**Predict the Number of Tips Needed to Attain one Point Higher than the Highest Rating in the Area**

**Matthew H. Murphy**

**April 29, 2020**

Background

An individual is looking to open a new coffee shop. They need recommendations for a location to setup their new coffee shop and approximately how many Tips they will need to attain one point higher than the highest rating in their area.

Description of the Problem

An individual has money to invest in the startup of a new coffee shop and would like to hire a consultant to use statistical analysis to advise them as to where a good location in Toronto, CA is to start a new coffee shop and use machine learning to tell them how many Tips they will need to attain the goal of one point higher than the highest rating in the area.

The stakeholders in this project are myself and an individual investor that would like to open a new coffee shop in the area of Toronto, CA.

Description of the Data

The data utilized will be from the Foursquare API. This data will be used to determine such things as the number of coffee shops in the area, their price point, which will be categorized as, “Cheap”, “Moderate” and “Expensive”. We’ll also look at several consumer sentiment metrics such as the “Rating” of the coffee shop and the number of, “Likes” the coffee shop received and finally, “Count of Tips” recorded by the API for each coffee shop. This information needed will be considered, a “premium call” and therefore the dataset size will be very small as a result, as we only have a limited amount of, “premium calls” per diem.

The, “Count of Tips” variable will be used as a feature to build a Simple Linear Regression (SLR) model to predict the, “Rating”. Then, the Simple Linear Regression prediction equation will be used to answer the question of, what is the estimated, “Count of Tips” needed to achieve a “Rating” that is one point higher than the highest “Rating” in the dataset?

Exploratory Data Analysis

We were limited to 50 “premium calls” per diem, so as a result we pulled back information on 50 different coffee shops in the Toronto, CA area. Then, it was determined, which of the 50 records were eligible to be used in the model.

For the consideration of data cleaning, if there was any missing data in any of the records returned back from the API, then the record was simply dropped. We also considering imputing missing values, but we felt like there wasn’t enough records to implement an algorithm such as a Maximum Likelihood Estimation (MLE). In exploratory data analysis we found that there were 27 viable records of the 50 returned back from the API that could be used for the model.

We set up a SLR using, “Rating” as the dependent variable and possibly using, “Tips\_Count” and/or, “Likes” as independent variables. Next, we looked at the count, mean, standard deviation, minimum value, 25%, 50%, 75% and the maximum value of the dataset for the, “Rating”, “Likes” and “Tips\_Count” variables.

|  | **Tips\_Count** | **Likes** | **Rating** |
| --- | --- | --- | --- |
| **count** | 27.000000 | 27.000000 | 27.000000 |
| **mean** | 20.444444 | 65.481481 | 7.029630 |
| **std** | 29.111633 | 94.191854 | 0.878292 |
| **min** | 1.000000 | 0.000000 | 5.300000 |
| **25%** | 5.000000 | 9.000000 | 6.300000 |
| **50%** | 13.000000 | 34.000000 | 6.900000 |
| **75%** | 21.000000 | 86.000000 | 7.950000 |
| **max** | 152.000000 | 449.000000 | 8.400000 |

Inferential Statistical Testing (Correlation Matrix)

We ran a Pearson correlation on, “Tips\_Count”, “Likes” and “Rating” and found that, “Tips\_Count’ and “Likes” are highly correlated with one another (i.e. the two variables have a Pearson correlation value of 0.945131). Therefore, we’ll only select one of the two to be in the model to predict “Rating”.

Furthermore, we looked at the correlation between “Rating” and these two variables and determined that, “Rating” and, “Tips\_Count” had the highest correlation value, which was 0.439460, therefore, “Tips\_Count” was selected to be in the model by itself as, “Likes” and, “Rating” only had a correlation of .419733.

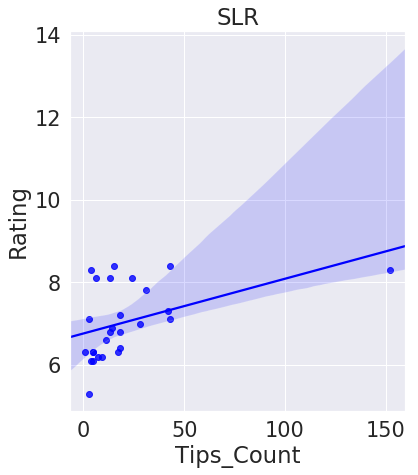
|  | **Tips\_Count** | **Likes** | **Rating** |
| --- | --- | --- | --- |
| **Tips\_Count** | 1.000000 | 0.945131 | 0.439460 |
| **Likes** | 0.945131 | 1.000000 | 0.419733 |
| **Rating** | 0.439460 | 0.419733 | 1.000000 |

Machine Learning Algorithm

Simple Linear Regression was used to train a model that would predict, “Rating” based on the number of, “Tips\_Count” for each coffee shop in the area.

Results

The prediction equation for SLR is .0132584(x) + 6.75856899 and we can see a visual of the prediction line graphed relative to the data points, which were used for training the model:



The max, “Rating” in our dataset is 8.4, since we would like to attain a rating of one above the highest how many, “Tips\_Count” will we need to attain such a rating?

First, one above the highest can be calculated as 8.4 + 1 = 9.4, so now we can plug in 9.4 into our prediction equation and solve algebraically to find out how many, “Tips\_Count” we’ll need to attain a rating of 9.4.

9.4 = .0132584(x) + 6.75856899.

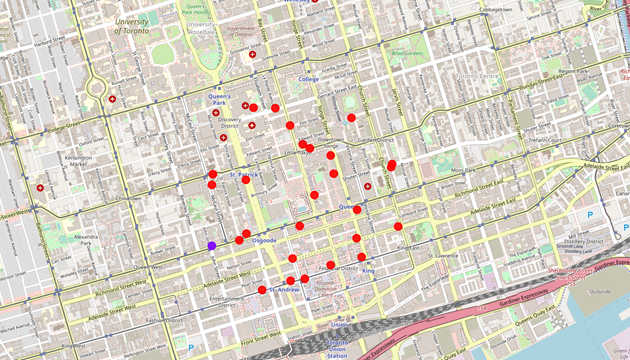
Once we solve the above equation we can determine that we need an x equal to 199.2269814. Or we can say that we need a, “Tips\_Count” value of 199.2269814 to get a, “Rating” of 9.4, according to our SLR model.

Discussion

We also wanted to make a recommendation of where the best location is for the investor to start a new coffee shop. First, we looked at the coffee shops that were classified as “Expensive” and there was only one coffee shop, which was named, “Coffee, Oysters, Champagne”, which seemed to be more fine dining, with coffee, so we omitted this as a possible location. Next, we looked at the coffee shops that were classified as “Moderate” and found eight that met this criteria. Of the eight that met this criteria, we then found the coffee shop with the lowest Rating, which was a rating of 6.3 and noted that coffee shop’s latitude and longitude. We want to be in the “Moderate” category with plenty of opportunity to, “Wow” new customers. As a result, we believe building a coffee shop in the vicinity of latitude 43.65001 and longitude -79.39099 is one of the best places in Toronto, Canada to do so.

Conclusion

We can conclude that based on our Simple Linear regression that we need a Tips\_Count of 199.2269814 to attain a rating of 9.4, which would be one point higher than any coffee shop in the general area of our study, which can be seen below visually. The visual below shows a plot of all of the coffee shops that we used as training data in our Simple linear regression model. The purple point represents the location that we recommend building the new coffee shop in the vicinity of, which had the lowest rating (6.3) in the, “Moderate” price point category. The red points represent the rest of the data points used to train the model.



References

<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

<http://cocl.us/Geospatial_data>

<https://api.foursquare.com>