

**DATA MINING**

**CUTOMER BEHAVIOR ANALYSIS THROUGH DATA MINING**

**SUBMITTED TO**

Dr. Neetu Singh

**SUBMITTED BY SUBMITTED ON**

Rohith Madamshetty 12/13/2018

[**rmada2@uis.edu**](mailto:rmada2@uis.edu)

**Abstract**

The world we know now is digital and every store is either moving or moved online. Catching the customer’s eye has become very important for any online store. To understand such customer’s behavior online, we applied data mining techniques to understand patterns and trends that customers tend to follow. We used 3 supervised - Regression analysis, decision trees, memory-based reasoning and 1 unsupervised – association rules techniques of data mining to predict and classify the rating and recommendations of a restaurant review application. Doing this we not only were able research on the techniques but also apply it to a real-time scenario. Results and findings showed how to predict the rating of a restaurant within a given location in India and recommend organizations what kind of cuisines are customer likely to prefer.

Table of Contents

[**1.** **Problem Statement** 3](#_Toc532507701)

[**I.** **Introduction** 3](#_Toc532507702)

[**II.** **Problem Addressed** 3](#_Toc532507703)

[**III.** **Research Questions** 3](#_Toc532507704)

[**IV.** **Literature Review:** 3](#_Toc532507705)

[**2.** **Methodology** 4](#_Toc532507706)

[**I.** **Data Source and Data Set** 4](#_Toc532507707)

[**II.** **Data Cleaning** 5](#_Toc532507708)

[**III.** **Data Preparation** 5](#_Toc532507709)

[**IV.** **Data Visualization to explore the dataset** 6](#_Toc532507710)

[**V.** **Reasoning for Model Selection** 7](#_Toc532507711)

[**VI.** **Comparison of each supervised technique and choosing the Best model** 8](#_Toc532507712)

[**i.** **Decision Trees** 8](#_Toc532507713)

[**ii.** **Regression** 12](#_Toc532507714)

[**iii.** **K-Nearest Neighbor** 13](#_Toc532507715)

[**VII.** **Exploratory Analysis using Association Rule Mining: -** 14](#_Toc532507716)

[**3.** **Best-chosen Model for our research question** 16](#_Toc532507717)

[**I.** **The best-chosen model for data analysis** 16](#_Toc532507718)

[**II.** **Scoring the Best Model** 16](#_Toc532507719)

[**III.** **Results using Data Visualization** 17](#_Toc532507720)

[**IV.** **Detailed Exploratory Analysis of Unsupervised Association Rule Mining: -** 18](#_Toc532507721)

[**4.** **Result Discussion** 20](#_Toc532507722)

[**5.** **References** 21](#_Toc532507723)

# **Problem Statement**

## **Introduction**

“Everything starts with the customer” –Lou Gerstner. In the digital world, that quote can be changed to, everything starts with the customer reviews. Approximately 95% of customers read online reviews before making a purchase. (Maslowska, Malthouse, & Bernritter, 2017). The purchase likelihood of a product increases by 15% when costumers tend to prefer verified buyer reviews over anonymous reviews. (Maslowska, Malthouse, & Bernritter, 2017). Categorizing these reviews and the patterns customers rate a service can help us understand customer behavior and understanding customer behavior can help build customer retailer relationship (Victor, Abimbola, Mercy, Esther, & Eloho, 2014). Data Mining tools are best suitable for classifying and predicting such patterns, this research paper uses both supervised and unsupervised techniques to mine the necessary data.

## **Problem Addressed**

The research problem here is important in terms of understanding customer needs, expectations which help in improving the relationship between customer and service providers. Our Research will help new aspiring restaurant start-ups who want to start their new restaurant in new preferable location, decide on price range and it will also help already established restaurant chains across India to add new cuisine in the menu card, also give additional or improved services based on the reviews provided by customers.

## **Research Questions**

* + Predicting on what requirements the restaurant should be established to be successful?
  + Predicting customer satisfaction based on the restaurant available in a location?
  + What better services can be given to customers for achieving high customer relationship?

## **Literature Review:**

Customer or consumer behavior can be defined as a study of a group of individuals, companies and the pattern they use to select products, services to satisfy their needs and the impacts these patterns in turn have on them(Hawkins & Mothersbaugh, 2013). These patterns or processes in which they select may vary depending on various factors such as online reviews, demographic locations, for example the Zomato dataset which we have selected was previously used for recommending best restaurants in a certain location also it has been used for user-based collaborative filtering by which users are able to rate different items (Saha & Santra, 2017). Some more research on the same dataset as observed on “Kaggle.com” were done on following topics -

1. Average Ratings on Countries
2. Top 15 Cuisines on Zomato
3. Restaurants by Price Range on Zomato
4. Top-Rated Cuisines on Zomato

Now a day’s social media is the biggest influencer and medium to express the opinion for all the customers also media and huge volumes of user comments are influential while creating a model for predictive analysis. The validation of the humongous amount of reviews is challenging because there can be bots which can eliminate all the negative comments and fake the customer reviews. However, using supervised methods, we can classify ratings as well as comments based on their authenticity, accuracy, and validity (Hemmatian & Sohrabi, 2017). Moreover, this data set and research problems have been used for developing the opinion mining application using implicit approach (Chandankhede, Devle, Waskar, Chopdekar, & Patil, 2016).

# **Methodology**

We have followed the SEMMA approach to analyze all the data mining technique to avoid any ambiguous situations (Singh, Bellathanda Kaverappa, & Joshi, 2018).

* **Sample:** Since out dataset is too large we have used sampling techniques. We have partitioned the data using data partition node and used the default 40 for train, 30 for validate and 20 for test.
* **Explore:** We have carried out both explanatory and exploratory analysis of the data by visualization to learn about the descriptive statistics and understand how one variable is related to the other.
* **Modify:** We have transformed the data set by eliminating variables that were did not help in data modelling. For example, we removed address, locality, locality verbose since this information could be retrieved from the longitude and latitude variables. This helped in increasing the accuracy of the model. We have created transaction matrix and binary incidence matrix to work on association rules.
* **Model:** We used supervised data mining techniques like Decision tress, Regression trees and Memory based reasoning and developed predictive models to answer our research question. We also used unsupervised data mining technique, association rules to make recommendations to the customers.
* **Asses:** We finally assessed the models by scoring the data and thus analyzed the actual and predicted ratings of individual restaurants in a given location

## **Data Source and Data Set**

The data set is about Zomato app which is very famous in India. We found the dataset from Kaggle.com It is a restaurant review application which shares customer feedback. It also has all the details of the restaurants which are collaborated with Zomato in different countries. The data set contains details of the locations of the restaurants, types of cuisines, price of the menu, ratings of different restaurants and all the delivery options. It consists of both structured and unstructured data making research gnarly and fascinating. The total number of variables in this dataset was 21 out of which 4 are Categorical, 7 are numerical and 9 are text variables. But have considered only 15 to execute the whole project 4 categorical, 7 numerical and rest text variables.

## **Data Cleaning**

Data Cleaning process began with exploring the data set and eliminating the variables that were not required (Erhard & Hong Hai Do, 2000). Next, deleted all the null values and removing all the extra spaces found in the entire dataset. Sql commands were run using the SQL developer to delete all the duplicate values. Additionally, special characters were removed and replaced. Stat Explorer helped to determine the skewness and kurtosis of the variables. The kurtosis and skewness levels were minimized so that they fall in the acceptable range of -3 to +3.

## **Data Preparation**

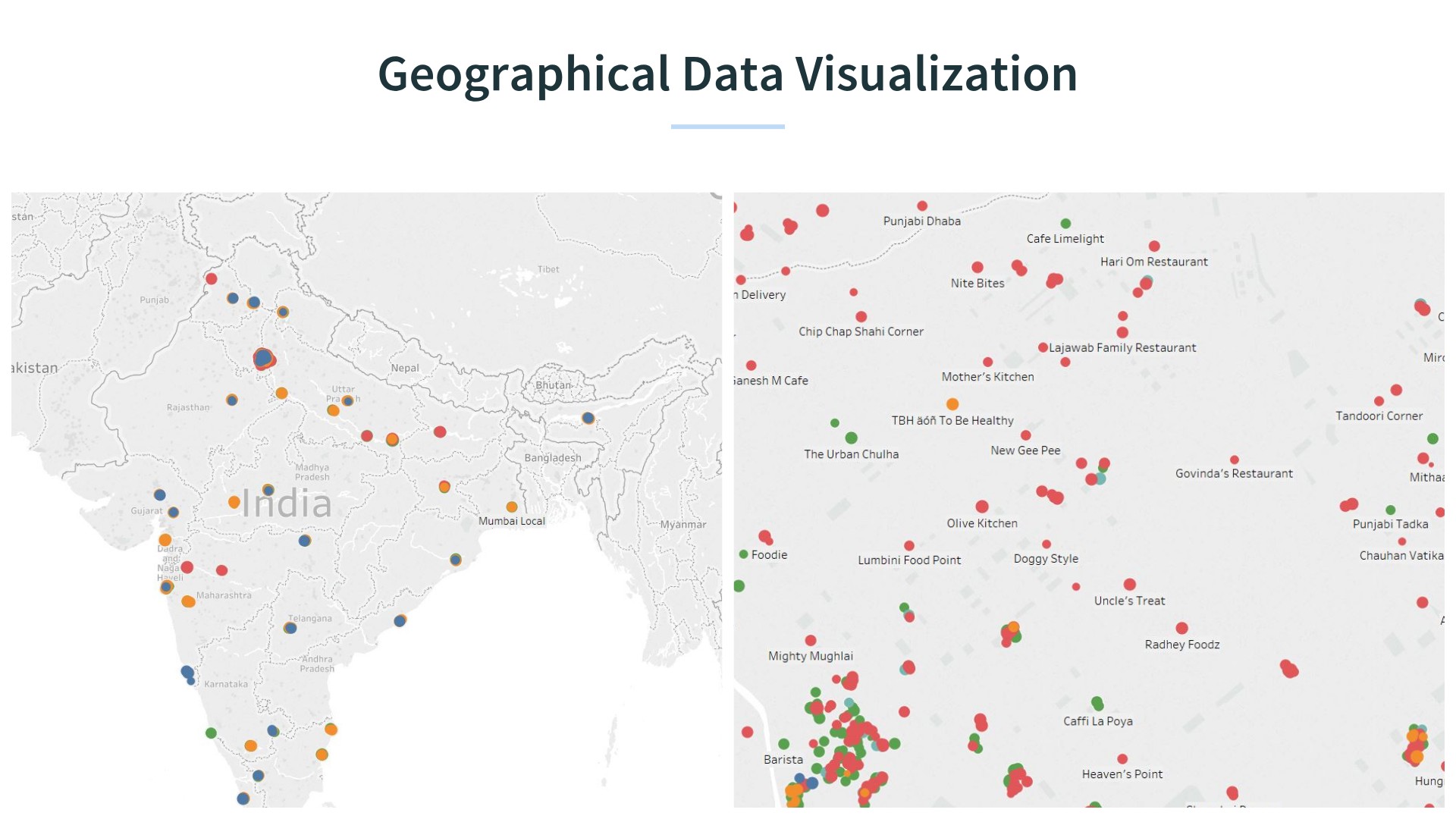
We have partitioned the data into score and train data. We have taken 80% of sample for the Train dataset and 20% sample for the score data set. The target variable assumed was first interval variable that was aggregate\_rating but later it was changed to a categorical variable rating\_text. There was an observation that the rating\_text and rating\_color was highly co-related which could have led to inappropriate results. Hence, the rating color was eliminated. The rating\_text was divided into 5 categories based on which prediction was done and research questions were answered. After data cleaning the following variables were considered to answer research questions.(Singh et al., 2018)

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Variable Type**  **(Numerical, Text, Categorical)** | **Variable Description** |
| Restaurant ID | Numerical | Id (primary key) |
| Restaurant Name | Text | Name of the Restaurant |
| City | Text | Name of the city |
| Longitude | Numerical | Location of the Restaurant |
| Latitude | Numerical | Location of the Restaurant |
| Cuisines | Text | Variety of Dishes |
| Average Cost for two | Numerical | Price |
| Has Table booking | Categorical | Reservation |
| Has Online delivery | Categorical | Door delivery |
| Is delivering now | Categorical | Status of delivery |
| Price range | Numerical | Different range of price |
| Aggregate rating | Numerical | Rating of the restaurant |
| Rating color | Text | Indication of rating |
| Rating text | Text | Comments on rating |
| Votes | Numerical | No. of likes and dislikes |

Figure 1: Dataset Variables

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

We used Tableau to explore and visualize our data set (Murray, 2013). We visualized the dataset with help of different graphs like bar graph, pie chart, bubble charts (Murray, 2013). We compared the restaurants that had good aggregate\_rating in different locations using geographical data visualization by using color, size. Color was used to represent best restaurants (rating\_text) and size was used to determine the restaurants with highest rating (aggregate\_rating). We also took votes into consideration and visualized using bar graphs to see which restaurants had maximum votes. Scatter plot was examined to identify any outliers. We used bubble charts to examine how the different cuisines are associated. Given below is restaurants in India based on their aggregate\_rating and rating\_text.

****

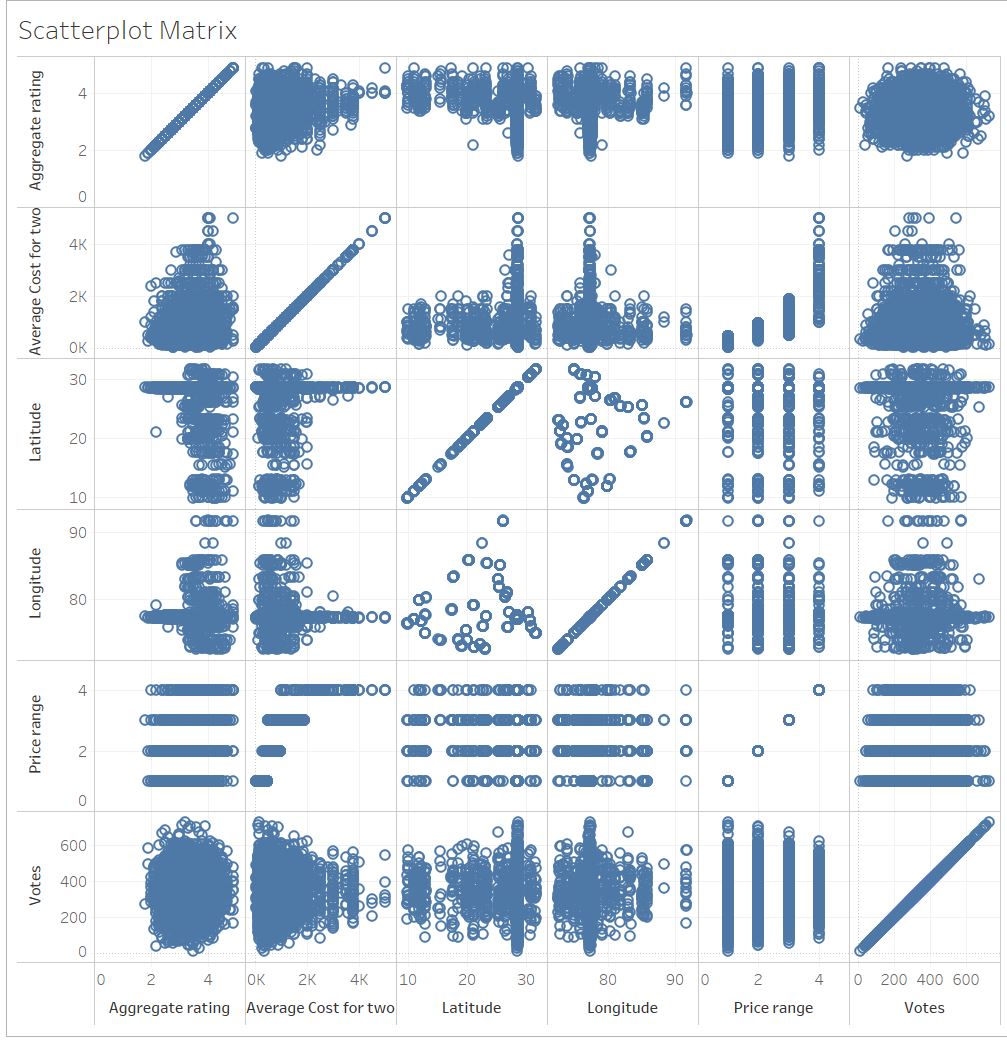


Figure 2 : Data visualization with Tableau

## **Reasoning for Model Selection**

**Decision trees:** This was helpful to predict customer satisfaction based on the restaurant available in a location. Classification trees helped to classify the prediction based on the target variable rating\_text which was categorical variable. Based on the price range, the location and other parameters, classification decision tree helped to predict the model in a better manner.

**Regression:** This method was used for forecasting and finding out cause and effect relationship between variables. Using regression, we were able to predict of the value of rating\_text using the independent variables

**Memory Based Reasoning or K-nearest neighbor:** Our dataset had real-time data also most of the data did not follow the theoretical assumptions like regression analysis. Since MBR it is non-parametric data-driven method, it was used for both classification and prediction

**Association:** To understand the relationships between different types of cuisines and make recommendations to the entrepreneurs, we used association rules and found the underlying hidden patterns between the items in the transaction data.

## **Comparison of each supervised technique and choosing the Best model**

### **Decision Trees**

To Predicting customer satisfaction based on the restaurant available in a location we have used Classification Decision Tree (Kumar & Arora, 2015). To answer this research question, we have used the target variable as rating\_text. This tree also helps us to find out the variable importance which will help in classification. We are using the Zomato cleaned data set to answer our research question one***.*** We have used B2D6, B2D4, and B3D6 for model comparison. Since the target variable is Rating\_text which is a categorical variable the recursive partition algorithm of decision tree will develop a Classification tree which will help us to classify and predict the customer satisfaction based on the Rating\_text***.***

The initial roles and levels for the Zomato train data set were as follows:

|  |  |  |
| --- | --- | --- |
| **Name** | **Role** | **Level** |
| Restaurant ID | ID | Interval |
| City | Rejected | Nominal |
| Longitude | Input | Interval |
| Latitude | Input | Interval |
| Cuisines | Text | Variety of Dishes |
| Average Cost for two | Input | Interval |
| Has Table booking | Input | Binary |
| Has Online delivery | Input | Binary |
| Is delivering now | Input | Binary |
| Price range | Input | Nominal |
| Aggregate rating | Rejected | Interval |
| Rating color | Rejected | Nominal |
| Rating text | Target | Nominal |
| Votes | Input | Interval |

Figure 3- Initial roles and levels of Variables

Further, when dataset is explored using the target variables we analyzed the following results:

|  |  |
| --- | --- |
| Number of Average Rating Restaurants? | 2864 |
| Number of Excellent Rating Restaurants? | 42 |
| Number of Good Rating Restaurants? | 1308 |
| Number of Poor Rating Restaurants? | 153 |
| Number of Very Good Rating Restaurants? | 432 |
| The majority class for your Rating\_text variable in the Zomato data set | Average |
| The percentage of the majority class. That is the baseline accuracy. | 4799/ (4799+2864) =62.61~ 63% |
| Baseline misclassification rate | 100-63=37% |

Figure 4- Target Variable Exploration

#### **Results for the Best Model**

**Best Mode1 ClassDecTreeB3D6:**

In this model the maximum branch is set to 3 and maximum depth is set to 6. After running this node let us analyze the Result window which is as shown below:

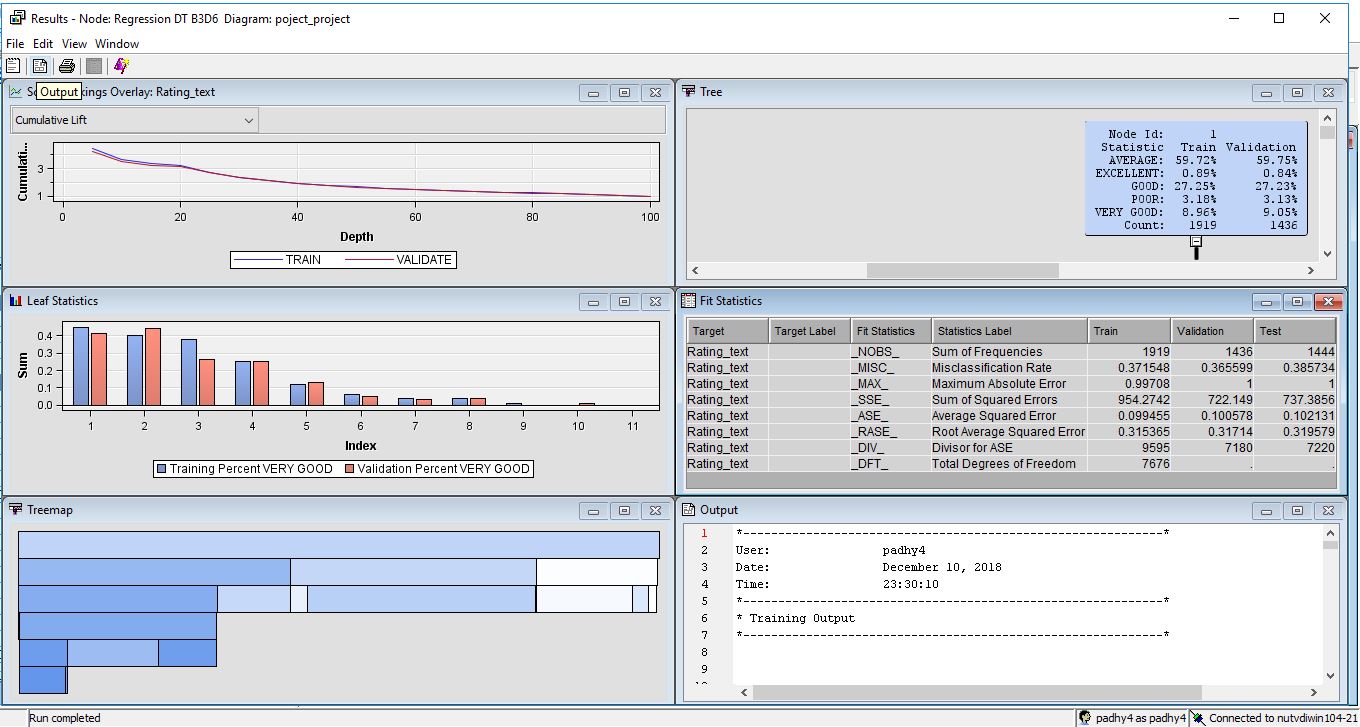


Figure 5 - Result Window of Classification Tree B3D6

|  |  |
| --- | --- |
| **Data Partition** | **Misclassification rate\_(\_MISC\_)** |
| **Training Set** | 0.37 |
| **Validation Set** | 0.36 |
| **Test Set** | 0.38 |

Figure 5 - Fit Statistic of B3D6 Classification Decision Tree

There is very slight overfitting in the classification model as the misclassification rate for the training set (0.37) is performing a little better than the validation set (0.36) which can be neglected. We can also see the Score Ranking overlay graph that training and validation curves are almost the same without much deviation hence overfitting chances are very less. In the leaf statistics window the classification is done based on the variable importance. The variables price\_range, latitude, longitude, has\_table\_booking, has\_online\_delivery are the important variables used to develop this decision tree. The variable with Validation Importance of 1 is the most important variable and the variable with Validation Importance of 0 is the least important variable. Compared to B2D6 and B2D4, B3D6 has the least misclassification rate hence it is chosen as the best model.

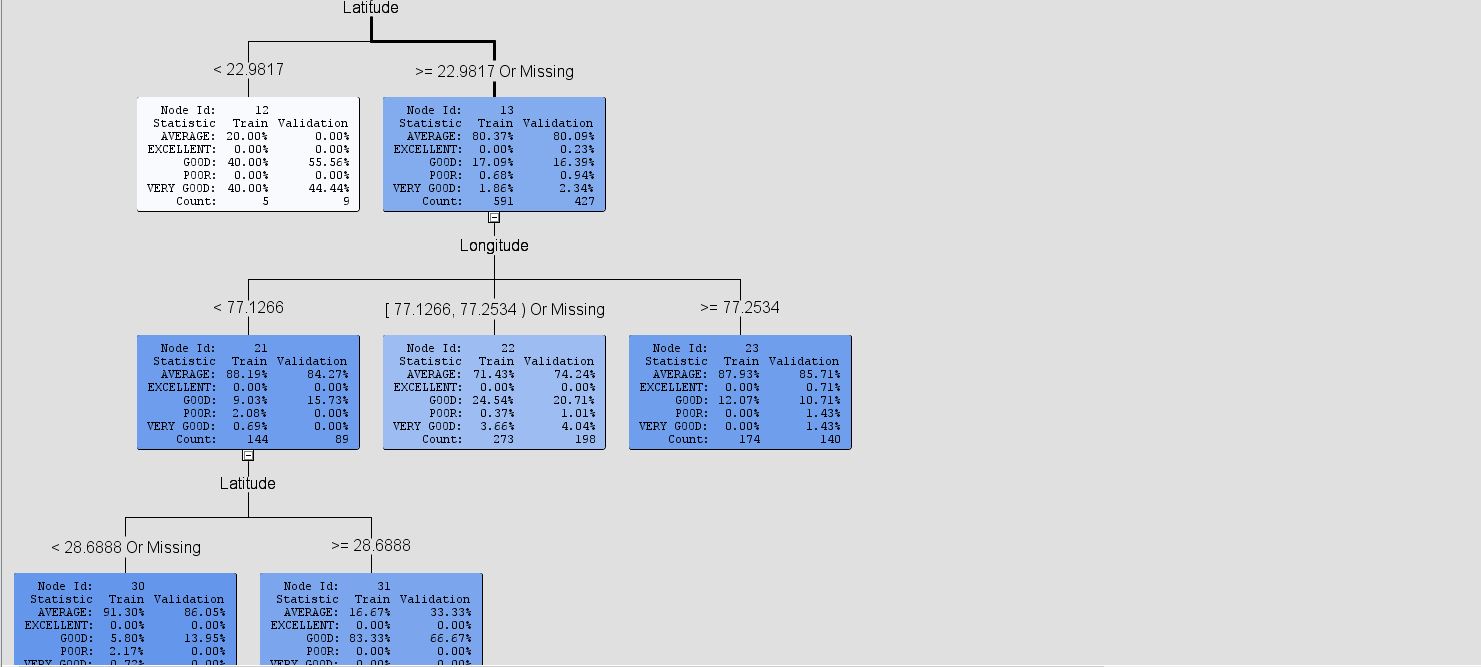
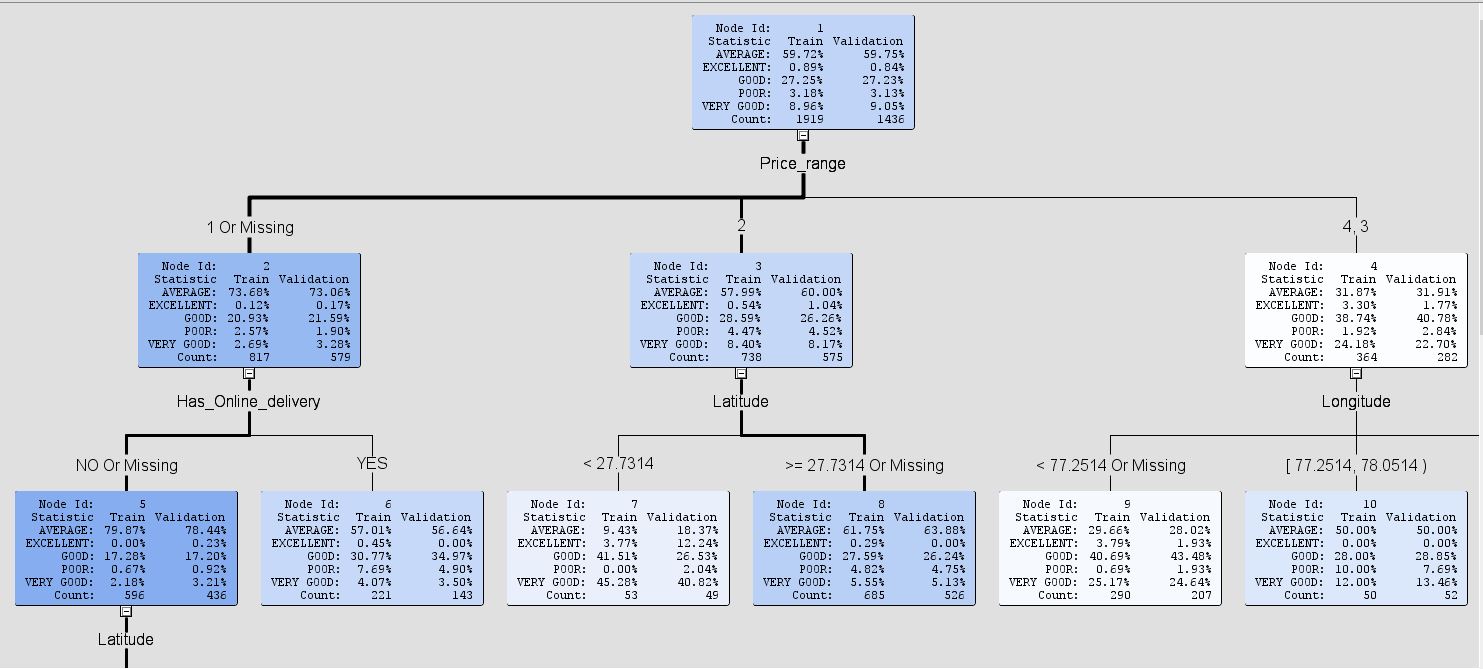


Figure 6- Leaf statistics window for Classification Decision Tree B3D6

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Misclassification rate\_(\_MISC\_)** | | |
| **Classification Decision Tree Models** | **Training Set** | **Validation Set** | **Test Set** |
| **Y B3D6** | 0.371 | 0.365 | 0.385 |
| **B2D6** | 0.376 | 0.368 | 0.386 |
| **B2D4** | 0.376 | 0.368 | 0.386 |

Figure 7: Model Comparison

The key takes away that from this technique is to classify and predict the record with the help of the target variable using classification tree, and finally assess it by scoring the model on the new data set. This was helped to predict customer satisfaction based on the rating in a given location. One strategy to get a better decision tree is to increase the depth of the trees to get better classification and homogeneity.

### **Regression**

This method is mostly used for forecasting and finding out cause and effect relationship between variables. Using regression, it is possible to predict of the value of dependent variables using the independent variables. In regression two or more independent variables are used to predict the value of a dependent variable.

We used 5 types of regression methods-

1. **Exhaustive search**- This algorithm considers of all possible subsets of predictors assessed.
2. **Forward selection**- Forward selection is a very attractive approach, because it's both tractable and it gives a good sequence of models. It starts with no predictor variables and adds the predictors one by one. This algorithm stops when the contribution of additional predictors is not statistically significant.
3. **Backward elimination**- This algorithm starts with all the predictors at once and then at each step it eliminates the least useful predictor which is least statistically significant.
4. **Stepwise** - This algorithm works same as forward but at each step we consider dropping predictors that are not statistically significant same as the backward regression.
5. **Interaction** - The value of "Two-factor Interactions" property is set to YES in the Interaction Node properties just to see the dependency interaction between the 2 or more variables which are related to target categorical variables.

#### **Results for the Best Model**

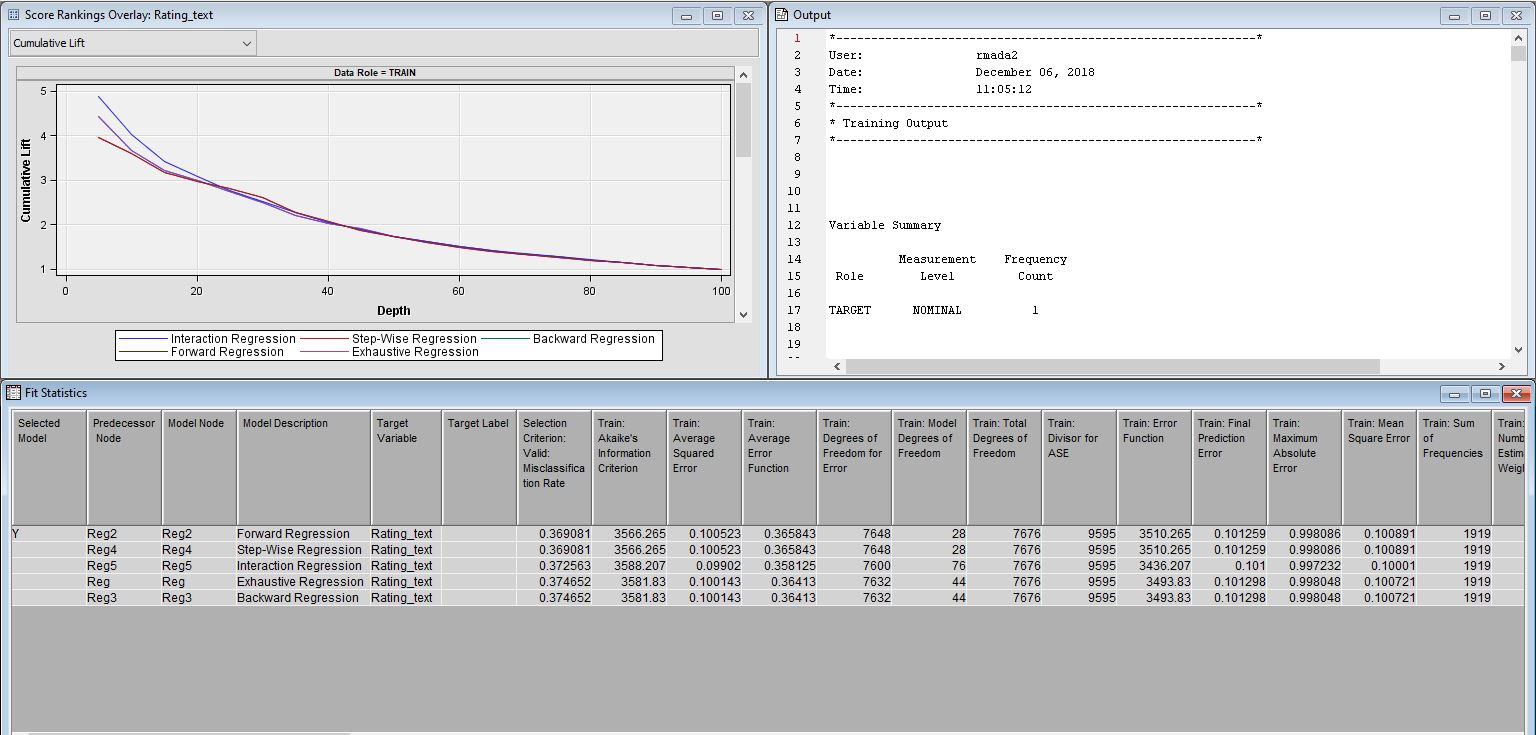


Figure 8: Result Window for Best Chosen Model in Regression

The best model chosen by model comparison node was Forward regression with Misclassification rate - 0.369081 and ASE-0.1

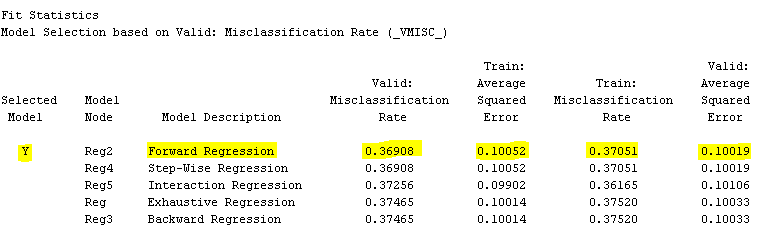


Figure 9: Fit Statistics Window for Regression

Based on this analysis we can say that model is performing best at validation data set.

### **K-Nearest Neighbor**

As our data consist of the real-world data where most of the data does not follow the theoretical assumptions like regression analysis. Therefore, KNN algorithm is one of the best classification algorithms used as there was very little knowledge we had for distribution of the data, also it is non-parametric data-driven method. It can be used for both classification and prediction. We chose all the positive integers for ‘K’ along with the samples. And then we selected K entered for our database which are close to the new sample. We are using Hierarchical KNN because our dataset is large, and it is complex for computation. Also, we have approximate value of k (16). Basic goal was to divide number of samples in pre- determined number of k (i.e.7,8,9,16) of non-overlapping clusters so that our clusters becomes homogeneous.

However, adding more clusters beyond 16 brings less improvement to cluster homogeneity for centroid clustering because when we entered k=17 then it did shrink the avg distance to the centroid but not as much as k=16. So, we picked values of k just to observe the number of clusters and its distance from centroid, when we got k=16 we stopped. KNN algorithm tries to determine classification of point combines the classification of K nearest neighbors. So, when increased the value of K we can see that MISC and ASE reduced. We got K=16 as optimal value.

#### **Results for the Best Model**

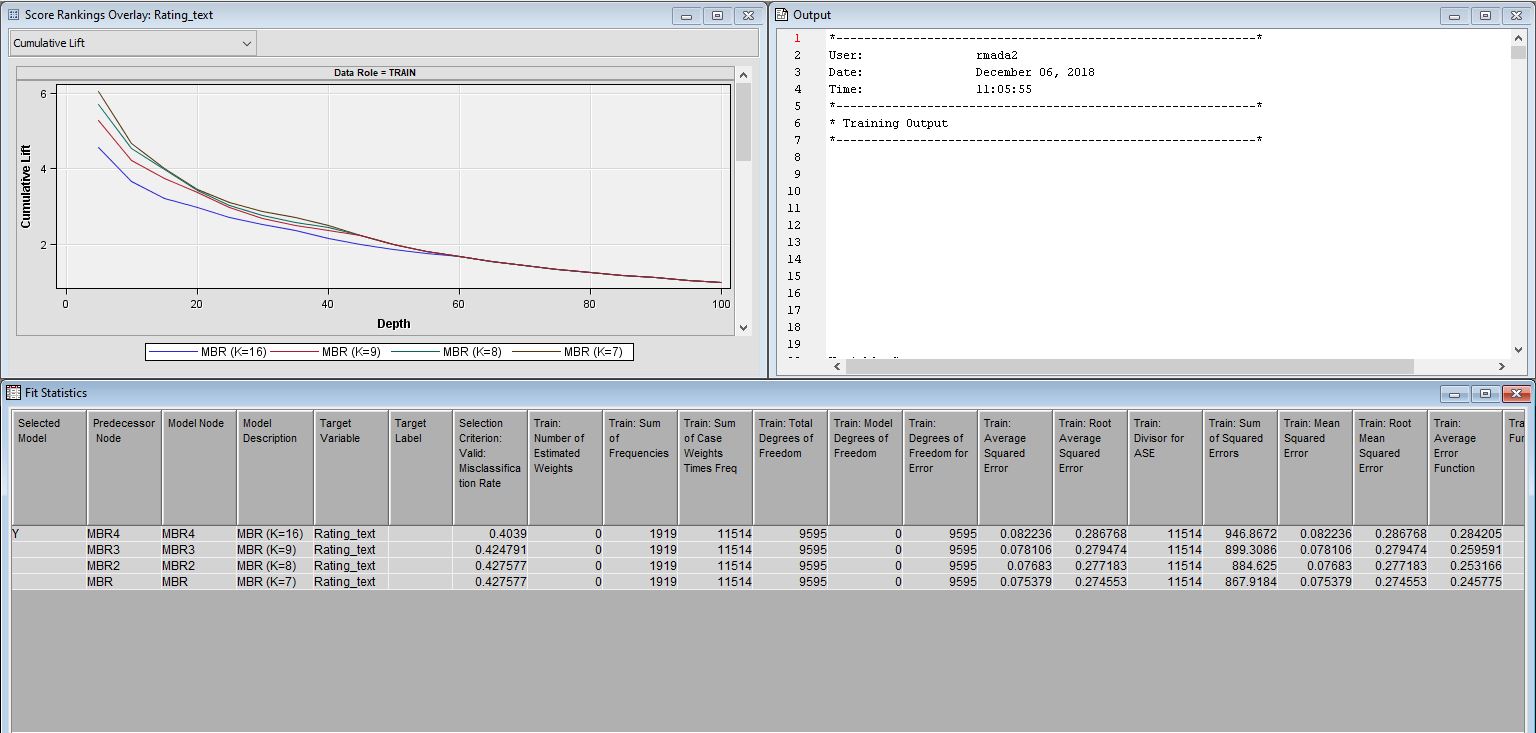


Figure 10: Result Window for MBR

After cross validation we came up with the best value of K=16. It has the least misclassification rate compared to the other nodes- 0.4039.

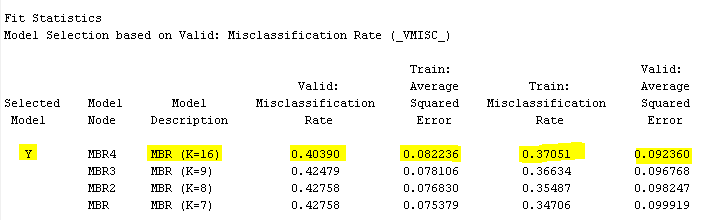


Figure 11: Fit Statistics Window for MBR

The smallest K with the lowest error rate in validation partition, ASE for this model is- 0. 0822.Based on this analysis we can say that algorithm is performing better at train data set. Low K is sensitive to outliers, the high value of K is resilient to outliers as it is considering more records to decide predictions. KNN algorithm searches its database to find ratings that are close to high values, and not close to the low values.

## **Exploratory Analysis using Association Rule Mining: -**

Association rule mining is an Unsupervised technique. It is also termed as affinity analysis and market basket analysis (Shmueli, Bruce, Yahav, Patel, & Lichtendahl, 2018). The general task in Association analysis is to find the underlying hidden patterns and relationships between the items in the transaction data (Sharma & Kumar, 2011). For example, if we consider a rule X 🡪 Y, X is called the antecedent which is on the Left-Hand Side and Y is the antecedent on the Right-Hand Side. We use Support, Confidence and Lift as metrics to measure the strength of the rule generated.

Support gives us the number of times the item repeated in the transaction dataset. Therefore Support (X 🡪 Y) = (number of times X and Y occurred together)/ (Total number of transactions) and we can say Confidence (X 🡪 Y) = (number of times X and Y occurred together)/ (number of times X occurred). The accuracy of the rule is measured by confidence (Victor et al., 2014). Lift is the ratio of the Support of X and Y occurred together to the Support of X and Y when they are independent. When the value of the lift>1 then we can say that the rule is useful (Shmueli et al., 2018).

In Association Mining there are two phases.

1. The frequent patterns are generated by considering a certain threshold level of support and
2. With desired level of confidence level frequent patterns are generated which gives meaningful insights of the data (Raorane & Kulkarni, 2011)

Here In this case study we have mined relationships between different types of cuisines in different restaurants.

**Apriori Algorithm: -**

The algorithm used in this case study to find the hidden association rules is apriori algorithm. It is executed in R studio using R programming language. Initially we must download and install ‘arules’ package to load the library for apriori algorithm. If we get an error in executing the commands, we must check whether the library package is installed or not in the studio.

**Input: -**

The transaction dataset which is extracted from the main dataset.

**Procedure: -**

The algorithm starts counting the items which are most frequently repeated starting from single item set and process is continued till no large itemset is identified. It eliminates item sets which are small.

**Output: -**

Finally, it gives the output item sets which are large and has minimum support and confidence level with lift>1 (Ahmad & Bhatnagar, 2014)

# **Best-chosen Model for our research question**

## **The best-chosen model for data analysis**

When we compare the misclassification rates of different models obtained in the final model comparison node we can see that Classification DT B3D6 is chosen as the best model as shown in the below diagram.

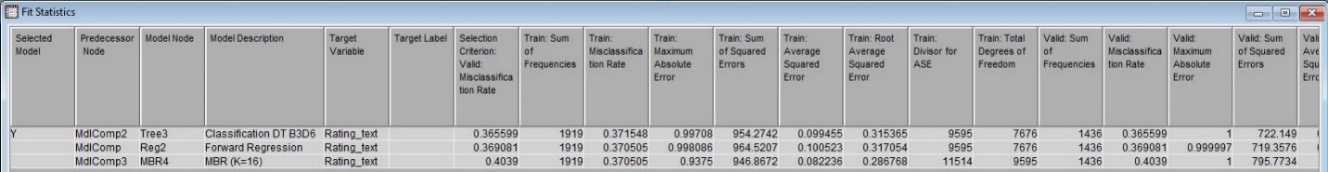


Figure 12: Best Model

If we see the accuracy of the Classification DT B3D6 which is 1- Misclassification Rate, we can see that it has 64% of accuracy.

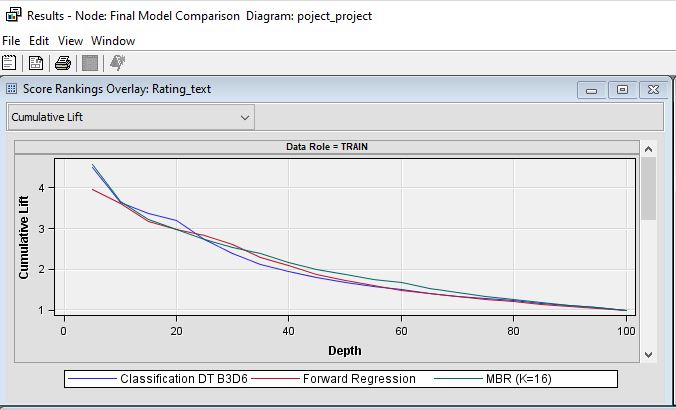


Figure 13: Score Rankings Overlay

## **Scoring the Best Model**

After developing the model, we scored the model on the new data set**.** The roles of all variables here were set as input variable and the levels kept as default. After running the score node, the input and target variables based on the best model identified by the model comparison node classified the new observations in the score data using the classification model developed by us. After the Score node was executed successfully we analyzed the results of the Score Node. In the Exported Data Score Wizard, we checked SCORE assessment. We compared the columns rating text and predicted\_rating\_test that helped help us to identify how many observations will be predicted accurately and/or inaccurately. With both the ratings having very less difference in their values we can predict the customer satisfaction on a restaurant in a location.

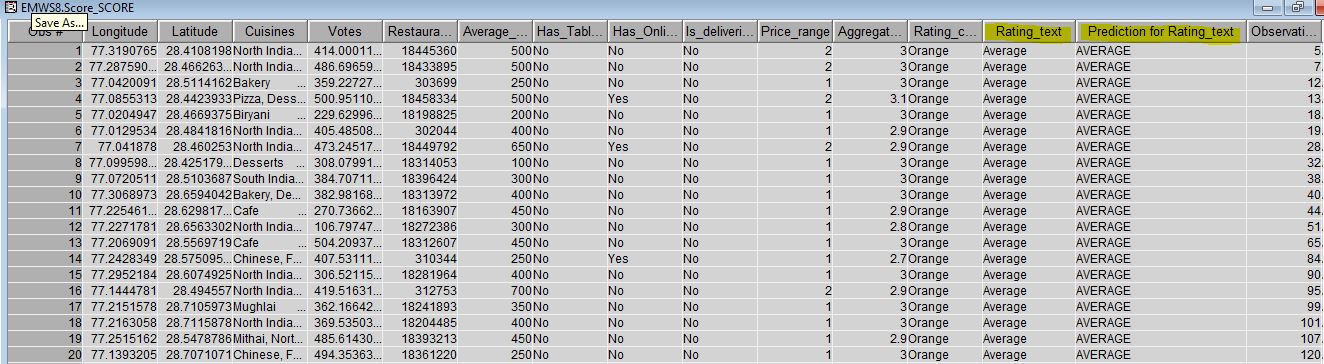


Figure 14: Score Results

## **Results using Data Visualization**

After predicting the Rating text for the restaurants in the score node Figure 15 shows the best restaurants in the given location. The color intensity shows the standard of the restaurant. In Figure 16 it shows that more number of restaurants are predicted to be average.

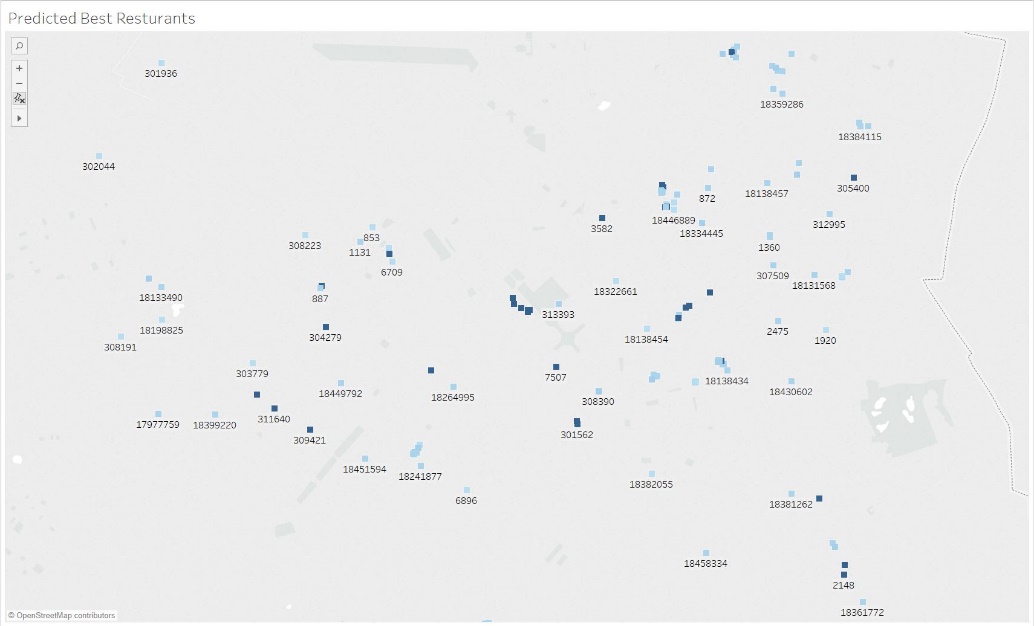


Figure 15: Data Visualization for predicted results

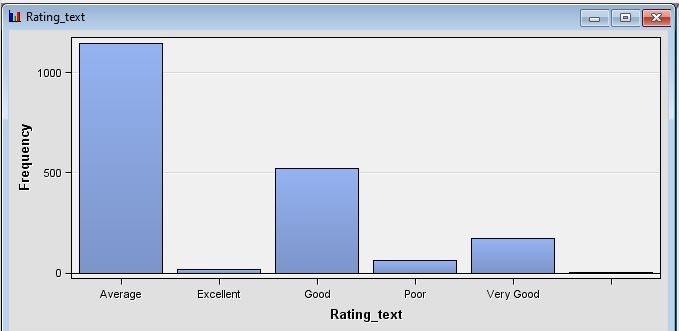


Figure 16: Final Predicted Rating

## **Detailed Exploratory Analysis of Unsupervised Association Rule Mining: -**

Association Rules when support = 0.02, confidence = 0.5



Figure 17:R Result Window

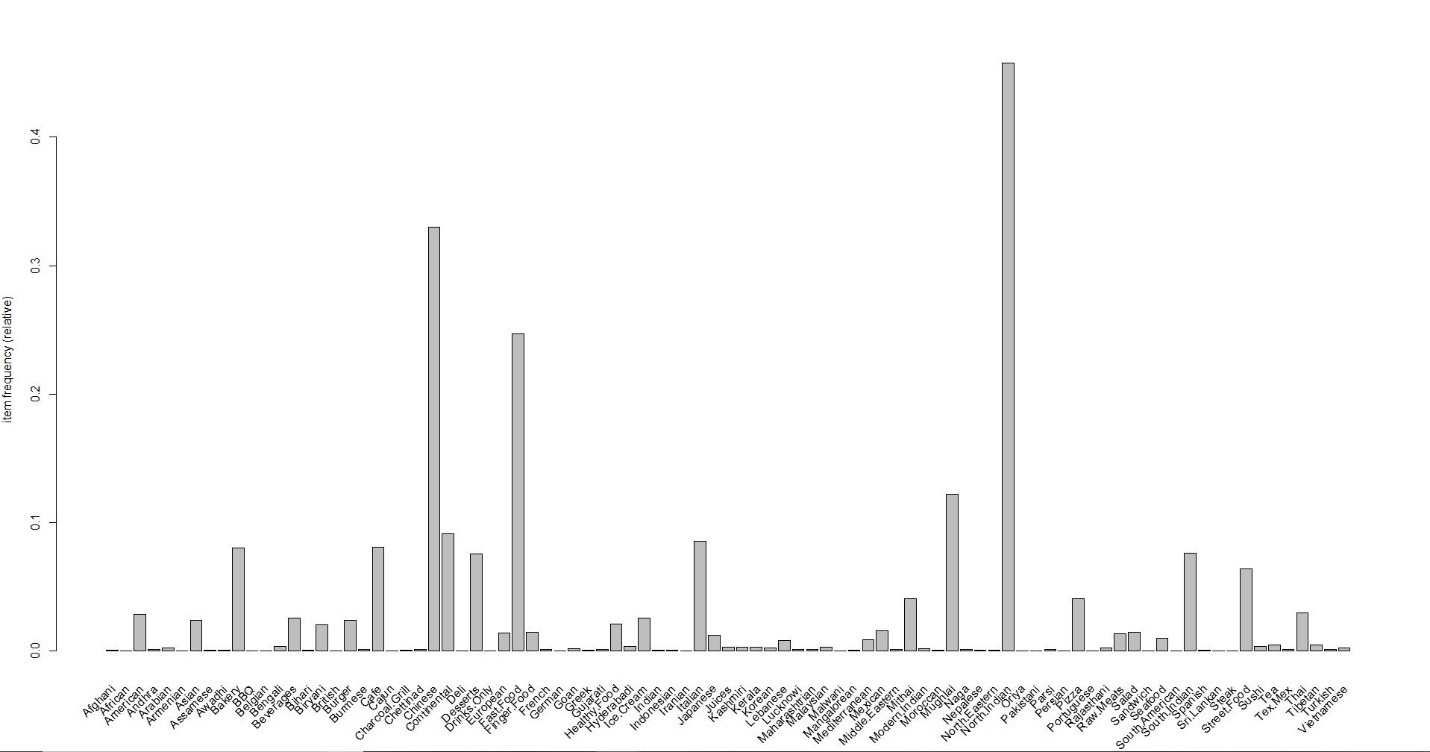


Figure 18:Frequency of Cuisines

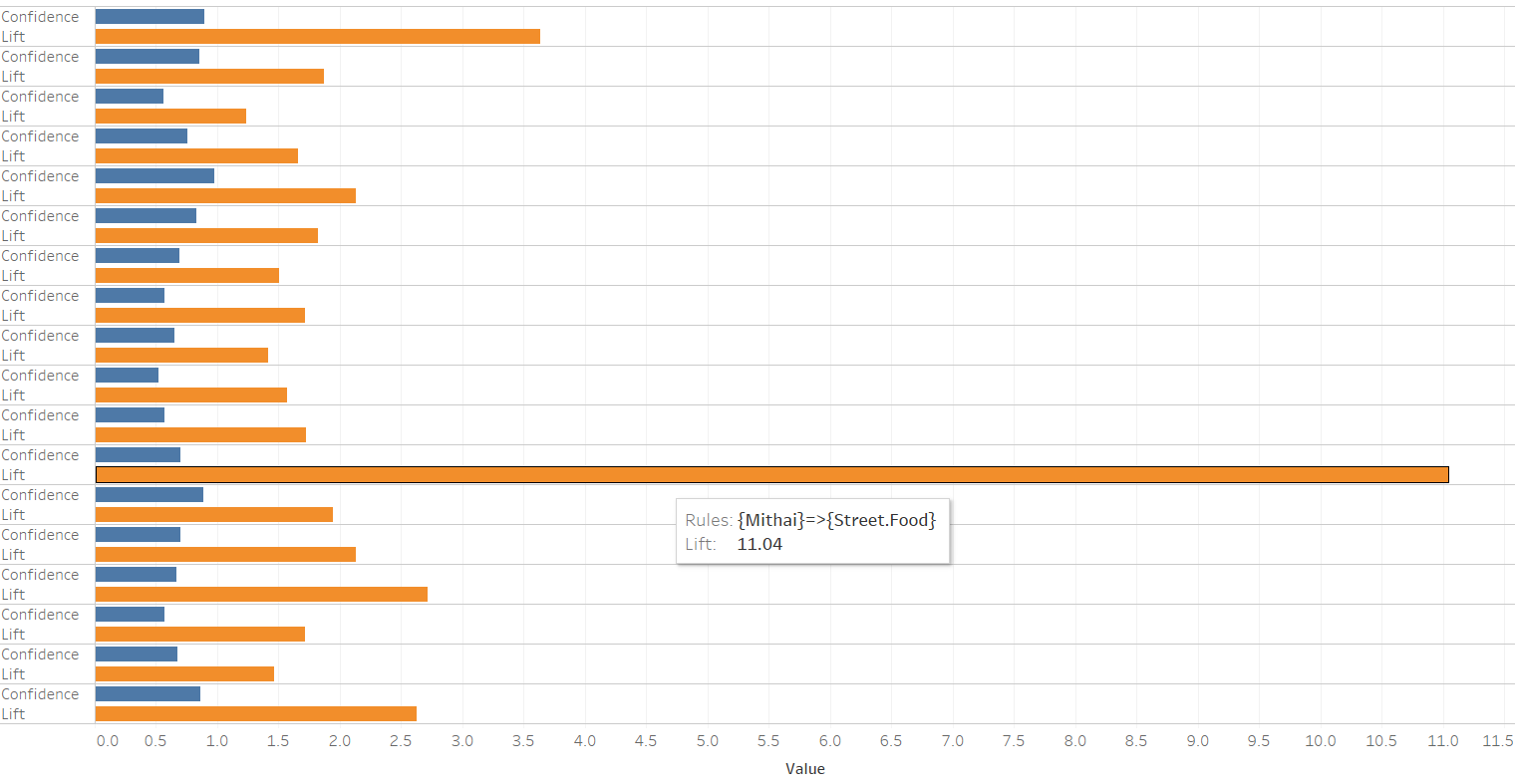
****

Figure 19: Visualization of Association Mining Results

Figure 18 Bar graph shows the frequency of the cuisine and it says that the cuisine North India is repeated most in the transaction dataset.

Considering the lift values greater than one we can suggest that if startup restaurants planning to have mithai cuisine then we recommend them to add Street Food to their cuisine list to attract more customers and based on the requirements of the owners we can give the suggestions from the above results.

Figure 19 shows the visualization of lift, confidence and rules of Association Mining. The orange bar represents the lift value and the blue bar represents the confidence and the rule is represented in the detail.

# **Result Discussion**

This research paper is useful to understand customer behavior, not only through the prediction of rating in a location but also recommend organizations/restaurants what costumers would like to prefer when they are looking for a restaurant. As, we can see in the figure 15, it clearly shows what restaurants have a better chances of customer satisfaction. In the real world, this analysis can be utilized to establish a restaurant that attracts customers, this way it can benefit both the customers and the business.

One interesting thought after doing the association analysis is that considering different cuisines any two random restaurants might have same cuisines so, how to attract the customers towards us will be the next question and exploring the differences by considering other facilities that a restaurant provides would be further interesting topic to get more accurate results.

This topic can be a really fascinating one when coupled with machine learning. Although we humans, do not really follow defined logic to decide what we want to buy, we do have repeated patterns. With the help of AI we can understand those patterns from previous experiences and predict the next buyer’s choices.

# **References**

Ahmad, M. A., & Bhatnagar, N. (2014). Fast Algorithms for Mining Association Rules. Retrieved from http://www.columbia.edu/~rd2537/docu/apriori(abstract).pdf

Chandankhede, C., Devle, P., Waskar, A., Chopdekar, N., & Patil, S. (2016). ISAR: Implicit sentiment analysis of user reviews. In *2016 International Conference on Computing, Analytics and Security Trends (CAST)* (pp. 357–361). https://doi.org/10.1109/CAST.2016.7914994

Erhard, R., & Hong Hai Do. (2000). Data Cleaning: Problems and Current Approaches.

Hemmatian, F., & Sohrabi, M. K. (2017). A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-017-9599-6

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

Saha, S., & Santra, A. K. (2017). Restaurant rating based on textual feedback. In *2017 International conference on Microelectronic Devices, Circuits and Systems (ICMDCS)* (pp. 1–5). https://doi.org/10.1109/ICMDCS.2017.8211542

Sharma, P., & Kumar, S. (2011). An Approach for Customer Behavior Analysis Using Web Mining, *1*(2231), 6.

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl, K. C. (2018). DATA MINING FOR BUSINESS ANALYTICS: Concepts, Techniques, and Applications in R, 577.

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). Cham: Springer International Publishing. 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.

|  |
| --- |
| E. Maslowska, E. Malthouse and S. Bernritter, "Features in Probability to Buy, International Journal of Advertising," *Too Good to be True: The Role of Online Reviews,* pp. 142-163, 2017. |
| A. V. Haastrup, O. A. Oladosu, F. M. Okikiola, O. E. Oladiboye and P. E. Ishola, "American Journal of Computation, Communication and Control," *Customer behaviour analytics and data mining,* 2014. |
| D. I. Hawkins and D. L. Mothersbaugh, Consumer behavior : building marketing strategy, McGraw-Hill Irwin, 2013. |