PSTAT 134 Project Memo

Group 14: Tobias Starling, Deanna Hu, Nabeel Vakil, Derek Gong

Overview of Dataset

Our dataset includes millions of yelp reviews on every type of business. We will be obtaining the data from this open source Yelp API. There are over 6.99 million reviews on yelp with 20+ predictors featured in the json files. We will be working with numerical variables as well as text. There is some missing data, as lots of star reviews don't also have written text reviews. We will handle this by having the model primarily look at star reviews, then use written reviews as a secondary variable to make a decision between comparable restaurants.

Overview of Research Goals

We are interested in predicting the best three restaurants that a user would enjoy most based on a variety of factors. The input variables we will be assessing include the cost, category of food, review rating, proximity to the restaurant (using zip code), and text reviews. The response variable is the restaurants we believe the user will enjoy best based on these inputs. The most useful predictor will be the rating as the preliminary way to rule out bad restaurants. The goal of our model is predictive. We want to predict the best restaurant for a user to visit based on the area they're in.

Overview of Project Timeline

We plan on starting this by week 5 so we can have ample time to work on the actual recommender system and attend office hours if we need to troubleshoot. None of us are super technical so we will probably also spend time learning more about working with APIs and re-familiarizing ourselves with data cleaning and Python as well. The timeline will look like this:

Week 4: Familiarizing ourselves with the data & selecting variables that we want to work with Week 5: Cleaning the data Week 6: Work on recommender system, flesh out basic code Week 7-8: Testing and training model Week 9-10: Polishing up project

Questions and Concerns

Some issues that we're worried about are handling a lot of data, overfitting our recommendation, accuracy of location (being able to factor in distance to our recommendation), actual usefulness of our recommendation system, learning how to use a natural language processing system for analyzing the business reviews data.