H1-B Visa petition data analysis and approval prediction

ORIE4741 Project Proposal

Chengnan Xu(cx223), Tian Yu(ty364)

1 Problem statement

H1-B visa, non-immigration visa category for international students who want to work in America, is highly desired. It requires petitioning employees with at least a bachelor's degree. The Citizenship and Immigration Services(USCIS) grants 85,000 H1-B visas every year, while the number of applicants far exceed that number. According to the 2018 Open Doors Report on International Educational Exchange, the number of international students in the United States surpassed one million for the third consecutive year, increasing by 1.5 percent to reach a new high of 1,094,792. The visa petition outcome for the applicants are of importance.

2 When the problems are important

In our project, we want to predict the outcome of H1-B visa petition based on the dataset. H1-B visa is a hot topic among the international non-residents in America. For international employees who want to continue working in America, it is of significant importance to predict whether their petitions can be approved because this information can help them plan their future life and career. For students who are in pursuit of a career in America, They also care about the statistical information of H1-B petition: What kinds of companies are most likely to sponsor applicants for H1-B petitions? Which states have the highest passing rate? Which kinds of jobs are more likely to be approved during H1-B visa review? Which employers file the most petitions each year? These information can guide them in career planning.

3 Dataset Description

The dataset is collected from Kaggle, based on the petition documents from 2011-2016. The dataset consists of more than three million applicants information. It contains 10 columns: case status, employer name, soc name for applicants who applied for visa and categorized into different job type, full time position, prevailing wage, filing year, worksite, longitude and latitude. This dataset is based on the USCIS and is the most direct detailed dataset for systems to decide whether the applicants would be certified or denied. Since the dataset is very clearly detailed and each features for the prediction don't have to much messy data. We can handle it successfully.