

**MACHINE LEARNING**

**PREDICTING CUSTOMER BEHAVIOUR THROUGH DATA MINING APPLICATION**

**SUBMITTED TO**

DR. RASSULE HADIDI

**SUBMITTED BY SUBMITTED ON**

Rohith Madamshetty [rmada2@uis.edu](mailto:rmada2@uis.edu) 05/09/2019

Contents

[Executive Summary 3](#_Toc6506959)

[Introduction 4](#_Toc6506960)

[Problem Statement 5](#_Toc6506961)

[Purpose and Research Questions 5](#_Toc6506962)

[Research Questions 6](#_Toc6506963)

[Scope, Assumptions, and Limitations 6](#_Toc6506964)

[Literature Review 6](#_Toc6506965)

[Research Methodology 16](#_Toc6506966)

[I. SCRUM Methodology Approach 16](#_Toc6506967)

[II. Data Cleaning 18](#_Toc6506968)

[III. Data Preparation 19](#_Toc6506969)

[IV. Data Visualization to explore the dataset 20](#_Toc6506970)

[V. Model Selection in Data Mining 25](#_Toc6506971)

[VI. XGBoost Model 26](#_Toc6506972)

[1. Creating Predictor Variables 26](#_Toc6506973)

[2. Creating Test Data for Scoring 29](#_Toc6506974)

[3. Applying the predictive model on the scoring data. 33](#_Toc6506975)

[VII. Apriori Algorithm 34](#_Toc6506976)

[1. Creating Transaction Data 34](#_Toc6506977)

[2. Creating and Executing Apriori Model 35](#_Toc6506978)

[Results and Discussion 35](#_Toc6506979)

[1. XGBoost Result 35](#_Toc6506980)

[2. Apriori Result 37](#_Toc6506981)

[Practical Recommendations and Conclusion 38](#_Toc6506982)

[References 40](#_Toc6506983)

# Executive Summary

The Heavy competition in retail industry and growth of e-commerce has brought many changes in the way their businesses are run. To gain competitive advantage and customer retention the retail companies must reinvent their business model. Predicting the customer behavior play important role in developing the business model of the company. The retail industries have advantage of face-to-face customer experience. The customer needs and purchase patterns are required to improve the customer services and gain their loyalty. Understanding purchasing pattern means finding the products which customers tend to order more frequently or products which are reordered frequently by the customers. This paper provides overview on the data mining techniques used to predict customer behavior and the report on the detailed analysis of patterns and demographics extracted from the data which can be used to develop the customer relation management and improve customer satisfaction.

We are using XGBoost and Apriori algorithm as prediction techniques for extracting information from data and data visualization to analyze the data. As a result, we expect to find the products which are likely to be reordered by the customers and demographics, such as which day at what time we can expects customer traffic in the store.

# Introduction

The retail sector is vital to the world economy. This Industry is growing at a rapid rate. Providing quality products at a low price has become the main objective. The main problem is cutthroat competition and a decrease in the profit margins that have led retailers to take advantage of data mining (Kumar & Arora, 2015). Large retail chains and grocery store contain a rich collection of data over years (Kumar & Arora, 2015). Analyzing customer behavior using this data is a challenging task. The cumulative process of finding and interpreting patterns from data involves many steps such as selection, preprocessing, transformation, data mining and interpretation (Kaur & Kang, 2016). Extracting this data to analyze customer behavior using data mining helps firms and organizations to improve their business and marketing strategies by (Raorane & Kulkarni, 2011)

1. Identifying the relationships between the products a customer purchases and providing a recommendation.
2. Arranging similar products on shelves after studying the customer purchase patterns.
3. Understanding the environment or demographic location that is influenced by the customer in choosing products.
4. By attracting the customers by offering them discounts and coupons based by studying their behavior.
5. Improving restocking inventory based on the recurring purchases.

Through customers’ behavior analysis, accurate profiles are being generated by specifying the needs and interest of the customer. This helps business by giving customers what they want, when they want, leading to better customer satisfaction and hence keeping them to come back for more (Victor, Abimbola, Mercy, Esther, & Eloho, 2014.).

This paper uses data mining technique, to develop a predictive model that will be helpful to discover the patterns from customers purchases and use those patterns for future prediction. By using predictive algorithms, patterns and trends of the customer data will be studied. This study will help us understand the psychological mindset of a customer by analyzing his buying behavior. This ultimately helps improve business, marketing strategy, and customer satisfaction.

# Problem Statement

In the current scenario understanding customer behavior is crucial to stay ahead in the competition. For instance, to predict which previously purchased products will be in a user’s next order will be an uphill and beneficial task. The main problem here is to analyze and study the data to maximize benefits to the customer thus improving the business sales. Also, making recommendations through customer behavior prediction to expedite their search while shopping. Analyzing customer trends and patterns that would help predict customers future actions, this, in turn, will help the organization build a relationship with the customer.

# Purpose and Research Questions

The main purpose of this research paper is to understand the data analytics which is used to predict customer behavior and effectively provide critical improvements in sales and operational efficiency. By predicting, companies would be able to save cost and effort by producing only those offerings that customers would instantly prefer. However, by using data mining methods we can predict what could be the next order from the past data and transactional data.

In this research, we want to understand how to generate recommendation system by we are understanding frequent item-sets of customers purchase.

# Research Questions

Q1: How to analyze customer behavior through customer retailer relationship management?

Q2: What products will customers tend to reorder from the store?

Q3: What better services can be given to customers for achieving high customer relationship?

# Scope, Assumptions, and Limitations

The research on this topic so far was limited to building customer retailer relationship and improve customer satisfaction (Oliveira, 2012a). We would like to extend this research to classify the customer behavior and predict the trends which a customer would prefer when he repeatedly purchases similar items. This can help us understand the customer needs and target the products which help to maintain the inventory levels. Additionally, how this prediction would impact the organizations supply chain and enhance business strategies.

Assumptions:

* Variables to be used in analysis are normally distributed.
* Relationship between the variables used are linear.

Limitations:

* Incorrect or incomplete data effects the output and predictive analysis.
* Accurate results within its own limits, cannot promise error free accuracy.

# Literature Review

The following is the research has already been done on this topic. This paper fills in the gap using Data mining techniques to address the current problem. It makes use of XGboot package to build the predictive model and Apriori for the recommendation system. It also explains customer relationship management as an important aspect to solve issues related to smooth conduct of business and increase sales.

**Predictive Analysis Using Datamining**

Predictive analytics is an area of statistics that deals with extracting information from data and using it to predict trends and behavior patterns. Change in time and increase in data had led to the expansion in modern data storage size (Rodpysh, 2012a). We know that data mining is extraction of information and inhibiting the knowledge from the massive amounts of data. We mostly use data to analyze the shopping patterns of customers or what product is brought most often in the stores. What is the association between those products? Is there any pattern for that purchase? In case if the data shows certain relation relationship then the store management will predict it as fast-moving product and arrange it accordingly (Rodpysh, 2012a). For instance, if the management introduces different promotions on the products that are brought most frequently it would boost sales. To make profits the store will not run any promotions during the busy days. Another practice is observing the weekly shopping habit of customer, what products they buy and which brand and quality. The use of product is very important because it will help marketer how can it be encouraged to increase the sales. This information will be used for stocking purpose and handling the inventory cost. To do monthly analysis, how many times is the product being purchased in a month from start to end. In all these cases data is mostly collected from various sources. Some of these are inventory, product sales, cost, accounting etc (Raorane & Kulkarni, 2011a).

Data Mining can be classified by the techniques adopted or employed. These patterns will describe what kind of approach is best for predicting the behavior of the sales. Data mining techniques are expected to be more effective in analyzing customer behavior (Victor, Abimbola, Mercy, Esther, & Eloho, 2014).

Data mining process involves five important stages as follows (Singh, Bellathanda Kaverappa, & Joshi, 2018).

* Set of training samples should be collected and chosen for trimming
* What are the expected data mining techniques used for the data will be specified
* Understand the data, transfer of existing knowledge about the process
* Hierarchy of concepts
* Evaluation criteria

After all these extracted data should be presented and is determined on the type of knowledge. It is a goal to predict what happens in the future.

We will learn data mining with predictive analytics to determine:

* Customer churn
* Recommendations for products
* Risk calculation
* What features can be added for product development
* Quality fulfillment and maintenance of product

A data mining system will adopt multiple techniques to combine the merits of individual approaches.

**Predictive methods**

As four dimensions act as closed cycle which creates the deeper understanding of the customer to

maximize the customer value in customer management system. Data mining techniques helps to extract the hidden customer characteristics and behavior from lager databases (Patel, Karvekar, & Mehta, 2014). The different predictive models to study customer behavior are:

1. Association: Association rules are event-based technique which is used to learn the relations between variables in the large database. It is also known as market-based analysis or affinity analysis. It mainly deals about What-Goes-With-What and we use Apriori Algorithm for generating frequenting item sets. Link analysis and Sequence mining is most commonly used techniques in association. Where link analysis is used to evaluate the relation among different nodes and sequence mining is used to examine the relationships in order of occurrence (Raorane & Kulkarni, 2011a).
2. Classification: Classification consists of a set of pre-classified samples to create a model which can classify the large set of data. It can derive important information about data and metadata. It is like clustering where the different segments of data sets are divided into classes and classifies the data depending on algorithm. Decision Trees is among the classification method where the data is classified into number of classes depending on the value of the input variable (Oliveira, 2012).
3. Clustering: Cluster analysis is used to form group or cluster of similar objects based on similar records present in data. Cluster analysis can be performed in two ways i.e., Agglomerative(hierarchical) clustering and k-means (non-hierarchical) clustering. Agglomerative clustering is a bottom-up approach, and which doesn’t need to specify the number of clusters. In k-mean clustering, we can specify number of clusters to review the different results depending on the clusters. The main difference between these are k-mean can be applied to larger database (Oliveira, 2012).
4. Artificial neural network: Neural computing is a machine learning strategy that uses pattern

recognition and the models that it produces are known as artificial neural networks (ANN) or just neural networks. These networks are inspired by the human brain and its ability to process information and solve problems in ways that modern computing formerly couldn’t compete with. ANN are made up of simple processing elements called artificial neurons that are interconnected and that operate collectively and in parallel. These neurons receive information from each other or from outside stimuli, transform the inputs into some useful format, and then share that information with other neurons or use it to generate some form of external output (Oliveira,2012).

1. Regression: Regression analysis is used to identify and analyze the relationship among the variables. It is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. Regression and correlation analyze linear relationships between variables, finding the regression line that best fits the data. Correlation analysis evaluates the rho value which provides the evidence for linear relationship among same pairs of variables. It is used to reveal the associations among continuous variables, but it doesn’t address non-linear relationships. The correlation coefficient may range from -1 to +1, +1 indicates perfectly positive relationship, -1 indicates perfectly negative relationship and 0 indicates no relationship among variables (Oliveira, 2012). If correlation coefficient is between 0 to -1 then the relationship is inclined towards negative and if the correlation coefficient is between 0 to +1 then the relationship is inclined towards positive (Oliveira, 2012).

**Customer behavior analysis using Data mining:**

Customer behavior analysis is observation of a group of customers buying patterns to gain insight to customers behavior and how similar they behave under similar circumstances. Customer behavior analysis help in understanding preference and tendencies of the customer and providing them preferred type of service which help in maintain efficient customer relationship management (Ngai, Xiu, & Chau, 2009). Customer retention and satisfaction is main aim of customer relationship management. This analysis helps in predicting future trends and customer behavior, this help in making knowledgeable decisions (S. Kumar & Arora, 2015). Data mining techniques can be used to analyze customer’s buying behavior.

Data mining techniques are considered more effective for customer behavior analysis (Raorane & Kulkarni, 2011b). Using customer data from past to understand the customers buying behavior in predicting the future trend can be achieved by applying data mining techniques. Data mining is combination of different disciplines like statistics, artificial intelligence, databases and information science (Victor, Abimbola, Mercy, Esther, & Eloho, n.d.). Data mining techniques use the raw data collected and stored by the organization over years and convert them into integrated data which can be accessed easily (Rodpysh, 2012b). Data mining is used to explore data to find standard patterns and symmetric relations comparing various variables available in the data (Victor et al., 2014).

There are three main stages in data mining process, which are

1. Data Exploration: Data cleaning, Data transformation and subset selection is done in this stage (Victor et al., 2014).
2. Model Selection: In this stage the data mining model to be used is selected comparing different data mining models and selecting the better performing one (Victor et al., 2014). Various data mining techniques are used to create a data mining model.
3. Deployment: In this stage data is applied on previously developed/selected model. Prediction are generated as result of this stage. The results are first evaluated and validated, then they are stabilized and consolidated into useful information (Rodpysh, 2012b).

**Customer relationship management**

A customer-centric focus has become one of the most significant areas in the retail industry (Anderson, Jolly, & Fairhurst, 2007). The main problem is that the retailers are striving hard to maximize their relationship with the customers. CRM is based on RM and is focused on the technology underlying the management of customers (Oliveira, 2012). Customer identification, customer attraction, Customer Retention, Customer development these four dimensions can be regarded as closed-loop customer relationship management (Ngai, Xiu, & Chau, 2009).

**Customer Identification:** CRM begins with customer identification, which is referred to as customer acquisition (Ngai, Xiu, & Chau, 2009). This phase is to target the population who will be the future customer and will be beneficial to the organization. It includes a complete analysis of what kind of customers like to purchase an item.

**Customer Attraction:** There must be ways to attract customer to purchase the products. Loyalty programs include such activities and build long-term relationships with customers (Rodpysh, 2012). These programs help analyze customers shunned, ranking credit, the quality of services or satisfaction (Rodpysh, 2012). Direct marketing through coupons or direct mail is a promotion process which motivates customers (Ngai et al., 2009).

**Customer Retention:** Customer satisfaction is the main concern when it comes to customer retention (Oliveira, 2012). For the company to grow it is essential for it to retain its customers. Customer retention mainly includes one-to-one marketing, loyalty programs, and complaints management.

One-to-one marketing this is the personalized marketing campaigns which are supported by analyzing, detecting and predicting changes in customer behaviors (Ngai et al., 2009).

**Customer Development:** Customer Development involves consistent expansion of transaction intensity, transaction value and individual customer profitability. Elements of customer development include customer lifetime value analysis, up/cross-selling and market basket analysis (Ngai et al., 2009).

Cross Selling is advertisements or promotions that are concerned to improve customer interest in a product or service the organization has to offer (Prinzie & Van den Poel, 2006).

Customer Lifetime value is the prediction of the total net income a company can expect from a customer (Drew, Mani, Betz, & Datta, 2001).

Market basket analysis aims at maximizing the customer transaction intensity and value by studying the purchase behavior of customers (Kaur & Kang, 2016).

**Gamification Concept:** To improve customer retention and enhance the services given Gamification concept can be used (Huotari & Hamari, 2011). It basically contains game-like mechanics to attract customers to buy more products and return to the store to get more rewards. One such example is explained in (Olsson, Hogberg, Wastlund, & Gustafsson, 2016). A treasure hunt map App in a real retail environment which consists of different challenges and on completion of these challenges the customer would be rewarded points (Olsson et al., 2016). This will make the shopping experience fun and the customer will be satisfied when they are rewarded. This overall improves the customer relationship with the retailers.

**XGBoost**

XGBoost is the abbreviation for eXtreme Gradient Boosting. This package is so efficient to use, and it takes less effort to implement gradient boosting framework. As we have a large data set to deal with in the project we have selected this model. In R we do not get this package by default. To install this, we must execute install. package (“xgboost”) command. This package comes with efficient linear model solver and tree learning algorithm. Various kinds of objective functions which includes regression, ranking and classification. It is made in such a way that the package can be enhanced. So that defining of the objectives will be easy for the user. Special features of xgboost are (T. Chen & He, 2019): -

1. It is faster to execute. It can run on Windows and Linux to do parallel computation automatically. It is 10 times faster than opm.
2. It also takes different kinds of input formats: -

* Dense Matrix, Sparse Matrix (dgCMatrix), Data File (Local Data File), xgb.DMatrix (xgboost’s own class).

1. For both linear booster and the tree booster xgboost accepts sparse input.
2. It supports customization of different functions like objective function and evaluation function.
3. The performance of xgboost is better on different datasets (T. Chen & He, 2019).

**Apriori Algorithm**

Association Rule mining plays a major role in studying customer behavior and helps to increase sales by giving recommendations in a way to attract customers and to retain them. The increase in sales can be done by observing customer requirements. Based on the movement of products, we can increase product sales by placing the goods in a particular position and also offering special promotion plans (D. V. S. Kumar & Renganathan, 2011). The reason for selecting the Apriori Algorithm is it satisfies the purpose by giving recommendations based on the transactions of the user by discovering the underlying patterns. Identifying the frequent sets from the database and generating the association rules is the main aim of the analysis. It is easy to implement and does not require much memory (Oliveira, 2012b).

Uses of Apriori algorithm: -

• It is used to build a customer classification model.

• It is used to do the process of customer segmentation.

• It is used to analyze the data from the variable perspective and provide useful insights (Mu, 2013).

• It also allows using of messy and unusable data.

For example, if we consider two products A and B the association rule generated between them is validated by support, confidence and lift (Aguinis, Forcum, & Joo, 2013a). Rules which are stronger between the products are recommended. We can use the recommendations to compare with the previous behavior of the customer. This study helps in explaining the changes in added patterns, emerging patterns and asses to provide new marketing strategy (M. Chen, Chiu, & Chang, 2005a).

# Research Methodology

We found a data set on Kaggle (Anhtony & Hamner, 2017) which is from Instacart store. This data set contains customer profiles, transaction data and customer purchases for over a year. This data set contains about 200,000 user profiles and 50,000 different products.

Using the available data to predict customer data analysis. we start with cite the source you have used to list these steps (Singh et al., 2018),

1. SCRUM methodology approach
2. Data Cleaning and Merging
3. Data Preparation
4. Data visualization to explore the dataset
5. Model Selection in Data Mining
6. XGBoost Model
7. Apriori Algorithm

## **SCRUM Methodology Approach**

We have followed the SCRUM methodology to complete our project. SCRUM is an agile method to manage the project development. It can be defined as a framework to manage a process. SCRUM mainly concentrates on teamwork that is required for software development. It also focusses on iterative progress towards the main goal. This is a small part in the agile development (Cohn, 2010).

Each team member played a specific role and focused on quick adoption of the new technologies. We started with whatever knowledge we already had in this area and kept working more on it till we reached our goal. Our SCRUM framework was very simple which consisted of leader, scrum master and the team members. It began with product backlog, followed by sprint planning and sprint. We then reviewed the process and the started the working on framework again.

Product Backlog: The project leader made a list of tasks to be done and assigned them to the team members. This was the main building of the project backlog.

Sprint Planning: With the tasks assigned in the product backlog the team members started working on the tasks. The team members were free to ask any questions to the team leader

Sprint: Face to face SCRUM Meetings were held every week and discussions took place almost daily. Each team member shared the progress of his work in the meetings and reported to the SCRUM master.

Oversight by the SCRUM Master: The SCRUM master was responsible to review the progress and combine all the work done and report it to the team leader. SCRUM master was good facilitator who advised and guided the team in the right direction throughout the project

Completed Sprint: Once the sprint was complete the work was ready to be delivered and issues were solved.

Review: Everything was reviewed to check if the process is going on smoothly.

Repeat: This was repeated for all the sprints throughout the project development.

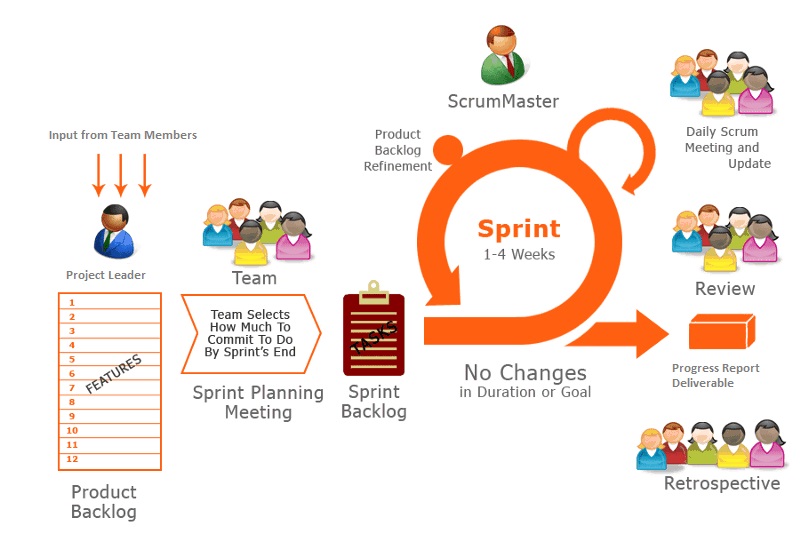


Figure 1: Scrum Model (Reprinted from Using Agile Scrum for Web Development., Retrieved from https://www.nefromonrain.com/agile-scrum-web-development/)

## **Data Cleaning**

After trying different tools like SAS enterprise miner, Excel, and R programming to analyze the data, we found that R programming is more flexible to handle large datasets. We have researched about 14 articles to come up with customer relationship management.

* At first, we saw the size of the data and decided on what tools we must clean the data.
* As the size of the data is more than 700 MB we are not able to open the file completely on Excel.
* Cleaning the data has become big challenge for all of us as we are not able to use Excel (a powerful productive tool which does not require any coding).
* We then tried to clean the data using SAS Enterprise Miner 14.3 which is available to us on citrix. As the limit for using citrix virtual machine is not more than 500 MB again we faced a failure in executing the data on SAS Enterprise Miner.
* Then our problem is resolved using R. As R executes the larger size of data sets easily we just cleaned the data by executing some simple code lines using R. We checked for NULL values and duplicate values and deleted them
* Initially we got the normalized data which requires us to load the different tables and execute them.
* We overcome the problem by de-normalizing the tables and eliminated the columns which are not required.
* Now we have a single data set which is approximately 1.3 GB. This gives us furthermore challenge to deal with the data.
* But R seems to be capable to handle the data size.

## **Data Preparation**

After cleaning the data, the following variables were considered for our research purpose:

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Variable Type**  **(Numerical, Text, Categorical)** | **Variable Description** |
| User ID | Numerical | Id (primary key) |
| Order ID | Numerical | Id of the Order |
| Product ID | Numerical | Id of the product |
| Product Description | Text | Name of Product |
| Days since prior order | Numerical | Days since prior order |
| Reorder | Categorical | If the product is reordered |
| order\_hour\_of\_day | Numerical | In which hour of the day product is ordered |
| Eval\_set | Categorical | Describes type of the order |
| Order\_number | Categorical | Describes order number |
| Day\_since\_prior\_order | Categorical | No.of days since prior order |
| Reorderd | Categorical | Describes whether product is reordered or not |

Figure 2: Important Variables from the Data Set

Since out dataset is extremely large we made use of sampling techniques. We have partitioned the data and used the default 40 for train, 30 for validate and 20 for test.

## **Data Visualization to explore the dataset**

We explored the data using R and Tableau. The smaller datasets were loaded in Tableau and large datasets were executed using R. We explored which hour of the day the product was ordered maximum. We also observed in which day of the week products were ordered maximum. We visualized the number of days customers is taking to order the products again. Lastly, we analyzed number of items people buy. The packages required to do data visualization in R are provided by the following libraries: -

* data.table, dplyr, ggplot2, knitr, stringr, DT

**Hour of Highest Sales**

The results of graph show that what products are ordered most frequently in a specific time of the day. As we can see that people buy groceries mostly at 10 a.m. in the morning and again after 3 p.m. in the noon. The other hours of the day seem to be little dull.

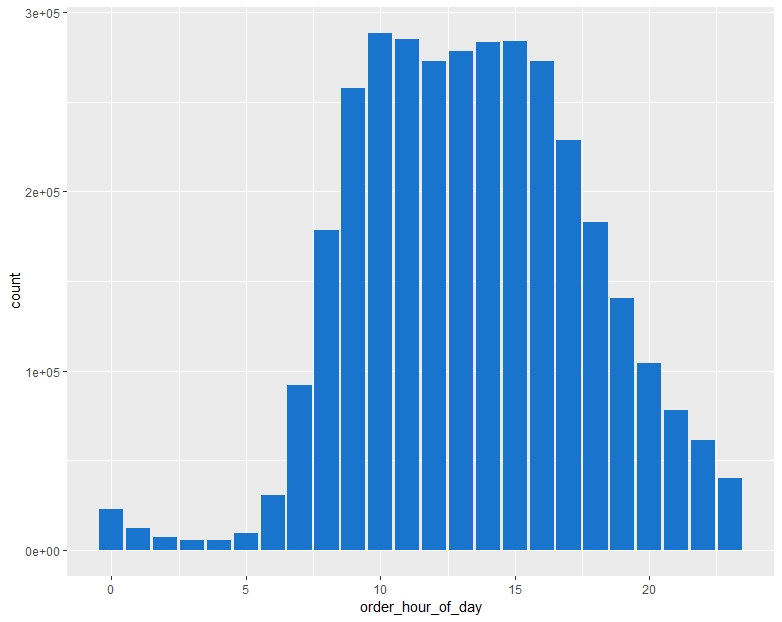


Figure 3: Hours of Highest Sales

**Products that are sold in High Margin**

When we look at graph BANANA is the bestseller. Among all the other products this is bought by most of the customers. This comparison is done with 10 other bestsellers.

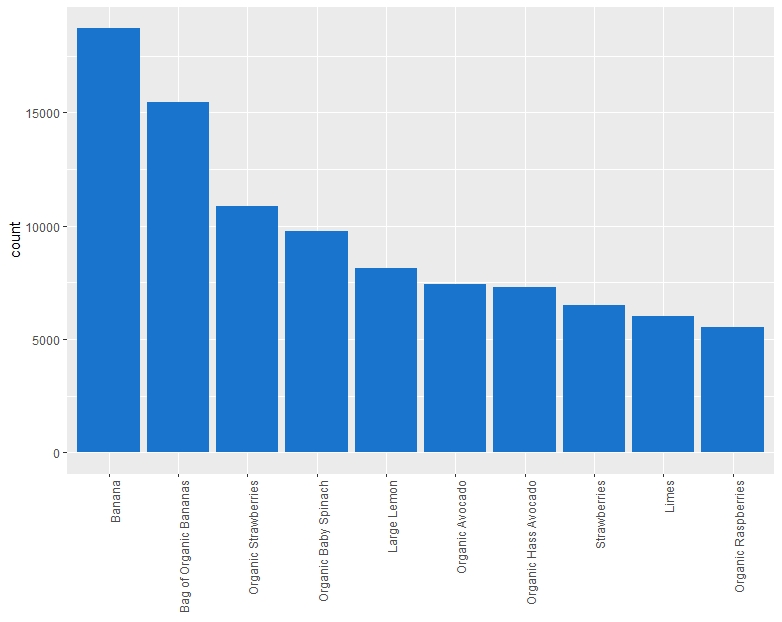


Figure 4: Products that are sold in High Margin

**Comparison of Re-Orders**

Insights show that 59% of the items are reordered. In the graph, on the axis of reordered there are two categories 0 and 1. 0 means the same item is not purchased whereas 1 shows the same item is purchased again. There is a high probability of reorders from the top 10 products which are purchased most often.

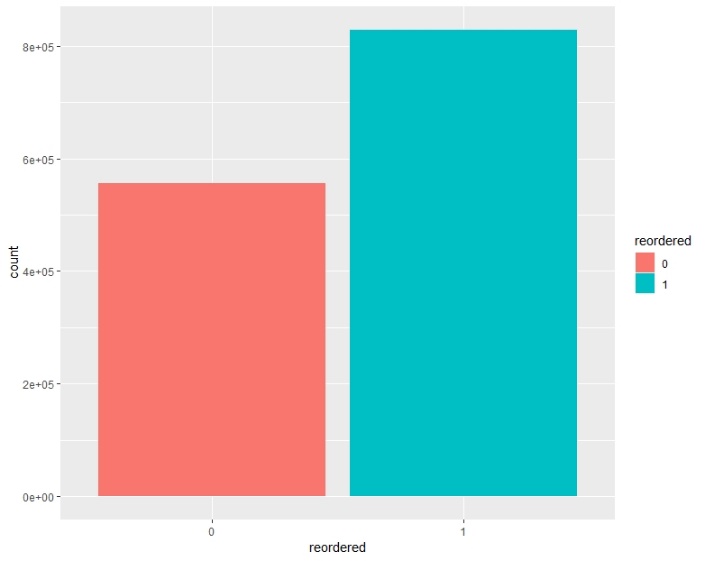


Figure 5: Comparison of Re-Orders

**Day of Highest Sales in a week**

Day plays a key role in grocery shopping. According to our results it shows most orders are 0 and 1. Apparently, 0 and 1 doesn’t represent which specific day. But we assume that the days are weekends because people will get free time from work.

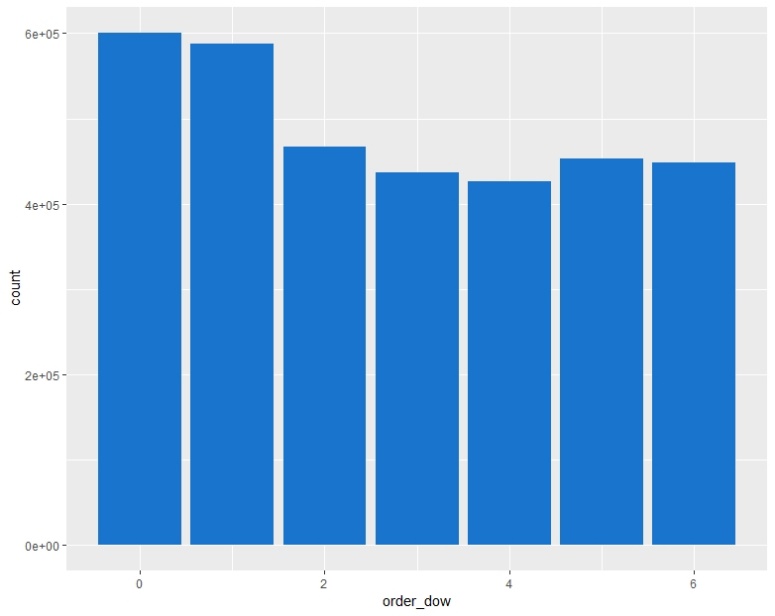


Figure 6: Day of Highest Sales in a week

**Days people take to re-order from prior order**

According to the analysis, it shows that most of the customers are re-purchasing groceries on the 7th and 30th day of their prior orders.

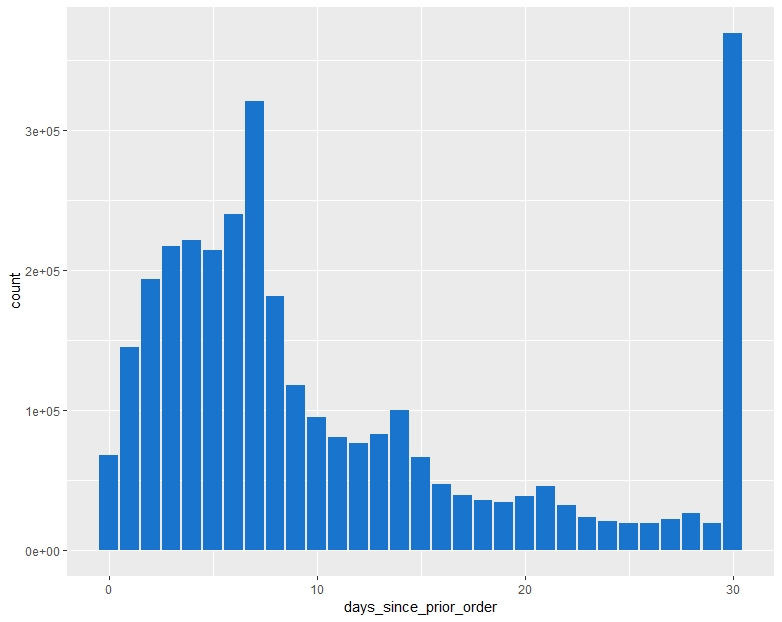


Figure 7: Days people take to re-order from prior order

## **Model Selection in Data Mining**

1. **XGBoost Model**

We have selected XGBoost method to build predictive model. We created predictor variables which helps to study the behavior of product and customers.

1. **Apriori Algorithm**

This algorithm gives recommendations based on the products purchased by the user by discovering the underlying patterns. Identifying the frequent item-sets from the database and generating the association rules is the main aim of the analysis. This method is simple to implement and does not require much memory (Oliveira, 2012b)

We can interpret the results by using data visualization techniques. Finding out the hidden benefits and suggesting them to customers for achieving high customer relationship.

## **XGBoost Model**

### **Creating Predictor Variables**

Based on the given data we can create many predictor variables. They can be product predictors to define the attribute of the product, user products that describe users’ behaviors, user product predictors that define the user behavior towards a product and datetime predictors.

Product Predictors: To define the products the predictor variable created are as follows:

1. orders\_per\_product – This is the total number of the orders per product.
2. prob\_reorder – This gives the probability if the product is ordered again after the first order.

prob\_reorder= (second\_time\_buyers / first\_time\_buyers)

1. reorders\_of\_product – This is the average number of times user has purchased the product that is already purchased once.

reorders\_of\_product= (1+ prod\_reorders/ first\_time\_buyers)

1. ratio\_of\_prodreorders – This is the ratio of products reordered on the total number of orders of product
2. first\_time\_buyers – These are the total number of customers who purchased the product atleast once
3. second\_time\_buyers – These are total number of buyers who bought the products more than once.

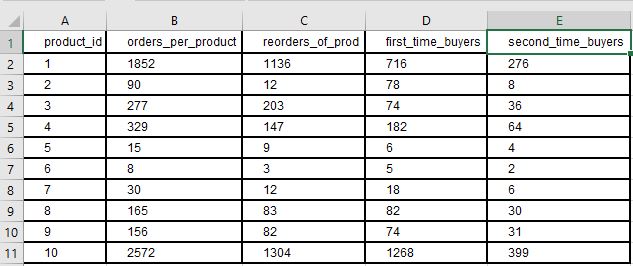


Figure 8: Product Predictors.1

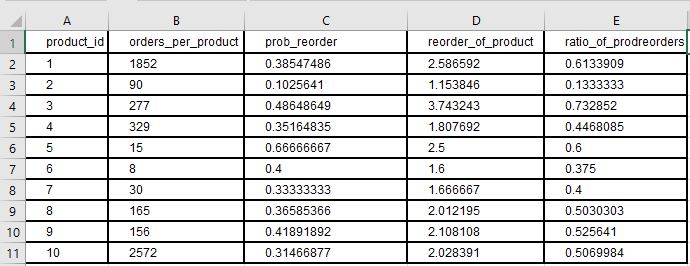


Figure 9: Product Predictors.2

User Predictors: The predictor variables that define the user behaviors are as follows:

1. no\_of\_orders\_user – This variable defines the total number of orders per user.
2. gap\_bet\_frst\_last\_order – This is the gap between the first and the last order
3. avg\_period\_of\_consecutive\_orders – This is the mean time between the two orders that are consecutive.
4. total\_basket\_items – These are the items included in the user’s basket
5. user\_reorder\_ratio – This is the ratio of reordered item for every user
6. user\_distinct\_products – These are the total number of unique items ordered by user
7. avg\_user\_items\_per\_order – This is the average items added to the basket for every order

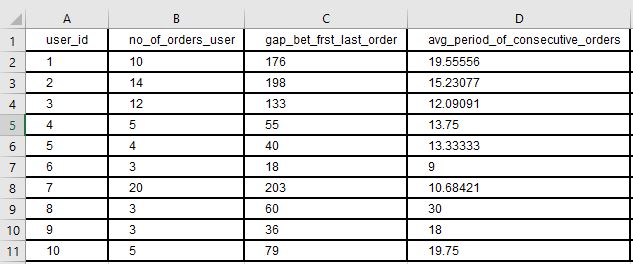


Figure 10: User Predictors.1

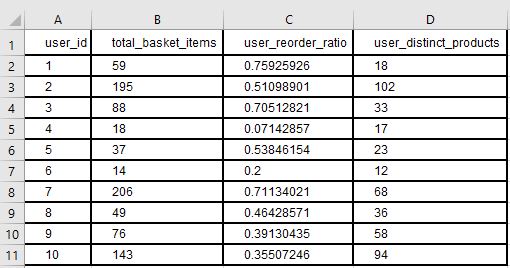


Figure 11: User Predictors.2

User Product Predictors: These variables define how a user behaves to a specific product

1. N\_orders\_product – The total number of times the product was ordered by the user
2. up\_first\_order – This is the first order of the product by the user
3. up\_last\_order - This is the last order of the product by the user
4. avg\_position\_of\_item – This is the average position of the item in the cart
5. rate\_of\_specific\_orders - This is the Percentage of customers’ orders that include a product
6. N\_orders\_sincelastorder – This is the total number of orders since the last time user ordered the product.
7. rate\_of\_specific\_product\_since\_firstorder – This is the Percentage of orders since first order of a product in which a customer bought this product.

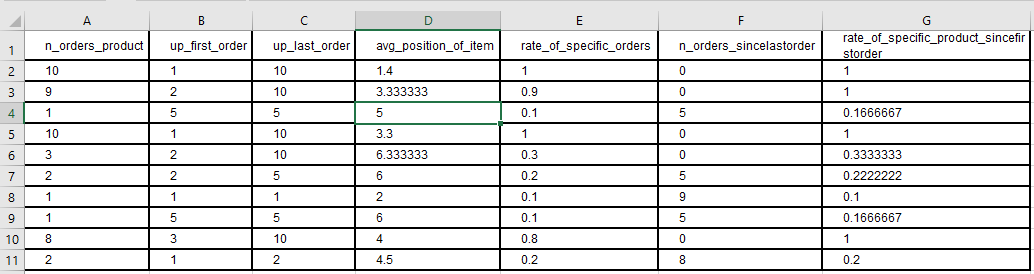


Figure 12: User Product Predictors

### **Creating Test Data for Scoring**

XGBoost is simplest method in evaluating and managing the large dataset efficiently. The original dataset consists of prior, train and test data. So, we are splitting the data into two parts where the prior data is one part, train and test data is the second part. Predictor variables are created using the prior data. After creating the predictor variables, they are added to the prior data set. The updated prior table is joined with the train and test orders data (second part) using Left Join function. With this join function we will able to distinguish and tell the re-ordered products from which the user had already ordered. As prior order set is larger than train and test data together, there are products in the prior data set which are not re-ordered and not there in train and test orders data. As there is no record for those product orders they have been noticed as NA in re-ordered column. One of the advantages of using XGBoost is that it handles the missing values. We have given the NA values of reordered to 0 in the train set to indicate that the products have not been reordered in future order. This re-ordered column is the target variable which we need to find using the XGBoost model. The resulted data now contains only the train and the test data. Again, this data set is divided into two parts which contains train orders as one part and the test orders data as second part. The main reason we are splitting the data so that we can develop the model for prediction using the train orders data and evaluate the predictions based on the test data.

**Creating a Predictive Model**

While creating the model we used training data and if there are any categorical values then they are transformed into null values because it has different values for different variables. So, XGBoost only accepts numerical values. We have created the boosted trees in this model where the data can be viewed as a whole set rather than an individual. To reduce the mis-classification rate the trees are generated one after the another. The parameters are the undetermined part where we need to specify the value which is necessary to predict the model or leave it as a default. The following are the parameters:

* Objective: We have specified the regression logistic i.e., reg:logistic which classifies the problem with the decision.
* eval\_metric: We have specified the logloss which indicates the negative log-likelihood which is used for validating the data.
* eta: It is the values which make the model more robust by shrinking the weights on each step. It can also be referred to as the learning rate and ranges from 0 to 1. In this model, we have given the value as 0.1.
* max\_depth: It refers to the size or depth of the decision tree. We used the default value i.e., 6.
* min\_child\_weight: It is the minimum sum of instance weights required in a child. The value we have specified is 10 because the larger the value the more changes will be in the algorithm.
* gamma: The value for gamma is given as 0.70 which specifies the minimum loss reduction which is required to make the split. The values are tuned when it has a positive reduction in loss function.
* Subsample: Subsample is used to prevent overfitting. As we have mentioned 0.76 as the value where XGBoost randomly samples ¾ of the data prior to growing trees and it happens only once in every boosting iteration.
* colsample\_bytree: It is a subsample ratio where the fraction of columns is constructed for each tree.
* Alpha and lambda: These two are regularization term on weights and can be used in case of higher dimensions which makes the algorithm run faster.

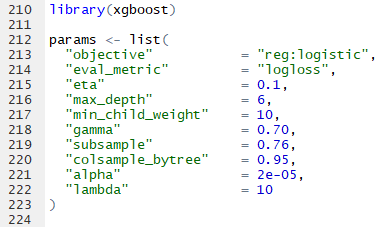


Figure 13: Predictive Model Parameters

After assigning values to the parameters we create a matrix X from the train data set. This matrix X is in executable model for the XGBoost Algorithm. By using this matrix X, we clearly mention the target variables, Input variables and the rejected variables while training the model.

XGBoost is not only used for predicting it also examines the importance of each feature in the dataset and visualize the result as a bar graph. The features are arranged in order according how many times they appear.

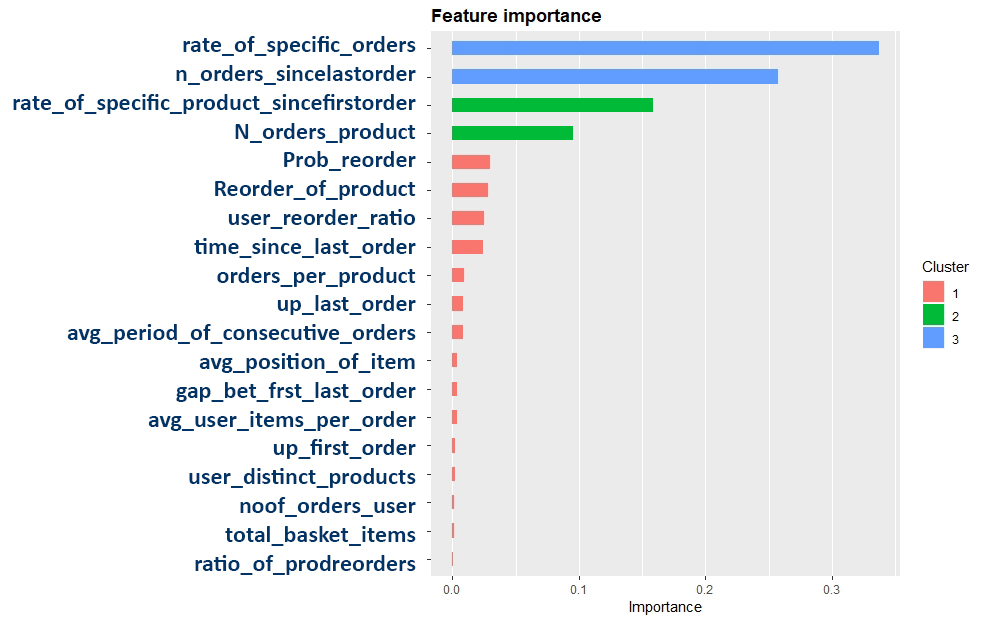


Figure 14: Variable Importance

### **Applying the predictive model on the scoring data.**

After creating the XgBoost model using the train data, we have applied the model to test data to score the model. Here in this step, we need to create a matrix X again for the test data which is in the executable format for XGBoost algorithm. As we mentioned in creating predictive model step, the X matrix is used to define the variables.

**X <- xgb.DMatrix(as.matrix(test %>% select(-order\_id, -product\_id)))**

Here X refers to name of the xgb.DMatrix.

Now we applied the prediction model to this matrix.

**test$reordered <- predict(model, X)**

By doing this, we are predicting the recorded variables for the test data. The model gave us estimated probabilities for reorder of every product in the test data. We decided to apply a threshold limit to the predictions to get more accurate results. We decided to consider every prediction with a probability over 0.21 as a reordered product.

A final table named is created as the output when we run the test data matrix on the prediction model. The table contains predictions of reordered products for the top 10 orders considered, order\_id is displayed in the first column of the table and different products (product\_id) are displayed in second column of the table adjacent to their respectable order\_id. We also created another table which contains orders which have no products reordered according to our prediction model.

We got the results in the final table, but we don’t have response values for the observations. This make evaluating the predictions of test data impossible. So, we decided to use to re-apply the prediction model on the train data which was used to create the model and compare our predictions to actual results. We followed similar steps as we used for the test data while re-applying the model to train data, the only change is we made sure that we get actual reorders in another column. As result we obtained the table containing order\_id, real\_products (actual reordered products) and products (predicted re-ordered products).

## **Apriori Algorithm**

### **Creating Transaction Data**

We can only use Binary Incidence Matrix or Transaction Matrix to execute Apriori algorithm (D. V. S. Kumar & Renganathan, 2011). The Binary Incidence Matrix is pulled out after combining the data sets of order\_products\_prior and products by using the left join function. From that, we take the order\_id, product\_name columns and with the use of pivot tables in excel we prepare Binary Incidence Matrix.

### **Creating and Executing Apriori Model**

The model is executed using R studio. The necessary library packages required to build the model are arules, arulesViz, plotly, IRdisplay, dplyr. Some of these packages won’t come by default. The generated rules are useful if the value of the rules is greater than one. The result after executing will display the rules with the defined threshold limit of support and confidence.

**Support: -** If we consider a rule C->D where C and D is antecedent and consequent respectively. The support tells us the number of times the product C and D bought together in the transactions.

**Confidence: -**Confidence defines the probability of ordering product D when product C is ordered.

Association Rules generated with support = 0.001, confidence = 0.4 and lift > 1



Figure 15: Threshold Values in Apriori Model

# Results and Discussion

## **XGBoost Result**

As said before, the final table obtained after applying the model to train data contain order\_id, real\_products and products. We added actual reorders to compare them directly with the predictions we obtained from the model.

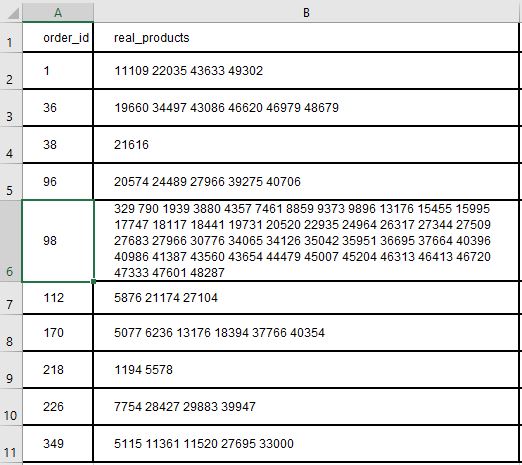


Figure 16: Re-ordered Products

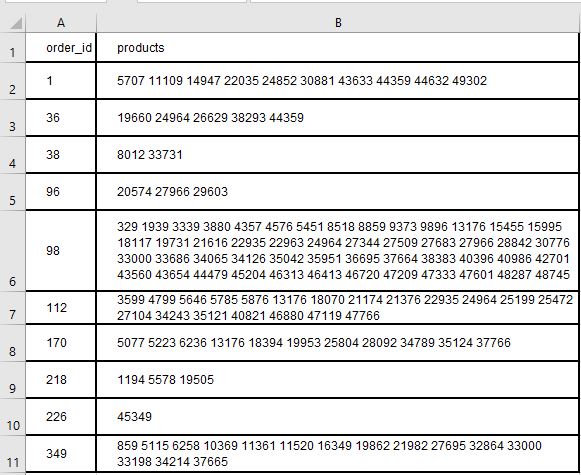


Figure 17: Predicted Products

From the table, we have observed that the results were not completely accurate. Most of the actual reorder products were predicted by our prediction model. So, we can say that the accuracy of the model is highly competitive, and the prediction model can be used.

## **Apriori Result**

Association Rules generated with threshold values of support = 0.001, confidence = 0.4

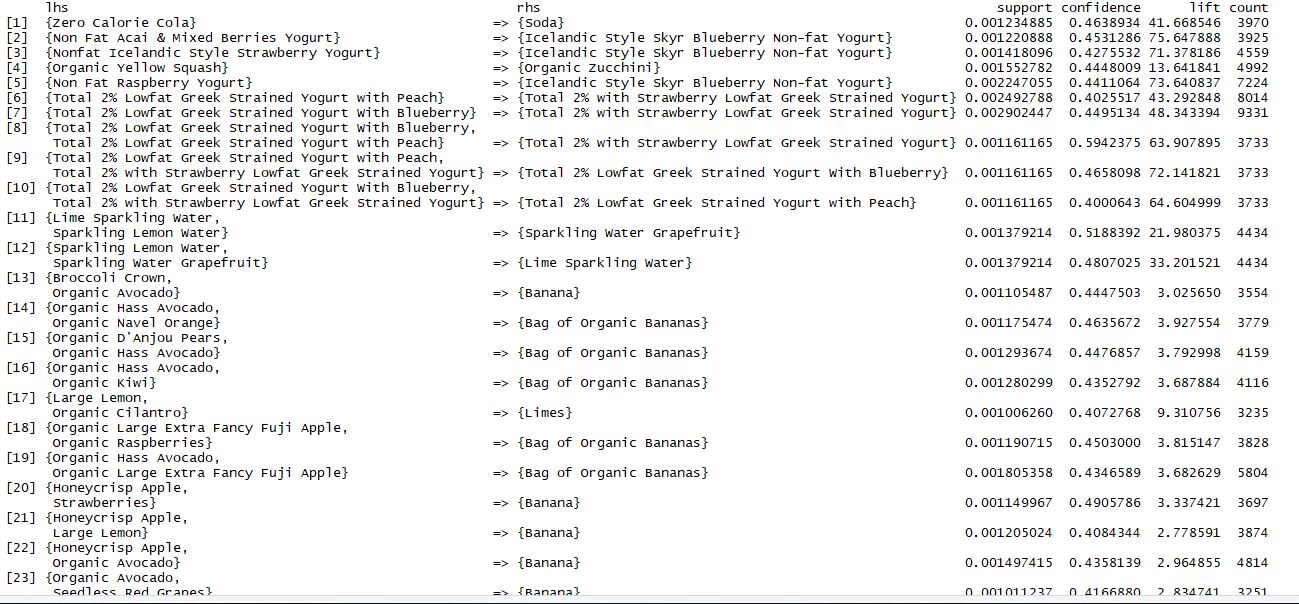


Figure 18: Rules generated

When you consider the first rule, the antecedent is Zero Calorie Cola and the Consequent is Soda. It means, the customer who bought Zero Calorie Cola might also willing to but Soda. Form the above table the rules generated with lift value greater then can be considered as recommendations. From all the rules with lift grater than one, we can tell that when a customer buys a product in the antecedent list he might also willing to buy the respective product from the consequent list. The rules generated will help product owners to give those as recommendations. In e-commerce website you can display the recommended products below the product selected. And, in retail store you can arrange the products closely based on the recommendations.

# Practical Recommendations and Conclusion

Advancement in analytics has led to massive growth in technology trends.

* First recommendation would be analyzing customer behavior for retailers. It is very important to predict the customer buying patterns in the store. For such analysis retailers are using big data to gather the information. To develop and execute the future profits of the company.
* In-store efficiency improvements are very important. It will help us to predict the traffic in stores. Also, outfit the stores with smart sensors and Wi-Fi to track how many customers are entering the store and near which products/ Aisles they are spending more time. For an instance, during the busy hours at store there might be long queues at the billing. They might install the self-checkouts to make the shopping experience more comfortable and enjoyable.
* It is always better to re-stock the bestsellers immediately instead of being out-of-stock for the long time. Customer behavior analysis will help in doing it quicker and efficiently.
* Personalization adds up a new shopping experience for the customers. For example, if a user is searching for an item and finds expensive in the store. In such situations our algorithm would be an added advantage, they can simply visit the online app/website and look for the better recommendations for much lesser price. This shows the customer loyalty and identifies their most important customers. This data would help in recognition of that customer and greeting them in the store would grab more attention.
* Improvements in Supply chain efficiency with the help of data analytics is a fruitful result. This helps to maintain clear logistics of the store. Detect inefficiencies from manufacturers and retailers to see the returns for damaged and returned products. Data collected from logs, machines, and other sources can highlight outliers and detect weakness and complications in the retailer supply chain.
* Pricing of the products plays key role. IT department can use different technologies such R, SAS, oracle to see the effect of price change for a demandable product. Retailers globally can check the consumer responses to prices. It helps to determine the correlation of new price has brought profits or loss.
* Soon Artificial intelligence will be given a completely different view of retail experience. With the drastic change in technology, retailer will implement machine learning to find out which products the customers are going to purchase with reference to historic and real-time behavioral data. Then a bot might engage the customer with other recommendations. The other interesting fact of artificial intelligence is sentimental analysis. This helps detailed overview of the product like how it is seen which implies feelings like good or bad light etc., which is developed by machine learning algorithms to process and determine the context on social media.

**Conclusion**

In this paper, we predicted which products the customer will reorder, purchase for the first time or add to their cart in the next order. We identified which previously purchased products will be in user’s next order. As a result, we identified the products which are likely to be reordered by the customers and demographics, such as which day at what time we can expect customer traffic in the store. Through Apriori Algorithm we studied purchasing patterns to provide recommendations to the customers. This research focused on the importance of customer relationship management in retail industry. We studied how Customer Relationship Management would help retailers boost sales and improve business. To survive in today’s competitive market customer satisfaction is very essential. This study helped us understand how to connect with the customers, increase productivity, retain customers and enhance business. Further research scope would involve having more enhanced features like including artificial intelligence system to better understand customer needs, navigate retail sites quicker and improve inventory management.

# References

Aguinis, H., Forcum, L. E., & Joo, H. (2013a). Using Market Basket Analysis in Management Research. *Journal of Management*, *39*(7), 1799–1824. https://doi.org/10.1177/0149206312466147

Aguinis, H., Forcum, L. E., & Joo, H. (2013b). Using Market Basket Analysis in Management Research. *Journal of Management*, *39*(7), 1799–1824. https://doi.org/10.1177/0149206312466147

Anderson, J. L., Jolly, L. D., & Fairhurst, A. E. (2007). Customer relationship management in retailing: A content analysis of retail trade journals. *Journal of Retailing and Consumer Services*, *14*(6), 394–399. https://doi.org/10.1016/j.jretconser.2007.02.009

Anhtony, G., & Hamner, B. (2017). *Kaggle: Your Home for Data Science*. Retrieved from https://www.kaggle.com/c/instacart-market-basket-analysis/kernels

Chen, M., Chiu, A., & Chang, H. (2005a). Mining changes in customer behavior in retail marketing. *Expert Systems with Applications*, *28*(4), 773–781. https://doi.org/10.1016/j.eswa.2004.12.033

Chen, M., Chiu, A., & Chang, H. (2005b). Mining changes in customer behavior in retail marketing. *Expert Systems with Applications*, *28*(4), 773–781. https://doi.org/10.1016/j.eswa.2004.12.033

Chen, T., & He, T. (2019). *xgboost: eXtreme Gradient Boosting*. 4.

Cohn, M. (2010). *Succeeding with Agile: Software Development Using Scrum*.

Drew, J. H., Mani, D. R., Betz, A. L., & Datta, P. (2001). Targeting Customers with Statistical and Data-Mining Techniques. *Journal of Service Research*, *3*(3), 205–219. https://doi.org/10.1177/109467050133002

Huotari, K., & Hamari, J. (2011). *“Gamification” from the perspective of service marketing*. 9.

Kaur, M., & Kang, S. (2016). Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining. *Procedia Computer Science*, *85*, 78–85. https://doi.org/10.1016/j.procs.2016.05.180

Kumar, D. V. S., & Renganathan, D. R. (2011). *Consumer Buying Pattern Analysis using Apriori Association Rule*. 10.

Kumar, S., & Arora, R. K. (2015). Analyzing Customer Behaviour through Data Mining. *International Journal of Computer Applications Technology and Research*, *4*(12), 884–888. https://doi.org/10.7753/IJCATR0412.1002

Mu, jiankang. (2013). *Application of Apriori Algorithm to Customer Analysis*.

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*(2), 2592–2602. https://doi.org/10.1016/j.eswa.2008.02.021

Oliveira, V. L. M. (2012a). *Analytical Customer Relationship Management in Retailing Supported by Data Mining Techniques*. 178.

Oliveira, V. L. M. (2012b). *Analytical Customer Relationship Management in Retailing Supported by Data Mining Techniques*. 178.

Olsson, M., Hogberg, J., Wastlund, E., & Gustafsson, A. (2016). In-Store Gamification: Testing a Location-Based Treasure Hunt App in a Real Retailing Environment. *2016 49th Hawaii International Conference on System Sciences (HICSS)*, 1634–1641. https://doi.org/10.1109/HICSS.2016.206

Patel, M., Karvekar, S., & Mehta, Z. (2014). *CUSTOMER BEHAVIOR MODEL USING DATA MINING*. (01), 8.

Prinzie, A., & Van den Poel, D. (2006). *Incorporating sequential information into traditional classification models by using an element/positionsensitive SAM.*

Raorane, A., & Kulkarni, R. V. (2011a). Data Mining Techniques: A Source for Consumer Behavior Analysis. *International Journal of Database Management Systems*, *3*(3), 45–56. https://doi.org/10.5121/ijdms.2011.3304

Raorane, A., & Kulkarni, R. V. (2011b). Data Mining Techniques: A Source for Consumer Behavior Analysis. *International Journal of Database Management Systems*, *3*(3), 45–56. https://doi.org/10.5121/ijdms.2011.3304

Rodpysh, K. V. (2012a). Applying Data Mining in Customer Relationship Management. *International Journal of Information Technology, Control and Automation*, *2*(3), 15–25. https://doi.org/10.5121/ijitca.2012.2302

Rodpysh, K. V. (2012b). Applying Data Mining in Customer Relationship Management. *International Journal of Information Technology, Control and Automation*, *2*(3), 15–25. https://doi.org/10.5121/ijitca.2012.2302

Singh, N., Bellathanda Kaverappa, C., & Joshi, J. D. (2018). Data Mining for Prevention of Crimes. In S. Yamamoto & H. Mori (Eds.), *Human Interface and the Management of Information. Interaction, Visualization, and Analytics* (Vol. 10904, pp. 705–717). https://doi.org/10.1007/978-3-319-92043-6\_55

Victor, H. A., Abimbola, O. O., Mercy, O. F., Esther, O. O., & Eloho, I. P. (2014). *Customer behaviour analytics and data mining*. 9.