H1_B Visa petition data analysis and approval prediction

ORIE4741 Project Proposal

Chengnan Xu(cx223), Tian Yu(ty364)

1 Dataset Description

H1_B visa, non-immigration visa category for international students who want to work in America, is highly desired. It is required petitioning employees at least a bachelor's degree. As more and more employees go into America, based on Citizenship and Immigration Services(USCIS) grants 85,000 H1_B visas every year, the visa outcome for the applicants are of importance.

The dataset is collected from Kaggle, baes 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.

2 Questions

In our project, we want to predict the outcome of H1B_visa applicants based on the dataset. Since for people who want to work here, it is of significant needs for then to know whether their petitions can be approved. Specifically, there are many questions proposed: What kinds of companies are easier to be approved and have more applicants for petitions? Which states have high ratio for the pass rate? Which kinds of jobs are more popular and have a higher rate of petitions? As the job in data science more and more popular, the question about which industry has the most number of Data Scientist positions can be attractive for applicants. Which employers file the most petitions each year?

3 Why the problems are important

Since more and more people in America to finish their bachelor's degree, master or even doctor's degree, predict the outcomes of petitions is of significance This dataset based on the USCIS and it 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 i successfully.