*Labor Certification in the United States: Bayesian vs. Frequentist Approaches*

Jack Crum

Immigration and the place of immigrants in our society is the focus of a significant amount of the political rhetoric and debate in our current government. With the current administration’s “America First” agenda being highly anti-immigrant, it may become more difficult for foreign workers to attain a permanent labor certification and work in this country legally. For many immigrants, this labor certification is the only means to a better, safer, more prosperous life. As the process of attaining US citizenship is paved with bureaucracy and long waiting periods, permanent labor certification and permanent residency (green card) provide the best opportunities for many hopeful immigrants. This project will attempt to determine the variables that increase or decrease the probability of a permanent labor certification being denied.

A permanent labor certification issued by the Department of Labor (DOL) allows an employer to hire a foreign worker to work permanently in the United States. The DOL must certify to the U.S. Citizenship and Immigration Services that there are not sufficient U.S. workers able, willing, qualified and available to accept the job opportunity in the area of intended employment and that employment of the foreign worker will not adversely affect the wages and working conditions of similarly employed U.S. workers.[[1]](#footnote-1)

The data covers 2012-2017 and includes information on employer, position, offered wage, job posting history, employee education and past visa history, associated lawyers, and final decision. Final decision, certified or denied, will be the variable the project is attempting to predict. As the data is mostly categorical, with the exception of wages, and there is a large amount of missing data, thorough cleaning will be required. Final decision will be converted into a binary value, combining certified and certified (expired) and dropping withdrawn, the categorical variables will be converted to dummy variables, and the numerical variables will be normalized.

To compare the Bayesian method to the frequentist method, I will first conduct the analysis using logistic regression as the problem is binary in nature. I will also use a random forest model to conduct a robust analysis using significant features. A confusion matrix and ROC curve will be used for accuracy evaluation. For the Bayesian analysis, formal analysis on a binomial probability will be conducted. As the likelihood is in Bernoulli form, a conjugate beta prior will be developed from the known data to calculate a beta posterior. I will also create a more efficient MCMC model and a JAGS model, as conjugacy is not required for the simulation for JAGS. Bayesian models will be compared using pseudo-priors and model averaging.

All analysis will be conducted using R. The caret package will be used on all frequentist models for preprocessing and splitting the data into training and testing datasets, creating GLM’s and random forests, and evaluating model performance using confusion matrices and ROC curves. Base R, caret, MCMCpack, BEST, and JAGS will be used for developing all Bayesian models. The dplyr package will be used for data manipulation and ggplot2 will be used for generating visualizations.

All cleaning, EDA, and modeling building will be conducted by myself.

<https://www.kaggle.com/jboysen/us-perm-visas>

1. Permanent Labor Certification. Accessed March 07, 2018. https://www.foreignlaborcert.doleta.gov/perm.cfm. [↑](#footnote-ref-1)