Something about Zomato

MATH 1298 Analysis of Categorical Data Project Phase I

Arion Barzoucas-Evans (s3650046) & Joshua Grosman (insert student number)

08/08/2018

Contents

1 Introduction

Zomato is a restaurant search and discovery service founded in 2008. Users can access a plethora of information about restuarants listed on Zomato, including information not available on the restaurant's own website. Such information includes the type of cuisine, opening hours, photos of the menu and the restaurant, pricing, and whetheror not the restaurant offers online delivery. Customers that visit the restaurants have the option of reviewing them and giving them a rating from 0 to 5. Higher rated restaurants receive more attention resulting in higher revenues. As such, understanding how users of Zomato rate restaurants is of great interest to restaurant owners in order to improve their business. The aim of this report is to explore the relationship of several factors like price and cuisine with Zomato ratings using publicly available data found on Kaggle. In phase I exploratory data analysis will be performed on the dataset, including data pre-processing, creation of new variables, and visual representation of the data. This will assist in observing and understanding any relationships present in the data. Following this, a logistic regression model will be fitted in phase II in order to predict the probability of receiving certain ratings according to the values of the chosen explanatory variables.

2 Data Set

The dataset used in this report was acquired from Kaggle. It contains 9,551 observations each of which corresponds to a different restaurant and contains the following information:

- *Restaurant.ID*: A unique ID assigned to each restaurant.
- *Restaurant.Name*: Name of the restaurant.
- *Country.Code*: Codes corresponding to countries listed in a separate dataset.
- *City*: Name of the city where the restaurant is located.
- *Address*: Address of the restaurant.
- *Locality*: General location of the restaurant (short description).
- *Locality.Verbose*: General location of the restaurant (long description).
- *Longitude*: Longitude of the location of the restaurant (geographic coordinates).
- *Latitude*: Latitude of the location of the restaurant (geographic coordinates).
- *Cuisines*: Type of cuisine offered by the restaurant.
- *Average.Cost.for.two*: Average cost for two people (in the countries respective currency).
- *Currency*: Currency which is used at each restaurant.
- *Has.Table.booking*: Whether the restaurant offers the option to book a table or not.
- *Has.Online.delivery*: Whether the restaurant offers delivery through the internet or not.
- *Is.delivering.now*: Whether the restaurant was delivering at the time the dataset was created or not.
- *Switch.to.order.menu*: Unclear variable with only one level (No) for all observations.
- *Price.range*: Categorised price (between 1 and 4).
- *Aggregate.rating*: Aggregated rating from user votes.
- *Rating.color*: Categorised rating into a colour code.
- *Rating.text*: Response variable. Categorised rating with 6 levels (between Poor to Excellent).
- *Votes*: Number of votes used in the aggregate rating.

3 Data Preparation

In this project, the following `R` packages were used.

```
library(mlr)
library(data.table)
library(plyr)
library(dplyr)
library(ggplot2)
library(broom)
library(vcd)
library(knitr)
library(kableExtra)
```

As Table 1 indicates, there are no NA values in any of the features. Additionally, according to Table 2 restaurant ID's are unique for every restaurant while there are appears to be some duplicate restaurant names due to the existance of restaurant chains. Furthermore, each restaurant has multiple cuisines which causes the `cuisine` feature to have so many levels. Finally, the `Switch.to.order.menu` feature only has one level ("No") for the entire dataset.

Table 1: Feature summary before data preprocessing.

name	type	na	mean	disp	median	mad	min	max	nlevs
Restaurant.ID	integer	0	9.051128e+06	8.791521e+06	6.004089e+06	8.900212e+06	53.00000	1.850065e+07	0
Restaurant.Name	factor	0	NA	9.913098e-01	NA	NA	1.00000	8.300000e+01	7446
Country.Code	integer	0	1.836562e+01	5.675055e+01	1.000000e+00	0.000000e+00	1.00000	2.160000e+02	0
City	factor	0	NA	4.269710e-01	NA	NA	1.00000	5.473000e+03	141
Address	factor	0	NA	9.988483e-01	NA	NA	1.00000	1.100000e+01	8918
Locality	factor	0	NA	9.872265e-01	NA	NA	1.00000	1.220000e+02	1208
Locality.Verbose	factor	0	NA	9.872265e-01	NA	NA	1.00000	1.220000e+02	1265
Longitude	numeric	0	6.412657e+01	4.146706e+01	7.719196e+01	1.506428e-01	-157.94849	1.748321e+02	0
Latitude	numeric	0	2.585438e+01	1.100794e+01	2.857047e+01	1.135080e-01	-41.33043	5.597698e+01	0
Cuisines	factor	0	NA	9.019998e-01	NA	NA	1.00000	9.360000e+02	1826
Average.Cost.for.two	integer	0	1.199211e+03	1.612118e+04	4.000000e+02	2.965200e+02	0.00000	8.000000e+05	0
Currency	factor	0	NA	9.412630e-02	NA	NA	20.00000	8.652000e+03	12
Has.Table.booking	factor	0	NA	1.212438e-01	NA	NA	1158.00000	8.393000e+03	2
Has.Online.delivery	factor	0	NA	2.566223e-01	NA	NA	2451.00000	7.100000e+03	2
Is.delivering.now	factor	0	NA	3.559800e-03	NA	NA	34.00000	9.517000e+03	2
Switch.to.order.menu	factor	0	NA	0.000000e+00	NA	NA	9551.00000	9.551000e+03	1
Price.range	integer	0	1.804837e+00	9.056088e-01	2.000000e+00	1.482600e+00	1.00000	4.000000e+00	0
Aggregate.rating	numeric	0	2.666370e+00	1.516377e+00	3.200000e+00	7.413000e-01	0.00000	4.900000e+00	0
Rating.color	factor	0	NA	6.087321e-01	NA	NA	186.00000	3.737000e+03	6
Rating.text	factor	0	NA	6.087321e-01	NA	NA	186.00000	3.737000e+03	6
Votes	integer	0	1.569097e+02	4.301691e+02	3.100000e+01	4.447800e+01	0.00000	1.093400e+04	0

Table 2: Variable summary for the Zomato dataset 9551 Observations of 21 Variables

Variable	Class	Cardinality	First Levels	First Values
Restaurant.ID	integer	9551	6317637, 6304287, 6300002, 6318506, 6314302, 18189371	6317637, 6304287, 6300002, 6318506, 6314302, 18189371
Restaurant.Name	character	6899	Le Petit S, Izakaya Ki, Heat - Eds, Ooma, Sambo Koji, Din Tai Fu	Le Petit S, Izakaya Ki, Heat - Eds, Ooma, Sambo Koji, Din Tai Fu
Country.Code	integer	15	162, 30, 216, 14, 37, 184	162, 162, 162, 162, 162, 162
City	factor	141	Abu Dhabi, Agra, Ahmedabad, Albany, Allahabad, Amritsar	Makati City, Makati City, Mandaluyong City, Mandaluyong City, Mandaluyong City, Mandaluyong City
Address	character	7626	Third Floo, Little Tok, Edsa Shang, Ground Flo, Building K, Building B	Third Floo, Little Tok, Edsa Shang, Third Floo, Third Floo, Ground Flo
Locality	character	1140	Century Ci, Little Tok, Edsa Shang, SM Megamal, SM by the , Sofitel Ph	Century Ci, Little Tok, Edsa Shang, SM Megamal, SM Megamal, SM Megamal
Locality.Verbose	character	1141	Century Ci, Little Tok, Edsa Shang, SM Megamal, SM by the , Sofitel Ph	Century Ci, Little Tok, Edsa Shang, SM Megamal, SM Megamal, SM Megamal
Longitude	numeric	8120	121.027535, 121.014101, 121.056831, 121.056475, 121.057508, 121.056314	121.027535, 121.014101, 121.056831, 121.056475, 121.057508, 121.056314
Latitude	numeric	8677	14.565443, 14.553708, 14.581404, 14.585318, 14.58445, 14.583764	14.565443, 14.553708, 14.581404, 14.585318, 14.58445, 14.583764
Cuisines	character	417	French, Ja, Japanese, Seafood, A, Japanese, , Chinese, Asian, Eur	French, Ja, Japanese, Seafood, A, Japanese, , Japanese, , Chinese
Average.Cost.for.two	integer	140	1100, 1200, 4000, 1500, 1000, 2000	1100, 1200, 4000, 1500, 1500, 1000
Currency	factor	12	Botswana Pula(P), Brazilian Real(R$), Dollar($), Emirati Diram(AED), Indian Rupees(Rs.), Indonesian Rupiah(IDR)	Botswana Pula(P), Botswana Pula(P), Botswana Pula(P), Botswana Pula(P), Botswana Pula(P), Botswana Pula(P)
Has.Table.booking	factor	2	No, Yes	Yes, Yes, Yes, No, Yes, No
Has.Online.delivery	factor	2	No, Yes	No, No, No, No, No, No
Is.delivering.now	factor	2	No, Yes	No, No, No, No, No, No
Switch.to.order.menu	factor	1	No	No, No, No, No, No, No
Price.range	integer	4	3, 4, 2, 1	3, 3, 4, 4, 4, 3
Aggregate.rating	numeric	33	4.8, 4.5, 4.4, 4.9, 4, 4.2	4.8, 4.5, 4.4, 4.9, 4.8, 4.4
Rating.color	factor	6	Dark Green, Green, Orange, Red, White, Yellow	Dark Green, Dark Green, Green, Dark Green, Dark Green, Green
Rating.text	factor	6	Average, Excellent, Good, Not rated, Poor, Very Good	Excellent, Excellent, Very Good, Excellent, Excellent, Very Good
Votes	integer	1012	314, 591, 270, 365, 229, 336	314, 591, 270, 365, 229, 336

To rectify the identified issues, all unnecessary features were removed. This includes `Restaurant.ID`

(unique identifier), `Restaurant.Name` ,`Address`, `Locality`, `Locality.Verbose`, `Is.delivering.now`, `Switch.to.order.menu`, `City`, `Currency`, `Rating.color`. The `Average.Cost.for.two` is in many different currencies and the concept of what is considered expensive would be affected by socio-economic factors in each country. For this reason this feature was standardised by currency. Furthermore, the `cuisines` variable was separated into 18 new binary features using the most prevalent levels within the original `cuisines` variable. Each of these features indicates the presence or absence of that particular cuisine in the restaurant. This way, restaurants can have multiple cuisines. Using these new binary variables, a new feature, `Cuisine_Range`, was created as the sum of all the cuisine binary variables. The `Cuisine_Range` would indicate the number of different cuisines present in a restaurant which may be of interest to the model in deciding a restaurant's rating. Finally, `Has.Table.booking` and `Has.Online.delivery` were recoded into binary variables, the country table was joined to the Zomato dataset through `Country.Code` and then aggregated into a `Continent` feature. Tables 3 and 4 show the data after preprocessing.

Table 3: Feature summary after data preprocessing.

name	type	na	mean	disp	median	mad	min	max	nlevs
Longitude	numeric	0	64.127	41.467	77.192	0.151	-157.948	174.832	0
Latitude	numeric	0	25.854	11.008	28.570	0.114	-41.330	55.977	0
Average.Cost.for.two.Std	numeric	0	0.000	0.999	-0.207	0.498	-1.328	12.384	0
Has.Table.booking	factor	0	NA	0.121	NA	NA	1158.000	8393.000	2
Has.Online.delivery	factor	0	NA	0.257	NA	NA	2451.000	7100.000	2
Price.range	integer	0	1.805	0.906	2.000	1.483	1.000	4.000	0
Aggregate.rating	numeric	0	2.666	1.516	3.200	0.741	0.000	4.900	0
Rating.text	factor	0	NA	0.609	NA	NA	186.000	3737.000	6
Votes	integer	0	156.910	430.169	31.000	44.478	0.000	10934.000	0
Seafood	numeric	0	0.018	0.134	0.000	0.000	0.000	1.000	0
Asian	numeric	0	0.318	0.466	0.000	0.000	0.000	1.000	0
European	numeric	0	0.104	0.305	0.000	0.000	0.000	1.000	0
Cafe	numeric	0	0.074	0.262	0.000	0.000	0.000	1.000	0
Fast Food	numeric	0	0.208	0.406	0.000	0.000	0.000	1.000	0
Bakery	numeric	0	0.078	0.268	0.000	0.000	0.000	1.000	0
Pizza	numeric	0	0.041	0.198	0.000	0.000	0.000	1.000	0
Desserts	numeric	0	0.078	0.269	0.000	0.000	0.000	1.000	0
Beverages	numeric	0	0.033	0.179	0.000	0.000	0.000	1.000	0
Burger	numeric	0	0.026	0.160	0.000	0.000	0.000	1.000	0
Indian	numeric	0	0.504	0.500	1.000	0.000	0.000	1.000	0
Finger Food	numeric	0	0.012	0.109	0.000	0.000	0.000	1.000	0
Continental	numeric	0	0.077	0.267	0.000	0.000	0.000	1.000	0
Street Food	numeric	0	0.059	0.235	0.000	0.000	0.000	1.000	0
Raw Meats	numeric	0	0.012	0.109	0.000	0.000	0.000	1.000	0
South American	numeric	0	0.024	0.152	0.000	0.000	0.000	1.000	0
Healthy Food	numeric	0	0.016	0.124	0.000	0.000	0.000	1.000	0
Other	numeric	0	0.058	0.234	0.000	0.000	0.000	1.000	0
Oceania	integer	0	0.007	0.082	0.000	0.000	0.000	1.000	0
Rest of World	integer	0	0.016	0.126	0.000	0.000	0.000	1.000	0
North America	integer	0	0.046	0.209	0.000	0.000	0.000	1.000	0
Asia	integer	0	0.923	0.267	1.000	0.000	0.000	1.000	0
Europe	integer	0	0.008	0.091	0.000	0.000	0.000	1.000	0
Cuisine_Range	numeric	0	1.740	0.900	2.000	1.483	0.000	8.000	0

Table 4: Variable summary for the Zomato dataset after preprocessing 9551 Observations of 21 Variables

Variable	Class	Cardinality	First Levels	First Values
Longitude	numeric	1624	78.012, 0, 77.998, 78.008, 78.044, 78.057	78.012, 0, 78.012, 77.998, 78.008, 0
Latitude	numeric	1490	27.162, 0, 27.161, 27.196, 27.202, 27.163	27.162, 0, 27.161, 27.196, 27.202, 0
Average.Cost.for.two.Std	numeric	279	0.38, 0.129, -0.207, -0.375, 0.632, 2.311	0.38, 0.129, -0.207, -0.375, 0.632, 2.311
Has.Table.booking	factor	2	0, 1	0, 0, 0, 0, 0, 0
Has.Online.delivery	factor	2	0, 1	0, 0, 0, 0, 0, 0
Price.range	integer	4	3, 2, 4, 1	3, 2, 2, 2, 3, 4
Aggregate.rating	numeric	33	3.9, 3.5, 3.6, 4, 4.2, 4.3	3.9, 3.5, 3.6, 4, 4.2, 4
Rating.text	factor	6	Average, Excellent, Good, Not rated, Poor, Very Good	Good, Good, Good, Very Good, Very Good, Very Good
Votes	integer	1012	140, 71, 94, 87, 177, 45	140, 71, 94, 87, 177, 45
Seafood	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Asian	numeric	2	0, 1	0, 0, 0, 0, 1, 0
European	numeric	2	0, 1	0, 0, 0, 0, 0, 1
Cafe	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Fast Food	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Bakery	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Pizza	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Desserts	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Beverages	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Burger	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Indian	numeric	2	1, 0	1, 1, 1, 1, 1, 1
Finger Food	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Continental	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Street Food	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Raw Meats	numeric	2	0, 1	0, 0, 0, 0, 0, 0
South American	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Healthy Food	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Other	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Oceania	integer	2	0, 1	0, 0, 0, 0, 0, 0
Rest of World	integer	2	0, 1	0, 0, 0, 0, 0, 0
North America	integer	2	0, 1	0, 0, 0, 0, 0, 0
Asia	integer	2	1, 0	1, 1, 1, 1, 1, 1
Europe	integer	2	0, 1	0, 0, 0, 0, 0, 0
Cuisine_Range	numeric	9	1, 2, 3, 5, 4, 6	1, 1, 1, 1, 2, 2

Table 5: Contingency tables for discrete and categorical variables.

Has.Table.booking	Freq
0	8393
1	1158

Has.Online.delivery	Freq
0	7100
1	2451

Price.range	Freq
1	4444
2	3113
3	1408
4	586

Rating.text	Freq
Average	3737
Excellent	301
Good	2100
Not rated	2148
Poor	186
Very Good	1079

Seafood	Freq
0	9377
1	174

Asian	Freq
0	6517
1	3034

European	Freq
0	8557
1	994

Cafe	Freq
0	8844
1	707

Fast.Food	Freq
0	7564
1	1987

Bakery	Freq
0	8807
1	744

Pizza	Freq
0	9162
1	389

Desserts	Freq
0	8802
1	749

Beverages	Freq
0	9234
1	317

Burger	Freq
0	9300
1	251

Indian	Freq
0	4738
1	4813

Finger.Food	Freq
0	9437
1	114

Continental	Freq
0	8815
1	736

Street.Food	Freq
0	8989
1	562

Raw.Meats	Freq
0	9437
1	114

South.American	Freq
0	9324
1	227

Healthy.Food	Freq
0	9401
1	150

Other	Freq
0	8994
1	557

Oceania	Freq
0	9487
1	64

Rest.of.World	Freq
0	9397
1	154

North.America	Freq
0	9113
1	438

Asia	Freq
0	736
1	8815

Europe	Freq
0	9471
1	80

Cuisine_Range	Freq
0	80
1	4483
2	3356
3	1239
4	295
5	68
6	25
7	3
8	2

4 Data Visualisation

5 Summary

References

Bischl, Bernd, Michel Lang, Lars Kotthoff, Julia Schiffner, Jakob Richter, Erich Studerus, Giuseppe Casalicchio, and Zachary M. Jones. 2016. "mlr: Machine Learning in R." *Journal of Machine Learning Research* 17 (170): 1–5. http://jmlr.org/papers/v17/15-066.html.

David Meyer, Achim Zeileis, and Kurt Hornik. 2017. *Vcd: Visualizing Categorical Data*.

Dowle, Matt, and Arun Srinivasan. 2018. *Data.table: Extension of 'Data.frame'*. https://CRAN.R-project.org/package=data.table.

Robinson, David, and Alex Hayes. 2018. *Broom: Convert Statistical Analysis Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Wickham, Hadley. 2011. "The Split-Apply-Combine Strategy for Data Analysis." *Journal of Statistical Software* 40 (1): 1–29. http://www.jstatsoft.org/v40/i01/.

———. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org.

Wickham, Hadley, Romain Francois, Lionel Henry, and Kirill Muller. 2017. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2018. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://CRAN.R-project.org/package=dplyr.

Zhu, Hao. 2018. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.