Mini-Project 3

1 Introduction

Nowadays, social media plays an important role in determining the success of businesses. This especially holds very well for restaurants and food establishments. The reviews about these establishments posted online determines consumer attraction or repulsion. Yelp dataset contains each business's number of reviews and information about the business spread over multiple states. This information can be used for predicting tastes of states and thus finding parameters that determines a restaurant's success.

2 Problem Statement

The objective is to determine a restaurant's success based on the cuisines (likely american, european, asian, etc), the number of attributes that corresponds to the restaurant (likely Happy Hour, Alcohol, Has TV, Good for kids). The success measure is determined by number of review counts of the business and its star rating. I'm fitting these features into Naive Bayes Model or Logistic Regression Model and compare the performance of both the models and choose the better model.

3 Data Preparation & Observations

Yelp dataset is "structured dataset" in json format. Business collection contains details about businesses that are using yelp. This collection includes information like state, neighborhood, its location, average star rating, number of reviews, etc. I processed this data by using a python script, which parses through this .json file and segregate the businesses based on states. The data had higher number of restaurants in the following states - AZ, NV, QC, NC and PA and businesses on other states were very less.

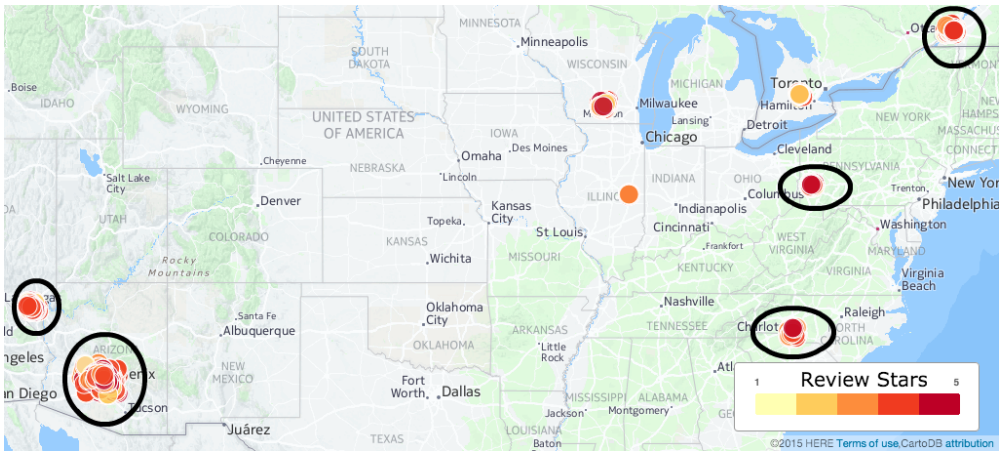

Figure 1: Distribution of businesses from the dataset [1]

Review collection contains reviews from users for each business. Again, using python script, I was able to collect the reviews from unsuccessful and successful businesses. By using text processing techniques, such as removing punctuations, removing stop words, etc, I was able to observe the most frequently used words by users of these two category of business.

Figure 2: Distribution of review stars for business [3]

From figure 2, we see a nice normal distribution of review star count. Average review star for the businesses is around 3.5. This seems to be about right, as people's opinion is dependent on various factors and their experience with the restaurant on that day, they are tending to give more generic ratings.

Number of business which are "Restaurants" and "Shopping" are at the highest, from figure 3. This seems quite obvious as most of the businesses which heavily depend on online reviews for their success. Next to this is the "Food" category but this can be combined along with restaurants.

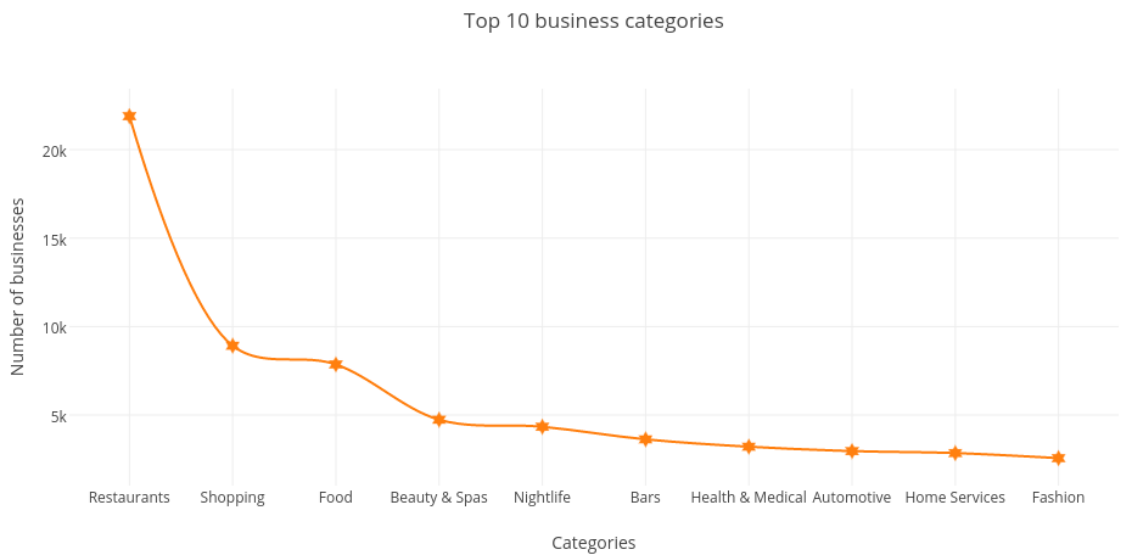

Figure 3: Distribution of business categories [3]

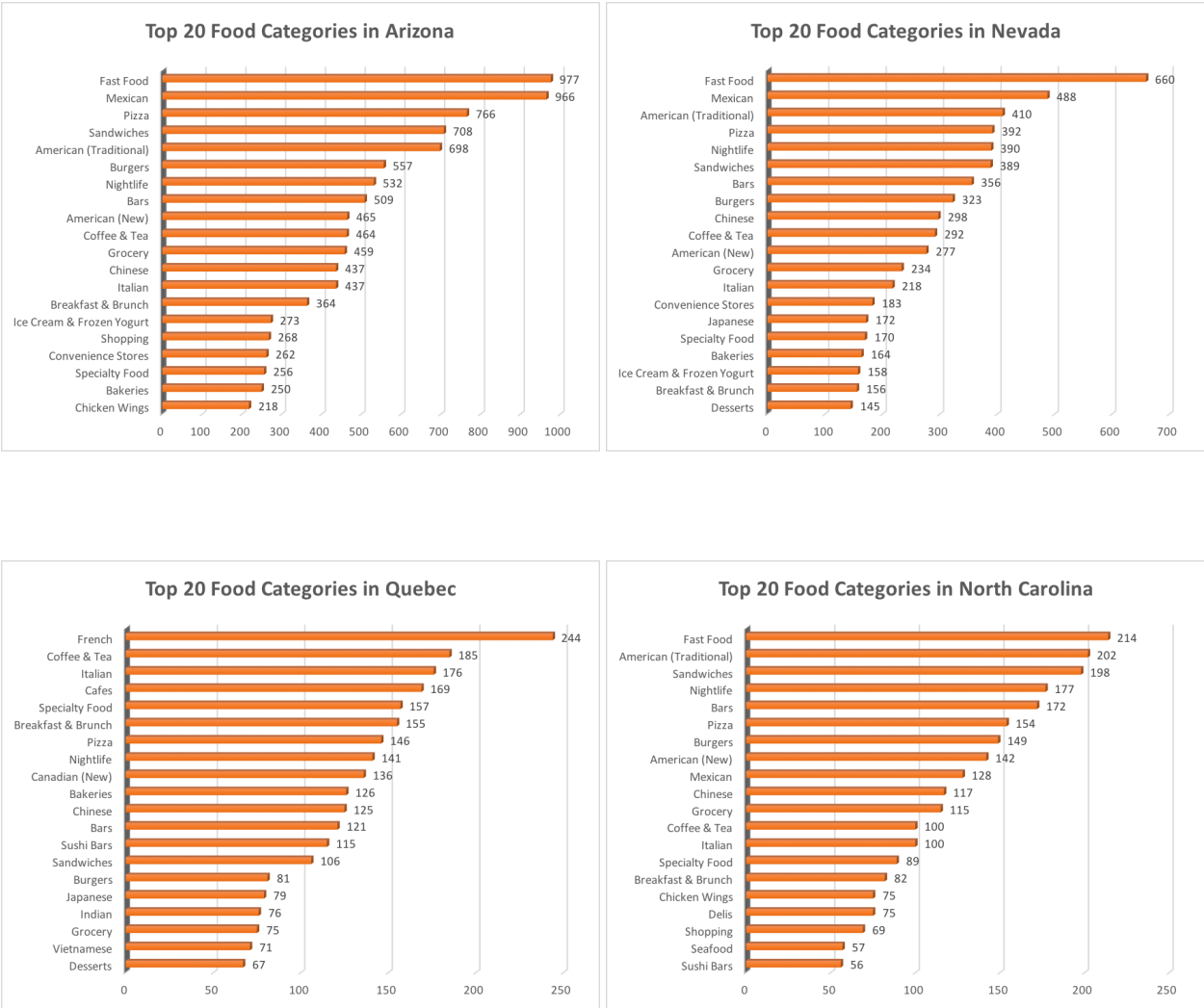

Figure 5: Top 20 categories in restaurant of various states [3]

Based on 5, I'm categorizing the cuisine of each restaurant to one of the following - "Fast-Food", "American", "European", "Latin American", "Asian", "Others". Also, we can see that state, Quebec, in Canada, has lot of "French" restaurants and the American states have lots of "Fast Food" restaurants.

4 Feature Selection

Based on exploratory data analysis and few sample runs of models, following features seems to produce better predictions of the success of the business.

1. **Price Range** - Integer value : Higher the value higher the price range
2. **Good for Kids** - Boolean value : Good for kids directly proportional to attracting families
3. **Good for Groups** - Boolean value : Good for groups attract corporates to hang out after work
4. **Happy Hour** - Boolean value : This is good strategy to attract students

5. **Alcohol** - Boolean value : Depends on user's preference actually. Included to consider varied opinions

6. **Has TV** - Boolean value : Again depends on user. Few love to watch TV while dining others not so much

7. **Success** - Boolean value : True on successful business; False on not so successful business

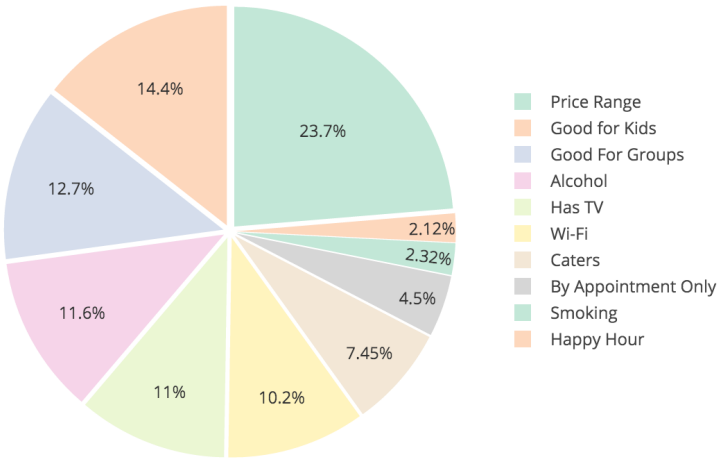

Figure 6: Distribution of restaurants attributes [3]

5 Prediction Model

By the nature of the values of the features, mostly "dichotomous", my first approach was to apply Logistic Regression. As expected, Logistic Regression performed well and achieved about 88% accuracy.

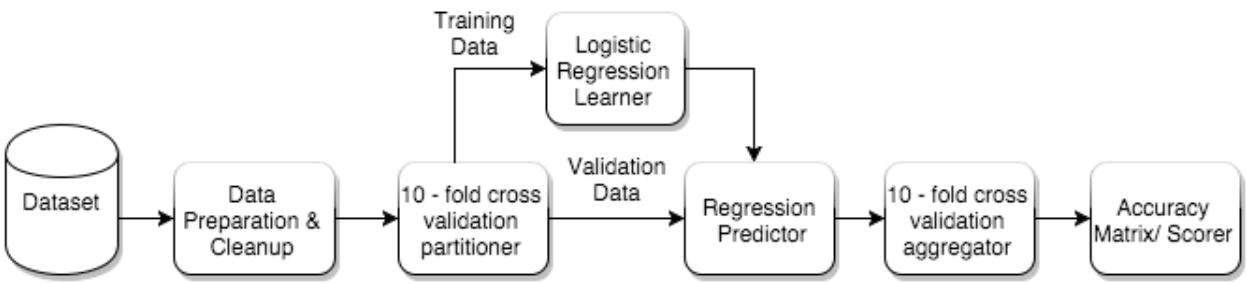

Figure 7: Logistic Regression Model Implementation[5]

6 Results

Following were the result obtained with 10-fold cross validation from Logistic Regression. The model's accuracy was 88%. The table below shows the distribution of error during each run.

Fold Num	Error in %	Size of test set	Error count
Fold 1	6.060606	264	16
Fold 2	10.26615	263	27
Fold 3	12.87878	264	34
Fold 4	11.40684	263	30
Fold 5	12.12121	264	32
Fold 6	13.30798	263	35
Fold 7	12.12121	264	32
Fold 8	12.92775	263	34
Fold 9	13.25757	264	35
Fold 10	12.16730	263	32
Average Error %	**11.651**	**Average Accuracy**	**88.349 ≈ 88%**

Below is the confusion matrix obtained which contains TrueNegative, FalseNegative, TruePositive and FalsePositive, Recall and Precision values.

Confusion Matrix [6]

TruePositive	FalsePositive	TrueNegative	FalseNegative	Recall	Precision
1972	50	356	257	0.88470166	0.975272008
356	257	1972	50	0.876847291	0.580750408

Quick visualization of reviews from unsuccessful business and successful business.

(a) Word cloud of unsuccessful restaurant reviews (b) Word cloud of successful restaurant reviews

Figure 8: Word cloud from reviews [3]

From the above visualization, "food" is predominant in both the cases. In unsuccessful businesses, it might have been used with negative connotations and vice versa. Similarly, the word "service" also pops out in both the clouds which again corresponds to "bad services" in case of unsuccessful restaurants and "good services" in case of successful restaurants. Words like "love", "delicious", "amazing", "friendly" pops out well in successful restaurants. Conversely, words like "time", "wait", "dont", "didnt", "minute", "bad", "horrible" pops well in unsuccessful restaurants.

7 Conclusion

Based on exploratory analysis and visualization, tastes of states varies from state to state. While predominantly Americans stick to "fast food", citizens from states like Nevada and Arizona choose "Mexican" over "Traditional American" as their next best choice. Whereas in North Carolina, "Traditional American" is preferred better. This is can highly contributed to the fact that these states are in close proximity to Mexico. While states like Quebec, prefer "French" cuisine over "American" or "Fast Foods".

As the preferences of people/consumers varies over states, it is necessary that the restaurants too have to adapt to the expectations of the people of those states. This model, with about 90% accuracy, will prove to be a good means for anticipating to the expectations of the state and thus improving the success rates of business.

References

[1] CartoDB `https://cartodb.com/`

[2] Yelp Challenge DataSet `http://www.yelp.com/dataset_challenge`

[3] Plot.ly - Plot charts online `https://plot.ly`

[4] PostgreSQL DB `http://www.postgresql.org/`

[5] Flow Charts Online `https://www.draw.io`

[6] Knime `https://www.knime.org/`