Capstone Project Proposal

Predicting Successful Business through Data Science

Muhammad Baqui

Low and medium sized businesses are essential for the success of middle class. The current election cycle of USA is seen as the revolt of middle class. The candidate with better economic agenda for middle class, has become the ultimate winner of the election. However, there is always a high risk involved in starting a new business since its success depends on various factors. The demographic and economic data of a particular neighborhood can be used to access the potential viability of a business in that area. This work aims to predict the business success for a particular area by relying purely on data science. The work employs two different datasets for prediction. The first dataset is the Yelp Academic dataset that offers business information for some selected neighborhood. Along with business name and location, the dataset also provides business ratings, user reviews, business categories etc. The second one is the US census that provides a wide variety of demographic and economic information of almost every places in USA.

The proposed work aims to study the relationship between demographic and economic data of a particular neighborhood to the business success of that place. The Yelp dataset lists all business information in terms of location, business success (ratings by star values), user reviews, currently in business or out of business etc. From census dataset, economic information, employment, population, amount of race diversity, income etc. can be extracted. To train the predictive model, the economic and demographic information along with business success (in terms of star ratings) will be taken as input features. During the testing phase economic and demographic information of a new area and category of the new business will be given as input. The model will predict the star rating of that particular business in that particular area.

To study the feasibility of implementing such project, a preliminary exploratory data analysis is performed with the obtained datasets. The analysis is performed with 'pandas' library of Python programming language. Initially the Yelp_academic_business dataset is read into the platform. Afterwards, it is explored that how many businesses per cities are listed in the dataset. It is found that the city of Las Vegas has the highest amount of businesses in this dataset. As a result, Las Vegas is chosen as the place to perform exploratory data analysis in this study.

The Yelp dataset has category columns that list the type of a particular business. First, which categories have highest number of entires is explored. It is found that many categories have redundant entries. For example 'Restaurant' category is found in 'Food', 'Chinese restaurant', 'Mexican' etc. It is also found that the categories column has many entires separated by comma in a single cell. For simplification these categories are separated into individual columns per each category. As a proof of concept, only category of 'Restaurants' and 'Fashion' are selected for this analysis. These two categories are chosen because they are found to have the most number of entries in the dataset. As the study is performed on successful businesses, the business with lower than star rating of 4 is discarded. The distribution of the successful fashion businesses are shown in Figure 1.

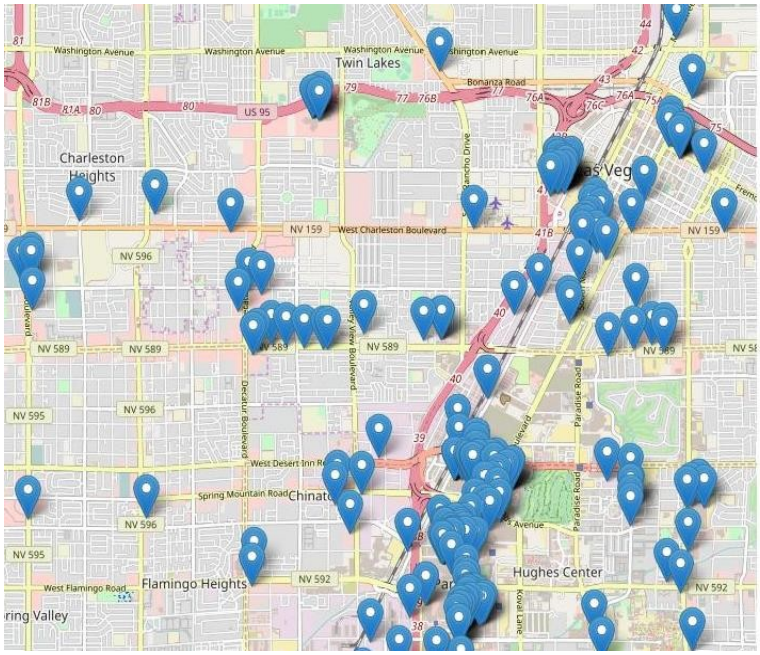

Figure1. Fashion business outlets distribution for Las Vegas

From US census, the income information is collected based on zip codes of Las Vegas. In order to view the areas and median house hold income, the longitude and latitude information of zip codes have been collected. Afterwards, a merge operation is performed in order to have income, longitude, latitude information based on zip code. Now, the income based on zip for Las Vegas can be viewed in a map (Figure 2.). The median household income of the locality is shown as yellow circle. The diameter of the circle shows a scaled value of the income. Higher the value of income larger the diameter would be.

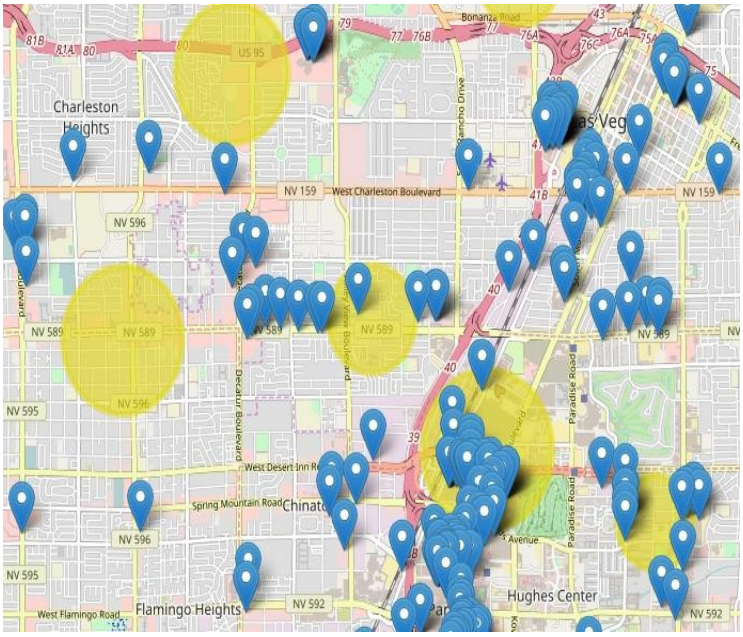

Figure2. Fashion outlets and median household income of Las Vegas

Similar operations can be performed to create a feature space of business rating, type, demographic, economic, diversity information of an area. A machine learning model will be trained with that feature space. Final model would be able to predict ratings of a new business for a particular area.