Regression on the Term Deposit Class 37

14. Format of Results from the Test Learners Widget Using Orange for Logistic Regression, Random Forest and SVM. The Top Shows the Results for the 'No' Category and the Bottom Shows Results for the 'Yes' Category. 39

15. Results for the Age Group Class. The Top Shows the F1, Precision and Recall Score and the Highest Value Is Highlighted in Each Column. Below Is a Histogram of the Categories and Also the Top Contributing Coefficients for All Algorithms Where Common Attributes Are Colored. 41

16. Results for the Employment Status Class. The Top Shows the F1, Precision and Recall Score and the Highest Value Is Highlighted in Each Column. Below Is a Histogram of the Categories and Also the Top Contributing Coefficients for All Algorithms Where Common Attributes Are Colored. 43

17. Results for the Marital Status Class. The Top Shows the F1, Precision and Recall Score and the Highest Value Is Highlighted in Each Column. Below Is a Histogram of the Categories and Also the Top Contributing Coefficients for All Algorithms Where Common Attributes Are Colored. 45

Figure Page

18. Results for the Education Level Class. The Top Shows the F1, Precision and Recall Score and the Highest Value Is Highlighted in Each Column. Below Is a Histogram of the Categories and Also the Top Contributing Coefficients for All Algorithms Where Common Attributes Are Colored. 46

19. Results for the Housing Loan Class. The Top Shows the F1, Precision and Recall Score and the Highest Value Is Highlighted in Each Column. Below Is a Histogram of the Categories and Also the Top Contributing Coefficients for All Algorithms Where Common Attributes Are Colored. 48

20. Results for the Personal Loan Class. The Top Shows the F1, Precision and Recall Score and the Highest Value Is Highlighted in Each Column. Below Is a Histogram of the Categories and Also the Top Contributing Coefficients for All Algorithms Where Common Attributes Are Colored. 50

21. Results for the Term Deposit Class. The Top Shows the F1, Precision and Recall Score and the Highest Value Is Highlighted in Each Column. Below Is a Histogram of the Categories and Also the Top Contributing Coefficients for All Algorithms Where Common Attributes Are Colored. 52

22. Number of Top Contributors Associated with Each Category for Each of the Attributes Tested. 54

CHAPTER 1

INTRODUCTION

1.1 Purpose and Motivation

Marketing strategies are utilized by banks to increase client subscriptions to investments. In turn, this strategy increases customer retention. One such selling technique is telemarketing. Phone calls made by banks help to gain investments and increase company profits. Although this is a working strategy, there is more that can be done to maximize profits. To gain a competitive edge, these marketing strategies can be coupled with statistical techniques that predict outcomes. Through the use of data mining classification algorithms, banks can make these predictions of client interest to refine their marketing strategies and customize them appropriately for their different customer base.

Data mining is the identification of patterns that enable derivation of meaningful information from a dataset. Such predictions provide a probable picture of the future using historical data. These futuristic outlooks can serve as a guide for making beneficial decisions in the present. Classification is a type of data mining algorithm that creates a model on which future records can be evaluated. A division of the dataset into two subsets is initially made. One part of the dataset is the training set and the other portion is the testing set. The training set is the portion of the data used to generate a model that is used to predict future values. The testing set, the set of data unseen by the model, is used to test the model with the idea that it is representative of the population and eventually also future instances.

A classification model can be utilized to improve bank decision-making. For example, predicting clients most and least likely to subscribe will allow a bank to prioritize the customers to contact for each subscription offer in order to maximize total number of subscriptions in less time. In addition, the ability to predict client information such as age group or education level will enable the bank institution to tailor telemarketing strategies to those customers. Overall, it will increase the bank's focus to areas that are likely to cause most efficient usage of company resources.

1.2 Scope

The scope of this thesis includes applying data mining classification techniques on bank client data to determine predictability of several classes related to the client's demographic and financial situation by the chosen algorithms. The demographic attributes include age, employment, marital status and education level; the financial attributes include housing loan, personal loan and term deposit. In addition, the attributes contributing most to the class will be derived. Predictability will be measured by precision, recall and F1 score. The algorithms used will include logistic regression, random forest and SVM.

1.3 Outline

The remainder of the thesis is organized in the following format. Chapter 2 covers the background information, which is the foundation knowledge on which the thesis is based. This includes the dataset analyzed in the thesis and the tool used for model generation. It explains the details of the three algorithms utilized for analysis. In addition, it covers the evaluation metrics used to compare the results. Chapter 3 summarizes the work that was performed previously by others on the dataset. Chapter 4 focuses on the steps performed to implement the thesis. It includes the preprocessing steps used to prepare the data as well as the actual model generation and model testing process. Chapter 5 discusses the process used for evaluation as well as the results obtained. Chapter 6 concludes the thesis work and proposes further research ideas.

CHAPTER 2

BACKGROUND

2.1 Introduction

This chapter discusses the foundational knowledge on which this thesis is based. It covers the details of the dataset analyzed and provides background information on the algorithms used for analysis. It discusses the data mining tool used to generate models for each of the algorithms as well as the evaluation metrics applied when comparing the results.

2.2 Data

The data used for this thesis consisted of a multivariate dataset from a Portuguese bank that contains client information as well as the result of telemarketing phone calls for subscription to a term deposit. The dataset contains 41,188 instances with twenty-one attributes of which the original prediction class is the client subscription to a term deposit [1].

All of the attributes can be categorized into five distinct categories: the client's demographic information (CD), the client's financial information (CF), items related to the current marketing campaign (CM), items related to a prior marketing campaign (PM) and the social economic situation (SE). The distribution of each of these category types is shown in Figure 1. Items from the previous marketing campaign is the least represented with 3 attributes and the social-economical situation as well as the current marketing campaign is the most with 5 attributes each. These category types are used for comparison during analysis.

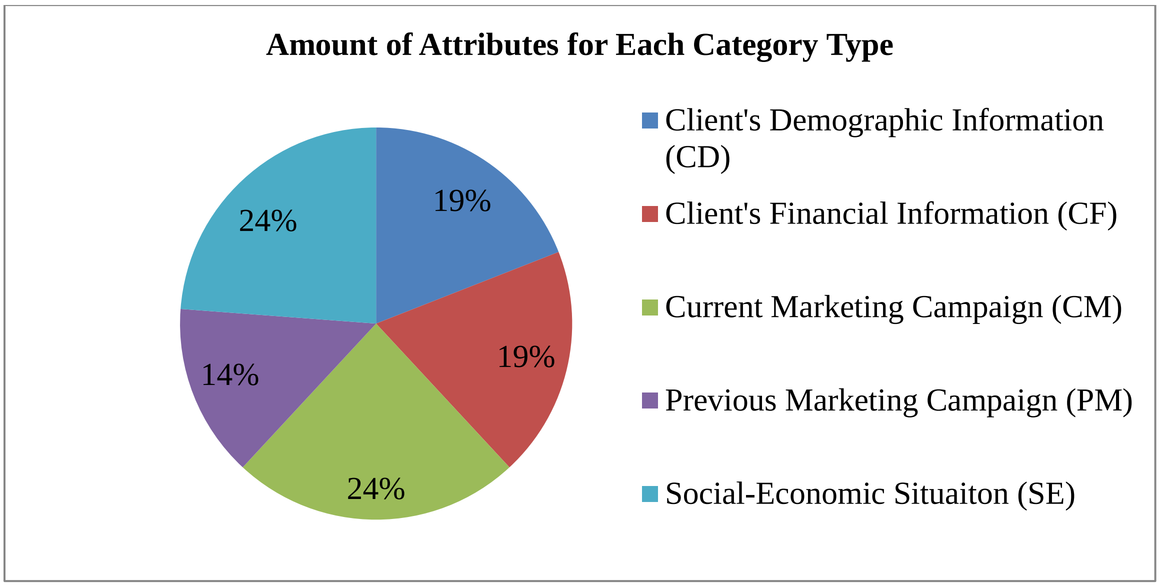


Figure 1. All Dataset Attributes Categorized into Different Types with the Percentage of Each Type Shown as a Pie Chart.

A detailed description of the dataset attributes is shown in Table 1. The first column in the table is a numerical value representing the column of the data in the dataset. The second column contains the name of the attribute. The third column shows a brief description of the attribute. The fourth column classifies the type of the data attribute as either numeric or categorical represented by an N or C respectively. The fifth column describes the category of the attribute. The last column describes the values that are contained in the attribute.

Several data instances in the dataset contain unknown values. These values will need to be imputed or ignored during evaluation. A few attributes are of the continuous or numerical type. These values will need to be discretized into a smaller number of categories. Also, several attributes have values that do not exhibit close to equal representation compared to other values in the attribute. Since these attributes are also used as the class variable, a measure suitable for imbalanced dataset will be required for more proper evaluation.

Table 1. Description of Attributes in the Original Dataset

	Attribute	Description	Type	Category	Values
1	age	age of the client	N	CD	[17, 98]
2	job	type of job	C	CD	{admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown}
3	marital	marital status	C	CD	{divorced, married, single, unknown} (divorced means divorced or widowed)
4	education	education level of client	C	CD	{basic.4y, basic.6y, basic.9y, high.school, illiterate, professional.course, university.degree, unknown}
5	default	has credit in default	C	CF	{no, yes, unknown}
6	housing	has housing loan	C	CF	{no, yes, unknown}
7	loan	has personal loan	C	CF	{no, yes, unknown}
8	contact	last contact communication type	C	CM	{cellular, telephone}
9	month	last contact month of year	C	CM	{jan, feb, mar, ..., nov, dec}
10	day_of_week	last contact day of the week	C	CM	{mon, tue, wed, thu, fri}
11	duration	last contact duration in seconds	N	CM	[0,4918]
12	campaign	number of contacts performed during this campaign and for this client including last contact	N	CM	[1, 56]
13	pdays	number of days after client was last contacted from previous campaign	N	PM	[0, 27], {999} (999 means client was not previously contacted)
14	previous	number of contacts performed to this client before this campaign	N	PM	[0,7]
15	poutcome	outcome of the previous marketing campaign	C	PM	{failure, nonexistent, success}
16	emp.var.rate	employment variation rate quarterly indicator	N	SE	[-3.4,1.4]
17	cons.price.idx	consumer price index monthly indicator	N	SE	[92.201,94.767]
18	cons.conf.idx	consumer confidence index monthly indicator	N	SE	[-50.8,-26.9]
19	euribor3m	euribor 3 month rate daily indicator	N	SE	[0.634,5.045]
20	nr.employed	number of employees quarterly indicator	N	SE	[4963.6,5228.1]
21	subscription	subscription to a term deposit	C	CF	{yes, no}

2.3 Algorithms

The classification algorithms examined in this thesis include support vector machine (SVM), random forest and logistic regression. All of these algorithms provide a different method to allow model generation for classification of future data instances.

Logistic regression is a regression technique that analyzes the relationship between the various attributes. The class may be continuous or categorical but predictions are made on a binary class. The data is first split into a positive and negative class and logistic regression is run. The goal of logistic regression is to find the best fitting model that will describe the relationship between the inputs and the class. The log odds of the outcome is modeled as a linear combination of the predictor variables. A prediction is made of the probability of the response based on several predictor variables that are independent. Logistic regression generates coefficients as well as standard errors and significance levels of a formula to predict a logic transformation. Instead of the selecting parameters that minimize the sum of squared errors as performed in ordinary regression, logistic regression estimation chooses parameters that maximize the likelihood of observing the sample values. [2]

Multinomial logistic regression is performed with more than two values in the class. This type of logistic regression can be done using the LIBLINEAR libraries. LIBLINEAR is a classifier library built for large datasets. It supports both binary and multi-class types of logistic regression, where multi-class is implemented using the one-vs-the-rest strategy. [3]

SVM is a classification technique based on the concept of decision planes that define decision boundaries. It is a supervised learning algorithm that aims to map the data into space and divide it with a maximized clear boundary. A training dataset identifies the decision boundaries and classifies each bounded area to a specific target value. New instances or records that fall into one of the classification bounded areas will then be categorized as the target value specified for that bounded area. Therefore, all new data points are predicted to belong to one of the divided sides. During training when boundaries are being identified there may be several decision boundaries that can be made to separate two different spaces that is expected to perform equally well on unseen data. In such instances, the decision boundaries with large margins are selected as they tend to have better generalization errors, than those with small margins. Classifiers that produce decision boundaries with small margins are more prone to model overfitting and tend to generalize poorly on unseen data. Therefore, SVM is an optimization algorithm which selects the boundary with the maximum margin. It does not use a greedy-based strategy, which typically finds the local optimal solution, but rather finds the global optimal solution. Depending upon the data, these boundaries may be linear or non-linear. Non-linear SVM is performed by the use of kernel tricks, which essentially enable the mapping of the inputs into a multi-dimensional feature space. SVM can be applied to categorical data by attributing each categorical value to a numerical value. [4] The LibSVM library enables SVM classification, regression as well as distribution estimation. It also supports multi-class classifications. The library provides several kernels for use including linear, polynomial, radial basis function and sigmoid. [5]

Random forest is a class of ensemble methods that generates multiple decision trees from the training set. Ensemble methods are techniques that improve classification accuracy by aggregating the predictions of multiple classifiers. An ensemble method creates a set of base classifiers using training data. It then performs classification by taking a vote on the predictions that are made by each base classifier. For an ensemble method classifier to outperform a single classifier, two conditions should be met. The base classifiers should all be independent of each other and the base classifiers should make predictions better than random guessing. Random forest combines predictions from many different decision trees with each tree constructed using values of an independent set of random vectors. First, the original training data is used and randomization is applied. Randomization in random forest helps to reduce the correlation among the decision trees so that the generalization error can be improved. For example, a set of random vectors may be created, where each will be independently used to create a decision tree. The second step is to use the randomized data to build multiple decision trees. Finally a combination of these decision trees yields the final predictions. [4]

2.4 Tool

Orange is an open source data mining tool developed by Bioinformatics Lab in the University of Ljubljana. [6] It is an accepted tool for data mining and predictive analytics. Its popularity stems from its ability to cater to both novice as well as expert users. It allows users to model the complete workflow of a typical data mining process as a diagram using a graphical user interface. This includes preprocessing data, applying an algorithm on it and performing the actual calculations and analysis. Each action of the process is represented by an object called a widget. Widgets exist for several data mining algorithms as well as common actions such as data imputation, discretization, data visualization via a distribution graph, and confusion matrix for the results. Once a diagram model is created, it can be saved for later use on any dataset. In addition to the visual application of data mining algorithms on datasets through the use of widgets, Orange also allows the user to create personalized Python scripts for specific tasks. Since Orange supports multiple classifications algorithms required for this thesis, this will be the data mining tool of choice.

The tool expects a specific format for the input dataset. It is capable of reading a tab-delimited text file that has three header rows. The first row has the names of the attributes, the second contains the domain type including continuous, discrete or string, and the last contains the type of attributes including class, meta or string. The dataset for this thesis needs to be converted to a format compatible with Orange since the current format has data instances in each row and each of the values for a data instance are semicolon delimited.

Orange allows various types of sampling to be performed. This includes cross-validation, leave-one-out, random sampling, testing on the train data and testing on the test data. Cross-validation is a technique that splits the data into a specific number of pieces called folds. The first fold is left out to be used for classification and the model is created from using the other folds. This is repeated for all folds until the full dataset has been classified. Leave-one-out is a similar process as cross-validation but the number of items in the fold is a single data instance. For this reason, this method is very slow. Random sampling splits the dataset into a testing set and training set where the size of both sets can be user specified. Then the full model creation process is repeated for a given number of times. Test on train data is a strategy that uses the full dataset for training and then also uses the same full dataset for testing. Since the full dataset is used both times, this technique gives very good results but may not be as successful on predicting previously unseen data instances. The test on test data strategy uses two separate datasets as input. One dataset is used for training and the other dataset is used for testing.

In addition to the variety of sampling techniques, several metrics are also provided for analyzing the result of a specific classifier. The metrics include sensitivity, specificity, area under the ROC curve, information score, F1 score, precision, recall, brier score and the Matthews correlation coefficient. Orange provides these values per category of each test class.

2.5 Evaluation Metrics

The results of predictive models can be viewed in the form of a confusion matrix. A confusion matrix is a table that displays the number of instances that are correctly and incorrectly classified in terms of each category within the attribute that is the target class. The positive class is with respect to the current category and the negative class includes all categories other than the current. The confusion matrix displays the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) values for a given attribute. TP is the number of values predicted to be positive by the algorithm and was actually positive in the dataset. TN represents the number of values that are expected to not belong to the positive class and actually do not belong to it. FP depicts the number of instances misclassified as belonging to the positive class thus is actually part of the negative class. FN shows the number of instances classified as the negative class but should belong to the positive class. Figure 2 below shows a confusion matrix where the columns represent the prediction and the rows are the actual classification.

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

Figure 2. Confusion Matrix Where the Columns Represent the Prediction and the Rows Represent the Actual Classification

A common evaluation metric for algorithms is classification accuracy, which is simply referred to as accuracy. Accuracy can be derived from the TP, TN, FP and FN values of a confusion matrix. The equation for accuracy, shown below in Equation 1, identifies the ratio of all values that were correctly classified based on both the positive and negative class over the total number of instances examined. Since the classification accuracy includes values from both the positive class as well as the negative class, the value is consistent for an attribute regardless of the category from which it is extracted.

$$\boldsymbol{Classification\ Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FP + FN}} \tag{1}$$

Accuracy exhibits a phenomenon known as the accuracy paradox. The accuracy paradox states that “predictive models with a given level of accuracy may have greater predictive power than models with higher accuracy” [7]. A useless model, one that predicts only the positive class or only the negative class, can have higher accuracy than a model with some predictive power. Predictive power is the power to make a good prediction. For example, if a model only predicts one class, it has extremely low predictive power. This can be illustrated by the following scenario. Consider the confusion matrices in Figure 3 below. Examining the matrix on the top, the accuracy of the model is accuracy = (100 + 10)/(100 + 50 + 5 + 10) = 66.7%. Now consider the confusion matrix on the bottom which always predicts the negative class. The accuracy of this matrix is accuracy = (150 + 0)/(150 + 0 + 15 + 0) = 90.9% which is 24.2% higher than from the confusion matrix with more predictive power. Thus, even though this has higher accuracy it is useless as a predictive model since it always predicts the same class. As a general rule, “when TP < FP, then accuracy will always increase when we change a classification rule to always output ‘negative’ category. Conversely, when TN < FN, the same will happen when we change our rule to always output ‘positive’.” [8]

	Predicted Negative	Predicted Positive
Actual Negative	100	50
Actual Positive	5	10

	Predicted Negative	Predicted Positive
Actual Negative	150	0
Actual Positive	15	0

Figure 3. Confusion Matrix of a Model with Some Predictive Power (Top) and a Confusion Matrix of a Model with Zero Predictive Power (Bottom) as Items Are Always Classified as Part of the Negative Class.

Thus, all models are not suitable to be evaluated using accuracy. Accuracy is more suited for datasets that contain balanced positive and negative classes. For imbalanced datasets, other metrics such as precision and recall are more desirable. [9] Precision represents the amount of results that are relevant while recall is a measure of the amount of relevant results returned. A value of 1 is the highest possible for both measures, while 0 is the lowest measure. Both these values are dependent on the category being analyzed within the target class. Precision is shown in Equation 2 and recall is shown in Equation 3 below. The concepts or precision and recall are illustrated in Figure 4.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}} \tag{2}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}} \tag{3}$$

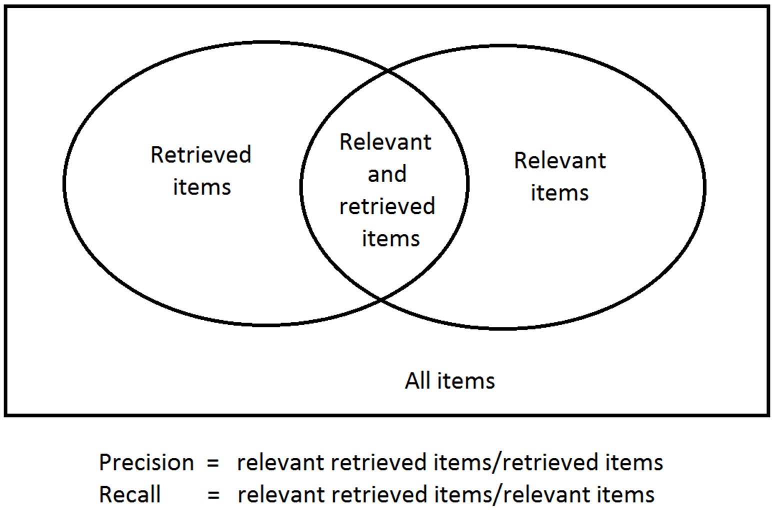


Figure 4. Precision and Recall

Precision says nothing about the data instances not correctly classified and recall says nothing about the data instances incorrectly labeled as the positive class. Thus both values are often examined as this information is more valuable. However it may be difficult to increase both values together. For example, if the TP of a minority class is increased the number of FP may also increase, which in turn reduces precision. [9] As a result, a single measure that is a combination of both measures is more ideal. This measure, known as the F1 score, is a harmonic mean of precision and recall where both precision and recall are weighted equally. The ideal classification algorithm will exhibit high precision, recall and F1 scores values. The equation for F1 score is shown in Equation 4 below.

$$\boldsymbol{F1\,Score = 2 * \frac{Precision * Recall}{Precision + Recall}} \qquad (4)$$

CHAPTER 3

RELATED WORK

3.1 Introduction

Other researchers have also used the same dataset for data mining analysis. This section describes the work performed in those papers.

3.2 Examination of Related Work

Moro, Cortez and Rita [10] used this bank dataset in addition to an external dataset to determine the best set of features and analyze different data mining models on the term deposit subscription class. Research was conducted by first combining the dataset with statistical data from a website belonging to the central bank of the Portuguese Republic. This external dataset allowed the inclusion of bank client information, product information as well as data related to social and economic information. With the combination of the two datasets, a total of 150 features were created. Feature selection was performed on different sets of features and compared by two metrics including area of the receiver operating characteristic curve (AUC) and area of the LIFT cumulative curve (ALIFT). The feature set of 22 features was used for further analysis to compare four algorithms. Logistic regression, decision trees, SVMs and neural networks were applied on the reduced set of features and the results showed that neural networks had the best value with an AUC value of 0.8 and ALIFT value of 0.7.

The term deposit subscription attribute of this dataset was also analyzed using a combination of business intelligence (BI) and data mining techniques. According to Moro, Laureano and Cortez [11] “BI is an umbrella term that includes architectures, tools, databases, applications and methodologies with the goal of using data to support decisions of business managers”. The CRoss-Industry Standard Process for Data Mining (CRISP-DM) model was used. This methodology defines the process of generating a model that can be used for predicting in real life. It has six phases which include business understanding, data understanding, data preparation, modeling, evaluation and deployment. The business understanding phase is used to define a business goal which needs to be achieved by generating a predictive model. The data understanding, data preparation, modeling and evaluation phases are similar to the data collection and preprocessing, model creation, and analysis phases followed in a typical data mining process. The last phase of this step is deployment of the model in the real world. Based on the application of the CRISP-DM methodology, SVM displayed the highest predictive power as compared to naïve bayes and decision trees when measured using AUC and ALIFT.

Vajiramedhin and Suebsing [12] compared three different sets of features to determine the best model of feature selection using the term deposit subscription attribute. The first comparison was done with the full dataset of 20 features and one target class with no techniques applied. This method showed an 88.4% ROC rate. The second comparison was made on the dataset with three features which were derived using a feature subset selection algorithm that is correlation-based. This method had a ROC rate of 91%, which is a 2.6% improvement from the first model. The last model combined the feature subset selection algorithm that is correlation-based with a dataset balancing technique to select eight features for the model. This technique yielded a ROC rate of 95.6% which was a 4.6% improvement from model 2 and a 7.2% increase from model 1. As a result, this method was the best prediction model based on the ROC metric.

Another paper by Elsalamony [13] used the dataset with the goal of determining influencing attributes on the term deposit subscription attribute. The algorithms used were multilayer perception neural network (MLPNN), Bayesian Networks, Logistic Regression, and C5.0. The metrics used for analysis included classification accuracy, sensitivity, and specificity. The results showed that the duration of the last conversation was the most influencing factor on success of the client’s subscription to the term deposit for C5.0, Logistic Regression, and MLPNN. According to Bayesian Networks the most influencing attribute was the client’s age.

CHAPTER 4

IMPLEMENTATION

4.1 Introduction

Typical steps involved in the data mining process generally include data collection, data preprocessing, model generation and evaluation. Data collection is the process of gathering all of the data instances to generate a dataset. Data preprocessing modifies the dataset to improve quality and provide more meaningful inputs to the data model. Data model generation includes creating a model by applying data mining algorithms onto the preprocessed dataset. The preprocessing and model generation step can be repeated or varied to extract more meaningful information from the dataset. Evaluation includes using metrics to compare the predictive power of the algorithms applied on a particular dataset.

Data preprocessing involves techniques such as aggregation, sampling, discretization, variable transformation and dimensionality reduction through feature subset selection and feature creation. Aggregation is the combination of data objects, which are the actual data instances, into a single data instance. One example where this would be useful is in the combination of multiple store transactions to a single data instance to represent the store when evaluating transactions of several different stores. Sampling means using a representative subset of the dataset most often to avoid the time and expense of utilizing the full dataset. Discretization and categorization involves reducing the number of categories associated with a categorical attribute and generating categories for continuous attributes. This is especially useful for algorithms which require only categorical attributes. Variable transformation is a transformation applied to each value of an attribute. An example of this is to take the absolute value of the values when only the magnitude is needed. Dimensionality reduction is a technique applied on a dataset with a large number of attributes in order to remove irrelevant features that do not aid in pattern identifying within the dataset. Feature subset selection achieves dimensionality reduction by utilizing only a subset of the features available in the dataset. Feature creation involves creating a completely new set of attributes from the current attributes. [4]

Data used in this thesis has already been collected by the banking institution. Some data preprocessing techniques were applied on the dataset to improve the quality of the data model generated. Models were generated for three classification algorithms including SVM, random forest and logistic regression. The resulting precision, recall and the F1 score were collected. This process was applied on several attributes include age, job, marital status, education, housing loan, personal loan and term deposit subscription.

4.2 Data Preprocessing

The original dataset was preprocessed to improve data quality. Preprocessing techniques of discretization and categorization were applied on several attributes after examining results from the initial iteration where models were generated using the original dataset. In addition, several attributes which were continuous were converted to categorical mainly because some algorithms, including logistic regression, can only be applied on categorical attributes. Also, attributes with a very large number of categories were combined into a single attribute. The sampling technique was not applied in this dataset since the data was of manageable size. Thus, the results are representative of the full dataset. Dimensionality reduction was also not applied since the number of attributes was significantly smaller than the number of data instances. In addition to preprocessing techniques, all data instances with unknown values in a class were imputed and all unknown values for attributes other than the class were imputed by the most frequently occurring value. Finally, the dataset required formatting in a way that is acceptable by the Orange data tool.

The original dataset for the Age Group attribute had age ranges from 17 to 98. Simply converting each of the age values into a single category would create 78 unique categories with some categories being represented by as little as a single instance. This is illustrated in Table 2 where all values in the dataset are listed and followed by the number of occurrences of that value in square braces. Having multiple categories makes it challenging to determine how the attribute affects the test class because some categories are significantly underrepresented.

Table 2. All Unique Categories for Age Group with the Number of Instances for Each Category in Square Braces. A Total of 78 Unique Categories Exist and the Representation of Those Categories Range from 1 Instance to 1947 Instances.

Age (Original Distribution)									
Value [# of occurrences in dataset]									
17[5]	25[598]	33[1833]	41[1278]	49[839]	57[646]	65[44]	73[34]	81[20]	89[2]
18[28]	26[698]	34[1745]	42[1142]	50[875]	58[576]	66[57]	74[32]	82[17]	91[2]
19[42]	27[851]	35[1759]	43[1055]	51[754]	59[463]	67[26]	75[24]	83[17]	92[4]
20[65]	28[1001]	36[1780]	44[1011]	52[779]	60[283]	68[33]	76[34]	84[7]	94[1]
21[102]	29[1453]	37[1475]	45[1103]	53[733]	61[73]	69[34]	77[20]	85[15]	95[1]
22[137]	30[1714]	38[1407]	46[1030]	54[684]	62[62]	70[47]	78[27]	86[8]	98[2]
23[226]	31[1947]	39[1432]	47[928]	55[648]	63[55]	71[53]	79[14]	87[1]	
24[463]	32[1846]	40[1161]	48[979]	56[704]	64[57]	72[34]	80[31]	88[22]	

Thus to reduce the high number of unique attributes in the original dataset the values were initially bucketed into 10 categories. The first category included all values that were less than 25 and the remaining categories were 5 year increments up to the last category which included values of 65 and greater. The evaluation metrics did not yield very high results with these initial categories. As a result, the data values were further bucketed into one of three categories called ‘young’, ‘working’ and ‘retired’ since these are well accepted age group divisions. Young individuals include those under the age of 25. Retired individuals are those who are 65 and over since that was the retirement age in Portugal during the years in which the data was collected [14]. The rest of the results belong to the ‘working’ category. The resulting data distribution after preprocessing is in Table 3.

Table 3. Data Distribution and Categories of Age Group after Preprocessing

Age Group (Preprocessed Distribution)		
Attribute Value	**Attribute Details**	**Number of Occurrences**
Young	<25	1068
Working	25-64	39457
Retired	65+	663
Total instances used		41188

For the Employment Status attribute there are 11 unique categories in the original dataset as shown in Table 4. When these were used directly for analysis the precision, recall and F1 scores were varying extremely for the different categories and several of the categories had a 0 value for those. To improve the precision, recall and F1 score for Employment Status, the data instances were bucketed into two categories which include 'employed' and 'unemployed'. Individuals with the retired or student status are assumed to be unemployed. Individuals from any other profession category are assumed to be employed. The data distribution of the final changes to this attribute is shown in Table 5. As expected, a significantly larger number of individuals are employed as compared to those who are unemployed.

Table 4. All Unique Categories for Employment Status and the Number of Instances for Each Category. A Total of 11 Unique Categories Exist.

Employment Status (Original Distribution)	
Attribute Value	**Number of Occurrences**
admin	10422
blue-collar	9254
entrepreneur	1456
housemaid	1060
management	2924
Retired	1720
self-employed	1421
Services	3969
Student	875
technician	6743
unemployed	1014
unknown	330
Total instances used	40858

Table 5. Data Distribution and Categories of Employment Status after Preprocessing

Employment Status (Preprocessed Distribution)	
Attribute Value	**Number of Occurrences**
Employed (admin, blue-collar, entrepreneur, housemaid, management, self-employed, services, technician)	37249
Unemployed (retired, student, unemployed)	3609
Unknown	330
Total instances used	40858

The original Marital Status attribute has three categories: ‘divorced’, ‘married’, and ‘single’ as shown in Table 6. The ‘divorced’ category includes those who are divorced or widowed. When using these categories, the results in the initial run yielded very low F1 score for the ‘divorced’ category. Therefore, the ‘divorced’ category was later combined with the ‘single’ category which provided a more balanced class. The preprocessed data distribution is shown in Table 7.

Table 6. Data Distribution and Categories of Marital Status

Marital Status (Original Distribution)	
Attribute Value	**Number of Occurrences**
divorced (includes widowed)	4612
married	24928
single	11568
unknown	80
Total instances used	41108

Table 7. Data Distribution and Categories of Martial Status after Preprocessing

Marital Status (Preprocessed Distribution)	
Attribute Value	**Number of Occurrences**
married	24928
unmarried (single, divorced and widowed)	16180
unknown	80
Total instances used	41108

The data distribution of the original dataset for the Education Level attribute is shown in Table 8. There are a total of seven categories. The ‘illiterate’ category has very low representation but the other attributes have a good representation. To reduce the number of categories some values were grouped together.

Table 8. All Unique Categories for Education Level and the Number of Instances for Each Category. A Total of 7 Unique Categories Exist

Education Level (Original Distribution)	
Attribute Value	**Number of Occurrences**
basic.4y	4176
basic.6y	2292
basic.9y	6045
high.school	9515
illiterate	18
professional.course	5243
university.degree	12168
Unknown	1731
Total instances used	39457

Table 9. Data Distribution and Categories of Education Level after Preprocessing

Education Level (Preprocessed Distribution)	
Attribute Value	**Number of Occurrences**
Lower.degree (basic.4y, basic.6y, basic.9y, high.school)	22028
professional.course	5243
university.degree	12168
Unknown	1731
Total instances used	39439

To increase the data quality, the categories in Education Level were bucketed into four categories: lower degree, university degree, professional course and illiterate. With only 18 occurrences total, the 'illiterate' category was not contribution to any useful information and thus this value was ignored. The final data distribution is shown in Table 9.

Data distributions for Housing Loan, Personal Loan and Term Deposit are shown in Table 10, Table 11, and Table 12 respectively. Each of these attributes contains only two categories including 'no' and 'yes' and both categories are well represented in the dataset. Thus, these values were not modified.

Table 10. Data Distribution and Categories of Housing Loan

Housing Loan	
Attribute Value	**Number of Occurrences**
no	18622
yes	21576
Unknown	990
Total instances used	40198

Table 11. Data Distribution and Categories of Personal Loan

Personal Loan	
Attribute Value	**Number of Occurrences**
no	33950
yes	6248
unknown	990
Total instances used	40198

Table 12. Data Distribution and Categories of Term Deposit

Term Deposit	
Attribute Value	**Number of Occurrences**
no	36548
yes	4640
Total instances	41188

4.3 Imputation Analysis

Different imputation methods were analyzed using a technique known as sensitivity analysis. Sensitivity analysis is a method in which pure black box testing is performed with different inputs and the results are used for determining parameters to use in the final analysis. The term deposit subscription class was used for these tests. Different conditions were compared for imputation of the dataset. The first method imputed the attributes by average/most frequent and also imputed the class; the second approach was to not impute the attributes and not impute the class. The result of not imputing the dataset as shown in Figure 5 did not show a difference from imputing it. This may be because the dataset is very large and imputing by the value that is already most common does not further add more information in determining patterns. Imputation was still chosen for all attributes to allow for running the python code that prints the weights of logistic regression and SVM.

	NO			YES		
	F1	Precision	Recall	F1	Precision	Recall
Random Forest	0.94	0.90	0.99	0.21	0.66	0.12
Logistic Regression	0.95	0.93	0.97	0.51	0.66	0.42
SVM	0.94	0.91	0.99	0.30	0.65	0.19

Figure 5. Results from Not Imputing the Attributes and Not Imputing the Class for Term Deposit.

4.4 Training and Testing

Once the data preprocessing part was complete, Orange was used for model generation and testing. The workflow model created for this dataset is shown in Figure 6 below. The File widget read in the preprocessed file and fed in the data to the Impute widget. This widget enabled data imputation of instances in the features as well as the class. The imputed results were read in by Logistic Regression, SVM and the Random Forest widgets. The learners obtained for logistic regression and SVM were sent to a Python Script Widget in order to print out the model details. The learner was passed on from the Python Script Widget to the Test Learners widget to evaluate the models and display the results in a tabular format. The learner for random forest was sent to the Test Learners widget directly and was also sent to a Classification Tree Graph widget to display the tree model generated. Results of the Test Learners widget were read by the Confusion Matrix widget in order to display the confusion matrix. This process was repeated for all of the test classes of interest.

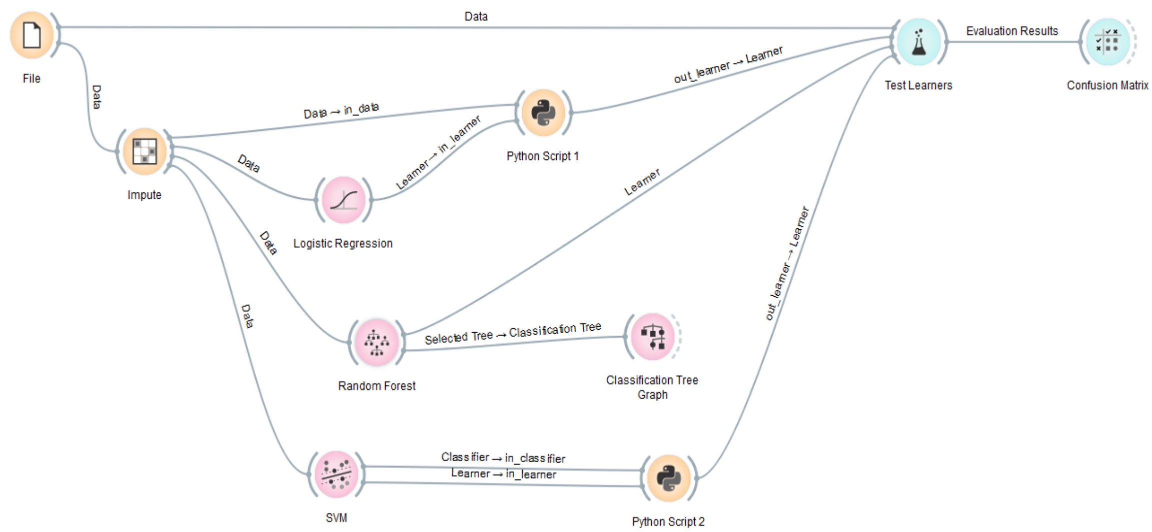


Figure 6. Data Flow Model Created Using Orange to Depict the Process Flow of the Data from Input to Evaluation.

Figure 7 on the left side shows an example of the inputs used in the File widget. This widget read in the preprocessed file in a tab separated format with the .tab file extension. Once it reads in the data file, it calculates the number of data instances as well as the number of features. It also determines the type of class, which can be discrete or continuous, and the number of different categories in the class. Symbols are used to represent missing values that are used in the Impute widget. Different symbols were used to differentiate between the "Don't care" and "Don't know" types of missing values. The recommended settings were applied for the new attributes selection, which allows the creation of new attributes when multiple files are used for input.

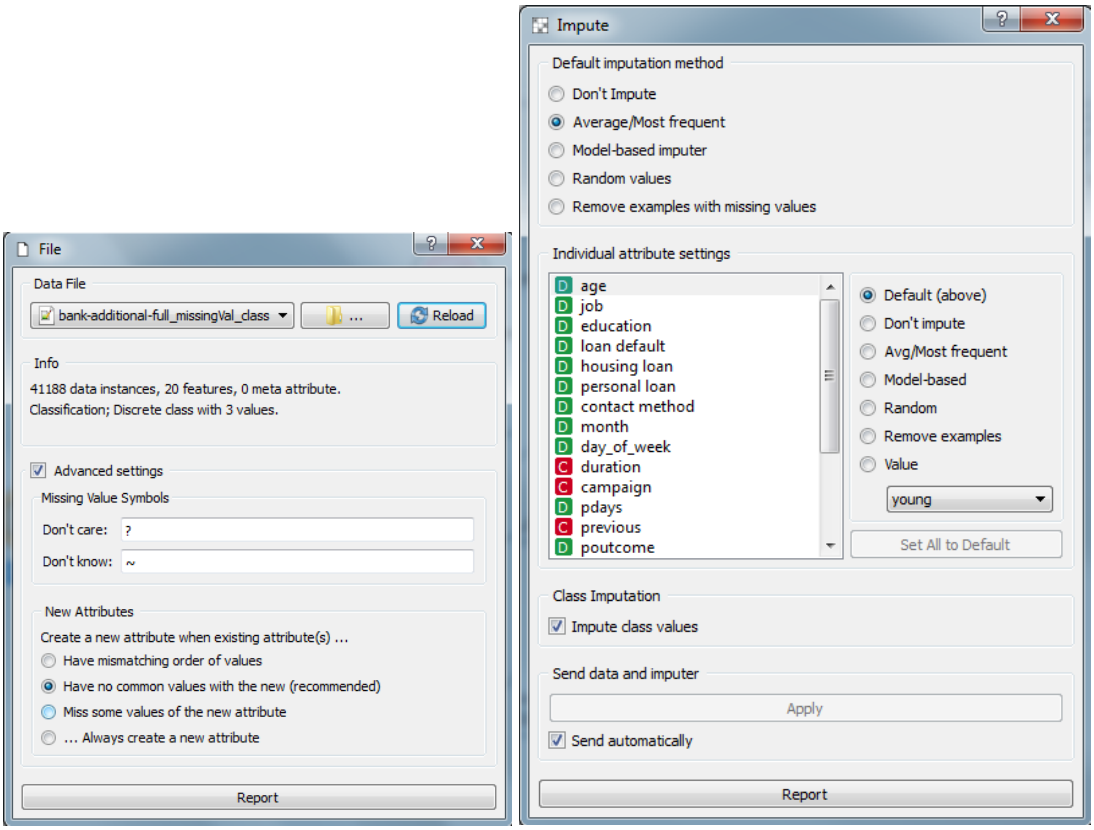


Figure 7. Left: Inputs to the File Widget for the Marital Status Test Class. Right: Settings Used for the Impute Widget.

The data instances were imputed using the Impute widget. Orange allows the imputation of missing values from both the features as well as the class. For each of the features, any instances with missing values are set to be replaced with the average or most frequent value in the attribute. Although this option can be customized for each category within each class, this was the setting used for all of the categories in all classes. All instances with missing values in the class were also imputed. The settings used are shown in Figure 7 on the right.

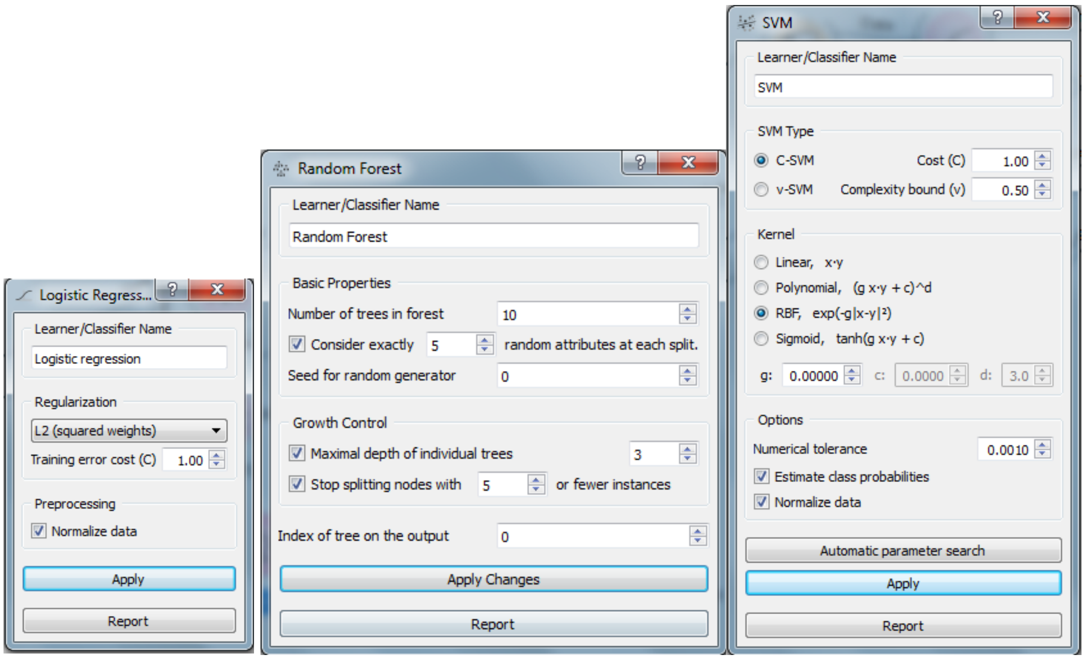


Figure 8. Parameters Used for Logistic Regression, Random Forest and SVM Widgets

Figure 8 shows the parameter inputs used for logistic regression, random forest and SVM respectively. Each of them read in the name of the classifier which is used when representing the results in the Test Learners and Confusion Matrix widgets. For logistic regression, L2 regularization was chosen with a training error cost of 1. The data was also normalized. Random forest was applied with ten trees in the forest where exactly five attributes were considered at each split. The growth was controlled by allowing up to three levels in each individual tree and not splitting nodes that have five or less instances. SVM was run with using C-SVM with cost value 1 and a RBF kernel. The data was also normalized.

```
# create the classifier out of the learner
learner = in_learner
c = learner(in_data)

# print all the features and the weights
print "\nFeature:Weight"
for feat, w in zip(c.domain.features, c.weights[0]):
    print feat.name, ":", w

# send the learner to the Test Learners widget
out_learner = in_learner
```

Figure 9. Code for Python Widget 1 to Print Model Information for Logistic Regression

```
from Orange.classification import svm

# get the classifier and weights
classifier = in_classifier
weights = svm.get_linear_svm_weights(classifier)

# print the attribute and its weight
print "\nAttribute:Weight"
for attr, w_attr in weights.items():
    print str(attr) + ":" + str(w_attr)

print "\nSorted weights:"
print sorted("%.4f" % w for w in weights.values())

# send the learner on to the Test Learner widget
out_learner = in_learner
```

Figure 10. Code for Python Widget 2 to Print Model Information for SVM

Python code was used to display the model information for logistic regression and SVM. Figure 9 shows the code for Logistic Regression. The code reads in the learner and generates the classifier based on the learner. The classifier is used to print each feature and weight. Finally, the unmodified learner used as input to the script was forwarded as output of the script. The code used for SVM follows a similar process as shown in Figure 10. Since the classifier is already provided as input, the weights are directly calculated. Each attribute and weight is printed to the console followed by the weights in sorted order. The learner fed into the widget is again forwarded to the next widget unmodified.

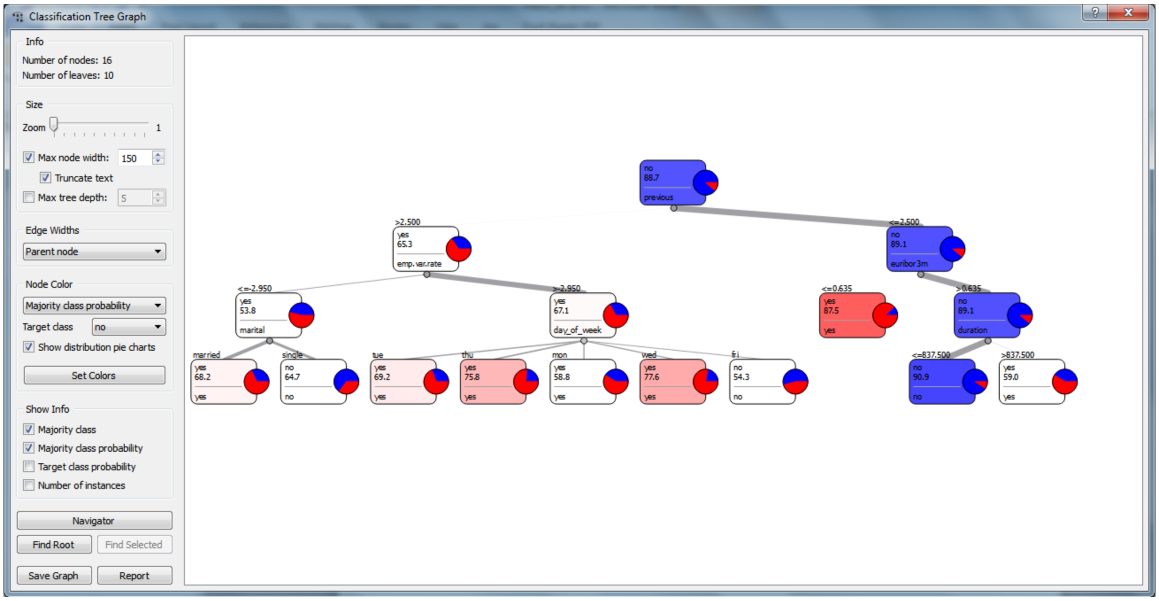


Figure 11. Classification Tree Graph Widget to Depict the Data Model Generated by Random Forest

The Classification Tree Graph widget shown in Figure 11 was used to view the tree that was created from Random Forest. The settings for this widget applied in this thesis are shown on the left side. The total number of nodes and leaf nodes are also shown on the top. Examining the tree visible on the left side, the nodes are color coded based on the majority class of that node. A pie chart of the instances in the node is also shown in each of the nodes. The first value in any node is the category belonging to the majority class in that node. The numerical value that follows it is the percentage of instances that belong to the majority class. The last value is the attribute used for splitting the node into further nodes. If a node is the leaf node, the last value matches the majority class. When a node is split, the category or range by which it is split is shown above the node.

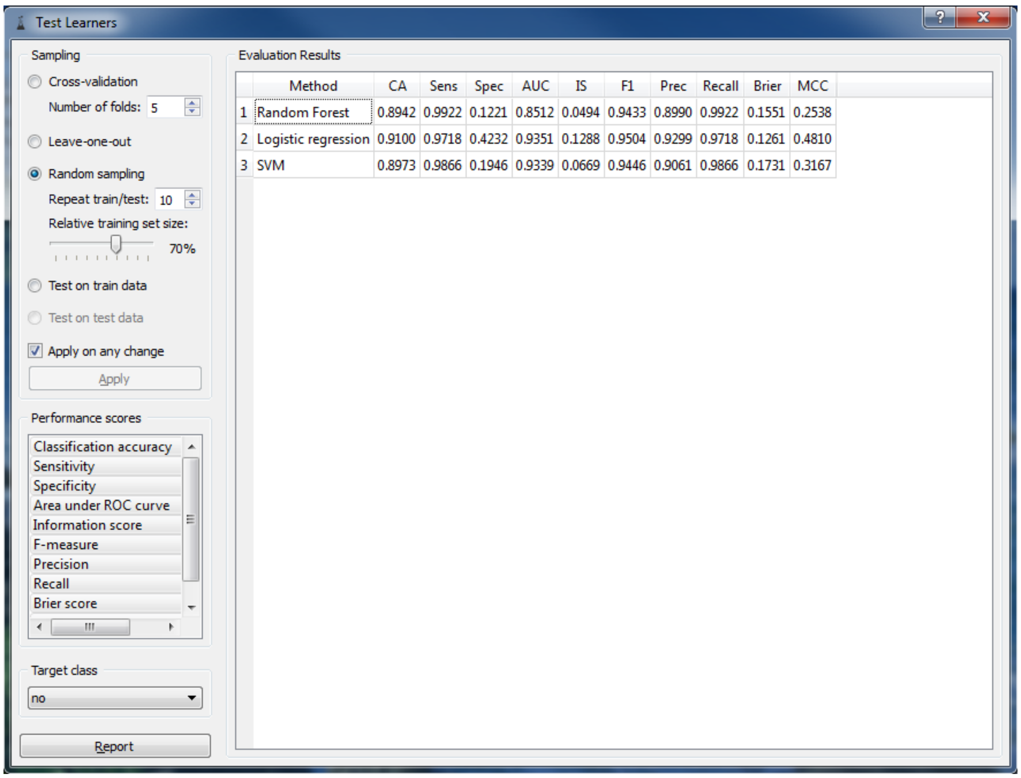
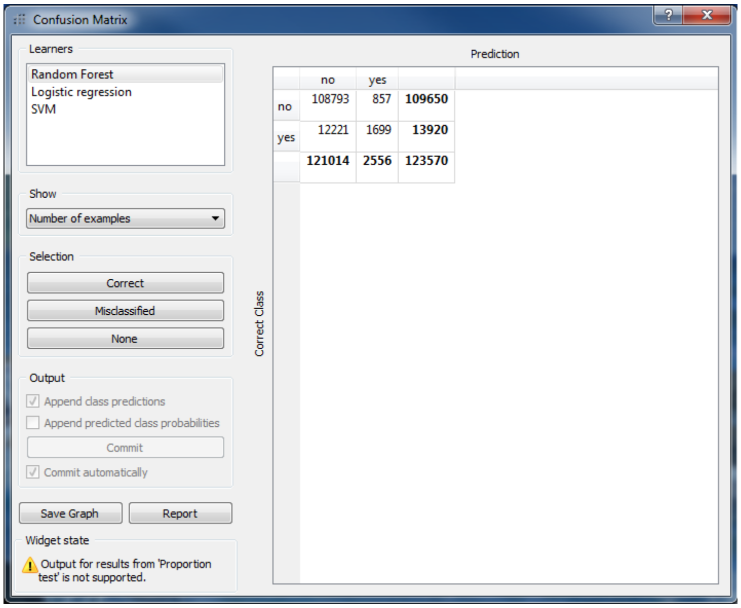


Figure 12. Sampling Settings for the Test Learners Widget and the Results for Logistic Regression, Random Forest and SVM on the ‘No’ Category of the Term Deposit Class

Results from analysis of the data model were shown in the Test Learners and Confusion Matrix widgets. The settings used for the Test Learners widget on the ‘employed’ category of the Term Deposit class is shown in Figure 12. The model was created and tested using a random sampling technique with a training set size of 70% and testing set size of 30% where the training and testing was repeated ten times. The results of applying this technique are shown on the right in tabular format. Figure 13 shows the corresponding confusion matrix using the Confusion Matrix widget.



	no	yes	
no	108793	857	**109650**
yes	12221	1699	**13920**
	121014	**2556**	**123570**

Figure 13. Confusion Matrix for Logistic Regression on the Term Deposit Class

CHAPTER 5

EVALUATION

5.1 Introduction

Results obtained from running logistic regression, random forest and SVM using the flow created in Orange were processed and analyzed. Comparisons were made on the performances of the different algorithms on each of the test classes examined. The influencing attributes for each algorithm was determined for each test class.

5.2 Format of Output from Orange

Orange returns all experimental results of a test class per category. Take for example the term subscription test class. Since it has two categories which include ‘no’ and ‘yes’, Orange will return a table for each of those target classes. Figure 14 is an example of the result from the Test Learners widget on the term subscription attribute. Results for the ‘no’ target class are shown on top and for the ‘yes’ target class are shown on the bottom. Values for precision (Prec), recall, and F1 score (F1) vary based on the target class for which it is calculated. This is beneficial because each of the categories can be analyzed individually as is desired for this dataset.

Test Learners

Validation method

Method: Random sampling
Repetitions: 10
Proportion of training instances: 70%
Target class: no

Data

Examples: 41188
Attributes: 20 (age, job, marital, education, loan default, housing loan, personal loan, contact method, month, day_of_week, duration, campaign, pdays, previous, poutcome, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed)
Class: subscription

Results

	CA	Sens	Spec	AUC	IS	F1	Prec	Recall	Brier	MCC
SVM	0.8973	0.9866	0.1946	0.9339	0.0669	0.9446	0.9061	0.9866	0.1731	0.3167
Random Forest	0.8942	0.9922	0.1221	0.8512	0.0494	0.9433	0.8990	0.9922	0.1551	0.2538
Logistic regression	0.9100	0.9718	0.4232	0.9351	0.1288	0.9504	0.9299	0.9718	0.1261	0.4810

Test Learners

Validation method

Method: Random sampling
Repetitions: 10
Proportion of training instances: 70%
Target class: yes

Data

Examples: 41188
Attributes: 20 (age, job, marital, education, loan default, housing loan, personal loan, contact method, month, day_of_week, duration, campaign, pdays, previous, poutcome, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed)
Class: subscription

Results

	CA	Sens	Spec	AUC	IS	F1	Prec	Recall	Brier	MCC
SVM	0.8973	0.1946	0.9866	0.9339	0.0669	0.2993	0.6476	0.1946	0.1731	0.3167
Random Forest	0.8942	0.1221	0.9922	0.8512	0.0494	0.2062	0.6647	0.1221	0.1551	0.2538
Logistic regression	0.9100	0.4232	0.9718	0.9351	0.1288	0.5144	0.6558	0.4232	0.1261	0.4810

Figure 14. Format of Results from the Test Learners Widget Using Orange for Logistic Regression, Random Forest and SVM. The Top Shows the Results for the 'No' Category and the Bottom Shows Results for the 'Yes' Category.

5.3 Experimental Results and Analysis

The results were analyzed to determine predictability in each test class with SVM, random forest and logistic regression. The final data models for each of the algorithms in each class can be found in APPENDIX A: FINAL DATA MODELS. The most influencing attributes for each class was determined for all algorithms. The number of influencing attributes that belong to each category was determined based on Table 1.

5.3.1 Age Group

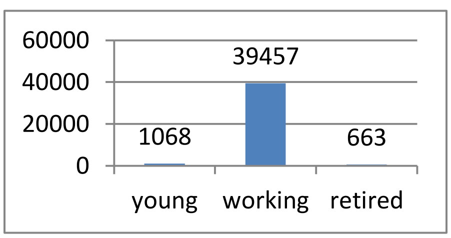
The F1 score, precision and recall values for the Age Group class are listed in Figure 15 along with a histogram of the categories in the class. Both precision and recall are very high for the ‘working’ category and in turn the F1 score is very high. Thus, this category has very strong predictability in the Age Group class with all three classification algorithms. Looking at the histogram, the Age Group class is dominated by this category and thus is expected to show high F1 scores.

The ‘young’ category has a much higher precision score than recall score and a very low F1 score for SVM and logistic regression. For these two algorithms the precision value is 0.41 which means that less items predicted to be in this category were actually a part of the category. Recall is also very low which means that neither of these two algorithms was successful in retrieving most of the values for this category. Random forest showed very poor precision, recall and therefore F1 score. Logistic regression has the highest F1 score for this category with a 0.11 value, but it is insufficient to allow good predictability of the category. This category is the second most represented in this class but is significantly less represented than the ‘working’ category.

The ‘retired’ category has the least number of instances but has higher evaluation scores than the ‘young’ category for SVM and logistic regression. The precision score means that greater than half of the instances classified as part of this category are actually part of this category for these two algorithms. The recall for these two algorithms is lower than precision so not many relevant items were selected. Random forest has zero precision, recall and F1 score in this category too. The overall F1 score for all algorithms is low, so this category has weak predictability. For this category SVM provided the highest predictability.

Age Group

	Young			Working			Retired		
	F1	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall
SVM	0.10	0.41	0.06	0.98	0.96	1.00	0.40	0.59	0.30
Random Forest	0.00	0.00	0.00	0.98	0.96	1.00	0.00	0.00	0.00
Logistic Regression	0.11	0.41	0.06	0.98	0.96	1.00	0.32	0.51	0.23



Top Contributors		
SVM	*Logistic Regression*	*Random Forest*
marital=unmarried	marital=unmarried	poutcome
N_euribor3m	job=unemployed	term deposit
N_nr.employed	month=nov	nr.employed
		cons.price.idx

Figure 15. Results for the Age Group Class. The Top Shows the F1, Precision and Recall Score and the Highest Value Is Highlighted in Each Column. Below Is a Histogram of the Categories and Also the Top Contributing Coefficients for All Algorithms Where Common Attributes Are Colored.

The top three coefficients contributing most to this class according to logistic regression include 'marital=single', 'job=unemployed' and 'month=nov' and for SVM it includes 'marital=single', 'N_euribor3m' and 'N_nr.employed' as is also shown in Figure 15. Marital status is common to both these algorithms. According to random forest the contributing attributes were poutcome, term deposit subscription, nr.employed and cons.price.idx. The nr.employed attribute was common to this and SVM. Overall this class is influenced by 1 attribute of the CD category, 2 of CF, 1 of PM, 1 of CM and 3 of SE category.

5.3.2 Employment Status

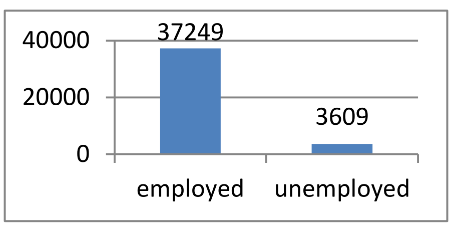
Figure 16 shows the F1, precision and recall values for Employment Status as well as the histogram for all the categories. The ‘employed’ category is the dominant category of this class, and both precision and recall are very high for this category. That means most of the relevant values were classified correctly and most of the values classified as this category were truly belonging to this category. The resulting F1 score was very high also, which means that this category has strong predictability using all three algorithms.

Even though the ‘unemployed’ category is not well represented in this class, it has approximately a 0.86 precision among the three algorithms. That means a very high number of values that were classified as belonging to this category were correctly classified. However, the recall value is low so all the values that should have been classified as positive were not classified correctly. Logistic regression showed the highest F1 score with a value of 0.34. Thus, the ‘unemployed’ category has weak predictability.

Figure 16 also shows the top three coefficients contributing most to this class. For logistic regression it includes 'age=retired', 'age=young' and 'month=dec' and for SVM it includes 'age=retired', 'age=working' and 'age=young'. Overall, the age attribute contributed significantly to this class based on both SVM and logistic regression. Based on random forest, the contributing attributes are poutcome, cons.price.idx, nr.employed and term deposit. This class is influenced by 1 attribute of the CD category, 1 of CF, 1 of PM, 1 of CM and 2 of SE category.

Employment Status

	Employed			Unemployed		
	F1	Prec	Recall	F1	Prec	Recall
SVM	0.96	0.92	1.00	0.27	0.87	0.16
Random Forest	0.96	0.92	1.00	0.08	0.86	0.04
Logistic Regression	0.96	0.93	1.00	0.34	0.85	0.22



Top Contributors		
SVM	***Logistic Regression***	***Random Forest***
age=retired	age=retired	poutcome
age=working	age=young	cons.price.idx
age=young	month=dec	nr.employed
		term deposit

Figure 16. Results for the Employment Status Class. The Top Shows the F1, Precision and Recall Score and the Highest Value Is Highlighted in Each Column. Below Is a Histogram of the Categories and Also the Top Contributing Coefficients for All Algorithms Where Common Attributes Are Colored.

5.3 3 Marital Status

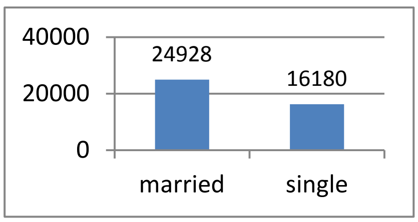
The marital status class was tested with two categories as shown in Figure 17 along with the corresponding histogram of all the categories. Both these categories are more evenly split as compared to other classes. The ‘married’ category ranges from a 0.61 to 0.64 precision score for the three algorithms, which means that a little more than half of the items classified as ‘married’ were actually married. The recall score is very high for random forest since 0.99 of instances that were truly belonging to the married category were classified correctly. The recall for SVM was similar to random forest. Logistic regression had the lowest recall score with a 0.9 value which is still very high. The resulting F1 score was highest for SVM and random forest with a 0.76 score.

The ‘single’ category had a 0.77 precision value for SVM, 0.87 for random forest and 0.59 for logistic regression. This means many selected items were relevant. The recall value was very low ranging from 0.03 to 0.22 with the highest value belonging to logistic regression, which means that many relevant items were not selected. The F1 score was also similarly low with a highest value of 0.32 which again was for logistic regression. The predictability for the ‘single’ category is very poor.

The top three coefficients contributing most to this class are also shown in Figure 17. They include 'age=young', 'age=retired' and 'education=university.degree' for logistic regression and 'age=young', 'month=mar' and 'age=retired' for SVM. Age is common to SVM and logistic regression. For the random forest, the contributing coefficients are poutcome, term deposit, campaign and pervious. This class is influenced by 1 attribute of the CD category, 2 of CF, 2 of PM and 2 of CM category.

Marital Status

	Married			Single		
	F1	Prec	Recall	F1	Prec	Recall
SVM	0.76	0.62	0.99	0.14	0.77	0.08
Random Forest	0.76	0.61	1.00	0.05	0.87	0.03
Logistic Regression	0.75	0.64	0.90	0.32	0.59	0.22



Top Contributors		
SVM	*Logistic Regression*	*Random Forest*
age=young	age=young	poutcome
month=mar	age=retired	term deposit
age=retired	education=university.d	campaign
		previous

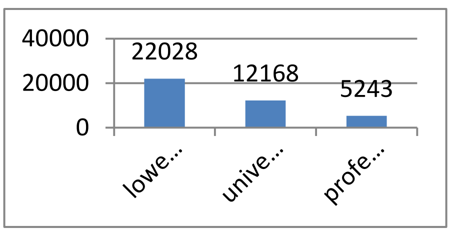
Figure 17. Results for the Marital Status Class. The Top Shows the F1, Precision and Recall Score and the Highest Value Is Highlighted in Each Column. Below Is a Histogram of the Categories and Also the Top Contributing Coefficients for All Algorithms Where Common Attributes Are Colored.

5.3.4 Education Level

The F1 score, precision and recall values for all categories of the Education Level class are shown in Figure 18 along with a histogram for the categories. This class has three categories which include ‘lower degree’, ‘university degree’ and ‘professional degree’. The ‘lower degree’ category is the most instances in this class. It has a precision value ranging from 0.56 to 0.62 with random forest being the lowest and logistic regression being the highest. That means a little over half of the selected instances were relevant. The recall value is lowest for logistic regression with a 0.87 value, followed by SVM with a 0.9 and then random forest with a 0.99 value. The F1 score was 0.72 for random forest and 0.73 for the other two algorithms. This category has the highest predictability as compared to the other two categories.

Education Level

	Lower Degree			University Degree			Professional Degree		
	F1	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall
SVM	0.73	0.61	0.90	0.36	0.49	0.28	0.00	0.00	0.00
Random Forest	0.72	0.56	0.99	0.04	0.46	0.02	0.00	0.00	0.00
Logistic Regression	0.73	0.62	0.87	0.42	0.49	0.36	0.00	0.33	0.00



Top Contributors		
SVM	*Logistic Regression*	*Random Forest*
month=aug	month=aug	poutcome
month=mar	N_euribor3m	campaign
month=may	age=young	term deposit
		previous

Figure 18. Results for the Education Level Class. The Top Shows the F1, Precision and Recall Score and the Highest Value Is Highlighted in Each Column. Below Is a Histogram of the Categories and Also the Top Contributing Coefficients for All Algorithms Where Common Attributes Are Colored.

The ‘university degree’ category had a lower overall predictability than ‘lower degree’. This category is the second most represented in this class. SVM and logistic regression had a precision value of 0.49 and random forest produced worse results with a 0.46 score. The recall value was very low for random forest with a 0.02 value but was better for SVM and logistic regression with a 0.28 and 0.36 value respectively. Thus, selected items were not very relevant and the relevant items were not selected well. The F1 score was very low for all algorithms. Thus this category was not predicted well.

The ‘professional degree’ category had the smallest representation of the three categories and performed the worst with a 0 F1 score for all three algorithms. The recall for all algorithms was also 0. The precision was 0.33 for logistic regression but 0 for the other two algorithms. This category had no relevant items selected by any algorithm and no selected items were relevant based on SVM and random forest.

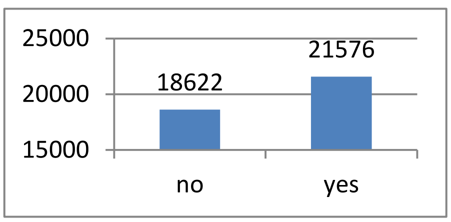
Figure 18 also shows the top three coefficients contributing most to this class. They include 'month=aug', 'N_euribor3m' and 'age=young' for logistic regression and 'month=aug', 'month=mar’ and 'month=may' for SVM. Month was common to SVM and logistic regression. According to random forest the contributing coefficients were poutcome, campaign, term deposit and pervious. This class is influenced by 1 attribute of the CD category, 1 of CF, 2 of PM, 2 of CM and 1 of SE category.

5.3.5 Housing Loan

Figure 19 shows the precision, recall and F1 scores for all categories of the Housing Loan class and also has a histogram of the categories. Both categories are well represented in this class, but the ‘no’ category had a slightly lower representation than the ‘yes’ category. The ‘yes’ category performed better than the ‘no’ category. The precision values were 0.54 for random forest and 0.57 for SVM and logistic regression. That means slightly more than half of the selected items were relevant. The recall value was similar for SVM and logistic regression with values of 0.75 and 0.77 respectively. Random forest showed the highest recall value for this category with a 0.93 value. The F1 score for 'yes' for SVM and logistic regression was 0.65 and for random forest was 0.69. This is because the recall value for random forest was significantly high. Based on the F1 score, this category did not show strong predictability.

Housing Loan

	No			Yes		
	F1	Prec	Recall	F1	Prec	Recall
SVM	0.41	0.54	0.34	0.65	0.57	0.75
Random Forest	0.16	0.53	0.09	0.69	0.54	0.93
Logistic Regression	0.40	0.55	0.32	0.65	0.57	0.77



Top Contributors		
SVM	*Logistic Regression*	*Random Forest*
N_cons.price.idx	personal loan=yes	poutcome
N_cons.conf.idx	month=aug	cons.price.idx
month=jun	N_cons.price.idx	term deposit
		nr.employed

Figure 19. Results for the Housing Loan Class. The Top Shows the F1, Precision and Recall Score and the Highest Value Is Highlighted in Each Column. Below Is a Histogram of the Categories and Also the Top Contributing Coefficients for All Algorithms Where Common Attributes Are Colored.

The 'no' category has approximately a 0.54 precision score for the three categories which means a little over half of the selected items were relevant. The recall value is extremely low for random forest and slightly higher for the other two algorithms, but it shows that many relevant items were not selected. It has a 0.16 F1 score for SVM, a 0.40 value for logistic regression and 0.41 for SVM. This category showed weak predictability.

As shown in Figure 19 top three contributing information to this class include 'personal loan=yes', 'month=aug' and 'N_cons.price.idx' for logistic regression and 'N_cons.price.idx', 'N_cons.conf.idx' and 'month=jun' for SVM. For random forest the coefficients were poutcome, cons.price.idx, term deposit and nr.employed. Month was common to SVM and logistic regression. Cons.price.idx was common to all algorithms. This class is influenced by 2 attributes of the CF category, 1 of PM, 1 of CM and 3 of SE.

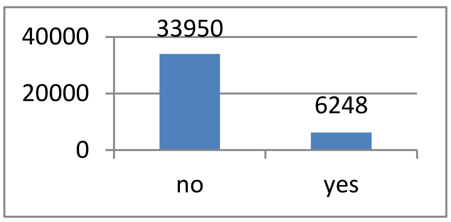
5.3.6 Personal Loan

The F1 score, precision and recall values for the Personal Loan class are shown in Figure 20 in addition to a histogram of the categories. There is a significant imbalance in the representation of the two categories in this class. The ‘no’ category performed very well in terms of precision and recall. For all algorithms the precision was 0.84, which means that more 84% of items classified as ‘no’ were truly belonging to this category. The recall for all algorithms was 1 which is the ideal. That means 100% of all values that were truly belonging to this category were classified. Since the overall F1 score was 0.92 for all algorithms in this category, the ‘no’ category shows strong predictability.

In contrast to this category, the ‘yes’ category has a 0 value for both precision and recall, and in turn F1 score, for all algorithms. That means no selected items were relevant and no relevant items were selected by any of the algorithms. Thus, the overall F1 score for algorithms is 0 so this category is poorly predicted by any of the algorithms.

Personal Loan

	No			Yes		
	F1	Prec	Recall	F1	Prec	Recall
SVM	0.92	0.84	1.00	0.00	0.00	0.00
Random Forest	0.92	0.84	1.00	0.00	0.00	0.00
Logistic Regression	0.92	0.84	1.00	0.00	0.00	0.00



Top Contributors		
SVM	*Logistic Regression*	*Random Forest*
pdays=21	housing loan=no	poutcome
pdays=999	month=oct	age
N_nr.employed	pdays=3	previous
		pdays

Figure 20. Results for the Personal Loan Class. The Top Shows the F1, Precision and Recall Score and the Highest Value Is Highlighted in Each Column. Below Is a Histogram of the Categories and Also the Top Contributing Coefficients for All Algorithms Where Common Attributes Are Colored.

The top three coefficients contributing to this class are also shown in Figure 20. For logistic regression the top three are 'housing loan=no', 'month=oct' and 'pdays=3' and for SVM are 'pdays=21', 'pdays=999' and 'N_nr.employed'. For random forest the top contributors were poutcome, age, previous and pdays. Pdays was a common attribute for all the algorithms. This class is influenced by 1 attribute of the CD category, 1 of CF, 1 of PM, 3 of CM and 1 of SE category.

5.3.7 Term Deposit

Figure 21 shows the precision, recall and F1 scores for the Term Deposit categories and also has a corresponding histogram of the categories. The ‘no’ category has a very high representation in the class. The precision values are very high for this category with a range of 0.90 to 0.93. That means most of the selected items were relevant. The recall value is even greater with a 0.97 value for logistic regression and a 0.99 value for SVM and random forest. That means nearly all relevant items were selected. As a result, the resulting F1 score for all values is very high, and the highest is from logistic regression. This category showed strong predictability by all three algorithms.

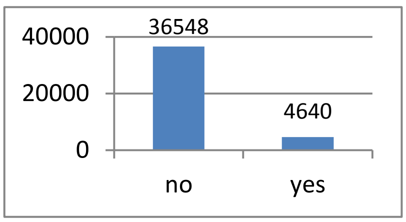
The ‘yes’ category performed worse than the ‘no’ category. The precision for SVM, random forest and logistic regression were 0.65, 0.66 and 0.66 respectively. That means most of the selected items were not relevant. The recall value was lowest for random forest and highest for logistic regression, but in all of the algorithms many of the relevant results were not selected. The F1 score was highest for logistic regression with a 0.51 score. This category did not show strong predictability.

As shown in Figure 21 the top three coefficients contributing to this class include 'month=mar', 'N_duration' and 'N_emp.var.rate' for logistic regression. For SVM it is 'N_duration', 'pdays=999' and 'poutcome=success' and for random forest it is previous, euribor3m and emp.var.rate. Duration was found by SVM and logistic regression.

Emp.var.rate was common to logistic regression and random forest. This class is influenced by 2 attributes of the PM category, 3 of CM and 2 of SE category.

Term Deposit

	No			Yes		
	F1	Prec	Recall	F1	Prec	Recall
SVM	0.94	0.91	0.99	0.30	0.65	0.19
Random Forest	0.94	0.90	0.99	0.21	0.66	0.12
Logistic Regression	0.95	0.93	0.97	0.51	0.66	0.42



Top Contributors		
SVM	***Logistic Regression***	***Random Forest***
N_duration	month=mar	previous
pdays=999	N_duration	euribor3m
poutcome=success	N_emp.var.rate	emp.var.rate

Figure 21. Results for the Term Deposit Class. The Top Shows the F1, Precision and Recall Score and the Highest Value Is Highlighted in Each Column. Below Is a Histogram of the Categories and Also the Top Contributing Coefficients for All Algorithms Where Common Attributes Are Colored.

CHAPTER 6

CONCLUSION AND FUTURE WORK

Data mining techniques were applied to bank telemarketing campaign data from a Portuguese bank. SVM, random forest and logistic regression were applied on seven attributes to determine predictability of each class as well as determine most contributing attributes. In each class the category with the smallest F1 score was not high enough to show strong predictability. Table 13 summarizes the results of the smallest category in descending order of the highest F1 score along with the corresponding precision and recall value and also the algorithm. The logistic regression model had the best F1 score for the majority of the classes tested.

Table 13. Best F1 Score for Each Attribute on the Smallest Category Followed by the Precision, Recall and Algorithm Used. Attributes Are Sorted by the F1 Score.

Performance on Smallest Category				
	F1	**Precision**	**Recall**	**Algorithm**
Term Deposit	0.51	0.66	0.42	Logistic Regression
Housing Loan	0.41	0.54	0.34	SVM
Age Group	0.4	0.59	0.3	SVM
Employment Status	0.34	0.85	0.22	Logistic Regression
Marital Status	0.32	0.59	0.22	Logistic Regression
Education Level	0	0.33	0	Logistic Regression
Personal Loan	0	0	0	All

Several attributes that contribute most were determined for each class. Figure 22 shows the number of attributes that contribute to each of the attribute categories. These attribute categories are based on Table 1 in Section 2.2 Data above. Age is least predicted by attributes of the bank-customer relation category. All categories influence employment status equally. Marital status is not influenced by social-economic situation. The bank-customer relation most influences the education level attribute of the client. Whether or not the client has a personal loan was more influenced by the bank-customer relation but a housing loan was more influenced by the social-economic situation. The client's decision to subscribe to a term deposit was most influenced by the bank-customer relation.

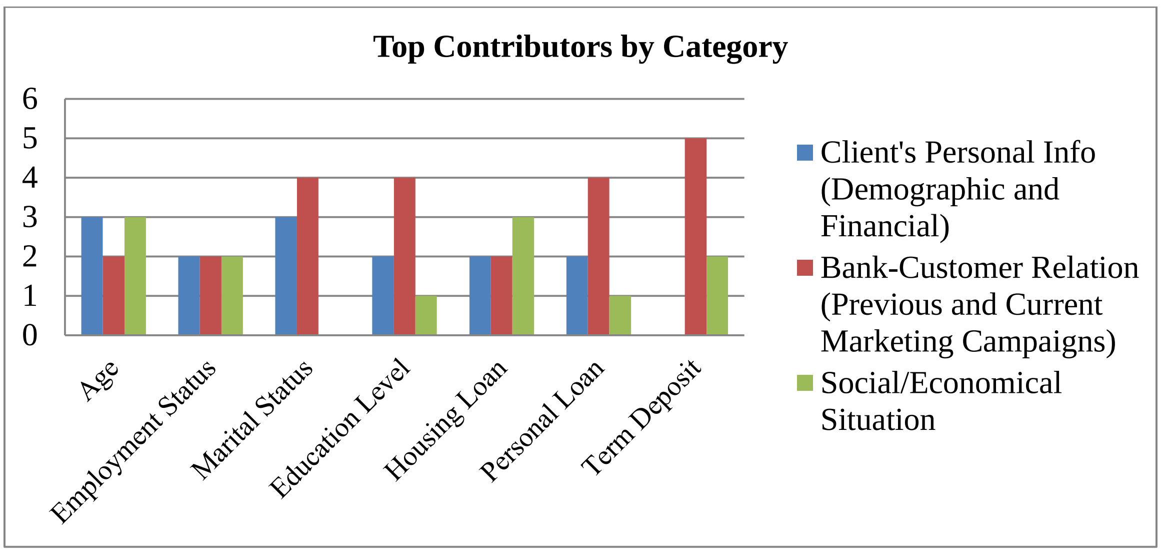


Figure 22. Number of Top Contributors Associated with Each Category for Each of the Attributes Tested.

Further research can be performed by combining this dataset with another dataset such as was done in Moro, Cortez and Rita [10] to find the result of predicting the same classes with the same algorithms and metrics but with more attributes. It would be interesting to see if similar conclusions are made when a larger dataset is used. In addition, adding more information may also allow more categories to be created for grouping the top contributors. Predictability of other attributes in the dataset or usage of different metrics and algorithms can also be analyzed.

REFERENCES

[1] [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014.

[2] "Logistic regression," 14 August 2015. [Online]. Available: https://www.medcalc.org/manual/logistic_regression.php.

[3] Machine Learning Group at National Taiwan University, "LIBLINEAR -- A Library for Large Linear Classification," [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/liblinear/.

[4] P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Pearson Education, Inc., 2006.

[5] C.-C. Chang and C.-J. Lin, "LIBSVM -- A Library for Support Vector Machine," [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[6] Demšar, J., Curk, T., & Erjavec, A. Orange: Data Mining Toolbox in Python; Journal of Machine Learning Research 14(Aug):2349−2353, 2013.

[7] Valverde-Albacete FJ, Peláez-Moreno C (2014) 100% Classification Accuracy Considered Harmful: The Normalized Information Transfer Factor Explains the Accuracy Paradox. PLoS ONE 9(1): e84217. doi:10.1371/journal.pone.0084217.

[8] Alan, "WHY ACCURACY ALONE IS A BAD MEASURE FOR CLASSIFICATION TASKS, AND WHAT WE CAN DO ABOUT IT," 25 March 2013. [Online]. Available: http://blog.tryolabs.com/2013/03/25/why-accuracy-alone-bad-measure-classification-tasks-and-what-we-can-do-about-it/.

[9] N. V. Chawla, "C4.5 and Imbalanced Data sets: Investigating the effect of sampling method, probabilistic estimate, and decision tree structure," Washington DC, 2003.

[10] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014.

[11] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS. [bank.zip].

[12] C. Vajiramedhin and A. Suebsing, "Feature Selection with Data Balancing for Prediction of Bank Telemarketing," *Applied Mathematical Sciences*, vol. 8, pp. 5667-5672, 2014.

[13] H. A. Elsalamony, "Bank Direct Marketing Analysis of Data Mining," *International Journal of Computer Applications*, vol. 85, no. 7, pp. 12-22, January 2014.

[14] "Portugal Retirement Age - Men," [Online]. Available: http://www.tradingeconomics.com/portugal/retirement-age-men.

APPENDIX A

FINAL DATA MODELS

Age Group

Logistic Regression

Feature:Weight
job=unemployed : 1.45957231522
marital=single : 2.30414056778
education=professional.course : -0.285010635853
education=university.degree : -1.18860256672
loan default=yes : -0.00134700455237
housing loan=no : -0.0872378349304
personal loan=yes : -0.0827678367496
contact method=telephone : -0.094453625381
month=jan : 0.0
month=feb : 0.0
month=mar : -0.326083928347
month=apr : 0.0301006790251
month=jun : 0.154620602727
month=jul : 1.04956662655
month=aug : -0.466708123684
month=sep : 0.268356114626
month=oct : -0.19288277626
month=nov : -1.23174083233
month=dec : -0.111696444452
day_of_week=mon : -0.213365629315
day_of_week=tue : -0.264860987663
day_of_week=wed : -0.0110015012324
day_of_week=fri : -0.171186670661
N_duration : -0.00812559667975
N_campaign : -0.0518000535667
pdays=0 : -0.121027685702
pdays=1 : 0.154464736581
pdays=2 : -0.244776874781
pdays=3 : 0.141835004091
pdays=4 : 0.130457475781
pdays=5 : 0.00827248953283
pdays=6 : -0.391178101301
pdays=7 : 0.0647301077843
pdays=8 : -0.0105945505202
pdays=9 : 0.116318546236
pdays=10 : 0.0365114696324
pdays=11 : -0.111566588283
pdays=12 : -0.00662565184757

pdays=13 : 0.169705182314
pdays=14 : 0.101252101362
pdays=15 : 0.0214226935059
pdays=16 : 0.0340120531619
pdays=17 : 0.106162428856
pdays=18 : -0.0517643578351
pdays=19 : -0.0337135381997
pdays=20 : -0.0465160124004
pdays=21 : -0.00565495342016
pdays=22 : -0.0278128180653
pdays=25 : -0.00661940313876
pdays=26 : 0.131618082523
pdays=27 : -0.00124768679962
N_previous : 0.118054918945
poutcome=failure : -0.269354224205
poutcome=success : -0.184637576342
N_emp.var.rate : -0.663087069988
N_cons.price.idx : 0.363447070122
N_cons.conf.idx : -0.0111547978595
N_euribor3m : -0.23866482079
N_nr.employed : 0.00335311936215
subscription=yes : 0.212411105633

SVM

Attribute:Weight
Orange.feature.Continuous 'education=lower.degree':1.67818554559
Orange.feature.Continuous 'education=professional.course':0.515366927017
Orange.feature.Continuous 'pdays=9':1.07560468048
Orange.feature.Continuous 'education=university.degree':2.10931863542
Orange.feature.Continuous 'pdays=10':0.587695267359
Orange.feature.Continuous 'loan default=yes':0.0
Orange.feature.Continuous 'pdays=11':1.01426875106
Orange.feature.Continuous 'pdays=8':0.162795890381
Orange.feature.Continuous 'housing loan=yes':0.135203869042
Orange.feature.Continuous 'pdays=12':1.2122757315
Orange.feature.Continuous 'personal loan=yes':0.257858382651
Orange.feature.Continuous 'pdays=13':0.537710047739
Orange.feature.Continuous 'contact method=telephone':1.07808205157
Orange.feature.Continuous 'pdays=14':0.924287868867
Orange.feature.Continuous 'month=jan':0.0

Orange.feature.Continuous 'pdays=15':0.140043900484
Orange.feature.Continuous 'month=feb':0.0
Orange.feature.Continuous 'pdays=16':0.859181179613
Orange.feature.Continuous 'month=mar':9.73326679584
Orange.feature.Continuous 'pdays=17':0.679470116
Orange.feature.Continuous 'month=apr':6.56910281852
Orange.feature.Continuous 'pdays=18':0.495022096794
Orange.feature.Continuous 'month=may':5.51441754234
Orange.feature.Continuous 'pdays=19':0.189566001296
Orange.feature.Continuous 'month=jun':2.40620030247
Orange.feature.Continuous 'pdays=20':0.211958006024
Orange.feature.Continuous 'month=jul':3.53922578404
Orange.feature.Continuous 'pdays=21':1.0
Orange.feature.Continuous 'month=aug':1.76865155242
Orange.feature.Continuous 'pdays=22':0.0
Orange.feature.Continuous 'month=sep':1.58276895868
Orange.feature.Continuous 'pdays=25':0.0
Orange.feature.Continuous 'month=oct':0.825633237093
Orange.feature.Continuous 'pdays=26':1.41421356237
Orange.feature.Continuous 'month=nov':1.97631115335
Orange.feature.Continuous 'pdays=27':0.0
Orange.feature.Continuous 'month=dec':0.717844341403
Orange.feature.Continuous 'pdays=999':0.84638328352
Orange.feature.Continuous 'day_of_week=mon':0.398582308264
Orange.feature.Continuous 'N_previous':1.17316160857
Orange.feature.Continuous 'day_of_week=tue':0.333975883091
Orange.feature.Continuous 'poutcome=failure':0.500457663111
Orange.feature.Continuous 'day_of_week=wed':0.698029624914
Orange.feature.Continuous 'poutcome=nonexistent':1.28655631321
Orange.feature.Continuous 'day_of_week=thu':0.219463354262
Orange.feature.Continuous 'poutcome=success':0.897823177377
Orange.feature.Continuous 'day_of_week=fri':0.328555163788
Orange.feature.Continuous 'N_emp.var.rate':22.5817937288
Orange.feature.Continuous 'N_duration':1.53203544679
Orange.feature.Continuous 'N_cons.price.idx':3.42220659734
Orange.feature.Continuous 'N_campaign':1.21898819761
Orange.feature.Continuous 'N_cons.conf.idx':2.9159503095
Orange.feature.Continuous 'pdays=0':0.590545920175
Orange.feature.Continuous 'N_euribor3m':27.6033182984
Orange.feature.Continuous 'pdays=1':1.15368185276
Orange.feature.Continuous 'N_nr.employed':22.9837680677
Orange.feature.Continuous 'pdays=2':0.447384178132
Orange.feature.Continuous 'subscription=yes':1.07349642029
Orange.feature.Continuous 'pdays=3':1.36897613178

Orange.feature.Continuous 'pdays=4':0.289809862675
Orange.feature.Continuous 'pdays=5':0.499696941815
Orange.feature.Continuous 'pdays=6':1.07166156651
Orange.feature.Continuous 'marital=single':71.8936097564
Orange.feature.Continuous 'pdays=7':0.296596114053
Orange.feature.Continuous 'job=unemployed':9.3831647987

Sorted weights:
['0.0000', '0.0000', '0.0000', '0.0000', '0.0000', '0.0000', '0.1352', '0.1400', '0.1628', '0.1896', '0.2120', '0.2195', '0.2579', '0.2898', '0.2966', '0.3286', '0.3340', '0.3986', '0.4474', '0.4950', '0.4997', '0.5005', '0.5154', '0.5377', '0.5877', '0.5905', '0.6795', '0.6980', '0.7178', '0.8256', '0.8464', '0.8592', '0.8978', '0.9243', '1.0000', '1.0143', '1.0717', '1.0735', '1.0756', '1.0781', '1.1537', '1.1732', '1.2123', '1.2190', '1.2866', '1.3690', '1.4142', '1.5320', '1.5828', '1.6782', '1.7687', '1.9763', '2.1093', '2.4062', '2.9160', '22.5818', '22.9838', '27.6033', '3.4222', '3.5392', '5.5144', '6.5691', '71.8936', '9.3832', '9.7333']

Random Forest

Tree size: 22 nodes, 12 leaves

<u>Employment Status</u>

Logistic Regression

Feature:Weight
age=young : -1.66950881481
age=retired : -3.36810064316
marital=single : -0.20413172245
education=professional.course : 0.294509083033
education=university.degree : 0.683849155903
loan default=yes : -0.0840618312359
housing loan=no : 0.0094482852146
personal loan=yes : 0.00998468697071

contact method=telephone : -0.243371859193
month=jan : 0.0
month=feb : 0.0
month=mar : -0.807804524899
month=apr : -0.449353337288
month=jun : -0.32230681181
month=jul : -0.550947248936
month=aug : -0.91481757164
month=sep : -0.58251708746
month=oct : -0.4940508008
month=nov : -0.476988166571
month=dec : -0.998300969601
day_of_week=mon : 0.0266277100891
day_of_week=tue : -0.0345024056733
day_of_week=wed : -0.0091074667871
day_of_week=fri : -0.0581567659974
N_duration : 0.005665384233
N_campaign : -0.0318745523691
pdays=0 : -0.140335604548
pdays=1 : 0.00183187855873
pdays=2 : -0.0403323173523
pdays=3 : -0.195524781942
pdays=4 : 0.225522115827
pdays=5 : 0.0677793398499
pdays=6 : -0.125854447484
pdays=7 : -0.0406650900841
pdays=8 : 0.101693540812
pdays=9 : -0.0984656736255
pdays=10 : -0.424441665411
pdays=11 : 0.0890951156616
pdays=12 : 0.278004109859
pdays=13 : 0.00205332436599
pdays=14 : -0.237113565207
pdays=15 : -0.180631577969
pdays=16 : 0.132690399885
pdays=17 : -0.0795172601938
pdays=18 : 0.0105093717575
pdays=19 : -0.0503020957112
pdays=20 : -0.0821955427527
pdays=21 : 0.00031484363717
pdays=22 : 0.0614402927458
pdays=25 : 0.0212668962777
pdays=26 : 0.052602943033
pdays=27 : 0.0133326109499

N_previous : -0.0305148568004
poutcome=failure : 0.240627884865
poutcome=success : 0.13153706491
N_emp.var.rate : 0.658731341362
N_cons.price.idx : -0.33080843091
N_cons.conf.idx : -0.0602298155427
N_euribor3m : -0.44801646471
N_nr.employed : 0.00423885695636
subscription=yes : -0.156523063779

SVM

Attribute:Weight
Orange.feature.Continuous 'day_of_week=wed':0.0142289875657
Orange.feature.Continuous 'loan default=yes':0.129425004125
Orange.feature.Continuous 'month=sep':0.411038121209
Orange.feature.Continuous 'pdays=13':0.0509810000658
Orange.feature.Continuous 'pdays=7':0.131005974486
Orange.feature.Continuous 'pdays=18':0.0819100141525
Orange.feature.Continuous 'pdays=999':1.14447129215
Orange.feature.Continuous 'pdays=25':0.0781619995832
Orange.feature.Continuous 'pdays=0':0.0179249946959
Orange.feature.Continuous 'month=aug':1.03784493706
Orange.feature.Continuous 'N_duration':0.113562063111
Orange.feature.Continuous 'pdays=8':0.0518469922245
Orange.feature.Continuous 'pdays=20':1.0
Orange.feature.Continuous 'month=oct':0.485666955821
Orange.feature.Continuous 'pdays=14':0.16323004663
Orange.feature.Continuous 'pdays=12':0.110114013776
Orange.feature.Continuous 'pdays=11':0.0838389918208
Orange.feature.Continuous 'pdays=21':0.160794973373
Orange.feature.Continuous 'pdays=3':0.224020929541
Orange.feature.Continuous 'N_campaign':0.254992951998
Orange.feature.Continuous 'N_previous':0.392733343231
Orange.feature.Continuous 'month=dec':0.228212084156
Orange.feature.Continuous 'pdays=5':0.0384339913726
Orange.feature.Continuous 'day_of_week=tue':0.188343713991
Orange.feature.Continuous 'education=professional.course':0.573713794816
Orange.feature.Continuous 'pdays=9':0.0216919686645
Orange.feature.Continuous 'education=lower.degree':1.36787975027
Orange.feature.Continuous 'day_of_week=thu':0.0688238157891

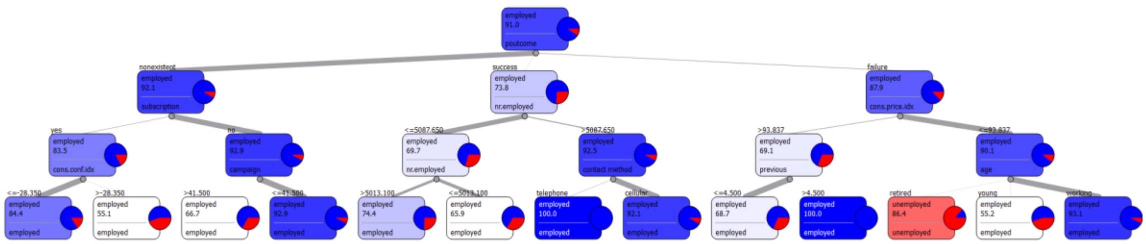
Orange.feature.Continuous 'marital=single':0.586129533185
Orange.feature.Continuous 'N_emp.var.rate':1.82609965207
Orange.feature.Continuous 'month=jul':0.789525063243
Orange.feature.Continuous 'month=nov':0.308148936834
Orange.feature.Continuous 'month=feb':0.0
Orange.feature.Continuous 'age=working':30.4472893867
Orange.feature.Continuous 'age=retired':47.7282970436
Orange.feature.Continuous 'pdays=2':0.025931943208
Orange.feature.Continuous 'day_of_week=fri':0.0178451170214
Orange.feature.Continuous 'day_of_week=mon':0.151587302797
Orange.feature.Continuous 'age=young':17.2810009569
Orange.feature.Continuous 'poutcome=nonexistent':1.25559631176
Orange.feature.Continuous 'education=university.degree':0.794159255456
Orange.feature.Continuous 'pdays=17':0.0403459668159
Orange.feature.Continuous 'pdays=27':0.0
Orange.feature.Continuous 'pdays=16':0.00138497725129
Orange.feature.Continuous 'housing loan=yes':0.00706990691833
Orange.feature.Continuous 'N_cons.conf.idx':1.25823134023
Orange.feature.Continuous 'pdays=22':0.125056996942
Orange.feature.Continuous 'personal loan=yes':0.220644143468
Orange.feature.Continuous 'pdays=4':0.02182803303
Orange.feature.Continuous 'contact method=telephone':0.181623381097
Orange.feature.Continuous 'pdays=1':0.074930020608
Orange.feature.Continuous 'month=jan':0.0
Orange.feature.Continuous 'N_nr.employed':2.32275310564
Orange.feature.Continuous 'N_euribor3m':2.46138773398
Orange.feature.Continuous 'pdays=10':0.101592999417
Orange.feature.Continuous 'pdays=26':1.0
Orange.feature.Continuous 'poutcome=success':0.99579797202
Orange.feature.Continuous 'month=mar':0.388117162976
Orange.feature.Continuous 'poutcome=failure':0.259805039736
Orange.feature.Continuous 'N_cons.price.idx':0.173756825154
Orange.feature.Continuous 'month=apr':0.175411989272
Orange.feature.Continuous 'subscription=yes':1.14850795205
Orange.feature.Continuous 'month=may':1.5805679874
Orange.feature.Continuous 'pdays=6':0.386879953847
Orange.feature.Continuous 'pdays=15':0.0712440004572
Orange.feature.Continuous 'month=jun':0.664340436691
Orange.feature.Continuous 'pdays=19':0.0468690171838

Sorted weights:
['0.0000', '0.0000', '0.0000', '0.0014', '0.0071', '0.0142', '0.0178', '0.0179', '0.0217', '0.0218', '0.0259', '0.0384', '0.0403', '0.0469', '0.0510', '0.0518', '0.0688', '0.0712', '0.0749', '0.0782', '0.0819', '0.0838', '0.1016', '0.1101', '0.1136', '0.1251', '0.1294', '0.1310', '0.1516',

'0.1608', '0.1632', '0.1738', '0.1754', '0.1816', '0.1883', '0.2206', '0.2240', '0.2282', '0.2550', '0.2598', '0.3081', '0.3869', '0.3881', '0.3927', '0.4110', '0.4857', '0.5737', '0.5861', '0.6643', '0.7895', '0.7942', '0.9958', '1.0000', '1.0000', '1.0378', '1.1445', '1.1485', '1.2556', '1.2582', '1.3679', '1.5806', '1.8261', '17.2810', '2.3228', '2.4614', '30.4473', '47.7283']

Random Forest

Tree size: 23 nodes, 13 leaves



Marital Status

Logistic Regression

Feature:Weight
age=young : -2.19580006599
age=retired : 0.672370791435
job=unemployed : -0.203482106328
education=professional.course : -0.258334815502
education=university.degree : -0.588413119316
loan default=yes : 0.011705798097
housing loan=no : 0.0248215049505
personal loan=yes : 0.00767252407968
contact method=telephone : 0.00543400738388
month=jan : 0.0
month=feb : 0.0
month=mar : -0.29718503356
month=apr : 0.162410825491
month=jun : 0.0159089621156
month=jul : -0.372277587652
month=aug : 0.0521575920284
month=sep : 0.230291858315
month=oct : 0.0634162649512

month=nov : -0.0276760216802
month=dec : 0.152484804392
day_of_week=mon : 0.0385625436902
day_of_week=tue : -0.00661889323965
day_of_week=wed : -0.0239796359092
day_of_week=fri : 0.0151863293722
N_duration : 0.0104096783325
N_campaign : -0.00576466787606
pdays=0 : -0.0219844598323
pdays=1 : -0.00820596143603
pdays=2 : 0.0502774938941
pdays=3 : 0.0138692446053
pdays=4 : -0.00479408912361
pdays=5 : 0.0280380621552
pdays=6 : 0.0334340557456
pdays=7 : -0.00527216261253
pdays=8 : 0.0374559834599
pdays=9 : -0.0197046101093
pdays=10 : 0.0302824433893
pdays=11 : 0.00240125181153
pdays=12 : -0.0315947160125
pdays=13 : 0.000820586283226
pdays=14 : -0.0290412548929
pdays=15 : -0.0194264799356
pdays=16 : 0.0177336428314
pdays=17 : -0.0295058116317
pdays=18 : 0.0147935068235
pdays=19 : -0.0185370370746
pdays=20 : -0.00586070865393
pdays=21 : 0.0110872676596
pdays=22 : -0.0189462844282
pdays=25 : 0.00616032257676
pdays=26 : 0.00949117448181
pdays=27 : 0.00622528837994
N_previous : 0.00809162948281
poutcome=failure : -0.0127778984606
poutcome=success : 0.0377215892076
N_emp.var.rate : 0.00831096339971
N_cons.price.idx : -0.0663036853075
N_cons.conf.idx : 0.0636059269309
N_euribor3m : 0.148737579584
N_nr.employed : 0.000841796456371
subscription=yes : -0.061363954097

SVM

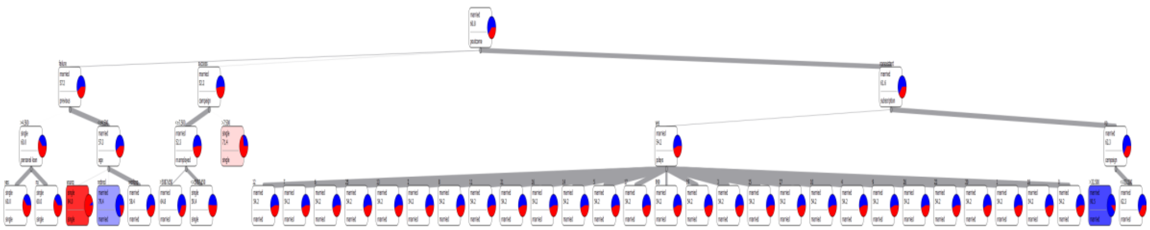
Attribute:Weight
Orange.feature.Continuous 'N_campaign':0.0679706157297
Orange.feature.Continuous 'pdays=21':1.72166001797
Orange.feature.Continuous 'pdays=22':3.0
Orange.feature.Continuous 'pdays=3':0.896272936836
Orange.feature.Continuous 'pdays=25':1.0
Orange.feature.Continuous 'pdays=26':1.0
Orange.feature.Continuous 'pdays=11':0.0
Orange.feature.Continuous 'day_of_week=tue':0.657229982782
Orange.feature.Continuous 'pdays=27':0.994740009308
Orange.feature.Continuous 'day_of_week=wed':0.717465911061
Orange.feature.Continuous 'month=mar':51.0
Orange.feature.Continuous 'pdays=999':0.145075853914
Orange.feature.Continuous 'N_duration':0.143403529015
Orange.feature.Continuous 'month=jan':0.0
Orange.feature.Continuous 'month=oct':5.8796689678
Orange.feature.Continuous 'pdays=0':3.0
Orange.feature.Continuous 'loan default=yes':0.223743993789
Orange.feature.Continuous 'poutcome=failure':0.53270702064
Orange.feature.Continuous 'pdays=9':0.428530953825
Orange.feature.Continuous 'age=working':25.136662107
Orange.feature.Continuous 'month=jul':4.81441737898
Orange.feature.Continuous 'month=apr':5.60785798542
Orange.feature.Continuous 'poutcome=success':0.404101153836
Orange.feature.Continuous 'education=university.degree':0.351655198261
Orange.feature.Continuous 'day_of_week=fri':0.527296838351
Orange.feature.Continuous 'month=may':5.67373125907
Orange.feature.Continuous 'N_cons.price.idx':0.804529311737
Orange.feature.Continuous 'N_cons.conf.idx':3.61929902568
Orange.feature.Continuous 'day_of_week=mon':0.131903866306
Orange.feature.Continuous 'pdays=15':1.0
Orange.feature.Continuous 'N_euribor3m':2.08148043848
Orange.feature.Continuous 'day_of_week=thu':0.335155030247
Orange.feature.Continuous 'pdays=12':0.137255996466
Orange.feature.Continuous 'pdays=18':1.0
Orange.feature.Continuous 'pdays=16':1.40025499463
Orange.feature.Continuous 'subscription=yes':0.088642292656
Orange.feature.Continuous 'N_emp.var.rate':0.848817749882
Orange.feature.Continuous 'pdays=5':1.41954600811
Orange.feature.Continuous 'month=jun':4.59706448112

Orange.feature.Continuous 'poutcome=nonexistent':0.128607880324
Orange.feature.Continuous 'month=aug':5.32112574484
Orange.feature.Continuous 'N_nr.employed':3.93172340483
Orange.feature.Continuous 'pdays=19':3.0
Orange.feature.Continuous 'job=unemployed':0.951067786198
Orange.feature.Continuous 'housing loan=yes':0.166918013245
Orange.feature.Continuous 'month=nov':6.03102422226
Orange.feature.Continuous 'pdays=2':1.64402198792
Orange.feature.Continuous 'pdays=8':2.0
Orange.feature.Continuous 'month=dec':6.37982200831
Orange.feature.Continuous 'pdays=7':0.875930964947
Orange.feature.Continuous 'pdays=13':0.967958025634
Orange.feature.Continuous 'pdays=17':4.0
Orange.feature.Continuous 'education=professional.course':0.92840191815
Orange.feature.Continuous 'education=lower.degree':0.576748733409
Orange.feature.Continuous 'month=feb':0.0
Orange.feature.Continuous 'pdays=20':1.0
Orange.feature.Continuous 'personal loan=yes':0.0131678320467
Orange.feature.Continuous 'pdays=10':0.877574980259
Orange.feature.Continuous 'pdays=4':0.743576928973
Orange.feature.Continuous 'age=retired':28.249096049
Orange.feature.Continuous 'pdays=6':1.03819097579
Orange.feature.Continuous 'pdays=14':2.0
Orange.feature.Continuous 'age=young':53.3857601695
Orange.feature.Continuous 'N_previous':0.00531913203837
Orange.feature.Continuous 'contact method=telephone':0.698687966447
Orange.feature.Continuous 'pdays=1':0.13113296032
Orange.feature.Continuous 'month=sep':6.69528593868

Sorted weights:
['0.0000', '0.0000', '0.0000', '0.0053', '0.0132', '0.0680', '0.0886', '0.1286', '0.1311', '0.1319', '0.1373', '0.1434', '0.1451', '0.1669', '0.2237', '0.3352', '0.3517', '0.4041', '0.4285', '0.5273', '0.5327', '0.5767', '0.6572', '0.6987', '0.7175', '0.7436', '0.8045', '0.8488', '0.8759', '0.8776', '0.8963', '0.9284', '0.9511', '0.9680', '0.9947', '1.0000', '1.0000', '1.0000', '1.0000', '1.0000', '1.0382', '1.4003', '1.4195', '1.6440', '1.7217', '2.0000', '2.0000', '2.0815', '25.1367', '28.2491', '3.0000', '3.0000', '3.0000', '3.6193', '3.9317', '4.0000', '4.5971', '4.8144', '5.3211', '5.6079', '5.6737', '5.8797', '51.0000', '53.3858', '6.0310', '6.3798', '6.6953']

Random Forest

Tree size: 46 nodes, 37 leaves



Education Level

Logistic Regression

Feature:Weight
age=young : 1.0207927227
age=retired : 0.982188940048
job=unemployed : 0.573116183281
marital=single : -0.494462132454
loan default=yes : -0.0284495167434
housing loan=no : 0.0344993770123
personal loan=yes : -0.0674560815096
contact method=telephone : 0.0393917262554
month=jan : 0.0
month=feb : 0.0
month=mar : -0.782980680466
month=apr : -0.491301506758
month=jun : -0.894448041916
month=jul : -0.334548324347
month=aug : -1.27970588207
month=sep : -0.496689736843
month=oct : -0.301488161087
month=nov : -0.595813333988
month=dec : -0.579795837402
day_of_week=mon : 0.0584962852299
day_of_week=tue : 0.101925000548
day_of_week=wed : 0.111271306872
day_of_week=fri : 0.0756566375494
N_duration : 0.045036431402
N_campaign : -0.0299973133951
pdays=0 : -0.0960081741214
pdays=1 : -0.00677893357351
pdays=2 : -0.27956956625
pdays=3 : -0.097979195416

pdays=4 : 0.24357996881
pdays=5 : -0.0828369110823
pdays=6 : 0.0524091720581
pdays=7 : -0.130580276251
pdays=8 : -0.00728510320187
pdays=9 : 0.0418233759701
pdays=10 : 0.249310150743
pdays=11 : 0.0183628480881
pdays=12 : 0.110766075552
pdays=13 : -0.0262498930097
pdays=14 : 0.132504358888
pdays=15 : -0.0126699423417
pdays=16 : 0.0738767832518
pdays=17 : 0.0786657556891
pdays=18 : -0.0150046991184
pdays=19 : -0.00790530722588
pdays=20 : 0.0234462842345
pdays=21 : -0.0123122986406
pdays=22 : 0.0128676760942
pdays=25 : 0.0263001490384
pdays=26 : -0.0428065024316
pdays=27 : -0.0272589940578
N_previous : 0.0302045289427
poutcome=failure : 0.0455319695175
poutcome=success : -0.110787294805
N_emp.var.rate : -0.264570116997
N_cons.price.idx : 0.947631716728
N_cons.conf.idx : 0.120204687119
N_euribor3m : -1.178358078
N_nr.employed : 0.0171953588724
subscription=yes : -0.19165968895

SVM

Attribute:Weight
Orange.feature.Continuous 'month=nov':12.529842545
Orange.feature.Continuous 'pdays=27':1.41421356237
Orange.feature.Continuous 'pdays=8':0.111575229528
Orange.feature.Continuous 'month=oct':3.41394718612
Orange.feature.Continuous 'pdays=7':8.85986421191
Orange.feature.Continuous 'poutcome=failure':2.10047484795
Orange.feature.Continuous 'month=sep':13.3688283558
Orange.feature.Continuous 'pdays=11':2.67178054233

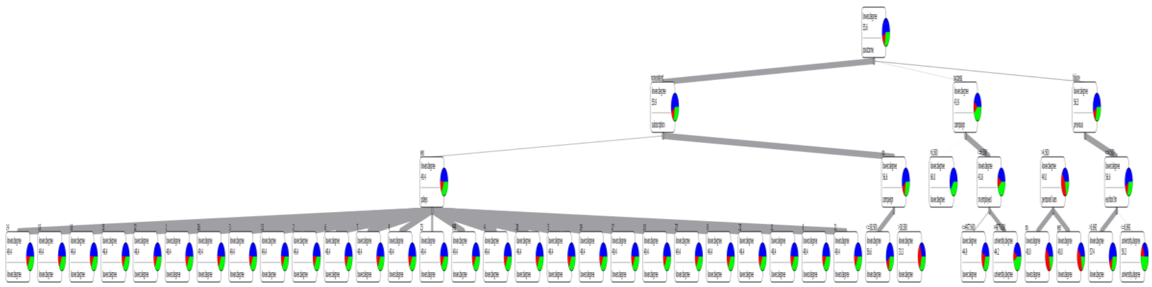
Orange.feature.Continuous 'pdays=6':2.27451768362
Orange.feature.Continuous 'month=aug':28.8905798974
Orange.feature.Continuous 'N_emp.var.rate':2.67528777411
Orange.feature.Continuous 'pdays=16':0.689460186421
Orange.feature.Continuous 'day_of_week=fri':0.336248253853
Orange.feature.Continuous 'month=jul':9.74342300785
Orange.feature.Continuous 'pdays=12':2.543087236
Orange.feature.Continuous 'month=jun':2.93804050559
Orange.feature.Continuous 'pdays=5':2.04652394145
Orange.feature.Continuous 'month=may':23.5374361924
Orange.feature.Continuous 'pdays=4':0.708631661545
Orange.feature.Continuous 'month=apr':18.6824263197
Orange.feature.Continuous 'pdays=3':5.96503026229
Orange.feature.Continuous 'pdays=999':2.22394700934
Orange.feature.Continuous 'month=mar':21.5770472381
Orange.feature.Continuous 'pdays=15':2.44407130003
Orange.feature.Continuous 'month=feb':0.0
Orange.feature.Continuous 'day_of_week=thu':1.18262530939
Orange.feature.Continuous 'pdays=2':4.02429837769
Orange.feature.Continuous 'month=jan':0.0
Orange.feature.Continuous 'pdays=1':1.08534366523
Orange.feature.Continuous 'N_nr.employed':25.3112825601
Orange.feature.Continuous 'contact method=telephone':0.800897907159
Orange.feature.Continuous 'pdays=0':1.07652810889
Orange.feature.Continuous 'N_cons.price.idx':4.63286144831
Orange.feature.Continuous 'personal loan=yes':1.61062464318
Orange.feature.Continuous 'N_previous':0.275756940144
Orange.feature.Continuous 'housing loan=yes':1.4126533277
Orange.feature.Continuous 'N_campaign':0.377044034153
Orange.feature.Continuous 'N_duration':0.497242300895
Orange.feature.Continuous 'pdays=14':1.11397369548
Orange.feature.Continuous 'age=retired':19.0707728633
Orange.feature.Continuous 'poutcome=success':2.97442471903
Orange.feature.Continuous 'pdays=10':4.15289745805
Orange.feature.Continuous 'age=working':21.4748682113
Orange.feature.Continuous 'N_cons.conf.idx':25.149437431
Orange.feature.Continuous 'age=young':2.52881292356
Orange.feature.Continuous 'job=unemployed':17.4502473869
Orange.feature.Continuous 'day_of_week=wed':0.185611319787
Orange.feature.Continuous 'marital=single':2.0015896363
Orange.feature.Continuous 'pdays=13':1.0754087695
Orange.feature.Continuous 'loan default=yes':2.2360679775
Orange.feature.Continuous 'day_of_week=tue':0.642684615204
Orange.feature.Continuous 'subscription=yes':0.459638025053

Orange.feature.Continuous 'day_of_week=mon':0.158077736313
Orange.feature.Continuous 'month=dec':7.00014030599
Orange.feature.Continuous 'pdays=9':2.45216485474
Orange.feature.Continuous 'pdays=26':1.41421356237
Orange.feature.Continuous 'pdays=25':1.04053615043
Orange.feature.Continuous 'pdays=22':0.0
Orange.feature.Continuous 'pdays=21':1.41421356237
Orange.feature.Continuous 'pdays=20':0.0
Orange.feature.Continuous 'poutcome=nonexistent':0.882618945376
Orange.feature.Continuous 'N_euribor3m':4.99572231474
Orange.feature.Continuous 'pdays=19':2.44948974278
Orange.feature.Continuous 'pdays=18':1.41431256336
Orange.feature.Continuous 'pdays=17':0.640112339499

Sorted weights:
['0.0000', '0.0000', '0.0000', '0.0000', '0.1116', '0.1581', '0.1856', '0.2758', '0.3362', '0.3770', '0.4596', '0.4972', '0.6401', '0.6427', '0.6895', '0.7086', '0.8009', '0.8826', '1.0405', '1.0754', '1.0765', '1.0853', '1.1140', '1.1826', '1.4127', '1.4142', '1.4142', '1.4142', '1.4143', '1.6106', '12.5298', '13.3688', '17.4502', '18.6824', '19.0708', '2.0016', '2.0465', '2.1005', '2.2239', '2.2361', '2.2745', '2.4441', '2.4495', '2.4522', '2.5288', '2.5431', '2.6718', '2.6753', '2.9380', '2.9744', '21.4749', '21.5770', '23.5374', '25.1494', '25.3113', '28.8906', '3.4139', '4.0243', '4.1529', '4.6329', '4.9957', '5.9650', '7.0001', '8.8599', '9.7434']

Random Forest

Tree size: 45 nodes, 36 leaves



<u>Housing Loan</u>

Logistic Regression

Feature:Weight
age=young : -0.0156814623624
age=retired : -0.000417107454268
job=unemployed : -0.012183397077
marital=single : -0.0219527091831
education=professional.course : -0.0770765542984
education=university.degree : -0.0278219562024
loan default=yes : 0.000593431992456
personal loan=yes : -0.173203796148
contact method=telephone : 0.12320253253
month=jan : 0.0
month=feb : 0.0
month=mar : -0.00926439184695
month=apr : -0.0285507254303
month=jun : 0.102639354765
month=jul : -0.0434854440391
month=aug : -0.13774317503
month=sep : -0.0112824328244
month=oct : -0.000276654551271
month=nov : -0.0920785665512
month=dec : -0.00508864829317
day_of_week=mon : -0.0549845807254
day_of_week=tue : 0.0163714662194
day_of_week=wed : -0.0329482741654
day_of_week=fri : 0.0394499786198
N_duration : 0.0175373069942
N_campaign : -0.00183928955812
pdays=0 : -0.00490589486435
pdays=1 : 5.28007185494e-05
pdays=2 : -0.00246004690416
pdays=3 : 0.00548577448353
pdays=4 : -0.00411972729489
pdays=5 : -0.00160876684822
pdays=6 : -0.00458016432822
pdays=7 : 0.00474422704428
pdays=8 : -0.000736225163564
pdays=9 : 0.000355183350621
pdays=10 : -0.00578114530072
pdays=11 : -0.00263172551058
pdays=12 : 0.00563037954271
pdays=13 : -0.00234475033358
pdays=14 : 0.00019631498435
pdays=15 : 0.00363853876479

pdays=16 : 0.00287951389328
pdays=17 : 0.000414824957261
pdays=18 : -0.000151195024955
pdays=19 : -0.000296829035506
pdays=20 : -0.000416501512518
pdays=21 : 0.000117548734124
pdays=22 : -0.000323337502778
pdays=25 : 0.000500903872307
pdays=26 : 0.000509820063598
pdays=27 : 0.00051512446953
N_previous : 0.00927280075848
poutcome=failure : 0.00762043148279
poutcome=success : -0.0104972422123
N_emp.var.rate : -0.00365947955288
N_cons.price.idx : 0.128829970956
N_cons.conf.idx : 0.0513025783002
N_euribor3m : -0.000919912126847
N_nr.employed : 0.000638687924948
subscription=yes : -0.0119651881978

SVM

Running script:

Attribute:Weight
Orange.feature.Continuous 'month=aug':12.1055119466
Orange.feature.Continuous 'pdays=22':1.0
Orange.feature.Continuous 'month=sep':8.43497002125
Orange.feature.Continuous 'pdays=25':1.0
Orange.feature.Continuous 'month=oct':4.61550202593
Orange.feature.Continuous 'pdays=26':1.0
Orange.feature.Continuous 'month=nov':8.77164692245
Orange.feature.Continuous 'pdays=27':1.0
Orange.feature.Continuous 'month=dec':5.76907398552
Orange.feature.Continuous 'pdays=999':0.618635613471
Orange.feature.Continuous 'day_of_week=mon':0.00837879069149
Orange.feature.Continuous 'N_previous':0.945823170213
Orange.feature.Continuous 'day_of_week=tue':0.491282155737
Orange.feature.Continuous 'poutcome=failure':0.733622902073
Orange.feature.Continuous 'day_of_week=wed':0.180648843758
Orange.feature.Continuous 'poutcome=nonexistent':2.29308131803
Orange.feature.Continuous 'day_of_week=thu':0.262990155257
Orange.feature.Continuous 'poutcome=success':1.55945704132

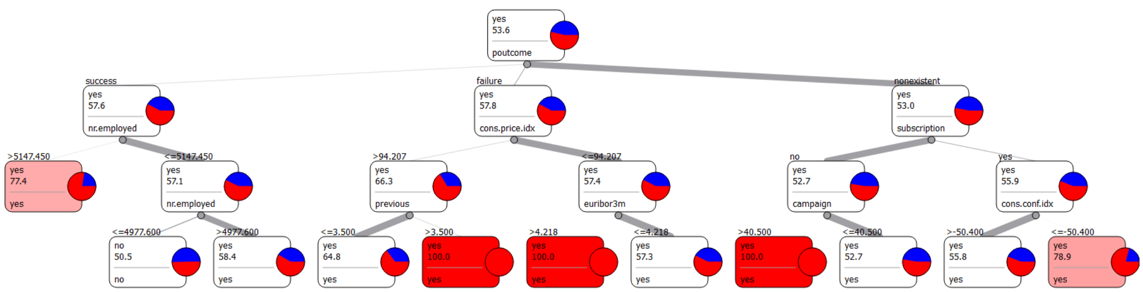
Orange.feature.Continuous 'day_of_week=fri':0.0392629913986
Orange.feature.Continuous 'N_emp.var.rate':1.00438915857
Orange.feature.Continuous 'N_duration':0.0711183659203
Orange.feature.Continuous 'N_cons.price.idx':33.4025142374
Orange.feature.Continuous 'N_campaign':0.199990845162
Orange.feature.Continuous 'N_cons.conf.idx':35.3283821735
Orange.feature.Continuous 'pdays=0':5.18028196692
Orange.feature.Continuous 'N_euribor3m':14.3543465117
Orange.feature.Continuous 'pdays=1':0.132715016603
Orange.feature.Continuous 'N_nr.employed':5.77117933781
Orange.feature.Continuous 'pdays=2':1.186381042
Orange.feature.Continuous 'subscription=yes':0.647874107584
Orange.feature.Continuous 'education=lower.degree':0.278525161557
Orange.feature.Continuous 'pdays=3':0.269212007523
Orange.feature.Continuous 'marital=single':0.0489259362221
Orange.feature.Continuous 'pdays=4':1.17778596282
Orange.feature.Continuous 'job=unemployed':0.540446069092
Orange.feature.Continuous 'pdays=5':1.26030503213
Orange.feature.Continuous 'age=young':0.742993012071
Orange.feature.Continuous 'pdays=6':1.83093697578
Orange.feature.Continuous 'age=working':1.29008848965
Orange.feature.Continuous 'pdays=7':4.0
Orange.feature.Continuous 'age=retired':0.547094102949
Orange.feature.Continuous 'pdays=8':1.0
Orange.feature.Continuous 'pdays=9':1.21610400081
Orange.feature.Continuous 'education=professional.course':0.377675935626
Orange.feature.Continuous 'pdays=10':2.18102699518
Orange.feature.Continuous 'education=university.degree':0.0991521487013
Orange.feature.Continuous 'pdays=11':1.48771800101
Orange.feature.Continuous 'loan default=yes':1.0
Orange.feature.Continuous 'pdays=12':6.0
Orange.feature.Continuous 'personal loan=yes':4.69693497755
Orange.feature.Continuous 'pdays=13':1.98855301738
Orange.feature.Continuous 'contact method=telephone':11.4740369283
Orange.feature.Continuous 'pdays=14':0.351562976837
Orange.feature.Continuous 'month=jan':0.0
Orange.feature.Continuous 'pdays=15':4.0
Orange.feature.Continuous 'month=feb':0.0
Orange.feature.Continuous 'pdays=16':5.0
Orange.feature.Continuous 'month=mar':15.0897010118
Orange.feature.Continuous 'pdays=17':0.0
Orange.feature.Continuous 'month=apr':7.46885704063
Orange.feature.Continuous 'pdays=18':1.0
Orange.feature.Continuous 'month=may':15.6527361926

Orange.feature.Continuous 'pdays=19':0.0
Orange.feature.Continuous 'month=jun':14.8734459812
Orange.feature.Continuous 'pdays=20':1.0
Orange.feature.Continuous 'month=jul':13.3880339498
Orange.feature.Continuous 'pdays=21':0.178094029427

Sorted weights:
['0.0000', '0.0000', '0.0000', '0.0000', '0.0084', '0.0393', '0.0489', '0.0711', '0.0992', '0.1327', '0.1781', '0.1806', '0.2000', '0.2630', '0.2692', '0.2785', '0.3516', '0.3777', '0.4913', '0.5404', '0.5471', '0.6186', '0.6479', '0.7336', '0.7430', '0.9458', '1.0000', '1.0000', '1.0000', '1.0000', '1.0000', '1.0000', '1.0000', '1.0000', '1.0044', '1.1778', '1.1864', '1.2161', '1.2603', '1.2901', '1.4877', '1.5595', '1.8309', '1.9886', '11.4740', '12.1055', '13.3880', '14.3543', '14.8734', '15.0897', '15.6527', '2.1810', '2.2931', '33.4025', '35.3284', '4.0000', '4.0000', '4.6155', '4.6969', '5.0000', '5.1803', '5.7691', '5.7712', '6.0000', '7.4689', '8.4350', '8.7716']

Random Forest

Tree size: 20 nodes, 11 leaves



Personal Loan

Logistic Regression

Feature:Weight
age=young : 0.0320375151932
age=retired : -0.0539315566421
job=unemployed : 0.00673824874684
marital=single : 0.0121101653203
education=professional.course : -0.0482903048396
education=university.degree : -0.0968355312943
loan default=yes : 0.00990772806108
housing loan=no : 0.197607859969
contact method=telephone : 0.0354083664715
month=jan : 0.0

month=feb : 0.0
month=mar : 0.137085929513
month=apr : 0.0610753186047
month=jun : 0.0745555981994
month=jul : -0.0659530982375
month=aug : 0.0697707086802
month=sep : 0.0275192856789
month=oct : 0.185503602028
month=nov : 0.0754924789071
month=dec : -0.122812569141
day_of_week=mon : -0.053557343781
day_of_week=tue : 0.0311442166567
day_of_week=wed : -0.011178961955
day_of_week=fri : -0.0705093443394
N_duration : -0.00476910918951
N_campaign : -0.00918189622462
pdays=0 : -0.017693458125
pdays=1 : 0.0494171567261
pdays=2 : -0.00989786349237
pdays=3 : -0.151905208826
pdays=4 : -0.0144122038037
pdays=5 : 0.0206113774329
pdays=6 : -0.0199571289122
pdays=7 : 0.0553008653224
pdays=8 : -0.0556034855545
pdays=9 : 0.00762492232025
pdays=10 : 0.0260684750974
pdays=11 : 0.035359274596
pdays=12 : -0.0542226806283
pdays=13 : -0.0417469255626
pdays=14 : -0.00396074401215
pdays=15 : 0.00843340810388
pdays=16 : 0.0355640165508
pdays=17 : 0.0272941291332
pdays=18 : 4.14461237597e-06
pdays=19 : 0.010343122296
pdays=20 : 0.00386681524105
pdays=21 : -0.0180152449757
pdays=22 : 0.0103676943108
pdays=25 : 0.00326311262324
pdays=26 : 0.00322048086673
pdays=27 : 0.00290901400149
N_previous : 0.00573470210657
poutcome=failure : -0.00794168747962

poutcome=success : -0.0170741267502
N_emp.var.rate : -0.0428452827036
N_cons.price.idx : 0.0183797832578
N_cons.conf.idx : 0.0136797754094
N_euribor3m : 0.086548730731
N_nr.employed : -0.000843588262796
subscription=yes : 0.0243089478463

SVM

Attribute:Weight
Orange.feature.Continuous 'pdays=2':0.0398549828678
Orange.feature.Continuous 'subscription=yes':0.0665944067296
Orange.feature.Continuous 'pdays=3':0.0047109471634
Orange.feature.Continuous 'pdays=4':0.00246111489832
Orange.feature.Continuous 'pdays=5':0.0355510190129
Orange.feature.Continuous 'age=young':0.0323977076914
Orange.feature.Continuous 'job=unemployed':0.0512758137193
Orange.feature.Continuous 'pdays=7':0.00447599077597
Orange.feature.Continuous 'marital=single':0.0641440171748
Orange.feature.Continuous 'pdays=8':0.0502289682627
Orange.feature.Continuous 'education=lower.degree':0.0484595418675
Orange.feature.Continuous 'pdays=9':0.0246319882572
Orange.feature.Continuous 'education=professional.course':0.0377906109206
Orange.feature.Continuous 'pdays=10':0.0359599571675
Orange.feature.Continuous 'education=university.degree':0.0106793222949
Orange.feature.Continuous 'pdays=11':0.0264100395143
Orange.feature.Continuous 'loan default=yes':0.0
Orange.feature.Continuous 'pdays=12':0.00606197305024
Orange.feature.Continuous 'housing loan=yes':0.0108843040653
Orange.feature.Continuous 'pdays=13':0.0542590916157
Orange.feature.Continuous 'contact method=telephone':0.0100892946357
Orange.feature.Continuous 'pdays=14':0.0503810350783
Orange.feature.Continuous 'month=jan':0.0
Orange.feature.Continuous 'pdays=15':0.00463998876512
Orange.feature.Continuous 'month=feb':0.0
Orange.feature.Continuous 'pdays=16':0.0182290002704
Orange.feature.Continuous 'month=mar':0.016680881381
Orange.feature.Continuous 'pdays=17':0.0364790000021
Orange.feature.Continuous 'age=retired':0.030250063166
Orange.feature.Continuous 'month=apr':0.0116962126922

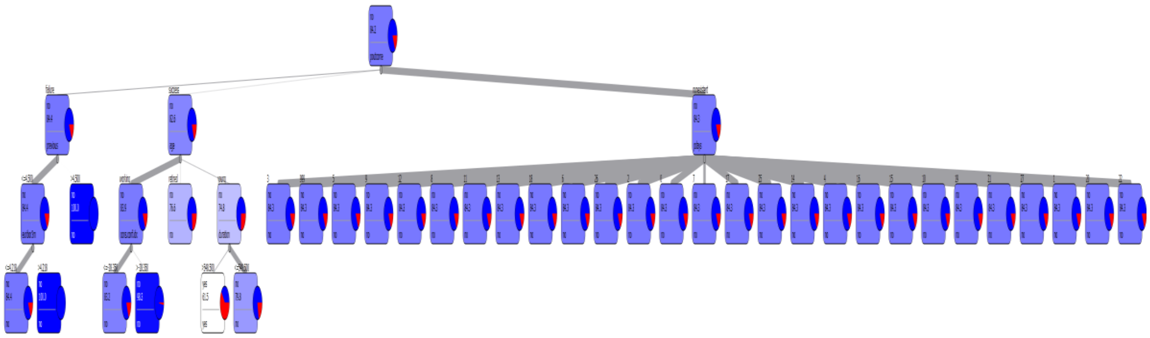
Orange.feature.Continuous 'pdays=18':0.0120670199394
Orange.feature.Continuous 'month=may':0.0463214349002
Orange.feature.Continuous 'pdays=19':0.0128739997745
Orange.feature.Continuous 'month=jun':0.0120318239788
Orange.feature.Continuous 'pdays=20':0.0
Orange.feature.Continuous 'month=jul':0.00512893893756
Orange.feature.Continuous 'pdays=21':0.144638001919
Orange.feature.Continuous 'month=aug':0.0151539663784
Orange.feature.Continuous 'pdays=22':0.0227680001408
Orange.feature.Continuous 'month=sep':0.000418924260885
Orange.feature.Continuous 'pdays=25':0.0199270006269
Orange.feature.Continuous 'month=oct':0.0112018296495
Orange.feature.Continuous 'age=working':0.0626373795094
Orange.feature.Continuous 'pdays=26':0.0415600016713
Orange.feature.Continuous 'month=nov':0.00286902068183
Orange.feature.Continuous 'pdays=27':0.0
Orange.feature.Continuous 'month=dec':0.0174540802836
Orange.feature.Continuous 'pdays=999':0.0887402853696
Orange.feature.Continuous 'day_of_week=mon':0.0131531972438
Orange.feature.Continuous 'N_previous':0.00327629918968
Orange.feature.Continuous 'day_of_week=tue':0.0593487218721
Orange.feature.Continuous 'poutcome=failure':0.0202787742019
Orange.feature.Continuous 'day_of_week=wed':0.0291630814318
Orange.feature.Continuous 'poutcome=nonexistent':0.0418211034266
Orange.feature.Continuous 'pdays=6':0.00735498638824
Orange.feature.Continuous 'day_of_week=thu':0.0107871657237
Orange.feature.Continuous 'poutcome=success':0.0215527205728
Orange.feature.Continuous 'day_of_week=fri':0.0278092175722
Orange.feature.Continuous 'N_emp.var.rate':0.0282111033048
Orange.feature.Continuous 'N_duration':0.0123315249789
Orange.feature.Continuous 'N_cons.price.idx':0.00851548450419
Orange.feature.Continuous 'N_campaign':0.0527776315703
Orange.feature.Continuous 'N_cons.conf.idx':0.0119128506385
Orange.feature.Continuous 'pdays=0':0.0199030525982
Orange.feature.Continuous 'N_euribor3m':0.0241888076089
Orange.feature.Continuous 'pdays=1':0.0218559876084
Orange.feature.Continuous 'N_nr.employed':0.0600524797974

Sorted weights:
['0.0000', '0.0000', '0.0000', '0.0000', '0.0000', '0.0004', '0.0025', '0.0029', '0.0033', '0.0045', '0.0046', '0.0047', '0.0051', '0.0061', '0.0074', '0.0085', '0.0101', '0.0107', '0.0108', '0.0109', '0.0112', '0.0117', '0.0119', '0.0120', '0.0121', '0.0123', '0.0129', '0.0132', '0.0152', '0.0167', '0.0175', '0.0182', '0.0199', '0.0199', '0.0203', '0.0216', '0.0219', '0.0228', '0.0242', '0.0246', '0.0264', '0.0278', '0.0282', '0.0292', '0.0303', '0.0324', '0.0356', '0.0360', '0.0365',

'0.0378', '0.0399', '0.0416', '0.0418', '0.0463', '0.0485', '0.0502', '0.0504', '0.0513', '0.0528', '0.0543', '0.0593', '0.0601', '0.0626', '0.0641', '0.0666', '0.0887', '0.1446']

Random Forest

Tree size: 42 nodes, 35 leaves



<u>Term Deposit Subscription</u>

Logistic Regression

Feature:Weight
age=young : -0.245843276381
age=retired : -0.360809653997
job=unemployed : -0.224647328258
marital=single : -0.0923094004393
education=professional.course : -0.160418212414
education=university.degree : -0.312785893679
loan default=yes : 0.00105567451101
housing loan=no : -0.0222909655422
personal loan=yes : 0.0688914358616
contact method=telephone : 0.689092218876
month=jan : 0.0
month=feb : 0.0
month=mar : -1.63970327377
month=apr : -0.453877449036
month=jun : -0.622804760933
month=jul : -0.566721498966
month=aug : -0.602766513824

month=sep : 0.227541357279
month=oct : -0.0898834094405
month=nov : 0.0486870221794
month=dec : -0.0771316960454
day_of_week=mon : 0.235879138112
day_of_week=tue : 0.0279677789658
day_of_week=wed : -0.0452449098229
day_of_week=fri : 0.111136101186
N_duration : -1.18916046619
N_campaign : 0.119098544121
pdays=0 : -0.0381787978113
pdays=1 : 0.0716212242842
pdays=2 : -0.11003806442
pdays=3 : -0.427813977003
pdays=4 : 0.00609106989577
pdays=5 : -0.0594638101757
pdays=6 : -0.317229688168
pdays=7 : -0.105529770255
pdays=8 : -0.018808202818
pdays=9 : -0.0347017273307
pdays=10 : -0.0583988465369
pdays=11 : -0.00120950385462
pdays=12 : -0.0102109434083
pdays=13 : -0.0886419564486
pdays=14 : 0.0169091522694
pdays=15 : -0.0909593254328
pdays=16 : 0.00446212012321
pdays=17 : 0.0455099083483
pdays=18 : -0.0133682452142
pdays=19 : 0.0218748040497
pdays=20 : 0.0141771035269
pdays=21 : -0.0168494097888
pdays=22 : -0.00476881489158
pdays=25 : -0.014628986828
pdays=26 : -0.00905222259462
pdays=27 : -0.00648868503049
N_previous : -0.054204184562
poutcome=failure : 0.6207010746
poutcome=success : -0.915919244289
N_emp.var.rate : 0.957746505737
N_cons.price.idx : -0.522663295269
N_cons.conf.idx : -0.129644051194
N_euribor3m : -0.429043233395
N_nr.employed : 0.00952477380633

SVM

Attribute:Weight
Orange.feature.Continuous 'month=jun':0.392322070897
Orange.feature.Continuous 'pdays=20':1.0
Orange.feature.Continuous 'month=jul':0.0609140852466
Orange.feature.Continuous 'pdays=21':2.0
Orange.feature.Continuous 'month=aug':0.571137018502
Orange.feature.Continuous 'pdays=22':1.0
Orange.feature.Continuous 'month=sep':1.02324098349
Orange.feature.Continuous 'pdays=25':1.0
Orange.feature.Continuous 'month=oct':0.128225058317
Orange.feature.Continuous 'pdays=26':1.0
Orange.feature.Continuous 'month=nov':0.822292964906
Orange.feature.Continuous 'pdays=27':1.0
Orange.feature.Continuous 'month=dec':0.610457986593
Orange.feature.Continuous 'pdays=999':30.6339498467
Orange.feature.Continuous 'day_of_week=mon':1.31530005112
Orange.feature.Continuous 'N_previous':0.292706302131
Orange.feature.Continuous 'day_of_week=tue':0.450274035335
Orange.feature.Continuous 'poutcome=failure':13.9019240951
Orange.feature.Continuous 'day_of_week=wed':0.417420107871
Orange.feature.Continuous 'poutcome=nonexistent':12.7320257516
Orange.feature.Continuous 'day_of_week=thu':0.374166072346
Orange.feature.Continuous 'poutcome=success':26.6339559434
Orange.feature.Continuous 'day_of_week=fri':0.0734459322412
Orange.feature.Continuous 'N_emp.var.rate':7.64534231825
Orange.feature.Continuous 'N_duration':46.0415652456
Orange.feature.Continuous 'N_cons.price.idx':6.114876625
Orange.feature.Continuous 'N_campaign':3.87632828864
Orange.feature.Continuous 'N_cons.conf.idx':2.85900341302
Orange.feature.Continuous 'pdays=0':2.90082299709
Orange.feature.Continuous 'N_euribor3m':1.23821359042
Orange.feature.Continuous 'pdays=1':10.0
Orange.feature.Continuous 'N_nr.employed':3.39324707971
Orange.feature.Continuous 'education=lower.degree':0.200107811717
Orange.feature.Continuous 'pdays=2':4.45513898134
Orange.feature.Continuous 'marital=single':0.502298877575
Orange.feature.Continuous 'pdays=3':5.39441198111
Orange.feature.Continuous 'job=unemployed':0.00127203762531

Orange.feature.Continuous 'pdays=4':2.0
Orange.feature.Continuous 'age=young':0.0870669856668
Orange.feature.Continuous 'pdays=5':4.08749395609
Orange.feature.Continuous 'age=working':0.429104884854
Orange.feature.Continuous 'pdays=6':4.71887002885
Orange.feature.Continuous 'age=retired':0.342043995857
Orange.feature.Continuous 'pdays=7':4.9036199972
Orange.feature.Continuous 'pdays=8':3.20216500759
Orange.feature.Continuous 'education=professional.course':0.337528951466
Orange.feature.Continuous 'pdays=9':1.77904999256
Orange.feature.Continuous 'education=university.degree':0.137415043078
Orange.feature.Continuous 'pdays=10':1.40852099191
Orange.feature.Continuous 'loan default=yes':0.0
Orange.feature.Continuous 'pdays=11':0.68905299902
Orange.feature.Continuous 'housing loan=yes':0.0185050470755
Orange.feature.Continuous 'pdays=12':6.0
Orange.feature.Continuous 'personal loan=yes':0.0455370573327
Orange.feature.Continuous 'pdays=13':5.05583100021
Orange.feature.Continuous 'contact method=telephone':1.80350393709
Orange.feature.Continuous 'pdays=14':0.805100023746
Orange.feature.Continuous 'month=jan':0.0
Orange.feature.Continuous 'pdays=15':4.23387798667
Orange.feature.Continuous 'month=feb':0.0
Orange.feature.Continuous 'pdays=16':0.0
Orange.feature.Continuous 'month=mar':2.68947494309
Orange.feature.Continuous 'pdays=17':4.0
Orange.feature.Continuous 'month=apr':0.462276889244
Orange.feature.Continuous 'pdays=18':1.0
Orange.feature.Continuous 'month=may':2.28462386504
Orange.feature.Continuous 'pdays=19':1.0

Sorted weights:
['0.0000', '0.0000', '0.0000', '0.0000', '0.0013', '0.0185', '0.0455', '0.0609', '0.0734', '0.0871', '0.1282', '0.1374', '0.2001', '0.2927', '0.3375', '0.3420', '0.3742', '0.3923', '0.4174', '0.4291', '0.4503', '0.4623', '0.5023', '0.5711', '0.6105', '0.6891', '0.8051', '0.8223', '1.0000', '1.0000', '1.0000', '1.0000', '1.0000', '1.0000', '1.0000', '1.0232', '1.2382', '1.3153', '1.4085', '1.7790', '1.8035', '10.0000', '12.7320', '13.9019', '2.0000', '2.0000', '2.2846', '2.6895', '2.8590', '2.9008', '26.6340', '3.2022', '3.3932', '3.8763', '30.6339', '4.0000', '4.0875', '4.2339', '4.4551', '4.7189', '4.9036', '46.0416', '5.0558', '5.3944', '6.0000', '6.1149', '7.6453']

Random Forest

Tree size: 16 nodes, 10 leaves