Analyzing and Predicting Features to Increase the user base for Zomato App

Final Project Report

Principal of Business Data Mining
INSY 5339, Sec-003
Professor Dr. Anam Sahoo

Group No: *Group-5*

Group Member's name:

1. *Faria Tasnim (1001967007)*
2. *Vishali Bairam (1002036395)*
3. *Chiriki Sai Goutham (1002018678)*
4. *Aditi Tibrewal (1001955627)*

Table of Contents

Executive Summary 1
Business Overview 2
Business Objectives 2
Source of Data Set 2
Brief Description of Dataset 2
Data Preparation 3
Data Visualization 4
Prediction Technique and Findings 9
Conclusion 13
Model Output 13
Reference 14

Executive Summary:

Online food ordering and delivery has become very popular in recent years. Specially during pandemic with work from home concept, food delivery business has grown significantly across the world and has a great future prospect in the business world.

For any online food delivery business their main revenue is earned through commission for every ordering done through their app or website and delivery fee. Without attracting additional revenue, a company cannot move forward for which they need to increase their user base. Food delivery companies also need to understand which features will attract new user base resulting additional revenue.

We have chosen a dataset collected by Zomato Bangalore for a particular day through their mobile app and we have performed few analyses to identify the key components which has significant impact in increasing their user base eventually contributing to revenue.

In addition to above, we have also identified an equation which online food delivery companies like Zomato can use to predict average cost based on the user selection of his location, total number of cuisines he is looking for, dinner ratings and reviews, delivery ratings and reviews, home delivery, indoor seating, veg only option on the Zomato app.

Business Overview:

Zomato is an Indian multinational restaurant aggregator and food delivery company.[1] Their business model is based on providing local restaurants search services and collecting data on food menus, contacts and providing relevant information to their customers on the mobile app and its online website.[2] Zomato focuses on the QAAA model, which promises and offers customers and partners quality, accessibility, accessibility, and portfolio. The following are the values of Zomato, making Zomato unique and admirable: Zomato works to maintain an impartial platform.[3]

Zomato earns their revenue in two folded way- one is through commission for every delivery from a restaurant and second one is through delivery fee.[4] Thus, user preferences and selection of choices assumes a vital part in deciding the development of the app which leads to increasing user base to the platform welcoming more revenues for Zomato.

Business Objectives:

For this group project, we are committed to analyze the data collected by Zomato through their mobile app of different user preferences compared to dinner and delivery ratings and reviews and average cost per restaurant.

- Selecting the most significant features/options that have strong relationship with the average cost.
- Based on our research and data analysis, we want to propose a best model for predicting average cost to attract new user base for the app.

Source of the dataset:

https://www.kaggle.com/datasets/vora1011/zomato-bangalore-restaurants-2022

Brief Description of the Dataset:

Our dataset was collected from Kaggle under the title "Zomato Bangalore Dataset" and has 19 variables and 8924 observations. This data set have 4 binary, 8 categorical, 2 continuous and 3 discrete variables. Our target variable is average cost. Data set contains dinner and delivery ratings and reviews, number of cuisines served and average cost. This data was collected on 26th March 2022 by Zomato Bangalore through their app. Below are the variable names and their respective descriptions:

Variable Name	Description
Name	Restaurant name
URL	Restaurant website url
Cuisines	Cuisine name for the restaurant

Areas	Area name where restaurant is situated
Timings	Open hours of restaurant
Full_Address	Location of the restaurant in detail
Phone Number	Contact Number of restaurant
isHomeDelivery	Does the restaurant have home delivery option-Yes/No
isTakeaway	Does the restaurant have takeaway option-Yes/ No
isIndoorSitting	Does the restaurant have indoor sitting option-Yes/No
isVegOnly	Is it a vegetarian restaurant – Yes/No
Dinner Ratings	Customer rating on food quality (Likert scale 5point)
Dinner Reviews	Total no of reviews given by customers after dining in a particular restaurant
Delivery Ratings	Customer rating on food delivery service (Likert scale 5point)
Delivery Reviews	Total no of reviews given by customers after taking food delivery services from a specific restaurant
KnownFor	Food items for which the restaurant is famous among customers
Popular_dishes	The most popular dish of the restaurant
PeopleKnownFor	Services for which a restaurant is known for
AverageCost	Average food cost for 2 people

Name	URL	Cuisines	Area	Timing	Full_Addr	PhoneNumber	IsHomeDe	isTakeawa	isIndoorSe	isVegOnly	Dinner Ra	Dinner Re	Delivery R	Delivery R	KnownFor	PopularDi	PeopleKn	AverageCos
Sri Udupi	https://w	South Ind	Indiranag	7am â€“ 1	273, Mon	919946000000	1	1	1	1	4	462	4.1	16000		Filtered C	Economic	450
Meghana	https://w	Biryani, A	Indiranag	Opens at	544, First	918041000000	1	1	1	0	4.3	1654	4.3	28600	Spicy Chic	Authentic	Boneless	700
Donne Bir	https://w	Biryani	Indiranag	11am â€“	8/ 9, 17th	918862000000	1	1	1	0	3.9	411	3.5	33200			Great Rec	300
Domino's	https://w	Pizza, Fast	Indiranag	10:57am	308, 2nd	919916000000	1	1	1	0	2.4	422	4.4	8205		Barbeque	Value for	400
KFC	https://w	Burger, Fa	Indiranag	11am â€“	38/1A, CM	919514000000	1	1	1	0	2.8	673	4	9148		Fiery Chic	Elegantly	400
The Cong	https://w	South Ind	Ulsoor, Bangalore		26, Vemea	918147000000	1	0	0	0	-	0	4.5	239			Fresh Foo	100
Burger Kir	https://w	Burger, Fa	Indiranag	9am â€“ 2	545, Next	918043000000	1	1	1	0	4.1	1499	4.1	22400	High-qual	Crispy Spi	Bang for t	400
McDonalc	https://w	Burger, Be	Indiranag	9:30am â€	539, Grou	918928000000	1	1	1	0	3.8	500	4.1	7508	Their reas	Mcspicy C	Casual Sea	400
Mani's Du	https://w	Biryani, K	Jeevan Bh	11am â€“	474, Grou	918060000000	1	1	1	0	4.2	708	4.1	12700	It's flavou	Paneer Fr	Serving Si	800

Data Preparation:

- The data preparation consists of converting categorical data into quantitative data.
- 7 out of 19 columns are in text and are not useful for the analysis. 4 of these 7 columns (KnownFor,PopularDishes, PeopleKnownFor, Timing) had more than 3000 rows of missing values and was removed from analysis.
- Found 5324 cells in dinner ratings and 1131 cells in delivery ratings having (-) dash as values which were replaced by the mean ratings.
- There was total 168 distinctive locations for which we created a new variable called area index.
- We converted no of cuisines into numerical data by calculating and aggregating how many cuisines each restaurant is serving.
- We have converted average cost into binary groups: if average cost is less than 1000 then it falls to 0 group, if average cost is greater than 1000 then it falls to group 1.

- We have removed URL, timing, full address, telephone, known for, popular dishes and people known for columns from the data set as those were having the greatest number of missing values.

We have identified below columns which we are planning to consider for further analysis.

Variable Name	**Description**
Name	Restaurant name
Cuisines	Cuisine name for the restaurant
Areas	Area name where restaurant is situated
isHomeDelivery	Does the restaurant have home delivery option-Yes/No
isTakeaway	Does the restaurant have takeaway option-Yes/ No
isIndoorSitting	Does the restaurant have indoor sitting option-Yes/No
isVegOnly	Is it a vegeterinian restaurant – Yes/No
Dinner Ratings	Customer rating on food quality (Likert scale 5point)
Dinner Reviews	Total no of reviews given by customers after dining in a particular restaurant
Delivery Ratings	Customer rating on food delivery service (Likert scale 5point)
Delivery Reviews	Total no of reviews given by customers after taking food delivery services from a specific restaurant
Average Cost	Average food cost for 2 people

Data Visualization:

In order to perform exploratory data analysis, we have done some visualization on the dataset using **Tableau** to analyze and understand the relationship between different variables of the dataset.

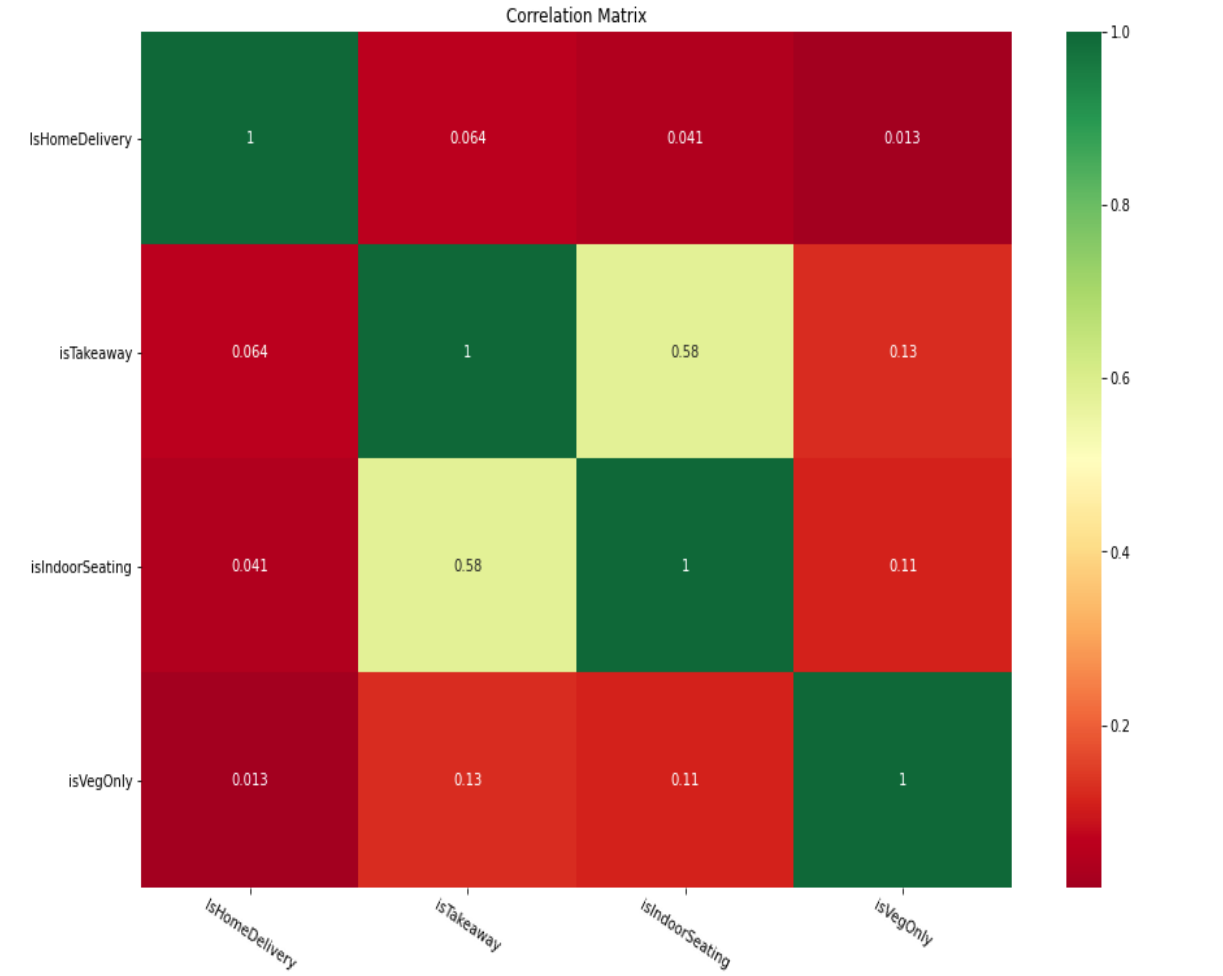

Figure-1: - Correlation matrix between isHomeDelivery, isTakeaway, isIndoorSeating, isVegOnly

In figure 1, we can see the correlation between isHomeDelivery, isTakeaway, isIndoorSeating, isVegOnly where 1 means very strong relationship and .2 indicating weak relationship between variables. After thorough analysis of the matrix, we can depict that there is some moderate relationship (0.58) between isIndoorSeating and isTakeaway only and for rest of the variables there is no such relationship.

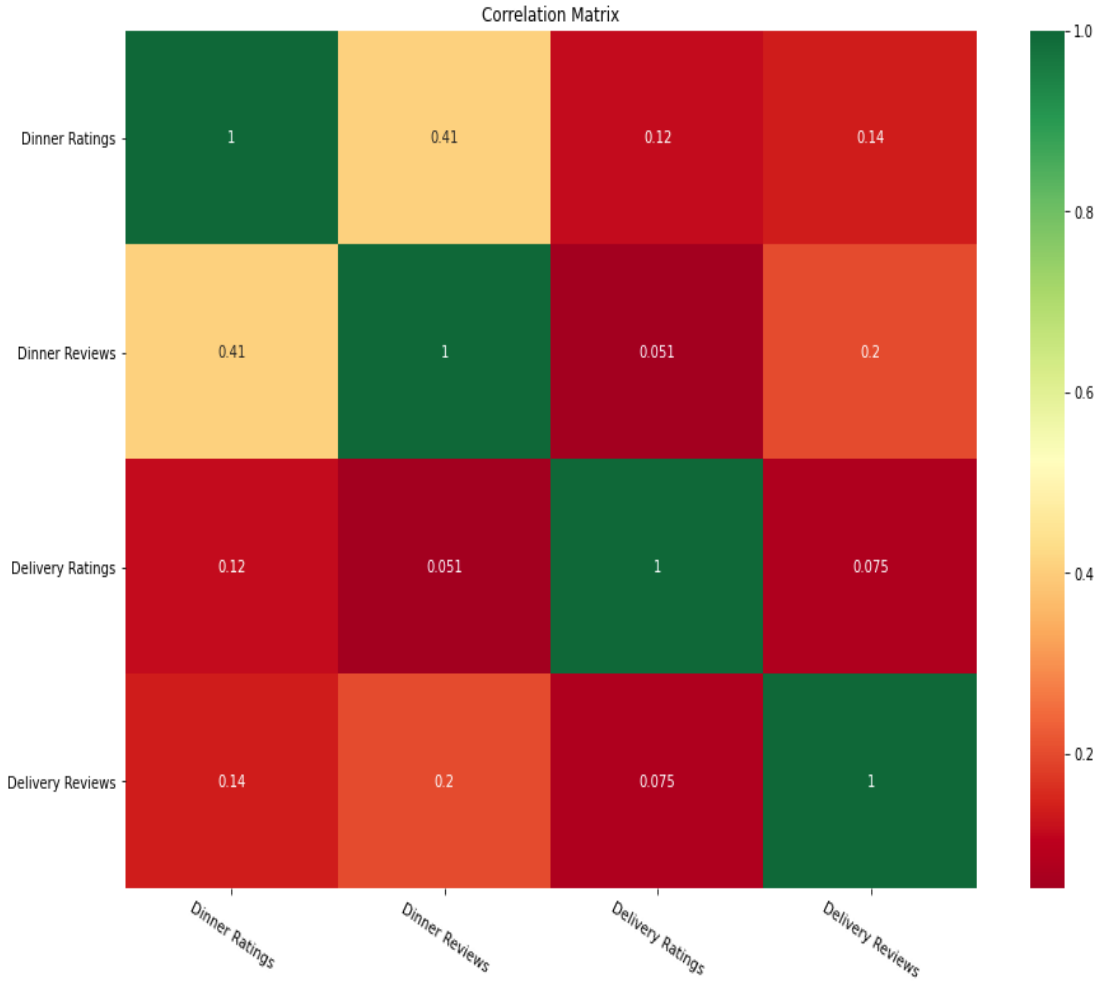

Figure-2: - Correlation matrix between Dinner Ratings, Dinner Reviews, Delivery Ratings and Delivery Reviews.

In figure 2, we have run correlation matrix between Dinner Ratings, Dinner Reviews, Delivery Ratings and Delivery Reviews. From the heatmap, we can summarize that only dinner ratings and dinner reviews have some correlation.

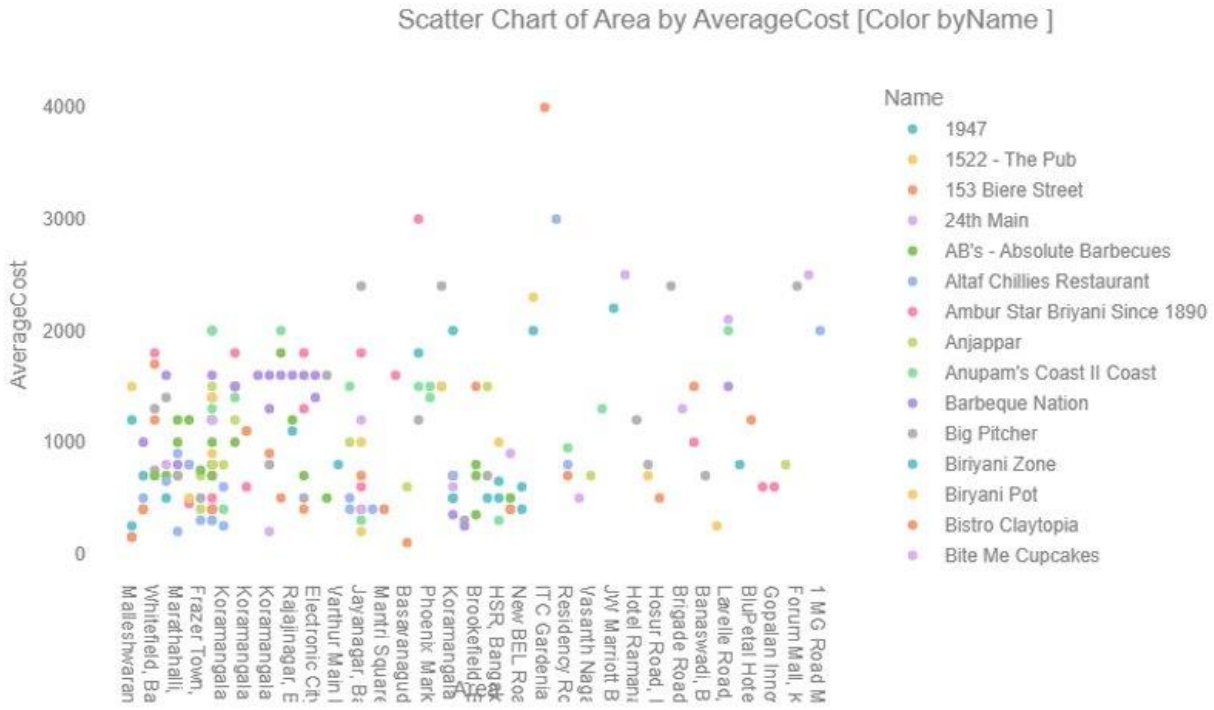

Figure-3: - Scatter plot of Area by Average Cost

We have used the variables Average Cost and Area to draw the scatter plot and observed the likelihood of getting a smaller number of restaurants with increase in cost. So, the probability of getting good business with cost higher that 2000 as shown in the scatter plot will have risks associated. According to the plot, restaurants with cost less than 2000 rupees are more likely to have customers which can be considered while planning out for adding a new restaurant in the app.

Figure-4: - Bubble Chart by Delivery Ratings by Dinner Ratings

This bubble map shows the correlation between Dinner and Delivery Ratings, and which is more likely to affect the likelihood of having good business. We can observe that dinner and delivery ratings are almost linearly increasing. So, when it comes to average cost, these two variables will be given more importance to determine which premium quote the customer gets.

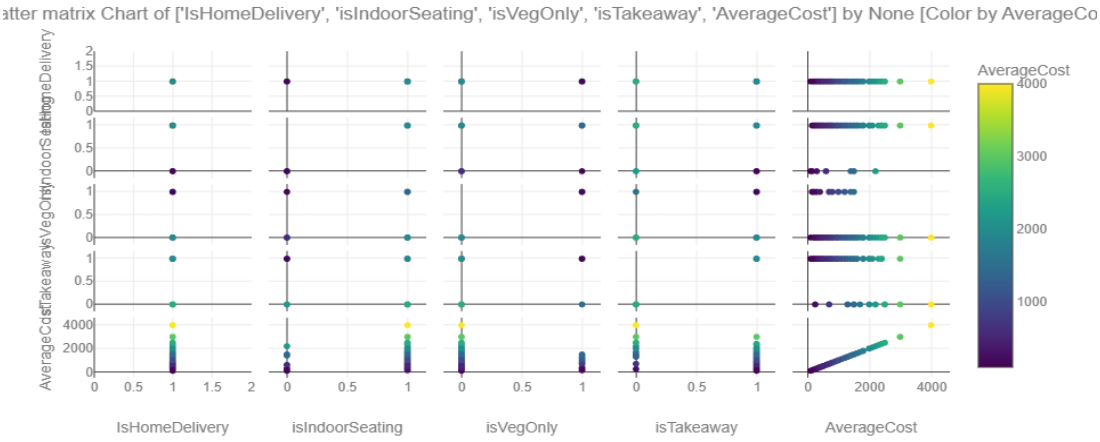

Figure- 5: - Scatter matrix chart of isHomeDelivery, isTakeaway, isIndoorSeating, isVegOnly, AverageCost

In figure 5, we have drawn a scatter matrix between isHomeDelivery, isTakeaway, isIndoorSeating, isVegOnly with our target variable AverageCost to have an in-depth insight.

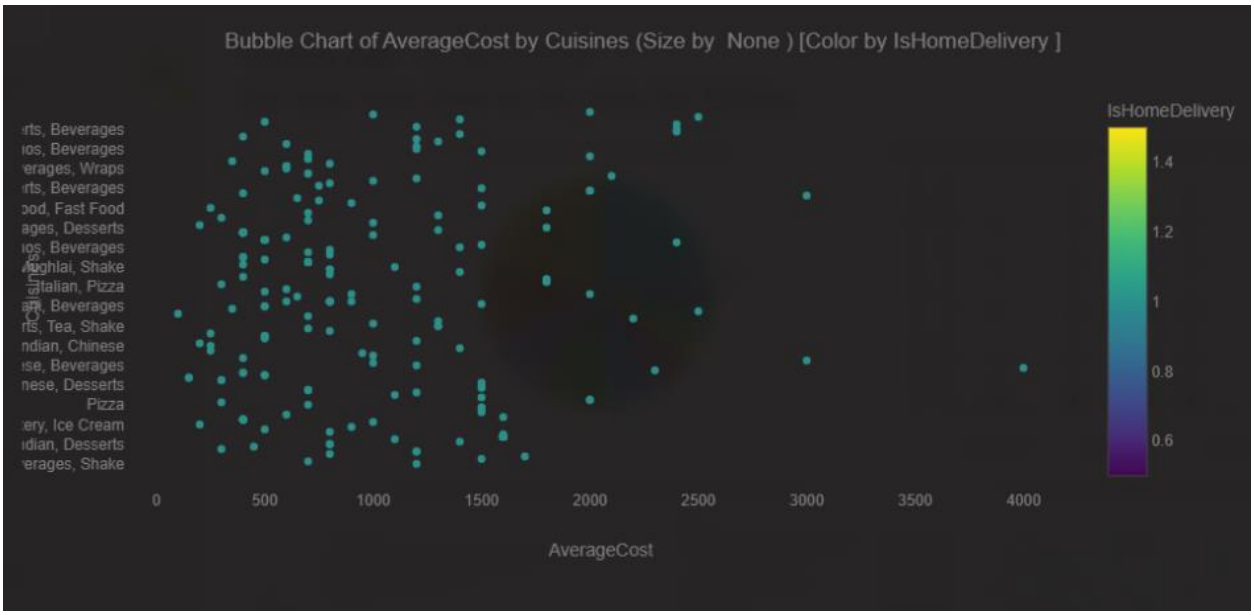

Figure -6: - Bubble Chart of AverageCost by Cuisines

In figure 6, we have drawn a bubble chart between our target variable average cost and different cuisines. After much evaluation of the above chart, we can conclude that majority of the cuisines have an average cost between 0 to 1500.

Prediction Techniques and Findings:

We have used SAS Enterprise Miner for predictive data analysis. In terms of prediction technique, we have used decision tress and multiple linear regression based on which we will predict the best model for Zomato.

Multiple Linear Regression: to determine which variables are significant and to find the ultimate model.

We ran multiple regressions to identify the best fit of model and to generate an equation in predicting average cost based on user's selection of choices.

R-Square: 0.6623

Adj R-Square: 0.6617

DeliveryRatings, DinnerRatings, Dinner Reviews, NoOfCuisines, IsIndoorSeating, IsVegOnly, IsHomeDelivery variables are significant in predicting average cost based on the training data set.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	347570298	34757030	1047.96	<.0001
Error	5343	177207416	33166		
Corrected Total	5353	524777715			

Figure- 7: - Analysis of Variance table from SAS

Model Fit Statistics

R-Square	0.6623	Adj R-Sq	0.6617
AIC	55742.3213	BIC	55744.3666
SBC	55814.7629	C(p)	11.0000

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
AreaIndex	1	24605.5780	0.74	0.3891
DeliveryRatings	1	2523678.97	76.09	<.0001
DeliveryReviews	1	187527.235	5.65	0.0174
DinnerRatings	1	8338080.66	251.40	<.0001
DinnerReviews	1	77777043.8	2345.06	<.0001
IsHomeDelivery	1	5903024.62	177.98	<.0001
NameIndex	1	28766.3088	0.87	0.3517
NoOfCuisines	1	12292950.5	370.65	<.0001
isIndoorSeating	1	12109761.5	365.12	<.0001
isVegOnly	1	913725.883	27.55	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	-442.9	34.4030	-12.88	<.0001
AreaIndex		1	0.0543	0.0630	0.86	0.3891
DeliveryRatings		1	86.5949	9.9271	8.72	<.0001
DeliveryReviews		1	0.00128	0.000537	2.38	0.0174
DinnerRatings		1	135.3	8.5339	15.86	<.0001
DinnerReviews		1	0.0602	0.00124	48.43	<.0001
IsHomeDelivery	0	1	134.3	10.0660	13.34	<.0001
NameIndex		1	0.00157	0.00169	0.93	0.3517
NoOfCuisines		1	36.0821	1.8742	19.25	<.0001
isIndoorSeating	0	1	-71.3519	3.7341	-19.11	<.0001
isVegOnly	0	1	25.5455	4.8669	5.25	<.0001

Figure- 8: - Findings of Regression model from SAS Enterprise Miner

Decision Trees: to evaluate the options available for user selection and find the accuracy of the model.

While running Decision tree model, we have considered all the variables except name, price range, isTakeaway, AverageCost.

If dinner reviews are < 2217 and dinner ratings is <4.05 then 99.47% of data falls into the 0 group where average cost is less than 1000. The rule of decision tree is self-explanatory as people are more prone to look for restaurant options that provides a good meal within an average cost of 1000 rupees on Zomato app.

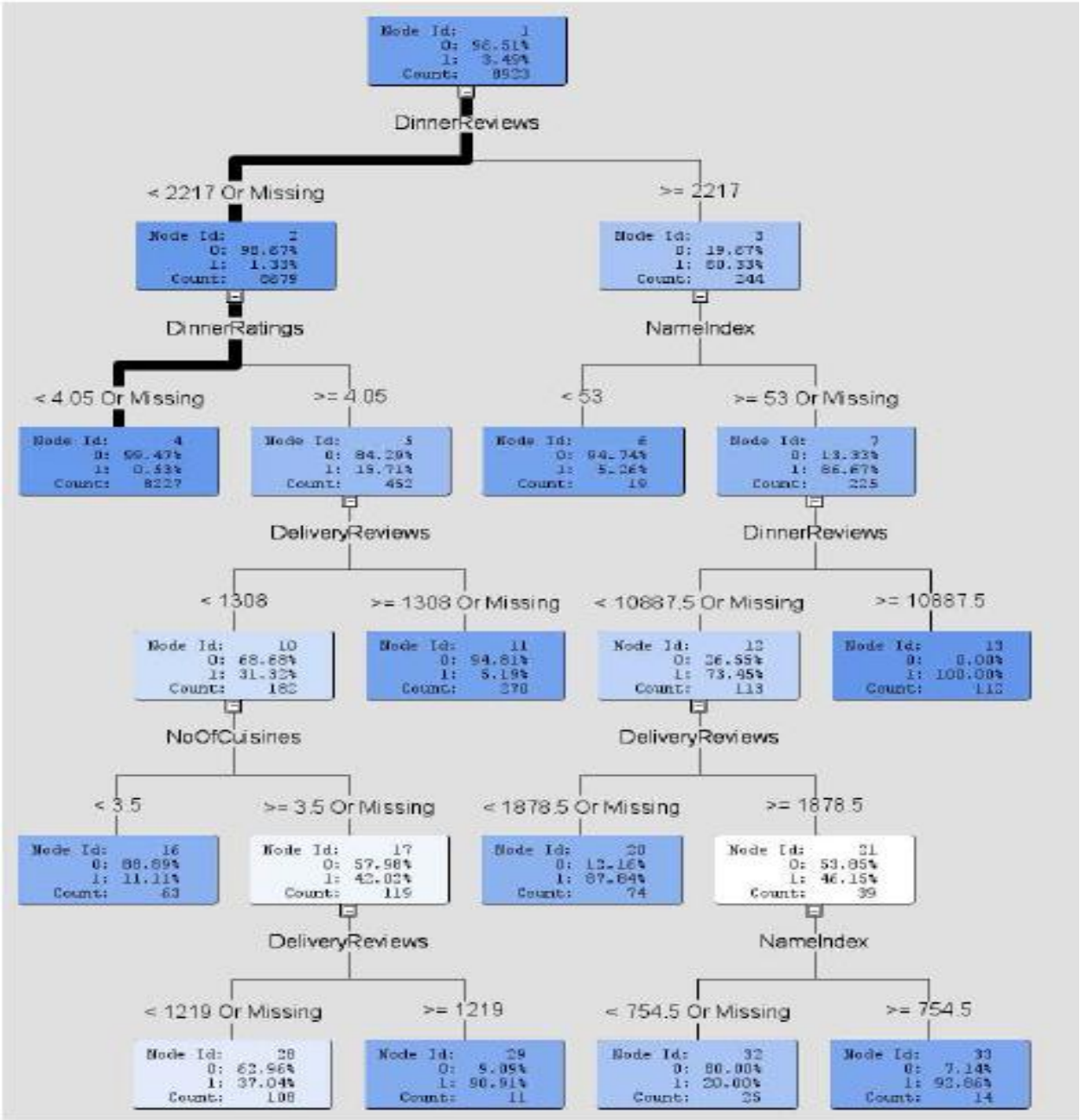

Figure- 9: - Final Decision Tree using SAS Enterprise Model

Fit Statistics	Statistics Label	Train
NOBS	Sum of Frequencies	8923.00
MISC	Misclassification Rate	0.01
MAX	Maximum Absolute Error	0.99
SSE	Sum of Squared Errors	206.27
ASE	Average Squared Error	0.01
RASE	Root Average Squared Error	0.11
DIV	Divisor for ASE	17846.00
DFT	Total Degrees of Freedom	8923.00

Figure- 10: - Error table for training dataset

Model Comparison: to determine the best model for recommendation.

Through model comparison, we can conclude that multiple linear regression is the best model so far.

Fit Statistics

Model Selection based on Valid: Average Squared Error (_VASE_)

Selected Model	Model Node	Model Description	Valid: Average Squared Error	Train: Average Squared Error
Y	Reg	Regression	31614.65	33098.14

Figure- 11: - Results of Model Comparison Node using SAS Enterprise Miner

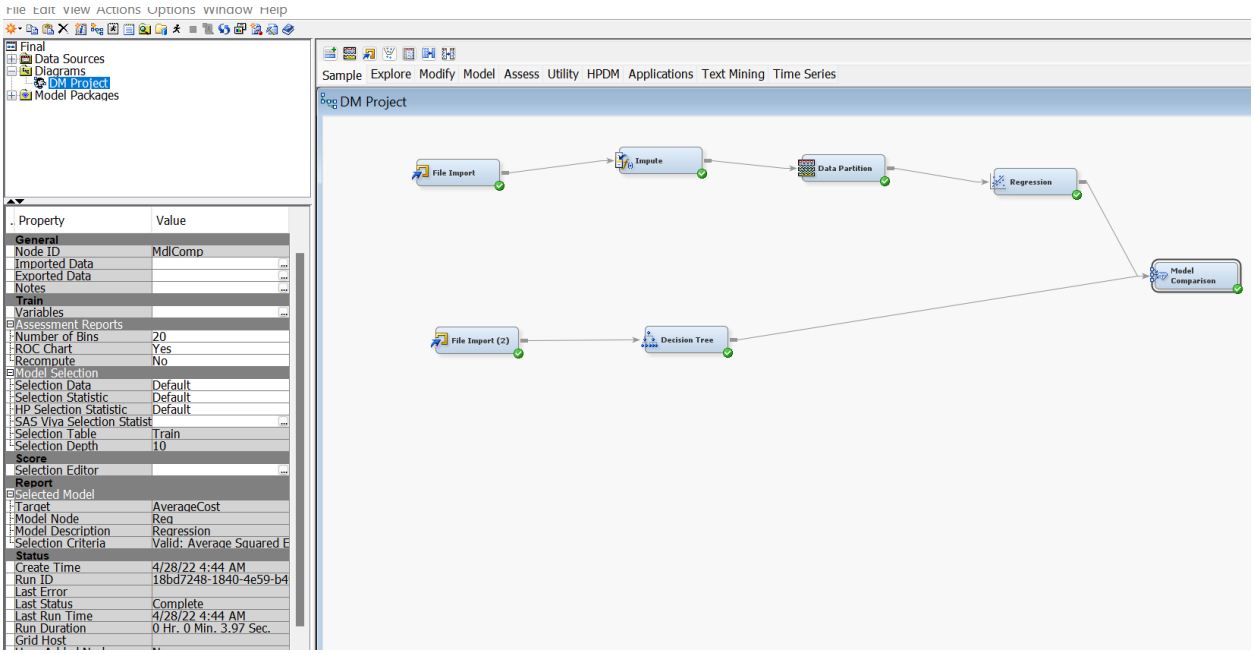

Figure – 12: - Final Node from SAS Enterprise Miner

Conclusion:

For any on demand food delivery company, getting desired level of user base is one of the main targets towards the end goal of earning revenue. By applying above mentioned visualization and prediction technique we can estimate the significant variables which brings new user base to Zomato app.

Model Output:

Equation to attract new user base by predicting average cost:

- Average cost = 86.5929* DeliveryRatings + 135.3*DinnerRatings + 0.0602*DinnerReviews + 36.0821*NoOfCuisines – 71.3519*IsIndoorSeating + 25.5455*IsVegOnly +134.3*IsHomeDelivery – 442.9

Reference:

1. **https://en.wikipedia.org/wiki/Zomato**
2. **https://startuptalky.com/business-revenue-model-zomato/**
3. **https://jungleworks.com/zomato-business-model-explained/**
4. **https://oyelabs.com/zomatobusinessmodel/#:~:text=the%20platform's%20retailers.,Food%20Delivery,the%20supplier%20and%20the%20business.**

INSY 5339: PRINCIPLES OF BUSINESS DATA MINING

Academic Integrity

In order for your Assignment/Homework/Project to be accepted you must read the following, sign this form and attach it to your papers (as the last page of your assignment).
Academic Integrity: Students enrolled in this course are expected to adhere to the UT Arlington Honor Code:

I pledge, on my honor, to uphold UT Arlington's tradition of academic integrity, a tradition that values hard work and honest effort in the pursuit of academic excellence.
I promise that I will submit only work that I personally create or contribute to group collaborations, and I will appropriately reference any work from other sources. I will follow the highest standards of integrity and uphold the spirit of the Honor Code.

Student Signature: Faria Tasnim, Vishali Bairam, Chiriki Sai Goutham, Aditi Tibrewal
Student Id Number: 1001967007, 1002036395, 1002018678, 1001955627