Predicting H1B Labor Condition Application Petition Success: Capstone 1 Milestone Report

# 1. Problem and Client

The problem I have chosen for my first capstone project is attempting to predict the outcome of the Labor Condition Application (LCA), a pre-requisite for H1B visa applications. The LCA is the final step before applicants may submit their H1B visa application to the US Department of Labor. Once the LCA is approved and the visa application is submitted, the next steps involve a lottery and therefore the decisions are essentially taken out of the hands of the applicants. For this reason, it is essential that applicants have their LCA approved as it is the last time that their choices are taken into consideration.

There are two prospective clients for this work. One would be a foreign citizen who would like to obtain employment in the United States. This person might want to know which types of jobs, what employment levels, and in what parts of the country they should be looking for employment to best maximize the chances that their H1B visa is ultimately approved. The other prospective client is an American company who may be looking to hire a specific foreign citizen. They may want to ensure that the combination of position, salary, employment status, etc. that they are offering their prospective employee is likely to succeed the application process.

The main dataset that I will use for this project consists of LCA petitions filed with the US Department of Labor between 2011 and 2016. I have obtained these data from the Kaggle H-1B Visa Petitions 2011-2016 dataset (<https://www.kaggle.com/nsharan/h-1b-visa>). This represents roughly 3 million LCA petitions. It should be noted that the data made available through Kaggle do not include all fields that are required on the LCA petition. However, they do include the final case status, employer name, standard occupational classification (SOC) system name, job title, whether the position will be full time or part time, the prevailing wage for that position, the city and state of the position, as well as the latitude and longitude of the employer. Other datasets include census block information and county population data available from the FCC Area API and the Census Population Estimates API.

# 2. Data Wrangling and Cleaning

The dataset provided by Kaggle was already relatively clean. With that said, it was necessary to do some cleaning and rearranging of the data to suit the needs of my analysis. After downloading the data and reconciling field names, my first step was to inspect the field “CASE\_STATUS.” This is the field that contains the outcome of the LCA process and is the outcome that I will eventually attempt to predict. An online search revealed that there are only four valid values for this field: certified, certified-withdrawn, withdrawn, and denied. For the purposes of this analysis, I decided only to keep certified, certified-withdrawn, and denied.

While it is helpful to know the outcome when an application was certified but then withdrawn, it is not helpful for prediction if all I know about the application is that it was withdrawn. Therefore, I believe that it is appropriate to drop records where CASE\_STATUS is “withdrawn.” I also wanted to condense these results so that I had a binary success-failure field. I created a new field (“CERTIFIED”) where all records with CASE\_STATUS values of “certified” or “certified-withdrawn” have the value “certified” and records with the CASE\_STATUS value of “denied” have the value “denied.”

Other cleanup included changing the datatype of YEAR from a float to an integer. I split the field WORKSITE into two separate fields so that I now have one field with cities (“CITY”) and one field with sates (“STATE”). Finally, I created a new column (LOG\_WAGE). This is populated with log10 value of the PREVAILING\_WAGE field. PREVAILING\_WAGE has values that span from 0 to over 10^9 so I determined that it would be easier to visualize this spread of the data on a log-scale.

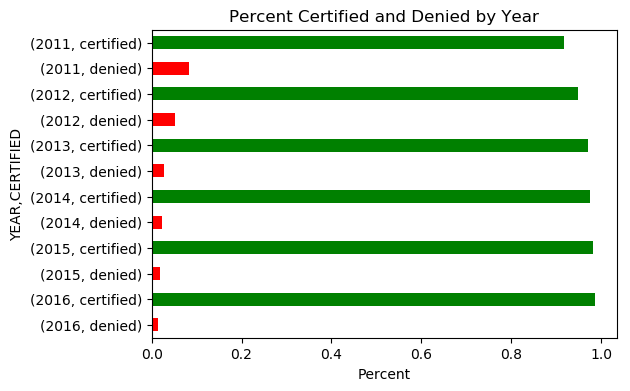
In addition to the cleanup of the data obtained from Kaggle, I generated two other datasets via calls to the FCC Area API and the Census Population Estimates API. First, I obtained a unique list of latitude and longitude pairs from the cleaned H1B dataset. I then used these to call the FCC Area API. This resulted in a dataset that contains latitude, longitude, census block Federal Information Processing Standard (FIPS) code, population of the census block, and other county and state identifiers including the county FIPS code. I then used this FCC census block dataset and, with the unique list of county-FIPS codes, called the Census Population Estimates API. The resulting dataset included the population of each county in my dataset. Finally, I merged these two datasets so that I had one overall csv containing location information including FIPS codes for counties and states, as well as county population, and census block population.

All code that I used to clean the Kaggle dataset as well as make the calls to the FCC and Census APIs can be found on git hub here: <https://github.com/Liptoni/Springboard/tree/master/H1B_Capstone/python>.

# 3. Initial Data Exploration

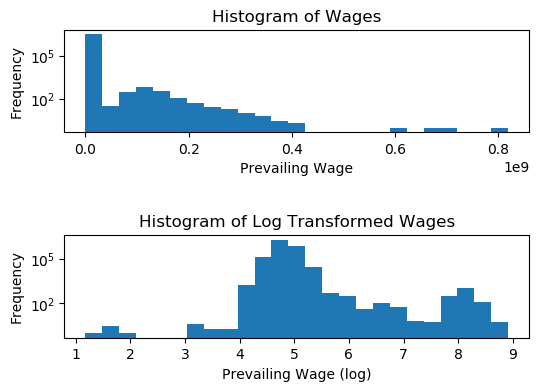
My initial exploration of the data revealed a number interesting items. The total number of certified applications is 2,818,282 while there were only 94,346 denied applications. This large discrepancy in the number of certified vs. denied applications leads me to believe that there may be some outside factor influencing which applications are submitted to the LCA process. It is possible that the service that I am proposing already exists in some form. When looking at the percent of applications that were certified by year, there was an increasing trend of certified applications from 2011 to 2016. In 2011, 91.6% of applications were certified. That number grew to 98.5% in 2016 (Figure 1).

Figure 1. Certified applications by year



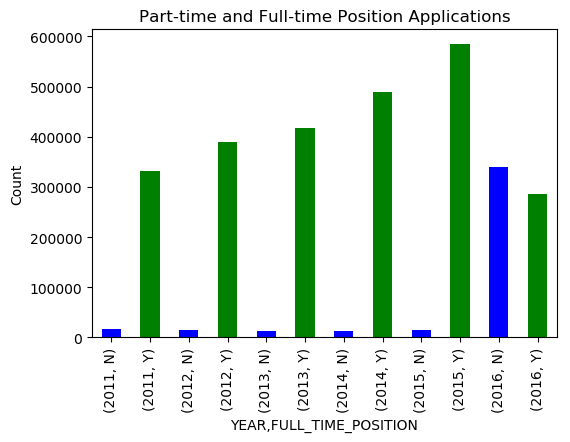
Next, I looked at the distribution of wages in general. There were two values over $1 billion dollars. While this may not be outside of the realm of possibility, the positions being applied for were physical therapist and software developer. I was forced to determine that these were erroneous records and removed them from future analysis. The mean wage, excluding wages that were reported as $0 was $141,144.20 while the median was $65,000. A histogram of prevailing wages revealed what the summary statistics indicate, which is wages that are heavily right skewed. This is not unsurprising as there would not be many incredibly-high paying jobs. Figure 2 shows the distribution of prevailing wage in general as well as on a log10 scale. The log transformation indicates that the data may be bi-modal with a cluster of very high wages.

Figure 2. Wage histograms.



Finally, I was interested in determining the number of full-time and part-time positions that were applied for each year. This turned out to be very interesting. In each year from 2011-2015, the number of full-time positions far exceeded the number of part-time positions. Then in 2016 the number of part-time positions being applied for suddenly exceeded the number of full-time positions (Figure 3). This extreme shift in the relationship between full-time and part-time positions being applied for indicates to me that there may have been some guidance issued by the Department of Labor indicating that certain positions would be more likely to be approved.

Figure 3



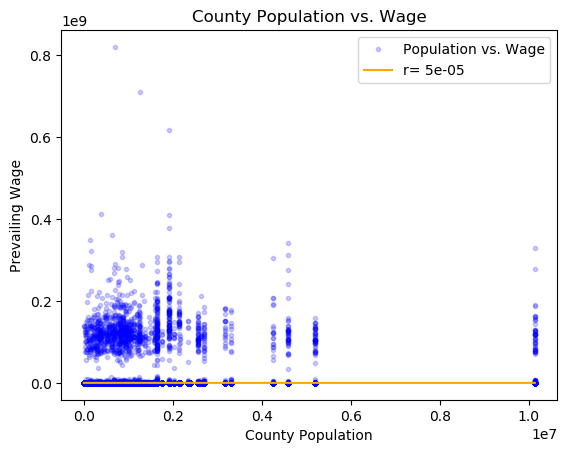
# 4. Preliminary Statistical Analysis

Following up this initial exploration I chose to do some preliminary statistical analysis of the dataset. The first test I ran was to determine whether the mean prevailing wage was the same for applications that were certified and denied. The results of the t-test provided a p-value of essentially zero. This led me to reject the null hypothesis that the means of the two sets of applications were the same. I ran a similar test to determine if the mean wages for full-time and part-time positions were equal and again obtains a p-value of essentially zero leading me to reject the hypothesis that the means of these two groups were equal.

I ran one more t-test. This time I was interested in testing whether the county populations for certified and denied applications were equal. The mean population for certified applications is 1,720,650 (n=2,721,739) and the mean for denied applications is 1,993,234 (n=85,107). The resulting p-value of the t-test was nearly zero leading me to reject the null hypothesis that the county populations for the two groups were equal.

Finally, I wanted to explore the relationship between county population and the prevailing wages. I plotted the two against each other and fit a least-squares line (Figure 4). The resulting Pearson correlation coefficient was 0.00005 indicating that there is essentially no correlation between county population and the wages of jobs being applied for in those counties.

Figure 4. County Population vs. Prevailing Wage.



All of the code I used for my data exploration and preliminary statistical analysis can be found in the jupyter notebooks here: <https://github.com/Liptoni/Springboard/tree/master/H1B_Capstone/jupyter_notebooks>.