Cross-Cultural Comparison of Food Appetite and Restaurant Features

BRAC
UNIVERSITY

Inspiring Excellence

MD SAMI UL ISLAM - 12201033

MD SHARIFUL HUQ ABIB - 14201059

Supervisor

Samiul Islam

Lecturer

Co-Supervisor

Hossain Arif

Assistant Professor

A THESIS

Submitted to the School of Engineering and Computer Science,
BRAC University, in partial fulfillment of the requirements for the degree of
Bachelor of Science in Computer Science and Engineering

Declaration

We hereby declare that this thesis is completely based on the results obtained from our own work. Proper acknowledgment has been made in this paper while using all other material. This thesis, neither in whole or in part, has been submitted to any other University or Institute for the award of any Degree or Diploma.

Signature of Supervisor

Samiul Islam
Lecturer
Computer Science & Engineering
BRAC University

Signature of Co-Supervisor

Hossain Arif
Assistant Professor
Computer Science & Engineering
BRAC University

Signature of Authors

Md Sami Ul Islam

Md. Shariful Huq Abib

ACKNOWLEDGMENTS

This paper is the result of our extreme hard work, patience, and thoughts. This journey was not easy after maintaining every other aspect of life and giving the proper amount of time here to establish this kind of work.

Our journey towards the Bachelor degree would not have been possible without the help of many people. It is our great pleasure to take this opportunity to thank them for the support and advice that we received.

We are grateful to our Supervisor Mr. Samiul Islam (Lecturer) and our Co-Supervisor Mr. Hossain Arif (Assistant Professor) who kept faith in us at the first place and guide us and pushed us to the next level to ensure the minimum quality.

We would like to thank our parents for inspiring us in every step of our lives, supporting us spiritually and providing us with the best educations, our families gave us the strength to complete this degree. This thesis is dedicated to them

To our parents, sisters and brothers

TABLE OF CONTENTS

List OF Figures

ABSTRACT

Food business has expanded drastically over the years. It’s safe to say that you never have to make a loss for a restaurant business today onwards. People have become food lovers and they are always in search of new dishes and different tastes. From business meetings to engagement proposals to school assignments- all are happening over a cup of coffee or during lunch or dinner. So, it’s worth a try to understand this food culture globally. Because food represents the culture. Taste and method represent the history of a country. Price-variation tells us about the overall economy. Watching these scenarios nowadays, we got encouraged to study on people around the world to find out where they are on the same page in terms of food consumption. Which types of cuisines got popular worldwide and what are the reasons. Are the prices and environments similar to the same type of cuisines everywhere? We have adequate data of great variation to analyze these factors. We will take the help of statistic formulas and several machine learning techniques to accomplish our project.

CHAPTER 1

1.1 Introduction

Nowadays food business is booming rapidly. There is practically no one in this era who does not like foods. Humans are the top of the food hierarchy, they always search for something different & exquisite. People have differences of choosing food of their daily life base on their culture, livelihood, region, weather etc. Some prefer spicy foods others prefer something more raw and fresh. As of today, there are more than 7 billion people is breathing in this world with individual taste buds which may change over time. So it is very difficult to know the exact food preferences of all. If we can make a system or process which can calculate on the peoples' taste behavior on a region so that it can help in to predict the food preferences. If a person wants to open up a restaurant or food parlor in a certain location, he needs to have a clear idea of what the people favor in that particular region. He even needs to know about the money people want to spend on foods. That's where our theory comes in. We are developing such a method which studies on peoples food behaviors based on their earning, weather, culture etc. and establish reliable statistics to predict the food pattern.

In an article [1] we observe that they did multiple surveys and quaternary to find out the result. Some researchers even tried conjoint methods [2] about customers' preferences or even executed SERVQUAL" scale [3] to measure the deviation of "Service Quality" between customers' expectation of service and the actual service provided. But these research were done in other developed countries. As per our region, weather & culture, our food choices are different from other. So in our country, we need a slightly different method to triangulate our food preferences more accurately.

Our main goal is to locate a certain area, where the restaurants are, then we study cuisines and their prices. Not only the food but also we take the services the restaurants are offering. We also consider some other elements like table booking, online deliveries etc. We will use statistic formulas and some machine learning techniques here. This gives us more accurate output than the previous researchers.

1.2 Aim of Study

Who doesn't love food? People's schedules are getting tight and it's really difficult most of the time for them to cook food in home after the amount of toiling they have in work. Hence, they look for online delivery in restaurants.

Food business is growing rapidly and to make a loss in this sector is very unlikely. People now want to try and taste different things and want to review them by sharing their thoughts. Here basically happens 2 things - (i) Customer can share his opinion, (ii) Other potential customers can come seeing the review.

This is why the starters or the restaurant owners have to understand the market and the present thought process of the customers. They need to analyze the major factors that make the most impact if they want to be successful in this business.

In this paper we have showed that how some factors can shift the rating of restaurants up and down. We believe this research should help both the customers and owners.

1.3 Thesis Outline

The rest of this report is organized as below:

- In chapter 2, previous works related to this research have been discussed.
- In chapter 3, data description, our proposed model and work processes have been explained briefly.
- Chapter 4 represents results, findings, important factors and challenges.
- Chapter 5 concludes the paper with future works and general remarks.

CHAPTER 2

2.1 Background Study

Lots of Multiculturalism and ethnic groups are present in the United States [1]. So, restaurateurs of US have to provide a variety of food from different countries. People nowadays also love to discover new foods and try different tastes. They are also interested in trying dishes of other countries and culture. Thai foods have become very famous and popular all over the world. This paper tried to show that the majority of the customers like to get the authentic taste of the food rather than the restaurant environment and the interior. They did a survey and distributed questionnaires to a total of 250 people including young, middle-aged and old people who had served in fast-food, buffet and sitting arrangements. They asked questions about customer perception, expectation, and acceptance of food and service. The correlation was used to determine relations between the factors. Their results showed that the higher the income, higher the tendency of people to eat outside. They also used mean and standard deviation to find out personal choices of what most people feel. They found that USA people prefer Thai food way more than other foods, though there were some limitations like people's tastes differ from location to location and it covered some segments of customers, locations and services.

In the second paper [2], they implemented the conjoint method to achieve the customer preference for restaurants. The survey was examined in Hong Kong but they believe it is similar for other countries. The respondents were of different ages and background. There are more than one factors which play a role in choosing a restaurant for customers. These factors work independently but they focused more on the total points of the factors. Because customers make trade-off while choosing a cuisine. So, they listed the preferred factors for customers in order that they like to see in a food court. The results showed some are happy just with the taste of the food while for some others the environment and interior are also important. Interestingly, it is also found that people are less likely to go out with family than business meetings and friends. This paper approached communal analysis based on some major factors such as discrete, linear, ideal and anti-ideal. This research's design and findings of the perception of descending was applied in the study to help reveal the restaurant preferences.

Service quality and behavior are fundamental aspects to get success in the restaurant business. Profit and expansion of business are involved with these two key factors. This [3] study observes the relationship among service quality, food quality, perceived value, customer satisfaction, and behavioral intentions. Besides, the environment and the reliability, friendliness, responsiveness and apologizing manner of employees are not avoided by the customers. So, they play a huge part in customer service. These variables were treated independently where the gender key was a dependent factor. Their accuracy was 66.00% in classifying gender-based customers. They used "SERVQUAL" scale to measure the deviation of "Service Quality" between customers' expectation of service and the actual service provided. It was a 7-point scale from 1 to 7 which took points from customers for different variables. The results show improvement of food quality boosts customer satisfaction.

Retaining customers is more profitable than gaining new customers. This paper [4] examined what pushes the customer to visit a restaurant again and how can this tendency be increased more. They noticed that some factors are involved here. Of course, food quality, price, service, atmosphere are key terms here. Social Media is a big phenomenon nowadays for information sharing of a variety of restaurants. Customers get interested in something by seeing the reviews and comments from visiting customer. It is also helpful for restaurant managers and owners to go through the reviews and improve customer service if there is any room for it. This paper used an original model called "DINESERV" as well as used text mining technology to analyze reviews of customers and understand their revisit intention. The ratio between price and benefit of the foods is considered as consumers' value. They developed 8 hypotheses which were based on the variables on which a customer decides to visit again. Then they collected online reviews and did regression analysis on it. In their result, service quality got most points; followed by the atmosphere, food quality, and price. Interestingly, from the reviews, more than 90% mentioned about the tastes of food where the rest of the tiny portion of people discussed the healthiness of food.

The last paper [5] that we studied identifies that people's food habit is being influenced over time. During consumer research, they found that people's implications & preferences of food changes. They explored that sub-cultural food consumption is different like fashion and trends. This paper is work in elaborating how cultural capital expresses different tastes. Europe plays a big role in food consumptions of people all over the world. They showed that food taste preferences are related to cultural development. The geographical, economic and historical contexts influenced people's culture and taste. Even people's wealth and consumption of sweets made an impact on their preferences and globalization. This paper covered the micro and macro-environmental attributes all over Europe. The ingredients used and the way food is cooked that certainly convey the lifestyle, thought process, progress and culture of a nation. Food habits also differ based on people's religion.

CHAPTER 3

3.1 Data Description

We have collected our dataset from Kaggle – a rich resource for the data scientists. The data are still very new and has been posted 4 months ago.

This data is about types of restaurants and cuisines all over the world. We can fetch exact location by the ‘longitude’ and ‘latitude’ that are available in our dataset along with the country and city names. We also found the average cost for 2 people in terms of currencies of respective countries.

We have observed our data and found a number of important features that will help us analyzing further. We tried to visualize the data along with its attributes by plotting some bar diagrams and pie charts. We considered 7 features to be most important to understand the distribution and range of our data.

Now, on the next page, we have shown the state of our data that we found out during our research.

3.2 Features

3.2.1 Countries

Variety of countries are there in the data. The number of restaurants in different countries is plotted here. Restaurants from 15 countries covering all the continents across the world are here. As we can see, this dataset largely contains restaurants from India of almost 9,000. The rest of them individually gather below 500.

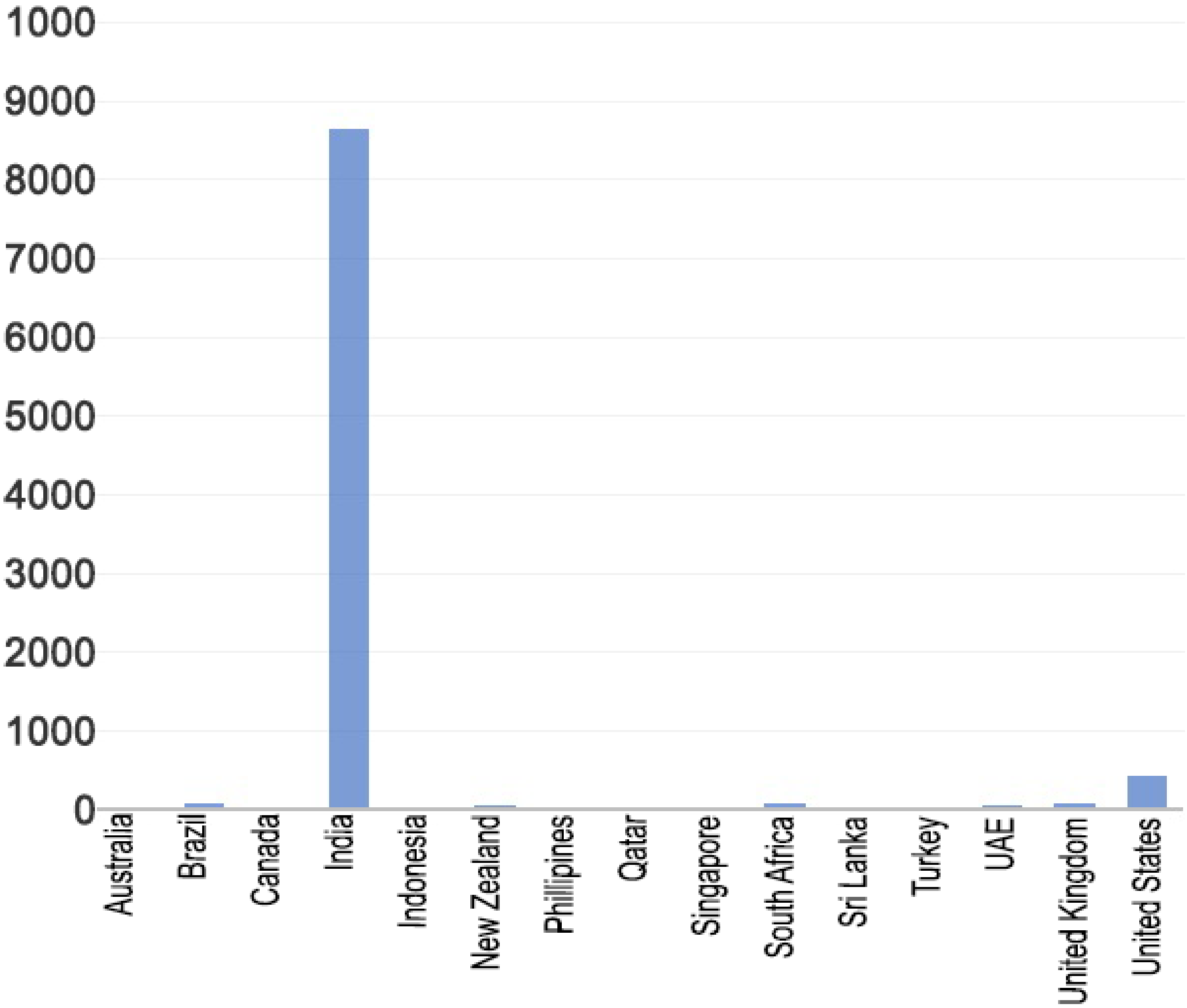


Fig 1: Bar Diagram of Data Distribution based on Countries

3.2.2 Cuisines

The types of restaurants which are available in our dataset are plotted here. No surprises, Indian cuisines are the most as we have seen this dataset massively contains restaurants from India and moreover Indian cuisines are also popular beyond India. The 2nd most cuisines we have all over the world are mainly local fast food shop.

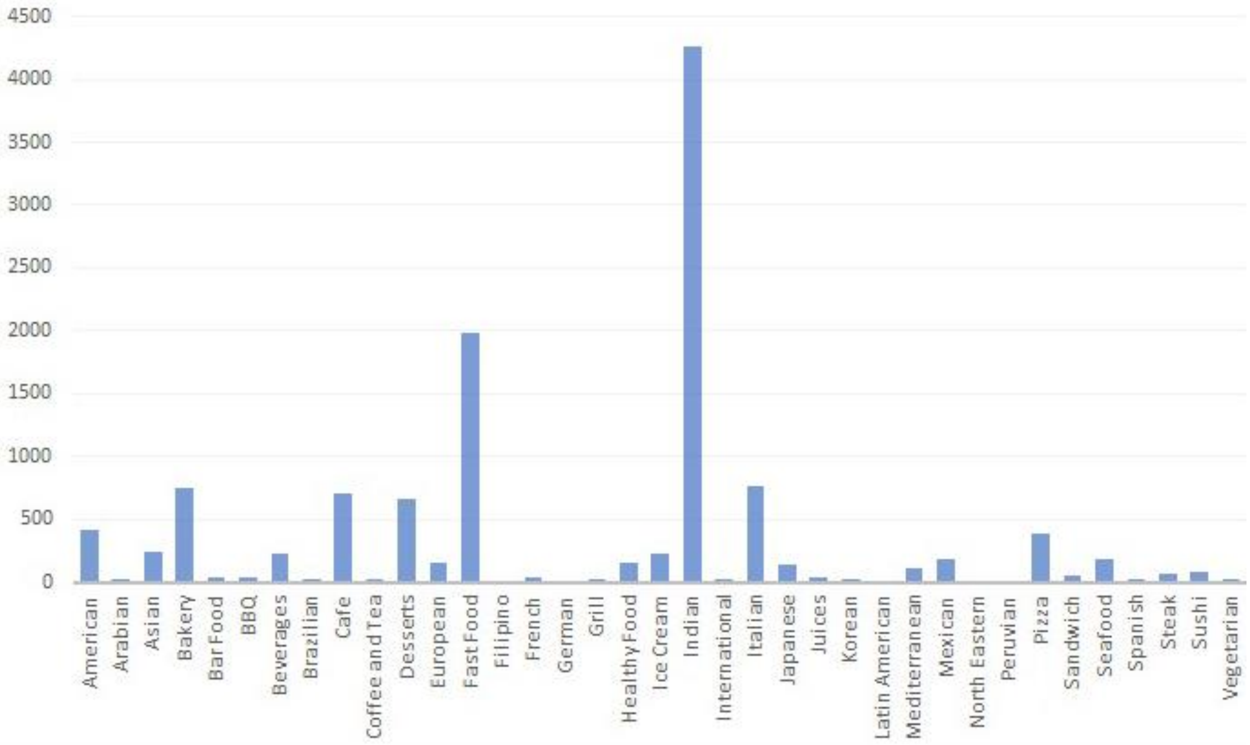


Fig 2: Bar Diagram of Data Distribution based on Cuisines

3.2.3 Price Range

The average cost for 2 people is here. Although our main data has currencies of respective countries, we converted them all into US Dollars to compare the actual value. We can picture that the major portion of our data has restaurants with price-range around and above 3-7 USD.

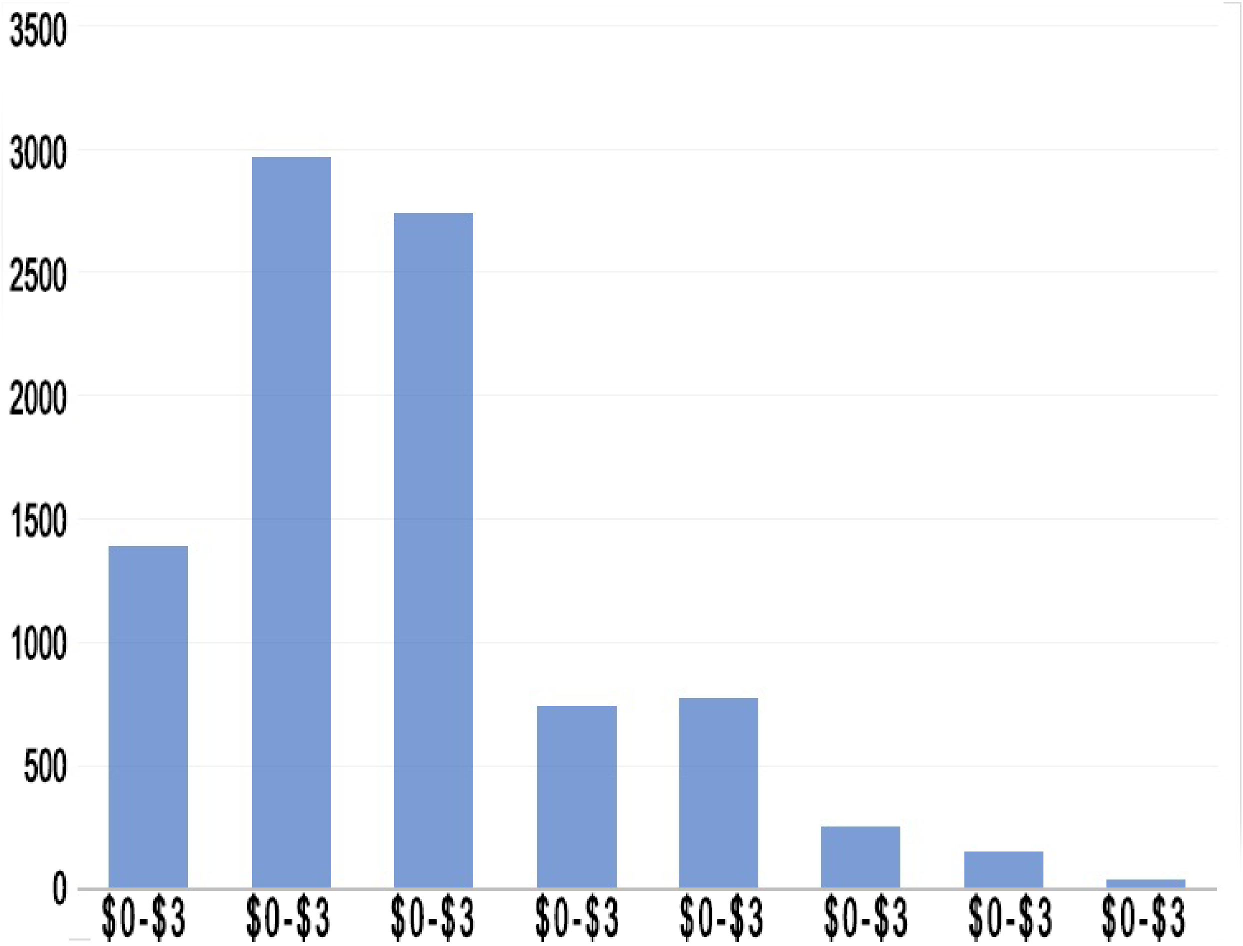


Fig 3: Bar Diagram of Data Distribution based on Price-range

3.2.4 Table Booking

The next important feature we considered is whether the restaurant has a booking facility or not. We see that a huge part of the restaurants of our data does not allow this service. Remember, this is because, in our dataset, 2nd most numbers of cuisines belong to local fast-food shops.

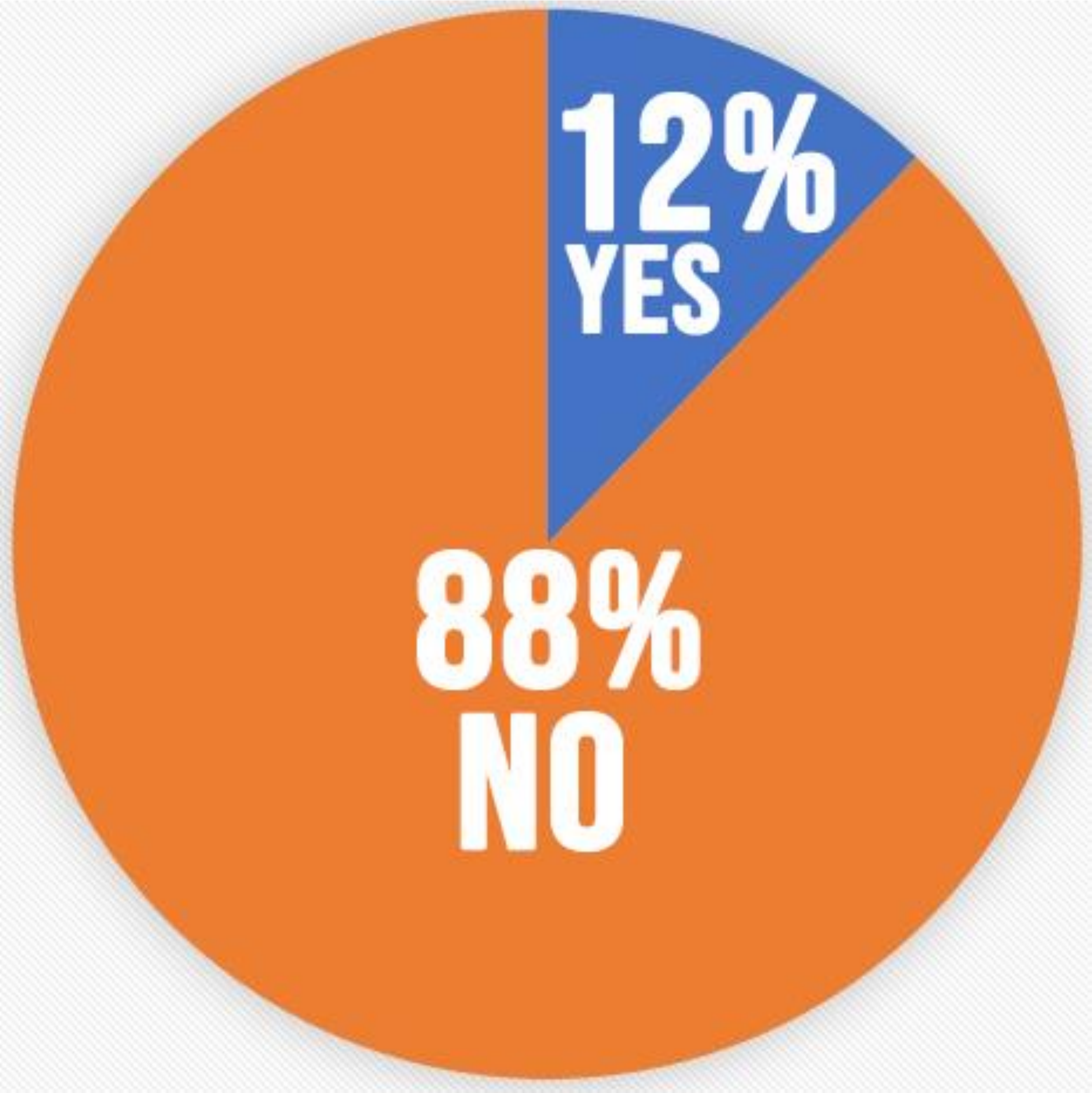


Fig 4: Pie Chart of Data Distributions of Cuisines based on Booking Facility

3.2.5 Online Delivery

Most of them don’t adopt online delivery with an amount of 74%.

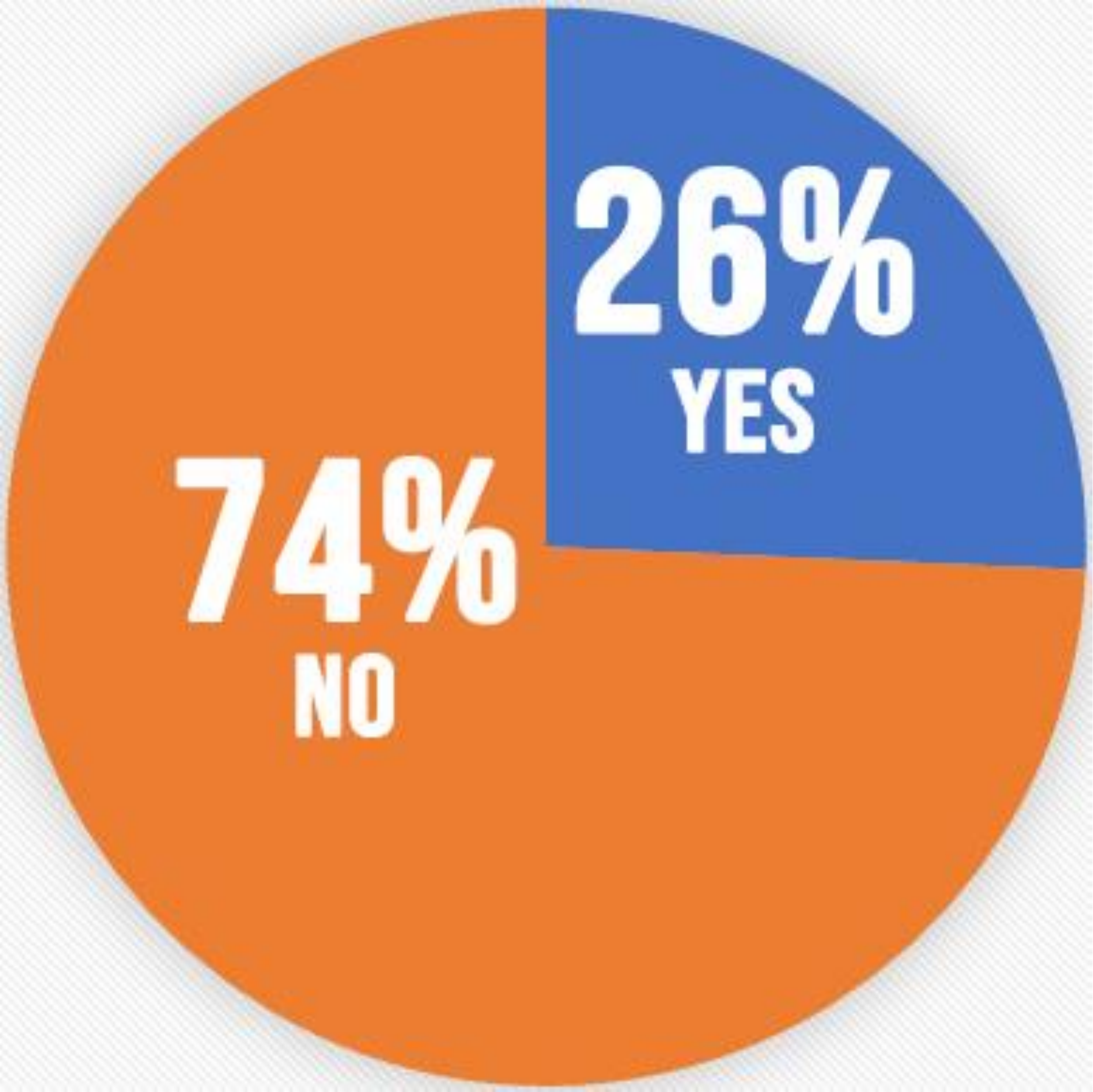


Fig 5: Pie Chart of Data Distribution based on Availability of Online Delivery

3.2.6 Rated or Not

This is the section where we can understand that the numbers that are not rated are the ones that don’t have an online social media profile to be rated. Easy enough to know obviously most of them will have the facility as 80% of the total cuisines of our dataset are rated by people.

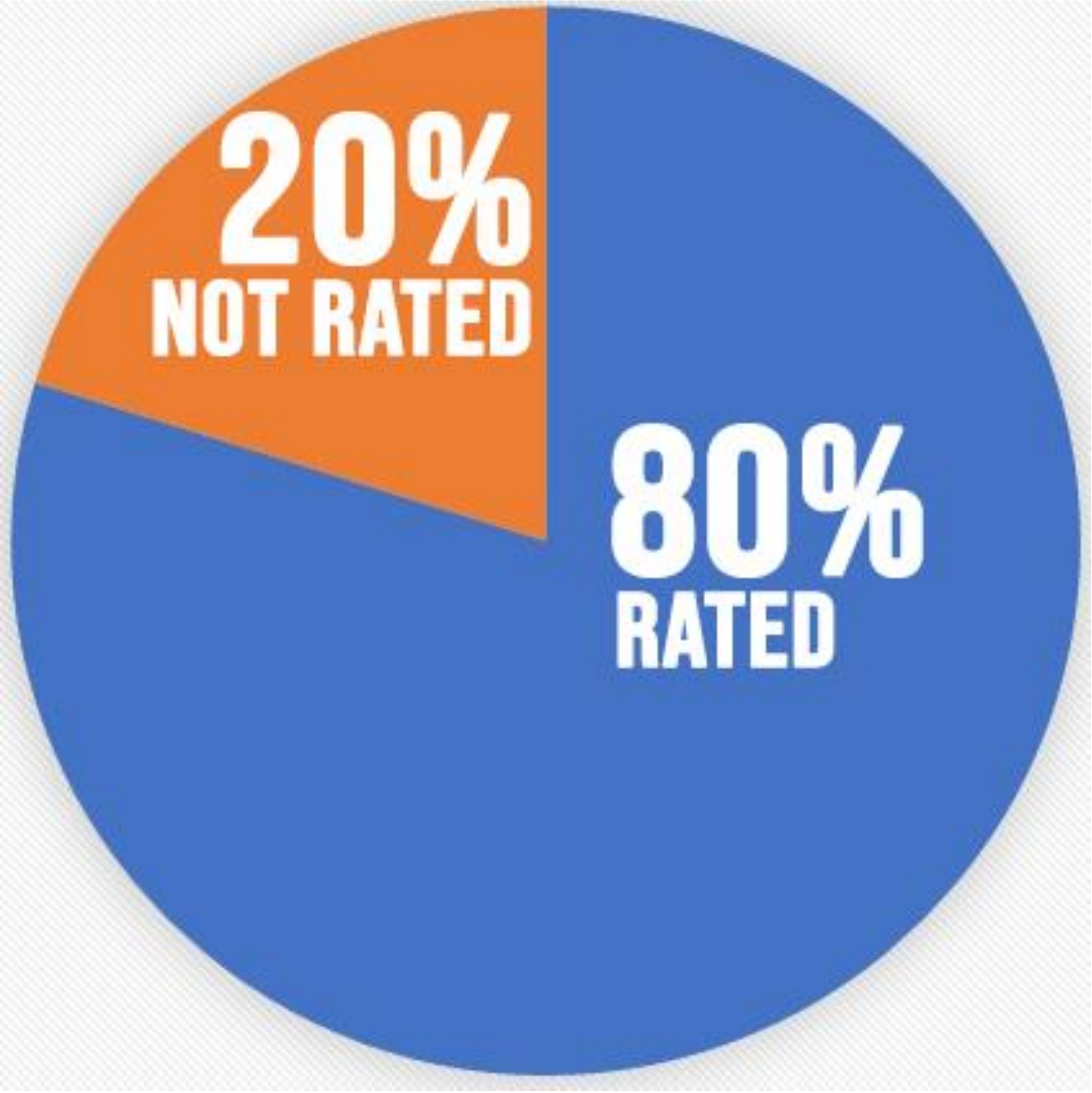


Fig 6: Pie Chart of Data Distribution of Cuisines whether Rated or Not

3.2.7 Rating Points

This is one of the most important findings of this research. Most of the restaurants got reviews as Average or good where the very little amount is Poor and some is Excellent. The dataset actually contains rating points in numbers and those are also available in texts. We just picked the rating-texts since they both mean the same.

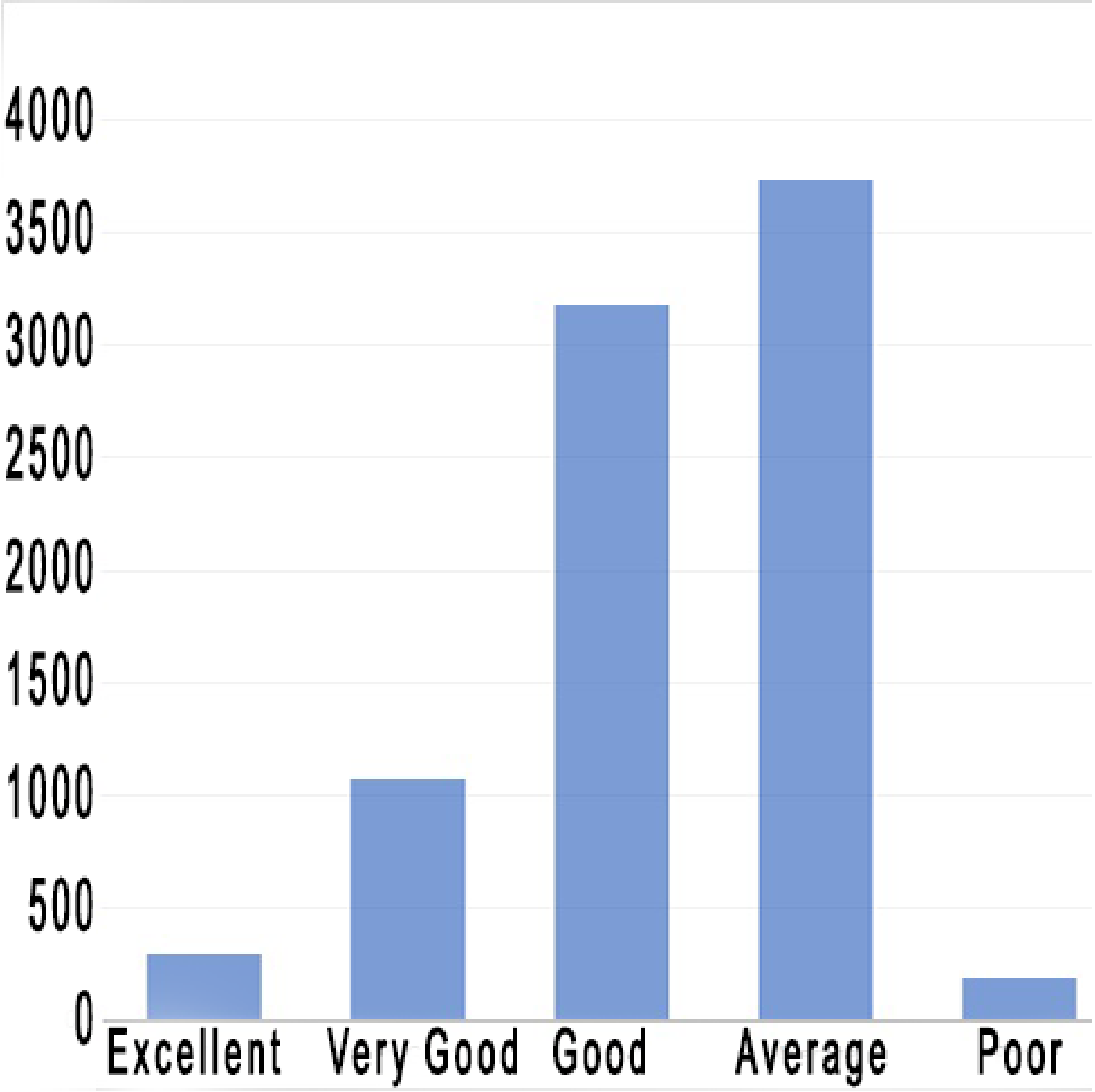


Fig 7: Bar Diagram of Data Distribution of Cuisines based on Ratings

CHAPTER 4

4.1 Hypothesis of System Implementation

This research is a part of data mining concept. So, we will implement machine learning techniques to achieve our goal.

Our scopes of findings are to evaluate comparisons:

1. Cultural-wise, 2. Cuisine-wise, 3. Price-wise, 4. Rating-wise

The workflow of our research is shown below that how we can do it to complete this journey.

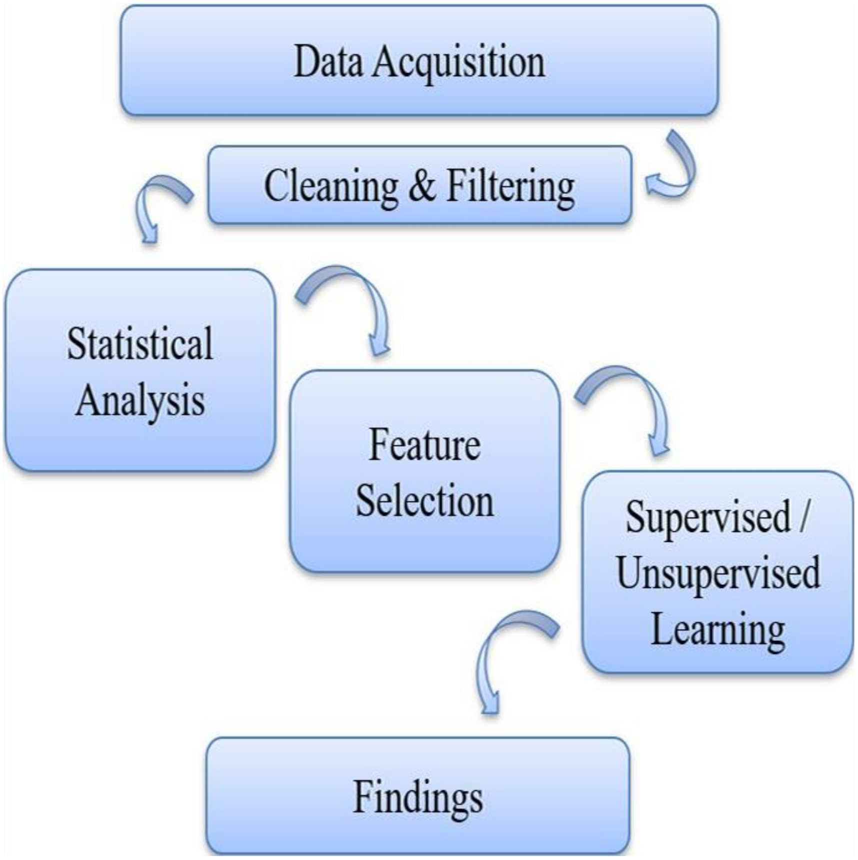


Fig 8: Proposed Model for Machine Learning Analysis

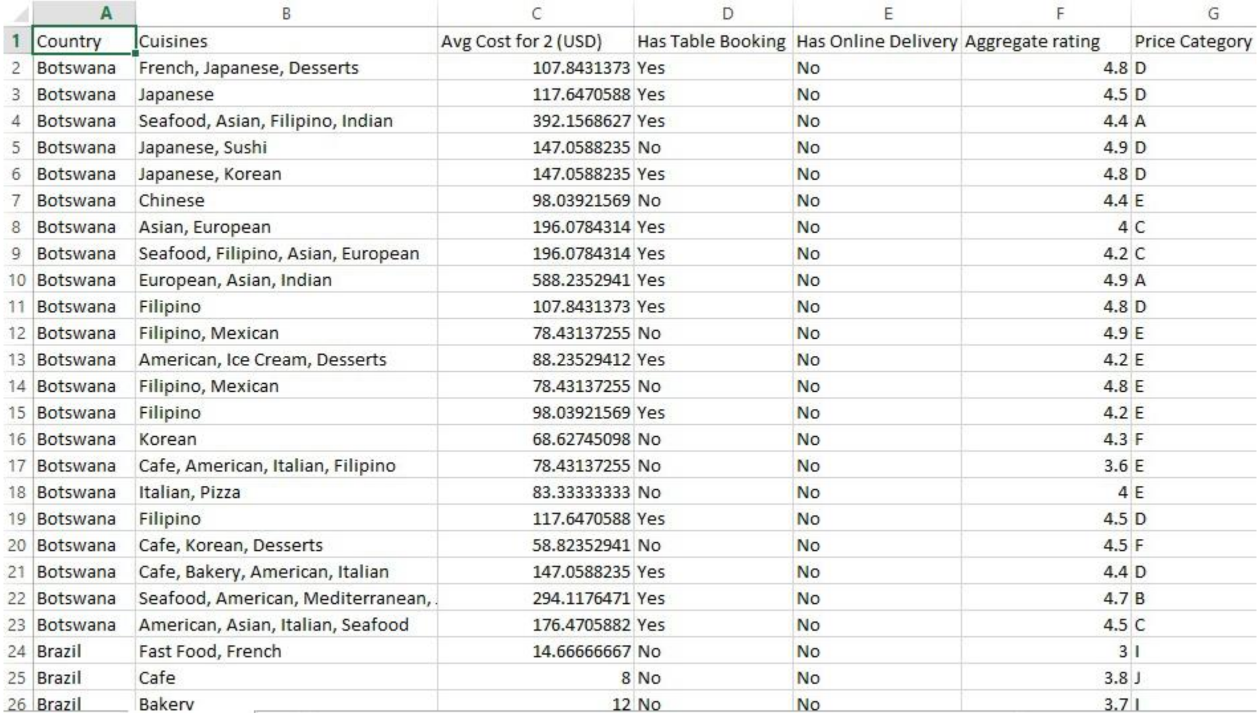
4.2 Data Handling

4.2.1. Data Cleaning

First thing first. We addressed that our data contains too much needless information those won’t help us or will have any impact on our research. So, we removed those parts and also removed the missing values and outliers. Finally, our dataset looked good to go.

4.2.2. Adding Extra Feature

We added a new column in our dataset naming “Price Category” which is associated with the column “Average Cost for 2 (USD)”. We basically categorized the costs in the new column so that it becomes easy to understand in which category the price has fallen since there is huge price discrepancy in our data.



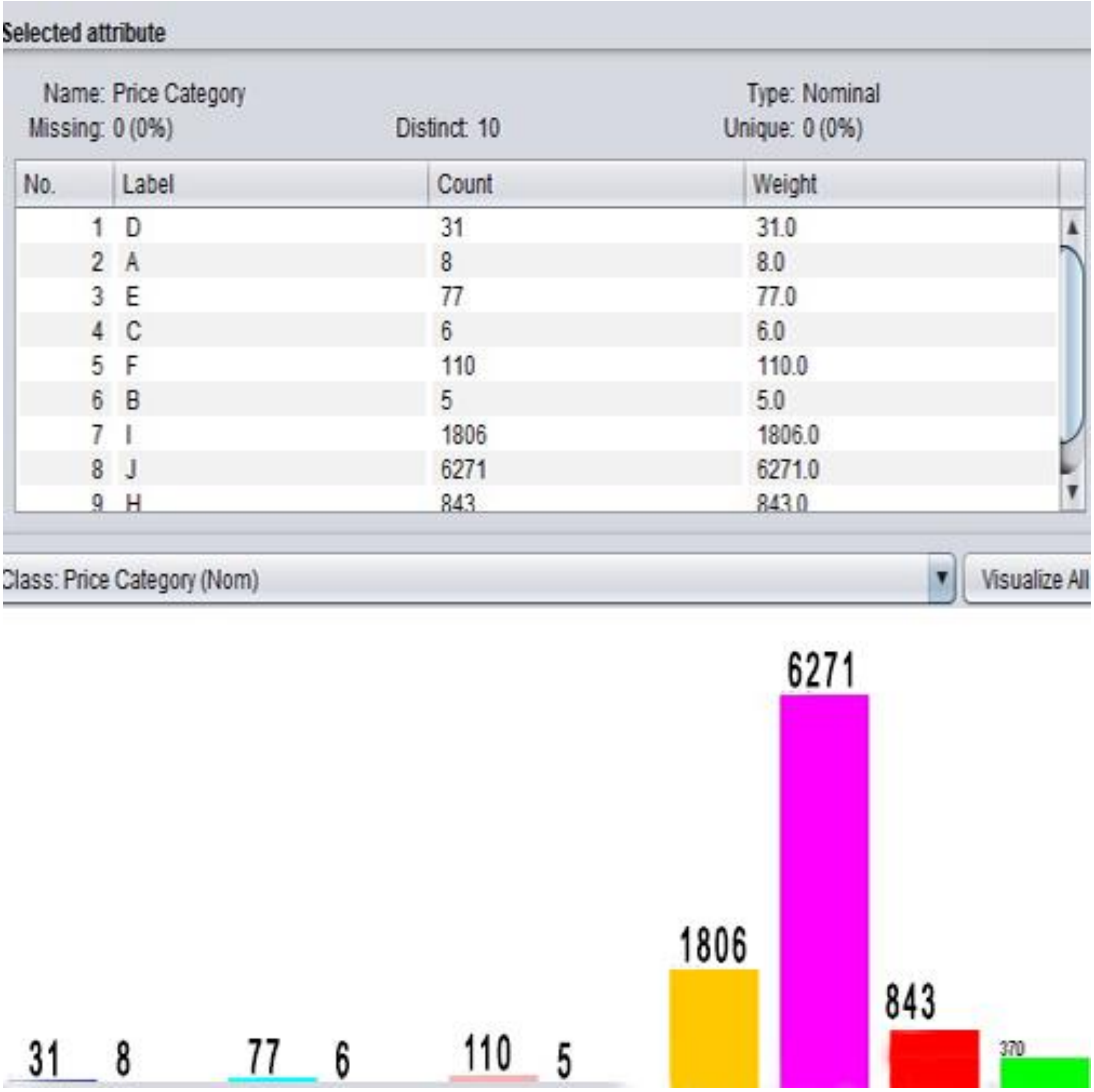
	A	B	C	D	E	F	G
1	Country	Cuisines	Avg Cost for 2 (USD)	Has Table Booking	Has Online Delivery	Aggregate rating	Price Category
2	Botswana	French, Japanese, Desserts	107.8431373	Yes	No	4.8	D
3	Botswana	Japanese	117.6470588	Yes	No	4.5	D
4	Botswana	Seafood, Asian, Filipino, Indian	392.1568627	Yes	No	4.4	A
5	Botswana	Japanese, Sushi	147.0588235	No	No	4.9	D
6	Botswana	Japanese, Korean	147.0588235	Yes	No	4.8	D
7	Botswana	Chinese	98.03921569	No	No	4.4	E
8	Botswana	Asian, European	196.0784314	Yes	No	4	C
9	Botswana	Seafood, Filipino, Asian, European	196.0784314	Yes	No	4.2	C
10	Botswana	European, Asian, Indian	588.2352941	Yes	No	4.9	A
11	Botswana	Filipino	107.8431373	Yes	No	4.8	D
12	Botswana	Filipino, Mexican	78.43137255	No	No	4.9	E
13	Botswana	American, Ice Cream, Desserts	88.23529412	Yes	No	4.2	E
14	Botswana	Filipino, Mexican	78.43137255	No	No	4.8	E
15	Botswana	Filipino	98.03921569	Yes	No	4.2	E
16	Botswana	Korean	68.62745098	No	No	4.3	F
17	Botswana	Cafe, American, Italian, Filipino	78.43137255	No	No	3.6	E
18	Botswana	Italian, Pizza	83.33333333	No	No	4	E
19	Botswana	Filipino	117.6470588	Yes	No	4.5	D
20	Botswana	Cafe, Korean, Desserts	58.82352941	No	No	4.5	F
21	Botswana	Cafe, Bakery, American, Italian	147.0588235	Yes	No	4.4	D
22	Botswana	Seafood, American, Mediterranean, .	294.1176471	Yes	No	4.7	B
23	Botswana	American, Asian, Italian, Seafood	176.4705882	Yes	No	4.5	C
24	Brazil	Fast Food, French	14.66666667	No	No	3	I
25	Brazil	Cafe	8	No	No	3.8	J
26	Brazil	Bakery	12	No	No	3.7	I

Fig 9: New Look of the Dataset

4.3 Data Visualization

4.3.1 Data Pre-processing

Now the time has come to visualize our freshly looked data. We imported our data in the Machine-learning tool “Weka” and selected the attribute “Price Category.



Selected attribute

Name: Price Category
Missing: 0 (0%)
Distinct: 10
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	D	31	31.0
2	A	8	8.0
3	E	77	77.0
4	C	6	6.0
5	F	110	110.0
6	B	5	5.0
7	I	1806	1806.0
8	J	6271	6271.0
9	H	843	843.0

Class: Price Category (Nom)

Fig 10: Data Points of “Price Category”

We can see that this feature has 10 different nominal values starting from A to J. We cleared this before that this feature is the subdivision of another attribute “Average Cost for 2 (USD)” from our dataset. Prices from 300 and above are categorized as ‘A’. 200 and above are set as ‘B’, 150 and above areas ‘C’, 100 and above are as ‘D’, 70 and above are as ‘E’, 50 and above are as ‘F’, 30 and above are as ‘G’, 20 and above are as ‘H’, 10 and above are as ‘I’ and finally below 10 are labeled as ‘J’.

Now, this is understandable that most of the prices will fall under category ‘J’ and ‘I’ which covers 6271 and 1806 instances respectively among about 9000 examples.

4.3.2 Selecting Data Points

Next, we selected all the attributes in terms of “Price Category” to observe the scene. It’s easy to understand that the Pink and Yellow colors are the Price Category of J and me which are the majority among other classes.

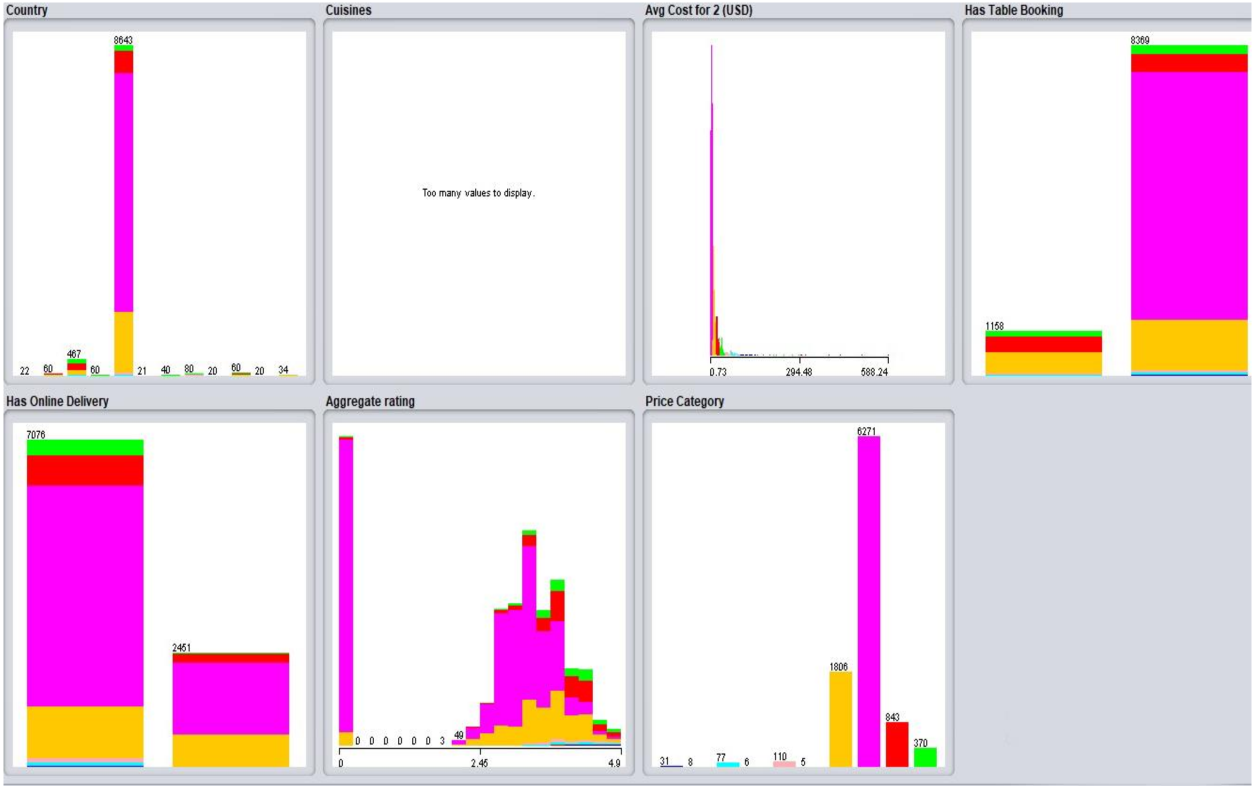


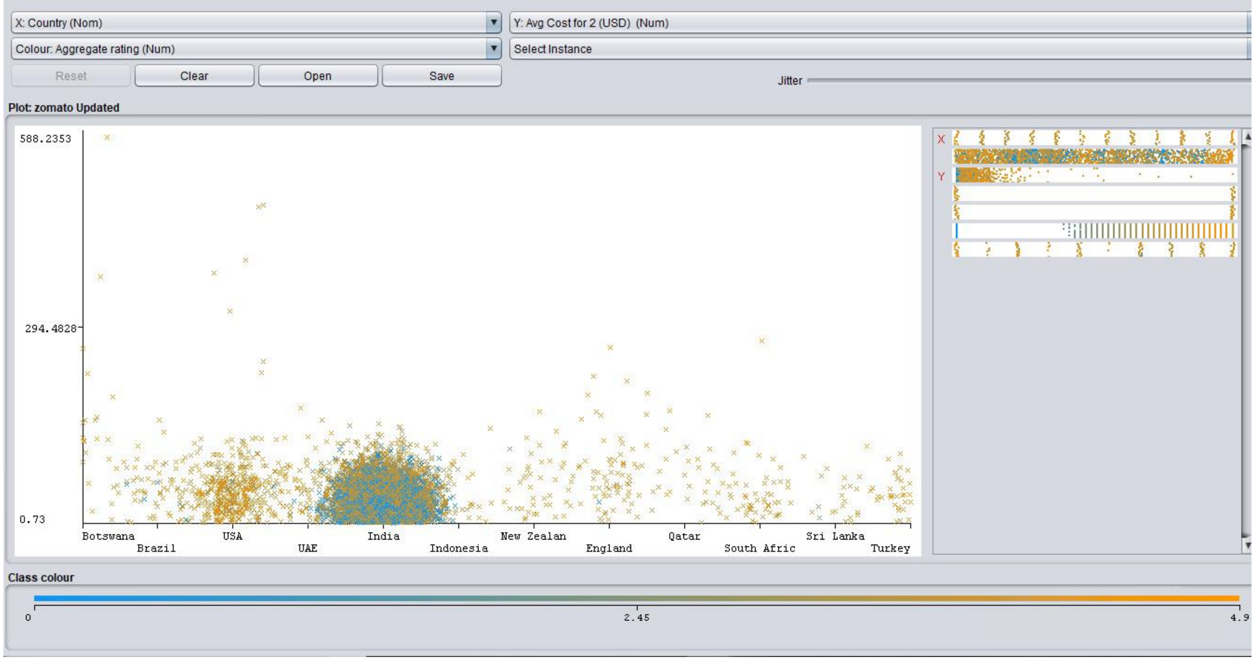
Fig 11: Data Points of all attributes based on “Price Category”

Here, one thing is noticeable that the box “Cuisine” is Empty. The reason for this is there are too many classes for this feature in the dataset. For example, you can imagine the type of restaurants and bars around the world. Japanese, French, Chinese, Italian, American, Indian, Pizza, Burgers, Desserts, Ice-creams, Kebab, Café, Bakery, Continental, Asian and so on. So, it was difficult to tie them down in a single frame.

4.3.3 Visualizing Data

We can now see the relationships among the attributes to gain more knowledge about our data.

In Fig-12, we have “Country” on the X-plane and “Average Cost for 2 (USD)” on the Y-plane. The colorful scattered dots which are visible are the “Aggregate Rating.” We can see most of the points are yellowish as most of the ratings is in and around 4. Interestingly, a huge portion of points is bluish too since a number of cuisines were not rated due to having any rating service and they eventually got the rating 0 in the dataset.



X: Country (Nom)

Y: Avg Cost for 2 (USD) (Num)

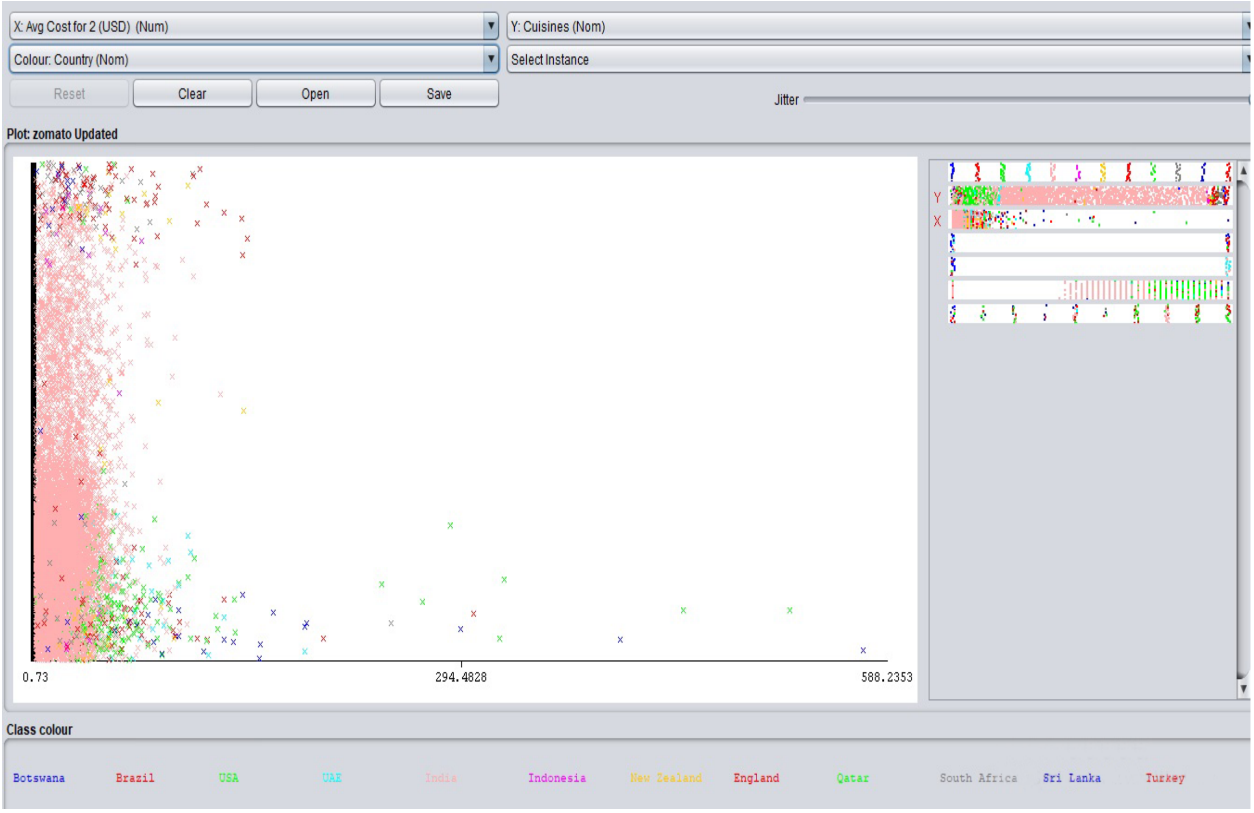
Colour: Aggregate rating (Num)

Select Instance

Plot: zomato Updated

Fig 12: Relating Countries with Cost in terms of Rating

In Fig-13, this time on the X-axis, we have “Average Cost for 2 (USD)” and on the Y-axis we have “Cuisines”. As we can see nothing shown in the picture beside Y-axis because there are too many values of cuisines to show. The colorful scattered points describe countries. Most of them are pink as they are denoting the country India. No surprises, the pink dots will take all over the place since 8,000 of 9,000 data in our dataset are from the country India.



X: Avg Cost for 2 (USD) (Num)

Y: Cuisines (Nom)

Colour: Country (Nom)

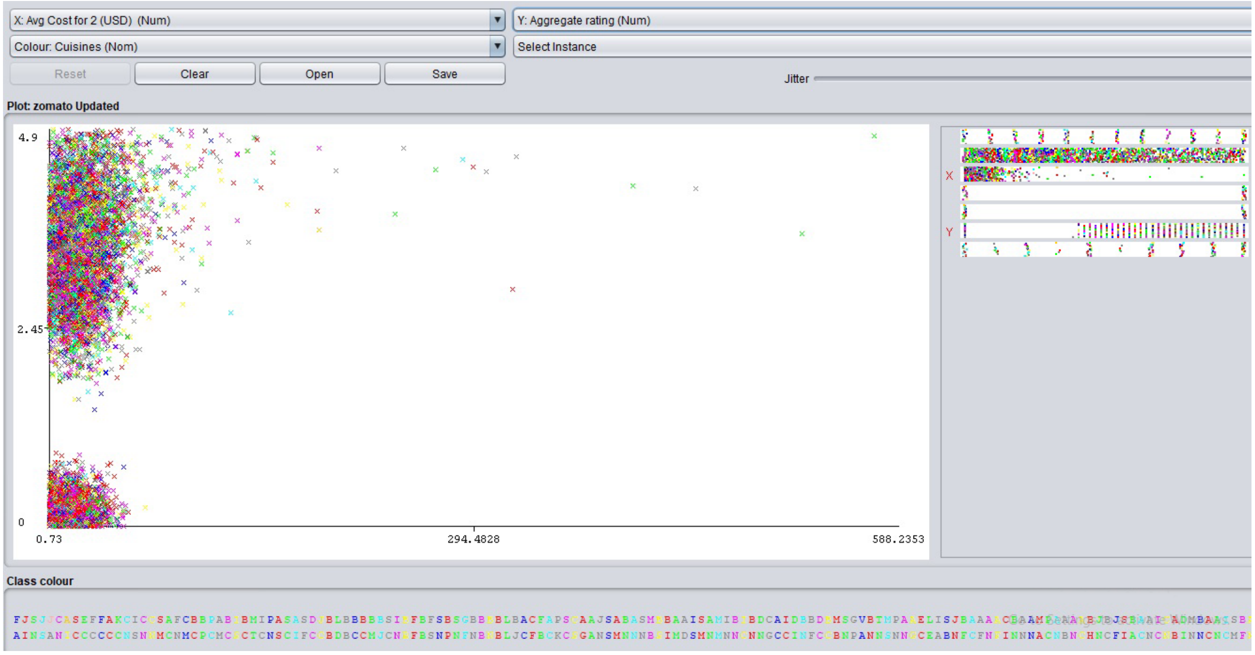
Select Instance

Plot: zomato Updated

Fig 13: Relating Cost with Cuisines in terms of Countries

In Fig-14, we have “Average Cost for 2 (USD)” on the X-axis and “Aggregate Rating” on the Y-axis. This time the colorful scattered points are representing the attribute “Cuisines.” As we can see, there are a huge amount of colors naturally because of the number of cuisines we have in our dataset.

Besides, you can imagine from the graph that most of the restaurant bills fall under under-10 to 30 USD for 2 people.



X: Avg Cost for 2 (USD) (Num)

Y: Aggregate rating (Num)

Colour: Cuisines (Nom)

Fig 14: Relating Cost with Rating in terms of Cuisines

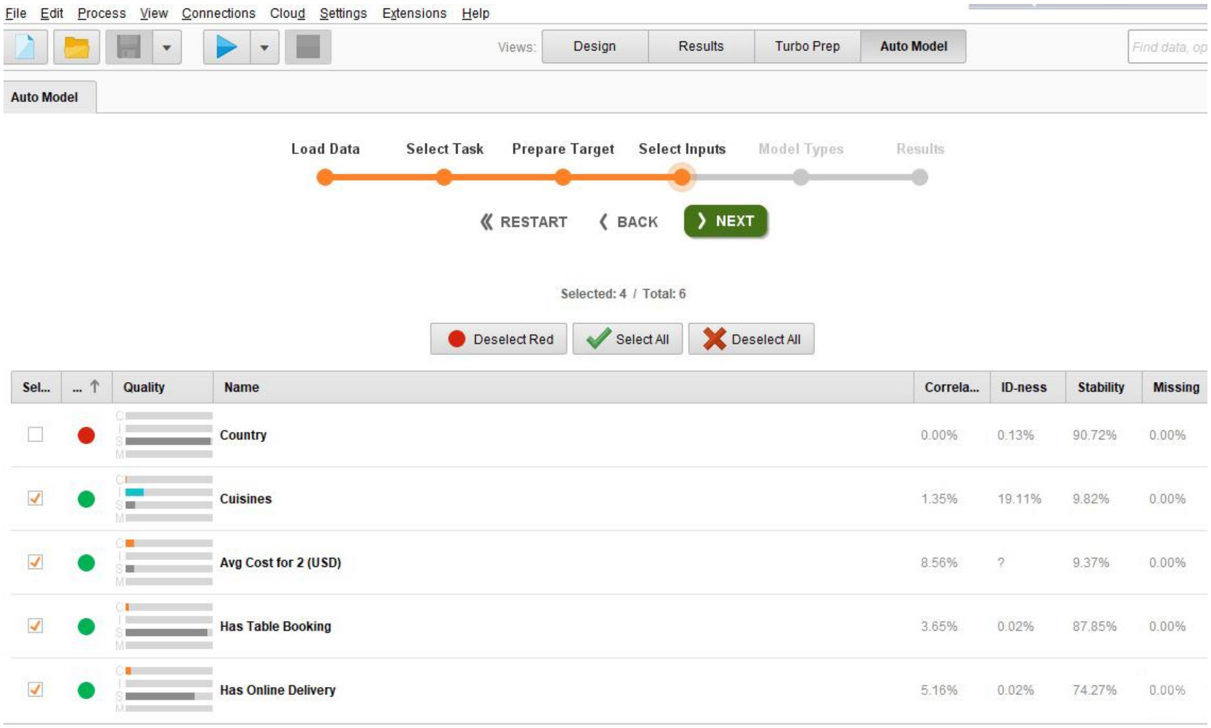
4.4 Technique & Algorithms

Now since we have full idea about our data, we can dig deep and find the hidden elements within our data.

We will try to predict the “Aggregate Rating” using correlation and regression in other attributes. We have taken the help of a strong and powerful tool “Rapidminer” to perform this task.

We have taken into account several methods to try things out and check which one performed the best.

First, we have imported our data into “Rapidminer” and selected the column that we want to predict. In this case, it was the numeric column “Aggregate Rating.” Next, this tool suggested which attributes could be ignored and which are most helpful.



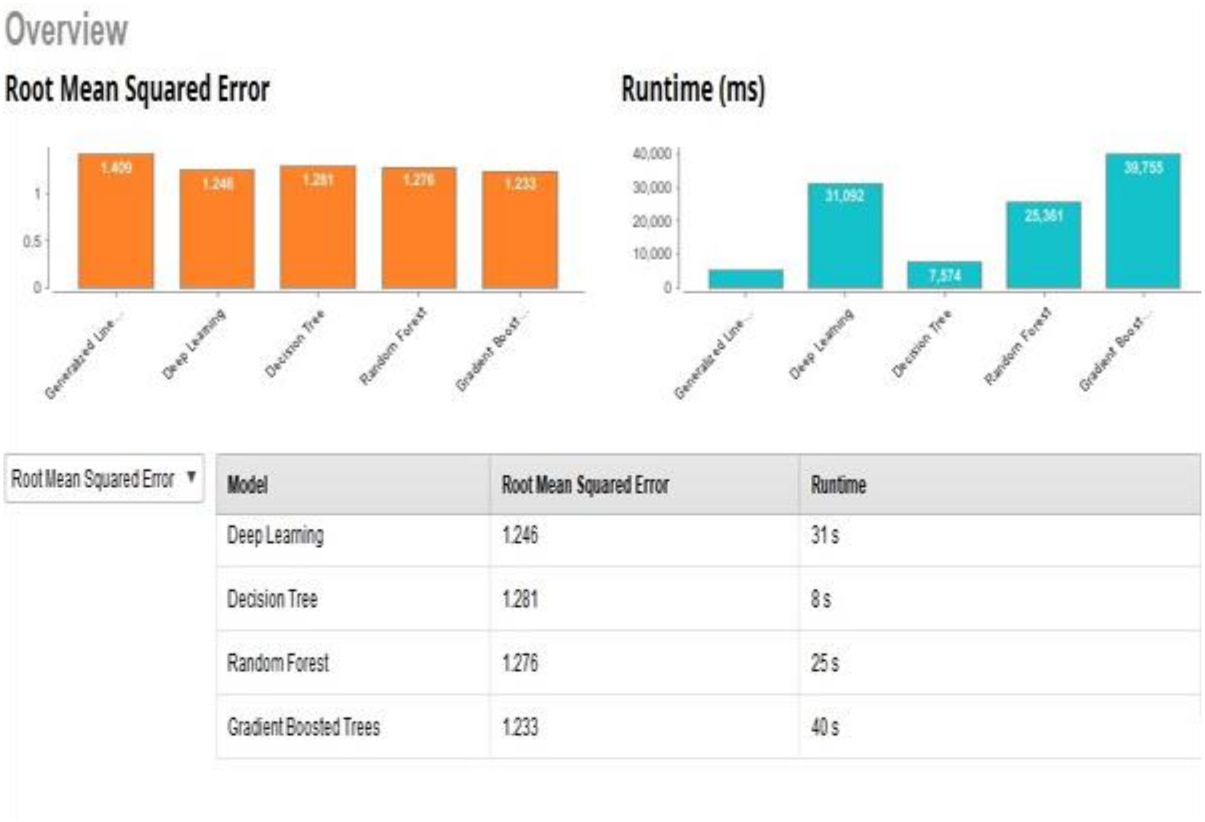
Selected: 4 / Total: 6

Sel...	... ↑	Quality	Name	Correla...	ID-ness	Stability	Missing
□	🟥	C I S M	Country	0.00%	0.13%	90.72%	0.00%
☑	🟩	C I S M	Cuisines	1.35%	19.11%	9.82%	0.00%
☑	🟩	C I S M	Avg Cost for 2 (USD)	8.56%	?	9.37%	0.00%
☑	🟩	C I S M	Has Table Booking	3.65%	0.02%	87.85%	0.00%
☑	🟩	C I S M	Has Online Delivery	5.16%	0.02%	74.27%	0.00%

Fig 15: Selecting Inputs

In “Fig 15,” the red circle is suggested as “Not Useful” or misleading feature. And the green circles are signaled as the right features or the important ones to be picked.

Then we selected the Model Types, played the “Run” button and waited for the results.



Model	Root Mean Squared Error	Runtime
Deep Learning	1.246	31 s
Decision Tree	1.281	8 s
Random Forest	1.276	25 s
Gradient Boosted Trees	1.233	40 s

Fig 16: Results of applied Algorithms

We applied Generalized Linear Model, Deep Learning, Decision Tree, Random Forest, and Gradient Boosted Trees in our data and among them, **Gradient Boosted Trees** performed best with the Root Mean Squared Error of 1.233. However, it had the most run-time of 40 seconds. But to get the most accurate result, we can compromise the run-time and wait for the best prediction rate.

4.5 Findings

We picked the result of "Gradient Boosted Trees" which gave us the best output. We analyzed the correlations among the attributes that the algorithm decided to be most suitable.

The term "Correlation" means how closely two things are related. Correlation can either be positive or negative. Correlation can be from -1 to +1. A correlation of -1 between 2 factors means they have a strong negative relation between them and +1 means they have a strong positive relationship. Any value around 0 means the attributes have a weak relation between them and if a value becomes 0, that means there is no relation between those features.

A positive correlation between two things means they are going in the same direction. For example, we can say, if the demand for any product becomes higher, the price gets higher. So, demand and price are positively correlated parameters here. That means if one increases, the other also increases or if one decreases the other one also decreases.

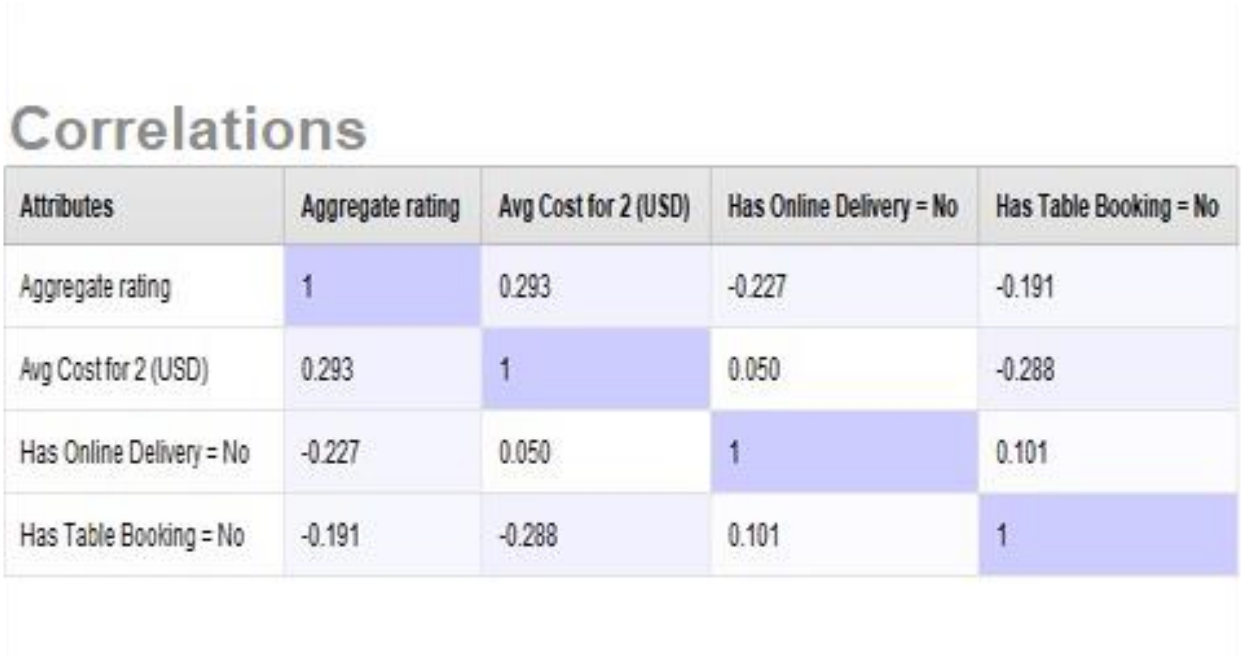
On the other hand, the negative correlation between two attributes means they are moving in the opposite direction. Which means if one increases, the other one decrease and vice-versa. For example, we can think of Supply-Price relation. If foods that rot away easily have a huge amount of supply, then the price for them gets dropped and if the supply is less than the demand, the price automatically gets higher.

In Fig-17, if we look at the 1st row next to the header section, we can see the correlation between "Aggregate Rating" and "Aggregating Rating" is 1 which describes a strong positive correlation. This is because these 2 factors basically are the same.

Next, between "Aggregate Rating" and "Average Cost for 2 (USD)" we got the best correlation value which is +0.293. This means ratings of restaurants are usually at the higher side if the cost is higher.

After that we get the relation of "Aggregate Rating" and "Has Online Delivery = No" which gives us a negative correlation value. This is because they are going the opposite direction as shown here if there is no online delivery, the rating will be less and when we have online delivery for customers, the rating will be clearly higher.

Similarly, when we have a table booking facility, the rating will go up and when there is no table booking, chances are the ratings are dropped.



Correlations

Attributes	Aggregate rating	Avg Cost for 2 (USD)	Has Online Delivery = No	Has Table Booking = No
Aggregate rating	1	0.293	-0.227	-0.191
Avg Cost for 2 (USD)	0.293	1	0.050	-0.288
Has Online Delivery = No	-0.227	0.050	1	0.101
Has Table Booking = No	-0.191	-0.288	0.101	1

Fig 17: Correlations among attributes for “Gradient Boosted Trees”

4.6 Important Factors That Worked

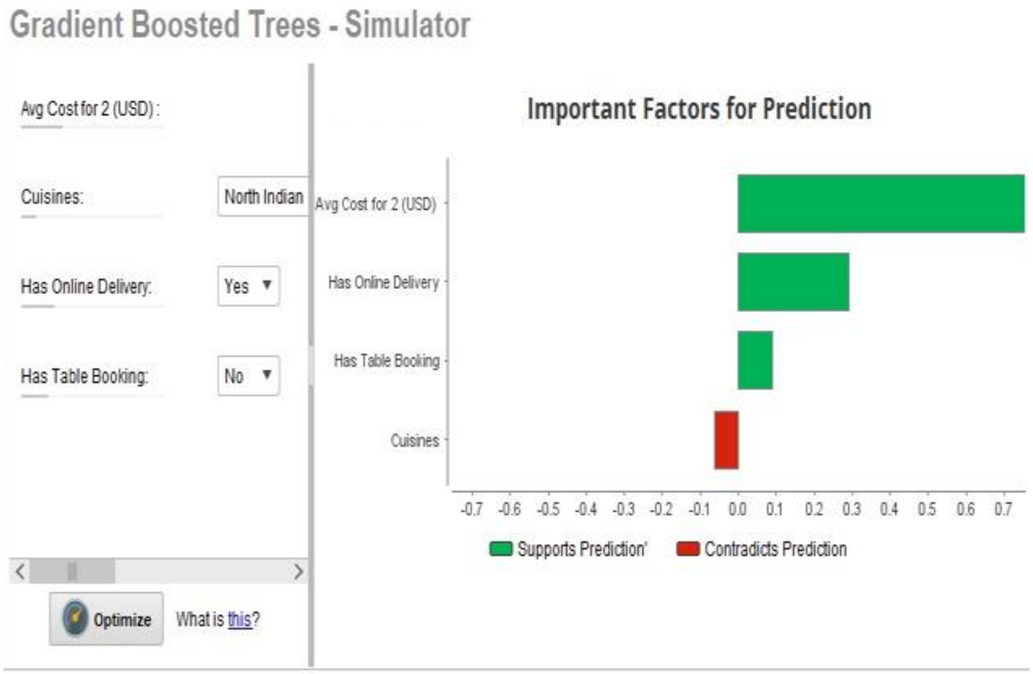
From Fig-18, as we can see the most important features that were decisive in this research are:

(1) Average Cost for 2 (USD)

(2) Has Online Delivery

(3) Has Table Booking

These 3 factors helped us in our research to make a prediction on “Aggregate Rating.” Whereas another feature “Cuisines” misguided the model to fetch the best output. This is why we skipped this feature just like the other ones which couldn’t help the model to perform the best way. The reason for ‘Cuisines” to be most contradictory is there are so many labels for this attribute and it was more than difficult to categorize it within a few labels.



Gradient Boosted Trees - Simulator

Avg Cost for 2 (USD) :

Cuisines: North Indian

Has Online Delivery: Yes ▼

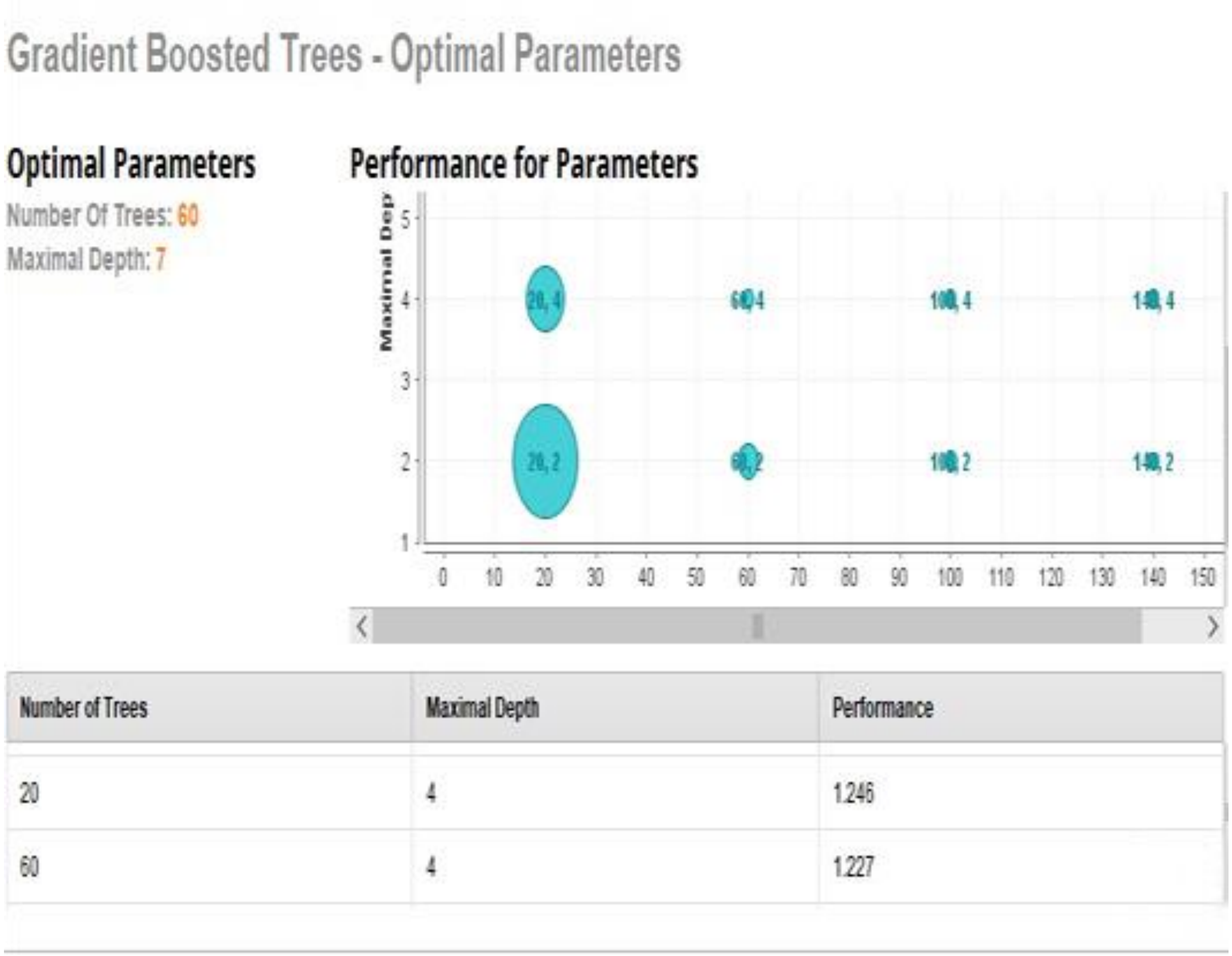
Has Table Booking: No ▼

Fig 18: Important Factors used in Gradient Boosted Trees

4.7 About the Model

The model that worked best for us is Gradient Boosted Trees. The model used 60 trees in total which had a maximum depth of 7.

From Fig-19, it is noticeable that with the number of trees increased, the performance of the model drops. The size of the circle in the picture defines the performance parameter. We can see that the circle became very small and hard to see as it travels with more number of trees towards the right side.



Gradient Boosted Trees - Optimal Parameters

Optimal Parameters

Number Of Trees: 60

Maximal Depth: 7

Performance for Parameters

Number of Trees	Maximal Depth	Performance
20	4	1.246
60	4	1.227

Fig 19: Performance Parameter for "Gradient Boosted Trees"

This is because with the number of trees increased, the model carries more information and it becomes time-consuming to process and extract the result. However, the more the information, the more accurate the model will produce a result.

This is why we often have to a trade-off between performance and the result or we can see between the run-time and the accuracy.

A part of the entire tree is shown below in Fig-20. As we can see the total number of trees in this model are 60 and it is quite impossible to fit them all together here and try to understand them.

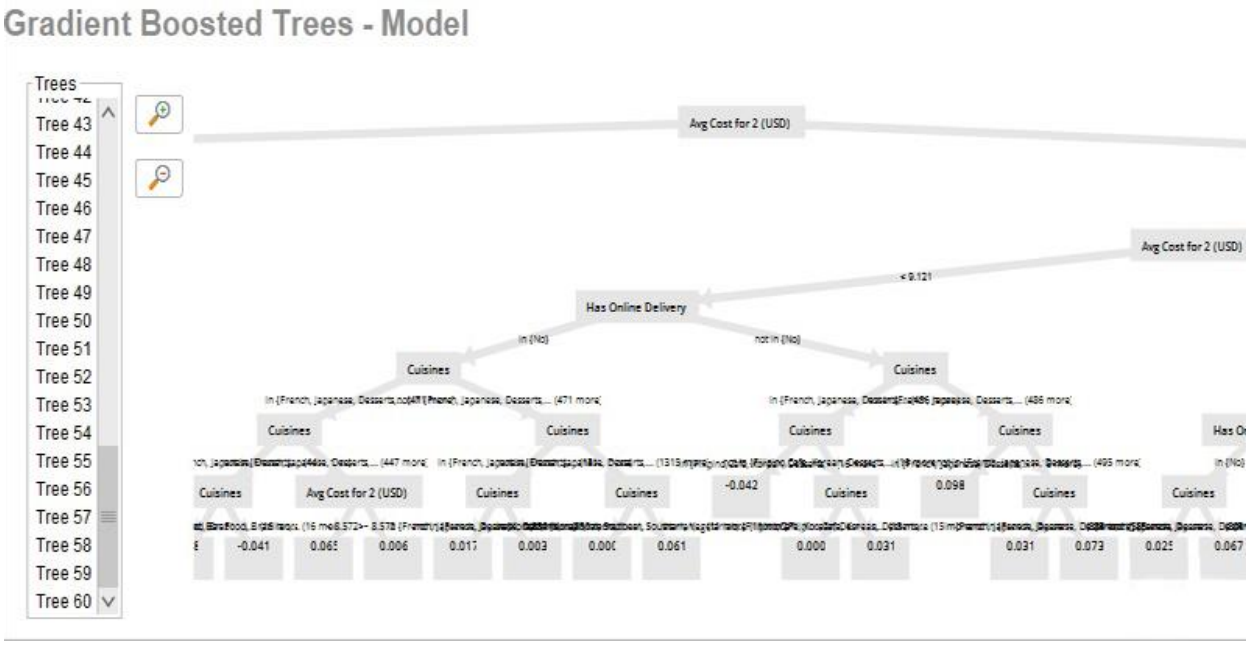
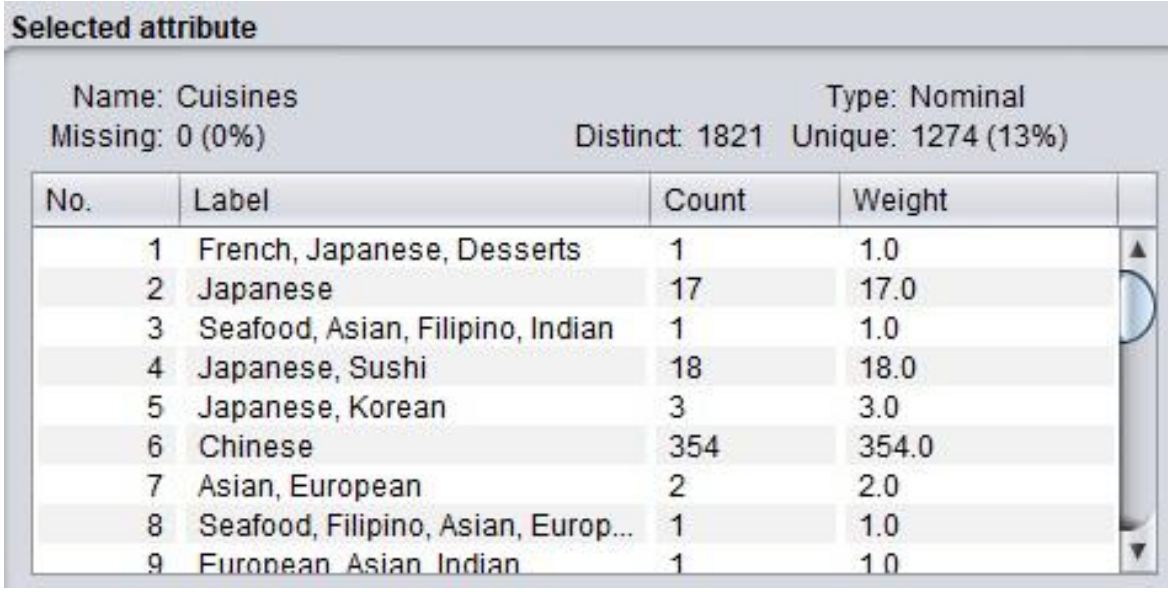


Fig 20: Image of the Gradient Boosted Trees (Total Trees 60, Depth 20)

4.8 Challenges

A major challenge was to categorize one of the most important features in our data - “Cuisines.” It could give us lots of interesting stats and more accurate predictions if it had limited numbers of labeling.



Selected attribute

Name: Cuisines
Missing: 0 (0%)
Distinct: 1821
Type: Nominal
Unique: 1274 (13%)

No.	Label	Count	Weight
1	French, Japanese, Desserts	1	1.0
2	Japanese	17	17.0
3	Seafood, Asian, Filipino, Indian	1	1.0
4	Japanese, Sushi	18	18.0
5	Japanese, Korean	3	3.0
6	Chinese	354	354.0
7	Asian, European	2	2.0
8	Seafood, Filipino, Asian, Europ...	1	1.0
9	European, Asian, Indian	1	1.0

Fig 21: Too many labels for the attribute "Cuisines"

Next challenge was "Aggregate Rating." Lots of cuisines are marked as rating 0 not because of the bad experience of the customers but because of not having the facility of the review system. So, it's a contradiction whether it is reviewed as 0 or there is no system to review. This was one of the reasons for which we can have a misleading or error-prone prediction.

CHAPTER 5

5.1 Future Works

Future works can be based on mitigating those challenges which are mentioned on the previous page. For example, trying to categorize and grouping the "Cuisines" under a few umbrellas so that it becomes easy to perform analytical tasks. Also trying to use other efficient algorithms can give more accurate outputs such as Naïve-Bayes Operator, KNN, Logistic Regression etc.

Besides, we have an ambitious plan to try and find out the economy of the countries that are in our dataset based on the cost of restaurant bills. Moreover, finding out the cultural differences based on the most popular cuisines in a country.

5.2 Conclusion

It's always interesting to analyze data and find out in-depth patterns which are not possible to understand in the naked eye or just visualizing the data. Performing data mining techniques can bring a new dimension and ideas. Prediction becomes easier if we understand and pick the correct inputs which are important and don't contradict with the output that we are trying to find out.

In this research, we basically did the correlation and regression method with the help of built-in statistical algorithms. We have identified the important features that are relevant to our goal which is always important to understand. After that cleaning, the missing data and misleading values are also the keys. The rest of the part is up to the machine to make the prediction based on our needs.

REFERENCES

[1] Piyavan Sukalakamala & Janice B. Boyce “Customer Perceptions for Expectations and Acceptance of an Authentic Dining Experience in Thai Restaurants”. First published: 09 May 2007; https://doi.org/10.1111/j.1745-4506.2007.00048.x

[2] L.C. Koo, Fredrick K.C. Tao & John H.C Yeung “Preferential Segmentation of Restaurant Attributes through Conjoint Analysis”. MCB UP Limited 1999; https://doi.org/10.1108/09596119910272784

[3] Hong Qin & Victor R. Prybutok “Service Quality, Customer Satisfaction and Behavioral Intentions in Fast-Food Restaurants”. Emerald Group Publishing Limited 2009; https://doi.org/10.1108/17566690910945886

4] Xiangbin Yan, Jing Wang & Michael Chau “Customer revisit intention to restaurants: Evidence from Online Reviews”. First online: 27 September 2013; https://doi.org/10.1007/s10796-013-9446-5

[5] Len Tiu Wright, Clive Nancarrow & Pamela M.H Kwok “Food Taste Preferences and Cultural Influences on Consumption”. British Food Journal, Vol. 103 Issue: 5, pp.348-357; https://doi.org/10.1108/00070700110396321

THE END