PPOL564 Final Project

Tianhui Cao

Word count: 2996

1 Introduction

H-1B visa is a non-immigrant temporary working visa essential for foreign workers to be lawfully employed in the US. Policy debates surrounding this area is fluctuating at this point. On the one hand, policymakers are tightening regulations to avoid employers from exploiting cheap external labor force that squeezes out local job applicants. But, on the other hand, bringing in high-skilled foreign workers should be a complement to industries where there is a gap in the U.S. workforce.[10]

This data science project is an effort to explore the truth behind these two conflicting opinions. We will approach this problem by understanding the profile of direct stakeholders of the potential policy change: H-1B visa applicants. Specifically, we will look at the Labor Condition Application (LCA) disclosure data from the United States Department of Labor (DOL). LCA certification is a prerequisite for H-1B visa approval that covers applicant-level information such as employers, locations, job titles, base salary, and case status. We will decompose the characteristics of the applicants in the year 2018 as a snapshot of the issue by investigating popular cities, job titles, base salaries and try to understand the factors that are predictive of an LCA certification.

This report will cover four sections: 1) problem statement and background 2) data collection and wrangling 3) methodology for analysis 4) results 5) discussion. In the first section, we will provide a concise overview of recent H-1B policy clashes and stakeholders. Following that, we will explain our data collection process as well as the underlying logic. In section 3) and section 4) we will unfold the story behind the data and finally, we will cover the policy implications in section 5).

The main findings can be concluded into three dimensions. First, companies with a large scale of business will sponsor more H-1B visa applicants. Applicants from those companies in turn will have a better chance to have a certified LCA. Second, the absolute amount of base salary is not predictive of an LCA certification. Finally, consulting, accounting, and architecture are three industries beyond the technology field whose applicants have the best chance to get LCA certified. These insights are critical to understanding the new policy proposals on the H-1B visas.

2 Problem Statement and Background

According to USCIS , an H-1B visa approval hinges on four criteria including specialty occupation, prevailing wage, employee qualification, and H-1B quota.[11]These four criteria aim at preventing H-1B visa to be an outsourcing vehicle where employers can exploit lower labor prices. In an essence, if foreign workers with any level of skills can easily enter the US job market and they can provide time and skill at a lower price, American workers will gradually be squeezed out of the job market because they are not the economic choice for employers. Whether from patriotism or the economic standpoint, such a job market will be unacceptable. In response, the Trump administration has been working on tightening immigration regulations to pressure employers only to

hire a limited number of high-skill foreign workers. But it looks like the definition of "high skill" has been pushed beyond the threshold, thus having a chance to bring side effects.

H-1B visa application is already a challenging process and the Trump administration's new proposal is only making it even more daunting. One of the fundamental requirements for an H-1B visa candidate is to hit the prevailing wage line depending upon both levels of the position and the location of the company. Trump administration has proposed to raise the requirement for level 1 position from 17th percent to 45th; level 2 position from 34th to 62nd; level 3 position from 50th to 78th; and level 4 position from 67th to 95th percentile in a released document on Oct.8th.[1] Suppose a new international college graduate with a computer science degree is entering the San Francisco job market, he or she will be most likely to get an entry-level job with a "level 1" or "level 2" label. How hard it will be to reach the average wage among all software engineers in San Francisco! Besides, a narrower definition for "specialty occupation" is adding another layer of challenge. For example, the Occupational Outlook Handbook by the US Bureau of Labor Statistics suggests a typical entry-level computer programmer position requires a bachelor's degree, but some may accept an associate's degree.[2] Under the new proposal, it won't be a specialty occupation for applicants with a bachelor's degree because of the latter part in the description. To prove the position meets the "specialty occupation" requirement, many will choose to find a level 2 job, which is harder to get, and with even higher wage requirements. Finally, the lottery aspect is controversial in the H-1B visa process to address the issue that the number of applicants is well above the 85000 applicants' cap.[7] With these challenges, many talented young people are screened out of the US job market, and the issue is intertwined with other areas such as recruiting international students to US higher education

institutions.

This project is an attempt to explore the historical data of LCA applicants to understand their profile, as well as to build a model to predict LCA certifications in 2018. The insights should be helpful to stakeholders such as international students, US companies, and policymakers to evaluate the implication of recent proposals.

3 Data Collection and Initial Preprocessing

As mentioned in the previous section, the data is applicant-level information from the LCA data set. The official disclosure is from the Office of Foreign Labor Certification(OFLC) under the Department of Labor(DOL).[2] However, downloadable tables on the related pages are very large and will be computationally expensive to conduct data analysis tasks on. As an alternative, the data source I'm extracting is a site named *H1B data info*. Records on this site are indexed from the official source and includes more than 3.8 million data entries of LCA applicants between 2012 and 2020. Variables in the tables include employer, job title, base salary, location, submission date, start date, and case status. Visitors can query the information by searching city, company, or year on the main page.[11]

In so far as the project is mainly concerned with application packages for the LCA certification, variables of interest are case status, employer, location(city), base salary and job title reflected by the original data set, and length of preparation period indicated by the gap between submission date and start date. Notice the H-1B selection process begins in April when employers start to report LCAs to DOL; employees can't start the job until October if not applying for the cap exemption program. In terms of the scope of the research, I decided to limit my analysis to the Top 50 cities with most *accumulative* filings over the last 9

years following the cities page. Besides, I restricted the time frame to the year 2018 because of data availability issues. As a reference, the cities page provides the full link of 2000 cities covering all 3.8 million entries.

To construct the main data frame, there are two general steps: web scraping and data preprocessing. For the first step, I used the "request" package to download the relevant pages, the "BeautifulSoup" module to parse the page, and the "Pandas" package to read the tables.[6][4][9] After scraping 50 links corresponding to the Top 50 cities, I obtained a concatenated data frame with 289871 observations with only 4 null values. It's a desirable sample size signals a good start for our analysis. However, the imbalanced proportion of certified cases (95.75%), compared with other uncertified cases (certified-withdrawn: 0.15%, denied: 1.21%, withdrawn 2.89%) brought extra obstacles for prediction tasks. We will downsize the majority class to keep a balanced data frame. Details will be discussed in section 4. At the same time, to explore our data visually, we will use the version that contains the original sample size. In addition to concerns about sample size, correct data types and information are also important. I wrangled the data by using "Pandas" and "Numpy" packages to achieve these goals.[6] [4] In practice, the "job title" variable includes messy codes in raw data. So, I cleaned the column to keep only the alphabets and spaces. Finally, a new variable indicating the length of the preparation period is created. Strings in the "submission date" and "start date" columns are converted to the "DateTime" type of variable, and the difference between the dates is calculated and saved in a new column. Now we have a complete and cleaned data frame with variables of interest. Another data frame that complements the analysis is the trend for a selected tech company. Many technology-related jobs have been supplemented by external labor. I took Amazon as an example to investigate the trend in so far as the *H1B data info* website holds the record. Specifically, I'm focusing on

the number of applicants and the median base salary from 2012 to 2020. The scraping and preprocessing methods are very similar to how I constructed the main data frame.

4 Methodology for Analysis

This project has two sections including exploratory data analysis and machine learning.

In the exploratory data analysis part, the goal is to visually present rankings, distributions, word frequencies, or trends in terms of different levels of information. Intuitively, we need to first process our data, and then present the information by plots. Packages utilized in the data processing step are "Pandas" and "Numpy".[6][4] Plots are mostly generated by the "Matplotlib" package and "Plotnine" package. [5]For a ranking task, we will use "group by", aggregate methods, and "sort values" functions provided by the "Pandas" package in a method chain. For a distribution task, we directly plot histograms or grouped boxplots as needed by using the "Plotnine" package. For a word frequency task, we will use the "WordCloud" package to illustrate popularity. Finally, for a trend task, we will use the "Matplotlib" package to generate line plots.[5]

In the machine learning part, we aim to examine the factors that are predictive of an applicant's obtaining an LCA certification. Since this is a classification task, we need to transform our data set, compare models, test performance, and finally, interpret the results. The packages we utilized are "sklearn" and "pdpbox".[8][3]

The first challenge for our data set is the imbalanced target variable of case status. To tackle this issue, we downsized the majority class by using the resample method provided by "sklearn".[8] In a nutshell, we are producing a sample without replacement for the majority class(certified) to keep the number

of observations the same as the minority class(uncertified). As a result, we examined 12320 observations for each class.

Following that, we are constructing the features based on insights from exploratory data analysis. Features selected are top companies, base salary, job titles, technology industry, and length of the preparation period. Except for base salary and length of preparation period, we are using dummy variables as representations. For example, the transformed data frame contains a column that indicates whether the company is in the top 100 company list of the highest number of LCA applicants. We applied similar logic to construct "job20", which indicates whether the job title of the applicant contains any of the 20 most frequent words among all applicants' positions, and finally, we construct a dummy variable that signals whether the applicant's position matches a list of words indicating the technology industry. Besides, we screened out outliers of applicants with a base salary greater than 500k.

Models included in the machine learning part are Naive Bayes, KNN, Decision Tree, and Random Forest. We used train-test split, cross-validation, tunning parameters, and grid search methods to ensure the best result and minimize the influence of randomness. As we already balanced the classes, the ROC AUC score will be the metric to evaluate the model performance. Comparison between models is conducted by the "sklearn" package.[8]

Finally, for a more interpretable result, we are using permutation functionality in "sklearn" to evaluate variables' importance. In other words, if we scramble the order of the variable of interest, how much will the model performance score change. Besides, we are using partial dependency plots(PDP) and individual condition expectation(ICE) plots generated by the "pdpbox" package to investigate the marginal effects of the most predictive feature.[3] At the end of our project, we are drawing insights from a global surrogate model where we put the

predictions in a decision tree. After we follow the criteria indicated by the tree, we can identify applicants with the best chance to get an LCA certification.

5 Results

5.1 Exploratory Data Analysis

Top companies, cities, and job titles

On the company level, Deloitte, Tata, and Cognizant are the top 3 companies with the highest number of LCA applicants in 2018. The following graph demonstrates a full list of the top 100 companies. As expected, technology companies have a large share.(See Figure 1)

On the city level, the most popular destination for LCA applicants in 2018 was New York, followed by San Francisco and Chicago. It is no surprise that New York as one of the financial centers in the U.S. has a glamour to attract external labor.(See Figure 2)

On the position level, the most frequent word that appeared in the applicants' job title is "engineer". As shown in the word cloud below, words that indicate the technology industry such as software, analyst, and system are common among LCA applicants.(See Figure 3)

Base salary distribution by case status

According to Figure 4, although reaching the prevailing wage is one of the requirements to get an H-1B visa, there is no dramatic difference in salary distribution by case status. One possibly is the survivor bias due to screening by employers or self-initiated checks by applicants. In terms of case status, we see 95% certified, which reinforces our inference about survivor bias in Figure 5. In the machine learning part, we combined the three categories of "certified-

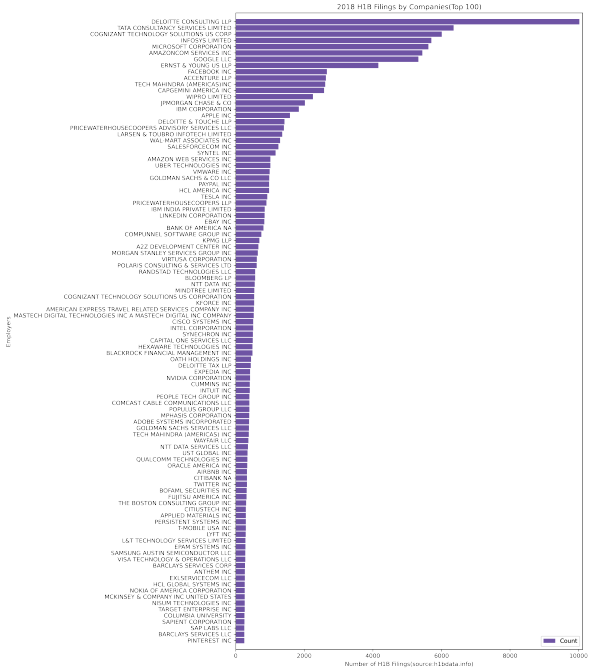

Figure 1: Top companies

withdrawn", "withdrawn" and "denied" as one category of uncertified.

Length of the preparation period

Figure 6 is consistent with the suggested process of an H-1B application with most applicants submit LCA in April (around 180 days towards October) or the same month as they start their jobs (October)

Highest median salary

For all positions, CIO, CEO, President have the highest median salary. However, this insight is not generalizable because of the limited sample size. After we specify the number of applicants to at least 10, the top 3 jobs changed to professors, anesthesiologists, and managing directors.(See Figure 7, Figure 8)

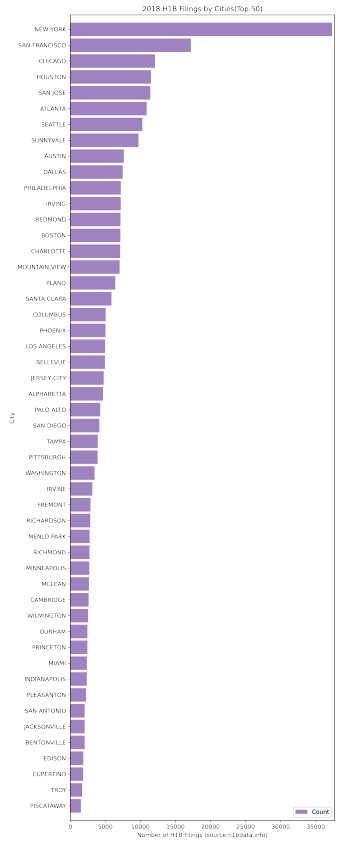

Figure 2: Top cities

The trend for median salary and number of applicants (Amazon 2012-2020)

Figure 9 and Figure 10 indicate a similar trend between the number of applicants and the median salary. Besides, in terms of salary distribution, the shapes are also comparable without noticeable skewness throughout all years, as seen in Figure 11.

Figure 3: Top job titles

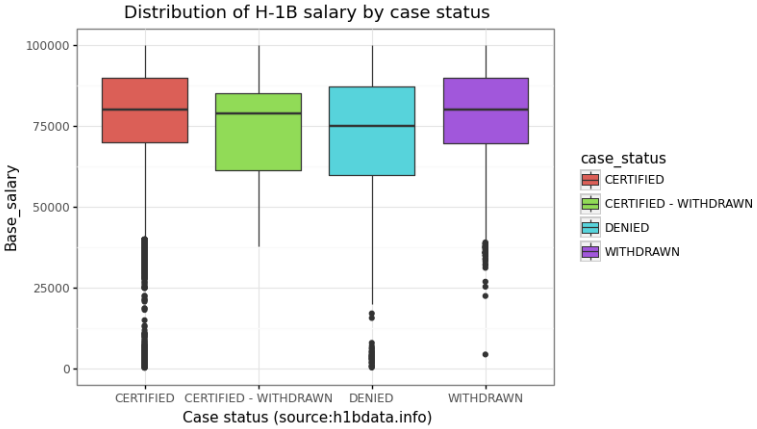

Figure 4: Distribution of base salary by case status

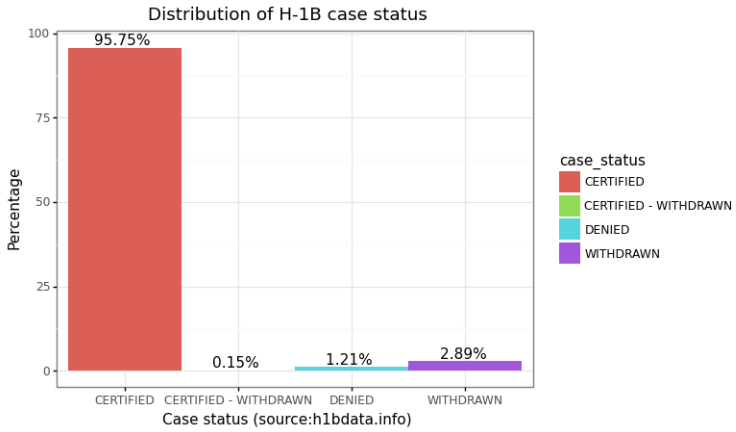

Figure 5: Distribution of case status

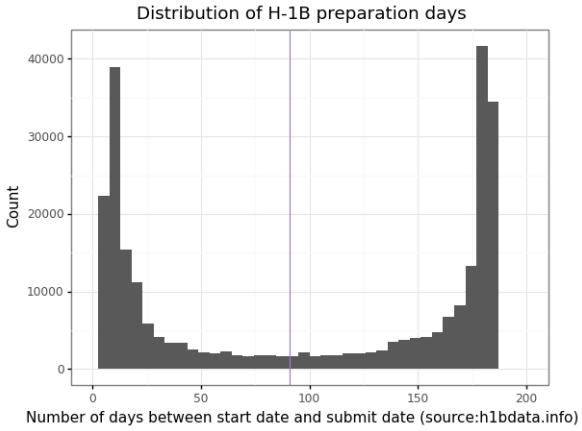

Figure 6: Distribution of preparation period

5.2 Machine Learning

Model

The most predictive model for our analysis is the random forest with a max depth of 4 and 1500 trees grown. The ROC curve is shown below indicating an overall score of around 0.63 for both the train and test data set. Although the performance score is not perfect, we still can get some insights into variable

job_title_clean	count	median_base_salary
CHIEF INVESTMENT OFFICER SOC	1	2134381.0
PRESIDENT CEO	1	1900000.0
CHAIRMAN PRESIDENT AND CHIEF EXECUTIVE OFFICER	1	1100000.0
PRESIDENT REAL ESTATE CHIEF CORPORATE DEVELOPMENT OFFICER	1	1000000.0
CHIEF FINANCIAL OFFICER AND PARTNER	1	992250.0
KFC GLOBAL CEO	1	850000.0
GLOBAL CHIEF MARKETING OFFICER	1	785000.0
EXECUTIVE VICE PRESIDENT CORPORATE MARKETING	1	700000.0
CHIEF SERVICES OFFICER	1	675000.0
SURGEON RESEARCHER	1	670000.0
GPC PARTNER	1	650000.0
PROFESSOR OF PEDIATRICS CLINICAL SCHOLAR	1	610000.0
ADVISORY PRINCIPAL	1	600000.0
PRESIDENT AND GM STUDIO OPERATIONS	2	600000.0
ADVISROY PRINCIPAL	1	600000.0
COHEAD OF TRADING	1	576000.0
DISTINGUISHED RESEARCH SCIENTIST	1	557680.0
ASSOCIATE PROFESSOR OF ORTHOPEDIC SURGERY	1	550000.0
TAX PARTNER	1	542000.0
CHIEF OPERATING OFFICER ASSISTANCE SERVICES WORLDWIDE	1	541200.0
VP OF VR	1	525000.0
PROFESSOR AND DIRECTOR OF PEDIATRIC HEPATOLOGY	1	520840.0
GASTROENTEROLOGIST	6	510000.0
ASSISTANT PROFESSOR OF CLINICAL NEUROSURGERY	1	506000.0
IT TRANSFORMATION MANAGER 5	1	500000.0
EXECUTIVE DIRECTOR OF ATHLETIC PERFORMANCE AND SPORTS MEDICI	1	500000.0
SPECIALIST ORAL SURGEON	1	500000.0
HEAD OF COMMODITIES	1	500000.0
MANAGING DIRECTOR HUMAN RESOURCES	1	500000.0
GLOBAL CHIEF TALENT OFFICER	1	480000.0

Figure 7: Top median salary for all applicants

job_title_clean	count	median_base_salary
ADJUNCT ASSISTANT PROFESSOR	17	275000.0
ANESTHESIOLOGIST	11	272000.0
MANAGING DIRECTOR	52	240771.0
STAFF PHYSICIAN	18	240000.0
PARTNER	31	235893.0
ASSOCIATE PARTNER	13	229468.0
HOSPITALIST	79	223200.0
EXECUTIVE DIRECTOR	64	217500.0
HOSPITALIST PHYSICIAN	41	217000.0
ASSISTANT PROFESSOR OF CLINICAL	12	216500.0
PRESIDENT	24	216500.0
SR MTS SOFTWARE ENGINEER	12	214408.0
ADVISORY DIRECTOR	77	214000.0
VICE PRESIDENT INTERMEDIATE FINANCE ASSOCIATE	27	210000.0
PSYCHIATRIST	23	208000.0
VICE PRESIDENT OF ENGINEERING	12	207500.0
ASSISTANT PROFESSOR OF MEDICINE	14	207500.0
NEONATOLOGIST	14	204842.5
CHIEF EXECUTIVE OFFICER	89	200304.0
CHIEF FINANCIAL OFFICER	33	200000.0
VICE PRESIDENT TRADER II	16	200000.0
CEO	13	200000.0
NEPHROLOGIST	32	195000.0
ENGINEERING DIRECTOR	16	194000.0
SENIOR MANAGER RD	12	192500.0
TEAM LEADER	61	192000.0
PRINCIPAL PROGRAM MANAGER	39	190460.0
DIRECTOR OF PRODUCT MANAGEMENT	12	190135.0
DIRECTOR PROGRAMMER LEAD MKTS MGR	11	190000.0
PRINCIPAL SOFTWARE ENGINEERING MANAGER	37	187313.0

Figure 8: Top median salary for more than 10 applicants

importance.(See Figure 12)

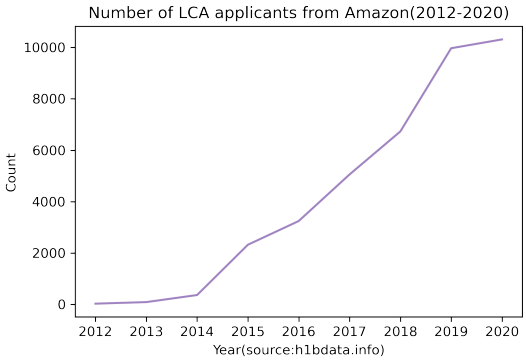

Figure 9: Number of applicants trend for Amazon LCA applicants

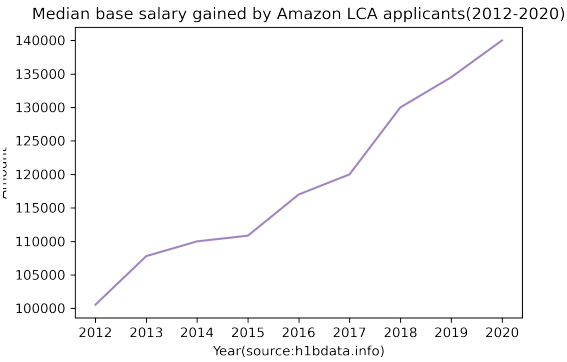

Figure 10: Salary trend for Amazon LCA applicants

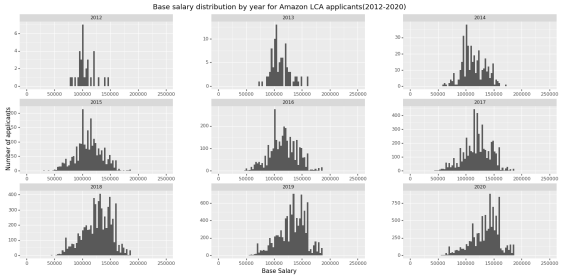

Figure 11: Distribution of Amazon LCA applicants' salaries

Variable importance

After permutation on each feature, the reduction of the ROC AUC score is shown below. Here, we observed our first and second key findings. First,

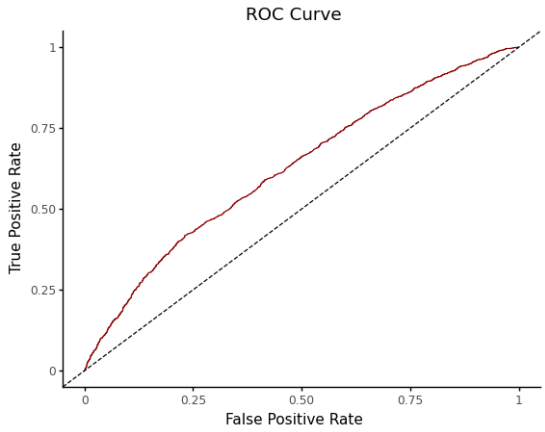

Figure 12: ROC curve for the random forest model

whether the applicant is from the top 100 companies with the most LCA applicants is the most important feature to predict LCA certification. If we randomly assign the value of this feature, we will observe a 0.1 reduction in the ROC AUC score. Second, the absolute number of base salaries is not predictive of an LCA application. After permutation, we didn't observe a significant drop in the performance metric. (See Figure 13)

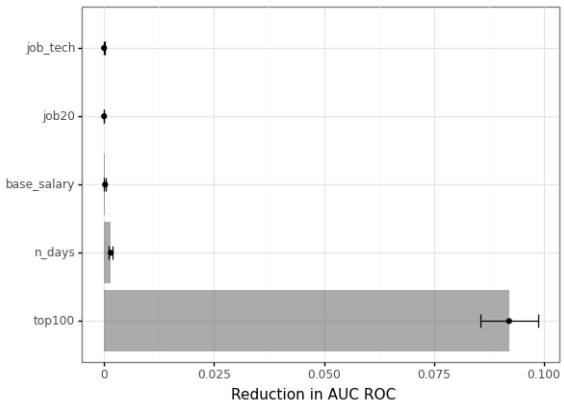

Figure 13: Permutation result

PDP and ICE

The strong positive differential effect of the "top100" variable is demonstrated in the following interaction PDP and ICE plots. Applicants with jobs in "top 100 companies" have a significantly higher chance to get LCA approved. One counterintuitive insight in these graphs is that non-technology jobs have a higher chance to be approved than its counterpart. We may further discuss the reason after we use the global surrogate model. (See Figure 14-16)

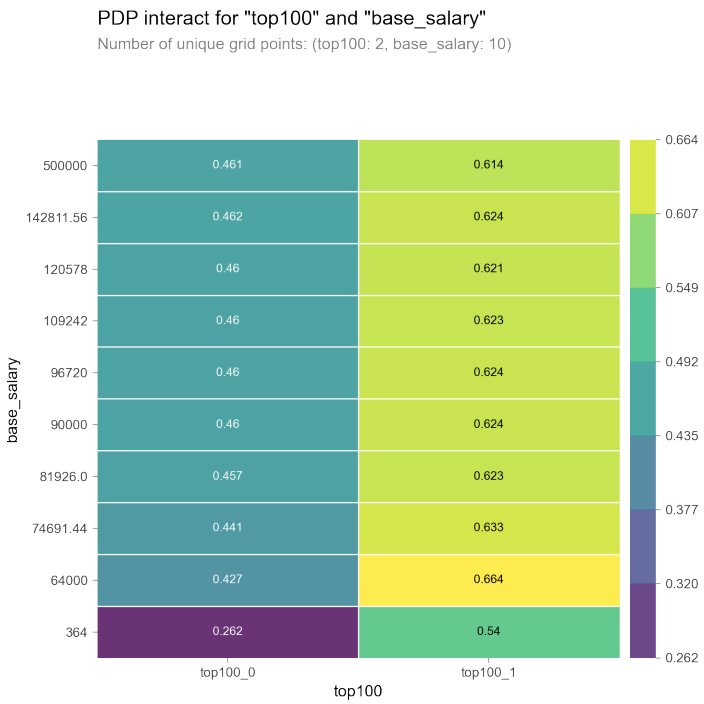

Figure 14: PDP Interaction between base salary and top 100 companies

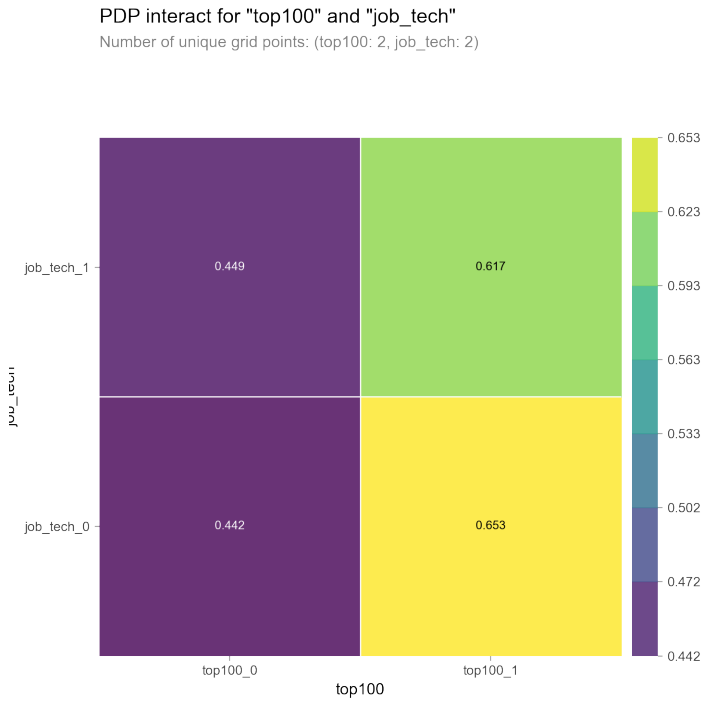

Figure 15: PDP Interaction between tech-related job and top 100 companies

Global surrogate model

The decision tree maps out our model predictions. Notice the best chance is 0.687 characterized by applicants from top 100 companies, with non-tech positions, lower than 74257 annual base salary. Following this route, we queried the original data set. Consulting, accounting, and architecture are three areas with most applicants meeting the requirements. The intuition behind it is the prevailing wage. If the non-tech position is offered by a top 100 company, the baseline of the salary could be high. Compared with the prevailing wage, there

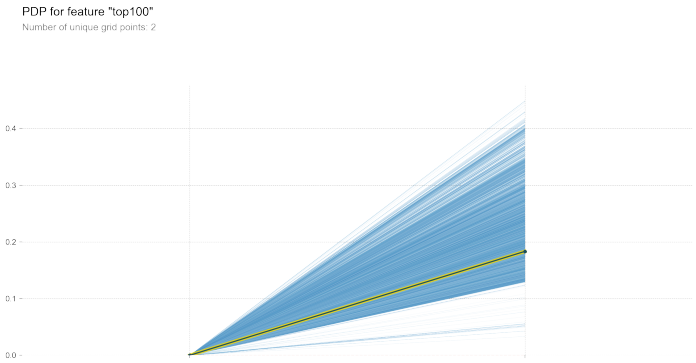

Figure 16: ICE plot

is a huge surplus. However, for tech-related jobs, the prevailing wage is already high. Entering a top 100 company won't ensure a certification. (See Figure 17,18)

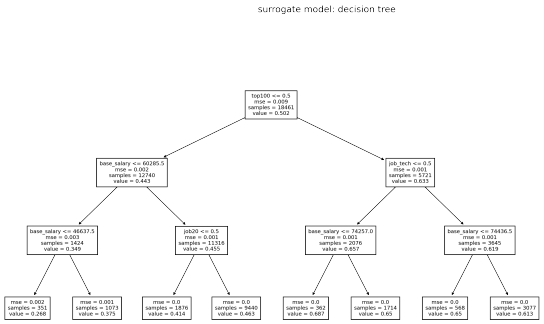

Figure 17: Decision tree of predictions

job_title_clean	count
CONSULTANT	2578
ACCOUNTANTS AND AUDITORS	420
ARCHITECT	387
ADVISORY CONSULTANT	219
TAX CONSULTANT	171
ASSURANCE ASSOCIATE	162
OPERATIONS RESEARCH ANALYSTS	162
ASSOCIATE	153
SENIOR CONSULTANT	140
AUDIT ASSURANCE ASSISTANT	138

Figure 18: Best chance for non-tech positions

6 Discussion

This project provides a snapshot of the H-1B visa issue. We observed some important trends such as most applicants are in the technology field, but non-tech positions have a better chance to get LCA certified. This is thought-provoking to design policies that interpret the real gap in the US job market. However, the scope of the project is limited. First, we set a cut off for accumulative number of applicants for cities we analyze. But in reality, smaller cities such as DC are also important to consider. We can expand our analysis to a longer time frame. Second, we have arbitrary definitions on technology industry. We provided a list of words, but a better practice may be unsupervised machine learning. Finally, we can map the new prevailing wage suggested by policy proposals on historical data to simulate the effect. This will indicate a more extensive project because we need a stricter definition on locations, industries and positions.

References

[1] Employment and Training Administration. Strengthening wage protections for the temporary and permanent employment of certain aliens in the united states, October 2020.

[2] U.S. Department of Labor Bureau of Labor Statistics. Computer programmers : occupational outlook handbook: : U. S. Bureau of labor statistics.

[3] Brandon M. Greenwell. pdp: An r package for constructing partial dependence plots. *The R Journal*, 9(1):421–436, 2017.

[4] Charles R. Harris, K. Jarrod Millman, St'efan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fern'andez del R'ıo, Mark Wiebe, Pearu Peterson, Pierre G'erard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[5] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[6] Wes McKinney. Data structures for statistical computing in python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56, 2010.

[7] Sara Ashley O'Brien. Tech's beloved H-1B visa is flawed. Here's why., February 2017.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Pas-

sos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[9] Leonard Richardson. Beautiful soup documentation. *April*, 2007.

[10] Andy J. Semotiuk. New fast-tracked U. S. Regulations threaten to restrict h1b visas again.

[11] USCIS. H-1b specialty occupations, dod cooperative research and development project workers, and fashion models | uscis, December 2020.