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:

```
address1 = 'Oakland, California'

geolocator = Nominatim(user_agent="foursquare_agent")
location1 = geolocator.geocode(address1)
latitude1 = location1.latitude
longitude1 = location1.longitude
print('The geograpical coordinate of {} are {}, {}.'.format(address1, latitude1, longitude1))

address2 = 'Emeryville, California'

geolocator = Nominatim(user_agent="foursquare_agent")
location2 = geolocator.geocode(address2)
latitude2 = location2.latitude
longitude2 = location2.longitude
print('The geograpical coordinate of {} are {}, {}.'.format(address2, latitude2, longitude2))

address3 = 'San Diego, California'

geolocator = Nominatim(user_agent="foursquare_agent")
location3 = geolocator.geocode(address3)
latitude3 = location3.latitude
longitude3 = location3.longitude
print('The geograpical coordinate of {} are {}, {}.'.format(address3, latitude3, longitude3))
```

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

```
# filter columns
filtered_columns1 = ['venue.name', 'venue.categories', 'venue.location.lat',
                     'venue.location.lng', 'venue.id']
nearby_venues1 = nearby_venues1.loc[:, filtered_columns1]

# filter the category for each row
nearby_venues1['venue.categories'] = nearby_venues1.apply(get_category_type, axis=1)

# clean columns
nearby_venues1.columns = [col.split(".")[-1] for col in nearby_venues1.columns]

# SECOND CITY

venues2 = results2['response']['groups'][0]['items']
nearby_venues2 = pd.json_normalize(venues2) # flatten JSON

# filter columns
filtered_columns2 = ['venue.name', 'venue.categories', 'venue.location.lat',
                     'venue.location.lng', 'venue.id']
nearby_venues2 = nearby_venues2.loc[:, filtered_columns2]

# filter the category for each row
nearby_venues2['venue.categories'] = nearby_venues2.apply(get_category_type, axis=1)

# clean columns
nearby_venues2.columns = [col.split(".")[-1] for col in nearby_venues2.columns]

# THIRD CITY

venues3 = results3['response']['groups'][0]['items']
nearby_venues3 = pd.json_normalize(venues3) # flatten JSON
```

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

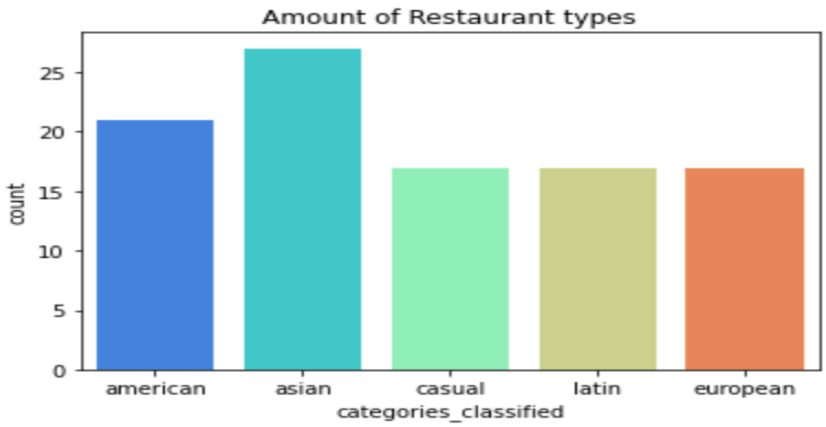
We filtered the data into new columns with the given information to analyze.

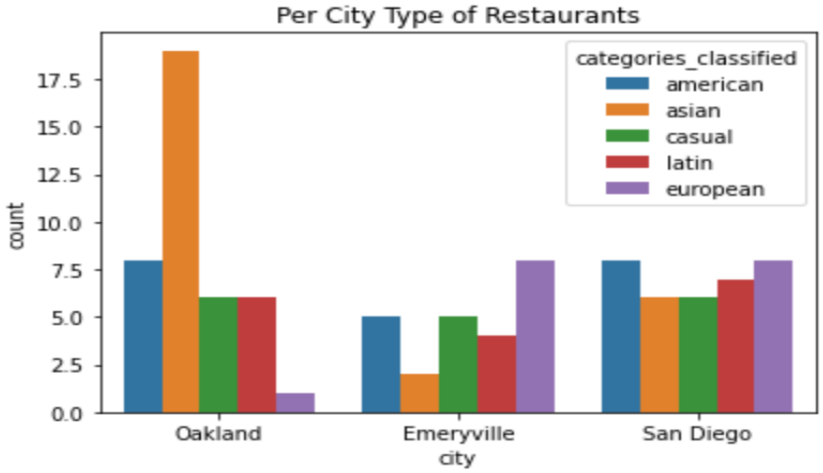
```
array(['Vegetarian / Vegan Restaurant', 'Japanese Restaurant',
       'Bagel Shop', 'Vietnamese Restaurant', 'Mexican Restaurant',
       'Seafood Restaurant', 'Chinese Restaurant', 'Caribbean Restaurant'
       'Kitchen Supply Store', 'Brazilian Restaurant', 'Sushi Restaurant'
       'Sausage Shop', 'Taco Place', 'Tapas Restaurant',
       'Cambodian Restaurant', 'Burger Joint', 'Sandwich Place',
       'American Restaurant', 'Hotpot Restaurant',
       'New American Restaurant', 'Thai Restaurant', 'Indian Restaurant',
       'Dumpling Restaurant', 'Dim Sum Restaurant', 'Breakfast Spot',
       'Falafel Restaurant', 'Ramen Restaurant', 'Diner', 'Pizza Place',
       'Mediterranean Restaurant', 'Scandinavian Restaurant',
       'Southern / Soul Food Restaurant', 'Filipino Restaurant',
       'Asian Restaurant', 'Food Truck', 'Wings Joint', 'Burrito Place',
       'Fast Food Restaurant', "Men's Store", 'Italian Restaurant',
       'Theme Restaurant', 'Hot Dog Joint', 'Turkish Restaurant',
       'Gastropub', 'Empanada Restaurant'], dtype=object)
```

```
# we can group some cuisines together to make a better categorical variable

european = ['Mediterranean Restaurant', 'Scandinavian Restaurant', 'Pizza Place',
       'French Restaurant', 'Falafel Restaurant', 'Italian Restaurant',
       'Turkish Restaurant']

latin = ['Mexican Restaurant', 'Taco Place', 'Brazilian Restaurant',
         'Burrito Place']

asian = ['Japanese Restaurant', 'Vietnamese Restaurant', 'Chinese Restaurant',
         'Hot Dog Joint', 'Hotpot Restaurant', 'Indian Restaurant',
         'Thai Restaurant', 'Dumpling Restaurant', 'Dim Sum Restaurant',
         'Asian Restaurant', 'Filipino Restaurant', 'Sushi Restaurant',
         'Ramen Restaurant']

american = ['Vegetarian / Vegan Restaurant', 'Seafood Restaurant', 'Caribbean Restaurant',
           'Burger Joint', 'American Restaurant', 'New American Restaurant',
            'Southern / Soul Food Restaurant', 'Diner']

casual = ['Bagel Shop', 'Sandwich Place', 'Fried Chicken Joint',
          'Breakfast Spot', 'Wings Joint', 'Fast Food Restaurant',
          'Theme Restaurant']
```

5. Analysis

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





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.

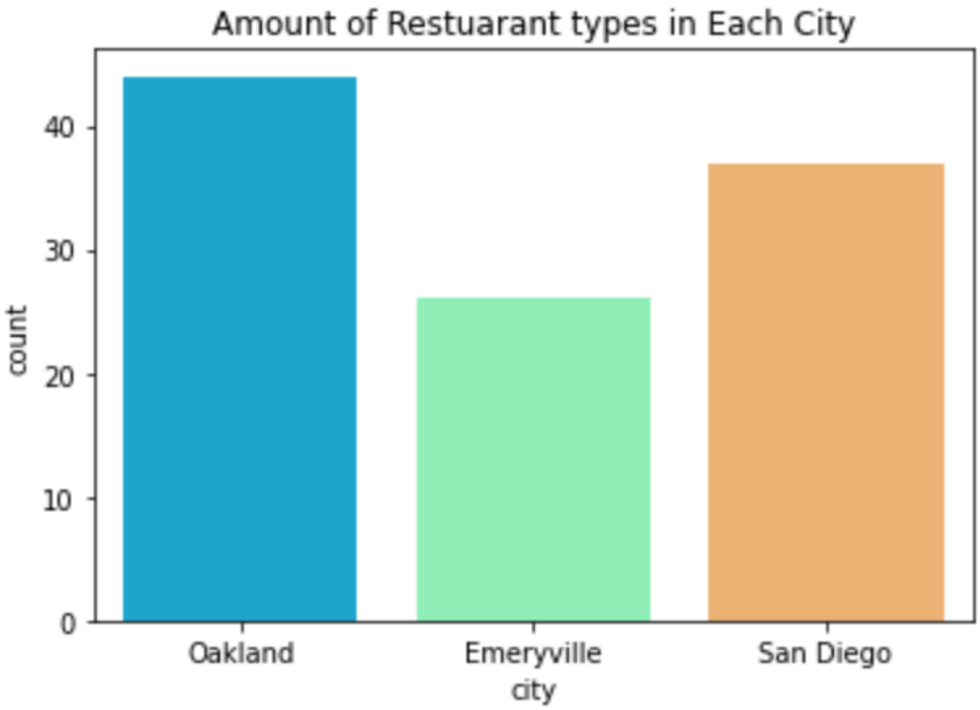
If owner wanted to open in **Emeryville**:

- European cuisine has the highest amount.

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

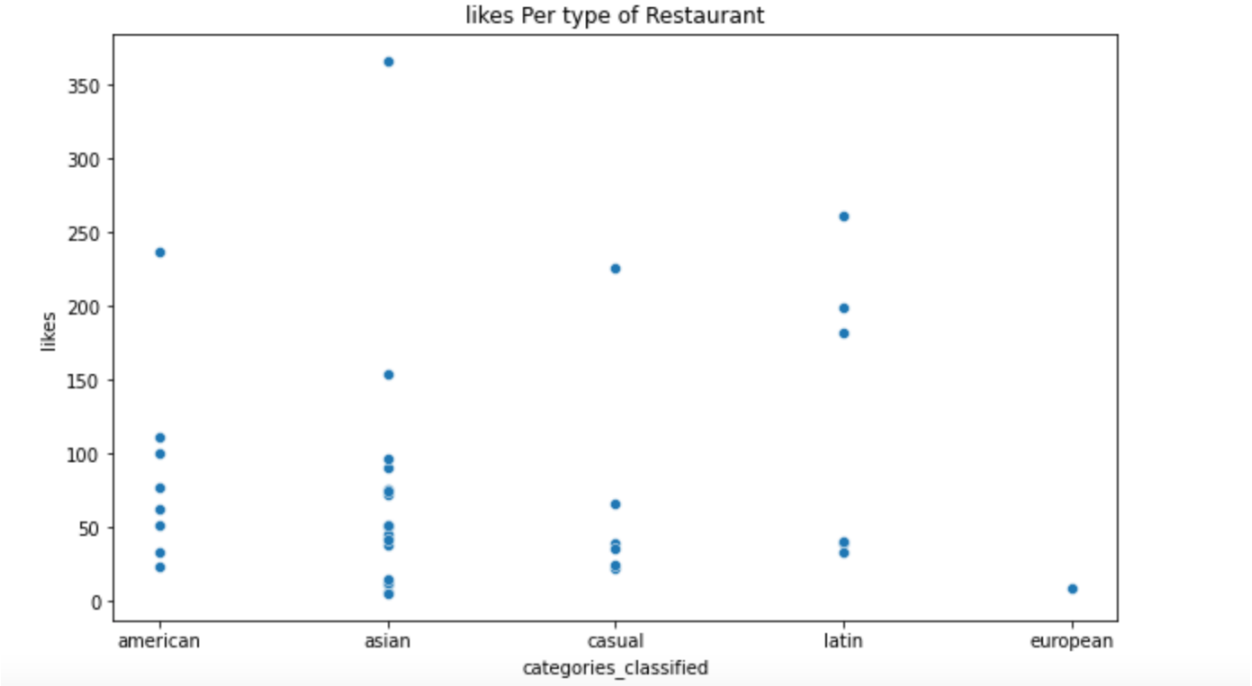
- American, European cuisine are the top two types.

5.1 Number of restaurants per city



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

5.2 Number of likes per restaurant type

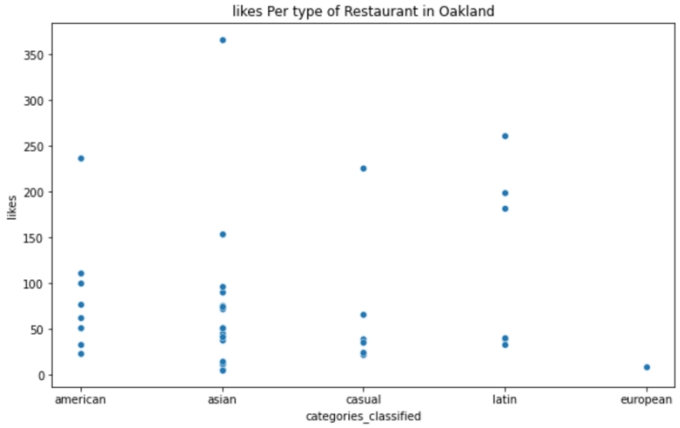


Even though one type of restaurant has more location, it does not mean it has the most likes. Let's investigate which type of restaurant in each city has the greatest number of likes.

5.3 Likes per City and Restaurant type

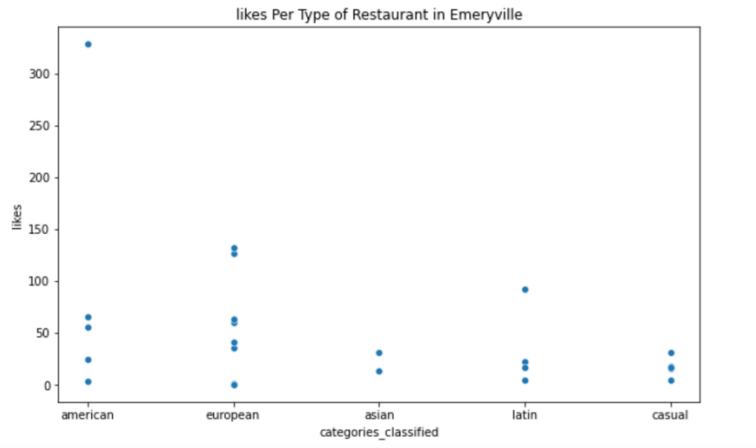
Let us explore the type of restaurant and the number of likes associated with each type of restaurant.

Oakland:



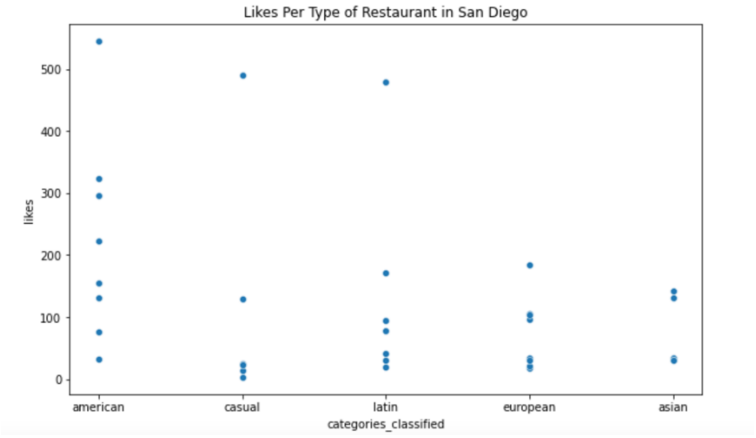
In Oakland, Asian food has the highest likes with over 350. European food ranks at the bottom with under 50 likes.

Emeryville:



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

San Diego:



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.

```
def rankings(df):

    if df['likes'] <= 60:
        return 3

    elif df['likes'] <= 100:
        return 2

    elif df['likes'] > 100:
        return 1
```

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

	name	american	asian	casual	european	latin	Emeryville	Oakland	San Diego	ranking	likes
2	Golden Lotus Vegetarian Restaurant	1	0	0	0	0	0	1	0	2	77
7	Abura-Ya	0	1	0	0	0	0	1	0	1	154
11	Beauty's Bagel Shop	0	0	1	0	0	0	1	0	3	22
12	Tay Ho Restaurant & Bar	0	1	0	0	0	0	1	0	2	72
14	Nature Vegetarian Restaurant	1	0	0	0	0	0	1	0	3	33

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.

	precision	recall	f1-score	support
1	0.00	0.00	0.00	7
2	0.00	0.00	0.00	3
3	0.47	1.00	0.64	8
accuracy			0.44	18
macro avg	0.16	0.33	0.21	18
weighted avg	0.21	0.44	0.28	18

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.