**Predicting the Success of Opening a Restaurant**

Hunter Sparrow

April 26th, 2021

**1. Introduction**

**1.1 Background**

A business owner is looking to invest in a new restaurant venture. The owner has selected California to be the next location. The owner wants to know what restaurant type is the most popular in the given areas (Oakland, San Diego, Emeryville). Also, to maximize the chances of success by choosing a type of cuisine that has a lot of ‘likes’ associated with it.

**1.2 Problem**

Analyze the data to determine what type of restaurants are in the area. The owner wants to specify in:

* European
* Asian
* American
* Latino
* Casual

To solve the problem, we need to analyze the data for the given areas and determine which type of restaurant is the most popular and has the highest accumulative *likes*. We will run this data through a Machine learning model to predict the type and area that should yield the highest success rate to open a new restaurant.

**2. Data acquisition and cleaning**

**2.1 Data**

Using Foursquare and the raw data scraped from each url. We are only focusing on:

* Name
* Category
* Latitude
* Longitude
* Id/City

Using another API we will get the **likes** data and applying that to the machine learning models.

Combined we will have an overview of all the information from each city and see the type of restaurant has the most likes.

# 3. Methodology

The owner wants to predict the 'likes' of a certain type of restaurant. The more likes it can predict the higher chance for success.

**Libraries needed:**

* Pandas
* Numpy
* Matplotib
* Seaborn
* Sci-kit learn

**Machine learning models:**

* Logistic regression: great for predicting classification. In this case **yes** like or **no** like.

**4. Data Cleaning**

Extracted the data from Foursquare for the given areas:

Graphical user interface, text, application

Description automatically generated

Web scrapped from the URLs: Venues names, category and location

A picture containing timeline

Description automatically generated

We used the foursquare API to extract the venues, location, and name for the surrounding areas. We decided to only focus on the type of restaurant that the owner wanted to open.

* European
* Asian
* American
* Latino
* Casual

We filtered the data into new columns with the given information to analyze.

Text

Description automatically generated

A picture containing calendar

Description automatically generated

**5. Analysis**

We will analyze what type of restaurant is the most popular in all three city locations:

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

**Analysis from Data**

If we wanted to open a restaurant in **Oakland**:

* Asian cuisine has the highest amount by double or triple the amount than other types.

If owner wanted to open in **Emeryville**:

* European cuisine has the highest amount.

If owner wanted to open in **San Diego**:

* American, European cuisine are the top two types.

**5.1 Number of restaurants per city**

Chart, bar chart

Description automatically generated

Oakland has the highest concentration of restaurant types the owner is looking for.

**5.2 Number of likes per restaurant type**

Chart, scatter chart

Description automatically generated

Even though one type of restaurant has more location, it does not mean it has the most likes. Let's investigate which type of restaurant in each city has the greatest number of likes.

**5.3 Likes per City and Restaurant type**

Let us explore the type of restaurant and the number of likes associated with each type of restaurant.

Oakland:

Chart, scatter chart

Description automatically generated

In Oakland, Asian food has the highest likes with over 350. European food ranks at the bottom with under 50 likes.

Emeryville:

Chart, scatter chart

Description automatically generated

In Emeryville, American cuisine restaurant has the highest likes with over 300. Casual has the lowest of around 50.

San Diego:

Chart, scatter chart

Description automatically generated

In San Diego, American with 500+ likes, casual 500+ likes and Latino 490+ likes types are the highest. Asian ranks the lowest with 150+ likes.

**6. Machine Learning Predictive Model**

We want to run the data through ha logistic regression model to predict based on the number of restaurants, and number of likes for each type of restaurant. This model will predict what kind of restaurant will have the highest success.

**6.1 Preprocessing Model**

Convert the data to numeric to run through the model.

A picture containing table

Description automatically generated

After converting it we can run it through the model to test:

A picture containing white

Description automatically generated

**7. Results**

From our first initial analysis we of Emeryville we knew Asian type restaurants ranked low in both location and number of likes. The coefficients we got show that opening a restaurant in Emeryville, or serving cuisine that is Asian, or casual, are negatively associated with 'likes'. This is a fairly accurate prediction.

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 26%. Although the prediction is not promising, a Jaccard score of 26% is somewhat reasonable. The classification report is included in the analysis.

Table

Description automatically generated

**8. Conclusion**

After analyzing restaurant 'likes' in California from the 300 restaurants, we can conclude that:

**3 Best type of restaurants to open**

* European
* Latino
* American

**Ranking of 3 cities**

* Oakland
* San Diego
* Emeryville

**Data-Driven decision**

* The owner would start looking into opening a European restaurant in Oakland.