**Texas Restaurant "Likes" Prediction Using Foursquare API and Machine Learning**

**Hafsah Ahmed**

Capstone Project

IBM Data Science Professional Certificate

# Introduction

An important aspect of marketing for a modern restaurant or any business for that matter is social media where the number of "likes" determines a company's public image, reputation and success (in terms of customer satisfaction).

For a business owner planning to open a new restaurant (or expanding an existing one) in a new city, knowing the restaurant type, cuisine and location with the potential to do well both physically and in social media ahead of time can be a game changer. This is because it could solve the problem of uncertainty surrounding performance when breaking into new markets.

In this analysis we will solve this uncertainty by leveraging data gathered from Four Square’s API, specifically "likes" data of different restaurants, their locations and category of cuisine.

The problem: How accurately can we predict the amount of "likes" a new restaurant opening in a certain location can expect to have based on the type of cuisine and where it will open.

For the purpose of this analysis we will focus on three of the most popular and heavily populated cities in Texas, namely Houston, San Antonio and Dallas. All three cities boast a very diverse restaurant scene mostly due to their culturally diverse residents.

The goal of this analysis is to aid a business owner make decisions regarding whether it is feasible to open a restaurant in a certain area and expect positive social media presence, what type of cuisines perform well in certain areas and the best area overall out of the three cities.

# Data

## 2.1 Data Scraping and Cleaning

1. First we will retrieve the geographical coordinates of the 3 cities i.e. Houston, San Antonio and Dallas.

2. We will then leverage the Foursquare API to obtain URLs that will lead to the raw data in JSON format.

3. It is important to note the Foursquare API will retrieve all venue data other than just restaurants i.e. concert halls. This means we will have to clean the data to remove all non-restaurant venue data.

4. The "id" is an important column because it is where we will pull "likes" from. After retrieving "likes" based on the restaurant "id" we will append it to our data frame.

## 2.2 Data Preparation

The raw data set we have thus far still needs more processing before we can use it to model. For example, the column 'categories' contains large variation of cuisines that will limit us from drawing very meaningful results due to the broadness. Therefore, as part of data preparation we will group these variations into groups of cuisines i.e. Asian, European, Latin, North American, Casual as in the case of cafes and drinking establishment as in the case of bars.

Our 3 different cities of focus and 6 different categories of cuisines are all categorical variables. Hence, we will require dummy variable encoding for meaningful analysis. We can accomplish this via one-hot encoding.

# Methodology

1. In our analysis, we will compare two machine learning models; linear and logistic regression. We will use the linear regression to attempt to predict the number of "likes" a new restaurant will have.

2. The logistic regression will be used as a classification method to possibly predict range of "likes" a new restaurant will have. Since we have multiple categories, we will use multinomial regression.

# Results

## 4.1 Linear Regression Results

The linear regression model was trained on a random subsample of 80% and tested on the remaining 20%. The residual sum of squares score and variance score were calculated.

## 4.2 Logistic Regression Results

1. A multinomial ordinal logistic regression model was trained on a random subsample of 80% of the sample and then tested on the other 20%.

2. The Jaccard similarity score and log-loss were calculated.

# Discussion

For the purpose of this project, we are assuming that 'likes' are a good proxy to show how well a new restaurant will do in a certain location, with a certain type of cuisine. Whether or not this assumption holds in real life scenario is subjective to the data available. It is Important to note however, that this analysis is limited in scope as far as the amount of data that can be fetched from the Foursquare API.

Using logistic regression we were able to obtain a Jaccard Similarity Score of 75%, which although not perfect, is more reasonable than the low variance score obtained from the linear regression. Therefore, given the data, logistic regression presents a better fit for the data over linear regression. Different binning methods for the classes were attempted, but the use of 2 bins by far yielded the best Jaccard Similarity Score.

# Conclusion

In conclusion, after analyzing restaurant "likes" from 300 restaurants in Texas 3 largest cities of Houston, San Antonio & Dallas, we can conclude that the best approach to take in regards to maximizing business performance (as measured by "likes") is to set up an 'Amercian' style restaurant in either Houston or San Antonio. However, the bar style restuarant also performed exceptionally well in Houston and San Antonio, so it could be a viable second option. Dallas had the lowest 'likes' in any cuisine and therefore would not recommend to set up there. Additionally, the predictive capabilities of the logistic regression prediction model are most accurate for classifying whether a restaurant will fall in either the best or worst classes when the data is binned into 2 classes.