Yelp Restaurant Success

Machine Learning Project

Maaz Qureshi

**Introduction**

U.S. restaurant industry sales will reach a record high of $863 billion this year, up 3.6% over last year, across more than 1 million businesses. The industry is continuously growing year after year, and Yelp is one of the biggest platforms of fueling the industry’s growth.

Yelp was founded in 2004 and their main focus is to help connect people with great local businesses. Yelp is an online directory for discovering local businesses ranging from bars, restaurants, gas stations, and all the way to spas or hairdressers. The listings can be filtered by business type, location, price range, and many more. My focus for the project is restaurant oriented data from Yelp.

**Business Objective**

Due to limited information on restaurant sustainability, my objective is to build a machine learning model that will predict restaurant success and closure. This will provide lenders or investors with crucial information which would help them make important business decisions when deciding to invest or lend to different restaurants. In order to achieve this, I needed to locate a dataset that would help us identify different factors that would help display accurate data which would then help these investors make valuable business decisions regarding investing or lending. Since Yelp is one of the largest platforms for the restaurant business, I knew I could rely on them for sufficient data to achieve my goal.

**Dataset Information**

The dataset used was one found on Kaggle provided by Yelp.

* The dataset is a subset of Yelp’s businesses, reviews, and user data
* The dataset contains 5,200,000 user reviews, information on 174,000 businesses, and spans 11 metropolitan areas in 4 countries
* The dataset contains 7 csv files: business, business attributes, business hours, check ins, review, tip, and user

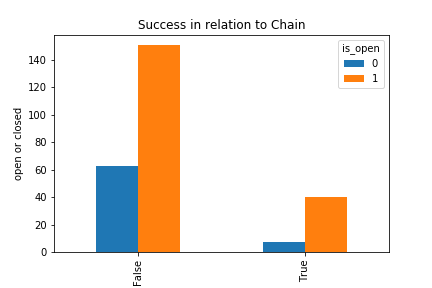
**Data Pre-Processing**

Since the dataset was large and included various types of restaurants in multiple countries, I decided that it was best to focus on a specific area to provide accurate information based on the location of the restaurants. The first step taken was loading the dataset into the Pandas dataframe. After loading the dataset, I filtered the data down to the province of Ontario and furthermore to Toronto. Next, I filtered the dataset to restaurants only as Yelp also has several other types of establishments within their dataset. The filtered dataset contained information about restaurants in Toronto. I then matched up the current Yelp Data with Yelp Data that was last updated two years ago. I compared the restaurants that were open then and closed now and updated the data I would be using to reflect that. This would essentially show the change over two years. I added two categories to the Yelp data. I then figured out which restaurants are a part of a chain using Excel functions and put that information into its own column. I also added a column that assigned a number to each category of restaurant.

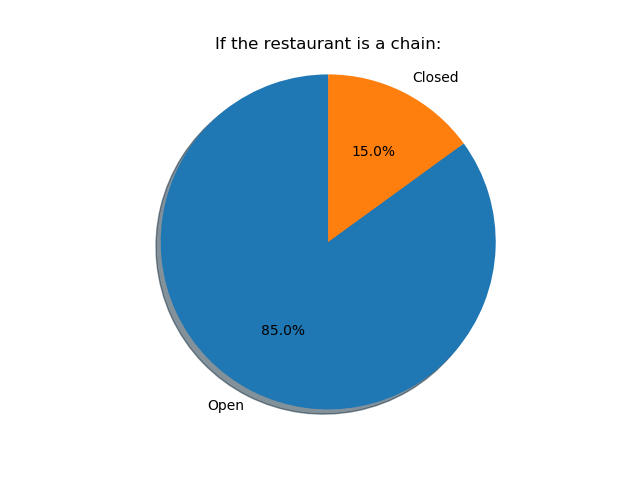
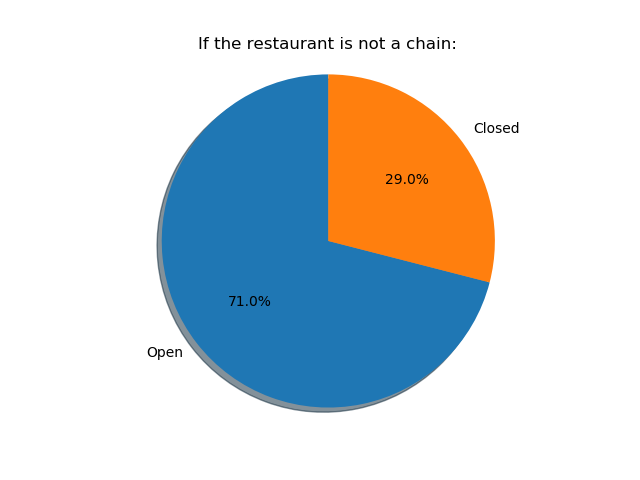
The factors included in the data are: 'business\_id', 'name', 'neighborhood', 'address', 'city', 'state', 'postal\_code', 'latitude', 'longitude', 'stars', 'review\_count', 'is\_open', 'categories', 'is\_chain', and 'category\_number'. To use the data in the machine learning models, I had to narrow it down even further to only numerical data. For the models I used 'latitude', 'longitude', 'stars', 'review\_count', 'is\_open', 'is\_chain', and 'category\_number'. The data associated with ‘is\_open’ and ‘is\_chain’ were made binary, with ‘is\_open’ being used as my target. As I tweaked my models further, I found that they all actually performed better when not including the ‘stars’ ratings. When looking at the factors individually, they could not provide accurate predictions on success. However when factors were combined, the data was able to give impactful information about the sustainability of restaurants in Toronto.

**Visualizations**

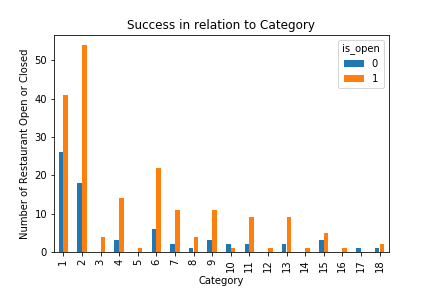
I used several visualizations in order to help illustrate some potential relationships between my data and what I describe as success. Generally, success would be measured in profit for a business. However, I do not have access to that information so I measured success in terms of staying open over a period of 2 years. This is the time frame I have decided to use to see if a restaurant can be sustained, thus making it worth an investment or loan.

****

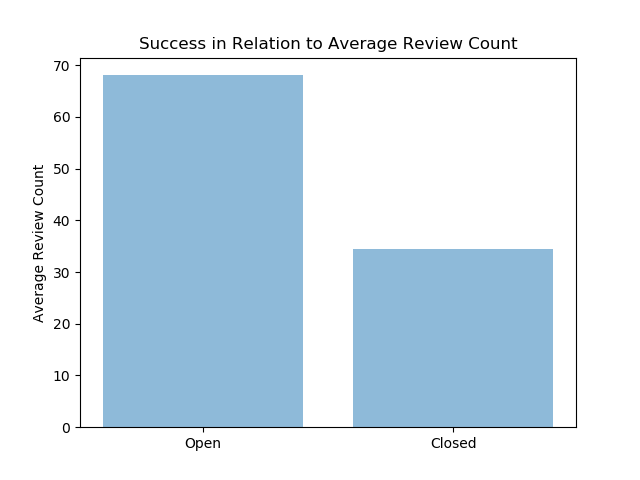
I generated a bar graph to show the relationship between a restaurant belonging to a chain and staying open. The chart data is hard to interpret at first because the sheer volume of restaurants that are not a part of a chain makes it hard to read. Therefore, I put the data into a pie chart to better illustrate ratios.

****

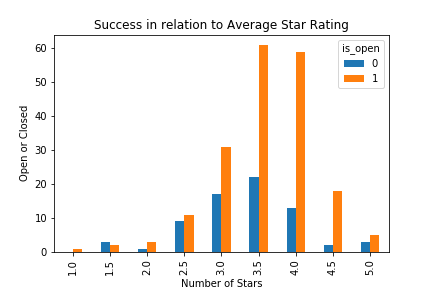
The pie charts show that there is a significantly higher ratio of restaurants staying open that are part of a chain, as opposed to those that are not.

****

I generated a bar graph to compare restaurant type (category) to the number of restaurants open or closed. This provides information about the level of success of each restaurant type. The data shows that certain types of restaurants do experience more success than others.

****

One way I theorized tracking relative number of patrons is through review count. The more people come, the more likely a review is to be written. To illustrate the relationship between review count and success as I define it, I found the average review count for open restaurants and closed restaurants. As the graph shows, the open restaurants have a significantly higher average review count.

****

I generated a bar graph to compare the star rating for restaurants to whether or not it is open or closed. This data is interesting because it does not follow an intuitive pattern. Rather than a positive linear relationship between star rating and success, it seems that most success occurs at 4.0 stars, with 5.0 stars actually having a poor ratio of open to closed. This manifests itself in my machine learning models, which I tweaked to not include the star rating and found the predictive capabilities of the models actually became more accurate.

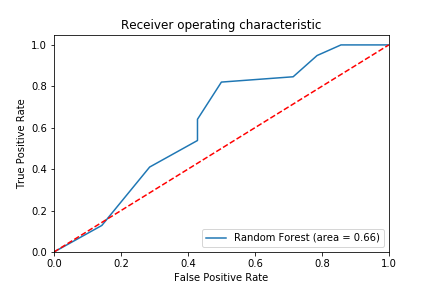
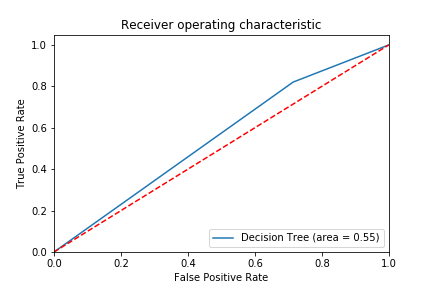
**Machine Learning Modeling**

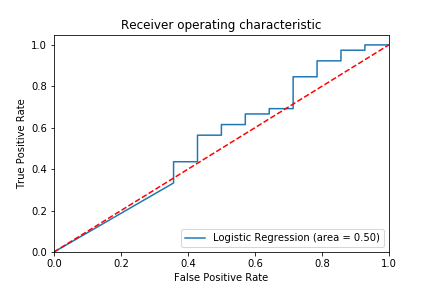
The three machine learning models used were random forest, logistic regression, and decision tree. Decision tree is a flowchart-like structure where each internal node represents an attribute. The branches represent the decision rule and each leaf node represents the outcome of the test. Decision tree focuses on classification. Random forest builds upon the logic of decision tree. It is essentially many decision trees working together. Logistic Regression is used to predict the probability of a categorical dependent variable. Since my project is predicting the probability on if a restaurant would be successful or not, I thought logistic regression would be a good model to use when measuring the level of success.

**Results**

Since decision tree and linear regression had the lowest scores out of the three models that I used, I decided to go with random forest model as it scored the best overall and is the one I have decided to be my model of choice. This is interesting because the logic between decision tree and random forest is similar. Random forest builds on top of decision tree because it makes many decision trees and produces an aggregate result of all of them. The random forest had an accuracy score of .736, precision score of .820, recall score of .820, F1 score of .820, and AUC for ROC of .66. Compared to decision tree and logistic regression, random forest scored highermost categories listed above. While logistic regression has an exceptional recall score of 1.0, its AUC for its ROC is very poor at .50. When looking at the ROC curve, a value of 1 is given to excellent ROC curve, and a value of .5 is considered a bad ROC curve. This, in addition to its other lower scores, eliminated it from contention. The decision tree also had a value around .5, scoring a .55, which indicates it is not a good model to use. The full breakdown of scores is given below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | **Precision** | **Recall** | **F1** | **Confusion Matrix** | **AUC for ROC** |
| **Random Forest** | .736 | .820 | .820 | .820 | [[ 7 7]  [ 7 32]] | .66 |
| **Logistic Regression** | .735 | .735 | 1.0 | .847 | [[ 0 14]  [ 0 39]] | .50 |
| **Decision Tree** | .679 | .761 | .820 | .790 | [[ 4 10]  [7 32]] | .55 |

****

****

The figures above show the ROC curves for logistic regression, decision tree, and random forest.

**Packages Used / Tools**

from sklearn.model\_selection

from sklearn.metrics

from sklearn.linear\_model import Linear Regression

from sklearn.tree import DecisionTreeClassifier

from sklearn.ensemble import RandomForestClassifier

Import pandas

from matplotlib.ticker import FuncFormatter

import matplotlib.pyplot as plt

import numpy as np

**Future Improvements**

According to the results random forest is the best model when predicting restaurant success and closure, but some future works that might help improve on this analysis could be:

* Using Census data to track the income and family demographics of a neighborhood and using that data in my analysis
* Coming up with more numerical values for data such as “neighborhood”
* Testing over a larger scale of area in order to get more data
* Testing other models as well to see if another model could work better with this dataset

**Work Cited**

Brownlee, Jason. “Logistic Regression for Machine Learning.” *Machine Learning Mastery*, 12 Aug. 2019, machinelearningmastery.com/logistic-regression-for-machine-learning/.

Gupta, Prashant. “Decision Trees in Machine Learning.” *Medium*, Towards Data Science, 12 Nov. 2017, towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052.

Yelp, Inc. “Yelp Dataset.” *Kaggle*, 6 Feb. 2018, www.kaggle.com/yelp-dataset/yelp-dataset/version/6.

Yiu, Tony. “Understanding Random Forest.” *Medium*, Towards Data Science, 14 Aug. 2019, towardsdatascience.com/understanding-random-forest-58381e0602d2.