Machine Learning Engineer Nanodegree
Capstone Proposal
Numan Yilmaz
February 4th, 2018

Proposal

Domain Background

H-1B is a type of visa that allows US employers to employ foreign workers to work in the USA for a certain time. For an employer, H-1B visa process starts with finding an employee outside of the USA. After that, employer offers the job to the employee and file a petition to the US Immigration Office. Response for a petition depends on the demand, job market, time of the submission, employee qualification etc and is about 2 - 6 months[1]. Entire process of bringing an employee from abroad could take up to 24 months. Duration of stay in the US is between 3 to 6 years after extensions.[2]

	Fiscal Year	Oct to Dec	Jan to Mar	Apr to Jun	Jul to Sep	Total
Petitions Filed	2013	40,048	39,433	159,380	60,606	299,467
	2014	45,211	42,781	158,623	72,209	318,824
	2015	51,964	46,088	176,042	74,575	348,669
Petitions Approved[9]	2013	76,720	52,859	79,813	77,381	286,773
	2014	64,526	58,121	91,779	101,431	315,857
	2015	64,799	44,217	84,233	82,068	275,317

Number of H-1B Petitions Filed and Number Approved[3]

Petitions are increasing every year. However, petition approvals are not increasing. US government is also encouraging employers to hire Americans. [4]

Problem Statement

H-1B visas are taking long time to process which could be 1-2 years. According to paysa.com research, employees are staying at the big tech companies less than 2 years.[5] After all, does it really make sense for big tech companies to go through visa process for an employee that may or may not be a fit for a position? To solve this problem, given the occupation name, job

title, position type and salary, a model could give us the chance of acceptance before even file a petition. That would save so much time and money for big companies.

Datasets and Inputs

Dataset for this project is from Kaggle.[6] The raw data can be obtained at The Office of Foreign Labor Certification (OFLC) website as well. The dataset contains five year's worth of H-1B petition data(2011 - 2016) which is approximately 3 million records.

The features in the dataset are case status, employer name, occupation name, job title, position type, prevailing wage, year, work site, latitude and longitude. Case status is the final decision and valid values include “Certified,” “Certified-Withdrawn,” Denied,” and “Withdrawn”. LCA : Labor Condition Application. DOL : Department of Labor.
Certified: Employer filed the LCA, which was approved by DOL
Certified Withdrawn: LCA was approved but later withdrawn by employer
Withdrawn: LCA was withdrawn by employer before approval
Denied: LCA was denied by DOL.
Certified statuses are the approval cases, others are not approved. Position type values are “Y” for full time, “N” for part time. Prevailing Wage is the annual salary for the job being requested for the foreign employee. Work site currently has city and state information but city information will be removed during data preparation process and only state information will be fed to the model.

Solution Statement

In order to predict if a foreign labor might get H-1B visa, I will create a model that takes occupation code, job title, prevailing wage, position type, work site as inputs and based on these inputs, the model is going to predict the case status as output. This is a two-class classification model.
I will make use of three different machine learning algorithms. The first will be the logistic regression. It is easy to implement and interpret. It is very fast to train and generally perform great on the test dataset. Second algorithm will be Random Forest. It is an ensemble model, meaning that it consist of multiple decision trees. It takes times to train the model. The nice thing about Random Forest is that it gives us the importance of each feature in the dataset. The third algorithm will be Support Vector Machines(SVM) which is a more sophisticated algorithm. It is hard to train and tune the model but gives us to find non-linear patterns in the dataset.

Benchmark Model

Being that this is a two class classification problem, we have 50 percent chance to guess the status of a case. The model should be more accurate than that.

Evaluation Metrics

Using accuracy as an evaluation metrics would the best if the dataset is normally distributed. Most of the cases the datasets are skewed. Therefore, precision, recall and f1 score will be used to evaluate the success of the model.

Project Design

This project will consist of six categories. Each category is explained below.

- **Data collection:** Sometimes data is a collection of image, audio or text data. The dataset I will be using in this project is from Kaggle.
- **Data Wrangling:** In this stage, dataset will be organized and cleaned to make it more usable for the next stages.
- **Data Exploring:** The data exploration stage is where some visualizations will be provided to understand the data. Some statistical methods will also be provided to gather information from data.
- **Data Transforming:** Most of the times, data has to be scaled using a scaler method in order to prepare it for the model.In order to visualize dataset, dimensionality reduction techniques will be used to reduced to number of features.
- **Data Modeling:** As it is mentioned above, three machine learning algorithms will be used to train a model. These are Logistic Regression, Random Forest and Support Vector Machines.
- **Model Evaluation:** This is where the model will be scored.

References

[1] https://www.quora.com/How-long-does-it-take-to-get-an-H1B-visa-and-be-able-to-work-How-can-it-be-done-faster
[2] https://en.wikipedia.org/wiki/H-1B_visa#Duration_of_stay
[3] https://www.uscis.gov/sites/default/files/USCIS/Resources/Reports%20and%20Studies/H-1B/H-1B-FY-2015-Petitions.pdf
[4] https://www.whitehouse.gov/presidential-actions/presidential-executive-order-buy-american-hire-american/
[5] https://www.paysa.com/blog/wp-content/uploads/2017/07/DisruptorsA8.png
[6] https://www.kaggle.com/nsharan/h-1b-visa