Predicting Zomato restaurant ratings with logistic regression

MATH 1298 Analysis of Categorical Data Project Phase II

Arion Barzoucas-Evans (s3650046) & Joshua Grosman (s3494389)

13/10/2018

Contents

1 Introduction

During phase I of this project, exploratory data analysis was performed on data pertaining to 7,403 restaurants listed on Zomato (a restaurant search and discovery service founded in 2008). This included data pre-processing, creation of new variables, and visual representation of the data. According to this, the number of votes and whether a restaurant offers table bookings or not were found to be variables of particular interest. This phase of the project will focus on fitting a logistic regression model to the Zomato data and examining the effects of the different variables included in the model on a restaurant's rating.

2 Methodology

Following some final minor data preparation, a multinomial logistic regression model will be fitted to the Zomato data. Firstly, the data will be split into training and test sets (80:20) to enable unbiased accuracy assessments. The response variable for the model will be `Rating.text` having the following levels: "Poor", "Average", "Good", "Very Good", and "Excellent". There are a total of 7 explanatory independent variables in the dataset out of which only the most significant variables will be included in the model. This will be accomplished by performing an exhaustive search of all possible regression models and selecting the model with the lowest BIC value. Due to computational constraints, feature selection will only consider one-way interactions between predictor variables.

The chosen predictors will then be included into a multinomial logistic regression model. Furthermore, as the response variable is ordinal, a proportional-odds cumulative logistic model will also be constructed. The effect of the chosen predictors and overall model performance will then be examined, and the best model will be identified based primarily on parameter significance, and the residual deviance of the models. The relationship between the predictor variables and the probability of each response level will be further explored via examination of odds probabilities and visual representations. Finally, the test data will be fed into the model, and the predictions will be compared to actual observations.

The following `R` packages will be utilised to accomplice these tasks:

```
library(nnet)
library(car)
library(glmulti)
library(knitr)
library(captioner)
library(stargazer)
library(data.table)
library(dplyr)
library(ggplot2)
library(png)
library(kableExtra)
```

3 Data Preparation

Table 1 displays the first 10 rows of the Zomato data. The levels for the `Rating.text` and `continent` were reordered so that "Poor" and "Rest of World" are the base levels for each variable respectively. This way 4 logistic regression models will be built, each comparing one level of the `Rating.text` variable to its base level ("Poor"). The data was then split into training and test sets via stratified sampling.

Table 1: First 10 rows of the Zomato dataset.

Has.Table.booking	Has.Online.delivery	Rating.text	Votes	Average.Cost.for.two.Std	continent	cuisine	Cuisine_Range
No	No	Good	140	0.3804579	Asia	Indian	1
No	No	Good	71	0.1286432	Asia	Indian	1
No	No	Good	94	-0.2071097	Asia	Indian	1
No	No	Very Good	87	-0.3749862	Asia	Indian	1
No	No	Very Good	177	0.6322726	Asia	Indian	2
No	No	Very Good	45	1.0000000	Asia	European	2
No	No	Very Good	133	1.0000000	Asia	Indian	1
No	No	Very Good	41	1.0000000	Asia	Indian	1
No	No	Good	59	0.2965197	Asia	Indian	1
No	No	Good	46	1.0000000	Asia	Indian	1

```
#Reordering factor variables to specify bases

#rating text (base=poor)
zomato$Rating.text<-factor(zomato$Rating.text,
                           levels = c("Poor", "Average", "Good","Very Good", "Excellent"))

#continent (base=rest of world)
zomato$continent<-factor(zomato$continent,
                           levels = c("Rest of World", "Europe",
                                      "Asia","North America", "Oceania"))

zomato$ID = seq(1,nrow(zomato),1) # for splitting dataset

#test/train

set.seed(57364)
zomato.train = zomato %>% group_by(Rating.text) %>% sample_frac(0.8) # stratified sampling
zomato.test = zomato[!(zomato$ID %in% zomato.train$ID),]

zomato.train = as.data.table(zomato.train)
zomato = zomato[,-c("ID")]
zomato.train = zomato.train[,-c("ID")]
zomato.test = zomato.test[,-c("ID")]
```

4 Feature Selection

Next, the most important features in the dataset will be selected to be used in the multinomial regression model. To do this, an exhaustive search was performed where all possible logistic regression models (excluding 2-way interactions) were fitted and their BIC values were recorded. Table 2 displays the top 5 models fitted in terms of BIC while Figure 1 shows all the fitted models and their BIC values. From this, it is clear that the best model to utilise should only include the `Has.Online.delivery` and `Votes` predictors. The small number of predictors is not surprising given that BIC favours smaller models.

Table 2: Top 5 multinomial regression models according to BIC.

model	bic	weights
Rating.text ~ 1 + Has.Online.delivery + Votes	1338.803	0.969
Rating.text ~ 1 + Has.Online.delivery + Votes + Average.Cost.for.two.Std	1347.048	0.016
Rating.text ~ 1 + Has.Table.booking + Has.Online.delivery + Votes	1347.490	0.013
Rating.text ~ 1 + Has.Online.delivery	1350.940	0.002
Rating.text ~ 1 + Has.Table.booking + Has.Online.delivery + Votes + Average.Cost.for.two.Std	1355.514	0.000

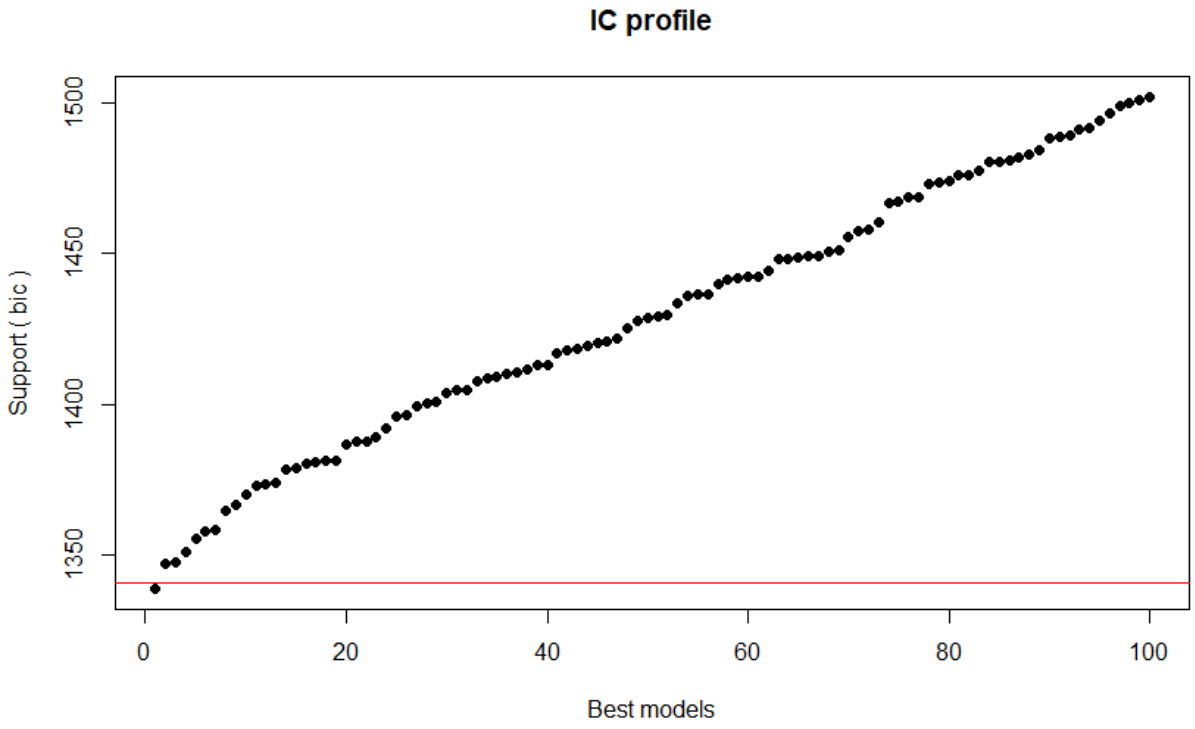

Figure 1: Fitted multinomial regression models and their BIC values.

5 Model Fitting

5.1 Nominal Model

A multinomial logistic regression model was fitting using the variables identified in the previous section.

```
fit.bic.nom<-multinom(formula = Rating.text ~ Has.Online.delivery  + Votes,
                      data=zomato.train)
```

The significance of the predictors was assessed below using an LRT for independence.

```
Anova(fit.bic.nom)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Rating.text
##                      LR Chisq Df Pr(>Chisq)
## Has.Online.delivery   169.48  4  < 2.2e-16 ***
## Votes                2566.99  4  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The predictors are shown to be highly significant with $p < .001$ in both cases. The coefficients of the 4 regression models are summarised below:

```
## Call:
## multinom(formula = Rating.text ~ Has.Online.delivery + Votes,
##     data = zomato.train)
##
## Coefficients:
##           (Intercept) Has.Online.deliveryYes        Votes
## Average    4.01425476              -1.179286 -0.006920626
## Good       2.26287043              -1.109830  0.005151466
## Very Good  1.34454414              -1.768064  0.006574436
## Excellent  0.03918896              -2.765905  0.006916414
##
## Std. Errors:
##           (Intercept) Has.Online.deliveryYes        Votes
## Average    0.08959406             0.08291565 0.0008868042
## Good       0.09016159             0.08353522 0.0008223652
## Very Good  0.09479803             0.09816355 0.0008258368
## Excellent  0.11023459             0.16840731 0.0008279014
##
## Residual Deviance: 11559.63
## AIC: 11583.63
```

5.2 Ordinal Model

A proportional-odds cumulative logistic model was also constructed using `Votes` and `Has.Online.delivery` as predictors.

```
fit.bic.ord<-polr(formula = Rating.text ~ Has.Online.delivery  + Votes,
                  data=zomato.train,  method = "logistic")
```

Again, an LRT was employed to check the significance of predictors.

```
Anova(fit.bic.ord)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Rating.text
##                      LR Chisq Df Pr(>Chisq)
## Has.Online.delivery    26.07  1  3.295e-07 ***
## Votes                1270.94  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both predictors are shown to be highly significant at $p < .001$. Interestingly, the p-value for `Has.Online.Delivery` is higher in this model than the nominal model assessed previously, however here the value is still very small.

The model summary is depicted below:

```
summary(fit.bic.ord)
```

```
##
## Re-fitting to get Hessian
```

```
## Call:
## polr(formula = Rating.text ~ Has.Online.delivery + Votes, data = zomato.train,
##     method = "logistic")
##
## Coefficients:
##                             Value Std. Error t value
## Has.Online.deliveryYes -0.273659  5.380e-02  -5.086
## Votes                   0.002654  9.176e-05  28.918
##
## Intercepts:
##                      Value    Std. Error t value
## Poor|Average         -3.4968   0.0853   -40.9998
## Average|Good          0.4493   0.0343    13.1162
## Good|Very Good        2.0849   0.0461    45.2454
## Very Good|Excellent   4.2296   0.0889    47.5650
##
## Residual Deviance: 13007.16
## AIC: 13019.16
```

5.3 Model Comparison

Comparison of the nominal and ordinal models specified above suggested that the nominal model was superior. This conclusion is based mostly on the differences in the residual deviance of the models. The nominal model had a lower deviance of 11,559.63 , while the ordinal model's was 13,007.16. Furthermore, it was seen that the `Has.Online.Delivery` predictor was slightly more significant within the nominal model compared to the ordinal one. As such, the nominal model was selected as the best model.

5.4 Final Model

The four logistic regression equations associated with the multinomial model defined previously are stated below:

- $log(\frac{Pr(Average)}{Pr(Poor)}) = 4.014 - 1.179 \times Has.Online.Delivery - 0.007 \times Votes$
- $log(\frac{Pr(Good)}{Pr(Poor)}) = 2.263 - 1.11 \times Has.Online.Delivery + 0.005 \times Votes$
- $log(\frac{Pr(VeryGood)}{Pr(Poor)}) = 1.345 - 1.768 \times Has.Online.Delivery + 0.007 \times Votes$
- $log(\frac{Pr(Excellent)}{Pr(Poor)}) = 0.039 - 2.766 \times Has.Online.Delivery + 0.007 \times Votes$

5.4.1 Parameter Confidence Intervals

The confidence intervals for the model's coefficients are shown below. It can be observed that all intervals are relatively narrow, and no interval contains 0.

```
## , , Average
##
##                               2.5 %       97.5 %
## (Intercept)             3.83865364  4.189855891
## Has.Online.deliveryYes -1.34179722 -1.016773833
## Votes                  -0.00865873 -0.005182522
##
## , , Good
##
##                               2.5 %       97.5 %
## (Intercept)             2.08615697  2.439583895
## Has.Online.deliveryYes -1.27355587 -0.946103834
## Votes                   0.00353966  0.006763272
##
## , , Very Good
##
##                                2.5 %       97.5 %
## (Intercept)             1.158743416  1.530344873
## Has.Online.deliveryYes -1.960461483 -1.575667441
## Votes                   0.004955825  0.008193046
##
## , , Excellent
##
##                               2.5 %       97.5 %
## (Intercept)            -0.176866857  0.255244782
## Has.Online.deliveryYes -3.095977663 -2.435833138
## Votes                   0.005293757  0.008539071
```

6 Effect of Predictor Variables on Prediction Outcomes

6.1 Odds

Using the above model, odds probabilities were generated to observe the effect of changing the two predictor variables independently on the response levels.

```
fit.e = exp(coef(fit.bic.nom))
stargazer(fit.bic.nom, type = "text", coef=list(fit.e), p.auto=FALSE)
```

```
##
## ================================================================
##                                   Dependent variable:
##                          ---------------------------------------
##                           Average      Good    Very Good  Excellent
##                             (1)        (2)        (3)        (4)
## ----------------------------------------------------------------
## Has.Online.deliveryYes  0.307***   0.330***   0.171***   0.063***
##                          (0.083)    (0.084)    (0.098)    (0.168)
##
## Votes                    0.993***   1.005***   1.007***   1.007***
##                          (0.001)    (0.001)    (0.001)    (0.001)
##
## Constant                55.382***   9.611***   3.836***    1.040
##                          (0.090)    (0.090)    (0.095)    (0.110)
##
## ----------------------------------------------------------------
## Akaike Inf. Crit.      11,583.630 11,583.630 11,583.630 11,583.630
## ================================================================
## Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

Here, it can be seen restaurants with online delivery are .307 times more likely than restaurants without online delivery to be rated average rather than poor. Furthermore, this trend is shown across all the levels, suggesting that generally a restaurant with online delivery is more likely to be rated poor than any other level. For the Votes variable, a one unit increase in votes is shown to increase the odds of a restaurant being rated excellent as opposed to poor by 1.007. In fact, an increase in votes essentially translates into a restaurant being more likely to be rated anything other poor across all levels, except for the Average level where the odds are less than 1.

6.2 Relationship Between Predictors and Response Probability

The effect of the predictor variables on the response level probabilities was further explored by feeding a dummy dataset into the model. The associated plot is shown below:

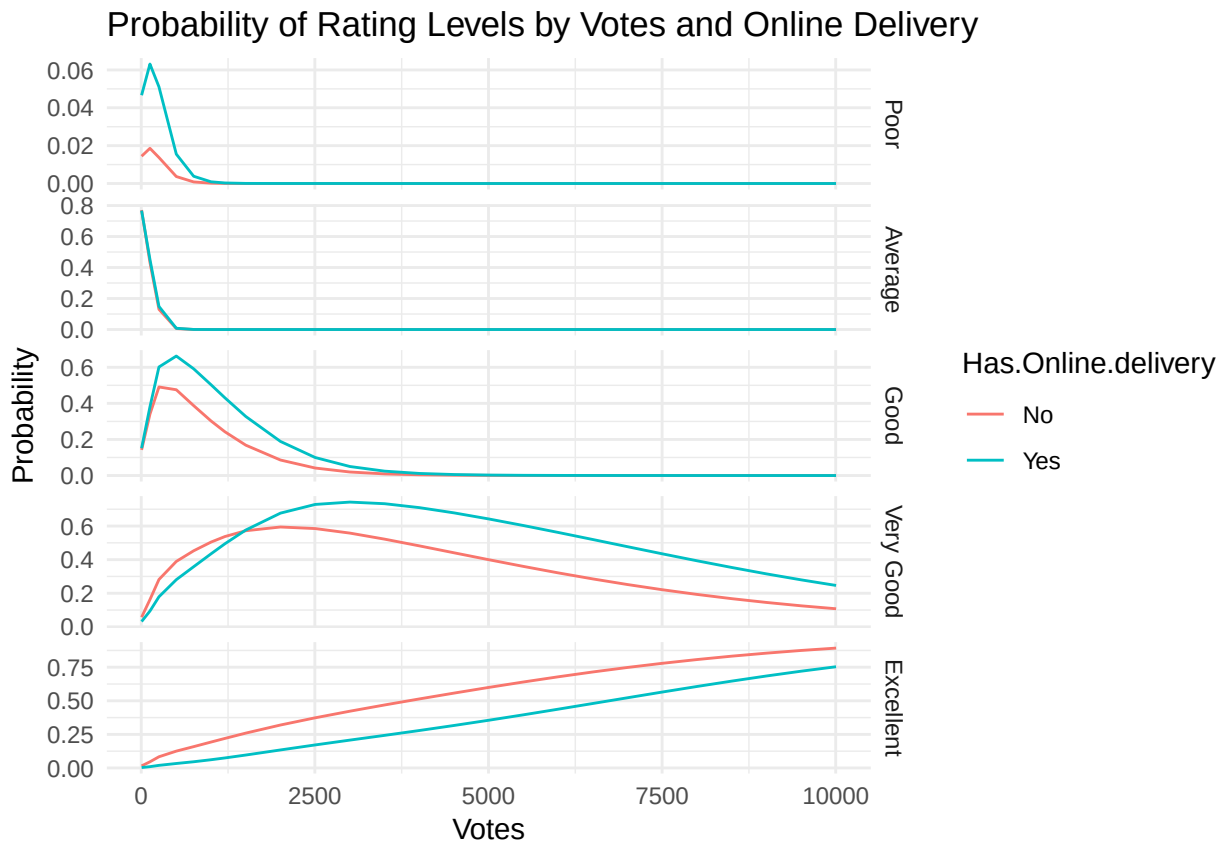

Figure 2: Probability of each rating level by number of votes and whether the restaurant offers online delivery or not according to the multinomial logistic regression model.

This visualisation provides valuable insight into how the predictors effect prediction outcomes. For example, it is clear that regardless of whether a restaurant has online delivery, if it has received over 5,000 votes, it is almost guaranteed to be rated as either very good or excellent. Furthermore, excellent restaurants are more likely to not offer online delivery regardless of how many votes are associated with it. This may be due to excellent-rated restaurant being "high-class" and therefore only offering in-house dining.

Another interesting insight relates to the "poor" facet. Evidently there is much higher relative probability of a poor-rated restaurant having online delivery than not - apparently much higher than any other response level. This might explain why the odds calculated in the previous section suggested that restaurants which did offer online delivery were more likely to be rated poor than other level.

7 Model Prediction

The test dataset was fed into the prediction model to obtain prediction responses for the `Rating.text` variable. These results were compared to the associated actual observations in the plot below:

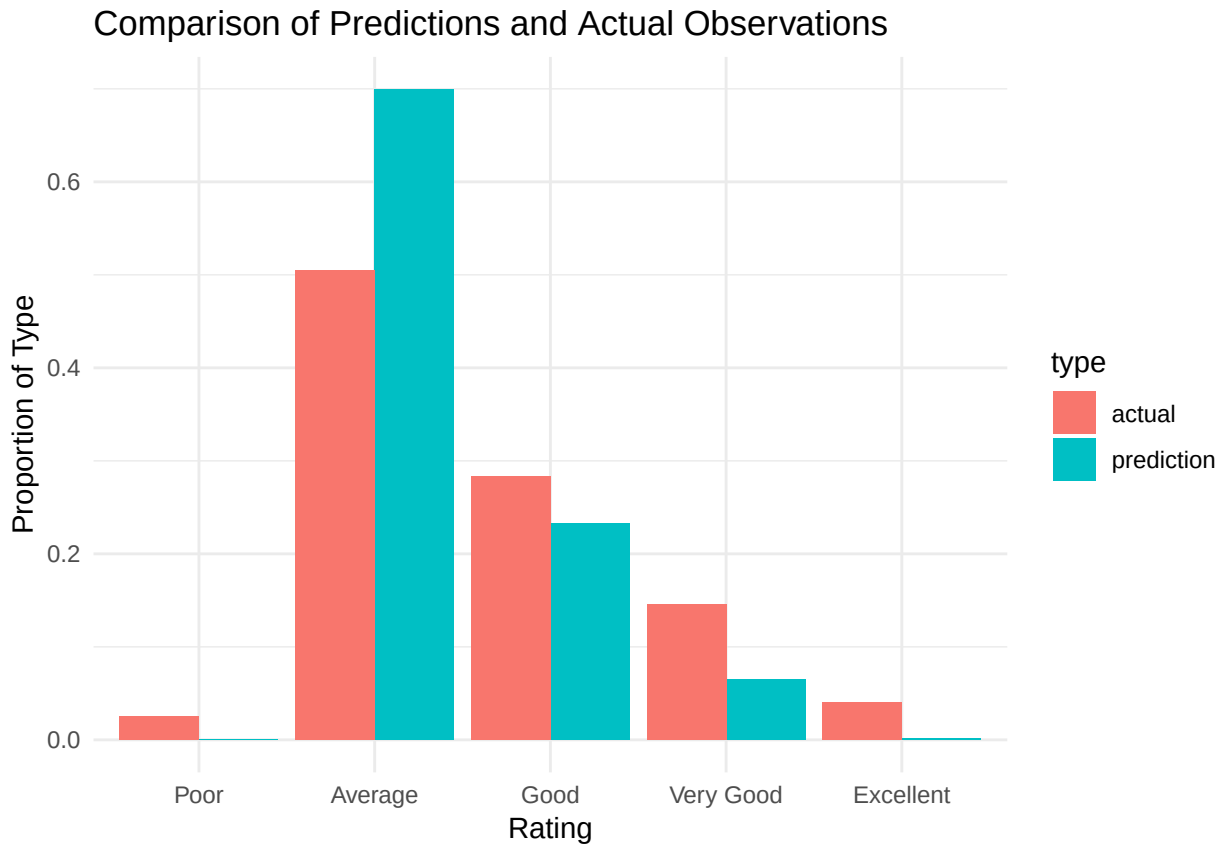

Figure 3: Bar plot comparing the proportion of the rating levels in the test dataset with the proportion of the rating levels of the predictions made by the model.

This plot highlights that the model is not performing very well. While the proportion of the predicted "good" rating level is fairly close to the actual proportion, the proportion of predicted average-rated restaurants is much higher than the true proportion. Similarly, the proportion of predicted "very good" and "excellent" levels are much lower than true levels. The model did not predict any poor-rated restaurants. This issue may be related to the imbalance across the levels. This is further highlighted in the confusion matrix below where it can be seen that the model has mostly predicted an "Average" Rating for restaurants while it never made a "Poor" prediction and very few "Very Good" and "Excellent" predictions. This results in a mmce of 0.63.

```
##            Predicted
## Actual      Poor Average Good Very Good Excellent
##   Poor         0      32    5         0         0
##   Average      0     705   41         1         0
##   Good         0     230  175        15         0
##   Very Good    0      62   99        53         2
##   Excellent    0       6   25        28         1
```

8 Summary

Two logistic regression models were considered in order to predict Zomato restaurant ratings; a multinomial and a proportional odds logistic regression model. For both models, the most important features of the dataset were used. These were found to be the number of votes a restaurant has received and whether the restaurant offers online delivery or not. The multinomial logistic regression model was found to perform slighlty better than the proportional odds model and was used for the rest of the analysis. The equations associated with the multinomial model are:

- $log(\frac{Pr(Average)}{Pr(Poor)}) = 4.014 - 1.179 \times Has.Online.Delivery - 0.007 \times Votes$
- $log(\frac{Pr(Good)}{Pr(Poor)}) = 2.263 - 1.11 \times Has.Online.Delivery + 0.005 \times Votes$
- $log(\frac{Pr(VeryGood)}{Pr(Poor)}) = 1.345 - 1.768 \times Has.Online.Delivery + 0.007 \times Votes$
- $log(\frac{Pr(Excellent)}{Pr(Poor)}) = 0.039 - 2.766 \times Has.Online.Delivery + 0.007 \times Votes$

According to this model, it was found that restaurants offering online delivery are more likely to be rated "Poor" than any other rating. Additionally, votes had a positive impact on a restaurant's rating as restaurants with more than 5,000 votes were almost guaranteed to have a "Very Good" or "Excellent" rating. However, in spite of these results, using 20% of the data as a test set, the model was found to be quite inaccurate in predicting the actual restaurant ratings. This could be attributed to the imbalance of the rating levels in the dataset as well as the potential existance of variables (not present in the dataset) more closely correlated to a restaurant's rating.

9 References

Data: https://www.kaggle.com/shrutimehta/zomato-restaurants-data

Alathea, Letaw. 2015. *Captioner: Numbers Figures and Creates Simple Captions*. https://CRAN.R-project.org/package=captioner.

Calcagno, Vincent. 2013. *Glmulti: Model Selection and Multimodel Inference Made Easy*. https://CRAN.R-project.org/package=glmulti.

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

Fox, John, and Sanford Weisberg. 2011. *An R Companion to Applied Regression*. Second. Thousand Oaks CA: Sage. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

Hlavac, Marek. 2018. *Stargazer: Well-Formatted Regression and Summary Statistics Tables*. Bratislava, Slovakia: Central European Labour Studies Institute (CELSI). https://CRAN.R-project.org/package=stargazer.

Urbanek, Simon. 2013. *Png: Read and Write Png Images*. https://CRAN.R-project.org/package=png.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Fourth. New York: Springer. http://www.stats.ox.ac.uk/pub/MASS4.

Wickham, Hadley. 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.