Trinity College Dublin

CS7DS3: Applied Statistical Modelling

Main Assignment

Conor O'Sullivan
18302737
osullc43@tcd.ie

April 12, 2019

1 Introduction

The Yelp academic dataset, which consists of two parts, was analysed. The first part contains information on many restaurants and the second contains reviews for all of these restaurants. The two parts are linked through a Business ID where each restaurant has a unique ID (Yelp 2019). The dataset has been narrowed down to consider only restaurants in Toronto. There are 3 sections to this analysis where each section focuses on a different question around the data.

Section 1 looks at comparing the average ratings of restaurants in the different neighbourhoods of Toronto. This was done using hierarchical models and Gibbs sampling. The ratings of open Indian restaurants in Scarborough and Etobicoke were compared. A less constrained analysis was also conducted which compared the ratings of all open restaurants across all the neighbourhoods.

Section 2 looks at constructing a model that can predict whether a restaurant is open or not. By constructing this model, it is also possible to determine what explanatory variables are strongly associated with the target variable. This was done using logistic regression and a simple heuristic was also discussed.

As an alternative to the neighbourhood groupings, section 3 looks at identifying alternative groupings of restaurants using latitude and longitude. This was done using a Gaussian Mixture Model. The resulting clusters were compared to an actual map of Toronto to gain a better understanding of their characteristics.

Before this analysis could occur, the data was cleaned and transformed resulting in several different datasets. This process is described in chapter 2. This is followed by chapter 3 which describes the analysis and results. Lastly, chapter 4 discusses these results as well as future work.

2 Data Handling

The raw Yelp data comes in JSON format. Namely, two JSON files: "business" and "review". The first file contains variables such as names, average star ratings (stars), the total number of reviews (review_count) and various other restaurant characteristics. The review file contains individual reviews for all the restaurants. These included a star rating and text description of the business. The two files are linked through the "business _id" variable.

These datasets have been filtered to obtain only restaurants in Toronto. This has resulted in two new JSON files: "Business_Toronto_Restaurant" (business file) and "Review_Toronto_Restaurant" (review file). The business file consists of information on 7148 restaurants in Toronto. The review file consists of 155300 individual reviews but this file seemed to be missing some reviews.

There are 7148 restaurants and therefore 7148 unique business_id's in the business file. However, there are only 7051 unique "business_id"s in the review file. In other words, reviews for 97 restaurants are missing. It was, therefore, necessary to extract the reviews for restaurants in Toronto from the original large review file.

This was done by obtaining the list of unique business_id values from the business file. Then the original review file was cycled through line by line. The 'business_id' value for each review was checked and the review was only kept if the ID was in the list of unique business_ids. This process resulted in a review file with 303780 individual reviews and 7148 unique 'business_ids. In other words, it appears as if all the restaurant reviews had been obtained. Further processing of this new review file as well as the business file was necessary for the 3 analysis sections.

Initial Data Preparation

The majority of the data preparation was done using Python and the Pandas package. "Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures

and data analysis tools for the Python programming language" (*pandas* 2019). Using the pandas function, json_normalize, the JSON files were transformed (*json_normalize* 2019).

The json_normalize function was used to convert the business file to a more structure dataset (business dataset). That is, each row of the dataset is a different restaurant and the columns give the restaurant's attributes (e.g. 'business_id, stars and review_count). Similarly, this was done to the review file and a snapshot of the resulting review dataset can be seen in Figure 1. Both the review and business datasets where then saved as CSV files.

	business_id	cool	date	funny	review_id	stars	text	useful	user_id
0	9_CGhHMz8698M9-PkVf0CQ	2	2012-05-11	0	ymAUG8DZfQcFTBSOiaNN4w	4	Who would have guess that you would be able to...	0	u0LXt3Uea_GidxRW1xcsfg
1	5r6-G9C4YLbC7Ziz57l3rQ	0	2013-02-09	0	w41ZS9shepfO3uEyhXEWuQ	3	Not bad!! Love that there is a gluten-free, ve...	1	u0LXt3Uea_GidxRW1xcsfg

Figure 1: Snapshot of Review Dataset

Subsequent Data Preparation

For section 1, 3 different datasets where constructed. For the first dataset (Dataset 1), the categories variable for the Business Dataset was transformed into an 'Indian' variable. This is an indicator variable that takes on the value 1 if the Indian category is in the categories of a restaurant and 0 otherwise. The business was then filter using 'Indian', 'is_open' and the 'neighborbood' variables to obtain only the average star ratings of open Indian restaurants in either Scarborough or Etobicoke. The unnecessary columns were then dropped. A snapshot of the resulting dataset can be seen in Figure 2. In the end, there were 48 rows in total - 34 from Scarborough and 14 form Etobicoke.

	business_id	name	neighborhood	review_count	stars
0	YLeAUkFFJri_21uTkyrVFg	Malaysian Garam Masala	Scarborough	4	4.0
1	E1qCzAOYLTnbT9c6TA-YjA	Royal Paan Rexdale	Etobicoke	3	2.5
2	A7lQT6RYVSWkvEAbiLCdtQ	Desi Spice	Etobicoke	30	4.5

Figure 2: Snapshot of Dataset 1

The second dataset (Dataset 2) for section 1 is very similar to Dataset 1. Except that the individual star reviews from the Review dataset are considered. Using the Review dataset, all the individual ratings for these 48 businesses were obtained. This was done by taking the 48 'business_id' values from Dataset 1 and matching them to the reviews in the Review dataset. The result was 976 individual star ratings - 687 for Scarborough and 289 for Etobicoke.

The final dataset (Dataset 3) for section 1 was obtained by filtering the Business dataset to obtain only open restaurants. There were 1257 rows where the 'neighborhood' variable was missing. These rows were dropped as this variable is necessary for the analysis in this section. In the end, there were 4028 average star ratings for 72 different neighbourhoods. Some of these neighbourhoods had only 1 restaurant (i.e. Cooksville and Meadowvale Village). Due to the fact that the variance of ratings is calculated in the modelling process, these two neighbourhoods were grouped together and relabelled as 'other'.

For section 2, a dataset was constructed (Dataset 4) by extracting potential explanatory variables from both Business and Review datasets. Most of the variables from the Business dataset where not considered as they were quite sparse. For instance, 23% of the 'Alcohol' variable was missing, 91% of 'smoking' was missing and only 7 restaurants gave information on their 'AgesAllowed' variable. However, from the Business dataset, 'latitude', 'longitude', 'review_count', 'stars' as well as the target variable, 'is_open' were selected. These variables could be extracted without any transformation and had no missing values. Additionally, 4 explanatory variables were derived from the 'categories' variable.

A restaurant's categories might be related to restaurant closure. For example, some restaurant categories may have become less popular over time leading to their closure. Each restaurant has a list of categories associated with it, for example, ['Italian', 'French', 'Restaurants']. It is necessary to transform this variable before it can be used in a logistic regression model. This could be done using indicator variables. However, the problem is that there are 323 unique categories which would result in 323 variables. To reduce the complexity of the model it is necessary to select only the most predictive variables. This was done by finding the top 10 most common categories for both open and closed restaurants. The categories that did not appear in both lists were selected. These restaurants are potentially popular for open restaurants and unpopular for closed restaurants and visa versa. In the end, the indicator variables, 'Italian', 'Japanese', 'Pizza' and 'Coffee & Tea' were selected where these variables have value 1 if the category is present for the restaurant and 0 otherwise.

Additionally, for each restaurant, the following potential explanatory variables were obtained from the Review dataset:

- $stars\_i$ for := the total number $i = 1, ..., 5$ star ratings
- $length$:= the average number of words of the text reviews
- $last\_year$:= the year of the last review

A snapshot of the final dataset can be seen in Figure 3. Section 3, used this same dataset but only considered the latitude and longitude columns.

	business_id	name	latitude	longitude	review_count	stars	is_open	Italian	Japanese	Pizza	Coffee & Tea
0	l09JfMeQ6ynYs5MCJtrcmQ	Alize Catering	43.71140	-79.39934	12	3.0	0	1	0	0	0
1	1K4qrnfyzKzGgJPBEcJaNQ	Chula Taberna Mexicana	43.66926	-79.33590	39	3.5	1	0	0	0	0
2	nbhBRhZtdaZmMMeb2i02pg	Sunnyside Grill	43.78182	-79.49043	3	5.0	1	0	0	0	0

stars_1	stars_2	stars_3	stars_4	stars_5	length	last_year
2	2	4	3	1	1475.750000	2012
0	4	11	20	4	643.153846	2017
0	0	0	0	3	218.333333	2017

Figure 3: Snapshot of Dataset 4

3 Analysis

3.1 Section 1

This section can be broken down into two parts. In both parts, the star ratings of different neighbourhoods are analysed. Part 1 specifically looks at open Indian restaurants in two different neighborhoods - Scarborough and Etobicoke. Part two is more general and compares the ratings of open restaurants in all the neighbourhoods.

To compare ratings it is important to distinguish between two different types. The first type is the average star ratings which were obtained from Dataset 1. Every restaurant has one average star rating. The second type is the individual star ratings obtained from Dataset 2. Each restaurant will have multiple individual star ratings and the average rating is obtained from these by taking the average. Part 1 considers both types whereas part 2 considers just average star ratings.

Part 1

Table 1 shows the summary of both the average star rating and individual star ratings of both neighbourhoods. It can be seen that, for the average star rating, the mean is 0.32 stars higher and the standard deviation is also lower for Scarborough. This suggests better average ratings for this neighbourhood. However, at this point, something should be said about the number of reviews per restaurant (i.e. 'review_count') as it tells us something about the reliability of an average rating. For example, see Figure 4. Both restaurants have an average star rating of 4.5 but the restaurant in Etobicoke has 10 times more reviews. The ratings for these restaurants could, therefore, be considered more reliable. Considering this, the weighted average of the average star rating for each neighbourhood was calculated where the weights are the number of reviews. In this case, Scarborough had a weighted average of 3.78 which was 0.02 stars lower than Etobicoke's weighted mean of 3.8.

Instead of considering a weighted average, individual reviews can be used. Using these reviews is similar to considering the weights as the higher a restaurant's 'review_count' the more individual reviews it will have. Table 1, shows that the mean individual star rating for Scarborough is 0.08 lower than that of Etobicoke. This is an interesting result as the average ratings for Scarborough seem to be better whereas the opposite is true for the individual ratings.

Another thing to notice is that the standard deviations of the individual ratings are both greater than for the corresponding standard deviations of the average ratings. This makes sense as we would expect some variance to be removed when considering means of ratings. This can also be seen in Figure 5 where the distributions of the individual stars are more spread out.

	Average Stars		**Individual Stars**	
	Scarborough	**Etobicoke**	**Scarborough**	**Etobicoke**
Count	34	14	687	289
Mean	3.68	3.36	3.78	3.86
Std.	0.5625	0.9492	1.2470	1.3209
Min	2.5	1	1	1
25%	3.5	2.75	3	3
50%	3.5	3.5	4	4
75%	4	4	5	5
Max	4.5	4.5	5	5

Table 1: Summary of Restaurant Average Stars and Individual Stars

	business_id	name	neighborhood	review_count	stars
2	A7lQT6RYVSWkvEAbiLCdtQ	Desi Spice	Etobicoke	30	4.5
11	zwMg7Qey7suZlw31S3jKmA	Araliya Take-out & Catering	Scarborough	3	4.5

Figure 4: Snapshot or Review Counts for Average Stars

A deeper understanding of the difference in the star ratings can be gained using a hierarchical model and Gibbs sampling. The same model was used for both average and individual ratings.

$$Y_{1i} \sim N(\theta_1, \tau) \; for \, i = 1, ..., n$$
$$Y_{2j} \sim N(\theta_2, \tau) \; for \, j = 1, ..., m$$

$$\theta_1 = \mu + \delta$$
$$\theta_1 = \mu - \delta$$

where Y_{1i} is a rating from Scarborough and Y_{2j} is a rating from Etobicoke. The model also has the following prior distributions and prior parameters :

$$\begin{aligned} \mu \sim N(\mu_0, \tau_0) \quad \mu_0 = 3 \quad \tau_0 = 1.1378 \\ \delta \sim N(\delta_0, \gamma_0) \quad \delta_0 = 0 \quad \gamma_0 = 1.0256 \\ \tau \sim G(\alpha_0, \beta_0) \quad \alpha_0 = 2 \quad \beta_0 = 1.7578 \end{aligned}$$

The above are weak informative prior parameters as they were chosen using minimal prior information. The star ratings, and so μ, can take on values in the range $[1, 5]$. μ_0 was set to the middle value and τ_0 was then selected so that roughly 95% of the interval was covered (i.e. $\mu_0 + 2\sigma_0 = 0.975(5)$ and $\tau_0 = 1/{\sigma_0}^2$). δ_0 and γ_0 were chosen in a similar way excepting the range of possible values for δ is $[-2, 2]$. A large α_0 indicates a high certainty of the priors and it is suggested that it should not be smaller than 2. This minimum value was selected due to the minimal prior information. Finally, β_0 was selected as $2\sigma_0$ so that $E[1/\tau_0] = \sigma_0$.

Gibbs sampling with 5000 iterations was then used to obtain the posterior distributions of μ, δ and τ. This was done using both the average star ratings and individual star ratings. Before these results are explained, the distributions in Figure 5 should be mentioned. The model is based on the assumption that the star ratings are normally distributed. However, looking at Figure 5 this is not necessarily the case. The average star ratings are potentially normally distributed although both neighbourhood distributions are slightly skewed to the right. Looking at the individual stars reviews it is unlikely that they have a normal distribution as the distributions seem significantly skewed to the right. Nonetheless, the model was still used but any results should be analysed with these potential assumption violations in mind.

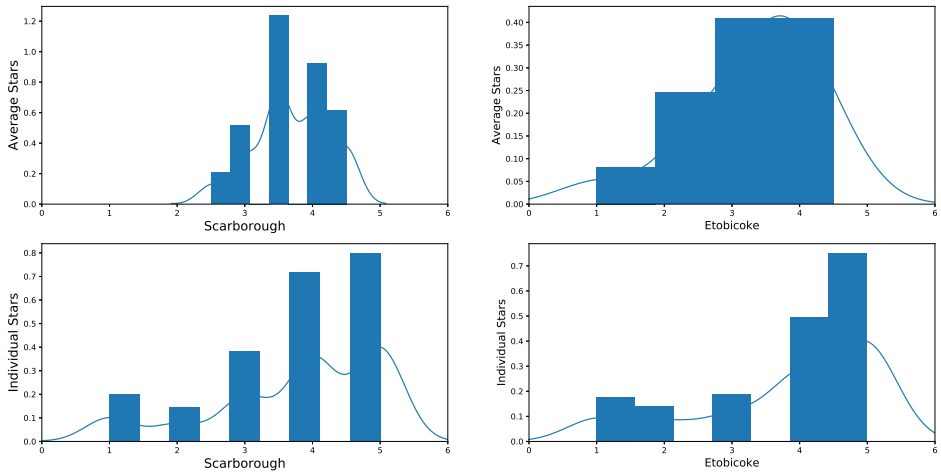

Figure 5: Distributions of Average Stars and Individual Stars

The average star ratings were analysed first. Table 2 shows the summary of the variables posterior distributions using these ratings. Importantly, the mean of δ is 0.1592 which is greater than zero. This suggests that the mean rating of Scarborough is greater than that of Etobicoke. Figure 6 shows the empirical posterior distribution of μ and τ. These variables seem to have a normal and gamma posterior distribution respectfully. Figure 7 gives the posterior distribution for δ. The three red lines give the distribution's empirical 2.5% quartile, mean and 97.5% quartile

respectfully. The quartile values can also be seen in Table 2 and they help determine whether δ is significantly different from 0. Notice that 0 does fall within this 95% empirical confidence interval. Additionally, by summing the number of δ samples less than zero and dividing by the total number of samples, it was determined that 0 is the 8.46% quantile. In other words, null hypothesis that δ is 0 cannot be rejected at the 5% confidence value. However, 0 is still fairly close to the edge of the confidence interval suggesting that with some more analysis we could say it is different from zero with certainty.

	μ	δ	τ	σ
count	5000	5000	5000	5000
mean	3.5120	0.1592	1.9464	0.7276
std	0.1160	0.1150	0.3877	0.0739
min	3.0697	-0.2619	0.8911	0.5210
2.50%	3.2811	-0.0665	1.2606	0.6008
50%	3.5131	0.1591	1.9149	0.7227
97.50%	3.7402	0.3853	2.7700	0.8907
max	3.9463	0.5889	3.6845	1.0593

Table 2: Summary of Simulation Results using Average Stars

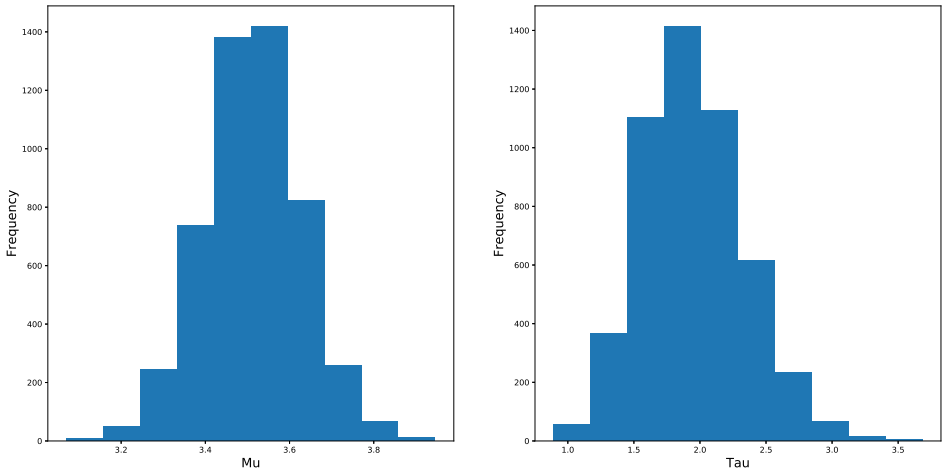

Figure 6: Posterior Distribution of Mu an Tau using Average Stars

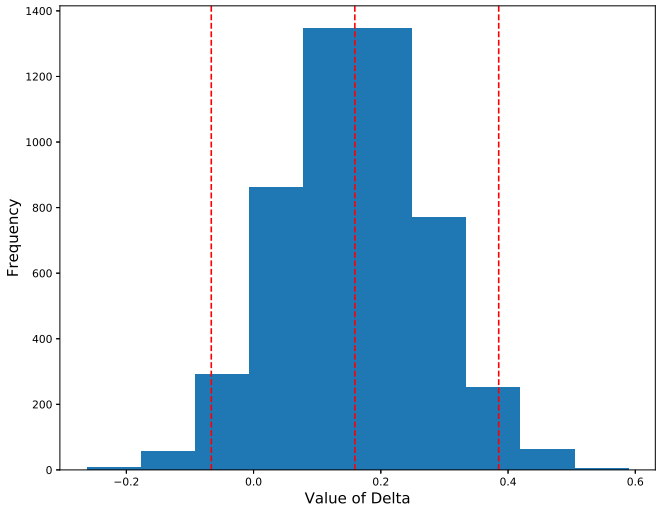

Figure 7: Distributions of Delta using Average Stars

The simulated values can also be used to estimate the probability that the mean rating of Scarborough is greater than that of Etobicoke. This is done by simulating 5000 ratings from each neighbourhood as follows:

$$\begin{aligned} Y_{1i} &\sim N(\mu_i + \delta_i, \tau_i) \\ Y_{2i} &\sim N(\mu_i - \delta_i, \tau_i) \\ for\, i &= 1, ..., 5000 \end{aligned}$$

The for $i = 1, ..., 5000$, the number of times $Y_{1i} > Y_{2i}$ is counted. This number is then divided by 5000 to obtain the estimated probability. In this case, it was estimated that the Scarborough's mean rating had a 62.76% chance of being greater than that of Etobicoke.

Next, a very similar analysis was conducted but using individual reviews. Table 3 shows the summary of the variables posterior distributions using these ratings. In this case, the mean of δ is now negative suggesting that the mean rating of Etobicoke is greater. Figure 8 shows that the empirical posterior distribution of μ and τ appear to be normal and gamma respectfully. Although, in this case, the distribution for τ appears to shifted more to the right than for the previous analysis. Figure 9, gives a similar plot for δ as before. Again we see that 0 does not fall out of the 95% confidence interval. As before, the probability that the mean rating of Scarborough is greater than that of Etobicoke was simulated. In this case, the probability was 0.4838 suggesting that Etobicoke had a higher probability of having a higher mean rating.

	μ	δ	τ	σ
count	5000	5000	5000	5000
mean	3.8211	-0.0389	0.6221	1.2688
std	0.0441	0.0446	0.0279	0.0285
min	3.6523	-0.1943	0.5227	1.1727
2.50%	3.7362	-0.1263	0.5673	1.2154
50%	3.8210	-0.0392	0.6218	1.2682
97.50%	3.9063	0.0482	0.6770	1.3277
max	3.9833	0.1224	0.7271	1.3831

Table 3: Summary of Simulation Results using Individual Stars

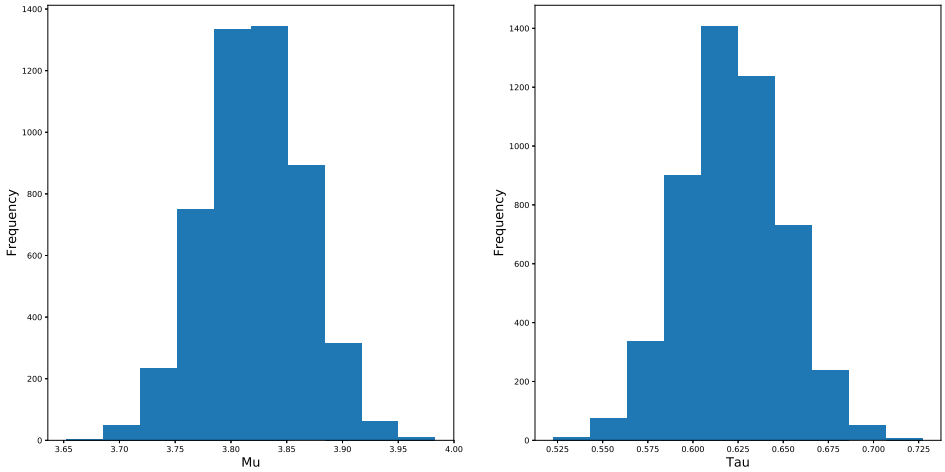

Figure 8: Posterior Distribution of Mu an Tau using Individual Stars

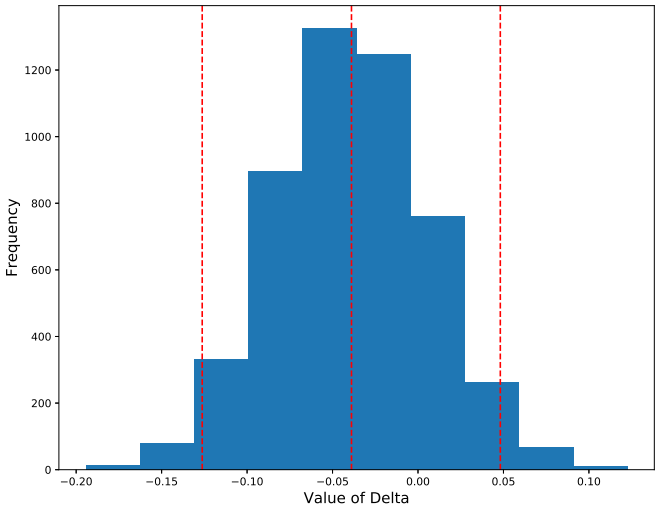

Figure 9: Distributions of Delta using Individual Stars

The results of the two different analysis seem to be in contradiction. The former set of results seem to favour the conclusion that Scarborough has a high mean rating while the latter seems to favour Etobicoke. However, in both cases it was not found that the difference in means was statistically significant.

Part 2

The boxplots of the average star ratings for the 71 different neighbourhoods can be seen in Figure 10. The neighbourhoods have been ordered by their mean average star rating. Most of the neighbourhoods have a median value of 3.5 stars. South Hill has the highest median value of 4.25 and Markland Wood has the lowest median of 2.75. The figure also gives an indication of the variance of the different groups.

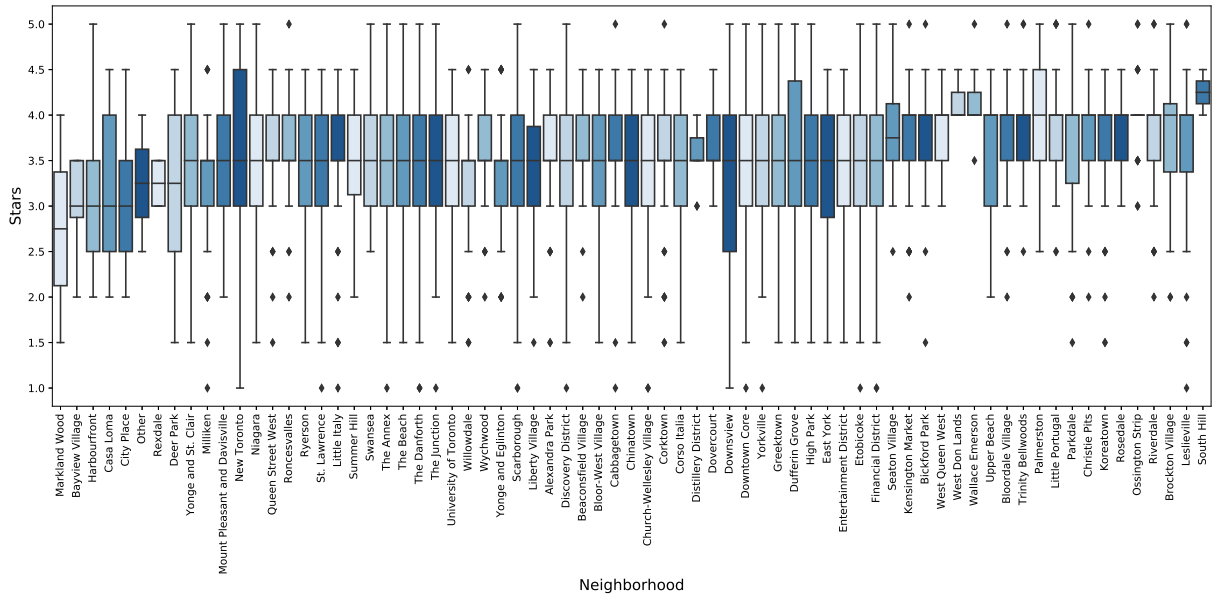

Figure 10: Prior Boxplot of Restaurant Average Stars

A hierarchical model and Gibbs sampling were used to compare the mean of the average star ratings of the different neighbourhoods. The hierarchical model is described below.

$$Y_{ij} \sim N(\theta_j, \tau_w) \quad for\, i = 1, ..., n_j \quad j = 1, ..., 71$$

where Y_{ij} is the ith rating from the jth neighbourhood. The model also has the following prior distributions and prior parameters:

$$\begin{aligned} \theta_j &\sim N(\mu, \tau_b) \\ \mu \sim N(\mu_0, \tau_0) \quad \mu_0 = 3 \quad \tau_0 &= 1.1378 \\ \tau_w \sim G(\alpha_0, \beta_0) \quad \alpha_0 = 2 \quad \beta_0 &= 1.7578 \\ \tau_b \sim G(\eta_0, t_0) \quad \eta_0 = 2 \quad t_0 &= 1.7578 \end{aligned}$$

Where θ_j is the mean rating for neighbourhood j, μ is the overall mean rating, tau_w is the within group precision and tau_b is the between group precision.

Again, weak informative prior parameters were chosen. μ_0 and τ_0 were chosen the same way as μ_0 and τ_0 in part 1. Similarly, α_0 and η_0 were chosen the same way as α_0 and β_0 and t_0 the same way as β_0 in part 1.

The summary of the posterior distributions can be seen in Table 4. The posterior mean of μ is 3.5165 which is only 0.0165 stars higher than the most common prior median of 3.5. It is easier to interpret the distributions of the within-group standard deviation, σ_w, and between group standard deviation σ_b, than the corresponding precision distributions. Notice that the mean value of σ_b is 11.7 times larger than that of σ_w. This suggests that there is more variance between groups than within groups. In other words, there should be some significant difference in the θ_j values. The posterior distributions of μ , tau_w and tau_b can also be seen in Figure 11. μ appears to have normal posterior distribution. tau_w and tau_b both appear to have gamma distributions although tau_b is more skewed to the left.

	μ	τ_w	τ_b	σ_w	σ_b
count	5000.0	5000.0	5000.0	5000.0	5000.0
mean	3.5165	1146.6889	8.5724	0.0295	0.3451
std	0.0408	26.0407	1.414	0.0003	0.0289
min	3.3382	1053.8715	4.1792	0.0285	0.2652
25%	3.4889	1129.1141	7.5856	0.0293	0.3248
50%	3.5167	1146.5131	8.4958	0.0295	0.3431
75%	3.5429	1164.1017	9.4812	0.0298	0.3631
max	3.6756	1232.4452	14.2157	0.0308	0.4892

Table 4: Multiple Gibbs Posterior Variables Summary

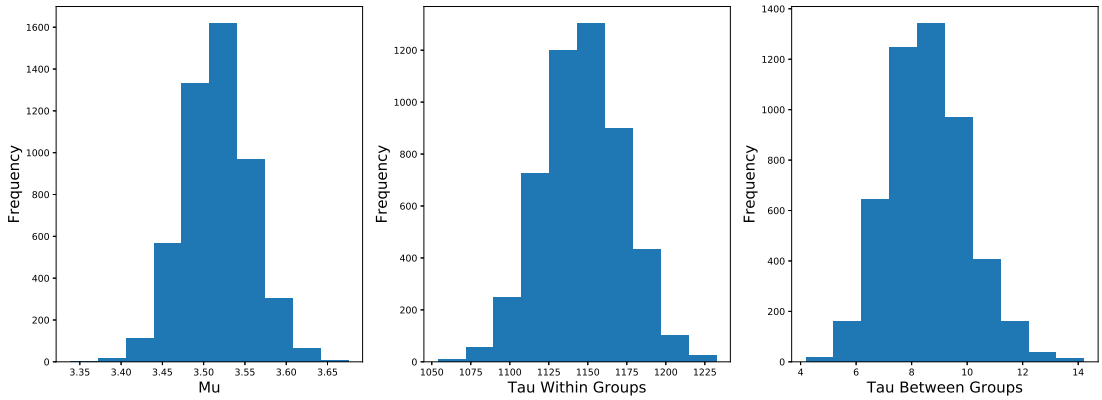

Figure 11: Posterior Distributions of Mu and Tau Parameters

Figure 12 visualises the simulated θ_j values. The width of the bar for neighbourhood j is determined by the mean of θ_j. The black line at the top of the bars is given by θ_j's empirical 95% confidence interval. If the confidence intervals for two neighbourhoods do not overlap it can be said that the mean θ values are significantly different. In other words, the average star ratings of these neighbourhoods are significantly different. The red line gives the overall mean (i.e. posterior mean of μ). This makes it easy to see which neighbourhoods have below and above average mean ratings.

The plot shows that South Hill has the best average rating and is significantly better than the second best neighbourhood, West Don Lands. Similarly, Markland Wood has the worst average rating. An interesting result is that from this analysis it can we determined that the mean rating of Etobicoke is significantly better than that of Scarborough. This is a conclusion that could not be made using the two means hierarchical model. Although, in this case, all

types of restaurants are considered and not just Indian restaurants.

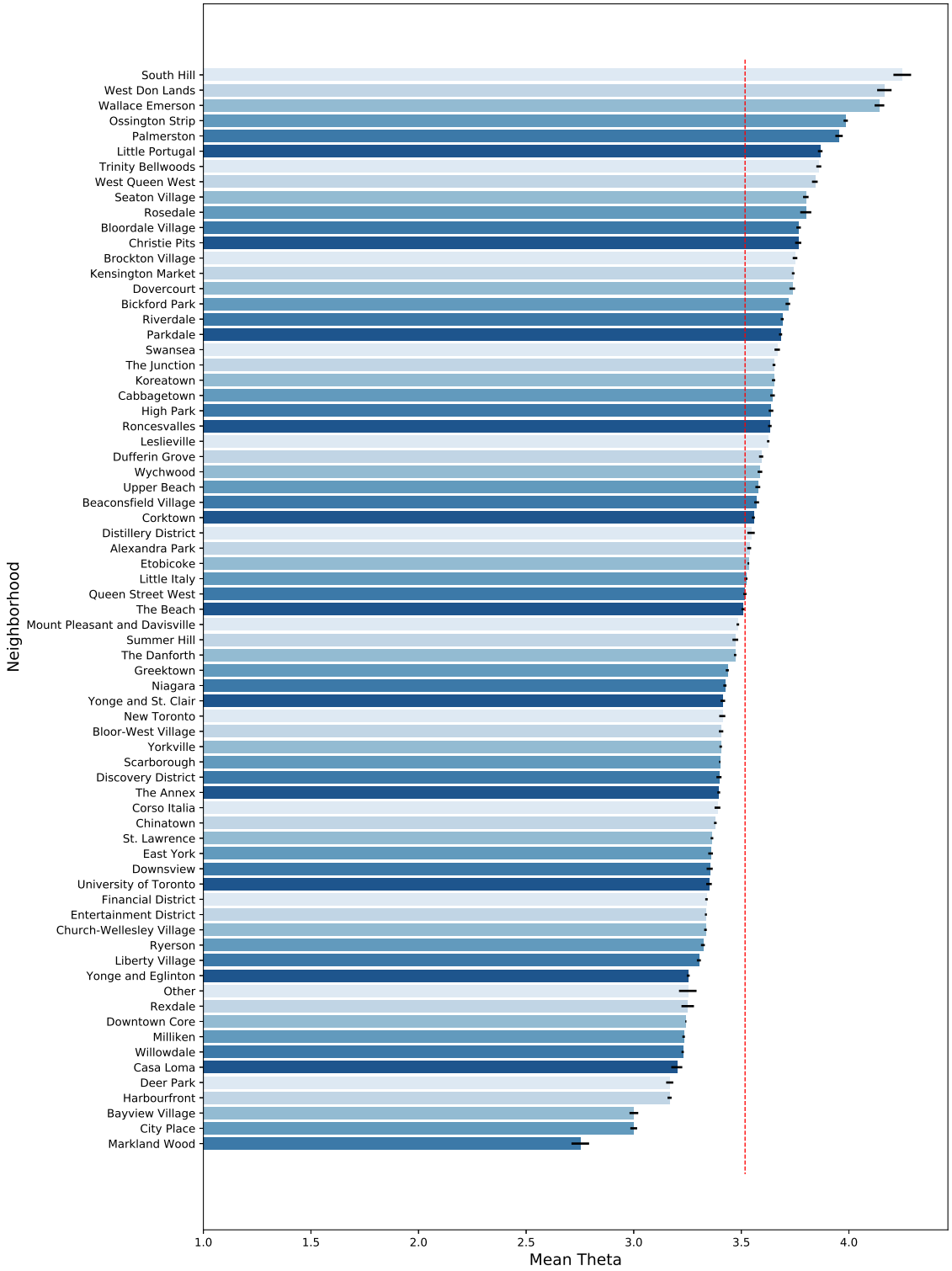

Figure 12: Barplot of Posterior Theta Means

3.2 Section 2

This section looks at creating a predictive model of the target variable, is_open. The aim was not only to create a predictive model. It was also necessary to gain an understanding of which

explanatory variable are good predictors and what sort of relationship these variables have with the target variable. Logistic regression was used to create these models. This was done by implementing the Python "Logit" package provided by StatsModel (StatsModels 2019). The "fit" function was used to train the models. This function fits a stand logistic regression model with no regularization. That is the log of the odds is modelled as a linear function of the explanatory variables.

Before models were fitted, a balanced dataset was constructed. Dataset 4 had 2180 closed restaurants and 4968 open restaurants. Subsequently, all 2180 closed restaurants and a random sample of 2180 open restaurants were selected. This lead to a dataset of 4360 restaurants which was then split into a training set (90%) and a test set (10%). Using the training set, the 10-fold cross-validation accuracy of the various feature combinations was calculated. The model with the highest accuracy was then refitted with all the training data and used to make predictions on the test set. The accuracy on the test set gives us a better indication of how the model would perform on out of sample data than the 10-fold accuracy. This is because, through the model fitting processes, the model may have become biased towards the training data. The model's results were also compared to the results of a simple heuristic.

Before all of this, an initial exploratory analysis was conducted. This involved using various figures and table to gain a better understanding of the data. Ultimately, the results of this section gave a good indication of which of the variables were potentially good predictors of the target variable. It also indicated which variables should not be considered due to high correlations with other explanatory variables. Using these insights, a few different combinations of the explanatory variables where tested.

Exploratory Analysis

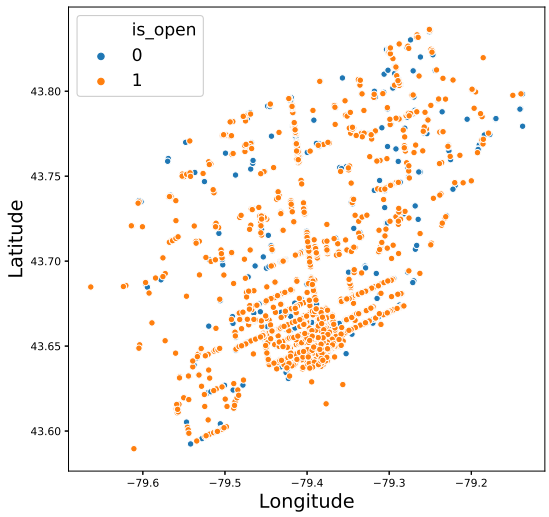

Figure 13: Scatter plot of Latitude and Longitude

Latitude and longitude were first analysed to determine whether there was a difference in the geographical location of the open and closed restaurants. This would be the case, if say for instance, certain areas of Toronto have become unpopular resulting in many closed restaurants. Figure 13 shows the distribution of restaurants where the blue and orange dots are closed and open restaurants respectfully. There seems to be a cluster of restaurants at a latitude 43.65 and longitude -79.4. It is also important to note that there are many blue dots hidden in this

cluster. From the figure, there doesn't seem to be any obvious relationship between location and the target variable. This suggests that closed restaurants are replaced by open restaurants in similar locations.

Table 5 shows the confusion matrices of two categorical variables constructed in chapter 2. We can see when a restaurant is not Italian (i.e. $Italian = 0$) the restaurant is more likely to be open (i.e. $2023 > 1970$). When $Italian = 1$ the restaurant is more likely to be closed (i.e. $210 > 157$). This suggests that Italian restaurants may have become less popular in Toronto. The opposite is true for the Pizza category. This is expected as it follows from the way the variables where constructed. That is Italian was more popular amongst closed restaurants and Pizza more likely amongst open. We should expect similar results for Japanese and Coffee & Tea categories.

		Italian		**Pizza**	
		0	**1**	**0**	**1**
is_open	**0**	1970	210	2060	120
	1	2023	157	2002	178

Table 5: Category Indicator Variable Confusion Matrices

Boxplots were used to investigate the average star and average review length variables. These are shown in Figure 14. From this figure, the avenger stars distribution seems to be the same for both open and closed restaurants. This is surprising as we have expected poor ratings to result in restaurants closing down. That is the average star would be lower on average for closed restaurants. For length there does appear to be a difference in the distribution. It appears as though the text review of closed restaurants are generally longer. Following from this figure, we can expect Length to be a potential explanatory variable and average stars to be a poor explanatory variable.

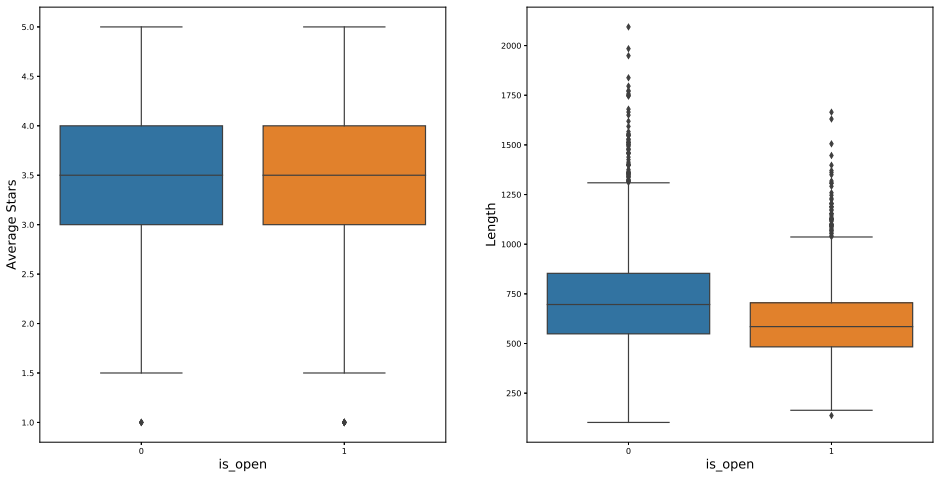

Figure 14: Boxplots of Average Stars and Length Variables

The individual star counts may provide more information than the average star rating above. Figure 15 shows the average number of star ratings (i.e. for ratings 1 to 5) as well as the average total number of reviews. For all the variables, the distribution for open and closed restaurants seem to be different. However, this appears to be simply because the total number of reviews

for open restaurants is on average greater than for closed restaurants. If you look closely the counts for all these variables appear to differ by a similar proportion. In other words, when the number of reviews is high the other variables will also have high counts. This suggests there may be significant multicollinearities between these variables and that it is only necessary to include one of these variables in the model.

Figure 16, confirms this. This figure shows the Pearson correlation matrix of these variables in the form of a heatmap. The redder a square the more correlated the two variables. We can see that review_count is highly correlated with star counts 2 to 5. It is not as highly correlated with star count 1. There also seems to be multicollinearities amongst the different star counts. Subsequently, only review_count was considered for the model fitting process.

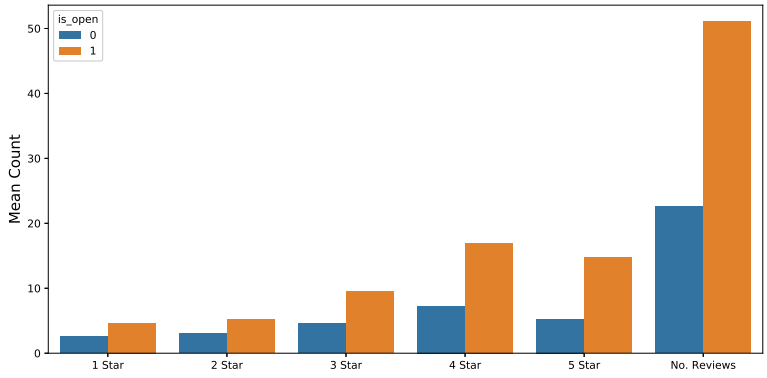

Figure 15: Barplot of Mean Counts of Star and Number of Review

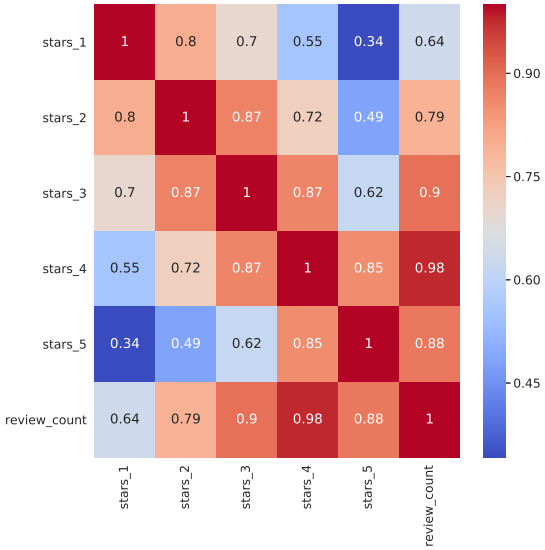

Figure 16: Star and Number of Review Counts Correlation Heatmap

Lastly, the last review year variable was investigated. The boxplot in Figure 17 shows that the distribution of this variable differs significantly for the open and closed restaurant. The last review year for the majority of the open restaurants in 2017 with a few outliers whereas the distribution for closed restaurants is more spread out. This makes sense as we would expect open restaurants to continually be receiving reviews resulting in the last review date close to

the current date (or the date the data was extracted). In contrast, once a restaurant is closed it can no longer receive reviews. This would result in a lower last review year on average for closed restaurants.

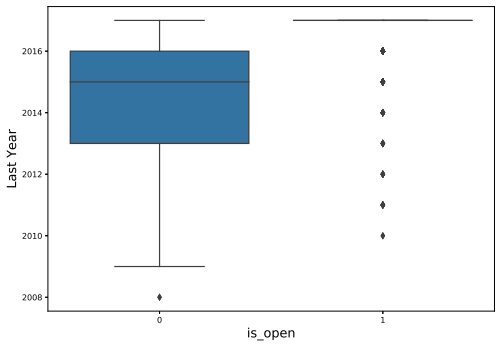

Figure 17: Boxplot of Last Year

Logistic Regression

Three different sets of parameters were tested and the 10-fold accuracy of these models can be seen in Table 6. Model 1 has all the variables discussed (specifically those selected for the model fitting process) above except 'stars' and 'last_year'. For Model 2, the 'stars' variable was added. This variable does not seem to have much effect on the accuracy of the model. Finally, for Model 3, we included the 'last_year' variable which appears to have significantly increased the accuracy of the model. The inclusion of this variable increased the accuracy by 19.06% percentage points compared to Model 2.

Model 3 was then retrained on all the training data and used to make predictions on the test set. Ultimately, the model achieved an accuracy of 0.8876 on the test set. The summary of the model can be seen in Tabel 7. The most important columns are 'coefficients' and 'P-value'. A variable's coefficients tell us the nature of the relationship that the variable has with the target variable. A variable's P-value tells us whether that relationship is statistically significant. A low P-value allows us to reject the hypothesis that the variable's coefficient is not different from zero.

Model	**Variables**	**Accuracy**
1	latitude, longitude, review_count, Italian, Japanese, Pizza, Coffee & Tea, length	0.6718
2	latitude, longitude, review_count, stars, Italian, Japanese, Pizza, Coffee & Tea, length	0.6715
3	latitude, longitude, review_count, stars, Italian, Japanese, Pizza, Coffee & Tea, length, last_year	0.8621

Table 6: 10-Fold Cross-validation of Logistic Regression Models

	Coefficient	Std Error	z	P-Value	[0.025	0.975]
const	-3124.7505	143.833	-21.725	0.000	-3406.659	-2842.842
latitude	1.6488	1.052	1.567	0.117	-0.414	3.711
longitude	-0.8720	0.786	-1.109	0.267	-2.412	0.668
review_count	0.0030	0.001	3.469	0.001	0.001	0.005
Italian	-0.4132	0.173	-2.391	0.017	-0.752	-0.074
Japanese	-0.4114	0.174	-2.367	0.018	-0.752	-0.071
Pizza	0.3596	0.186	1.931	0.053	-0.005	0.725
Coffee & Tea	-0.2182	0.176	-1.241	0.215	-0.563	0.127
length	-0.0012	0.000	-5.249	0.000	-0.002	-0.001
last_year	1.4802	0.056	26.446	0.000	1.370	1.590

Table 7: Logistic Regression Results

Looking at equations 1-3 below, we can see how the model's coefficients can be interpreted. Equation 1 shows a standard logistic regression model where π is the probability that $is\_open = 1$ and so $\pi/(1-\pi)$ are the odds that a restaurant is open. $x_1, ..., x_n$ and $b_0, ..., b_n$ are the model's variables and coefficients respectfully. In case of Model 3, $n = 9$ and, for example, x_1 is the restaurant's latitude and $b_1 = 1.6488$.

Moving on to Equation 2, Equation 1 can be rearranged so that the odds can be written as an exponential function of the linear equation. Then to get Equation 3, an arbitrary continuous variable x_p is increased by 1 unit. This results in the product of the exponential function in Equation 2 and e^{b_p}. In other words, holding all else constant, a one unit increase in variable x_p increases/decreases the odds by a factor of e^{b_p}. For example, suppose there were two identical restaurants (R1 and R2) except that R2 had a latitude 1 unit higher. Based on the model, it is expected that the odds that R2 is open is $e^{1.6488} = 5.2$ times greater than those of R1. The interpretation of indicator variables is very similar. In this case, $x_p = 0$ or $x_p = 1$ and so the odds are either increase/decreased by factor e^{b_p} or not.

$$log(\frac{\pi}{1-\pi}) = b_0 + b_1x_1 + ... + b_nx_n = X^T\beta \tag{1}$$

$$\frac{\pi}{1-\pi} = e^{b_0+b_1x_1+...+b_nx_n} = e^{X^T\beta} \tag{2}$$

$$\begin{aligned} \frac{\pi}{1-\pi} &= e^{b_0+b_1x_1+...+b_p(x_p+1)+...+b_nx_n} \\ &= e^{b_0+b_1x_1+...+b_px_p+...+b_nx_n}e^{b_p} \\ &= e^{X^T\beta}e^{b_p} \end{aligned} \tag{3}$$

It follows that, if a coefficient is positive, the corresponding variable will have a positive relationship with the odds. For instance, we see that 'last_year' has a positive relationship with the target variable. In other words, the greater the year of a restaurant's last review to greater the odds that that restaurant is open. This is consistent with the results from the exploratory analysis. Similarly, if a coefficient is negative, the corresponding variable has a negative relationship with odds. Both 'Italian' and 'Japanese' have a negative relationship with the odds. This is unsurprising as it was already observed that these categories were more popular amongst closed restaurants.

One unexpected result is that 'Coffee & Tea', has a negative relationship with the odds. However, the P-Value of this parameter is 0.215 and so we cannot say with certainty that this coefficients value is not 0. This is also reflected in the coefficient's 95% confidence interval (-0.563, 0.127) which includes 0. This suggests that 'Coffee & Tea' may not have any relationship

with the odds rather than a negative relationship. Similarly, it also appears that the latitude and longitude coefficients are not significantly different from 0. This was reflected in the exploratory analysis where no clear relationship between location and the target variable was observed.

Ultimately, the sign of the coefficient gives the direction of the relationship and the P-Value gives the significance of the relationship. Additionally, the size of the coefficient gives us an indication of how much the variable affects the odds. However, it is important to understand that the different variables have different scales and so the size of the parameters may be misleading.

Heuristic

The accuracy of Model 3 is significantly higher than the accuracy of random guess (50%). However, it is also possible to compare this model to a simple heuristic derived from the 'last_year' variable:

$$prediction = \begin{cases} 0 \text{ if } last\_year < 2017 \\ 1 \text{ if } last\_year = 2017 \end{cases}$$

This simple heuristic gives an accuracy of 0.8991 on the test set which is 0.0115 percentage points higher than Model 3. This is an interesting result as it suggests that the more complicated linear regression model is unnecessary to make accurate predictions.

3.3 Section 3

In question 1, the restaurants were grouped according to their neighbourhoods. However, this may not be the most appropriate grouping. Some neighbourhoods may have no or few restaurants and some important clusters of restaurants may overlap different neighbourhoods. This section attempts to identify important clusters of restaurants based on their latitude and longitude coordinates.

In order to find these clusters, a Gaussian Mixture Model was used. This was done by implementing the GaussianMixture package provided by scikit-learn. The 'covariance_ type' type parameter was set to 'full' which meant that 'each component has its own general covariance matrix'(sklearn 2019). This algorithm can sometimes converge on a solution that is not globally optimal. To avoid this, the number of initialisations performed was set to 5. Lastly, the number of cluster groups had to be specified.

The model will overfit the data if too many clusters are specified. If too few are specified, the model may miss out on some important clusters. To determine the appropriate number of clusters, the model's BIC value was considered. A lower BIC indicates that the model fits the data better. However, the BIC will always decrease as the number of clusters increases. To avoid overfitting, we look at the gradient of the BIC and select the number of clusters where the gradient levels out. Figure 18 show the BIC gradient for this model. For each number of clusters, the model is trained 5 times (with 5 iterations each time) and the mean BIC is calculated. The gradient at a particular cluster is obtained by subtracting the BIC value for the previous number of clusters. As expected all the Gradient values are negative and it appears as the gradient levels out at 6 clusters. Subsequently, this was chosen as the number of clusters for the model.

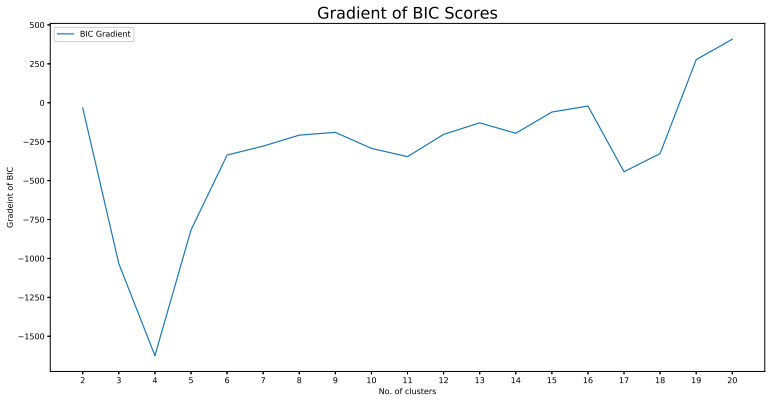

Figure 18: Gradient of BIC for Latitude and Longitude Clusters

Figure 19 visualises the results of the model where each cluster is given a different colour. The very long and narrow yellow cluster appears to be some road in Toronto. The red and grey cluster appear to be quite dense. These may be highly populated/ commercial areas. Lastly, the purple, pink and blue clusters appear to be less dense and more widely spread out. These may be more residential areas or suburbs.

We can gain a better understanding of these clusters if we look at the map of Toronto. Figure 20 shows the restaurant from the yellow cluster in Figure 19 plotted on such a map. Nearly all of these points fall on Young Street. A quick search on tourist websites shows that this is a very popular and some consider it the most famous street in Toronto (Padykula 2019, wikitravel 2019). Taking another look at Figure 20 shows that the red and grey cluster would fall on the areas around the harbour. These do appear to be dense areas. Moving away from these areas, the map seems to become less dense.

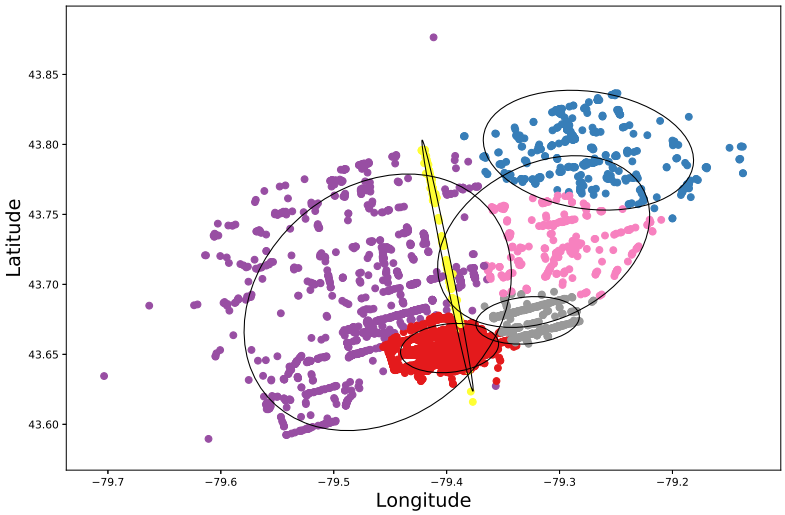

Figure 19: Latitude and Longitude Clusters

Figure 20: Map of Yellow Cluster Group

An attempt was made to fit a Gaussian mixture model to a few other variable pairs. However, none of these yielded any interesting results. For instance, the number of 5 star and 1 star ratings were analysed. It was hypothesised that the model may be able to determine 4 areas. The first being restaurants with a high number of 5 star ratings and a low number of 1 star ratings. These would be uncontroversially good restaurants. Another could be uncontroversially bad restaurants, with a high number of 1 star rating and a low number of 5 star ratings. There could also be a controversial group with high counts for both ratings and a group with simply low counts for both ratings. However, as seen in 21, this is not the case. A group of low counts for both ratings is obtained. This group would likely just simply have a small number of reviews in total. There does not seem to be a clear interpretation for the other clusters.

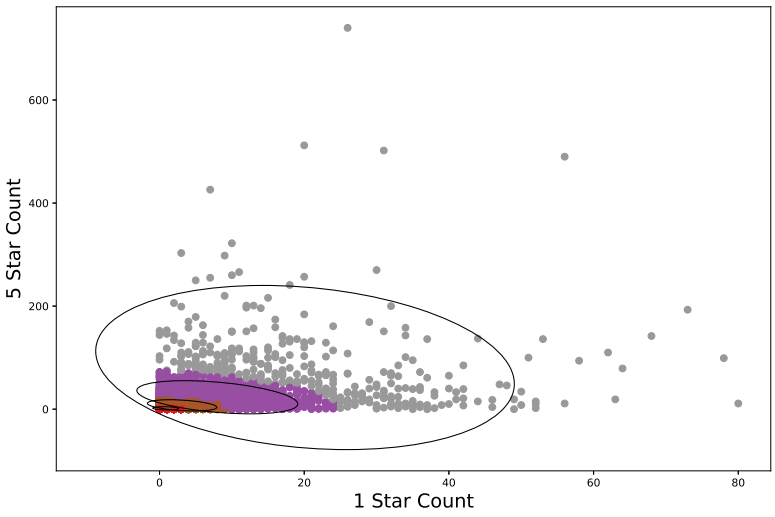

Figure 21: 1 Star and 5 Star clusters

4 Discussion

For the first section, a clear difference between the ratings of Indian restaurants for Scarborough and Etobicoke could not be determined. Additionally, contradictory conclusion seemed to be favoured when the different star ratings (average or individual) were used. Although, these results should be viewed in the light of possible assumption violations. Specifically, that the

ratings have a normal distribution. In the future, it may be necessary to redo the analysis using different assumptions for the distributions of star ratings. Using a hierarchical model based on different assumptions may lead to different results. In other words, it may be possible to show one neighbourhood has statistically better reviews than the other. In terms of the more general analysis, the analysis did show that some restaurants had significantly different ratings. This analysis only used the average star ratings and so it could be done again using the individual star ratings. Like, the above, this may lead to some contradictory results.

For section 2, it was shown that a logistic regression could predict whether a restaurant was closed with 88.76% accuracy. Although, it was shown that this could be outperformed by a simple heuristic. In other words, there is room for improvement. For instance, the 'last_year' variable could be converted to an indicator variable (i.e. 1 if $last\_year = 2017$ and 0 otherwise). This variable is essentially the same as the heuristic used and, if combined with the other significant explanatory variables, it could produce a model with greater accuracy. An alternative Natural language Processing (NLP) approach could also be used.

It was shown that the average length of a text review was a significant explanatory variable and other properties of the text may also have a relationship with the target variable. For instance, n-gram features could be extracted from the text. The rationale behind this is that these textual features may provide important information for why the restaurants are closed. For example, the bi-gram 'poor service' may become common among closed restaurants.

Section three revealed some interesting clusters based on the latitude and longitude of the restaurants. Additionally, by looking at a map of Toronto it could be shown that these clusters correspond to actually features on the map. Restaurants could be labelled with these new groups which could then be used in future models. A few other variables pairs were analysed but no interesting results were found. However, this was not exhaustive and it may still be possible to obtain some other interesting results with this method.

References

json_normalize (2019). Accessed: 2019-04-05.
URL: *https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.io.json.json_normalize.html*

Padykula, J. (2019), 'What to see and do on yonge street in toronto'. Accessed: 2019-04-05.
URL: *https://www.tripsavvy.com/things-to-do-yonge-street-4159756*

pandas (2019). Accessed: 2019-04-05.
URL: *https://pandas.pydata.org*

sklearn (2019), 'sklearn.mixture.gaussianmixture'. Accessed: 2019-04-05.
URL: *https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html*

StatsModels (2019), 'statsmodels.discrete.discrete_model.logit documentation'. Accessed: 2019-04-05.
URL: *https://www.statsmodels.org/devel/generated/statsmodels.discrete.discrete_model.Logit.html*

wikitravel (2019), 'toronto/yonge street - wikitravel'. Accessed: 2019-04-05.
URL: *https://wikitravel.org/en/Toronto/Yonge_Street*

Yelp (2019), 'Yelp open dataset'. Accessed: 2019-04-05.
URL: *https://www.yelp.com/dataset*