Restaurant Success Prediction

Final Project Report

Name: Trinath Reddy Inturi

SU ID: 476506643

Contents

Introduction

- In current society restaurants acts as got to place for all kinds of social gatherings. Owning a restaurant might be an exciting experience but it involves lot of effort like managing finance, maintaining food quality, and marketing your restaurant.
- The success of the restaurant depends on how well it is satisfying the needs of the customers. It is also one of the important factors to attract new customers.
- Now a days we can easily find good restaurants near us through review sites like google, yelp and Facebook where people share their experiences.
- Therefore, by utilizing reviews like these I would be using classification techniques to predict if a restaurant would be successful or not.
-

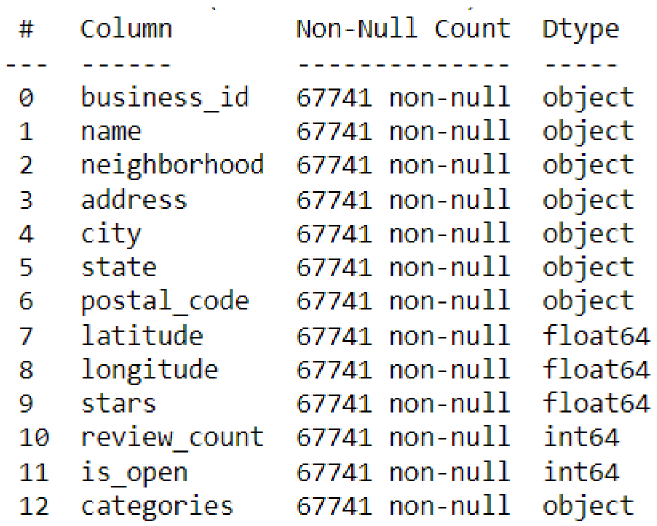
Prior Work

- Based on my research I have found that sentiment analysis is performed on yelp data to find out if the restaurant reviews are positive or negative
- Sentiment analysis was also used to see what users like about restaurants and what they don't
- But using these reviews I would like to find If a restaurant would be successful or not.
- Some of the reference I have found are as below:
  - Text Mining and Sentiment Analysis for Yelp Reviews of A Burger Chain | by Elva Xiao | Towards Data Science
  - Sentiment Analysis of the Yelp Reviews Data | Kaggle

Algorithm/ Prediction Method

Step 1: Data Preprocessing

- To implement Restaurant Success Prediction, First I've acquired datasets from Yelp
- For this project I am using 2 datasets "Business" and "Reviews" where "Business" contains the data about all business as below:



```
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   business_id   67741 non-null  object
 1   name          67741 non-null  object
 2   neighborhood  67741 non-null  object
 3   address       67741 non-null  object
 4   city          67741 non-null  object
 5   state         67741 non-null  object
 6   postal_code   67741 non-null  object
 7   latitude      67741 non-null  float64
 8   longitude     67741 non-null  float64
 9   stars         67741 non-null  float64
 10  review_count  67741 non-null  int64
 11  is_open       67741 non-null  int64
 12  categories    67741 non-null  object
```

- "Reviews" dataset contains reviews given by users for a business which stored with property "business_id" as reference
- Initially Business Data set has lot of unwanted data and some features which needed to be removed.
- As Yelp has lot of business which are open and closed. There are also lot of unwanted business which are not related to business therefore we need to remove all the business which are closed, and which are not restaurants.
- After cleaning the dataset now, we need to combine the reviews with the cleaned business dataset and remove all the null values in as missing values can affect performance and accuracy

Step 2: Text Sentiment Analysis

- We can determine the sentiment of the review based on the stars given for the review.
- For each review we need to remove special characters to improve the accuracy and performance of the text processing
- We can find polarity and subjectivity of each review, where polarity would indicate the emotions of the review and subjectivity quantifies the amount of opinion and information contained in the review.
- Then we need to apply lemmatization to all the cleaned words which would convert all the words into proper vocabulary which can later be used to train the model

- After lemmatizing the words, we need vectorize all the reviews into a spare matrix which would index every word in the review built with the vocabulary of all the words in the reviews
- Now we need to create a column called success class which take either 1 or 0 which represent that if a review would contribute to the success of a restaurant.

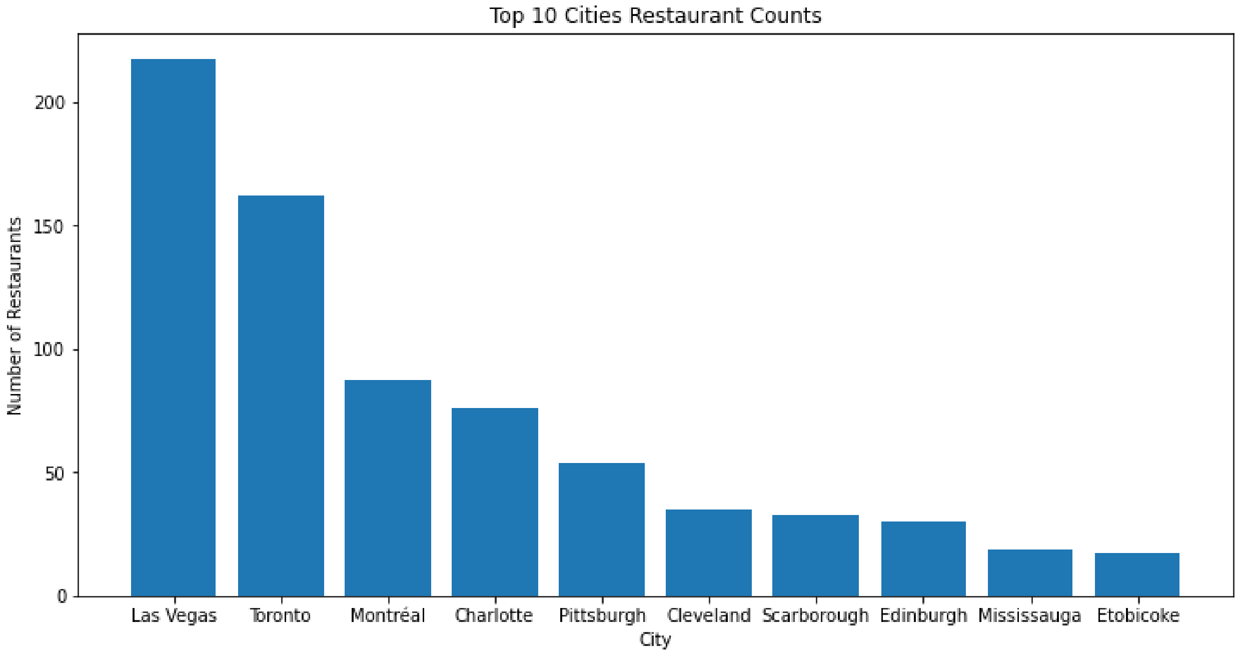
Step 3: Prediction

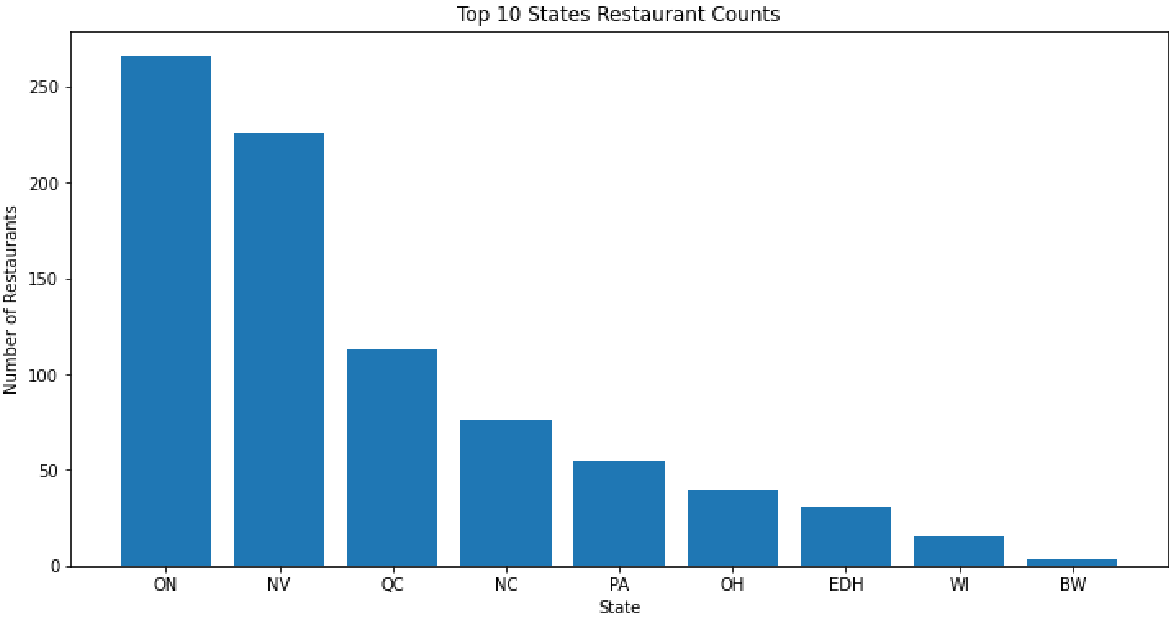
- After obtaining the sparse matrix X and the success column y we need to split the data into test (80%) and train (20%)
- Out of Naïve Bayes, Logistic Regression, Decision Trees and Random Forest Classifier Logistic Regression was more accurate which was used to train the model
- I have passed the training data of X and y and then used the model to predict the success class of the restaurants and the probability of success.
- After obtaining the predicted data for each review I have grouped the data which provided with better understanding of success rate of the restaurant.
- More explanation on results and evaluation can be found in next section

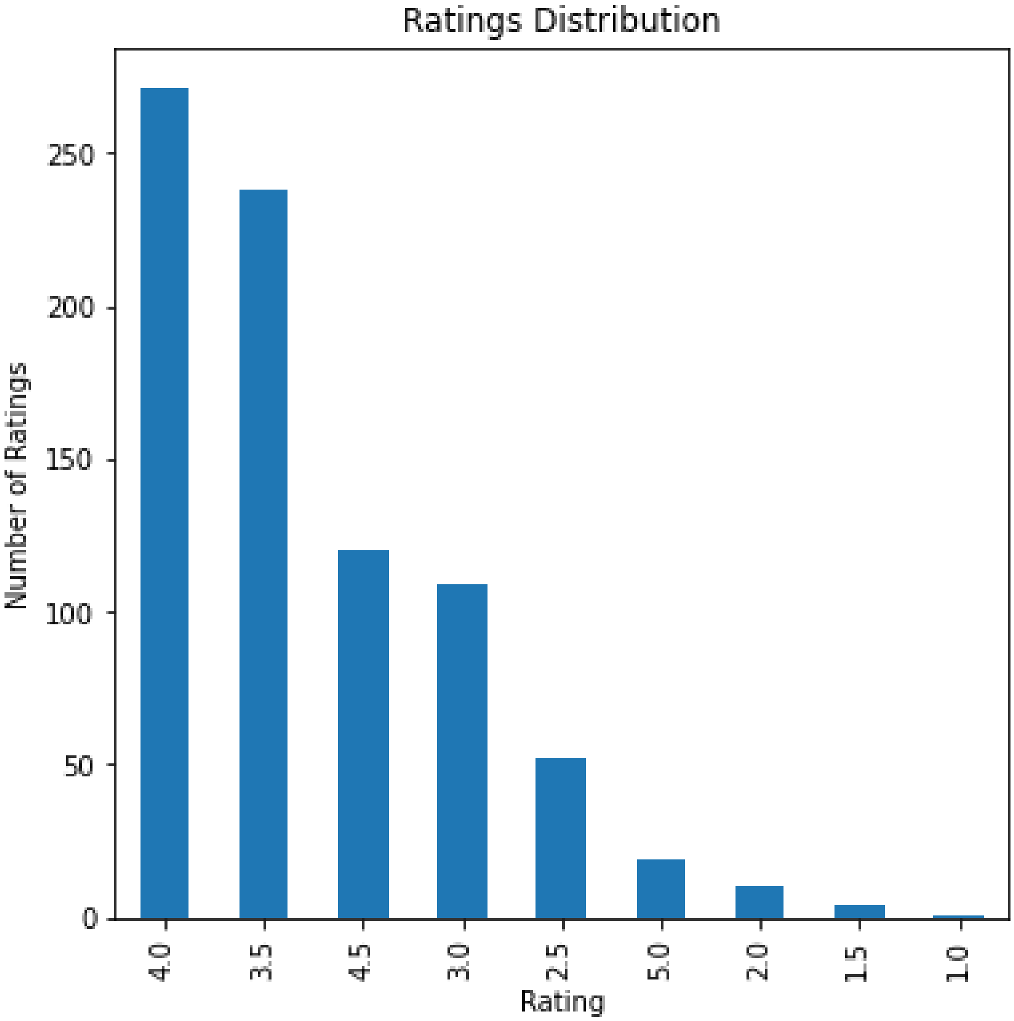
Results and Findings

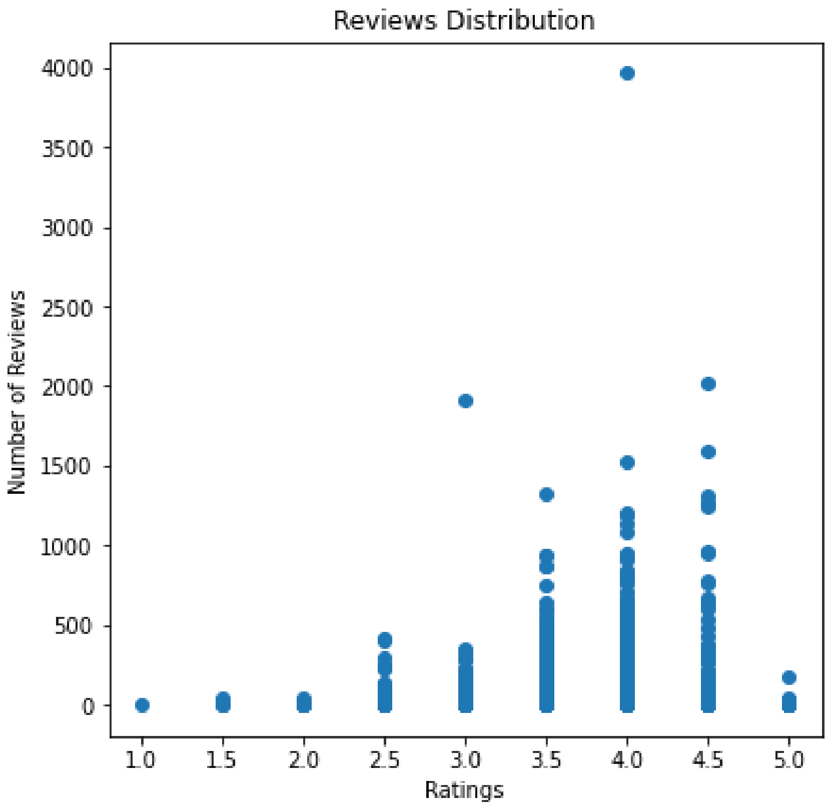
Analysis of Data:

- After cleaning the raw data obtained from yelp, these are some the highlights of the data sets used

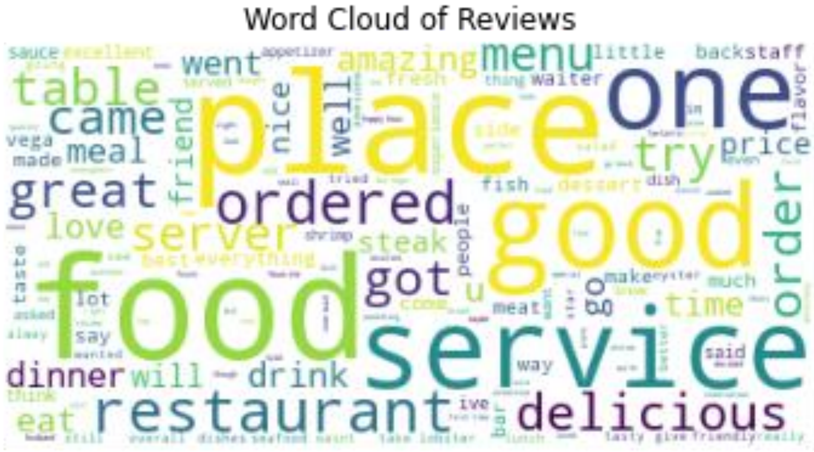



Top 10 States Restaurant Counts
Number of Restaurants
0
50
100
150
200
250
ON
NV
QC
NC
PA
OH
EDH
WI
BW
State


Ratings Distribution
Number of Ratings
0
50
100
150
200
250
4.0
3.5
4.5
3.0
2.5
5.0
2.0
1.5
1.0
Rating

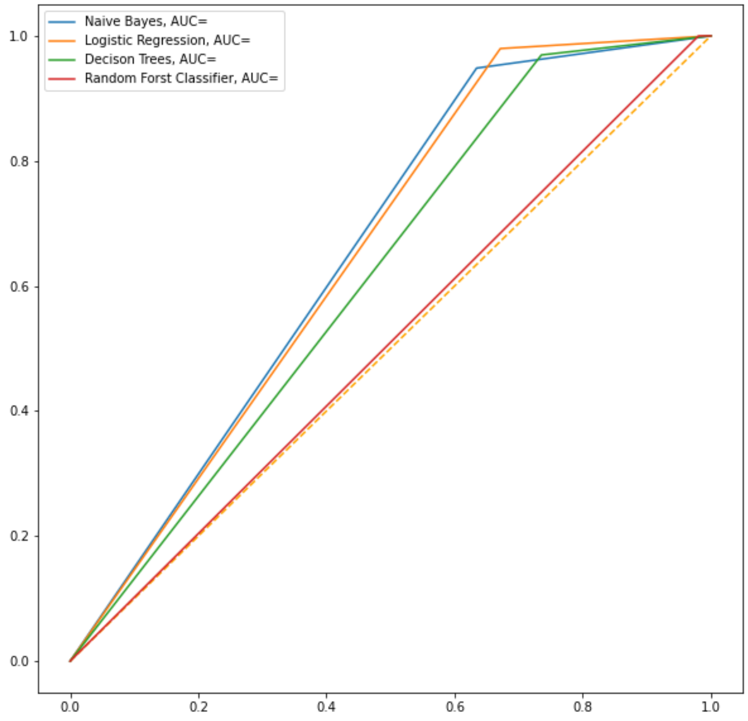


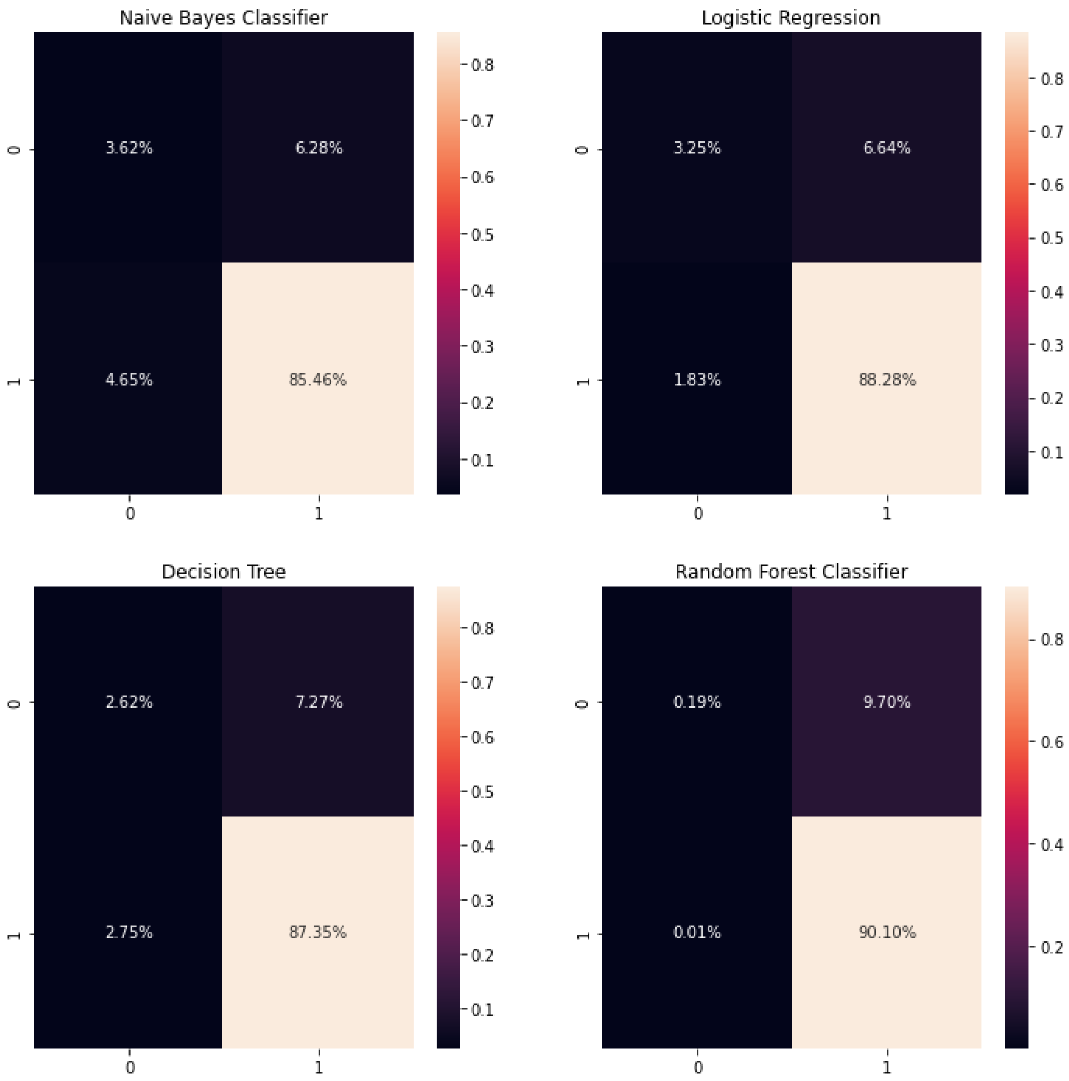
- In the above images we can observe which cities and states have most restaurants and what are the total stars given for all the restaurants and their distribution (box plot)
- This helps us to understand the data and its features like locality and where do most of the ratings lie.
- This will help us to understand which data we are dealing with and what might be the expected outcome of the prediction
- After cleaning the reviews, I have created a word cloud based on the reviews which provides a quick glance of what words are used in reviews by users to describe their experience in restaurants



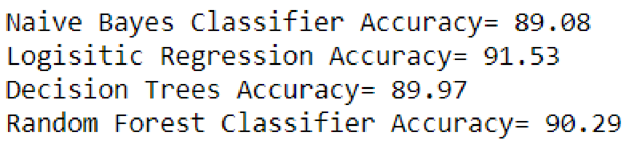
Evaluation Methods

- To predict the success of the classifier I have added features to the raw data such that I has each review vectorized and a success class representing the Boolean value.
- I have used this to train the data, but as we know there are lot of classification methods available which can serve the purpose.
- Therefore, I've picked four classification models of which I've trained the model and tested them to check which works better with my data.
- I have used Naïve Bayes Classifier as it works best with discrete features. It creates a tag for each review and tries to predict the tag with highest probability. Through this we will be able to identify if the given review would be positive or negative.
- Logistic regression vectorizes all the reviews and divides the data into positive and negative reviews which will enable the model to predict. I will not be making any assumptions about the distributions of the features.
- Decision trees are easy to read and interpret. They simply try to learn the classification data to make predictions
- I have also test Random Forest Classifier which words well with the high dimensional data. It is parallelizable therefore training speed is more optimized. It is also robust to outliers.
- Out of the 4 classifiers used below are the results of the performance of classifiers for the given input.
- Below is the ROC curve representing the diagnostic ability of the binary classifier.





- Above represent the confusion matrix of the 4 classifiers
- Below we can see the accuracy scores of the classifiers

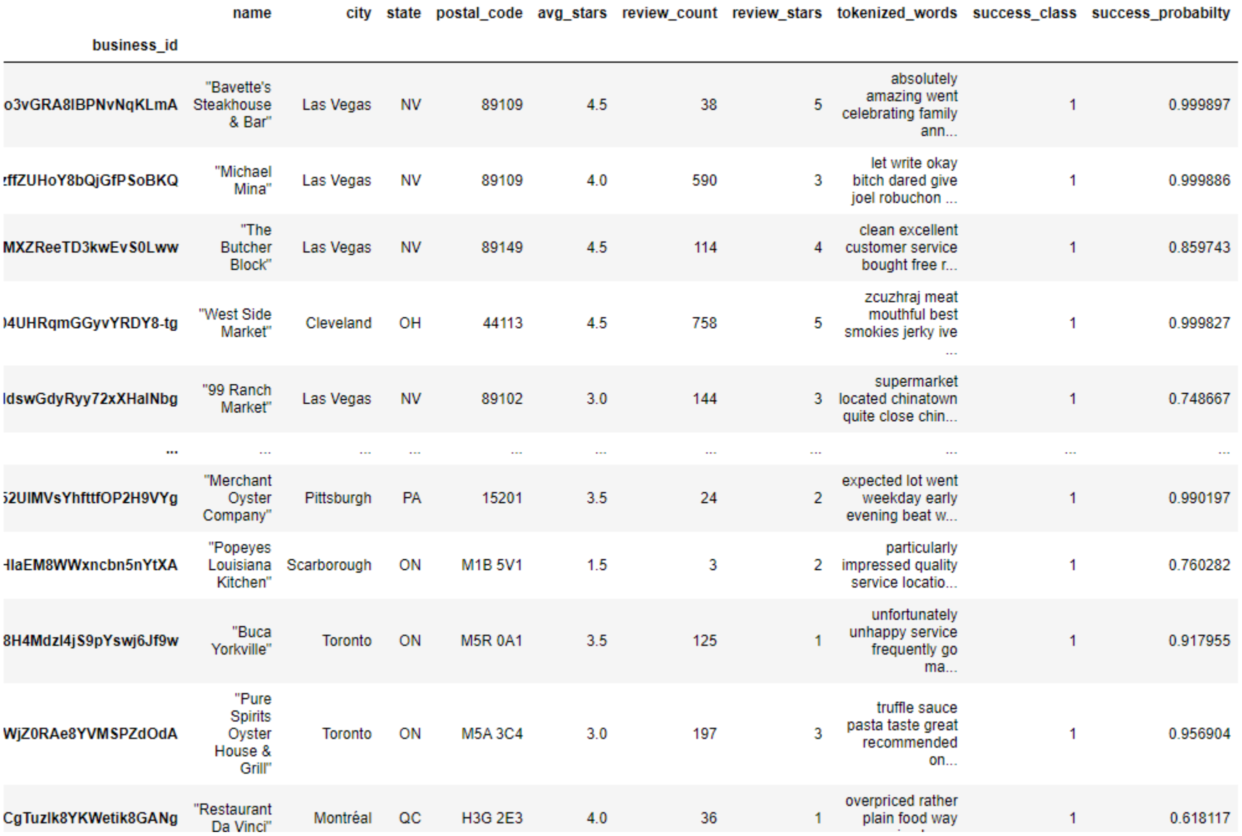


```
Naive Bayes Classifier Accuracy= 89.08
Logisitic Regression Accuracy= 91.53
Decision Trees Accuracy= 89.97
Random Forest Classifier Accuracy= 90.29
```

- After considering all the above results using Logistic Regression for the prediction would Ideal

Output

- After applying logistic regression to the test data below are the final outputs



	name	city	state	postal_code	avg_stars	review_count	review_stars	tokenized_words	success_class	success_probabilty
business_id										
o3vGRA8IBPNvNqKLmA	"Bavette's Steakhouse & Bar"	Las Vegas	NV	89109	4.5	38	5	absolutely amazing went celebrating family ann...	1	0.999897
:ffZUHoY8bQjGfPSoBKQ	"Michael Mina"	Las Vegas	NV	89109	4.0	590	3	let write okay bitch dared give joel robuchon ...	1	0.999886
MXZReeTD3kwEvS0Lww	"The Butcher Block"	Las Vegas	NV	89149	4.5	114	4	clean excellent customer service bought free r...	1	0.859743
)4UHRqmGGyvYRDY8-tg	"West Side Market"	Cleveland	OH	44113	4.5	758	5	zcuzhraj meat mouthful best smokies jerky ive ...	1	0.999827
IdswGdyRyy72xXHalNbg	"99 Ranch Market"	Las Vegas	NV	89102	3.0	144	3	supermarket located chinatown quite close chin...	1	0.748667
...	...	...	...	...	...	...	...	...	...	...
52UIMVsYhfttfOP2H9VYg	"Merchant Oyster Company"	Pittsburgh	PA	15201	3.5	24	2	expected lot went weekday early evening beat w...	1	0.990197
HlaEM8WWxncbn5nYtXA	"Popeyes Louisiana Kitchen"	Scarborough	ON	M1B 5V1	1.5	3	2	particularly impressed quality service locatio...	1	0.760282
8H4Mdzl4jS9pYswj6Jf9w	"Buca Yorkville"	Toronto	ON	M5R 0A1	3.5	125	1	unfortunately unhappy service frequently go ma...	1	0.917955
WjZ0RAe8YVMSPZdOdA	"Pure Spirits Oyster House & Grill"	Toronto	ON	M5A 3C4	3.0	197	3	truffle sauce pasta taste great recommended on...	1	0.956904
CgTuzIk8YKWetik8GANg	"Restaurant Da Vinci"	Montréal	QC	H3G 2E3	4.0	36	1	overpriced rather plain food way	1	0.618117

- These are predictions of the test data made. The complete result can be found in "output.csv" file
- In the test data set the model predicted 717 restaurants would be successful and 57 restaurants would be unsuccessful

```
1    717
0     57
Name: success_class, dtype: int64
```

- More detailed view can be found in Jupyter Notebook File Submitted