**Which City and what kind of Restaurant**

**Eric**

**11.22.20**

1. **Introduction/Business Problem**

For this project I decided I wanted to explore the restaurant scene in places I've lived in the past. I decided to explore the cities of Austin, Houston, and San Antonio to get a better glimpse into potential suggestions for investors and entrepreneurs looking to make a sound investment in a restaurant. I was looking for what type of restaurant would be the best and in which city to focus on when scouting for potential property.

The question we attempted to address by conducting this project was to discern how accurately we can predict the amount of "likes" a new restaurant opening in this region can expect to have based on the type of cuisine it will serve and which city in Texas it will open in. For this project I analyzed and modeled the data using machine learning by comparing both linear and logistic regressions to see which method yielded better predictive capabilities after training and testing

1. **Data**

Once this is complete, we finally name the dataframe 'raw\_dataset' as it is the most complete compiled form before needing any processing for analysis via machine learning.

First we retrieved the geographical coordinates of the three cities (Austin, Houston, and San Antonio). We then went ahead and leveraged the Foursquare API in order to obtain the URLs that gave us the raw data in JSON form. Each respective URL was then scraped for the columns: 'name', 'categories', 'latitude', 'longitude', and'id' for each city. The city column will help us when separating where the restaurants are from.

For this project, I decided to focus on those restaurants found within a 1000km radius from the coordinates that were provided by the geolocator. The Foursquare API provides us with more venue categories than we need, and therefore we had to make sure to clean our results by removing non-restaurant rows. Pulling the 'likes' data is necessary for us to make our final decisions. We don't want to be pulling information that will be discarded anyways and is of no valuable for our analysis.

We used the 'id' column in order to pull the 'likes' using the API and append the information into the dataframe. We concluded by naming the dataframe 'raw\_dataset', which we used in the machine learning portion of the project.

## Methodology:[¶](https://dataplatform.cloud.ibm.com/data/jupyter2/runtimeenv2/v1/wdpx/service/notebook/conda2py3733305d4735c344a6b5e5b6e8d3bd3daa/dsxjpy/9P3PVo4H7CC_CLKL9i9Hxw:fwWpykSLV_8n_IqRrZoozGVp-P6yaGs2vxEAdli-2eAlfkJxukr8TyTDBkF4OR-q-6kUpNA/container/notebooks/ad634d57-c183-49e6-a892-5d175e8996aa?api=v2&project=33305d47-35c3-44a6-b5e5-b6e8d3bd3daa#Methodology:)

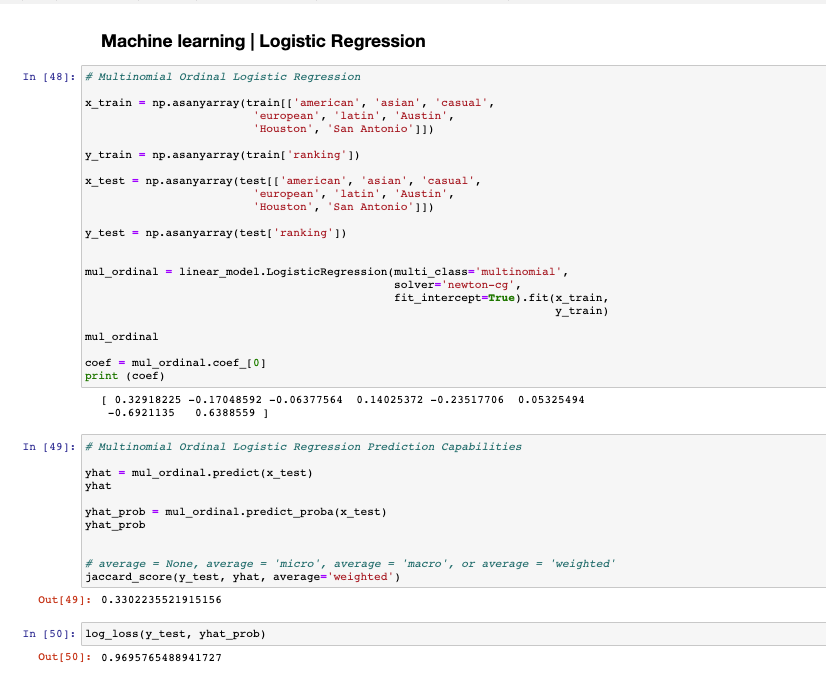
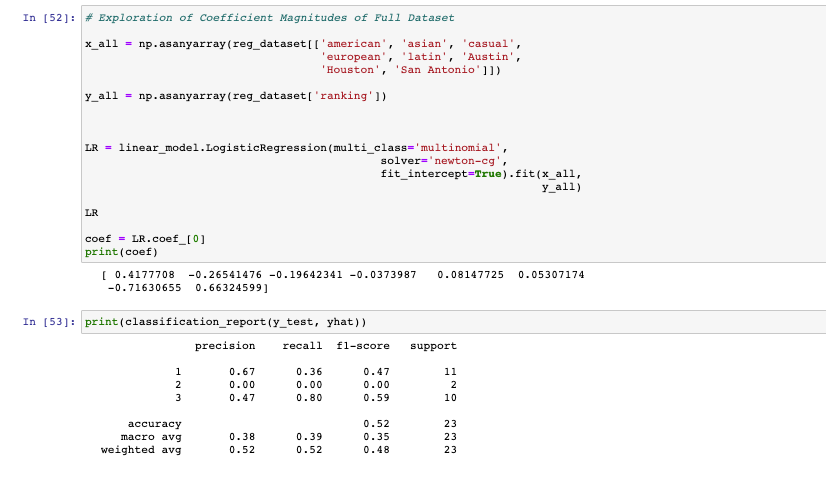
Both linear and logistic regression were used to train and test the data. Linear regression was used to predict the number of 'likes' a new restaurant in this region will acquire. Sci-Kit Learn was employed for this stage.

Logistic regression was used as the classification method. Since, we used binning when classifying by number of 'likes', we essentially made use of multinomial logistic regression to perform our analysis. Although the ranges are discrete categories, they can be considered ordinal in nature. The logistic regression will need to be specified as being both multinomial and ordinal. The Sci-Kit Learn package is perfect for this.

1. **Results**

A linear regression model was trained on a random subsample of 80% and then the other 20% was used for testing purposes. In order to evaluate if the model is reasonable, the residual sum of squares and variance score were both calculated (16023.25, 0.01). The variance score is quite low, which means that is not a good way of modeling our data. Therefore, we moved on to logistic regression for our analysis.

The multinomial ordinal logistic regression model was also trained on a random subsample of 80% and then tested on the remaining 20%. The jaccard score and log-loss were both calculated (33.02% and .9695 respectively). Although the prediction is not promosing, a jaccard score of 33% is somewhat reasonable. The classification report is included in the analysis.

Given the modestly accurate ability of this mode, we have the ability to run the model on the complete dataset. The coefficients we got show that opening a restaurant in Austin, or serving cuisine that is asian, or casual, are negatively associated with 'likes'. **  
  
**

1. **Discussion**

The first thing to note is that given the data, logistic regression presents a better fit for the data over linear regression. Using logistic regression we were able to obtain a Jaccard Score of 33.02%, which although not perfect, is more reasonable than the low variance score obtained from the linear regression. As stated before, please note that for the purposes of this project, we are assumming that likes are a good proxy for how well a new restaurant will do in terms of brand, image and by extension how well the restaurant will perform business-wise. Whether or not these assumptions hold up in a real-life scenario is up for discussion, but this project does contain limitations in scope due to the amount of data that can be fetched from the FourSquare API.

As such, to obtain insights into this data, we can proceed with breaking down the results of the logistic regression model. The results showed that the precision score for classifying whether the new restaurant would fall into classes 1, 2, or 3 (highest, medium, lowest) were 40%, 0%, and 50%. Therefore, the model is better at predicting if a restaurant will fall into the best or worst percentile of likes. This is good as we are mostly concerned with whether the restuarant will perform well or not so the high accuracy of predictions for the two extremum is a welcome feature. This allows us to fairly accurately predict the general performance of the business opportunity. Different binning methods for the classes were attempted, but the use of 3 bins by far yielded the best Jaccard Score.

Additionally, not only are we attempting to predict the general business performance but also pull insights to inform on business strategy. In this case strategy insight can be gleamed from the coefficient values from running the logistic regression on the full dataset. As such, we can see that opening a restaurant in Austin, or serving cuisine that is asian or casual in nature, are associated Positively with "likes." This suggests that the business opportunity should be opening a restaurant in either Austin, with a cuisine that is American, Casual, Asian, or Latin in nature would be the best approach for maximizing likes.

1. **Conclusion**

In conclusion, after analyzing restaurant 'likes' in Texas from the 300 restaurants, we can conclude that the approach to best take when looking to maximize business performance (as measured by 'likes') is to open a restaurant that is either American, Casual, Asian, or Latin and that opening the venue in Austin would be the best approach. Additionally, the predictive capabilities of the logistic regression prediction model proved to be the most accurate for classifying whether a restaurant fell in either the best or worst classes when the data was binned into their 3 respective classes.