Capstone Project Final Idea

Jun Ho Lee
08/25/19

H1B Visa Application Data

Background: *International students (non-US citizens / non Green Card holders) must hold a visa to legally work in the United States. Among various types of visas sponsored by the US government, H1B is the most popular form of work visa. The approval process however is quite stringent, with higher ratio of applicants getting their application denied. Through this dataset, I aim to understand if there are certain unique features that can predict which applicants eventually get approved for the H1B visa.*

1. **What is the problem you want to solve?**

- Can we predict which variables contribute the most in determining the approval rate of the H1B Visa?

- This could be both a regression and a classification problem in that:

A. we can analyze which predictors have the highest coefficients in determining visa approvals

B. we can analyze which predictors contribute the most in determining visa approvals versus visa rejection.

- End goal of this project is to inform future visa applicants on which factors to focus on when filling out their visa applications and when going through the job search so that they may increase their chances of visa approval.

2. Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn't have done otherwise?

- I am personally an international student myself and would require a H1B down the line. Having a prediction model based on this data would be helpful when doing the job search. I would be able to make informed decisions based on the results of the model.

3. What data are you using? How will you acquire the data?

- The data I will be using comes from ***United States Department of Labor on Foreign Labor Certification.***
URL: https://www.foreignlaborcert.doleta.gov/performancedata.cfm
The dataset is from Fiscal Year 2018 with a reporting period from October 1, 2017 through September 30, 2018. To access the dataset from the URL provided above, click on the ***Disclosure Data*** *tab and download the* ***'H-1B_FY2018.xlsx'*** *dataset under LCA Programs (H-1B, H-1B1, E-3). The dataset dictionary is downloadable under the* ***File Structure*** *column.*

4. Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.

- **A.** Since the data is managed and distributed by the government, I expect that the dataset is relatively clean. I would still search for any missing values in the dataset and treat them accordingly (drop/impute values). Additionally, the dataset is pretty big (over 200MB), which would take up a lot of resources on my local device. Therefore, I would either chunk the dataset or convert the datatypes of each column so that my device uses less memory, allowing me to analyze the dataset more quickly.
- **B.** I would do an exploratory data analysis to visualize the correlation between the features (columns). If I find any colinear variables, I'll drop those variables to reduce my dimensionality.
- **C.** I will perform a logistic regression model with the remaining variables to see if I get good metrics. I will also try out SVM (Support Vector Machines) to classify between two states (rejection vs. approval)

5. What are your deliverables? Typically, this includes code, a paper, or a slide deck.

There would be mainly **two** deliverables for this project:

1. Jupyter notebook that includes all my raw code and reasoning for the decisions I made.
2. PowerPoint presentation that summarizes the key results from the project and future directions that would be interesting to pursue – both by myself and other researchers.