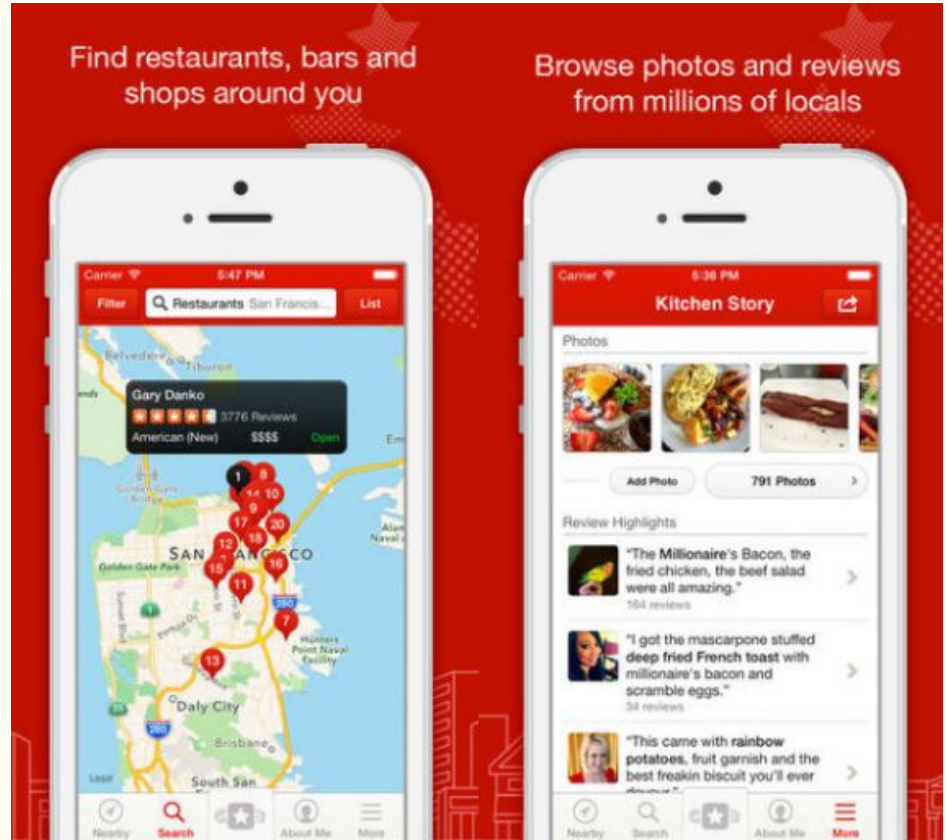# Yelp Recommendation System

DS-1 Project
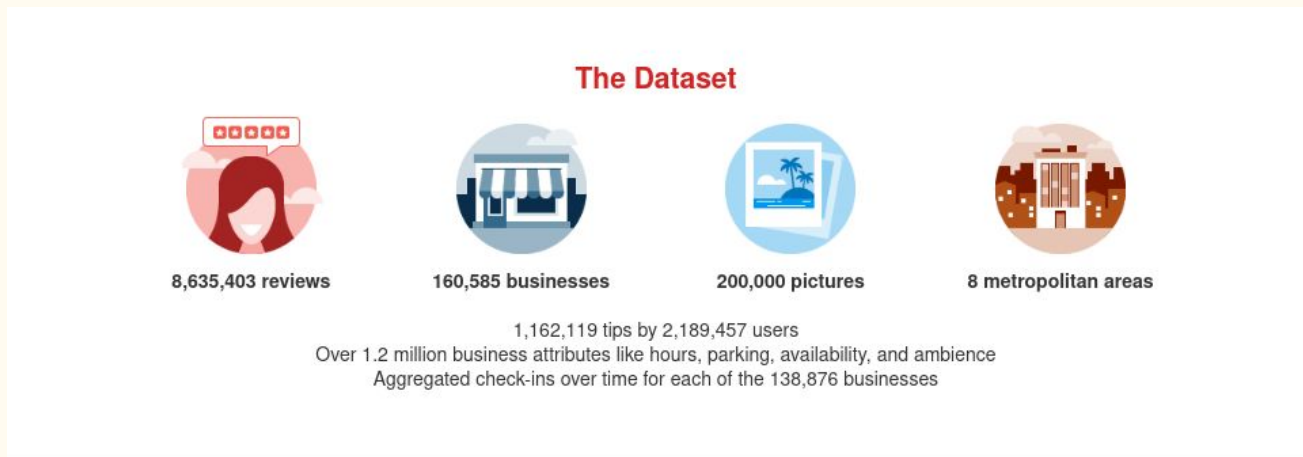
Sai Chaitanya Sanivada, Aayush Birla, Himanshu Kandpal, Samyak R Jain

Problem Statement

Building a recommendation system for businesses including restaurants, bars and shops using collaborative filtering by SGD.

Dataset

The Dataset

8,635,403 reviews   160,585 businesses   200,000 pictures   8 metropolitan areas

1,162,119 tips by 2,189,457 users
Over 1.2 million business attributes like hours, parking, availability, and ambience
Aggregated check-ins over time for each of the 138,876 businesses

Yelp Dataset

- Business data containing location data, attributes and categories

- Restaurant reviews data

- User data and metadata

- Check-ins on a business

- Tips written by user on a business

- Photo data including the caption and classification

Exploratory Data Analysis

Most Reviewed Restaurants
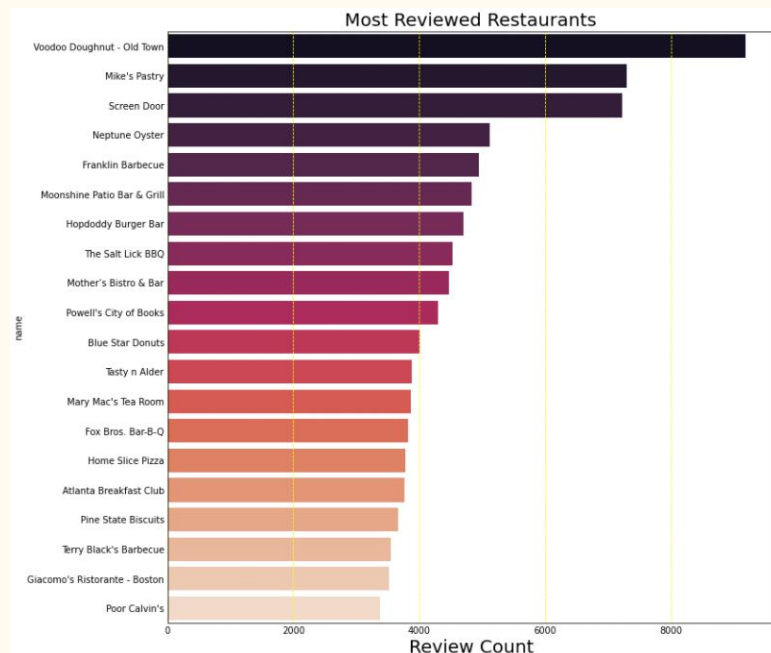
We see that the most reviewed restaurant has over 8000 reviews

Mean Rating by Cuisine
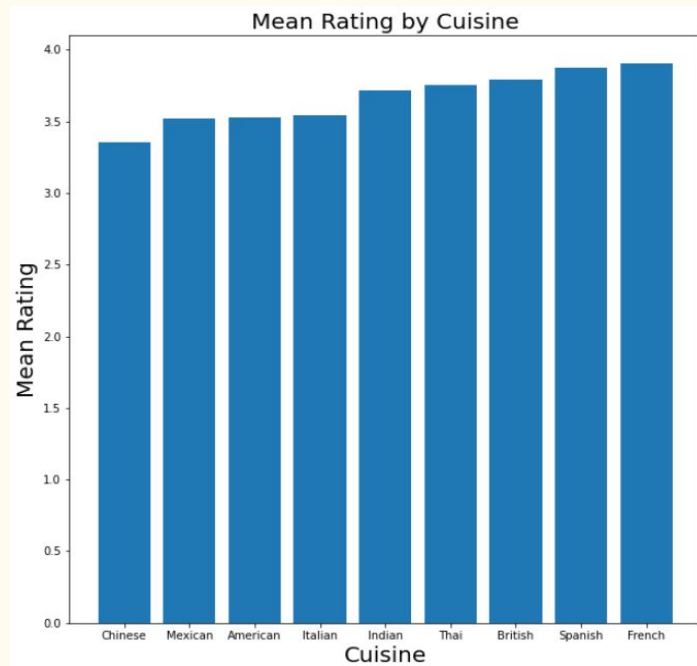
We observe that the French, Spanish and British cuisines are the most popular by average rating

Exploratory Data Analysis

Number of businesses with minimum of 5000 reviews/ratings: 5
Number of businesses with minimum of 1000 reviews/ratings: 408
Number of businesses with minimum of 500 reviews/ratings: 1741
Number of businesses with minimum of 100 reviews/ratings: 15791
Number of businesses with minimum of 50 reviews/ratings: 29037
Number of businesses with minimum of 10 reviews/ratings: 82888
Number of businesses with minimum of 5 reviews/ratings: 113898

Number of Users and Businesses w.r.to Total size of Data
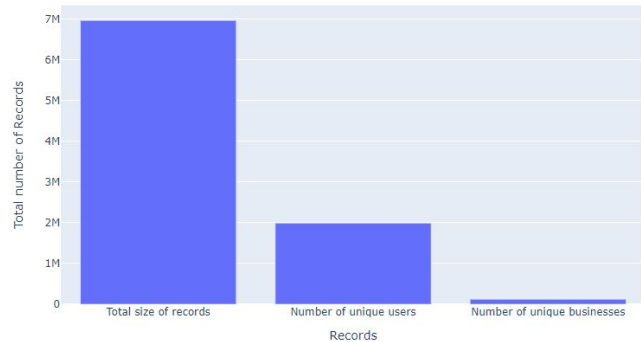
Ratings given by user
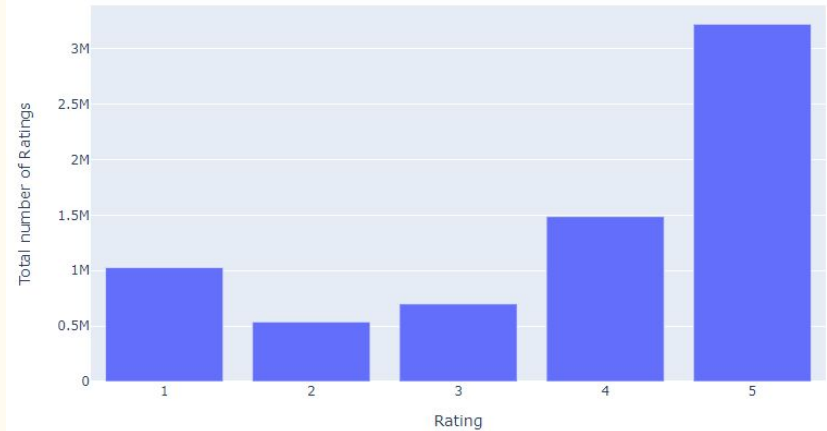
5 is the most common rating given by a user followed by 4

7 Million records with approx 2 Million unique users and around 0.2 Million businesses

Recommender System

- From e-commerce to online streaming platforms, recommender systems employ the past purchase patterns on it's user to predict which other products they may in interested in and likely to purchase.
- **Recommending the right products** gives a **significant advantage** to the business.

- Collaborative filtering uses **similarities between users and items simultaneously** to provide recommendations.
- **Similar users will rate similar restaurants (items) similarly**
- One of the requirements to be able to apply collaborative filtering is that sufficient number of products need ratings associated with them.
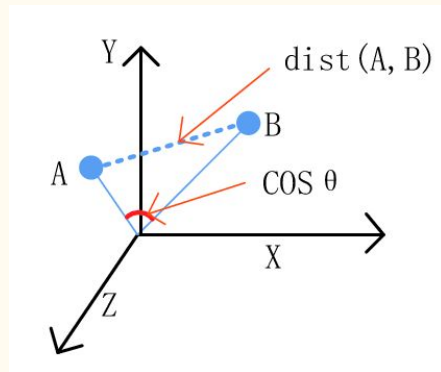
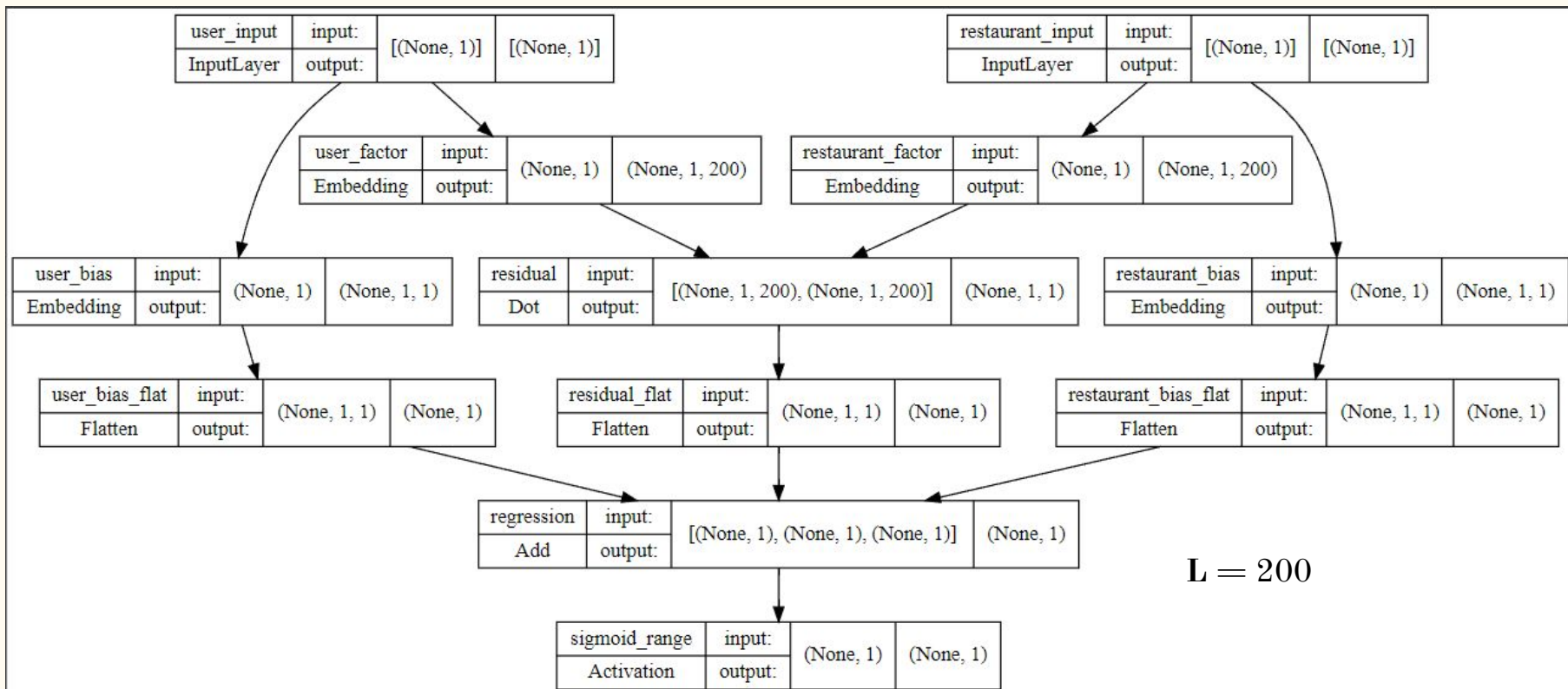# Collaborative Filtering Model

Estimate Function:

$$Y_{um} = \mu + \bar{\theta} \cdot I_u + \bar{\gamma} \cdot I_m + \bar{q}_m^T \cdot \bar{p}_u$$
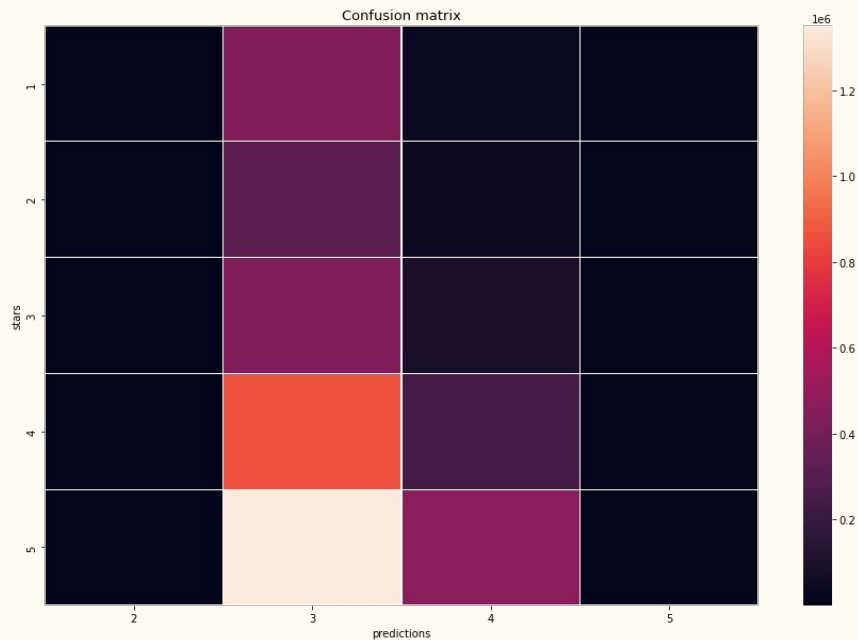

Loss Function:

$$\sum_{u,m} \left(Y_{um} - \mu - \bar{\theta} \cdot I_u - \bar{\gamma} \cdot I_m - \bar{q}_m^T \cdot \bar{p}_u \right)^2 + \alpha \left(\theta_u^2 + \gamma_m^2 + \|\bar{q}_m\|^2 + \|\bar{p}_u\|^2 \right)$$
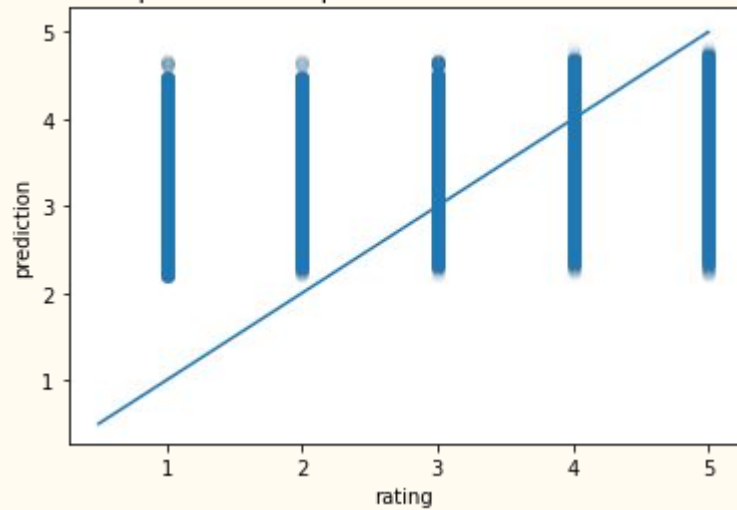
Collaborative Filtering Model

user_input	input:	[(None, 1)]	[(None, 1)]
InputLayer	output:		

restaurant_input	input:	[(None, 1)]	[(None, 1)]
InputLayer	output:		

user_factor	input:	(None, 1)	(None, 1, 200)
Embedding	output:		

restaurant_factor	input:	(None, 1)	(None, 1, 200)
Embedding	output:		

user_bias	input:	(None, 1)	(None, 1, 1)
Embedding	output:		

residual	input:	[(None, 1, 200), (None, 1, 200)]	(None, 1, 1)
Dot	output:		

restaurant_bias	input:	(None, 1)	(None, 1, 1)
Embedding	output:		

user_bias_flat	input:	(None, 1, 1)	(None, 1)
Flatten	output:		

residual_flat	input:	(None, 1, 1)	(None, 1)
Flatten	output:		

restaurant_bias_flat	input:	(None, 1, 1)	(None, 1)
Flatten	output:		

regression	input:	[(None, 1), (None, 1), (None, 1)]	(None, 1)
Add	output:		

$$L = 200$$

sigmoid_range	input:	(None, 1)	(None, 1)
Activation	output:		

Results

Confusion matrix

Scatter plot of model predictions vs actual stars (rating)

Conclusion and Future Work

- In conclusion, we have used **matrix factorization techniques with regression** to predict the user - item rating.
- **Recurrent Neural Networks** can be used in the future as **incorporating time** in the model is important.
- A few restaurants serving seasonal foods can have difference in their ratings corresponding to the current season.
- We can include other businesses such as saloons, Hotels along with restaurants in our model.
- We can use the business reviews text data available in the dataset to perform sentiment analysis and predict business ratings
- We can also use the photos available in the dataset to find similarities between different restaurants.

Thank You