CUSTOMER BEHAVIOR ANALYSIS FOR A RETAIL OUTLET

Anand Niranjan
College of Computing & Informatics
University of North Carolina at Charlotte
Charlotte, NC USA
aniranja@uncc.edu

***Abstract*—Customer behavior analysis is one of the advanced analytics technique i.e. Predictive Analytics to interpret complex behavior based on a specific conceptual framework. It is a data driven marketing method to directly reach out to people and potential market. The process consists of predicting the future behavior of a customer for a retail outlet based on historical quotes of data to help the retailer to develop cross-promotional programs, attract new buyers and improve the overall business. Thus, the evaluation of the overall system is based on the accuracy of the predictive model designed and the result of the knowledge discovered from the model.**

Predicting behavioral patterns is a Customer Relationship Management practice that any company uses to manage and analyze customer interactions and data throughout the customer lifecycle for enhancing the business relationship with customers. Machine learning algorithms Support Vector Machines (Linear Kernel) and Apriori are implemented to achieve better accuracy in predictive analysis.

The proposal is to analyze the data of a retail outlet using R programming, a powerful statistical language and to predict the purchase behavior of the customers. The raw data is in unstructured form and it is pre-processed and cleaned for predictive analysis. The cleaned data will be analyzed using regression models and the report is presented in a structured form which will be plotted in graph for better understanding of the retailer's data. The analysis report will be used by the retailer to understand the customer behavior and to promote his products. The analysis will be represented in an online shiny dashboard.

Keywords—predictive analytics, machine learning, support vector machines, shiny, apriori

I. INTRODUCTION

Customer behavior analysis in retail market generally deals with identifying the customers and reasoning their behavior patterns. The analysis not only helps in identifying the buyer's choice but also to learn and understand the customer response towards sales and promotions. The focus is only on the customer behavior and not on the buying preference of any customer. Most of the organizations maintain their historical data for merchandising results and these records are very valuable for analysis purpose. The workflow that defines the customer behavior pattern is based on selected attributes such as Transaction ID, Product Name, Unit Price, and Quantity. These are analyzed using predictive modelling techniques and insights are presented in the form of graphs and statistical measures in R.

Two main modelling machine learning algorithms implemented are Support Vector Machine (SVM) for linear classification and Apriori algorithm for association rule generation. SVM is used to transform the input data and identify the optimal boundary between the possible outputs. Apriori algorithm analyzes the association between each item and determines the frequency of each items that are purchased together.

II. BUSINESS INVOLVEMENT

The main business problem is predicting the customer behavior and deriving conclusion that reflects in real world applications. Predicting behavioral patterns is one of the Customer Relationship Management (CRM) practices that any company uses to manage and analyze customer interactions and data throughout the customer lifecycle and improve business relationship with customers. The field of CRM has huge potential for data mining as large amount of historical data exists and mining approaches are better when compared to human inspection.

III. PROBLEM STATEMENT

A retail outlet wants to understand the purchase behavior of a buyer. This information will help the retailer to understand the buyer's needs and refine the store's design accordingly, develop cross-promotional strategies, or attract new buyers. The analysis might tell a retailer that customers often purchase milk, bread & butter together, so putting all the items on promotion at the same time would not create a significant increase in profit, while a promotion involving just one of the items is likely to drive sales of the other.

IV. RETAIL TRANSACTION DATASET

The dataset contains the transactions of purchases made by customers in a retail outlet in comma separated version with 982 records [3]. The attributes of the original data are Row ID, Order ID, Order Data, Order Priority, Order Quantity, Sales, Discount, Ship Mode, Profit, Unit Price, Shipping Cost, Customer Name, Province, Region, Customer Segment, Product Category, Product sub category, Product Name, Product Container, Product base margin and Ship Date. All the attributes are not used in the analysis, so the data is cleaned and preprocessed to retain only the attributes from which required amount of information can be extracted to perform analysis and solve the problem. Figure 1 shows a preprocessed and cleaned dataset.

Attributes of interest are as follows:

i. Order ID

Numerical unique value for all transactions by any customer

ii. Customer Name:

First name & last name of customer.

iii. Order Date

Item purchased date (Range is between 1/5/2009 and 12/29/2012)

iv. Order Priority

The priority can be Low, Medium, High or Critical and it's not specified for some orders

v. Product Name

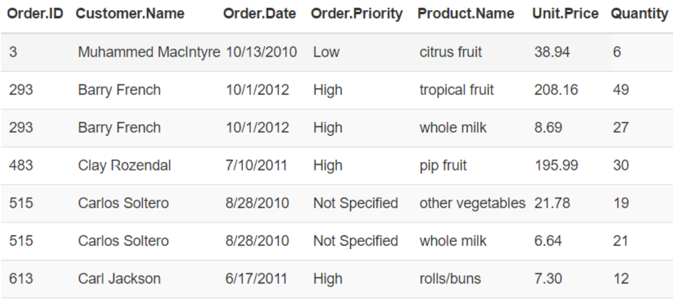
Name of item purchased.

vi. Unit Price

Cost of item per unit.

vii. Quantity

Total number of items purchased.



Order.ID	Customer.Name	Order.Date	Order.Priority	Product.Name	Unit.Price	Quantity
3	Muhammed MacIntyre	10/13/2010	Low	citrus fruit	38.94	6
293	Barry French	10/1/2012	High	tropical fruit	208.16	49
293	Barry French	10/1/2012	High	whole milk	8.69	27
483	Clay Rozendal	7/10/2011	High	pip fruit	195.99	30
515	Carlos Soltero	8/28/2010	Not Specified	other vegetables	21.78	19
515	Carlos Soltero	8/28/2010	Not Specified	whole milk	6.64	21
613	Carl Jackson	6/17/2011	High	rolls/buns	7.30	12

Figure 1: Sample Dataset

V. DATA PREPERATION PROCESS

A. *Data Selection*

The data is collected by selecting subsets of all the available resources that would address the problem and finding the extent of data availability, data that is not available that may be included for better results and what parts of data can be excluded.

B. *Preprocess Data*

- The data is now brought to a form to start working on it. Data is formatted by sub-setting only the attributes for initial analysis and cleaned to fix the missing data in the dataset.
- Duplicates are removed to fit the data in Apriori algorithm.
- Sampling of the resultant data is done by taking a smaller representative sample to explore the problem.

C. *Feature Engineering*

- This is the process of extracting features and transforming them into a format that the machine learning algorithm supports. So, the preprocessed data is transformed to fit into Support Vector Machine to build a predictive classifier [4].
- Scaling is done for Order Priority attribute as it has mixed scales. It is transformed to binary value of zeros (Priority: LOW) and ones (Priority: HIGH) as SVM classifier only accepts binary values for predictions [1].
- Aggregation and decomposition of Order Date is done to extract the year of purchase as it is only required in the analysis. Total Cost is calculated from Unit Price & Quantity.

D. *Training and Testing*

- The resultant data after applying feature engineering is divided into training & testing data sets, 80% is used for training and 20% for testing the accuracy.

- SVM builds a classifier by learning from training data and finds out the labels for input features in testing data [1], [5].
- The model gives a table of confusion matrix that describes the performance of support vector classifier and contains details of actual and predicted values.
- Performance of classifier is calculated using the data in the matrix.
- Statistical measure of performance like accuracy of model, precision, recall, sensitivity, specificity, f-1 score and so on are calculated.

VI. STRATEGIES

Machine Learning algorithms are implemented to achieve better accuracy in the analysis of the given data set. These algorithms have a fixed strategy approach to a specific problem. The two algorithms used are:

i. Apriori Algorithm
ii. Support Vector Machine

A. *Apriori Algorithm*

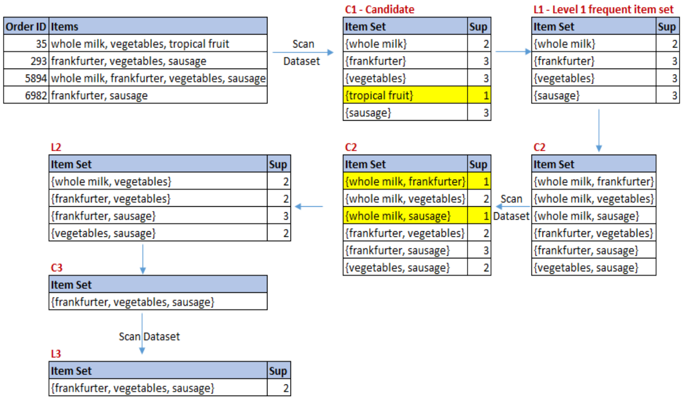
Apriori [2], [6] is a classic algorithm for learning association rules. It is designed to operate on databases containing transactions such as collection of items purchased by customers, or details of any website.

The algorithm attempts to find subsets which are common to at least a minimum number C which can be termed as cutoff or confidence threshold of the item sets.

Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time. This step is known as candidate generation for item sets and groups of candidates are tested against the test data (see Figure 2).

The algorithm terminates when no further successful extensions are found. Apriori uses breadth-first search and a hash tree structure to count candidate item sets efficiently. Figure 2 demonstrates step by step process of Apriori algorithm.

Consider Minimum Support = 50%



Order ID	Items
35	whole milk, vegetables, tropical fruit
293	frankfurter, vegetables, sausage
5894	whole milk, frankfurter, vegetables, sausage
6982	frankfurter, sausage

C1 - Candidate

Item Set	Sup
{whole milk}	2
{frankfurter}	3
{vegetables}	3
{tropical fruit}	1
{sausage}	3

L1 - Level 1 frequent item set

Item Set	Sup
{whole milk}	2
{frankfurter}	3
{vegetables}	3
{sausage}	3

C2

Item Set
{whole milk, frankfurter}
{whole milk, vegetables}
{whole milk, sausage}
{frankfurter, vegetables}
{frankfurter, sausage}
{vegetables, sausage}

C2

Item Set	Sup
{whole milk, frankfurter}	1
{whole milk, vegetables}	2
{whole milk, sausage}	1
{frankfurter, vegetables}	2
{frankfurter, sausage}	3
{vegetables, sausage}	2

L2

Item Set	Sup
{whole milk, vegetables}	2
{frankfurter, vegetables}	2
{frankfurter, sausage}	3
{vegetables, sausage}	2

C3

Item Set
{frankfurter, vegetables, sausage}

L3

Item Set	Sup
{frankfurter, vegetables, sausage}	2

Figure 2: Apriori Algorithm Demonstration

a.

The dataset has at least 90 unique items and around 900 transactions. Figure 2 shows a randomly selected four transactions which help us to understand the functionality of the Apriori algorithm. The dataset is scanned to calculate the support of all the items and forms the candidate set C1.

The frequency of occurrence of the selected items is known as SUPPORT (Sup). The item set that doesn't satisfy the minimum support (50%) is removed and the rest of the item set forms the level 1 frequent item set. Candidate C2 forms the possible item set combinations. These are the current potential candidates and there is no other possibility because of downward closure property. This means that if the item set {tropical fruit, sausage} is present in the candidate C2 then each element of {tropical fruit, sausage} should be a part of level L1. So, it just combines every element with every other element.

The dataset is again scanned and the support of the candidate C2 is calculated. The item set that doesn't satisfy the minimum support condition is eliminated and the remaining sets forms the Level 2.

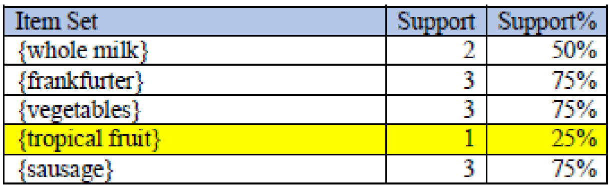
Similar process is followed and the potential candidates become the item set combinations.

i. *Support*

Support is defined as the frequency of occurrences of items divided by the size of the transaction [6] (see Figure 3).

From the dataset, the support % for some of the items is found to be: -

$$\text{Support (A -> B)} = \frac{\text{\# of Tuples containing both A \& B}}{\text{\# of Tuples}}$$



Item Set	Support	Support%
{whole milk}	2	50%
{frankfurter}	3	75%
{vegetables}	3	75%
{tropical fruit}	1	25%
{sausage}	3	75%

Figure 3: Illustration of Support of the items

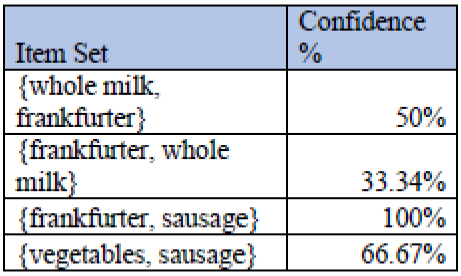
ii. *Confidence*

Confidence [6] is the probability or the likelihood of the combination of item sets being purchased by the customers (see Figure 4).

The Confidence is calculated for the item sets as follows: -

$$\text{Confidence(A->B)} = \frac{\text{\#Tuples containing both A \& B}}{\text{\# Of Tuples containing A}}$$

Sample results from the dataset: -



Item Set	Confidence %
{whole milk, frankfurter}	50%
{frankfurter, whole milk}	33.34%
{frankfurter, sausage}	100%
{vegetables, sausage}	66.67%

Figure 4: Illustration of Confidence of item sets

B. *Support Vector Machines*

SVM [4], [5] are supervised learning models that analyze data and recognize patterns. Classification and regression analysis is performed using SVM. An SVM model is a representation of the examples or observations as points in space which are mapped such that the observations of distinct categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

SVM can also be used for non-linear classification using kernels, however it is a linear classification model using SVM and predicting the quantity of sales of the most frequently purchased item. In the dataset, the most frequently purchased item is found to be frankfurter. The sales of frankfurter in classified into two classes HIGH and LOW depending upon its quantity being purchased in various months of the year.

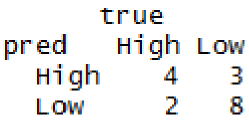
Now, this is a two-class separable dataset so there are many possible linear separators. SVM draws a decision boundary in the middle of the void between data items of the two classes. The SVM defines the criterion to be looking for a decision boundary such that it is maximally far away from any data point in a feature space. This distance from the decision surface to the closest data point determines the margin of the classifier. This method of construction means that the decision function for an SVM is completely specified by a subset of the data which defines the position of the separator. These points are referred to as the support vectors i.e. in a vector space, an observation can be thought of as a vector between the origin and the observation point.

i. Confusion Matrix

A confusion matrix [7], [8] is a nxn matrix (minimum value is 2) that describes the performance of a classification model built by any machine learning algorithm for which the true values are known. Each column of the matrix represents the instances in a predicted class while each row represents the instances in the actual class i.e. results from structured dataset.

ii. Matrix Table Snippet

Table 1: Confusion Matrix Table



```
      true
pred   High Low
  High    4   3
  Low     2   8
```

Table 1 shows a confusion matrix which depicts that there are 6 (4+2) HIGH instances in test set out of which 4 are predicted correctly and 2 are wrongly predicted as LOW

Similarly, there are 11 (3+8) LOW instances in the test set out of which 8 are correctly predicted and the rest 3 instances are wrongly predicted as HIGH.

Note: Terms High & Low indicates whether the sales of any item in a month is high or low.

The table of confusion has two rows and two columns that reports the number of false positives, false negatives, true positives, and true negatives. This allows more detailed analysis than any proportions of correct estimates. Accuracy is not a reliable metric for the real performance of a classifier, because it will yield misleading results if the number of samples in different classes vary profoundly in the dataset.

Some of the statistical terms [8] of confusion matrix:

1. True Positive
   Correctly predicted instances termed as 'LOW'.
2. True Negative
   Correctly predicted instances termed as 'HIGH'.
3. False Positive
   The 'HIGH' instance predicted as 'LOW'.
4. False Negative
   The 'LOW' instances predicted as 'HIGH'.

VII. STATISTICAL MEASURES OF PERFORMANCE

The result of the confusion matrix in Table 1 is used to calculate various statistical measures [8] which gives deeper analysis about the performance of classification system. The measures derived from this model are as follows:

A. Accuracy

It measures the statistical bias and tells the proportion of correct answers that any classifier can produce based on the model built.

B. Misclassification Rate

It is the proportion of incorrect results achieved by the classifier.

C. Recall

It calculates the chances of getting all the answers correct i.e. the proportion of positives that are correctly identified as such. It is also referred as fraction of relevant instances that are retrieved. Statistical measures Sensitivity and True Positive Rate are same as recall.

D. Precision

It measures how many correct answers can be produced by the classifier on an average i.e. fraction of retrieved instances that are relevant.

E. F-1 Score

It is a measure to test the accuracy of model and is the weighted average of precision and recall. Its values lie between 0 and 1. A score closer to 1 indicates best accuracy where as a score tending towards 0 indicates a poor accuracy of model.

F. False Positive rate

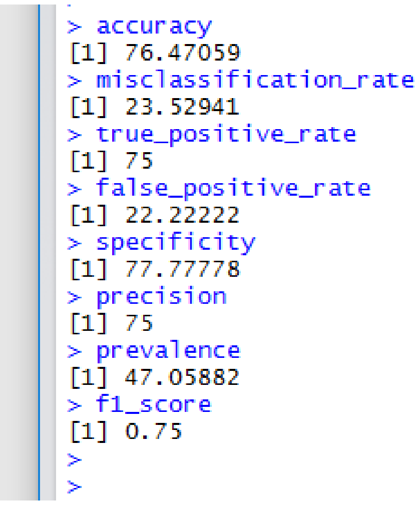
This measure indicates the rate at which a given event has occurred by the classifier whereas the event actually has not happened i.e. mistakenly a positive effect has taken place.

G. Prevalence

It tells how often the correct situation has actually occurred in the sample data.

H. Measures calculated in RStudio:

Mathematical formulas are implemented in R code to calculate the above mentioned statistical measures to understand the performance of the predictive model.



```
> accuracy
[1] 76.47059
> misclassification_rate
[1] 23.52941
> true_positive_rate
[1] 75
> false_positive_rate
[1] 22.22222
> specificity
[1] 77.77778
> precision
[1] 75
> prevalence
[1] 47.05882
> f1_score
[1] 0.75
>
>
```

Figure 5: Statistical Measures

VIII. RESULTS

There are various results that can be deduced and used by the retailers from this analysis report. They are explained as follows:

A. Customer Name vs Quantity

First step in this analysis process is to analyze the given data against the respective attributes. The scatter plot relating the customer vs quantity shown in figure 6 explains the quantity of products bought by various customers. This graph can give an idea about the customers who have purchased large quantity of products.

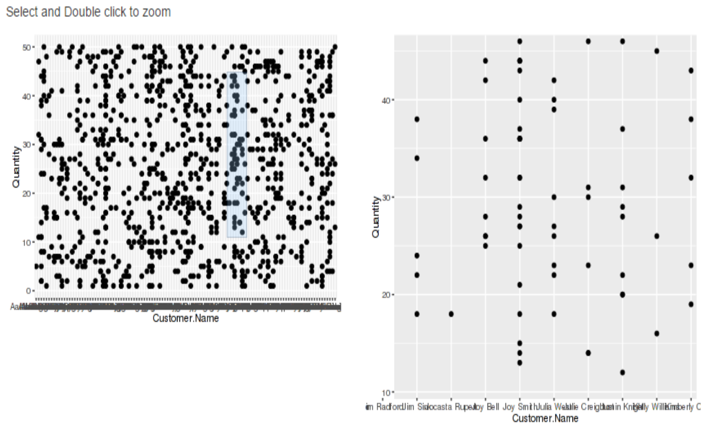
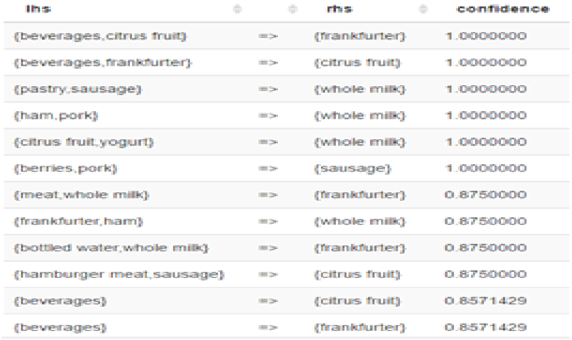


Figure 6: Scatter Plot of Customer Name vs Quantity

B. Association Rule Sets

Second result obtained in figure 7 is the rule set generated from apriori algorithm [2], [6] in relation with the given dataset. These rule set explains how the products are related to other purchased items. This characteristic is explained using support and confidence defined by apriori. A sample of the rulesets are explained below:

- The rules have three columns – lhs (left hand side), rhs (right hand side) and confidence. Lhs & rhs are the items or item sets and confidence is the probability of the item purchased in rhs when the item in lhs is purchased.
- Considering the third row for example, if a customer buys pastry & sausage then there is 100% probability that he will also purchase whole milk.
- Similarly, the combinations of sets can be any number of item sets.
- The retailer can use this analysis to decide right items for promotion and which one to exclude, ultimately maximizing profit of the store.



lhs		rhs	confidence
{beverages,citrus fruit}	=>	{frankfurter}	1.0000000
{beverages,frankfurter}	=>	{citrus fruit}	1.0000000
{pastry,sausage}	=>	{whole milk}	1.0000000
{ham,pork}	=>	{whole milk}	1.0000000
{citrus fruit,yogurt}	=>	{whole milk}	1.0000000
{berries,pork}	=>	{sausage}	1.0000000
{meat,whole milk}	=>	{frankfurter}	0.8750000
{frankfurter,ham}	=>	{whole milk}	0.8750000
{bottled water,whole milk}	=>	{frankfurter}	0.8750000
{hamburger meat,sausage}	=>	{citrus fruit}	0.8750000
{beverages}	=>	{citrus fruit}	0.8571429
{beverages}	=>	{frankfurter}	0.8571429

Figure 7: Association Rules between items

C. Frequently Purchased items

After deciding the rulesets, the next step of the analysis process is to guess the most frequent item [2]. The most frequent item graph is plotted based on how frequent an item occurs in the ruleset. For instance, the item purchased maximum number of times is frankfurter (see figure 8).

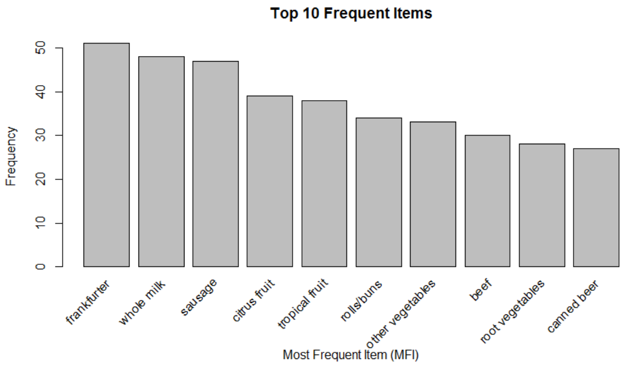


Figure 8: Top Frequently Purchased Items

D. Lift ratio

Lift ratio [2] tells how much significant is the relationship between the antecedent (lhs) and consequent (rhs) item sets. It's the ratio of confidence to the expected confidence. It is visualized as a horizontal bar plot shown in figure 9.

- Large lift ratio implies higher significance in the association rules.
- The relationship between lhs & rhs is more significant if the lift ration is more than 1.0.
- From the lift ratio plot, it can be deduced that the items like frankfurter. Hamburger meat, other vegetables & citrus fruit have higher lift ratio, hence there are greater changes of these items being purchased together.
- Similarly, items with very low lift ratio like yoghurt, root vegetables, chicken or bottled water have a relatively lower chance of forming an effective association rule set of items.

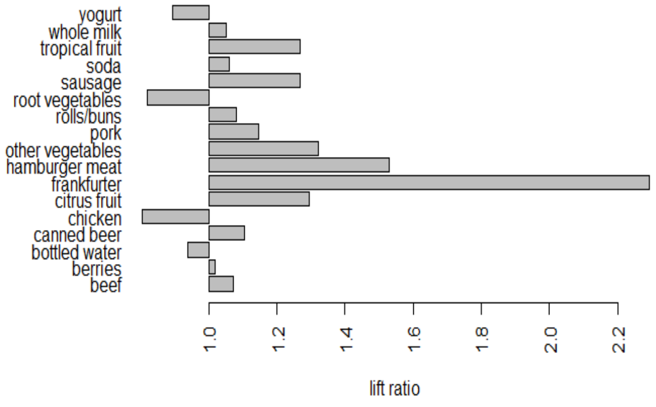


Figure 9: Lift Ratio

E. Relative Items Plot

The plot of frequently purchased items relative to the item being purchased [2] by most of the customer is shown in figure 10.

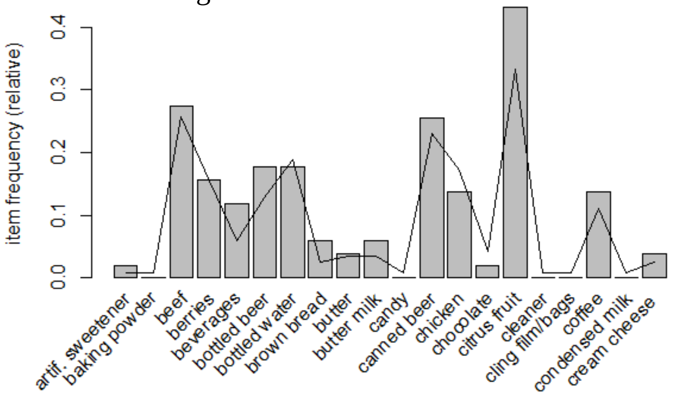


Figure 10: Items relative to Frankfurter

F. Support Vector Machines

SVM [1], [4] is used to create a decision boundary between the high and low values. This is used to predict the future behavior of the product. The item being considered is frankfurter as it is the most frequently purchased item. The output is analyzed using confusion matrix table discussed in section IV. SVM Classification is shown in figure 11.

- The decision boundary segregates the purchase quantity of ‘frankfurter’ item into two classes as ‘LOW’ & ‘HIGH’.
- The colored symbols black & red ‘X’ represent the high quantity that is correctly predicted as HIGH & low quantity being wrongly predicted as HIGH respectively.
- Similarly, black & red ‘O’ indicates low quantity of item correctly predicted as LOW and high quantity wrongly predicted as LOW respectively.

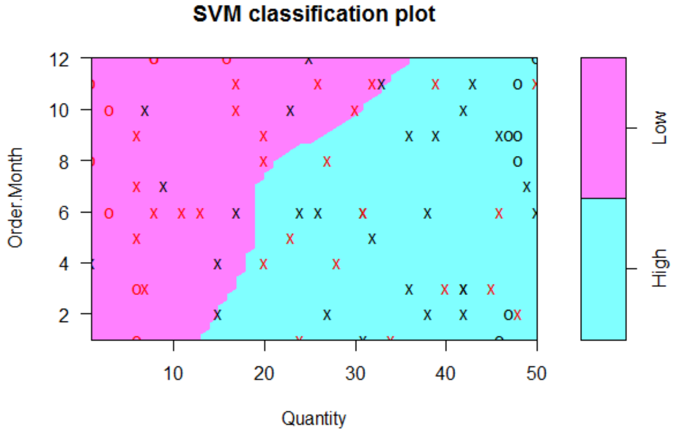


Figure 11: Support Vector Machine Classification

IX. Shiny Dashboard

Shiny is a web application framework for R that combines the computational power of R with the interactivity of the web [9]. Once the application is developed, it can be hosted and deployed on the web either on a shiny serve or in the cloud with shinyapps.io.

Shiny dashboard is hosted at https://niranjanrshiny.shinyapps.io/Prediction_App/

X. Conclusion

This report explains how the behavior analysis is done with the help of machine learning algorithms, association rule mining and visualization libraries.

The results achieved are:

- The relation between various products in the super market are related and a ruleset is generated with the help of apriori algorithm.
- The retailer can use the item set rules to conduct cross-promotional programs & promote items.
- The analysis of the most frequent item is established with the help of linear classification using SVM.
- SVM classification plot helps the retailer to decide that in which month of the year an item should be promoted depending on the sales of its item sets derived from the association rules.
- The accuracy achieved through this prediction ranges from 68% – 79%.

REFERENCES

[1] David Meyer, 'Support Vector Machines – The Interface to libsvm in package e1071', FH Technikum Wien, Austria, 2015

[2] Michael Hahsler, Christian Buchta, Bettina Gruen, Kurt Hornik and Christian Borgelt, 'Package arules', 2016.

[3] Ni.com, 'NI Vision 2010 Concepts Help', 2016. [Online]. Available: http://zone.ni.com/reference/en-XX/help/372916J-01/nivisionconcepts/supportvectormachines/

[4] Cran.r-project.org, 'arules: Mining Association Rules and Frequent Itemsets', 2016. [Online]. Available: https://cran.r-project.org/web/packages/arules/

[5] R-bloggers.com, 'R News & Tutorials', 2015. [Online]. Available: https://www.r-bloggers.com/

[6] Dataschool.io, 'Simple Guide to Confusion Matrix Terminology', 2014. [Online]. Available: http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

[7] Github.com, 'Data Science for Political & Social Phenomena', 2016. [Online]. Available: https://github.com/chrisalbon/superstore_sales

[8] Youtube.com, 'MIT OpenCourseWare – Support Vector Machine', 2014. [Online]. Available: https://www.youtube.com/watch?v=_PwhiWxHK8o

[9] Shiny.rstudio.com, 'Teach Yourself Shiny'. [Online]. Availble: http://shiny.rstudio.com/tutorial/