Sri Lanka Institute of Information Technology

Fundamentals of Data Mining IT3051

B.Sc. (Hons) in Information Technology

DATA SCIENCE

Declaration

We declare that this project report or part of it was not a copy of a document done by any organization, university any other institute or a previous student project group at SLIIT and was not copied from the Internet or other sources

Project Details

Project Title	**Fundamentals of Datamining Mini Project**
Project ID	**Group_4**

Group Members

Reg. No	**Name**
IT 19080840	Kovishwakarunya K
IT 19021430	Hillary J.R
IT 19162010	Yadushika S.K
IT 19175058	Mariyam M.S.S
IT 19955896	Tennakoon T.M.S.N
IT 19058474	Vithya Shagar T

Contents

1. Introduction

I. Introduction to the Dataset

The initial project requirement is to find two real world problems that could be solved by applying Data mining and Machine learning techniques. In order to carry out an in depth Predictive and Descriptive analysis, 2 datasets were chosen. The data sets are chosen in the aim of illustrating and solving business problems within the scope of Food Delivery systems. Therefore, two similar datasets in that scope were chosen.

https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants/version/1

This dataset contains the information of restaurants in Bengaluru which focuses on features such as the different locations of restaurants , approximate price of food , Online/Table booking details, overall reviews and votes obtained by the restaurants and the different cuisines served in them. This data approximately includes 50000 records of data in various data types such as float, string etc. starting from the year 2017 to 2019.

https://www.kaggle.com/henslersoftware/19560-indian-takeaway-orders?select=restaurant-2-orders.csv

This data set contains the order details of centers of two Indian restaurants at UK, which includes approximately 20000 rows of data starting from year 2016 to 2019. Each record describes about a single product within the order. There are nearly 10 features summarized in the data set among which the Order ID, items, quantity, order date and the product price could be considered as some notable features. Its created mainly with the motive of analyzing the takeaway orders in the restaurant and the distribution of products in each order in a daily basis.

II. Problem Identification

A restaurant had established various branches in various locations and there is a need to implement machine learning and data mining techniques to reveal hidden story and insights from their data in order to carry on the future activities of the restaurants in a pre planned manner.

Two real world scenarios of a food delivery system were identified.

The Top-level management has a requirement to identify the association between the food sold in the restaurants, so that appropriate marketing strategies and recommendations could be applied to improve the sales further while making strategic decisions on the least sold food combinations too.

The lower-level management has a need to pre identify the behavior of the customer and their ordering patterns thereby taking measures to provide their need so that the restaurant served ends up with a good overall rating.

III. Introduction to the project

Initially the story behind the overall project is built with the assumption that the two datasets chosen contains data regarding the same restaurant and that the data specified for the three financial years is approximately similar to the current financial year. It is also taken into consideration that the audience of the final outcome of the project are people with proper knowledge of identifying the descriptive and predictive analysis done using the data mining techniques. Initially a usable dataset is generated where necessary preprocessing are done ,then as the next step appropriate data mining techniques are applied for these datasets in order to bring out a predictive analysis which is classification and sentiment analysis and descriptive analysis which is association rule mining respectively.

Based on the predictions and observations from the model fine tuning of models and further enhancements are being done. Jupyter notebook ,Google Collaboratory and Visual studio code are the development environments majorly used for the model development in python whereas streamlit is used for the Web Application implementation. Throughout the project development an agile practice is being followed up by the team.

2. Business Objectives

The main business objective of the project is to increase the sales of the restaurants by implementing better marketing strategies while keeping up the standard of service in the restaurant. The application assists both the top and lower-level management for this purpose. When a customer places an order, the company is aware of the frequently bought together items by implementing an association rule mining technique therefore special offers and discounts are imposed on those items in advance which can further increase sales while enhancing customer and restaurant relationship. On the other hand, by the overall transactions happening in the restaurant, overall ratings of the restaurant can be predicted in advance where the management branch responsible for each restaurant can take necessary measures in advance to prevent any unsatisfaction among customers and implement steps to increase the positive rating. A sentiment analysis done helps the business to understand the social sentiment of their service. Through this the business party will be able to gauge how the customers feel about the different areas of service provided by the company without having to read thousands of customer comments at once.

3. Methodology

a. Predictive Analysis - Classification

I. Data Preprocessing

The ultimate goal of this process is to analyze the dataset and clean data making it appropriate for the Data mining problem to be solved while paving way to train high accurate models to accomplish the business objectives.

Two distinct data sets were used to solve two defined business problems. Therefore, data preprocessing was done to each dataset with regard to the requirement to be fulfilled by the dataset.

Dataset 1 - https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants/version/1

Process	**Column Name**	**Reason**	**Code segment**
Rename Columns	Cost_two	For precise understanding	rename()
Remove comma separations	Cost_two	To validate data for manipulation	replace()
Convert numeric to appropriate data types	Cost_two	To validate data for manipulation	apply(pd.to_numeric)
Remove irrelevant data points	Rate	“NEW”,”-” values are converted to NaN for data validity.	replace()
Denominators of fractions removed	Rate	Presence or absence doesn’t influence the column	replace()
Drop null values in rows and columns	Rate	To increase the model accuracy	dropna()
Converting decimals to rounded values	Rate	To increase the validity of data	Condition based user defined function
Drop unrelated Columns	Dish_Liked ,Menu_Item	Changes due to preference and cannot influence a common problem.	dropna()
Label Encoding	Location,online_order,book_table,type and city,	For further analysis	fit_transform()

II. Exploratory Data Analysis

In order to carry out the classification, various analysis were done prior to it , to identify the relationship between the features and the influence of them towards the interested target.

1. Analysis of a univariate distribution of data points of ‘Rate’ was done against the density distribution in the aim of analyzing how the Rate provided by the customers are distributed along the restaurants considered. Once Rate column is discretized a clear insight on the rate density is observed

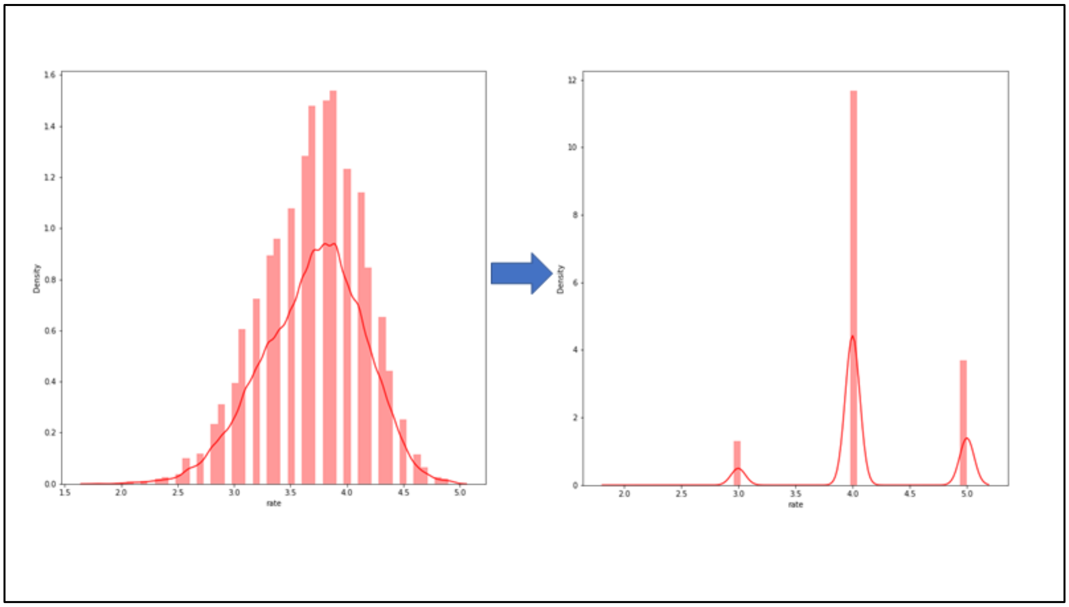
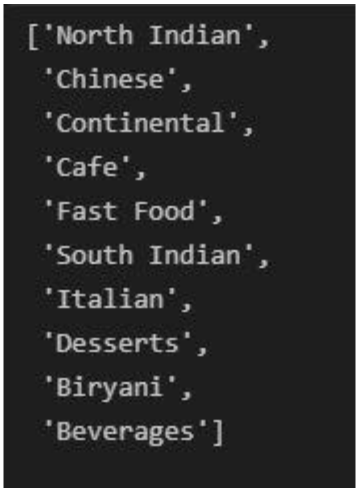


Figure 1 Rate Distribution Analysis

It is observed that the overall rates observed are highly distributed in the range of 3.0 - 5.0

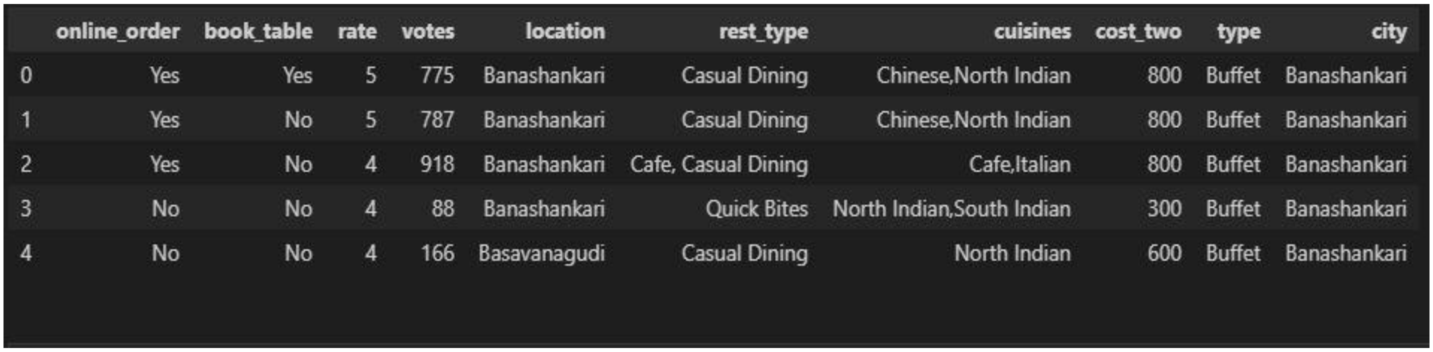
2. The cuisines addressed in the data are analyzed. It is found to be in separate lists. By an in-depth analysis, we could observe 97 different types of cuisines that the restaurant deals with. Rather than making predictions from all these cuisines, only the top 10 cuisines are taken further for better analysis.



```
['North Indian',
 'Chinese',
 'Continental',
 'Cafe',
 'Fast Food',
 'South Indian',
 'Italian',
 'Desserts',
 'Biryani',
 'Beverages']
```

Figure 2 Top 10 Cuisines

The below given overview of the data set shows the cuisines after filtering out the top 10 cuisines dealt in within the restaurant



	online_order	book_table	rate	votes	location	rest_type	cuisines	cost_two	type	city
0	Yes	Yes	5	775	Banashankari	Casual Dining	Chinese,North Indian	800	Buffet	Banashankari
1	Yes	No	5	787	Banashankari	Casual Dining	Chinese,North Indian	800	Buffet	Banashankari
2	Yes	No	4	918	Banashankari	Cafe, Casual Dining	Cafe,Italian	800	Buffet	Banashankari
3	No	No	4	88	Banashankari	Quick Bites	North Indian,South Indian	300	Buffet	Banashankari
4	No	No	4	166	Basavanagudi	Casual Dining	North Indian	600	Buffet	Banashankari

Figure 3 Data Overview

4. Further ahead of our analysis , the feature “book_table” is analyzed to identify its influence on the rate provided to the restaurant.

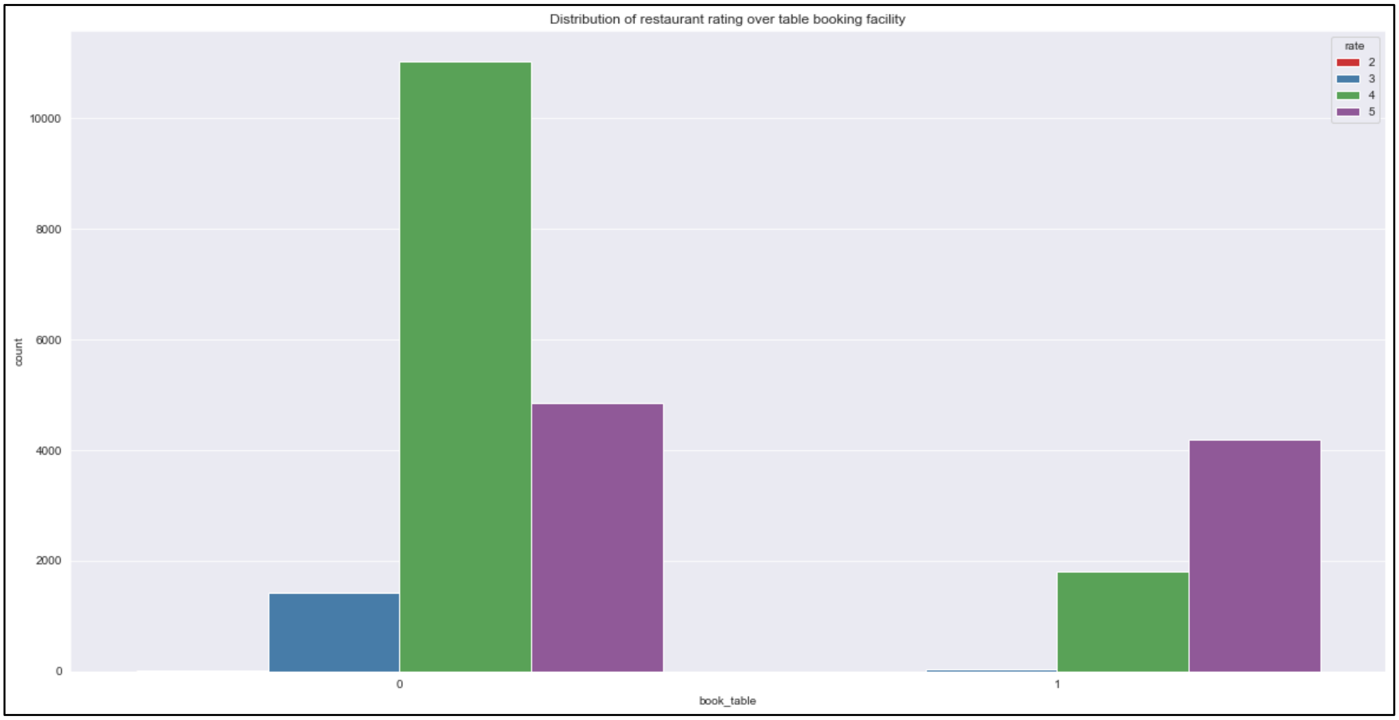


Figure 4 Rate vs Book Table facility

Through this analysis it is observed that book_table facility has a high influence on the rating obtained. When there is a table booked, the possibility of getting a high rate is less. So it is evident that the concern on a customer making a prior booking should be high in order to improve overall rating.

5. Further ahead of our analysis , the feature “online_order” is analyzed to identify its influence on the rate provided to the restaurant

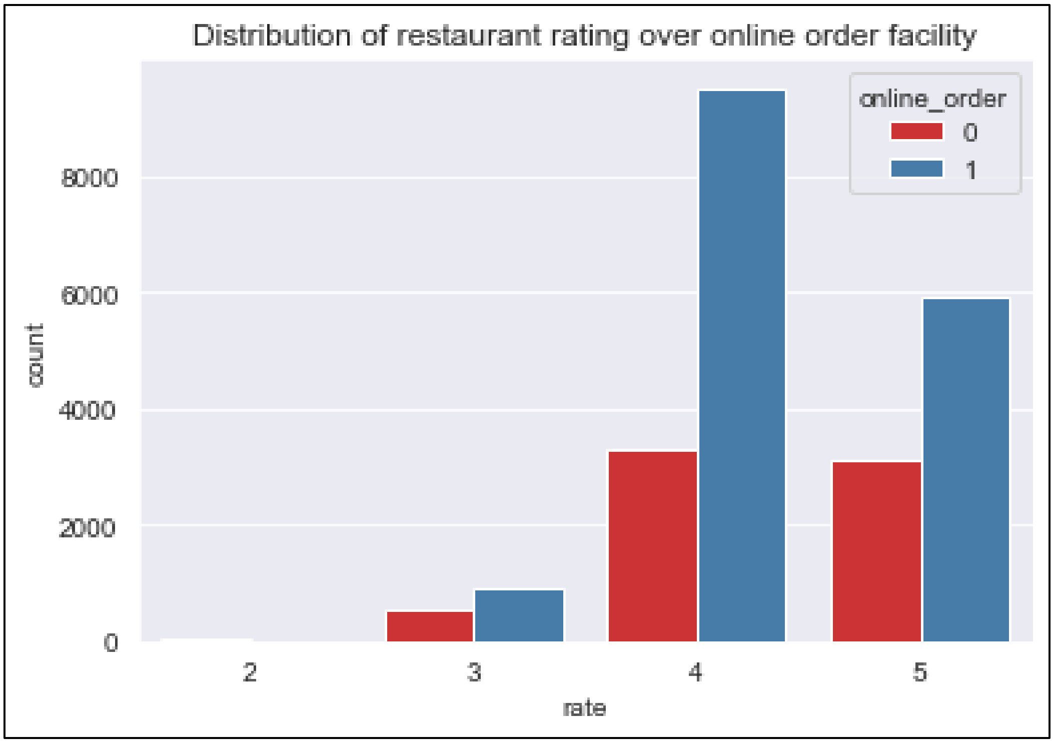


Figure 5 Rate Vs Online Order Facility

Through this analysis it is observed that “online_order” facility has a high influence on the rating obtained. When there is an online order made, the possibility of getting a rating of 3 to 5 is higher compared to that obtained when it is not an online order. So it is likely to conclude that the restaurants having an online order facility are more likely to get a high rating than that of the ones which do not have an online order facility.

III. Multi class Classification

Having done an **exploratory data analysis**, we were able to come up with good insights on which features highly influence the target of prediction.

The features online order, table booking , votes obtained , location , average cost of expense for two people and the cuisines in the restaurant highly influence the overall rating that a restaurant can obtain. So, these features are considered for the prediction to be done.

Before model evaluation, data is scaled and normalized as required to remove the bias of the data present in the independent features considered.

The data is split into train and test data where the train data was used to build the model and the accuracy of prediction was tested on the test data.

A series of Regression models were built to analyze the predictions we get in the aim of providing an accurate model that could allow predicting the continuous variable that we are interested in thus solving the business problem defined.

Shown below are the accuracies obtained for the model built after the data preprocessing step.

Model	**Accuracy**
Random forest Regression	0.89
Extra Tree Regressor	0.87
Gradient Boosting Regressor	0.69
Decision Tree Regressor	0.81

It is seen that Random Forest regression, Extra Tree Regressor and Decision Tree Regressor tend to provide better accuracies for our prediction.

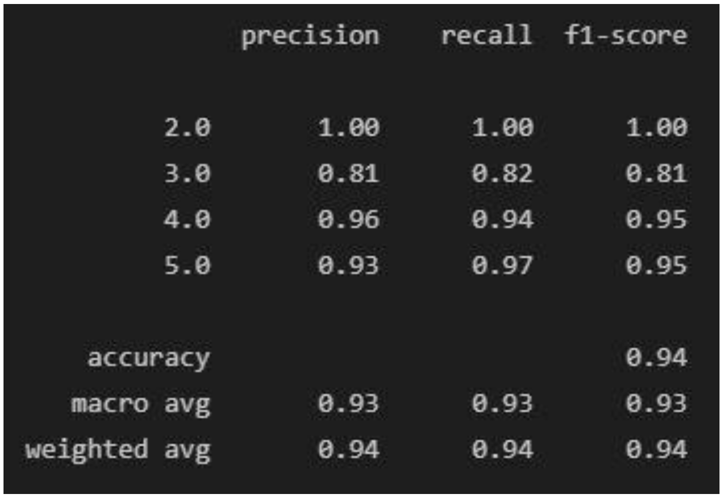
In order to increase the accuracy of prediction, each model was evaluated using **Cross validation** methods in **sklearn** model and **hyper parameter tuning** to identify the most suitable accuracy for the multi class classification performed.

Shown below are the result obtained after the model tuning process.

Model	**Accuracy**
Random forest Regression	0.96
Extra Tree Regressor	0.98
Gradient Boosting Regressor	0.77
Decision Tree Regressor	0.94

When deciding on the best prediction models , in addition to accuracy other measures such as precision , recall and f1 score were also examined. In order to analyze this a classification report visualizer from the sklearn library is used.

Shown below are the classification report visualizers for each model built.

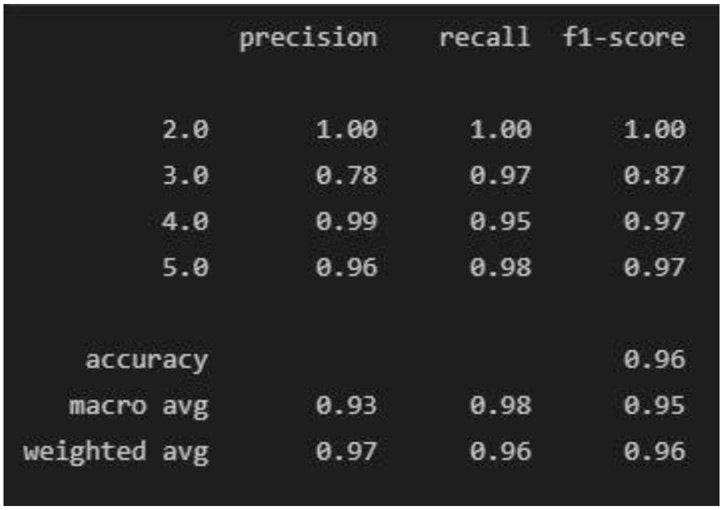


```
              precision    recall  f1-score


         2.0       1.00      1.00      1.00
         3.0       0.81      0.82      0.81
         4.0       0.96      0.94      0.95
         5.0       0.93      0.97      0.95

    accuracy                           0.94
   macro avg       0.93      0.93      0.93
weighted avg       0.94      0.94      0.94
```

Figure 9 Decision Tree Regressor visualizer

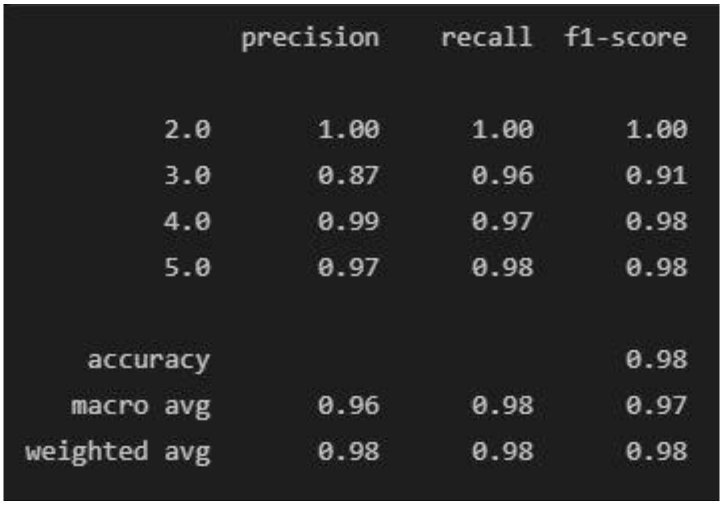


```
              precision    recall  f1-score


         2.0       1.00      1.00      1.00
         3.0       0.78      0.97      0.87
         4.0       0.99      0.95      0.97
         5.0       0.96      0.98      0.97

    accuracy                           0.96
   macro avg       0.93      0.98      0.95
weighted avg       0.97      0.96      0.96
```

Figure 8 Random Forest Regression Visualizer

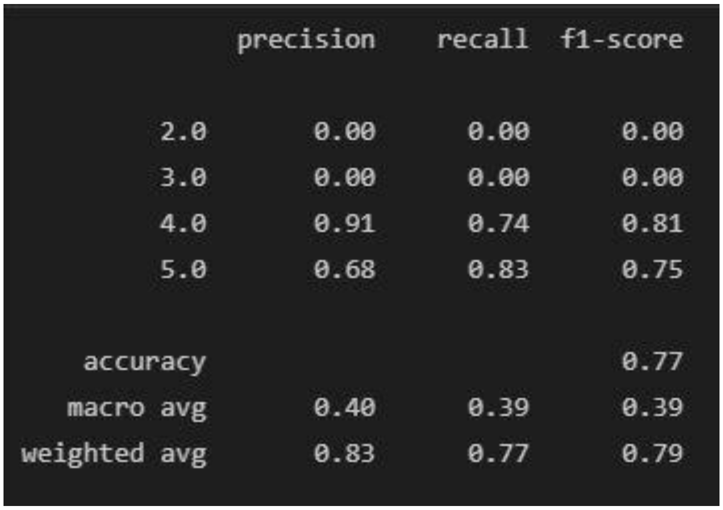


```
              precision    recall  f1-score


         2.0       1.00      1.00      1.00
         3.0       0.87      0.96      0.91
         4.0       0.99      0.97      0.98
         5.0       0.97      0.98      0.98

    accuracy                           0.98
   macro avg       0.96      0.98      0.97
weighted avg       0.98      0.98      0.98
```

Figure 7 Extra Tree Regression Visualizer



```
              precision    recall  f1-score


         2.0       0.00      0.00      0.00
         3.0       0.00      0.00      0.00
         4.0       0.91      0.74      0.81
         5.0       0.68      0.83      0.75

    accuracy                           0.77
   macro avg       0.40      0.39      0.39
weighted avg       0.83      0.77      0.79
```

Figure 6 XGB Regression Visualizer

Precision basically provides the accuracy of the positive prediction. (Case is positive and obtaining positive result and vise versa) . From the above visualizer it can be observed that the three models that gave a higher accuracy are having a **higher precision** too.

Recall tells about what percentage of the positive values were caught by our model. High recall values imply that our model is doing really well and hasn't been under fitting.

In addition to this, **Macro averaging** is obtained through the classification report, where the multi class predictions are broken down into multiple sets of binary predictions where the corresponding metric for each binary case is found and averaged.

Also **Weighted average** is obtained in the aim of obtaining the metrics with respect to each of the label's proportion in the data analyzed.

Despite of the accuracies obtained, in order to get an estimate of how well the model is performing, further analysis was done. On that basis, **Mean absolute percentage error (MAPE)** is used as the measure.

In order to do it we defined a function that gets the actual and predicted values as input parameters. It implies that, lower the value it has, the difference between actual and predicted value is less thus proving that our model is performing really well and is not under fitting / over fitting.

Given below is a summary of the MAPE obtained for our models.

Model	**MAPE**
Random forest Regression	0.9096746575
Extra Tree Regressor	0.6274971461
Gradient Boosting Regressor	5.6367722602
Decision Tree Regressor	1.3930507990

From the above statistics provided, it is evident that the Random Forest Regression model and the Extra Tree Regression model having **lesser MAPE** are found to be better performing models with high accuracy.

Once the fine-tuned models are analyzed in the above-mentioned aspects, we were able to come up to a conclusion that the **Extra Tree Regression model** is the best model to be used to predict the target that we are concerned about.

b. Descriptive Analysis – Association rule mining

Dataset 2 - https://www.kaggle.com/henslersoftware/19560-indian-takeaway-orders?select=restaurant-2-orders.csv

I. Data Preprocessing

Process	Column Name	Reason	Code segment
Group data	Item amount is grouped by OrderID, Item	To unstack for Association rule mining.	groupby() and unstack()
Convert data to Boolean values.	Quantity	To indicate association of items in the same order	.applymap()
Drop columns	ProductPrice , TotalQuanity	Process only important features	.drop()
Rename the columns	OrderID,OrderName	Precise understanding	.rename()

II. Exploratory Data Analysis.

1. Initially the number of transactions done in each part of the year is analyzed. It is observed that the number of transaction increase as time advances but tends to decrease at the end of the year.

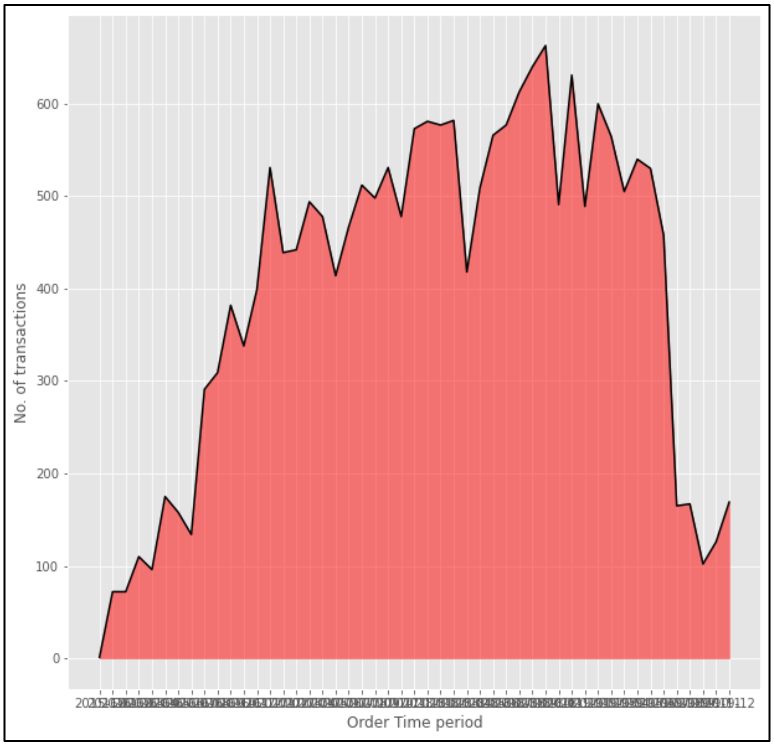


Figure 10 Time Period VS No of Transactions

2. As the next analysis, the most selling item in the market was analyzed based on the maximum revenue that each item brings.

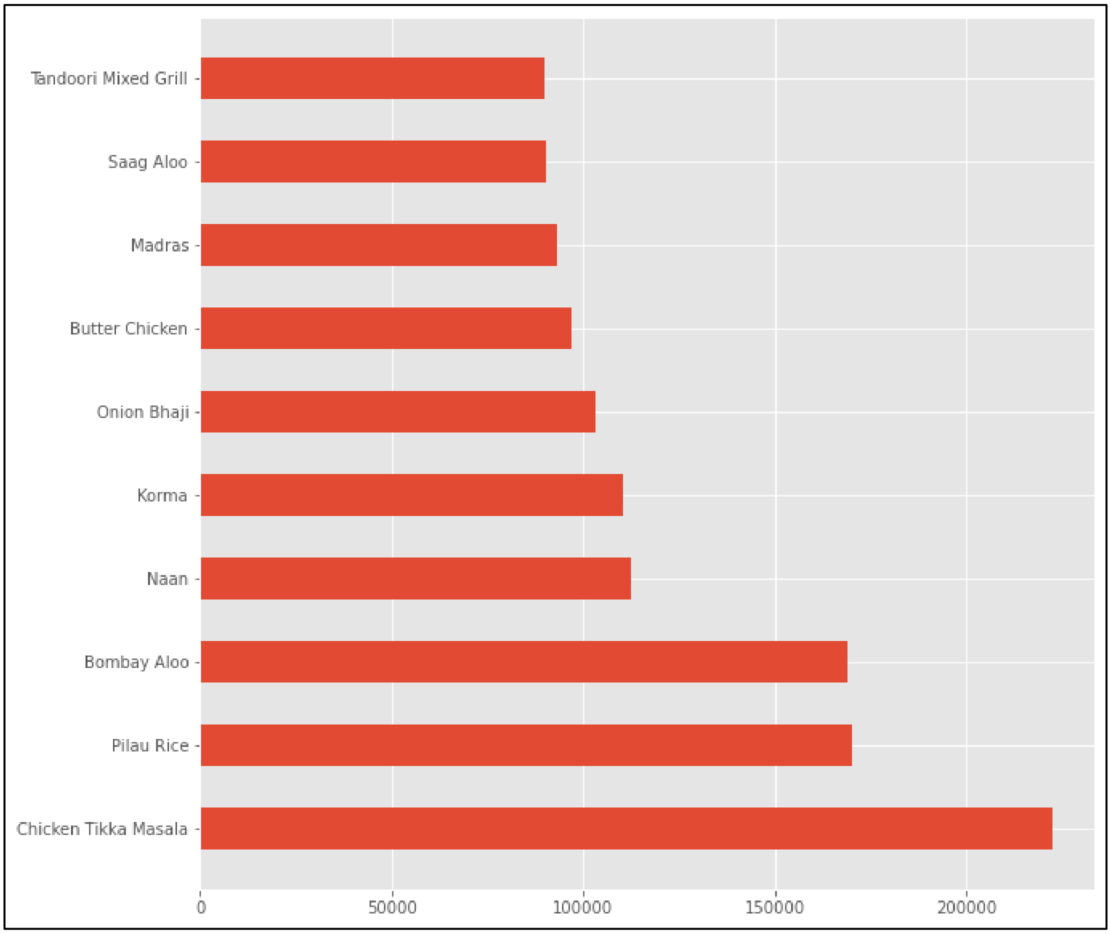
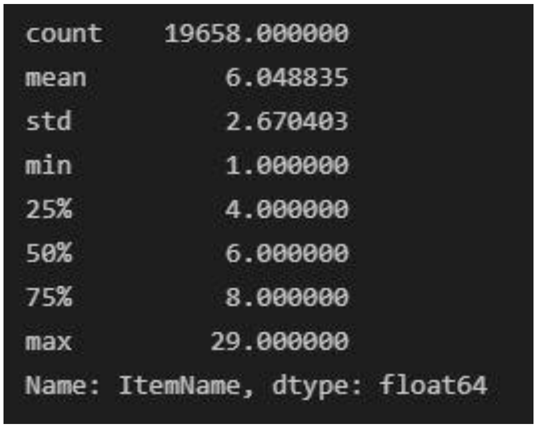


Figure 11 Most Selling Items

It is observed that the items such as Chicken Tika Masala, Pilau Rice and Bombay Aloo are the ones that bring up a high revenue to the company.

3. The number of items bought in each transaction are analyzed. The minimum no of items in a transaction is 1 and the maximum number of items are 29



```
count    19658.000000
mean         6.048835
std          2.670403
min          1.000000
25%          4.000000
50%          6.000000
75%          8.000000
max         29.000000
Name: ItemName, dtype: float64
```

.

Figure 12 Statistics on Number of Items

III. Apriori Algorithm Implementation

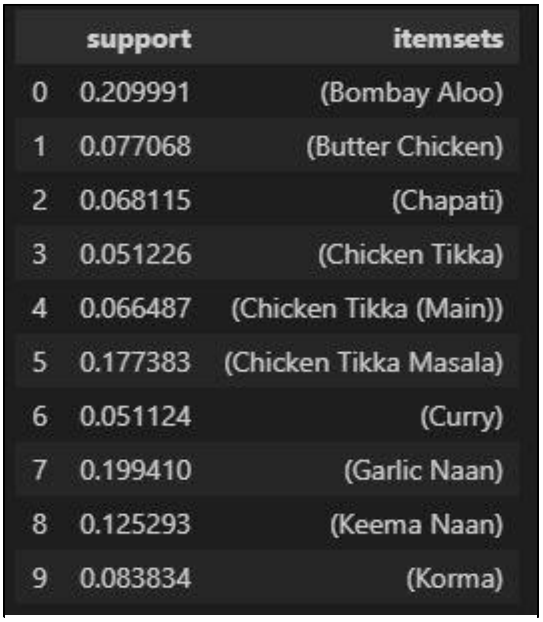
As the main purpose of Association rule mining for the restaurant was to analyze the food patterns within the orders done for a time period thereby helping business management to pre plan marketing strategies and offerings accordingly.

Once the data preprocessing and analysis is done, the data frame which summarizes about each transaction of item with the total number of items bought in each transaction are created as the final set of data for association rule mining.

Step 1 - Frequent item-set generation.

In order to generate accurate frequent itemset the main metric used is the 'Support'. A minimum threshold is defined thereby helping us validate the actual support of the food items in the order against this threshold.

The Minimum Support is defined as 0.05.We get the following itemset.



	support	itemsets
0	0.209991	(Bombay Aloo)
1	0.077068	(Butter Chicken)
2	0.068115	(Chapati)
3	0.051226	(Chicken Tikka)
4	0.066487	(Chicken Tikka (Main))
5	0.177383	(Chicken Tikka Masala)
6	0.051124	(Curry)
7	0.199410	(Garlic Naan)
8	0.125293	(Keema Naan)
9	0.083834	(Korma)

Figure 13 A view of itemset

To generate the frequent item sets the Apriori module from the mlxtend library is used.

Shown below is the overall analysis of the item sets obtained with respect to their support.

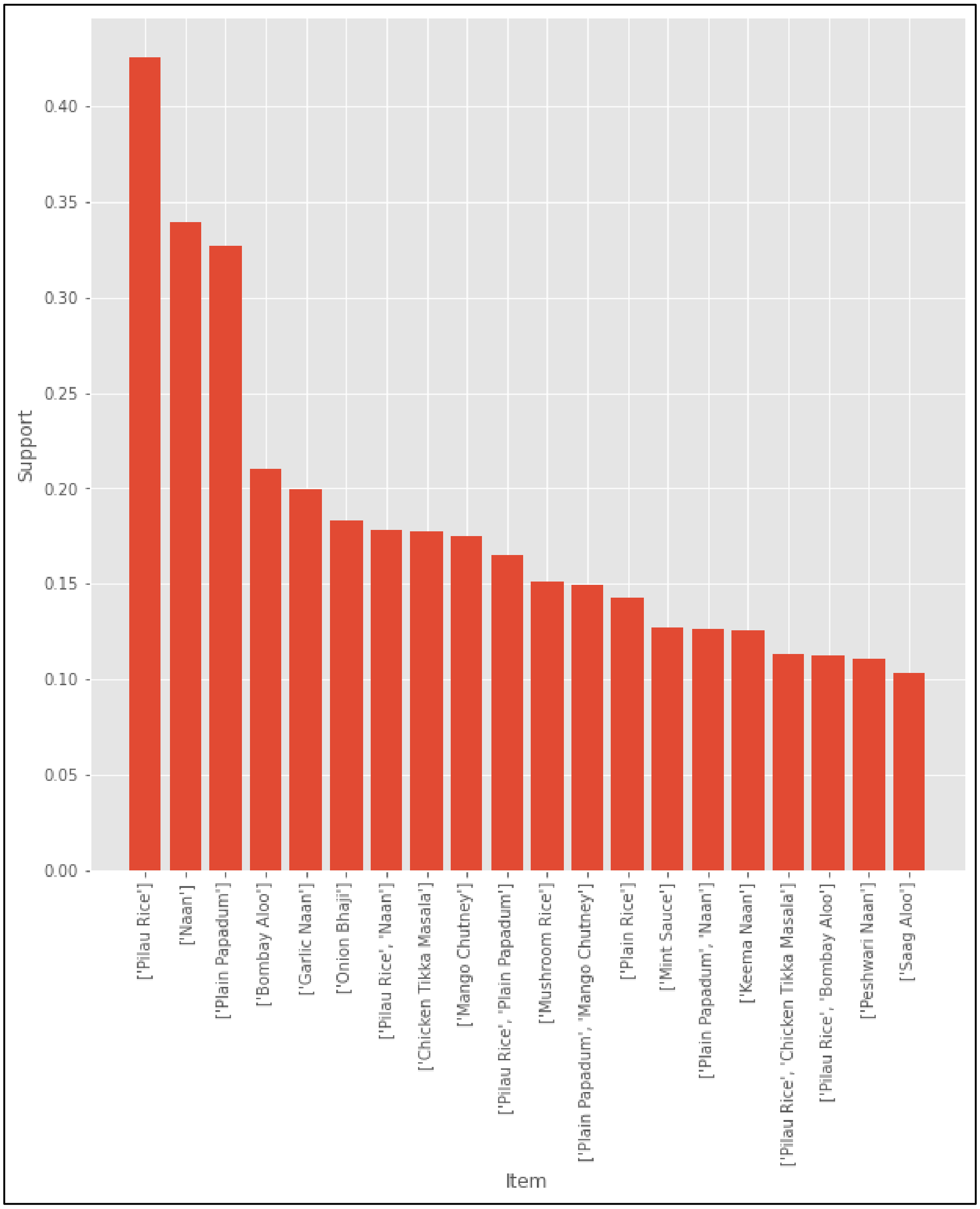


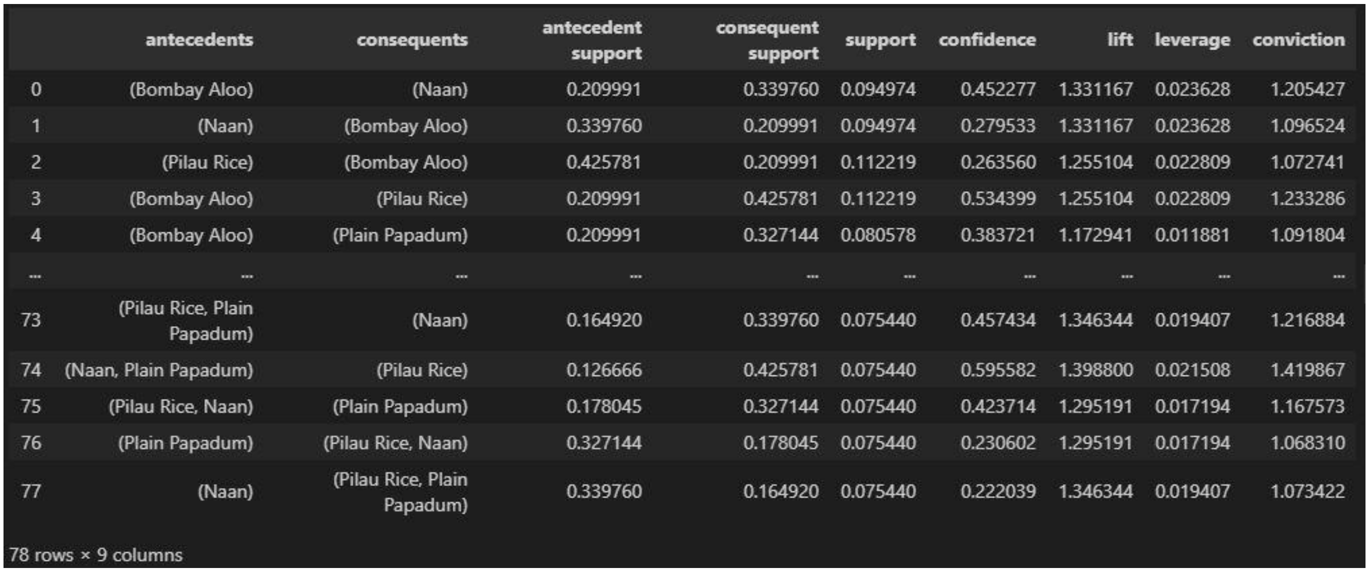
Figure 14 Support Vs Item Set

Step 2 – Association rule generation.

Once the frequent itemset are generated , high confidence rules are created from each frequent item set where each rule is a binary partitioning of a frequent itemset. For this another distinct metric is used which is known as 'Confidence'.

The Minimum Confidence is defined as 0.2.

By following the above two steps antecedents and consequents of a n-item sets were obtained.



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(Bombay Aloo)	(Naan)	0.209991	0.339760	0.094974	0.452277	1.331167	0.023628	1.205427
1	(Naan)	(Bombay Aloo)	0.339760	0.209991	0.094974	0.279533	1.331167	0.023628	1.096524
2	(Pilau Rice)	(Bombay Aloo)	0.425781	0.209991	0.112219	0.263560	1.255104	0.022809	1.072741
3	(Bombay Aloo)	(Pilau Rice)	0.209991	0.425781	0.112219	0.534399	1.255104	0.022809	1.233286
4	(Bombay Aloo)	(Plain Papadum)	0.209991	0.327144	0.080578	0.383721	1.172941	0.011881	1.091804
...	...	...	...	...	...	...	...	...	...
73	(Pilau Rice, Plain Papadum)	(Naan)	0.164920	0.339760	0.075440	0.457434	1.346344	0.019407	1.216884
74	(Naan, Plain Papadum)	(Pilau Rice)	0.126666	0.425781	0.075440	0.595582	1.398800	0.021508	1.419867
75	(Pilau Rice, Naan)	(Plain Papadum)	0.178045	0.327144	0.075440	0.423714	1.295191	0.017194	1.167573
76	(Plain Papadum)	(Pilau Rice, Naan)	0.327144	0.178045	0.075440	0.230602	1.295191	0.017194	1.068310
77	(Naan)	(Pilau Rice, Plain Papadum)	0.339760	0.164920	0.075440	0.222039	1.346344	0.019407	1.073422

78 rows × 9 columns

Figure 15 Rules generated with given confidence

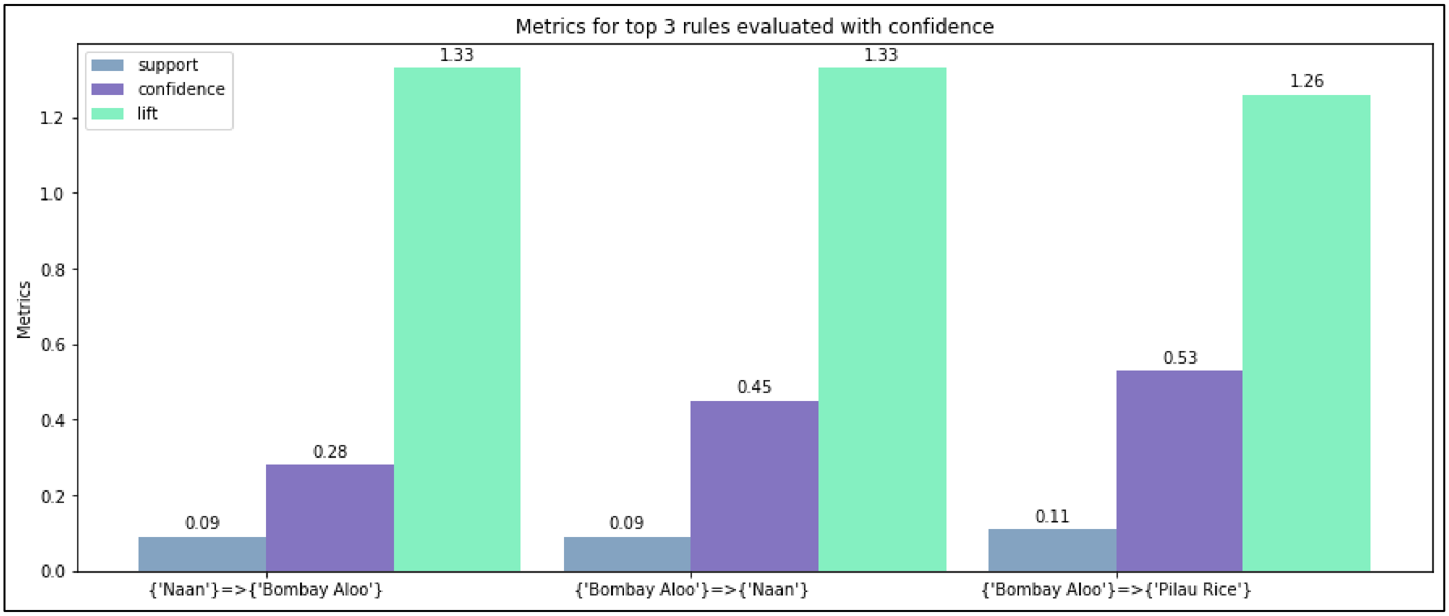


Figure 16 Top 3 rules

Shown above are the observations of the top 3 rules obtained among the 78 rules generated with respect to the metrics support, confidence, lift by implementing association rule mining on the food order combinations of the data analyzed

By this insight the low-level management can make decisions on what type of offers they can provide for the food in the order made and what strategies should be implemented to enhance the low-level food combinations while maintaining the demand for the top rules discovered.

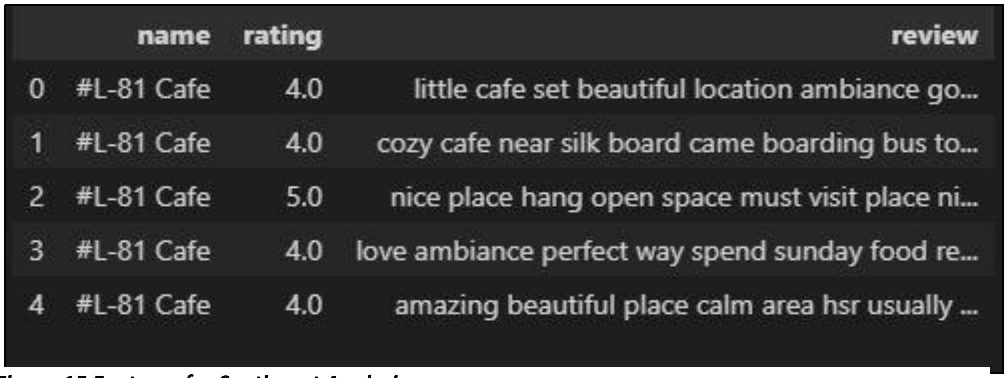
c. Sentiment Analysis.

I. Data Preprocessing

Process	Column Name	Reason	Code segment
Group and aggregate duplicate columns	Name and Address	Data Validation	.groupby() and .merge()
Drop rows with duplicates	Listed_in(type)	Data Validation	.drop_duplicates()
Drop null values	Rate	Doesn’t influence prediction	.dropna()
Remove irrelevant data points	Rate	“NEW”,”-” values are converted to NaN for data validity.	replace()
Remove stopwords	Review	For accurate prediction	Stopwords.words()

In addition to the preprocessing mentioned above, text preprocessing with the usage of **nltk** library is done. As the prominent text preprocessing more attention was given to **tokenization**, removal of **stop words** and **lemmatization**.

II. Sentiment Analysis Implementation



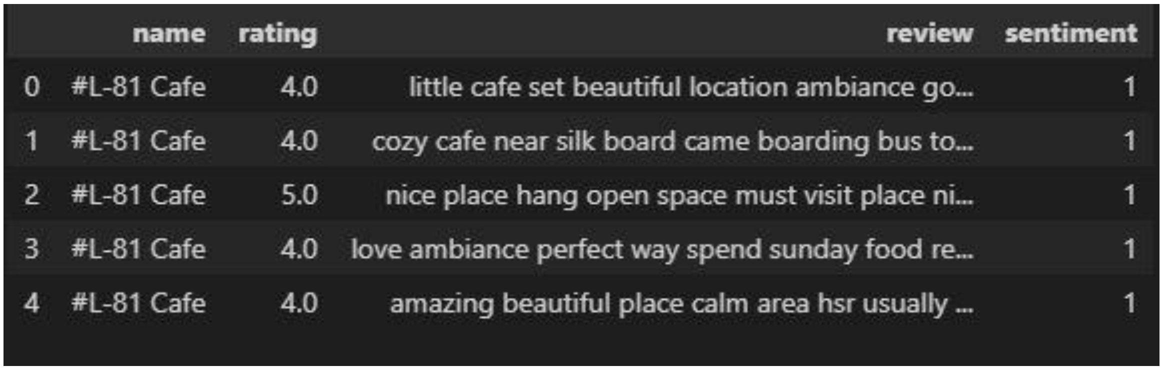
```
        name  rating                                             review
0  #L-81 Cafe     4.0   little cafe set beautiful location ambiance go...
1  #L-81 Cafe     4.0  cozy cafe near silk board came boarding bus to...
2  #L-81 Cafe     5.0   nice place hang open space must visit place ni...
3  #L-81 Cafe     4.0  love ambiance perfect way spend sunday food re...
4  #L-81 Cafe     4.0    amazing beautiful place calm area hsr usually ...
```

Figure 15 Features for Sentiment Analysis

The above figure shows the view of the final data resulting after the data preprocessing step. So further ahead depending on the rating a new column of labels were inserted to define the sentiment that we need to predict.

If the rating obtained is greater than 3, it is taken as a positive sentiment (1) while the others are considered to be negative.

Shown below is the final train data created for the model building process ahead.



	name	rating	review	sentiment
0	#L-81 Cafe	4.0	little cafe set beautiful location ambiance go...	1
1	#L-81 Cafe	4.0	cozy cafe near silk board came boarding bus to...	1
2	#L-81 Cafe	5.0	nice place hang open space must visit place ni...	1
3	#L-81 Cafe	4.0	love ambiance perfect way spend sunday food re...	1
4	#L-81 Cafe	4.0	amazing beautiful place calm area hsr usually ...	1

Figure 16 Training Data for Prediction

Word clouds are visualized to identify the positive feedbacks and negative feedbacks received by the restaurant.

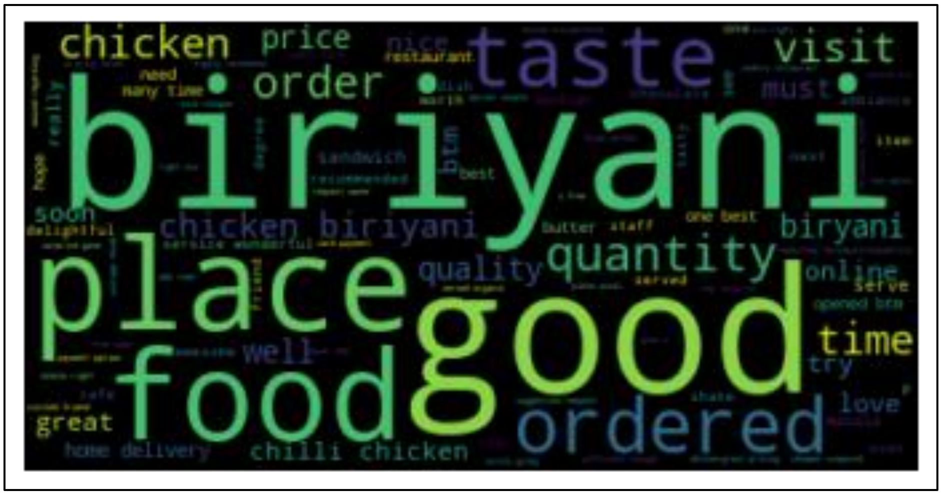


Figure 17 Positive word Cloud

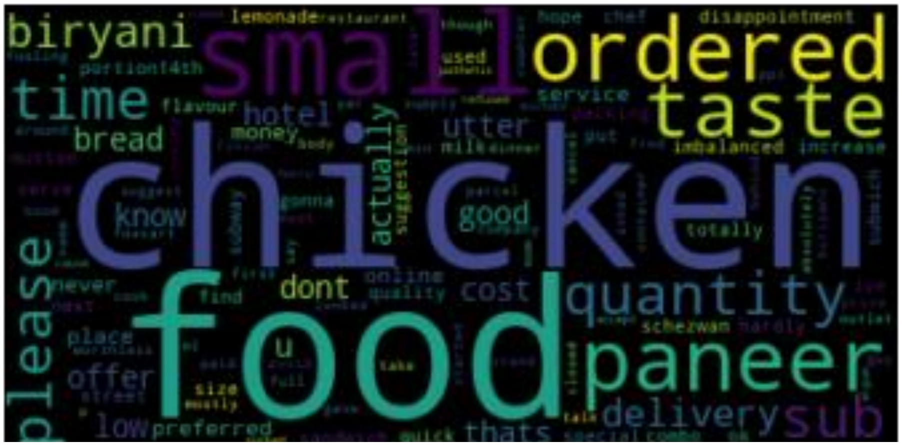


Figure 18 Negative Word Cloud

An analysis was done to obtain an idea on the overall customer sentiments with regard to the service of the restaurant.

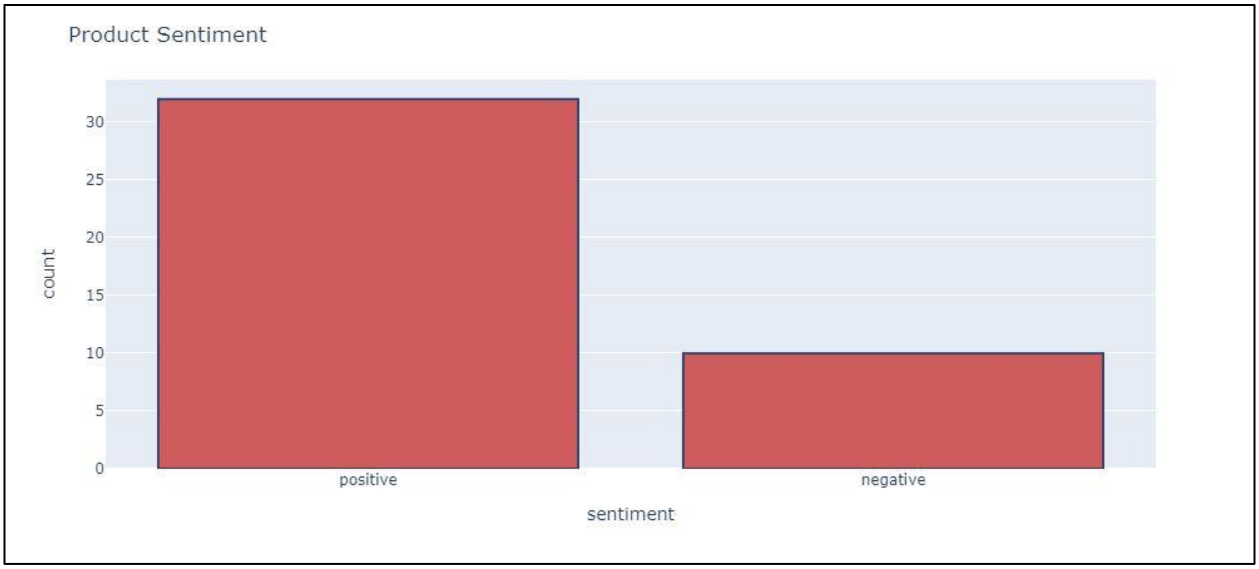


Figure 19 Customer Sentiment

The data is split into train and test data where the train data was used to build the model and the accuracy of prediction was tested on the test data.

A logistic Regression model is built to analyze the desired predictions and we ended up with an accuracy of 92%.

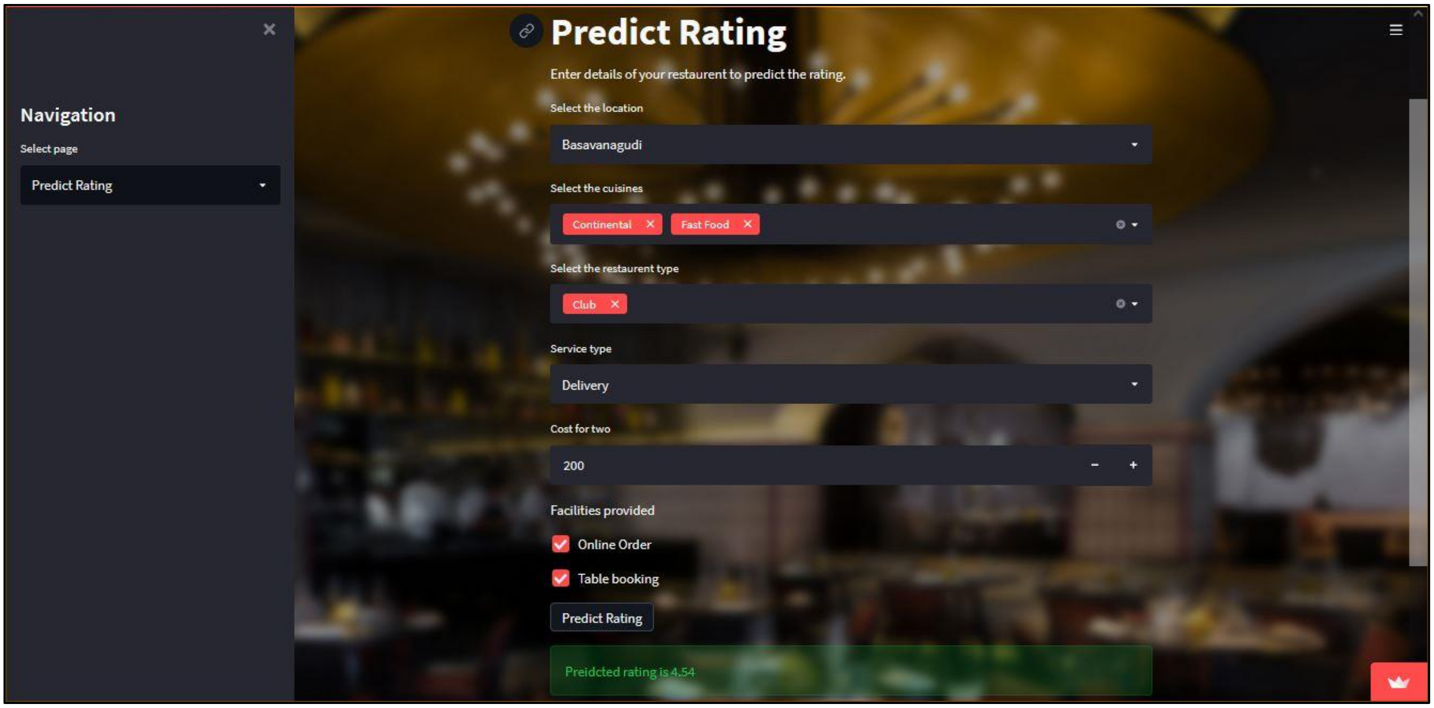
The main motive behind this prediction is to classify whether the given feedback or review in textual context is positive or negative. Reviews can be given to the model, and it classifies the review as a negative review or a positive review. This shows the satisfaction of the customer or the experience the customer has experienced which can help the business improve their service in target areas.

4. Web Application Implementation

Once the models are trained , the trained models are converted to a pickle file in order to feed the data points for the web application implementation

I. Classification Model Implementation

For the purpose of prediction, the application accepts a series of inputs as shown in the figure below.



Navigation

Select page: Predict Rating

Predict Rating

Enter details of your restaurent to predict the rating.

Select the location: Basavanagudi

Select the cuisines: Continental, Fast Food

Select the restaurent type: Club

Service type: Delivery

Cost for two: 200

Facilities provided

☑ Online Order

☑ Table booking

Preidcted rating is 4.54

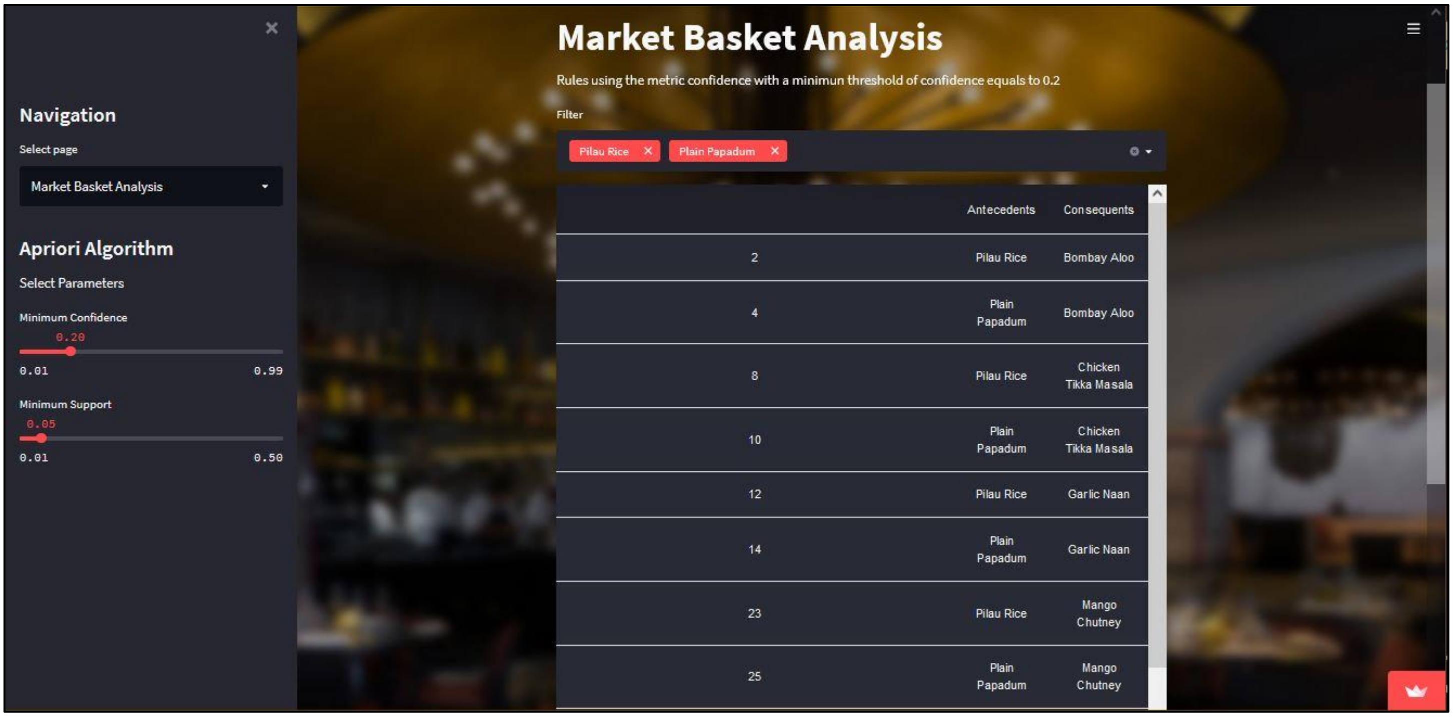
Figure 20 Rate Prediction interface

The system will predict the overall rating depending on the inputs given. If the rating obtained is low the management can switch to different appropriate combinations to identify how a better rating can be obtained and then appropriate strategies can be followed.

II. Market Basket Analysis.

To analyze the associations between the food ordered in a single order ,the market basket analysis allows the user to select the respective food they need to make an analysis on. By doing so the user can see the consequents of those food items.

We have provided sliders to adjust the support and confidence that could adhere with the changing management requirements.



Navigation

Select page

Market Basket Analysis

Apriori Algorithm

Select Parameters

Minimum Confidence 0.20 (0.01 – 0.99)

Minimum Support 0.05 (0.01 – 0.50)

Market Basket Analysis

Rules using the metric confidence with a minimun threshold of confidence equals to 0.2

Filter: Pilau Rice, Plain Papadum

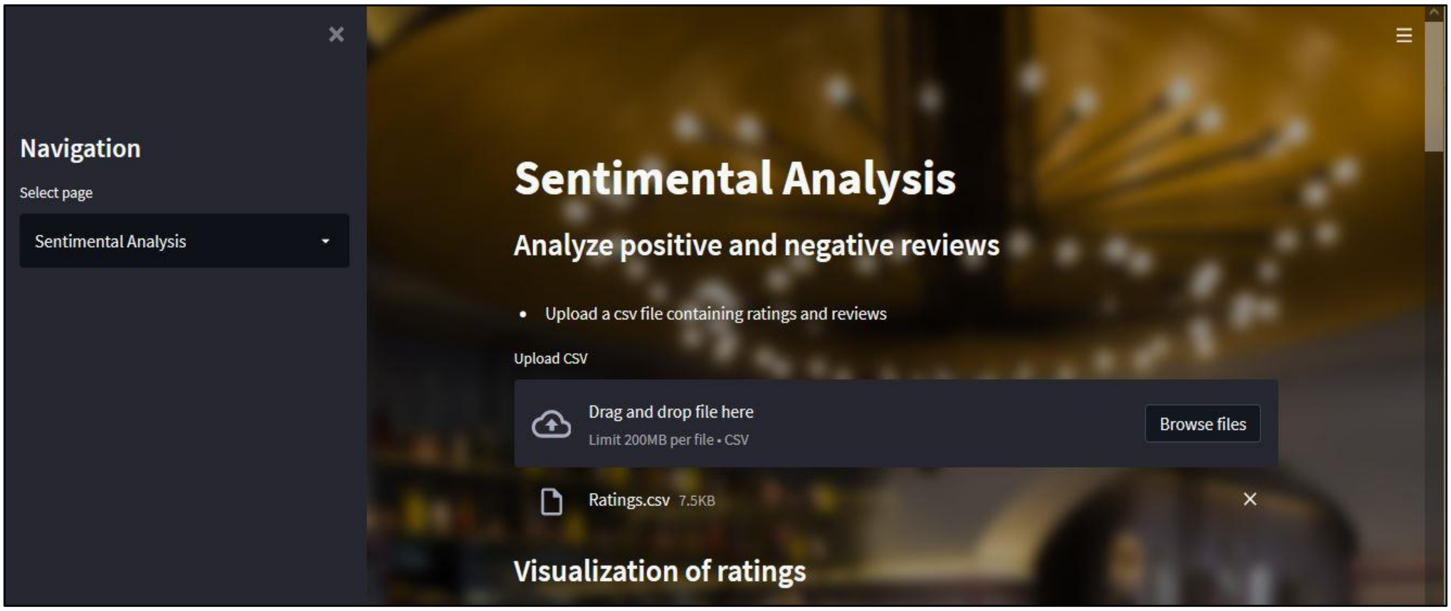
	Antecedents	Consequents
2	Pilau Rice	Bombay Aloo
4	Plain Papadum	Bombay Aloo
8	Pilau Rice	Chicken Tikka Masala
10	Plain Papadum	Chicken Tikka Masala
12	Pilau Rice	Garlic Naan
14	Plain Papadum	Garlic Naan
23	Pilau Rice	Mango Chutney
25	Plain Papadum	Mango Chutney

Figure 21 Apriori implementation interface

This can help the management to easily define the offers and implement other marketing strategies to enhance the sales of business

III. Sentiment Analysis.

To carry on a sentiment analysis, a CSV file which contains the rate and review are uploaded as an input.



Navigation

Select page

Sentimental Analysis

Sentimental Analysis

Analyze positive and negative reviews

- Upload a csv file containing ratings and reviews

Upload CSV

Drag and drop file here

Limit 200MB per file • CSV

Browse files

Ratings.csv 7.5KB

Visualization of ratings

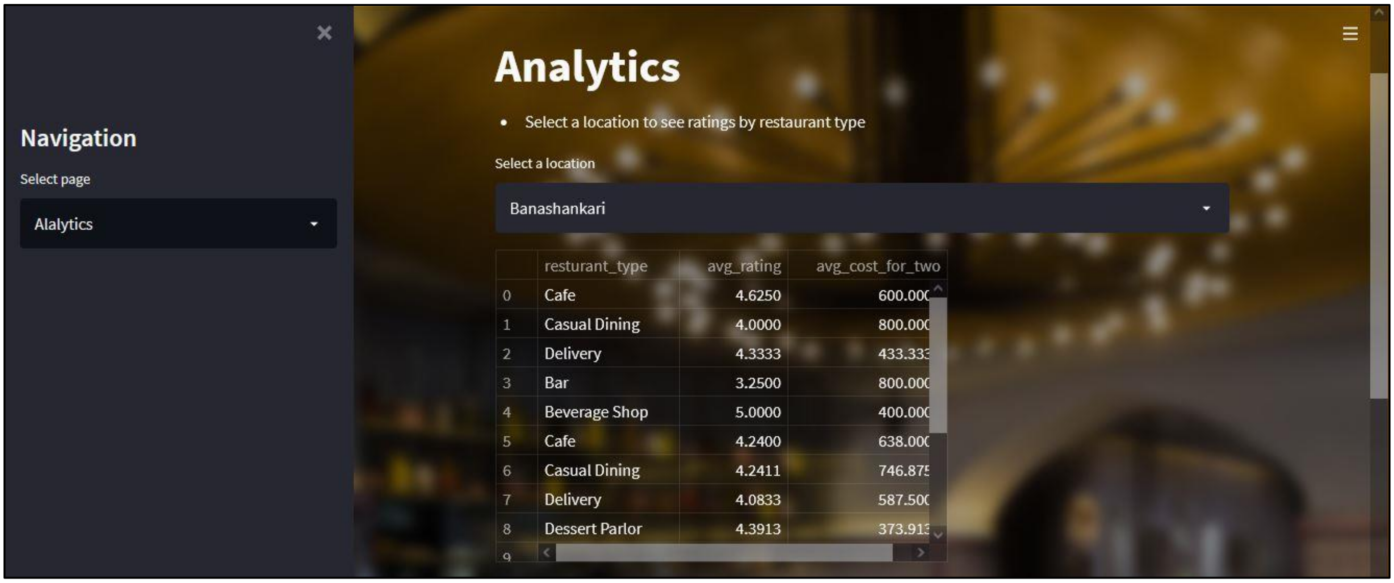
Figure 22 Sentiment Analysis Interface

By doing so the management can visualize the count of each rate category. In addition, the positive and negative reviews are highlighted here.

Word cloud representations help the business have a quick glimpse of the positive words represented in an enlarged manner whereas the negative words can be seen as small representations.

IV. Analytics

User can input the location of the restaurant to get some useful insights regarding the restaurant.



Analytics

- Select a location to see ratings by restaurant type

Select a location

Banashankari

	resturant_type	avg_rating	avg_cost_for_two
0	Cafe	4.6250	600.000
1	Casual Dining	4.0000	800.000
2	Delivery	4.3333	433.333
3	Bar	3.2500	800.000
4	Beverage Shop	5.0000	400.000
5	Cafe	4.2400	638.000
6	Casual Dining	4.2411	746.875
7	Delivery	4.0833	587.500
8	Dessert Parlor	4.3913	373.913
9			

Figure 23 Analytics Interface

By this analysis , the management can get an idea on the average cost incurred and the average rating they tend to receive thereby an overall idea on the business strategies that needs to be implemented can be obtained even before establishing a new restaurant in that location

5. Final Outcome

The final outcome of the project is an application which allows the top-level management to identify solutions for their business requirements as stated in problem identification. The application provides the capabilities for the top-level manager to input the required details and view the respective solution.

Initially the application enables the user to select the restaurant location simply to display the rates and costs by which the management can decide on which type of outlet can be opened at respective location to expand the restaurant chain. Further the application provides the capabilities of analyzing predictions and associations as well.

Accordingly, to address the requirement of analyzing the food patterns, the user can select the option of market basket analysis from the navigation pane and then view all the similar food patterns. The user also has the capability to filter the rules generated. Moreover, if the user has knowledge on data analysis, they can also adjust the metrics such as confidence and support thresholds to generate rules with variations. Advantage of this is that the management can decide on promotional packages and offers that they can provide to increase the customer base. If the management requires to increase the ratings provided by the customer, the application helps in identifying the areas to work on. The user can provide the inputs required and view the ratings predicted according to the given inputs. This helps the management to decide on how each features influence the rating which is predicted by the classification model. The application also has the capabilities of addressing the sentimental analysis too. Simply by uploading a file, the management can view the number of positive and negative reviews. Moreover, the app generates a word cloud image with keywords of reviews which will enable the decision makers to focus more easily on features influencing the positivity and negativity of the reviews.

As a conclusion, the web application is very much important for the organization to implement marketing strategies and identify the areas to improve on to increase the customer base. Ultimately, it signifies the importance of the high-level business development through data mining using predictive and descriptive techniques.

6. GitHub Link

Data Repository - https://github.com/kovisha/FDM_2021

Streamlit implementation - https://github.com/salitha10/Restaurant-Analytics

7. Deployed Link

https://share.streamlit.io/salitha10/restaurant-analytics/main/app.py

8. Video Link

https://1drv.ms/v/s!Apan3A7DSx97gzfHca_f1HdMynjR?e=Pdengj

9. References

https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397

https://towardsdatascience.com/machine-learning-multiclass-classification-with-imbalanced-data-set-29f6a177c1a

https://www.scikit-yb.org/en/latest/api/classifier/classification_report.html

https://towardsdatascience.com/machine-learning-multiclass-classification-with-imbalanced-data-set-29f6a177c1a