Exploratory Data Analysis of H1B-LCA Petitions

Link to notebook on Github:
https://github.com/sopaturi/Springboard/blob/master/Capstone%20Project/Capstone%20Project-EDA.ipynb.

Below is a summary of analysis found in notebook.

Background

The H1B LCA application is a highly competitive application process that rewards 85,000 applicants per year with an H1B visa out of approximately 650,000 applications annually. An H1B visa is a temporary work visa granted to foreign workers. It is commonly given to individuals who have just finished their bachelors or post graduate degrees. A preliminary step needed to apply for the H1B visa is to apply for an H1B-LCA which is done by the employer.

The H1B-LCA application's main goal is to ensure that the employer is providing fair work conditions for the foreign born worker who is applying for the H1B compared to a citizen who would be applying for the same job. The H1B-LCA application contains information that pertain solely to the employer, not the employee. An application varies in many categories such as job title, employer name, prevailing wage, and worksite city. The goal is to predict the outcome of the H1B-LCA application in order to inform an employer what they can change in their application to increase acceptance rate. Currently there is not a published deterministic cause for application rejection but we hope to find a relationship between responses to certain fields of the application to application outcome.

The goal of the this EDA notebook is to go field by field and identify which fields contain values that may be predictors for application outcome. If the field may be a predictor because its values display a type of correlation to application outcome, we can feed the values into a predictive model. The dataset involves all of the H1B-LCA applications from fiscal year 2017.

The dataset was downloaded from, https://www.foreignlaborcert.doleta.gov/performancedata.cfm. The dataset consists of more than 600,000 data points. The data labels are divided into 4 classes: (1) Certified (2) Denied (3) Withdrawn (4) Certified Withdrawn. In the data wrangling step, the withdrawn application were removed because this action was taken by the employer and does not result in an application outcome determined by the Department of Labor.

The goal is to predict application outcome. It is evident, that the dataset is highly imbalanced towards certified applications as seen in a graph of application count. A smaller sample of the certified applications may need to be taken when comparing to the denied applications so that the machine learning model has a more balanced dataset.

There are 70,537 unique Employer-names in the H1B-LCA dataset. The different employers were sorted according to their acceptance rates. The acceptance rate is defined as the number of accepted applications per employer over the total number of applications per employer. From the graph of employer vs acceptance rate is not a uniform distribution. And hence, it can be a good predictor of case-status in our prediction task.

There are 862 unique occupation codes in the H1B-LCA dataset. There are more unique job titles than occupation codes because each occupation code can have many job titles and this is seen in the fact that the unique job title count is at 91,646, approximately 100x greater than occupation code count. The different occupation codes and job titles were sorted according to their acceptance rates. The acceptance rate is defined as the number of accepted applications per occupation/job titles over the total number of applications per occupation/job title. From the of occupation vs acceptance rate and job title vs acceptance rate is not a uniform distribution. And hence, occupation and job title can be a good predictor of case status in our prediction task.

Similarly, it was found that industry, agent representing employer, prevailing wage, wage offered, prevailing wage difference (difference between prevailing wage and database), wage difference (difference between wage offered and prevailing wage), missing value count, and worksite city had non uniform acceptance rates across the range of values for each field.

Wage difference was a big factor in determining if an application was denied. To reiterate, 2010 out of 2012 applications with negative wage difference between wage offered and prevailing wage were denied. A hypothesis test was performed below to check if the mean wage difference for accepted applications was different than denied applications. If the test shows than the means are different, that means we can say difference in mean wage difference for accepted and denied applications is statistically significant. In addition a difference in mean mean wages will confirm that wage difference is a differentiating factor between accepted and denied applications. A hypothesis test was used for checking if the mean of wage difference for accepted vs denied applications is significantly different.

A two sample z-test was used because the population variance was known and the sample size taken was larger or equal to than 30. The significance level was set to 95%=significance level.

The z score was -3796.9. This value is well below the cut-off 0.05. So, we can reject the null hypothesis that there is no difference between the mean of wage difference between accepted and denied applications.

To conclude the EDA section, the following fields: Employer name, Agent representing employer, Occupation code, Job title, Industry, Prevailing Wage, Wage offered by employer, Wage difference, Prevailing wage difference, Worksite city, and
Missing value count will be used in the predictive task as the distribution of these field's acceptance rates were not uniform. Once a prediction is made, we can determine which fields

most influence application outcome and inform employers which fields and the values in the fields may be causing application rejection.