Machine Learning Report

Team Members: Shefali Satpathy, Sakada Lim

Problem formulation

The objective of this project is to determine whether a restaurant will be successful if it is located near a college campus. There are various factors that impact the business of a restaurant such as location, ratings, reviews, price, food, ambience etc. All these factors contribute in creating a successful restaurant but in this project, we will predict the success or the failure of a potential restaurant based on its proximity to a college campus.

Review of past techniques

For this project, while we did not specifically find any other machine learning work that solves the exact same problem we are trying to do but we read a paper on determining the success of a movie based on various factors such as US Gross Revenue, IMDB ratings, critic score etc. published by Jeffrey Ericson & Jesse Grodman from Stanford University. We have not yet found any example code in relation to this project.

Additionally, we reviewed all the past labs we have covered during the semester in our Machine Learning class and used some of the techniques and tools provided in those labs.

Solution (Approach to the problem):

Datasets used:

College_Location.csv from NCES

Business.json.zip from Yelp Dataset

In Data_Processing_State.ipynb document, we first extracted the business.json file that contains data for 156 639 restaurants in the US and then converted the json file to a pandas object. We extracted the following attributes from the dataset: business_id, latitude, longitude, review_count, stars and state and exported it to a csv file ("rest_state_data.csv"). Both Data_Processing_State and rest_state_data.csv can be found in the repository.

In Process_distance.ipynb document, we extracted data for all the colleges in the US from the dataset and extracted the following attributes: id, name, latitude, longitude, and state. Using geopy python package, we wrote a python program to calculate the distance for each restaurant to the nearest 5 college campuses and exported it to a csv file ("rest_sample_state.csv").

In ML_Project_Final.ipynb document, we try to determine if there is any form of relationship between the rating of a restaurant and its proximity to a college campus.

Evaluation:

We split the data into training and testing and after plotting our results, we found out that there is no form of correlation between the ratings of a restaurant and its proximity to a college campus. To further analyse this relationship and see if we could use a different kind of

model to find a relation, we decided to plot the average star given to a restaurant against its proximity to college campus by binning the data yet that produced no desire results and confirmed that there is no relationship between the ratings of a restaurant and its distance to college.

While our results showed no plausible relation between the stars/ratings of a restaurant and its proximity to a college campus, some possible extensions to this project is comparing the relationship between the average price of food at a restaurant and its proximity to college campus. Since college students often prefer cheap food, it would be interesting to see if there is any correlation between the cost at a restaurant and its proximity to college campuses. Another possible extension to the project is making a recommendation system and thus, it would be interesting to make restaurant recommendations closest to college campuses based on ratings and price.