Identifying Job Opportunities in the US for International Students

1 INTRODUCTION - MOTIVATION

The United States of America is the most preferred destination for international students to pursue both undergraduate and graduate studies. This is evident from the number of international students, which crossed the one million mark in 2018. Most international students attempt to find career opportunities in the USA post graduation. Changing visa policies places restrictions on who can be employed and for how long a candidate can work in the USA before his/ her visa expires. Consequently, there has been a drop in the number of students by around 7% since 2017, which is a direct result of the exogenous shock given by the policy changes.

H-1B visas are the most preferred category of non-immigrant visas to obtain temporary work authorization in the United States. For an international student to obtain an H1-B visa, the US employer must submit a petition with the US immigration department. The petition details reasons why the specific international worker will be more suitable to perform the job tasks than an American citizen. Each year, the U.S. Citizenship and Immigration Service (USCIS) opens a total of 85,000 H-1B visas. In 2019, the number of applications for H-1B visa hit the 200,000 mark within the first week. According to the National Foundation of American Policy (NFAP), United States Citizenship and Immigration Services (USCIS) is continually increasing the number of visa denials and imposing stringent conditions for petition requests. These policy changes cost millions of dollars to the hiring companies. As a result, companies are reluctant in hiring an international student due to the increased complexities and challenges posed by the petition filing process.

Data on H-1B visa applications will be used to develop a visualization tool that will provide interstate and intrastate differences between job functions, number of H-1B visas accepted, average salaries per job function and other variables. The visualization tool and the predictive model will assist international students in making a more informed decision on the choice of educational institutions and targeted job search strategies. We expect that this project will help a student to maximize their chances of getting an H-1B visa approved in the future.

One of the datasets used in this project contains six years of historical data detailing H-1B applications from 2011-2016, with approximately 3 million records. With this dataset, we will develop a classification model for predicting the approval probabilities of new H-1B applications. The other dataset used in this project contains one year of historical data detailing H-1B applications in 2017, with approximately 600,000 records. This is our testing data. Using both datasets, we will identify the companies, states, and job categories that attract the highest number of visa petitions and will identify the potential sources of misuse of this provision.

2 PROBLEM DEFINITION

Development of an interactive web application to assist international students targeting job opportunities. Integration of a predictive feature that allows users to predict the likelihood of successfully obtaining a H-1B visa.

3 LITERATURE REVIEW

STEM Workers, H1B Visas and Productivity in US Cities This paper [9] evaluates the impact of H1-B visas issued to STEM workers using H1-B data from 1990 to 2010. The statistical tools and frameworks used in this paper can be replicated for analysing the H1B dataset. Authors highlight that only major metropolitan cities across the United States were affected significantly due to the policy changes. However, this study analyses data only till the year 2010 due to which, it is outdated therefore necessitating a more recent study to account for the recent policy changes.

Immigration and the Economy of Cities and Regions This book chapter [12] examines the impact of immigrant workers on local economies of cities and region. It examines that immigrant workers are distributed unevenly and give way to a larger variety of skills into the market. It is useful for this project because it gives us an idea of how H1-B visa holders will be distributed and their likely effect on the economy. The downside of this paper is that it covers a broad group of immigrants, not just H1-B holders. [16] analyses the effect of high-skilled immigration on the wages of U.S.

Conference'17, July 2017, Washington, DC, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

born college graduates. The model reports that the wages of native STEM majors will decrease compared to other majors as skilled immigration increases. The paper uses data from the 2010-202 American Community Survey and categorizes workers into different skill groups based on college major and U.S. labor market experience.

The Effect of the H-1B Quota on Employment and Selection of Foreign Born Labor In dynamic decision-making environment, a social process such as screening of H1B applications can be analysed using a statistical algorithm [4] that will pre-screen the applicants who do not meet the minimum requirements for H1B certification. The selected pool can then be reviewed by the visa officers for a more subjective analysis prior to granting the approval. [11] presents an empirical analysis of the various effects of reducing the cap on new H-1B employment relative to what would have occurred under a non-binding constraint. These effects include the number of native workers, the composition of H-1B workers, worker quality, and firm participation. This paper is useful for this project because it provides an understanding of the economic impact of the H-1B program.

Previous Predictive Modeling Approaches This analysis [5] has been conducted on a dataset containing information such as employers, employee demographics, wages, and occupations from accepted H-1B applications between fiscal years 2002 and 2009. The analysis [15] compares the performance of 7 machine learning models predicting a classification outcome for the H1-B process using the same dataset, achieving a performance of 94.6 with C5.0 decision tree. Another paper [17] presents a predictive analysis on the same dataset we are using as our training data. 70% of the dataset was allocated for training while 30% was allocated for testing. While many supervised learning algorithms were explored, the study found bagged random forests to be the best performing algorithm. Also using the dataset we are using as our training data, this paper [10] presents a predictive analysis that uses K-means clustering and decision trees. Clustering was used to find the top 10 cities, and decision trees were used to predict H-1B visa application outcome for the 10 most common job titles. Job title, location, and salary were used as the predictor variables. The rows corresponding to years 2011-2015 were used as the training dataset, and the rows corresponding to year 2016 was used as the testing dataset. [8] used the same dataset, but compared the performances of K-nearest neighbors, random forest, and multilayer perceptron classifier. Another difference was that they used position type (full time versus part time) as one of the predictor variables instead of job title. The authors found the best classifier to be dependent on the metric used for comparison. [15] used a random forest model to predict the state of a petition as positive or negative, using H-1B applications between fiscal years 2001 and 2016 and achieving an accuracy of 99%. [3] uses several machine learning models to predict the classification of a H1-B petition status as certified, denied, withdrawal or certified withdraws. It uses the same dataset as used in this study, with petitions filled from 2015 to 2017 to propose an ensemble model with accuracy of 95.4%.

Demography, Demand, and Job placement of International Students. How to Combat "Settling" for Jobs That Students Do Not Want Immigration and the representation of international students in the United States workforce is mitigated by a variety of factors, such as the availability of visas [6]. Regression analyses reveal a strong positive correlation between the number of international students moving to the United States from a given country and the number of H1-B visas allocated to that specific county [14]. The limitation and allocation of H1-B visas has impacted not only the number of international students immigrating annually, but also the careers that they decide to pursue [1]. Most notably, there has been an increase in the number of international students entering academia, even if it is outside of their area of study [1]. This demonstrates how the H1-B landscape, and its respective cap, has changed the way international students choose their career paths. Our project aims to help fill the gap between students' need to assess demand in their career of interest and the feasibility of visa sponsorship for their industry of choice. Empowering students with more knowledge about the opportunities available per industry along with the likelihood of sponsorship will better prepare students to tailor their job search efforts and make more informed decisions with regard to their career of choice. Our goal is to provide students with actionable insights through our visualization to help students combat the results proposed in "Settling for Academia? H-1B Visas and the Career Choices of International Students in the United States" [1] so that less students have to settle for careers outside of their preferred area of interest.

The need for an innovative job search tool targeted at international students:

In Pitre 2017[13], the author explores how acculturative stress can have a severe impact on the career outcomes of international students. Specifically, the author points out that visa regulations and cultural differences make the job search process specially stressful when compared to American students and concludes that "international students have a heightened need for guidance and support in vocational tasks". This validates our hypothesis that international students would benefit from a specialized job search tool.

The 2015 report by McKinsey[7] provides an in-depth analysis on how the job search and talent acquisition process has dramatically changed in the digital age. While it is focused on the perspective of recruiters, it provides important insights on how online talent platforms have altered the job market landscape. It evidences that the best tools are highly interactive and leverage Data Analytics. Finally, Anantharamu 2014[2] actually implements a Job Search tool highly focused on data visualization. Some features present in *OneCareer* will be similarly implemented in our website, like the geographical visualizations of job listings. However, the data sources and overall objective of the tools will be very different.

4 PROPOSED METHOD

4.1 Intuition

We believe our web-application adds value for international students for the following reasons:
New predictive modeling approach - Although there have been similar studies done with the dataset we are using, we have not found any studies that investigated the combination of three different classification algorithms.

4.2 Description of approaches

Predictive Model
To predict the outcome of a H-1B visa application, the relevant independent variables from our datasets are annual wage, type of position (full time versus part time), job category, and work-site state. The predictive model combines logistic regression, SVM, and K-nearest neighbors. The predicted values are the majority votes of these classification models.

- Logistic regression - This model uses the logistic function to get the probability of being in a group (accepted) for all data points. A point with a probability greater than or equal to 0.5 will be predicted to be accepted and a point with a probability less than 0.5 will be predicted to be rejected.

$$p(X) = \frac{e^{a+b*X}}{1+e^{a+b*X}}$$

  a and b represent the coefficients which are obtained after the training phase, and X represents the values of each of the independent variables of the data.
- Support vector machines (SVM) - This model determines an optimal hyperplane which linearly separates the data points into two classes. It uses the vector equation for a line to define the negative classification boundary as $w * x_n + b = -1$ and the positive classification boundary as $w * x_p + b = 1$
  x_n is the input vector from class 0 and x_p is the input vector from class 1. With the distance between these two boundaries as $\frac{2}{\|w\|}$, the size of the margin is $\frac{1}{\|w\|}$. Therefore, the SVM optimization model is:

  Minimize $\|w\|$
  Subject to $y_i(w * x_i - b) >= 1$ for $i = 1, ..., n$

  We minimize $\|w\|$ because the optimal hyperplane is the one which maximizes the margin between the two classes, and the size of the margin is $\frac{1}{\|w\|}$. i indexes the n dimensions.
- K-nearest neighbors - This model takes a certain number of nearest neighbors (k) as a parameter and assigns to a data point the label of the majority vote of its k nearest neighbors. A common measure of distance between two points is Euclidean distance. i indexes the n dimensions.

$$d(p,q) = \sqrt{(}\sum_{i=1}^{n}(q_i - p_i)^2)$$

The implementations of these models has been done using R. Python and Excel were used to help with some parts of the modeling process.

5 PREDICTIVE MODELLING

Dataset Preparation
The following data cleaning tasks with both training and testing datasets were done in OpenRefine.

- Removed columns which contained irrelevant and/or redundant information.
- Removed rows with missing values.
- Removed rows which provided hourly, weekly, biweekly, or monthly wages. Most rows provided annual wages.
- Removed rows which pertained to other types of visas.
- Removed rows which pertained to certified-withdrawn or withdrawn statues.
- Resolved spelling inconsistencies within the job category and work-site state columns.
- Grouped similar categories within the job category column.

The cleaned training dataset comprises of 2597278 rows with 80 job categories and 50 states. Desktop with Intel i5 processor operating on a 64bit windows machine with 8 gigabytes of installed random access memory was used for developing the statistical models. Limitations on computing capabilities compounded with the size of the training data made it imperative to perform feature engineering and data manipulation to ensure that the computation is feasible at the same time the model is able to explain the trends observed in the data. Three component datasets were created from the master training dataset, which comprises of petition description and outcomes from 2011-2016.

The first dataset involves creation of new job and state categories. Four new job categories were created based on the percentage of petitions certified by UCIS. The first category has more than 96% acceptance, the second category has an acceptance rate between 91% and 95%, the third category has job categories with an acceptance rate between 86% and 90% and the fourth job category comprises of job categories that have an acceptance rate less than 86 percent. Similarly, two new categories of work-sites were created. The first category comprises of states that have an acceptance rate above 96% and the second category has states with acceptance rates less than 96%. The second dataset comprises of 12.5% of the entire training dataset. The data points were picked randomly from the training dataset. 98.8% of the training data comprises of petitions that have been certified. Hence, there is an inherent bias towards the petition being accepted due to the skewed training data. In this scenario, there is a higher probability for the statistical model to predict the outcome of a petition as accepted than the actual test scenario. The third dataset was created by having equal number of certified and denied petitions for each new state category as created in the first dataset.

Model Development

Dependent Variable

The aim of this project is to evaluate the outcome of each H-1B visa petition based on the characteristics of the job and geography. The outcome has been modeled as a binary variable named certified, which takes the value of one if the visa petition has been certified and zero if the petition has been denied.

Control Variables

In order to account for confounding factors affecting the outcome, we have considered a variety of variables that can be used as proxies to mathematically represent a petition. Full time is a dummy variable that takes a value of one if the job under consideration is full-time and zero otherwise. State categories have been included to account for geographical idiosyncrasies associated with acceptance of an application. Similarly, job categories have been included to factor professional seniority of an applicant. This project also considers employee salary, which is a quantitative variable. Salary is an important parameter used to determine whether the international employee possesses highly specialised skills for the particular job.

Model Description

Nine different statistical models were used to predict the outcome of an H-1B visa petition. Logistic regression, support vector machine, and K-nearest neighbors were trained using the three datasets leading to 9 models in total. The aim was to minimize the number of false positives so that the final model does not classify a petition as certified when it was supposed to be denied. The final ensemble model of all the component models would have a result as certified only when at least 7 out of 9 component models have certified as their answers. The dataset on H-1B visa applications for 2017 was used as the testing dataset. The ensemble model achieved an accuracy of 88.5%.

6 DISCUSSION AND CONCLUSION

With this interactive and user-friendly web application, international students can be more informed about their chances of obtaining an H-1B visa based on annual wage, type of position (full time versus part time), job category, and work-site state. The predictive model is robust because it combines various classification algorithms (logistic regression, support vector machine, and k-nearest neighbors) and corrects for bias towards certified H-1B visa applications. In addition to providing international students with a data-driven approach for seeking employment within the United States, the popular use of this application will likely facilitate communication among users and thus bring about a sense of community. The predictive model can assist decision makers to automate evaluation of H-1B petition. Thereby, saving them financial and human resources that can be used for other important activities. This model will also help the companies to estimate if the particular employee will be granted the visa. This will enable the company to allocate their international talent pool across geographies where the chances of visa acceptance will be higher. Although this project considers the major confounding factors impacting the H-1B visa decisions, there are opportunities for future work. Other classification methods such as naive Bayes, neural networks, and/or random forest can be used. Robustness of the model can be improved by incorporating company details, city locations, and exact job titles in the predictive models.

REFERENCES

[1] Catalina Amuedo-Dorantes and Delia Furtado. 2016. Settling for Academia? H-1B Visas and the Career Choices of International Students in the United States. (2016). http://jhr.uwpress.org/content/54/2/401.short

[2] Avinash Anantharamu. 2014. OneCareer-A Visualization approach to job search process. (2014).

[3] Dhanasekar Dhanasekar, Palm Nabarun, and Misraa Aashish. 2017. An analysis of nonimmigrant work visas in the USA using Machine Learning. *International Journal of Computer Science and Security(IJCSS), Vol. 6* (2017).

[4] Susan Gasson and Katherine M Shelfer. 2007. IT-based knowledge management to support organizational learning: Visa application screening at the INS. *Information Technology & People* 20, 4 (2007), 376–399.

[5] Beliz Gunel and Onur Cezmi Mutlu. [n.d.]. Predicting the Outcome of H-1B Visa Applications. ([n. d.]).

[6] Lesleyanne Hawthorne. 2010. "Chapter 5: Demography, migration and demand for international students" in Globalisation and tertiary education in the Asia-Pacific: The changing nature of

a dynamic market. (2010). https://books.google.com/books?hl=en&lr=&id=ee5pDQAAQBAJ&oi=fnd&pg=PA93&dq=USA+international+students+stem+job+rates&ots=RAIa07EZ0R&sig=6UyKJN5SwI9Yu6Gs1clc3m_IQBI#v=onepage&q&f=false

[7] Kelsey Robinson John Valentino Richard Dobbs James Manyika, Susan Lund. 2015. A LABOR MARKET THAT WORKS: CONNECTING TALENT WITH OPPORTUNITY IN THE DIGITAL AGE. (2015).

[8] Alishah Dholasaniya Habeeb Hooshmand Jonathan Arauco, Bhavik Bhatt and Joseph Martinsen. 2018. An Exploration of H-1B Visa Applications in the US. (2018).

[9] Ethan Lewis and Giovanni Peri. 2015. Chapter 10 - Immigration and the Economy of Cities and Regions. 5 (2015), 625–685.

[10] Renchi Liu and Jinglin Lu. 2017. H-1B Visa Data Analysis and Prediction by Using K-Means Clustering and Decision Tree Algorithms. (2017).

[11] Peri Shih Sparber Mayda, Ortega. 2017. The Effect of the H-1B Quota on Employment and Selection of Foreign-Born Labor. *National Bureau of Economic Research* (2017). http://www.nber.org/papers/w23902

[12] Giovanni Peri, Kevin Shih, and Chad Sparber. 2013. STEM Workers, H1B Visas and Productivity in US Cities. *IDEAS Working Paper Series from RePEc* (2013). http://search.proquest.com/docview/1698809799/

[13] Sneha J. Pitre. 2017. International Students Career Development: Acculturative Stress and Career Outcomes. (2017).

[14] Kevin Shin. 2015. Labor Market Openness, H-1B Visa Policy, and the Scale of International Student Enrollment in the United States. (2015). https://onlinelibrary.wiley.com/doi/abs/10.1111/ecin.12250

[15] Pooja Thakur, Mandeep Singh, Harpreet Singh, and Prashant Rana. 2018. An allotment of H1B work visa in USA using machine learning. *International Journal of Engineering Technology* (2018). https://www.sciencepubco.com/index.php/IJET

[16] Patrick S. Turnet. 2017. High-Skilled Immigration and the Labor Market: Evidence from the H-1B Visa Program. (2017). http://www.sites.google.com/a/colorado.edu/psullivant/Turnerh1b.pdf

[17] Sharmila Vegesana. 2012. Predictive Analysis for Classification of Immigration Visa Applications: A Discriminative Machine Learning Approach. (2012).