Data Science for

Small Businesses

by

Aveesha Sharma

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

Approved April 2016 by the
Graduate Supervisory Committee:

Arbi Ghazarian, Chair
Ashraf Gaffar
Srividya Bansal

ARIZONA STATE UNIVERSITY

May 2016

ABSTRACT

This reports investigates the general day to day problems faced by small businesses, particularly small vendors, in areas of marketing and general management. Due to lack of man power, internet availability and properly documented data, small business cannot optimize their business. The aim of the research is to address and find a solution to these problems faced, in the form of a tool which utilizes data science. The tool will have features which will aid the vendor to mine their data which they record themselves and find useful information which will benefit their businesses. Since there is lack of properly documented data, One Class Classification using Support Vector Machine (SVM) is used to build a classifying model that can return positive values for audience that is likely to respond to a marketing strategy. Market basket analysis is used to choose products from the inventory in a way that patterns are found amongst them and therefore there is a higher chance of a marketing strategy to attract audience. Also, higher selling products can be used to the vendors' advantage and lesser selling products can be paired with them to have an overall profit to the business. The tool, as envisioned, meets all the requirements that it was set out to have and can be used as a stand alone application to bring the power of data mining into the hands of a small vendor.

DEDICATION

I dedicate this text to the memory of my late mother, Alka Sharma, who would have been very proud to watch me graduate. May this text serve as a symbolic presence of her at this prestigious university.

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Arbi Ghazarian for giving me this opportunity to work with him. I would also like to thank my parents Ravi Sharma and Alka Sharma for being ever so supportive of everything.

I express my gratitude to my friend and an ASU alumni Rachit Agarwal for his constant support throughout graduate school and to my friend and colleague Mandar Patwardhan for unknowingly being an amazing figure of inspiration.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER 1

INTRODUCTION

"Well-run small businesses naturally form learning relationships with their customers. Over time, they learn more and more about their customers, and they use that knowledge to serve them better. The result is happy, loyal customers and profitable businesses" [1].

The statement holds true when a said small business is 'well-run' and then may develop relationships with their customers 'over time'. However, in the growing world of commerce, where larger firms employ data scientists to convert customer data into customer knowledge to improve their customer relationships and subsequently their businesses, a small vendor is powerless with little or no knowledge of data science and possibly oblivious to the benefits of data science solutions to their business.

Although analytics tools like Watson Analytics [2] and Mixpanel [3] exist, there is a large variety of vendors across the world, hailing from developed countries with sophisticated technology for accepting electronic payments and infrastructure to capture customer data, to small vendors from developing countries who primarily deal with customers through cash and no electronic means to keep a track of their purchase history.

According to a study of the UN, in the 48 of the UN designated Lease Developed Countries (LCD), 90% of the population does not have any kind of internet connectivity. This counters the possibility of using aforementioned analytics tools. Other studies [4] show us that inexperienced or ill informed business owners do not trust these tools and others welcome changes in their routine business dealings.

This report investigates the posing complication and find means to benefit those traditional, old school vendors by allowing them to build relevant data on their own, in due time, and help them recognize and address problems in marketing, customer relationships and retention, maintaining a successful business and eventually grow it further.

CHAPTER 2

PRE REQUISITES

2.1 Background Overview

The very concept of data mining dates back to the times when computers did not exist, in the sense that manual extraction of patterns or new information from data has been trending for centuries using techniques like Bayes' Theorem in the 1700s and Regression Analysis in the 1800s [1][5].

There are many definitions to describe this vast field of data science and data mining particularly. Generally speaking, the misnomer 'data mining' refers to the process of extraction or 'mining' of knowledge from large amounts of data [6]. Authors Kantardzic and Wiley [7] suggest that in many domains, with the flourishing computer industry and its usage, a huge amount of data is being generated that cannot be mathematically formalized owing to its complexity. Therefore, the paradigm of analyzes is shifting from classical modeling based on first principles to developing models to analyze the data directly.

Rightfully said, data mining is an iterative process and the discovery of new and valuable information, through automatic or manual methods, marks its progress. It is essential that no notions are fixed in advance about an outcome to be deemed interesting for data mining to be useful [7]. Authors in [7] deftly describe the two main goals of data mining to be prediction, which creates the model of the system as described by the data, and

description, where new, nontrivial information is produced based on the available data. They also list out the following primary data mining tasks:

1. Classification - Discovery of a predictive learning function that classifies a data item into one of several predefined classes or labels.

2. Regression - Discovery of a predictive learning function that maps a data item to a real-value prediction variable.

3. Clustering - A common descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data.

4. Summarization -An additional descriptive task that involves methods for finding a compact description for a set (or subset) of data.

5. Dependency Modeling - Finding significant dependencies or correlations between variables or between the values of a feature in a data set or in a part of a data set.

6. Change and Deviation Detection - Discovering the most significant changes in the data set.

"The success of a data-mining engagement depends largely on the amount of energy, knowledge, and creativity that the designer puts into it. In essence, data mining is like solving a puzzle" [7]. The research ahead entails why small business vendors would benefit from engaging time in maintaining and updating customer records and harnessing the power of analyzing data for profits.

2.2 Literature Review

The internet is full of articles and journals buzzing about the word 'data mining' for business and everyone seems to be talking about how bigger companies have used this for their own and customers' benefits [8][9]. There seems to be little literature on how relevant data mining can be, to small business and how to use it.

Need for the research

Huang and Brown [10] remark that the comprehension of the problems faced by small businesses is of practical importance for those involved in it, particularly the owners or managers. Their paper investigates the types of problems and their relative significance. The sample on which they base their study on predominantly consists of average performers and survivors, that is businesses that are neither growing too fast nor failures. The data they studied was availed from a contact log of a manager who participated in The Business Grow Program in 1995 in Western Australia. The main of the program was to provide consultancy services and training to small business operators. The data in the log was accumulated over a period of 2 years and 4 months. The data encompassed detailed information of 973 small business owners, including the problems faced by them, every contact made with clients with names, date of contact and actions taken to resolve problems aforementioned.

They analyzed the number of problems encountered in each type of problems faced and the results interesting. The most common area of problems faced was in the domain of

Sales and Marketing with a whopping 40.2% of the data. The subcategories for the domain included Market Research, Low Sales, Dependence on a few customers, Promoting/advertising, increased competition. Small owners have little appreciation for the marketing concept and hence lack the required skills. Since marketing is known to be the most vital of all business activities and important for the growth and sustenance of small businesses [11] [12], it is essential for small business owners to pay attention to selection of promotional media, content design and format of the promotional materials, market size and location among other factors that came up in the study [10]. Another significant area where problems were encountered was General Management which took up 14.3% of the results. The domain included Lack of management experience, Only one person/no time, Administrative problems and Planning, which was the most frequent issue among others.

Another study that was undertaken was the above problems faced across industries. The authors [9] found that the highest amount of Sales/Marketing problems occurred in the Wholesale and Retailing sector, which turned out to be 59% of the documented cases. Also, the same sector surfaced as the second most troubled industry in terms of General Management with 25.3% of the total cases.

When compared by the number of employees, 61.8% cases of Sales/Marketing problems occurred when the employees in a business were 4 or less in number. The statistics were 21.3% for the General Management domain. Our research attempts to aid the business owners to handle these issues better.

Selection of Method

The One Class Classification (OCC) or the Unary Classification is different from the mainstream binary and multiclass classification problems, which aim to classify unknown objects into several pre-defined classes or labels [13]. The difference lies in the training data that is used to build the classification model. Conventionally, training data is allegedly known to have instances from all the classes or labels (in the case of multiclass classification) or both the classes (in the case of binary classification). However, in OCC the negative class is either absent or not properly sampled/labeled.

Hence, the task in OCC is to classify positive cases when the negative cases or the outliers are not present. We can conclude from studies [10] [14] [15] that there the prospects of small business owners having access to an appropriately characterized training data sets are not promising. Therefore, OCC seems like a good solution to the problem of having little or no instances that would be a sample of a negative concept or non-target class(es). However, the same fact makes OCC harder than multiclass or binary classification too. The main aim in OCC is to set out or clarify a classification boundary around the target class, such that it acquires as many data objects as possible from the same class and minimizes the acceptance of negative or outlier objects [13].

It is difficult to decide how tight the boundary should be for the fitting of the model and which attributes should be used to find the best distinction between positive and negative class objects. Therefore, it can be presumed that the required number of training instances will be relatively more in OCC than compared to multiclass classification [16].

Authors in [13] have devised a new taxonomy in their paper which gives out three broad categories for the study of OCC problems. They are not mutually exclusive and the authors claim that key contributions in most OCC research fall into one of the categories mentioned. They are summarized in Fig 1.

1. Availability of Training Data: Learning with positive data only, or with a few instances of negative data samples, or learning with positive data and unlabeled data.
2. Methodology Used: Algorithms based on One Class Support Vector Machines (OSVMs) or methodologies based on other algorithms (OSVMs)
3. Application Domain Applied: OCC applied for text or document classification or in other application domains.

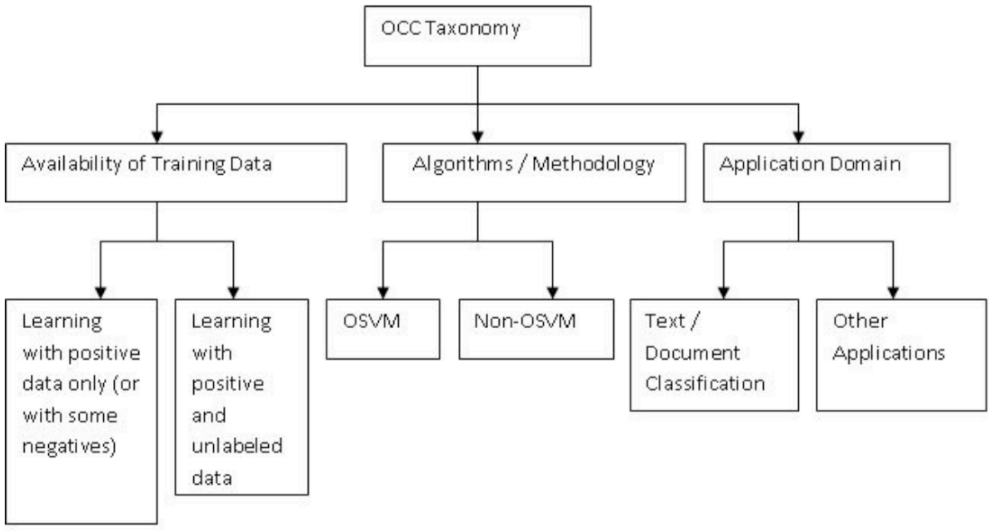


Figure 1. Taxonomy for study of OCC Techniques

Many approaches to OSVMs have been discussed in [13]. The authors pitch the theories of Tax and Duin [17] [18] against Scholkopf et al. [19] which are both approaches to address the OCC problem using positive examples only. Tax and Dunn [17] [18] attempted to solve the problem by considering a hyper sphere around the positive class data that includes most points in the data with a minimum radius. The method is known as Support Vector Data Description (SVDD). This model poses the risk of rejecting some measure of the positive training objects. Scholkopf et al [19] [20], on the other hand, propose another possible approach and suggest using a hyper plane. They try to segregate the region contacting data from the area containing none. The hyper plane is formed maximally distant from the origin, with all data points lying on the far side from the origin and such that the margin is positive. In OCC, as the negative data is not available, in most cases, the outliers are assumed to be uniformly distributed and the posterior probability can be estimated. Tax [17] mentions that in some OCC methods, distance is estimated instead of probability for one class classifier ensembling. Tax observes that the use of ensembles in OCC improves performance, especially when the product rule is used to combine the probability estimates. However, classifier ensembles have not been exploited much for OCC problems. Depending upon the data availability, algorithm use and application, appropriate OCC techniques can be applied and improved upon [13].

One Class Support Vector Machine is also known to be used for customer churn prediction [20]. Churn is described as the customer's decision to end the relationship with a company and switch to a competitor. In today's aggressive commercial era, churn prediction is a pressing concern. It is important to take appropriate steps to retain the

customers, especially in a small business, where the dependence on a few customers can be high as concluded from the study in [10]. One of the major characteristics of customer churn prediction is that the amount of churn customers, which constitute the negative samples in a data set, are insignificant [21]. Therefore, the standard SVM will not work well as it will for a standard binary classification problem, which is what customer churn prediction is generally considered as. Authors in [21] have presented an improved one class SVM method to predict churn. The experiment of the study was based on the fact that churn hits when customers are discontented with the price or quality of service when compared to competitors. The authors categorized their input variables as follows:

1. Demographics: Geographic and population data of a given region.

2. Usage level. Call detail records (date, time, duration, and location of all calls), peak / off-peak minutes used, additional minutes beyond monthly prepaid limit etc.

3. Quality of Service (QOS): Dropped calls (calls lost due to lack of coverage or available bandwidth), and quality of service data (interference, poor coverage).

4. Features / Marketing: Details of service bundle such as email, instant messaging, paging, rate plans offered by the carrier and its competitors, recent entry of competitors into the market, advertising campaigns, etc.

The dataset used included the description of 100,000 customers provided by a wireless telecom company that included 171 variables. The data was collected over a period of three months and the aim was to predict the churn in the fifth month. The authors administered the experiment on 2958 examples out of which 2134 were for training data set with 152 churn examples and 824 were for testing set with 67 churn examples.

The authors used different Kernel function in SVM and obtained different accuracy rates for each one The Linear Kernel yielded a 72.28% accuracy of prediction whereas the Polynomial Kernel yielded 77.65% accuracy. The Gaussian Kernel faired better than the others with 87.15% prediction accuracy.

The authors also compared the results of Gaussian Kernel SVM with other algorithms such as ANN, which yielded 78.12% accuracy, Decision Tree with 62% accuracy and Naïve Bays with 83.24% accurate results. As mentioned before, the number of the negative examples is too small and that forms the reason why the generalization performance of SVM classifier is weak, and the error rates unacceptable [21]. However, from the experiment, it is evident that use of One-class SVM for customer churn prediction is promising.

Another known use of One class classification using SVM is information retrieval. Opinion mining is a growing area of information gathering [22] [23]. Online review sites and personal blogs can be a good source of feedback from customers. Therefore, One

Class Classification using SVM seems a very promising method for implementing in this research project.

CHAPTER 3

PROBLEM STATEMENT

We can comprehend, from the literature study that was undertaken, that small businesses, particularly vendors in retailing sector, face a lot of problems in sales or marketing. They also face problems in general management of the business owing to the their less number of employees. Other sources [24] [25] tell us that cash is, by far, the most used method of payment. On an international scale, the percentage of cash transactions among all transactions range from 75% to as high as 90%, particularly when the payments are low value.

All the above mentioned factors cause a problem for vendors who have no means to collect data on their regular customers electronically through their purchase history or tracking their expenses through credit/debit card details. The research in this document will entail a scheme to help the small businesses or vendors in the retailing sector to use the potential of data mining to boost their sales and organize their business better, targeted principally at those vendors who do not feel comfortable with existing paid analytical tools or are looking for a challenge to break their monotonous routine [26] [27].

The proposed solution to this problem is to create an intuitive tool that may be used by the retail vendors to create datasets themselves and use them to determine the target audience for marketing (any deals or offers) and encourage their sales. A UI is ‘intuitive’

when users understand its behavior and effect without use of reason, experimentation, assistance, or special training. The tool will also be able to analyze data to predict the likelihood of the customer to respond positively to a marketing offer and help the vendor come up with customized marketing strategy for the audience. We can also aim to help with the general management of the business by mining for association rules, or more specifically to perform market basket analysis.

The methodology chosen for this tool is One-class classification using Support Vector Machine (SVM) and the reason for this selection is the possibility of less or unavailability of negative examples, as studied in the literature review and the promising results SVM offers. Market Basket Analysis or Association Rules and the GUI are also implemented using R language.

CHAPTER 4

DATA MINING TOOL - OCCUR

R first appeared 23 years ago in 1993 and was designed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand. It was an implementation of the language S and is now maintained by the R Development Core Team. It is a programming language and and environment for statistical computing and is widely used for data analysis. Surveys indicate R's increasing popularity in recent years.

Written primarily in C, Fortran and R, it is a freely available for use under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems. It in an interpreted and highly extensible language and has a command line interface. An integrated development environment (IDE) for R named RStudio was used to program the tool. RStudio, written in C++, is free and open source and was developed by JJ Allaire, the creator of the language ColdFusion.

4.1 Overview of the tool OCCUR

The tool is named OCCUR, an acronym for One Class Classification Using R. The tool aims to assist small business vendors apply data mining to the data they record themselves and improve their business by applying data analysis techniques to market products better.

The GUI of the tool is developed using the library gWidgets, which provides a toolkit

independent API for building GUIs for the R programming language. It provides a high-level wrapper over R packages that interface gtk, tcltk and qt. The GUI toolkit used in this tool is the Tcl/Tk. The Tcl programming language was created in 1988 by John Ousterhout at the University of California, Berkeley.

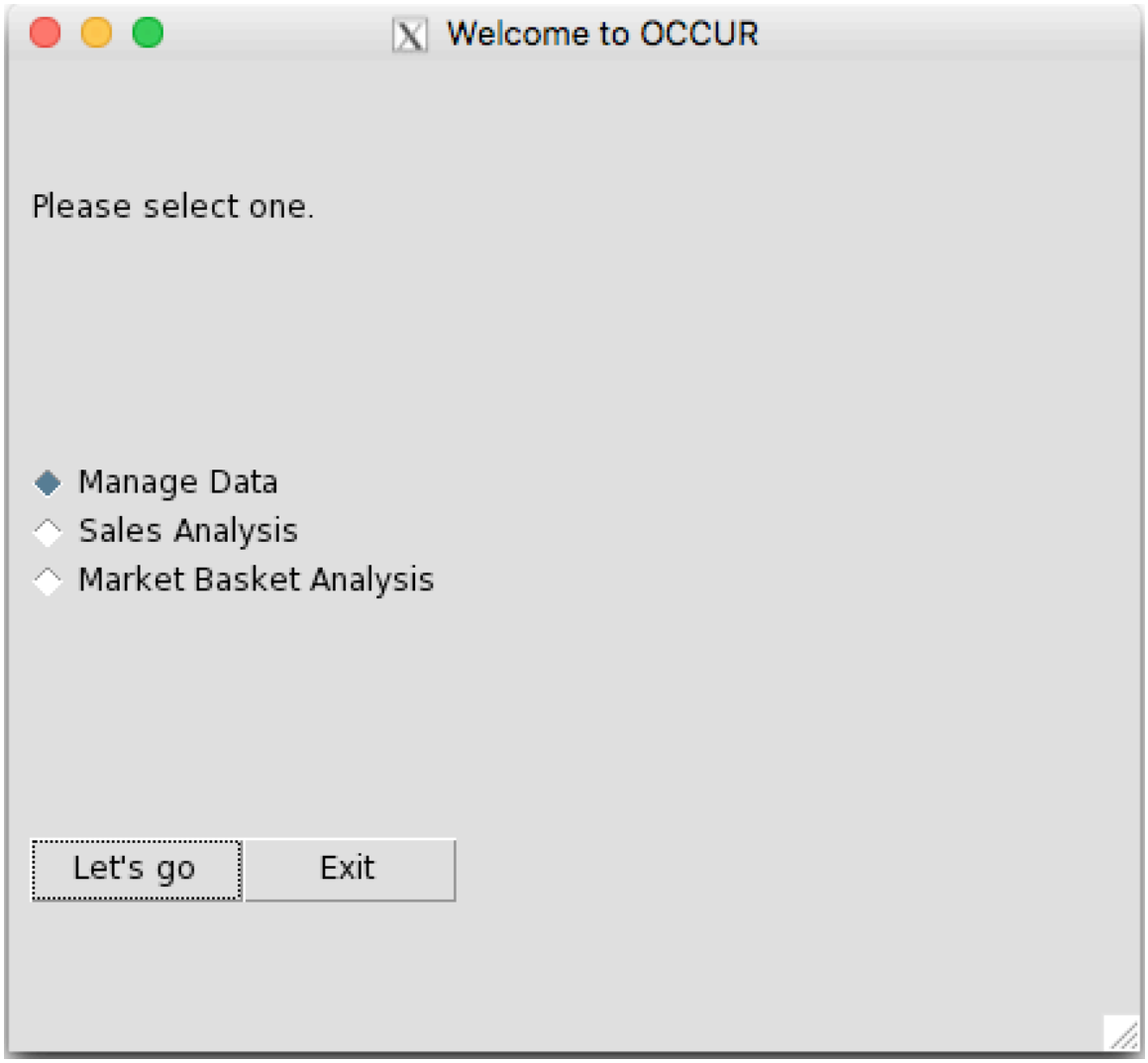


Figure 2. Main window/screen of the tool OCCUR

The main screen of the tool is fig 2. There are three options available to the user through a radio button. Each option, when selected, and the button clicked opens a new resizable window with individual functionality. The exit button quits the tool.

4.2 Module 'Manage Data'

This module of the tool OCCUR aims to create an interface for the user from where he can manage his customer details as well as his inventory of items and the sale made. The window is divided clearly under two categories of Manage Customer Data and Manage Items.

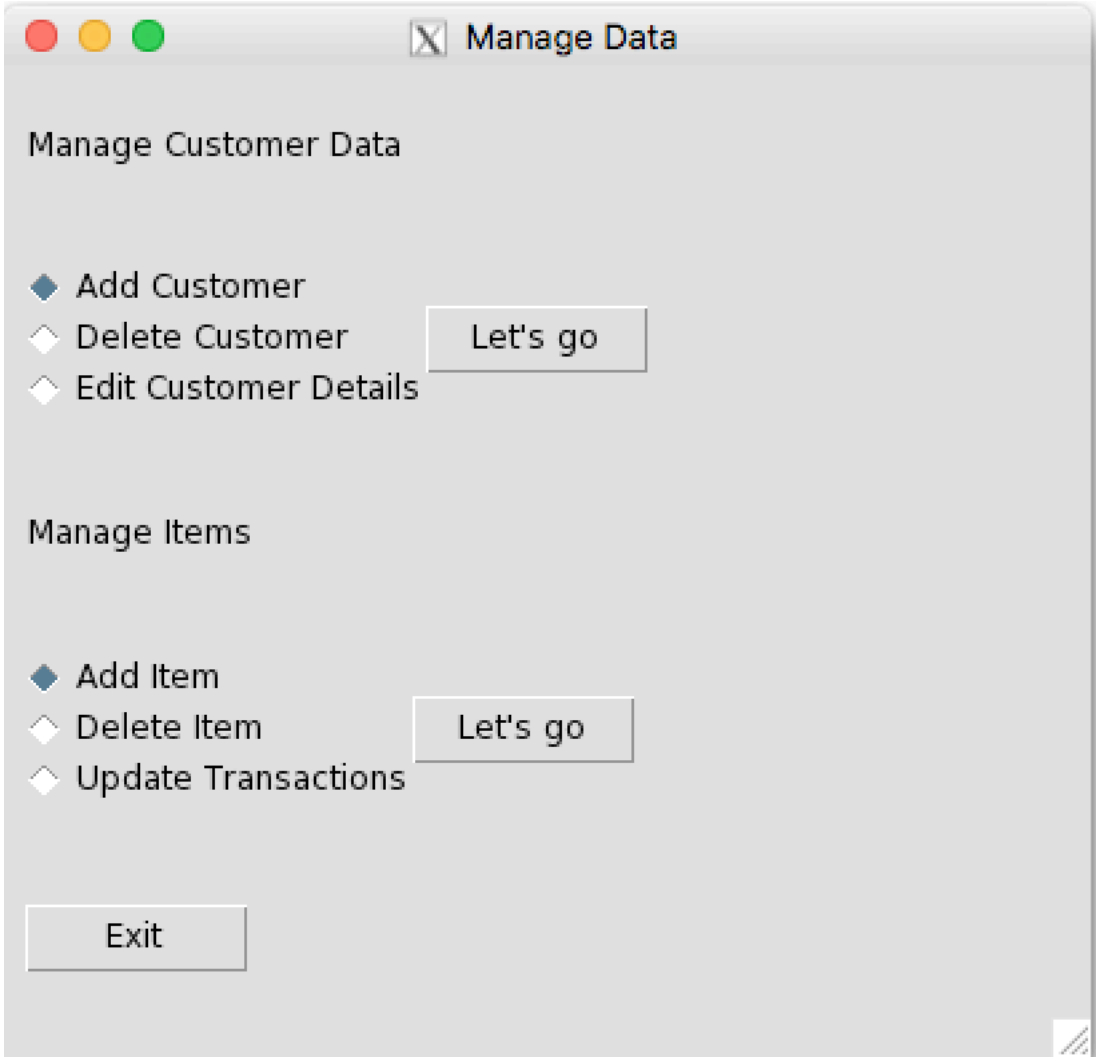


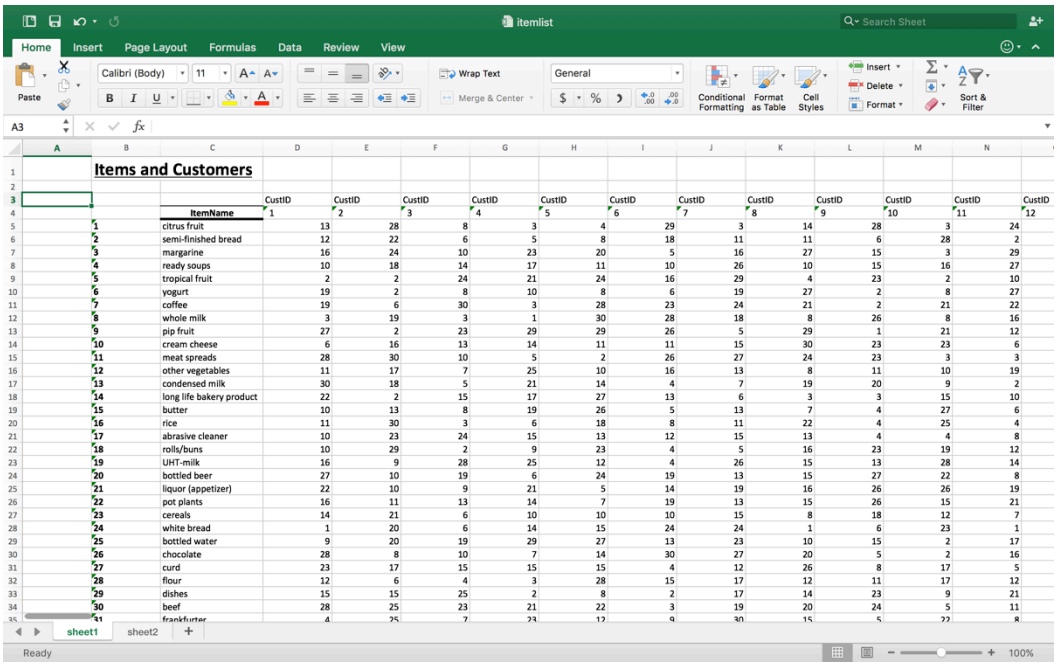
Figure 3. The 'Manage Data' screen of the tool OCCUR

1. Add Customer

On selecting this option, an MS Excel sheet opens up which contains the details of all the customers who shop with the vendor, as shown in Figure 4. The vendor can mark them with a unique ID and record their items purchased and increment the corresponding cell.

This data will contribute to building models and further analysis.

If a customer wishes to be registered with the vendor to receive offers or deals, they can opt to do so. The vendor can store their details in another sheet shown in Figure 5. The vendor will have an easy time corresponding registered customers with their customer ID. All the data is expected to be maintained and recorded by the vendor or user of the tool.



Items and Customers

		CustID	CustID	CustID	CustID	CustID	CustID	CustID	CustID	CustID	CustID	CustID	CustID
	ItemName	1	2	3	4	5	6	7	8	9	10	11	12
1	citrus fruit	13	28	8	3	4	29	3	14	28	3	24	
2	semi-finished bread	12	22	6	5	8	18	11	11	6	28	2	
3	margarine	16	24	10	23	20	5	16	27	15	3	29	
4	ready soups	10	18	14	17	11	10	26	10	15	16	27	
5	tropical fruit	2	2	24	21	24	16	29	4	23	2	10	
6	yogurt	19	2	8	10	8	6	19	27	2	8	27	
7	coffee	19	6	30	3	28	23	24	21	2	21	22	
8	whole milk	3	19	3	1	30	28	18	8	26	8	16	
9	pip fruit	27	2	23	29	29	26	5	29	1	21	12	
10	cream cheese	6	16	13	14	11	11	15	30	23	23	6	
11	meat spreads	28	30	10	5	2	26	27	24	23	3	3	
12	other vegetables	11	17	7	25	10	16	13	8	11	10	19	
13	condensed milk	30	18	5	21	14	4	7	19	20	9	2	
14	long life bakery product	22	2	15	17	27	13	6	3	3	15	10	
15	butter	10	13	8	19	26	5	13	7	4	27	6	
16	rice	11	30	3	6	18	8	11	22	4	25	4	
17	abrasive cleaner	10	23	24	15	13	12	15	13	4	4	8	
18	rolls/buns	10	29	2	9	23	4	5	16	23	19	12	
19	UHT-milk	16	9	28	25	12	4	26	15	13	28	14	
20	bottled beer	27	10	19	6	24	19	13	15	27	22	8	
21	liquor (appetizer)	22	10	9	21	5	14	19	16	26	26	19	
22	pot plants	16	11	13	14	7	19	13	15	26	15	21	
23	cereals	14	21	6	10	10	10	15	8	18	12	7	
24	white bread	1	20	6	14	15	24	24	1	6	23	1	
25	bottled water	9	20	19	29	27	13	23	10	15	2	17	
26	chocolate	28	8	10	7	14	30	27	20	5	2	16	
27	curd	23	17	15	15	15	4	12	26	8	17	5	
28	flour	12	6	4	3	28	15	17	12	11	17	12	
29	dishes	15	15	25	2	8	2	17	14	23	9	21	
30	beef	28	25	23	21	22	3	19	20	24	5	11	

Figure 4. Details of Items Sold

A	B	C	D	E
	Customer Details			
CustID	**First**	**Name**	**Email**	**Phone**
1	JOHN	SMITH	John.Smith@gmail.com	7149982345
2	ROBERT	JOHNSON	Rob.J@gmail.com	4801238879
3	MICHAEL	WILLIAMS	M.L@gmail.com	5023409989
4	WILLIAM	JONES	Will.Jones@gmail.com	
5	DAVID	BROWN	D.Brown@gmail.com	7142321119
6	RICHARD	DAVIS	Richard@gmail.com	4808732198
7	JOSEPH	MILLER	Miller.J@gmail.com	

Figure 5. Customer Details

The items names displayed in Figure 4 have been obtained by selecting unique names from a set of more than 9000 transactions collected anonymously by Salem Marafi [28]. The quantities purchased have been generated randomly and the customer names in Figure 5 have been picked up from a list of most common American names released by the Social Security Administration. The contact details used are fictitious.

2. Delete Customer and Edit Customer Details

   Selecting these options will lead to the same MS Excel sheet as Figure 5 and the vendor can get their contact information and keep adding the customer details to be kept at one place for easy access. The vendor can manage all the details of the customer from this excel sheet and send out alerts for any offers or deals.

3. Add and Delete Item

   Selecting these options will lead to the same MS Excel sheet as Figure 4. Any item that is added to the inventory or removed permanently can be updated in the records from this window. It gives the vendor an easy track of his stocked items.

4. Update Transactions

   Figure 6 shows the window that pops up when this button is clicked. This feature is meant to be used to keep record of each and every transaction made, involving registered or unregistered customers. This data is recorded to perform further analysis and derivation of rules for market basket analysis.

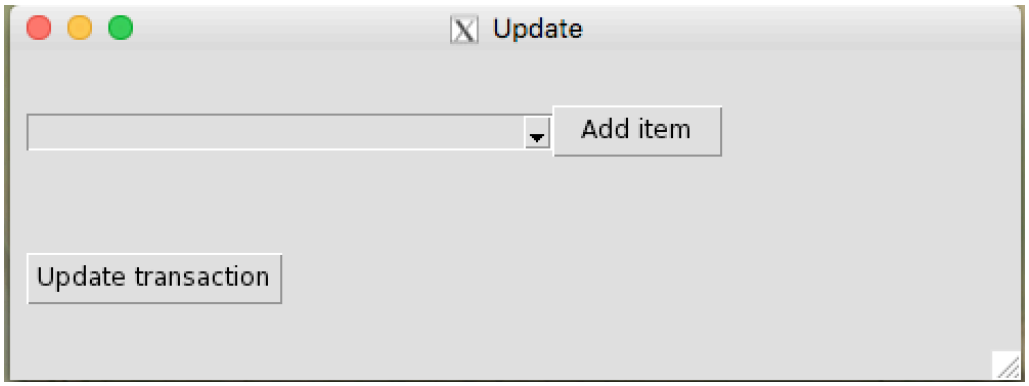


Figure 6. Update Transaction window

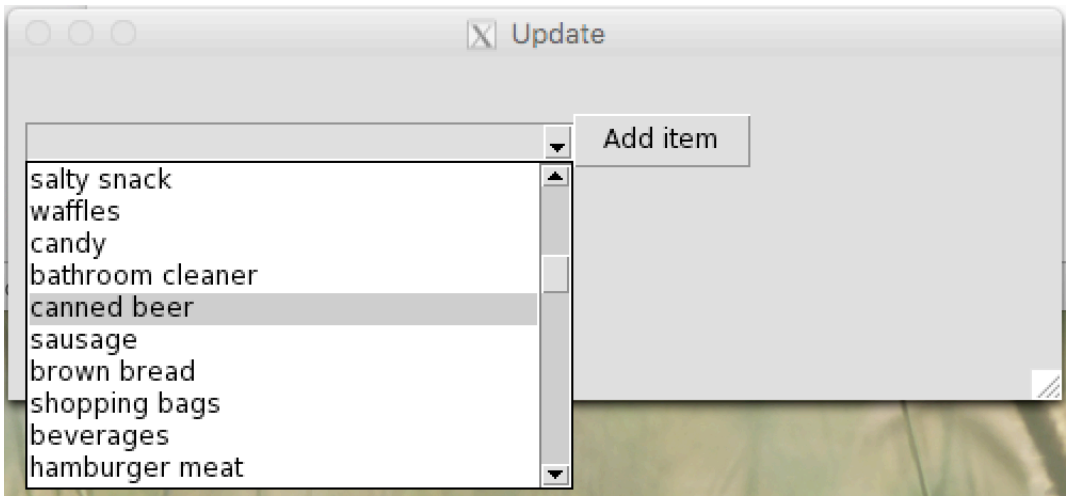


Figure 7. Update Transaction window drop down list

Figure 7 shows the list of items available with the vendor, which is populated from the MS Excel sheet from Figure 4. Every item from the transaction can be selected and 'Add Item' clicked to save the item. Clicking 'Update Transaction' will commit and write the transaction to the CSV file containing all the transactions, as shown in Figure 8.

9828	citrus fruit	herbs	other vegeta	dessert	sugar	shopping bags						
9829	frankfurter	tropical fruit	other vegeta	whole milk	frozen meals	rolls/buns	detergent	napkins	newspapers			
9830	sausage	butter	rolls/buns	pickled veget	soda	fruit/vegetat	waffles					
9831	tropical fruit	other vegeta	domestic egg	zwieback	ketchup	soda	dishes					
9832	sausage	chicken	beef	hamburger n	citrus fruit	grapes	root vegetab	whole milk	butter	whipped/sou	flour	
9833	cooking chocolate											
9834	chicken	citrus fruit	other vegeta	butter	yogurt	frozen desse	domestic egg	rolls/buns	rum	cling film/bags		
9835	semi-finishec	bottled wate	soda	bottled beer								
9836	chicken	tropical fruit	other vegeta	vinegar	shopping bags							
9837	tropical fruit	yogurt										
9838	yogurt											
9839												

Figure 8. CSV file containing all the transaction history

4.3 Module ‘Sales Analysis’

Figure 9 shows the window that opens up for Sales Analysis. There are only two buttons that are included in this window. The reasons for this are simplicity and abstraction for the user.

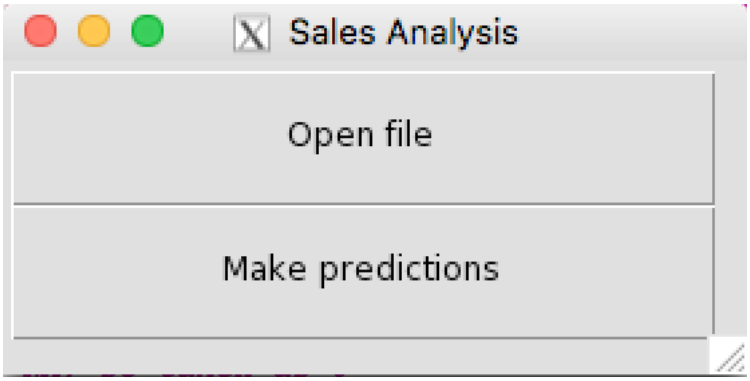


Figure 9. Sales Analysis window

The file that opens up is shown in Figure 10. This is one sample file for a deal to which we have actual responses from the customers. The values indicate lowered prices. As discussed in Chapter 2, One Class Classification using SVM was implemented.

	A	B	C	D	E	F
1	**Deals/Offers Responses**					
2						
3	CUSTID	WHOLE MILK	CEREAL	BUTTER	ROLLS/BUNS	Result
4	49	5.1	3.5	1.4	0.2	yes
5	144	4.9	3	1.4	0.2	no
6	39	4.7	3.2	1.3	0.2	yes
7	77	4.6	3.1	1.5	0.2	no
8	188	5	3.6	1.4	0.2	
9	128	5.4	3.9	1.7	0.4	yes
10	156	4.6	3.4	1.4	0.3	yes
11	39	5	3.4	1.5	0.2	yes
12	197	4.4	2.9	1.4	0.2	no
13	151	4.9	3.1	1.5	0.1	
14	172	5.4	3.7	1.5	0.2	no
15	36	4.8	3.4	1.6	0.2	yes
16	85	4.8	3	1.4	0.1	yes
17	136	4.3	3	1.1	0.1	no
18	10	5.8	4	1.2	0.2	yes
19	16	5.7	4.4	1.5	0.4	no
20	167	5.4	3.9	1.3	0.4	yes
21	64	5.1	3.5	1.4	0.3	yes
22	118	5.7	3.8	1.7	0.3	
23	49	5.1	3.8	1.5	0.3	no
24	188	5.4	3.4	1.7	0.2	no
25	24	5.1	3.7	1.5	0.4	
26	80	4.6	3.6	1	0.2	no
27	16	5.1	3.3	1.7	0.5	yes
28	74	4.8	3.4	1.9	0.2	yes
29	49	5	3	1.6	0.2	yes
30	123	5	3.4	1.6	0.4	yes
31	98	5.2	3.5	1.5	0.2	yes
32	47	5.2	3.4	1.4	0.2	no

Figure 10. Excel sheet for recording deals' responses

A model was trained and tested using these 150 observations for which the statistics are given in Figure 11. As expected with less number of observations, the accuracy of the model is approximately 43%, which is said it grow with time as the number of observations grow. The vendor can test other data with this model using the 'Make Predictions' button and all the positive fields will be marked 'TRUE' indicating these customers are likely to respond to a deal.

```
Confusion Matrix and Statistics

          Reference
Predicted FALSE TRUE
    FALSE    26   11
    TRUE     47   18

               Accuracy : 0.4314
                 95% CI : (0.3337, 0.5332)
    No Information Rate : 0.7157
    P-Value [Acc > NIR] : 1

                  Kappa : -0.0168
 Mcnemar's Test P-Value : 4.312e-06

            Sensitivity : 0.6207
            Specificity : 0.3562
         Pos Pred Value : 0.2769
         Neg Pred Value : 0.7027
             Prevalence : 0.2843
         Detection Rate : 0.1765
   Detection Prevalence : 0.6373
      Balanced Accuracy : 0.4884

       'Positive' Class : TRUE

>
> print(confTrain)
          Reference
Predicted TRUE
    FALSE    7
    TRUE    41
> print(confTest)
          Reference
Predicted FALSE TRUE
    FALSE    26   11
    TRUE     47   18
```

Figure 11. Statistics for One Class Classification using SVM

4.4 Module ‘Market Basket Analysis’

Figure 12 shows the window that opens for Market Basket Analysis. Market Basket Analysis, also known as Association Analysis or Frequent Itemset Mining, is a modelling technique based upon the theory that if you buy a certain group of items, you are more (or less) likely to buy another group of items. The set of items a customer buys is referred to as an itemset, and market basket analysis seeks to find relationships between purchases. Typically, the relationship is in the form of a rule, called association rule. Association rules are made up of antecedents and consequents and take the following form: $A \rightarrow B$.

Support count is the count of how often a given itemset appears across all the transactions. Frequency of its appearance is given by a metric called support. The most basic measure of rules is confidence, which tells us how often a given rule applies within the transactions that contain the ante. A given rule applies when all items from both antecedents and consequents are present in a transaction, so it is the same thing as an itemset that contains the same items. We can use the support metric for the itemset to compute confidence of the rule.

In a brute force method, you would calculate the support and confidence of all possible itemset combinations, but that would be computationally expensive, because the number of candidates grows exponentially. The Apriori algorithm, as used here, addresses this issue by generating candidates selectively. To get an intuition, think about the frequency of an itemset that contains some infrequent items. That itemset will never be more

frequent than the least frequent item it contains. So if you construct your candidates by combining the frequent itemsets only, starting from 1-itemset and continue to higher levels, then you avoid creating useless candidates.

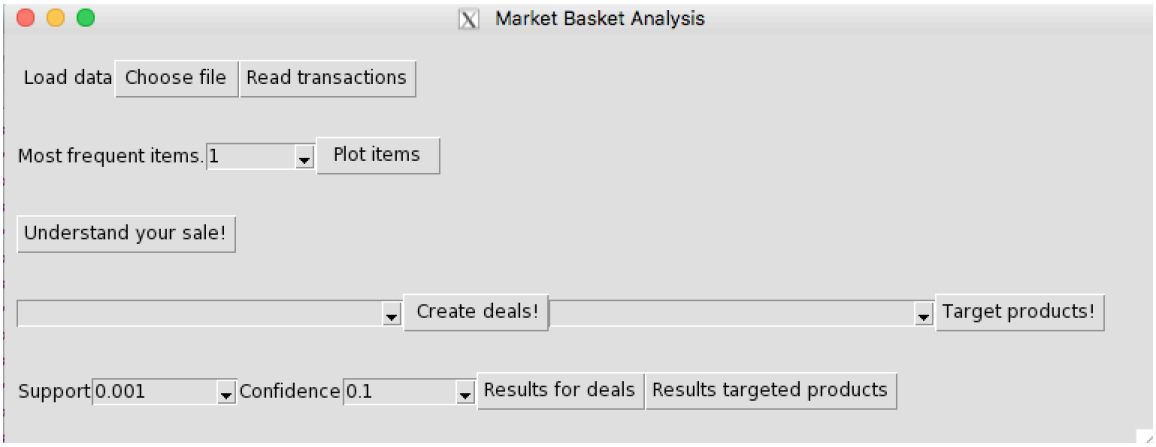


Figure 12. Market Basket Analysis window

The first two buttons of Choose File and Read Transactions are done to read in the transaction history which is created using 'Update Transactions' in Module 'Manage Data'. Figure 13 shows the dialog box for choosing a CSV file and uploading it in the tool through 'Read Transactions'. The data is public and was collected from [28] for this report. It has over 9000 transactions, containing almost 170 unique items.

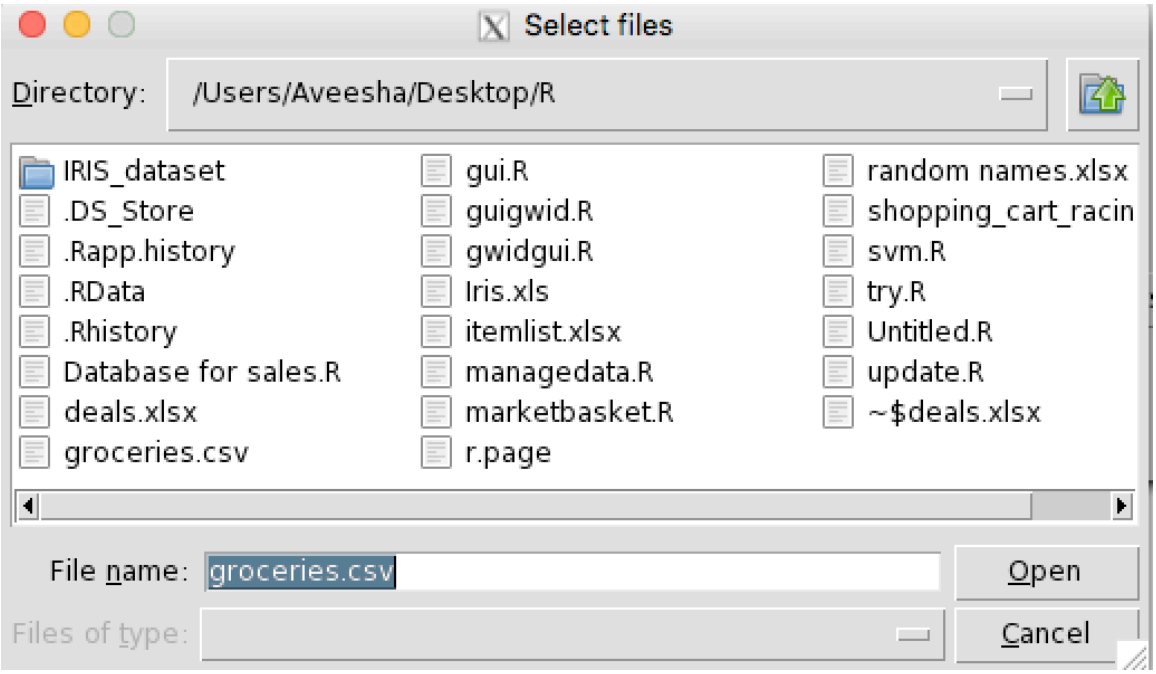


Figure 13. Open File Dialog Box

The vendor can plot top 50 most frequently sold items, or high selling items to have a better understanding of how to create deals and promote lesser selling products. He can select how many top items he wants to view, through a drop down list and their quantities sold so far. Figure 14 shows the plot of the results that is displayed to the user.

The button ‘Understand your sale’ gives details to the user in simple and concise language on how to use and select ‘support’ and ‘confidence’ required for further analysis. It opens a new window with a simple text explaining the terms and their meaning with examples.

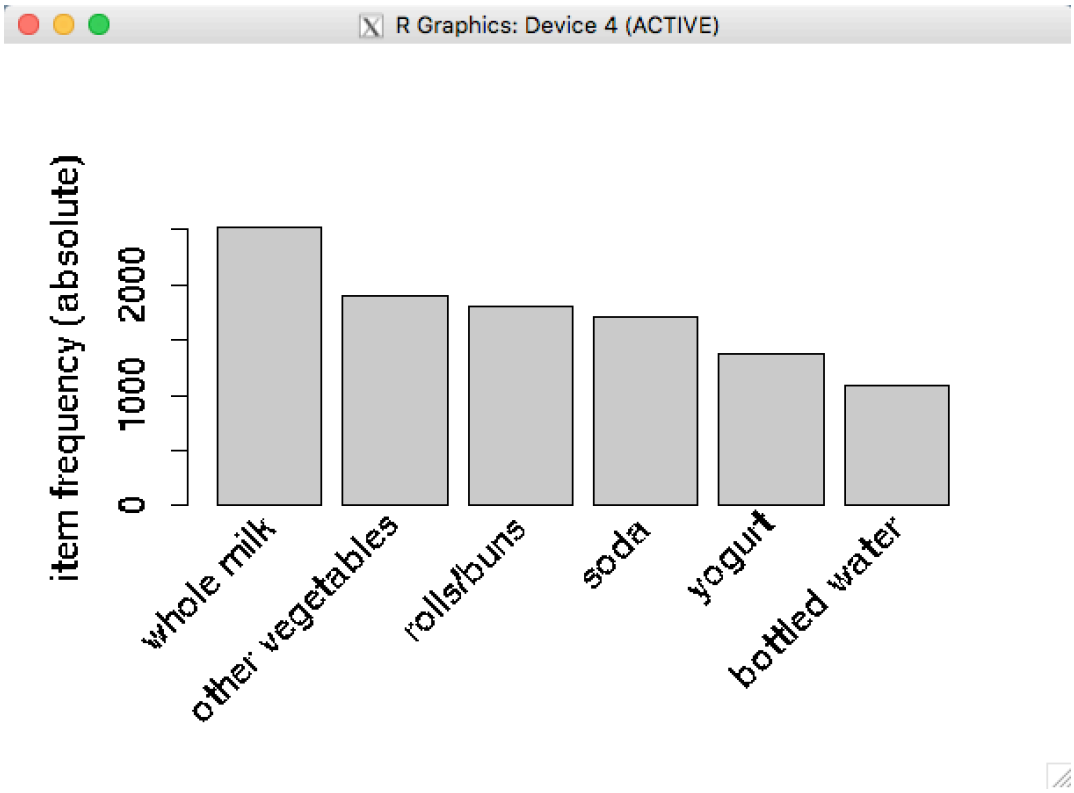


Figure 14. Plot of the highest selling items

The ‘Create Deals’ button is created to add multiple items from the drop down list, choose a value for support and confidence and see if any pattern emerges. This feature was included to keep the vendor informed of the patterns that occur in customer purchases and to benefit some products from promotion. From a technical point, we know what we want on the left side of the association rule. Figure 15 shows the interactive graph that is plotted if ‘rolls/buns’ and ‘pastry’ are included in the itemset with support 0.001 confidence chosen as at least 0.25. The bigger the yellow circle, higher the confidence. The highlighted arrow shows that if ‘rolls/buns’ and ‘pastry’ are sold along, one might also be likely to buy ‘other vegetables’. Other deals also indicate

the likelihood of ‘soda’ and ‘yogurt’ being sold.

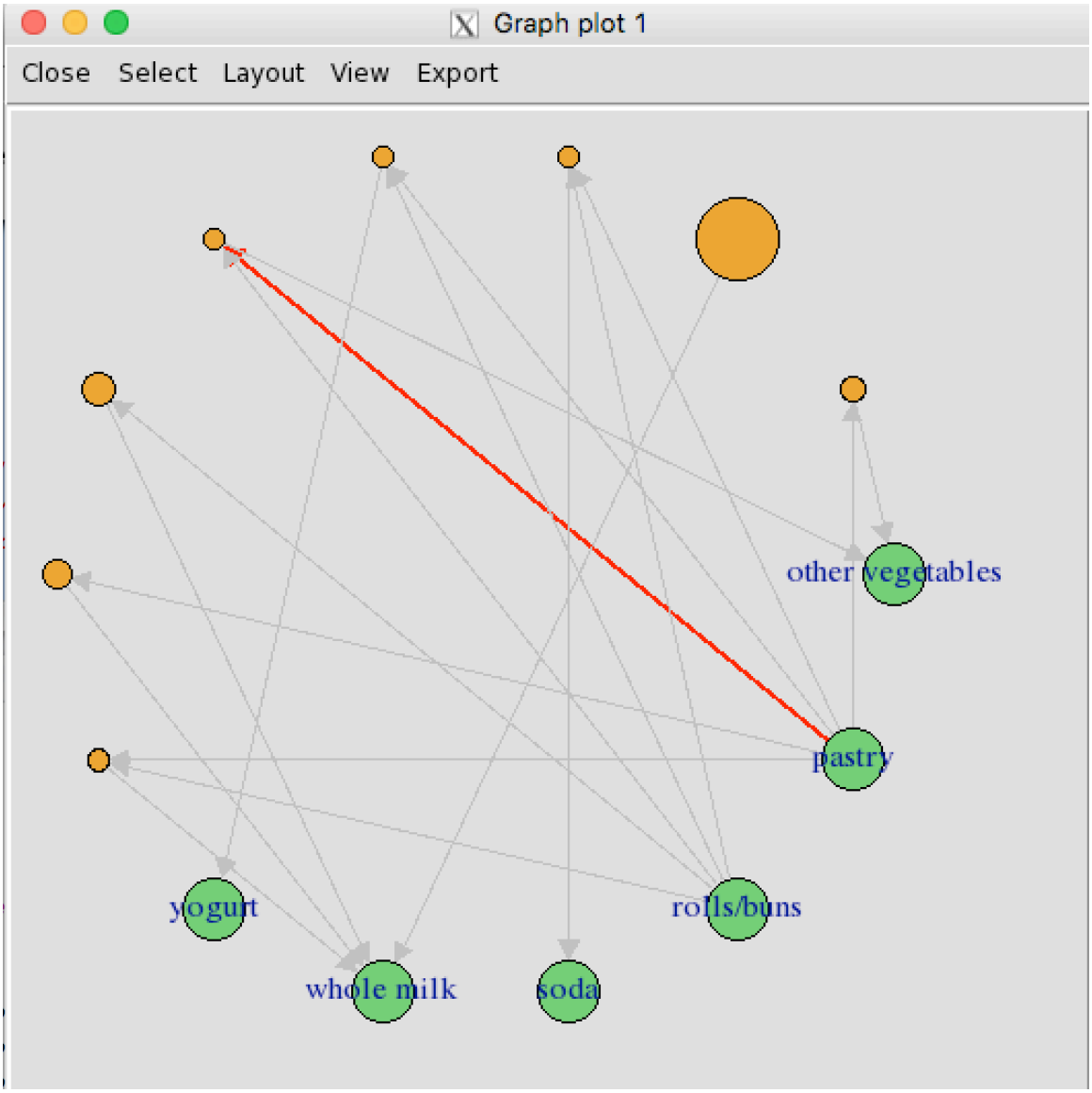


Figure 15. Interactive graph to create a promotional deal

Other way to improve business is to take advantage of most frequently sold items. High selling products can be paired with lesser selling products and tested with the ‘Sales Analysis’ module for tentative results. Or for general layout of the store, products can be placed besides higher selling products.

Thus, the lesser selling products get benefited from the practice of this feature. From a technical point, we know what we want on the right side of the association rule, therefore we are 'targeting' products. In Figure 16, from the highlighted arrow we can see, there's a strong likelihood that 'grapes', 'fruit/vegetable juice' and 'other vegetables' give way to tropical fruit, which happens to be the 8th most frequently sold item. Tropical fruit was the target with 0.75 confidence and 0.001 support.

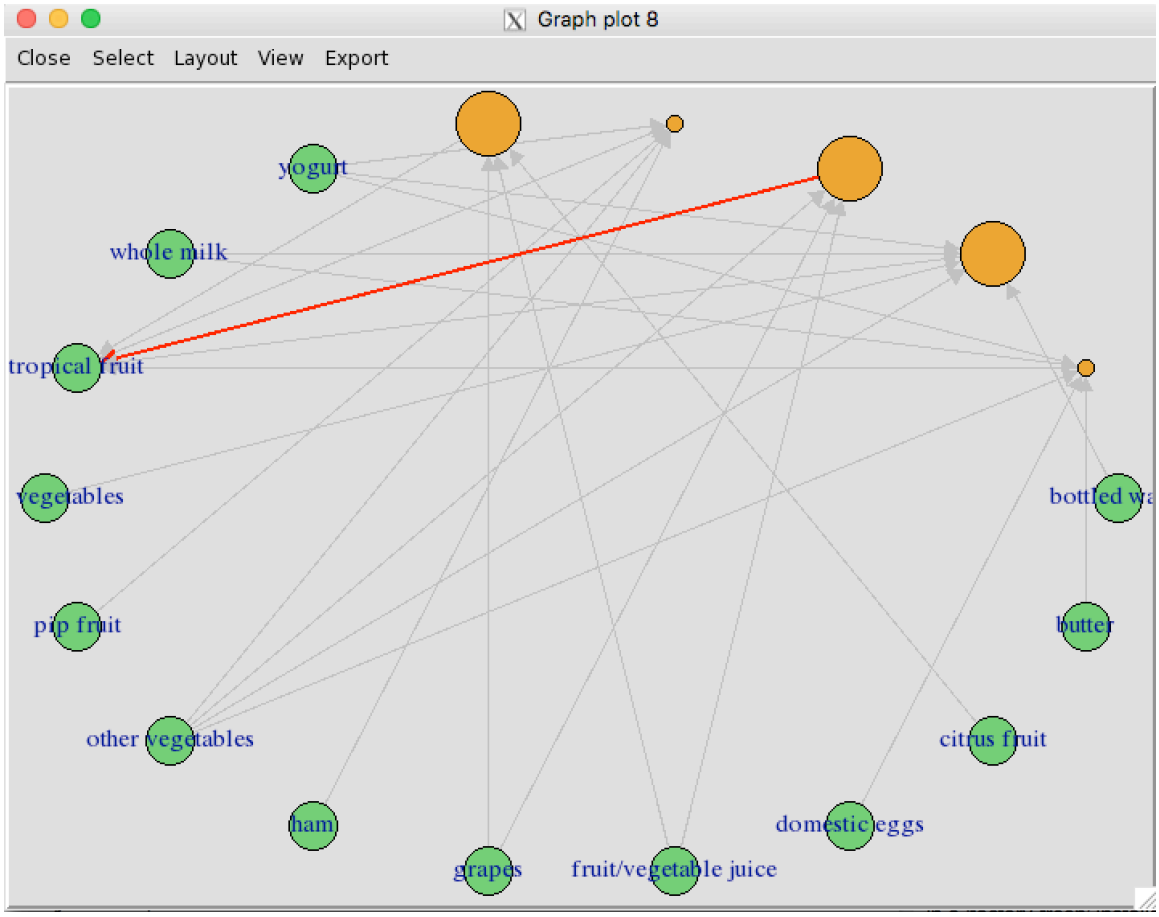


Figure 16. Interactive graph to target highest selling products

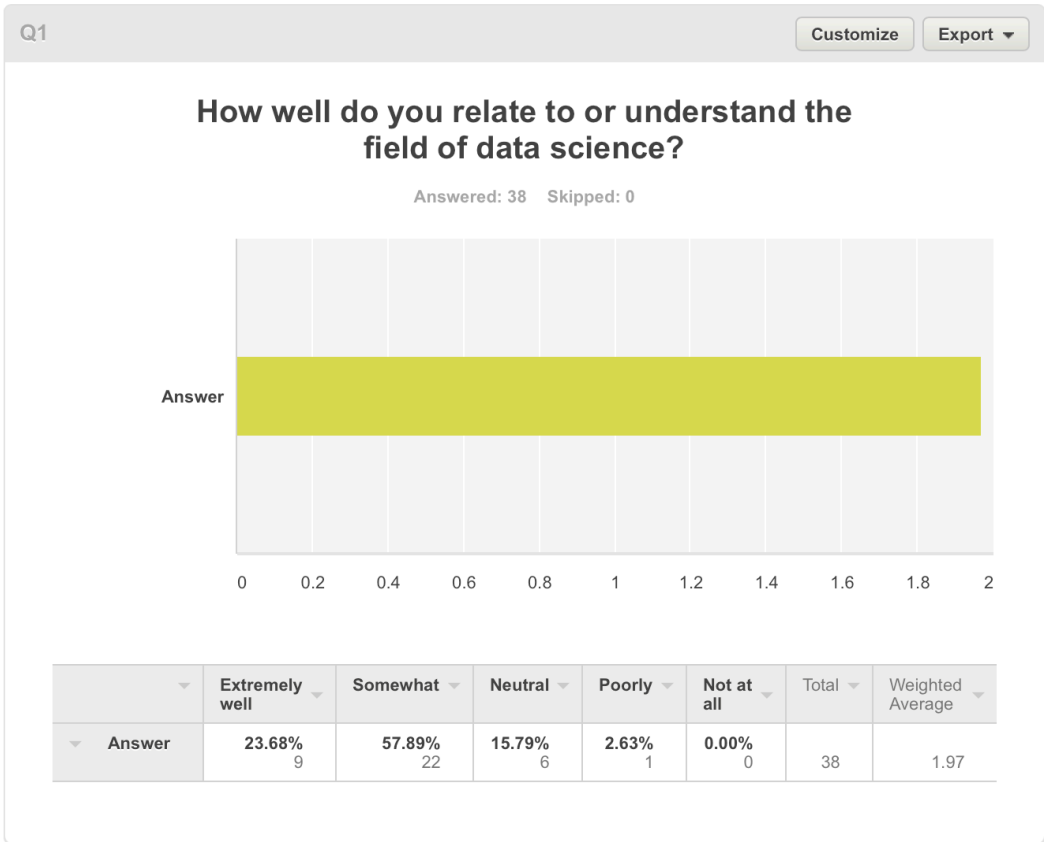
In case no rules exist for a certain product with the chosen support or confidence, an alert window is popped, as shown in Figure 17.

Figure 17. Alert window for no existing rules

CHAPTER 5

USER STUDY

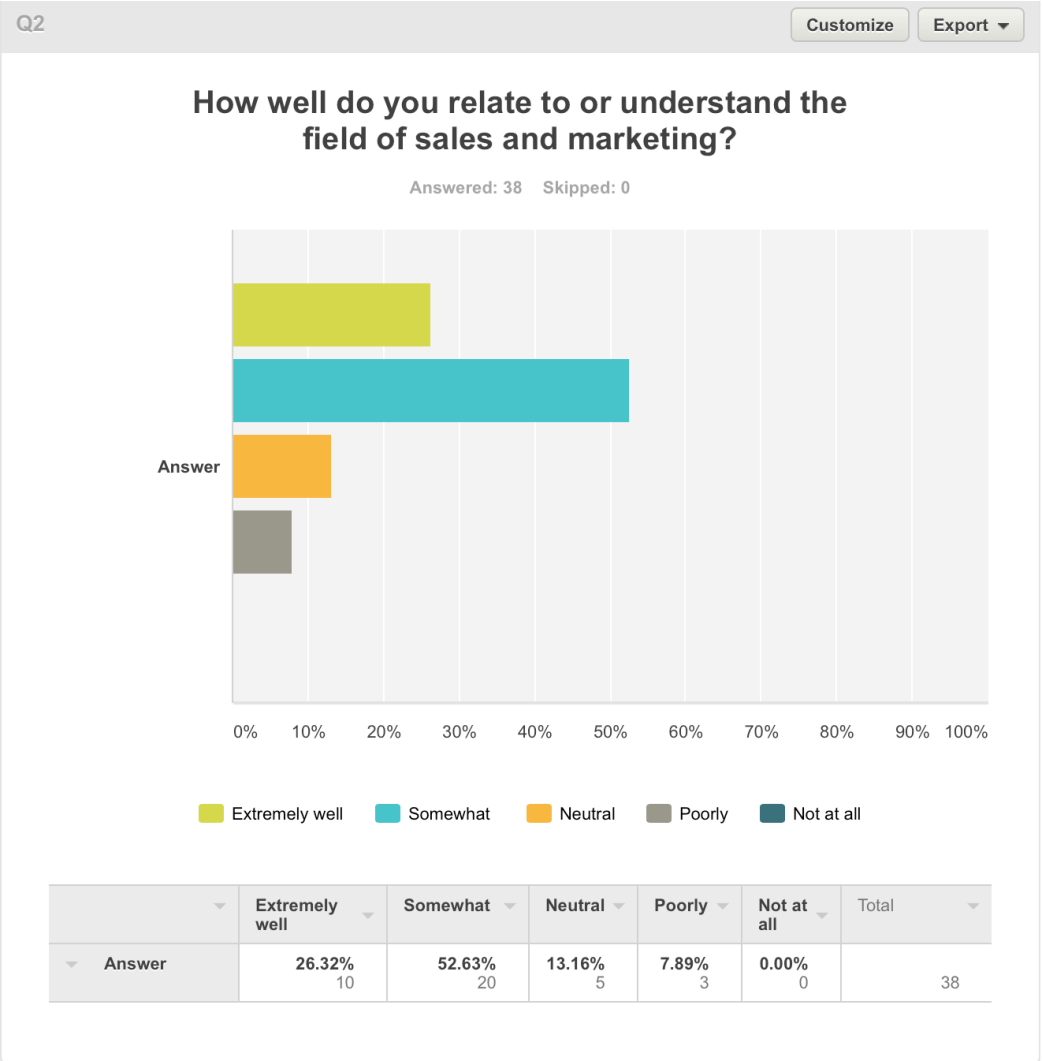
A study was conducted to validate the utility and feasibility of the tool for its specific audience. The study was done on 38 participants who were given a short verbal introduction to the tool and then were made to use the tool. It was followed by a survey and the participants answered a series of questions which reflect the feedback on the tool. The questions and the responses are given as follows.



	Extremely well	Somewhat	Neutral	Poorly	Not at all	Total	Weighted Average
Answer	23.68% 9	57.89% 22	15.79% 6	2.63% 1	0.00% 0	38	1.97

Figure 18. Question 1 of Survey

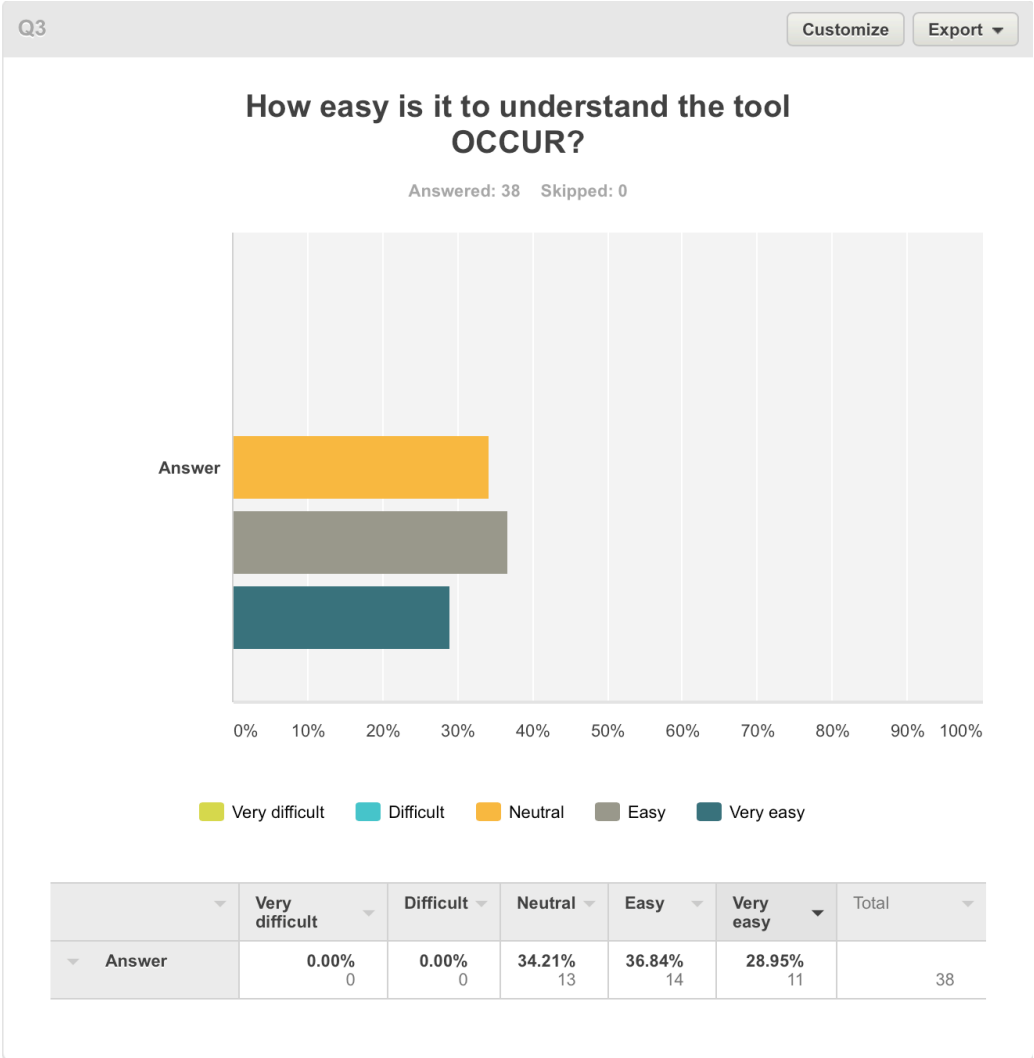
Participants were asked if they had any knowledge of data science in general and more than 80% of the participants responded that they had reasonable or good knowledge of the domain.



	Extremely well	Somewhat	Neutral	Poorly	Not at all	Total
Answer	26.32% 10	52.63% 20	13.16% 5	7.89% 3	0.00% 0	38

Figure 19. Question 2 of Survey

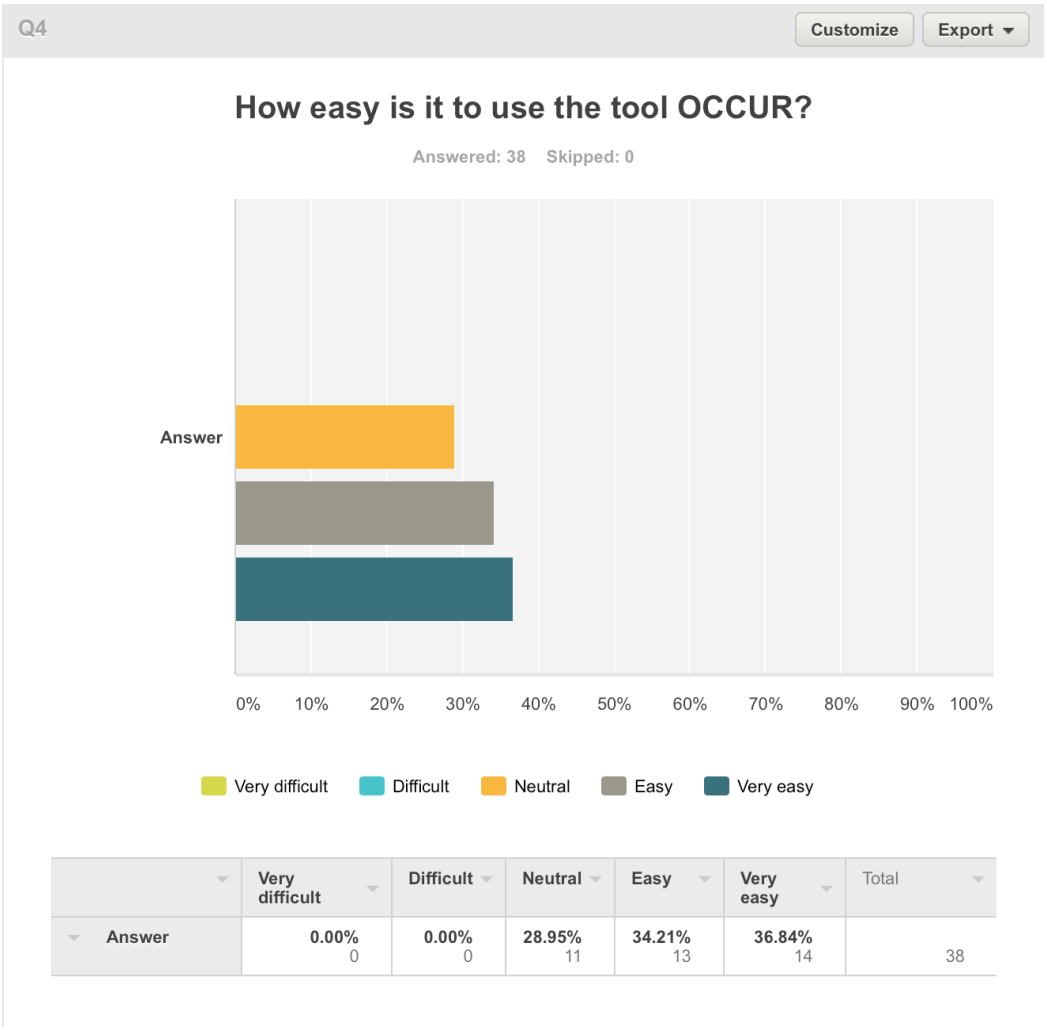
Participants were asked if they had any knowledge of marketing in general and close to 80% of the participants responded that they had reasonable or good knowledge of the domain.



	Very difficult	Difficult	Neutral	Easy	Very easy	Total
Answer	0.00% 0	0.00% 0	34.21% 13	36.84% 14	28.95% 11	38

Figure 20. Question 3 of Survey

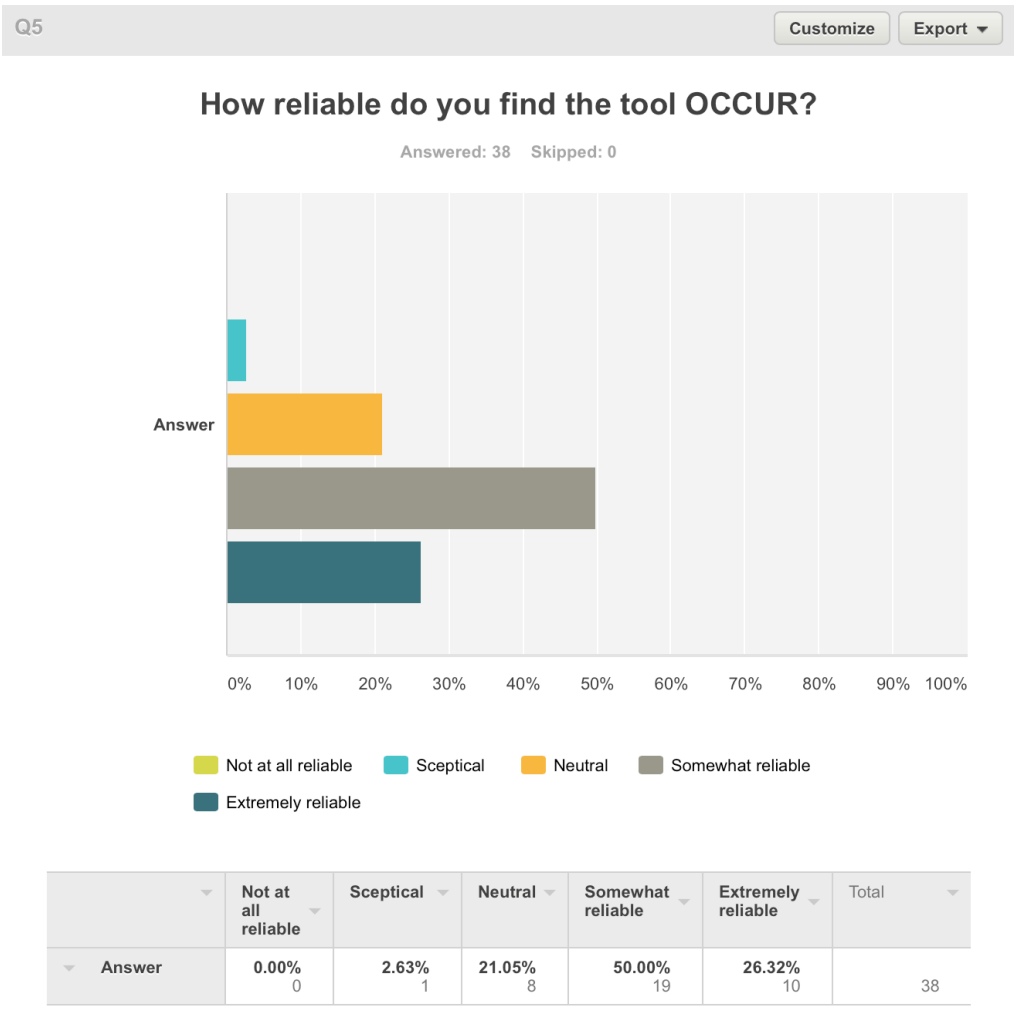
Participants were asked about the ease of understanding the tool OCCUR and most population found it easy or were neutral towards it.



	Very difficult	Difficult	Neutral	Easy	Very easy	Total
Answer	0.00% 0	0.00% 0	28.95% 11	34.21% 13	36.84% 14	38

Figure 21. Question 4 of Survey

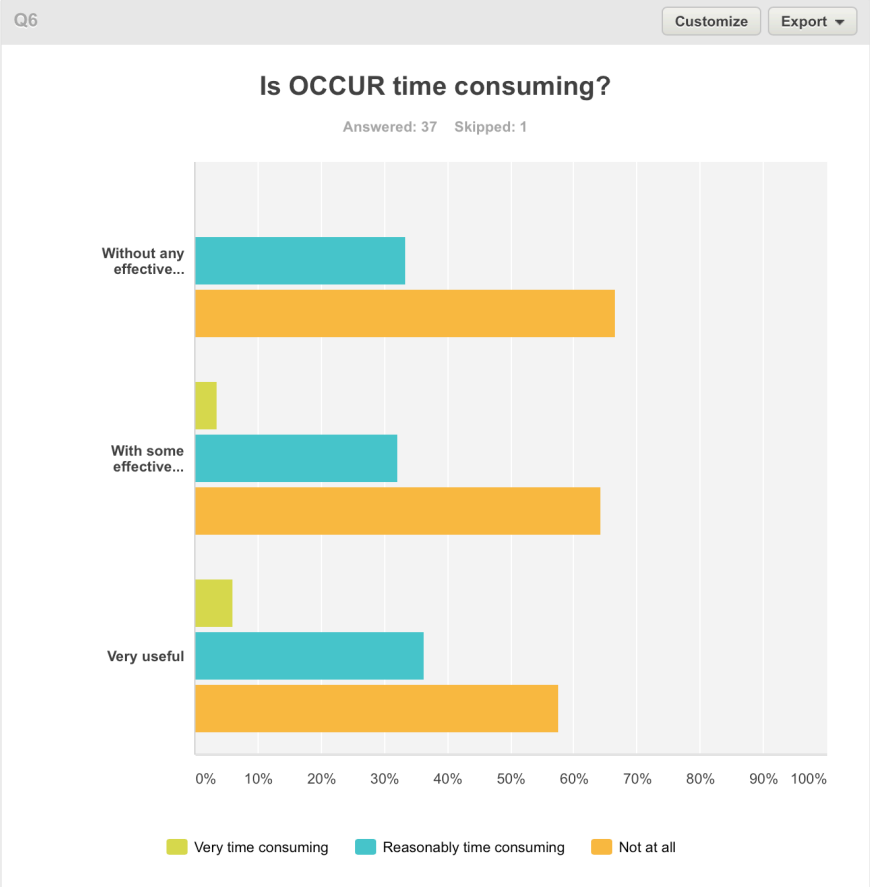
Participants were asked about the ease of using the tool OCCUR and most population found it easy or were neutral towards it.



	Not at all reliable	Sceptical	Neutral	Somewhat reliable	Extremely reliable	Total
Answer	0.00% 0	2.63% 1	21.05% 8	50.00% 19	26.32% 10	38

Figure 22. Question 5 of Survey

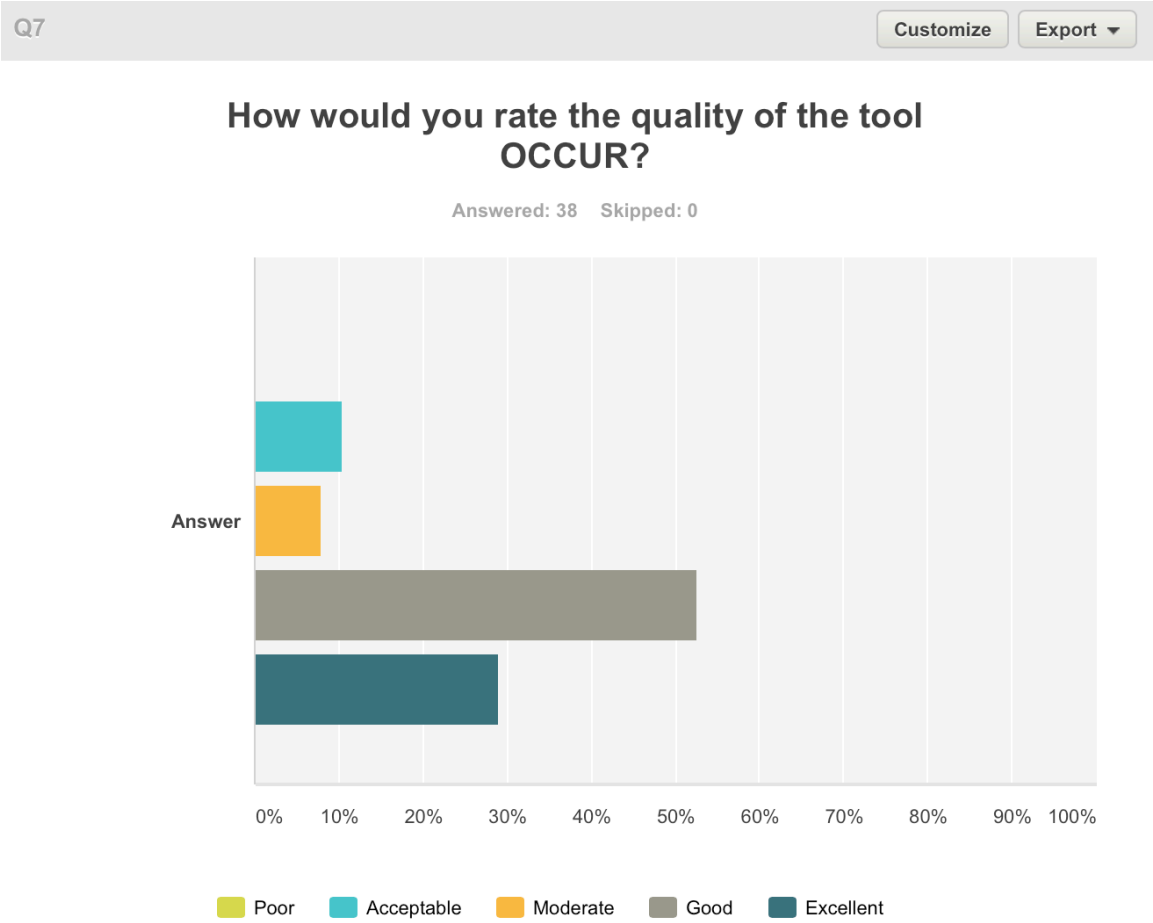
Participants were asked about the how reliable they found the tool OCCUR and most population found it reliable or trustworthy.



	Very time consuming	Reasonably time consuming	Not at all	Total
Without any effective usefulness	0.00% 0	33.33% 9	66.67% 18	27
With some effective usefulness	3.57% 1	32.14% 9	64.29% 18	28
Very useful	6.06% 2	36.36% 12	57.58% 19	33

Figure 23. Question 6 of Survey

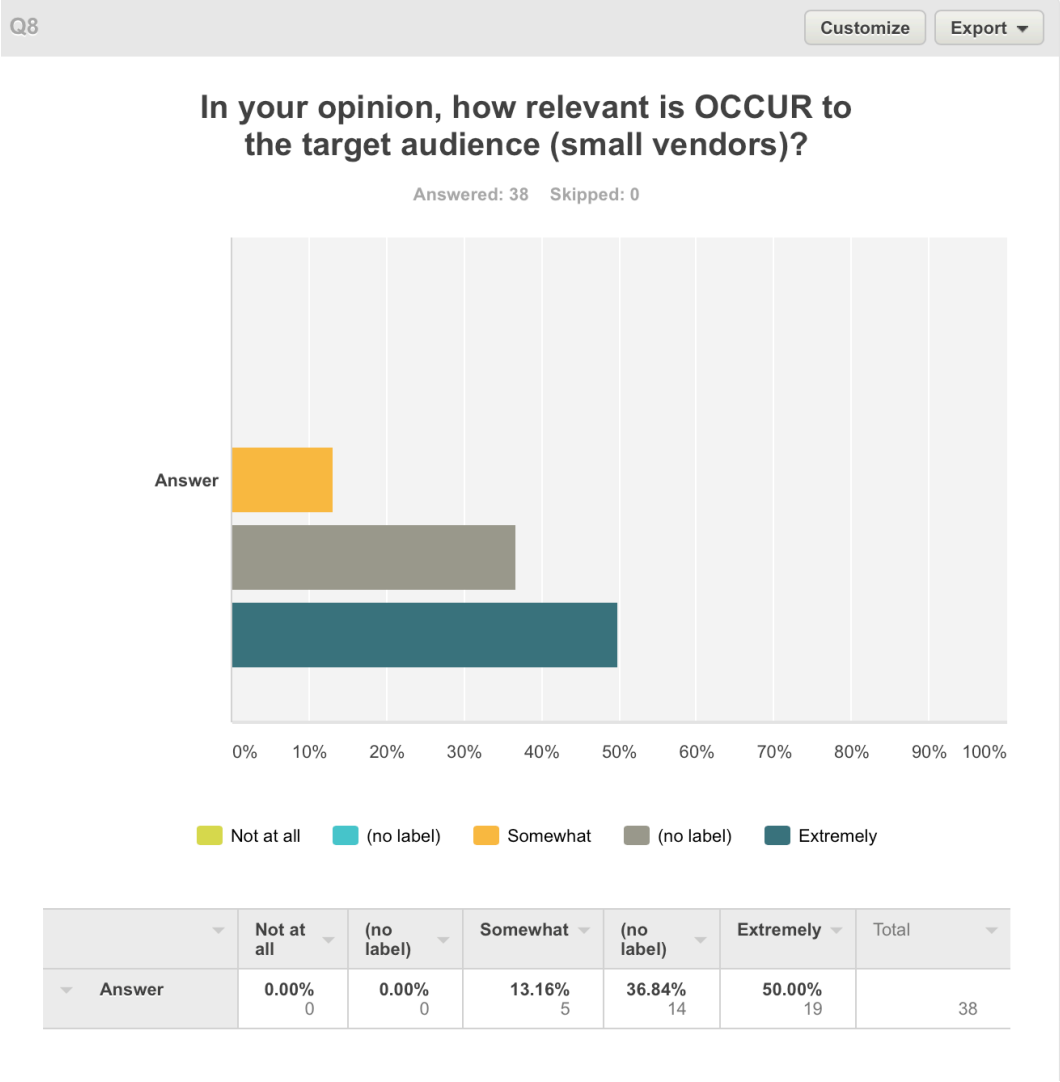
Participants found the tool little or reasonably time consuming with the results of the time being consumed very useful.



	Poor	Acceptable	Moderate	Good	Excellent	Total
Answer	0.00% 0	10.53% 4	7.89% 3	52.63% 20	28.95% 11	38

Figure 24. Question 7 of Survey

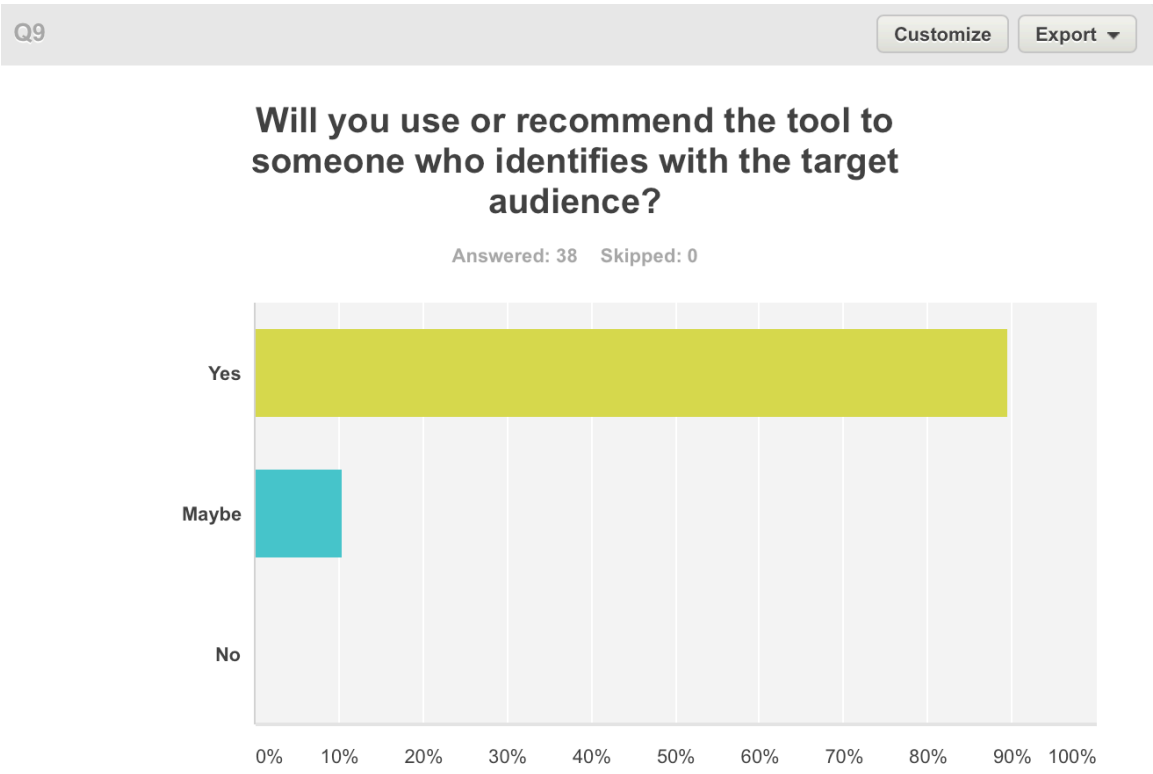
More than 80% of the participants found the tool to be of good quality as against less than 20% who thought the tool was of average quality.



	Not at all	(no label)	Somewhat	(no label)	Extremely	Total
Answer	0.00% 0	0.00% 0	13.16% 5	36.84% 14	50.00% 19	38

Figure 25. Question 8 of Survey

Participants were asked about the relevance of the tool OCCUR and most population found it to be relevant to the audience it is made for, that is the small vendors who have no internet connectivity and use cash as primary method of payment.



Answer Choices	Responses	
Yes	89.47%	34
Maybe	10.53%	4
No	0.00%	0
Total		38

Figure 26. Question 9 of Survey

Participants were asked if they would use or recommend the tool OCCUR and almost 90% of the population responded positively.

CHAPTER 6

CONCLUSION AND FUTURE WORK

The mining tool OCCUR was developed keeping in mind the marketing and management problems faced my small businesses, especially vendors, who do not have the means to invest in online analytics tools and who deal in cash as a primary method of payment. The tool that is developed fulfills the posed problem and gives the user facilities to manage customer and inventory data, thus aiding the management problems and also facilitates better understanding of the sales and hence market their products better, making use of Association Mining or Market Basket Analysis in particular. One Class Classification using SVM addresses the problem where there are no or less number of negative examples to build a classification model.

The tool, as intended for the population for lesser developed countries with no or little internet connectivity, brings the benefit of data mining to small vendors without any investment of money or significant time, and therefore addressing their basic problems effectively. The survey that was conducted yields encouraging results which indicate that the tool was seen as easy to understand and use with a reasonable amount of time invested in the tool. It was also deemed relevant to the intended audience by a population who consider themselves to have reasonable knowledge of the domains it pertains to. However, a more intuitive GUI with smoother controls could improve the quality of the tool greatly, as inferred from the user study and the feedback from the survey.

REFERENCES

[1] Linoff, Gordon S. Berry, Michael J.. (2011). Data Mining Techniques. *Wiley Computer Publishing. Retrieved January 25, 2016, from http://www.myilibrary.com.ezproxy1.lib.asu.edu/?ID=366313

[2] Guided and automated analytics from the cloud. (n.d.). Retrieved January 28, 2016, from http://www.ibm.com/analytics/watson-analytics/

[3] Actions speak louder than page views. (n.d.). Retrieved January 28, 2016, from https://mixpanel.com/

[4] Billions of people in developing world still without Internet access, new UN report finds. (2015, September 21). Retrieved March 18, 2016, from http://www.un.org/apps/news/story.asp?NewsID=51924#.VuydAMdnRcY

[5] Jabban, T. (n.d.). Retrieved January 28, 2016, from http://www.contrib.andrew.cmu.edu/~tjabban/datamining.html

[6] Han, J., and Kamber, M. (2006). Data Mining: Concepts and Techniques (2nd Edition). Burlington, MA, USA: Elsevier Science & Technology. Retrieved from http://www.ebrary.com

[7] Kantardzic, M., and Wiley, Inter Science (Online service). (2011). Data mining: Concepts, models, methods, and algorithms (2nd ed.). Hoboken, NJ: Wiley-IEEE Press.

[8] Waxer, C. (2013, October 28). How data mining can boost your revenue by 300%. Retrieved January 28, 2016, from http://money.cnn.com/2013/10/28/smallbusiness/data-mining/

[9] 10 Ways Data Mining Can Help You Get a Competitive Edge. (n.d.). Retrieved January 30, 2016, from https://blog.kissmetrics.com/data-mining/

[10] Huang, X., and Brown, A. (1999). An analysis and classification of problems in small business. International Small Business Journal, 18(1), 73-85. doi:10.1177/0266242699181004

[11] McKenna, P. (1991), Marketing is Everything. Harvard Business Review, (January – February), pp65-79.

[12] Romano, C. and Ratnatunga, J. (1995), The Role of Marketing: Its Impact on Small Enterprise Research, European Journal of Marketing, 29(7), pp9-30.

[13] Khan, S. S., and Madden, M. G. (2010). A survey of recent trends in one class classification. (pp. 188-197). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-17080-5_21

[14] Classification with data for only one class • /r/statistics. (2015, April). Retrieved January 31, 2016, from https://www.reddit.com/r/statistics/comments/32lajr/classification_with_data_for_only_one_class/

[15] Stackoverflow, Unary class text classification in weka (2012, May 1). Retrieved January 31, 2016, from http://stackoverflow.com/questions/10394615/unary-class-text-classification-in-weka

[16] Tax, D. (2001), One Class Classification. Ph.D. thesis, Delft University of Technology, Netherlands.

[17] Tax, D., Duin, R. (1999). Data domain description using support vectors. In: Proc. ESAN 1999, Brussels, pp. 251–256.

[18] Tax, D., Duin, R. (1999). Support vector domain description. Pattern Recognition Letters 20, 1191–1199.

[19] Scholkopf, B., Williamson, R., Smola, A., Taylor, J., Platt, J. (2000). Support vector method for novelty detection. In: Solla, S.A., Leen, T., Muller, K. (eds.) Neural Information Processing Systems, pp. 582–588.

[20] Scholkopf, B., Williamson, R., Smola, A., Taylor, J. (1999). Sv estimation of a distributions support. In Advances in Neural Information Processing Systems.

[21] Zhao, Y., Li, B., Li, X., Liu, W., & Ren, S. (2005). Customer churn prediction using improved one-class support vector machine. In Advanced data mining and applications (pp. 300-306). Springer Berlin Heidelberg.

[22] Bo Pang and Lillian Lee (2008), "Opinion Mining and Sentiment Analysis", Foundations and Trends® in Information Retrieval: Vol. 2: No. 1–2, pp 1-135. http://dx.doi.org/10.1561/1500000011

[23] Manevitz, L. M., & Yousef, M. (2002). One-class SVMs for document classification. the Journal of machine

[24] Humphrey, David B. and Pulley, Lawrence B. and Vesala, Jukka M. (Nov 1996). Cash, Paper and Electronic Payments: A Cross-Country Analysis. In Journal of Money, Credit and Banking, Volume 28, Issue 4, Part 2: Payment Systems Research and Public Policy Risk, Efficiency, and Innovation.

[25] Marcus, J. (2014, May 06). Cash Beats Debit, Credit, Checks and Mobile as Payment Choice. Retrieved February 11, 2016, from http://thefinancialbrand.com/39408/consumer-cash-usage-banking-payment-research/

[26] Getting bored of being a vendor; what to do? (2015, December). Retrieved February 11, 2016 from https://www.reddit.com/r/electronic_cigarette/comments/3wl35q/getting_bored_of_being_a_vendor_what_to_do/

[27] For small business owners: If you are not investing in marketing, can you explain why? Bad experiences? Lack of faith? I'd like to know your thoughts. - Clarity Answers. (2013, February). Retrieved February 11, 2016, from https://clarity.fm/questions/370/for-small-business-owners-if-you-are-not-investing-in-marketing-can-you-explain

[28] Salem : Salem Marafi. (n.d.). Retrieved March 22, 2016, from http://www.salemmarafi.com/author/salem/