Likelihood of a Work Visa Approval

Hitesh Vyas
hitvya@Stanford.edu

Siddhartha Prakash
sidp@Stanford.edu

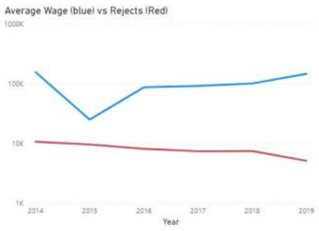
Abstract

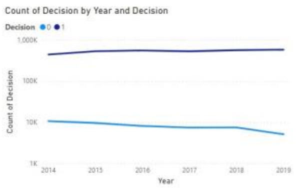
H1-B Visa is the most sought-after non-immigrant visa that allows foreign workers to work in United States in specialty occupation. In 2019, more than 1 million applicants applied to get an H-1B visa. From this pool a total of 500 thousand were approved and 3000 were denied after considering a wide range of data related to candidates.
The data is anonymized and publicly released. Data science techniques can be applied to study them and create a prediction model for approval which can be used by employers for risk estimation.

Dataset

The data used is from 2014 to 2019, with the study relying heavily on the data of 2019.
Out of a total of 73 features published, the ones used for the study include employer name, whether applicant is placed in secondary location, if another firm is representing employer, job title, SOC code, NAICS code, whether this is new employment for the candidate, whether the candidate is dependent on H1-B, whether the employer is a willful violator in past, etc.

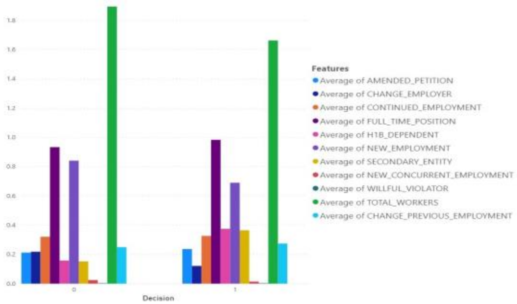
Exploratory Analysis





Experiments and results

Logistic regression by splitting the data in 70:30 ratio for testing and training
Naïve Bayes by combining a set of columns with textual data and then splitting it for testing and training
Random Forest* model generated using smaller dataset having equal subsample of rejected and approved cases and then testing it on entire dataset.



All logistic features average between accepted and rejected

	Logistic Regression		Naïve Bayes		Random Forest*	
	TN	TP	TN	TP	TN	TP
PN	1466	3679	14554	34311	1831	1010
PP	6660	577609	150135	3075885	180609	323916

Conclusion and Future Work

Logistic Regression model predicts with high accuracy but hides the true negatives as it tries to fit the data. So, visa outcome is not as dependent on employer and job profile as presumed, it could have random behavior in the decision. With individual company names, job titles and job categories as input to Naïve Bayes model, there is a drop in total accuracy but increase in predicting true negatives. Random forest model generated using a subset of equal number of randomly sampled approved and rejected cases and then applied to entire dataset substantially improves the prediction of true negatives but has lower overall accuracy.
A great future work will be to ensemble these three options and create a neural network for potential better accuracy in predictions

References

1. https://www.foreignlaborcert.doleta.gov/pdf/PerformanceData/2019/H-1B_Selected_Statistics_FY2019_Q4.pdf
2. https://www.uscis.gov/forms/h-and-l-filing-fees-form-i-129-petition-nonimmigrant-worker
3. https://github.com/hv5451/H1B
4. https://www.foreignlaborcert.doleta.gov/pdf/PerformanceData/2019/H-1B_Disclosure_Data_FY2019.xlsx
5. https://www.foreignlaborcert.doleta.gov/docs/py2014q4/H1B_FY14_Record_Layout.doc
6. https://blog.myyellowroad.com/using-categorical-data-in-machine-learning-with-python-from-dummy-variables-to-deep-category-66041f734512

Video Link

https://1drv.ms/v/s!Av-7kBjGzwHTgcUG3FIF9h2RM3_dZA?e=RbAsq0