




# Capstone Project

## Team Lucky 7

Caleigh Teahan  
Joe Demler  
Jongwook Choe  
Nilusha D.G.  
Vidhya Lakshmi



---

Which food do **Americans** prefer:  
Italian or Mexican food?



# Introduction

In this project, we analyzed restaurant data from across the United States to try and draw conclusions on consumer trends, specifically their preference in cuisines. We later took this data and created machine learning models to attempt to predict consumer restaurant ratings.



# Methods Used

**Data Preparation:** We extracted data from the Yelp API under specified location parameters. Then, we converted, cleaned and dropped unnecessary columns to create a usable Pandas dataframe.

**Data Visualization:** Using Tableau and Mathplotlib, we created visualization to explain our data and to show data biases.

**Machine learning Modeling:** Using both supervised and neural network models to train our data to predict a restaurant's rating based on the parameters of price, location, cuisine and customer reviews.

# Extracting the Data

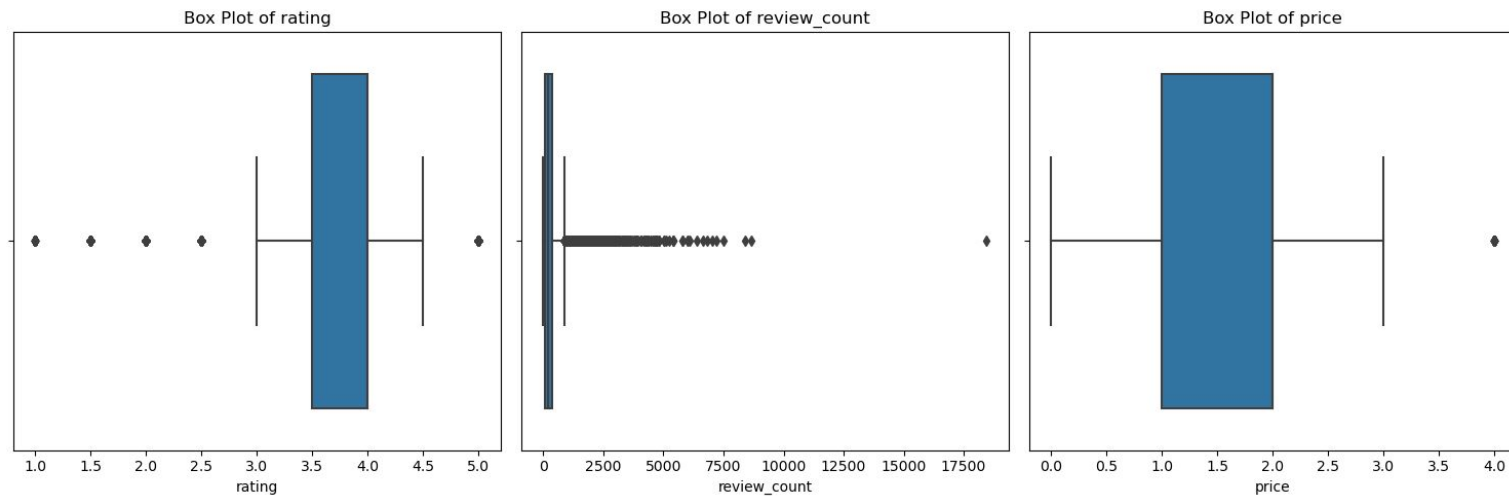


- Objective: To collect restaurant data from Yelp's API, perform analysis and create a model to predict restaurant rating.
- Filters added to request:
  - Location- Top 40 cities based on population
  - Cuisines- Mexican, Italian
  - Offset - To avoid duplicates while getting a large set of data
- Total = 19145 restaurants
- Columns = id, alias, name, image\_url, is\_closed, url, review\_count, categories, rating, coordinates, transactions, price, location, phone, display\_phone, distance, group\_city

# Data Cleaning

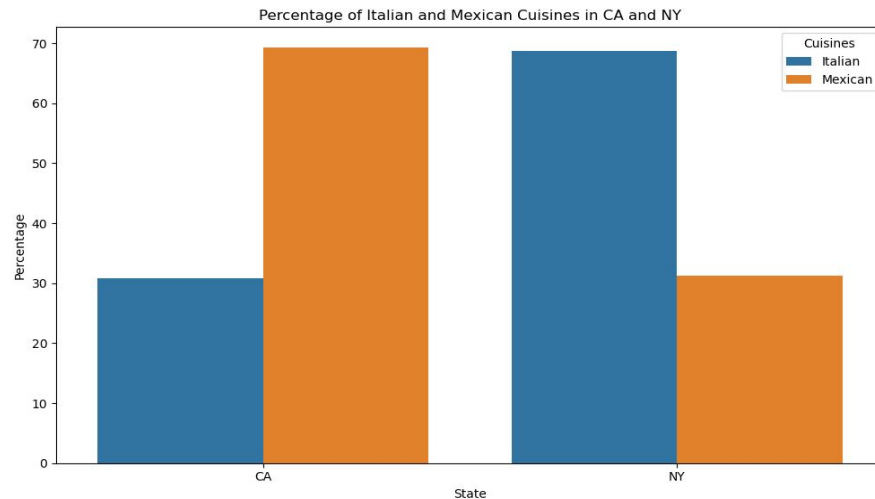
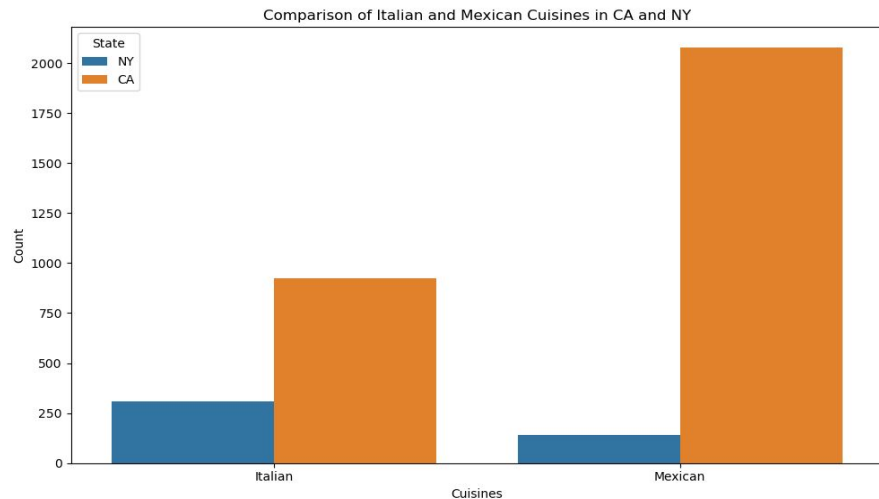
- Dropped columns('alias,' 'phone,' 'display\_phone,' and 'distance') irrelevant to analysis
- Creating a new column for cuisines, converting categories column from string to list of dictionaries and applying lambda function to extract cuisine titles.
- Extracting latitude and longitude from coordinates column
- Converted the 'price' column from symbols ('\$' to '\$\$\$\$') to integers (1 to 4) for price related analysis
- columns= id, name, image\_url, is\_closed, url, review\_count, rating, transactions, price, group\_city, cuisines, latitude, longitude, state

# Data Exploration



These plots show the spread of our data based on rating, review count and price. Some of the outliers can be explained by chain restaurants

# California and New York



These plots show a comparison between **west and east coast** cuisine preferences.



# Supervised Learning Models

We used two different types of supervised learning models to predict a restaurant's rating based off our data:

- Random Forest Regressor and Classifier: 63% accuracy and a root mean square of 0.48
- Nearest Neighbor Regressor and Classifier: 61% accuracy and a root mean square of 0.61

## Random Forest Results:

Mean Absolute Error (MAE): 0.41

Mean Squared Error (MSE): 0.23

Root Mean Squared Error (RMSE): 0.48

	precision	recall	f1-score	support
0	0.62	0.60	0.61	2892
1	0.63	0.65	0.64	2995
accuracy			0.63	5887
macro avg	0.63	0.62	0.62	5887
weighted avg	0.63	0.63	0.63	5887

## Nearest Neighbor Results:

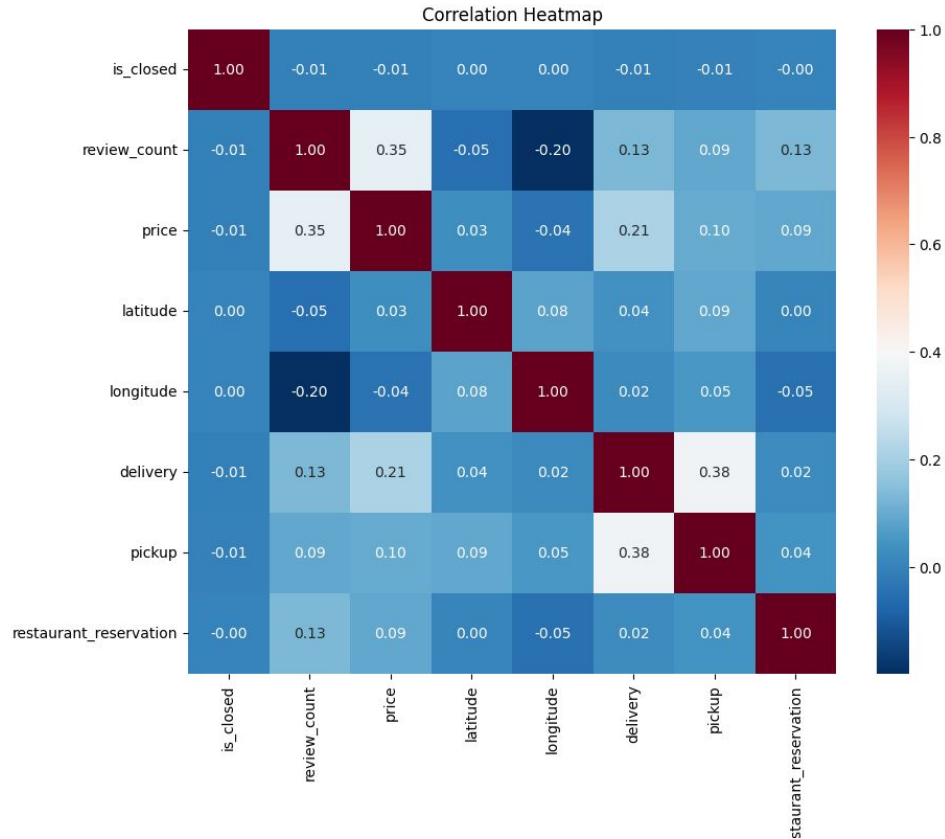
Mean Absolute Error (MAE): 0.39

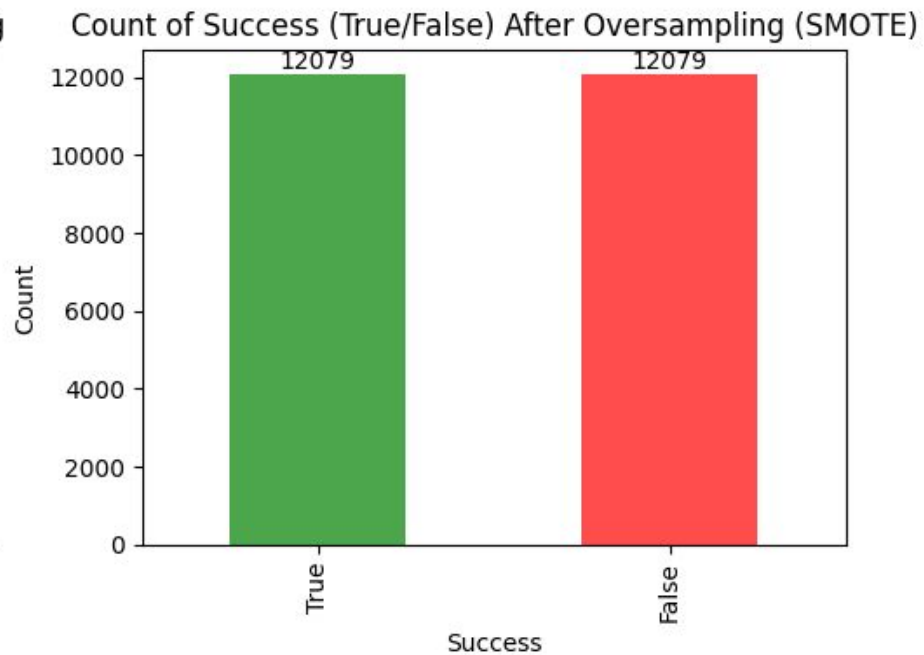
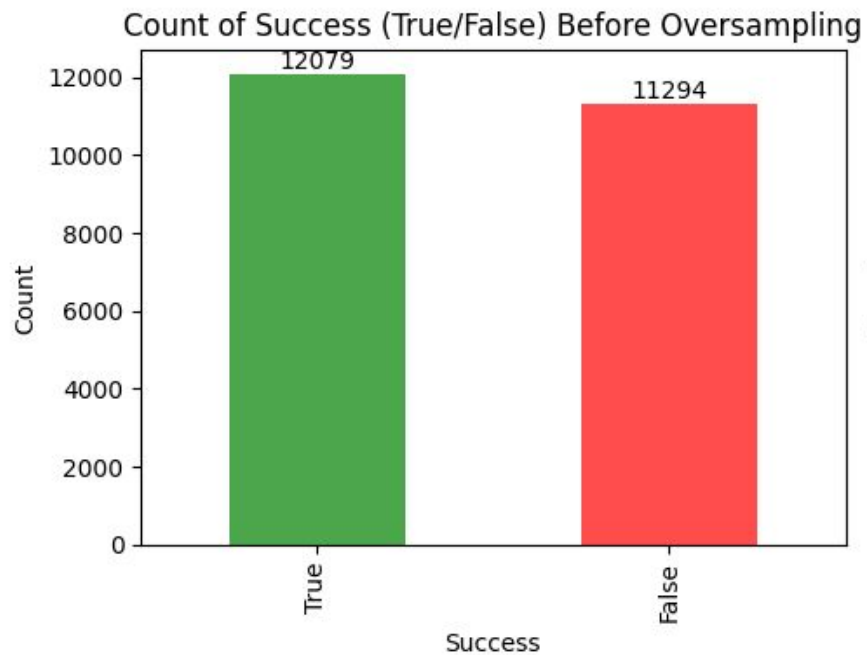
Mean Squared Error (MSE): 0.39

Root Mean Squared Error (RMSE): 0.63

	precision	recall	f1-score	support
0	0.60	0.62	0.61	2892
1	0.62	0.59	0.61	2995
accuracy			0.61	5887
macro avg	0.61	0.61	0.61	5887
weighted avg	0.61	0.61	0.61	5887

# Neural Network Model





Layer (type)	Output Shape	Param #
dense (Dense)	(None, 400)	158800
dense_1 (Dense)	(None, 150)	60150
dense_2 (Dense)	(None, 5)	755
dense_3 (Dense)	(None, 1)	6
Total params: 219711 (858.25 KB)		
Trainable params: 219711 (858.25 KB)		
Non-trainable params: 0 (0.00 Byte)		

567/567 - 2s - loss: 0.5863 - accuracy: 0.6863 - 2s/epoch - 3ms/step  
 Loss: 0.5863198637962341, Accuracy: 0.6862961649894714



# Tableau Visualization

- 1) Dining Across America
  - a) Average Price by State
  - b) Average Price, Rating, and Review Count by Cuisines
  - c) Review Count by State
- 2) Delight in Details
  - a) Top 10 Restaurants by Average Reviews and Average Price
  - b) Average Review Count by Cuisines
- 3) City Spotlight
  - a) Cities with the Most Ratings and Reviews
  - b) Cities with the Most Expensive and most affordable food
- 4) State of reviews
  - a) Average Review Count by State



[https://public.tableau.com/app/profile/nilusha.dg/viz/Project4\\_16955975765120/DashboardP4?publish=yes](https://public.tableau.com/app/profile/nilusha.dg/viz/Project4_16955975765120/DashboardP4?publish=yes)

## Concluding thoughts & what we could have done to improve

- Our project dived into the extensive analysis of restaurant data across the United States, focusing on **Consumer Preference for Italian and Mexican Cuisine**. Through data preparations and visualizations we gained valuable insights.
- Our supervised machine learning models yielded accuracy rates of **61%** and **63%** providing valuable insights on predicting restaurant ratings.
- Using deep learning neural network model achieved an accuracy of **68%** on the test data.
- Collect data more data with greater context (restaurant attributes, etc.) to improve our machine learning model accuracy.
- To answer the question, **which food do American prefer...**

Italian!





Questions?