Project Proposal

Problem Statement:

Yelp is commonly used in our life when we want to find or checkout if a place is good to go for food or drink. As a heavy Yelp user, it occurred to my friends and me that stars rating doesn't truly reflect whether a restaurant is good or not. We also need to consider the numbers of reviews, the average number of reviews and the contents of reviews to determine if it is a good restaurant to try out. However, it would take us more than few minutes, even an hour, to find good places and decide where to go. When we don't have enough time, we often just pick a 4-5 stars rating restaurant, and often turns out not as good as we expect.

Goal:

The project aim to predict good food businesses(restaurants, coffee shops, bars, etc.) on Yelp based on population of the area, number of reviews and stars rating by utilizing US cities and towns population data and Yelp data. The result should be able to help Yelp users to find 4-5 stars rating food businesses that are actually good in a short period of time.

Dataset:

1. Yelp dataset from Kaggle

https://www.kaggle.com/yelp-dataset/yelp-dataset#yelp_academic_dataset_business.json

2. Population data from Government Census

https://www.census.gov/data/tables/2017/demo/popest/total-cities-and-towns.html

Approach:

- Determine the range base for population to help finding the average number of reviews, and further determine the range of the number for the best result
- Research and consider in other potential factors to increase the accuracy of the model
- Develop the model using classification
- Construct different subdataset from Yelp dataset to test the model manually

Potential Challenges:

- Decide neccesary inputs for the model besides stars, location and review count
- Finding the optimal range of review numbers to generate the best result
- Choose data to form testing subdatasets

Deliverables:

Google Doc, PPT