Road to H1-B

Ann Tsai
Courant Institute of
Mathematical Sciences
New York University
New York, New York
mt4050@nyu.edu

Ko-Chun, Chiang
Courant Institute of
Mathematical Sciences
New York University
New York, New York
kc4200@nyu.edu

Jiancheng Wang
Courant Institute of
Mathematical Sciences
New York University
New York, New York
jw5865@nyu.edu

***Abstract*— While researching the H1-B acceptance rate, we found that past related studies only focused on the prediction of H-1B results, no work has been done on providing a recipe for international students about what they should learn at school to achieve the highest H1-B acceptance rate. In this paper, we explore what can be learned about H1-B applications from H1-B employer data, Linkedin Job Description, and H1-B salary dataset. We utilize Spark to process our data, statistical methods to verify if the higher salary you have, the higher H1-B acceptance chance you get, and we predict the H1-B acceptance rate by Machine Learning(ML) model. Except for existing inputs[1], we add some new inputs such as Salary and Initial Approvals to our model. Lastly, we visualize our result and develop an application, which allows users to input their skills, to see what kind of jobs they can do, and the corresponding H1-B acceptance rate.**

Keywords*— *Machine learning, Statistical analytics, Data analysis, Analytical models, Data visualization, Apache Spark, H-1B working visa

I. Introduction

H-1B is a type of non-immigrant visa that allows US employers to hire foreign workers in specialty occupations that require theoretical or technical expertise in specialized fields such as IT, finance, accounting, engineering, science, medicine, etc. The job must meet some criteria to qualify as a specialty occupation such as minimum entry requirement for the position as bachelor's or higher degree or its equivalent. Additionally, the employer should attest that it will pay the beneficiary a wage which is no less than the wage paid to similarly qualified workers or higher than the prevailing wage for the position in the geographic area.

There are about a million international students in the U.S. per year, and H-1B is the most commonly used working visa after one to three years of OPT (Optional Practical Training). The USCIS (US Citizenship and Immigration Service) grants 85,000 H-1B visas per year which includes 20,000 additional visas for workers with advanced degrees (i.e., MA, PhD) from an accredited U.S. academic institution.

However, most studies [1][2][3] focused on the prediction of H-1B results based on the application data which provided limited information for students. This paper applies empirical studies on how students can maximize their opportunities to get H-1B visas. To provide students a clearer guide to increase the possibility of getting H-1B visas, this paper will analyze the 10-year trend of H-1B acceptance rate for the top ten H-1B petition companies that hire computer science students. Additionally, this paper will analyze the salary of popular positions for computer science graduates and the skillsets of the positions. With the result, computer science students could have a clearer view of what skills are needed to get into a company that has a higher H-1B acceptance rate and a competitive salary.

This paper will also build a prediction model using the skill sets dataset, salary dataset, and company H-1B acceptance dataset. Also, the predictor and analytics result would be on our website, where students could leverage it for their course selection or career development.

II. Motivation

Computer science has become the most popular degree in the United States, especially for international students due to the high volume of positions and salaries in a career of the technology industry. However, students in computer science are confused about what skills or programming languages they should invest their valuable time in. Accordingly, we would like to provide analytics results among different skills and programming languages for computer science students to maximize their career.

The rapid development of the tech industry has spawned a vast number of skill sets and we want to figure out which subset of skills make great contributions to the salary, and indirectly impact the H1-B. For new grads, especially those just switched major to computer science in their last one or two years in academy, they can use this insight to pick up the skills that are most cost-efficient and rule out the skills that are no longer popular in the industry. For experienced developers, they can also cite the result when they want to challenge themselves into getting a higher pay.

III. Related Work

First, the only thing we know about salary and H1-B application is that you meet the application condition requirement as long as your salary reaches the prevailing wage. However, we want to verify if the higher salary you have, the higher H1-B acceptance chance you get. Jena, Olenski, and Blumenthal [4] applied Sensitivity Analysis to analyze the relationship between genders and salaries of physicians. This inspired us to use the same method to verify the relationship between salary and H1-B acceptance rate. Besides, we also utilize linear regression models to get the regression equation. Lastly, we will utilize hypothesis tests to verify that the result we get is credible.

Ke and Qiao [1] applied machine learning on H-1B case data. The H-1B case data is the application data of H-1B cases from 2008 to 2018 with attributes including wage, date of application, employer, location, economic title ,and citizenship. Lasso regression was applied to examine the impact of different factors with R^2 : 0.68, and logistic regression with L1 penalty was applied to build a predictor with AUC: 0.676. The result showed that applicants working in the healthcare industry or majored in healthcare-related majors usually have the highest wages. In this paper, we would apply the same methodology to examine which programming language and skills are most impactful to H-1B acceptance.

IV. Datasets

A. H1-B employer dataset

The dataset is obtained from the U.S. Citizenship and Immigration Services (USCIS) of the Department of Homeland Security which provides the information on employers petitioning for H-1B workers from 2009 to 2019 [5]. The size of the dataset is 45 MB with around 61 thousands of petitions from 28 thousands employers. Table 1. shows the schema of the dataset.

Table 1. Schama of H-1B employer dataset

Fiscal Year	Int	The fiscal year in which USCIS first recorded an approval or denial (adjudicated date) in the electronic systems. A fiscal year covers Oct. 1 of one year to Sept. 30 of the next year.
Employer	String	Petitioner's firm/employer name.
Initial Approvals	Boolean	H-1B petitions with "New employment" or "New concurrent employment" whose first decision is an approval.
Initial Denials	Boolean	H-1B petitions with "New employment" or "New concurrent employment" whose first decision is a denial.
Continuing Approvals	Boolean	H-1B petitions with anything other than "New employment" or "New concurrent employment" whose first decision is an approval.
Continuing Denials	Boolean	H-1B petitions with anything other than "New employment" or "New concurrent employment" whose first decision is a denial.
NAICS	Int	North American Industry Classification System Code: A character string that stands for an industry classification.
State	String	Petitioner's state. This is the State indicated in the mailing address of the employer and is not necessarily the beneficiary(ies) work location.

B. Linkedin Job Description dataset

The job descriptions are scraped from Linkedin.com [6] with keyword searches: roles set to "Software Engineer", "Data Scientist", etc.; locations set to "San Francisco", "New york", etc. The dataset is further normalized and job skills are extracted with hand-written rules.

Table 2. Schema of Linkedin Job Description dataset

employer	String	Name of the employer: "Facebook", "Google", etc.
job_titles	String	Including software engineer, web developer, ML engineer, data scientist, and data analyst
skill_sets	Int	Bitmask where each bit represents if a certain skill is required: Java, Python, C++, etc.

degree	Int	Bitmask where the first three LSB represents if BS/MS/PHD degrees are required in this role

C. H1-B salary dataset

The dataset is obtained from the up-to-date official H1-B data disclosed by the United States Department of Labor [7], which provides information on salary, company, job title, location, … and so on, from 2012 to 2019. The size of the dataset is 36 MB. Table 3. shows the schema of the dataset.

Table 3. Schama of H1-B salary dataset

year	Int	H1-B application submit year. Software engineer / Data analyst: 2012 ~ 2019, Data engineer : 2013 ~ 2019, and Web developer / Data scientist : 2014 ~ 2019.
employer	String	Petitioner's firm/employer name
job_titles	String	Including software engineer, web developer, data engineer, data scientist,and data analyst
salary	Int	annual salary[range : $923 ~ $1350001]
visa_status	String	WITHDRAWN/ CERTIFIED/DENIED.
state	String	Petitioner's state. This is the State indicated in the mailing address of the employer and is not necessarily the beneficiary(ies) work location
month	Int	H1-B application submit month

V. Description of Analytic

(Describe the analytic, which is the back-end of your application. What are the findings? What actionable insights does it provide?)

H1-B acception

To understand the H1-B acceptance status, we analyzed the H1-B employer dataset using Spark in Scala. First of all, we cleaned the data by removing null, duplicate and typing error rows (6521 rows in 622684 rows). Since we would like to provide information for students who have not had H1-B yet, we would neglect "Continuing Approvals/Denials" which is for the people who switched employers. Among "Initial Approvals/Denials", we used the "groupBy" and "pivot" function to get the H-1B approval rate and counts from 2011 to 2019 among each "State", "Employer", and NAICS (North American Industry Classification System Code). For "State" and "Employer" analysis, we concentrated on those jobs in the field of Professional, Scientific, and Technical Services, so we used a "filter" function to select those employers with NAICS: 54.

Programming Language and Software Skills

Each text file was first processed with two preprocessing stages -- tokenization and lower case conversion. After that, we applied rule-based NLP approaches to extract the keywords: programming languages and software skills. To be exact, we wrote a regular expression suite to cover different scenarios and irregular cases. An example of the issues we observed is when extracting language "C", we need to differentiate it from "C++", or letter "C" embedded in a word such as "conference". This was resolved by adding restrictions to the Regex that either "C" shows up in the beginning of the sentence, or there need be one or more stop words/white space before it. The similar restriction applies to the position after it. Nine most popular languages are extracted such as "C", "C++", "Java", and they are not mutually exclusive with regard to a single job qualification. We used a bitmap to store the intermediate result. For example, 0x0001ff indicates that all nine programming languages are mentioned in this job post.

In addition, top-trend software skills are extracted in the same fashion. We first generated candidates in each field of software engineering. For example, ["Django", "Flask", "Spring", ".Net", "Express.JS", "Nod.JS", "Ember.JS", etc.] are spawned as the candidates in the field of "Back-end framework". We then applied map-reduce to count the occurence of each skill specified in the job description and only selects the top 1/2 popular frameworks in its own language. Other fields of software engineering are "Front-end", "AWS", "Container", "Big Data", "Database", and "DL/ML".

We further used aggregate methods to analyze the popularity of different skills and their impact on salary.

Join

H1-B employer dataset is joined with other two datasets with "company name". Linkedin Job Description dataset and H1-B salary dataset are joined with the primary key of (company name, job_title). Company names are written in different formats in all our three datasets and therefore need to be taken care of. An example for normalization is converting "Amazon Web Service Inc." to "Amazon". We did not handle irregular words, since it rarely shows up in our datasets. We also did not deal with subsidiary relationships, for example, “A9” is a subsidiary of “Amazon”, as this cannot be solved without knowledge graph. The company name is put through a pipeline, built with regular expression patterns, in which five categories of noisy words are filtered out, including stopwords, location, industry identifier, and etc., based on our observation.

Salary and H1-B acceptance with Programming Language and Software Skills

To understand the salary and H-1B acceptance difference among different programming languages and software skills, we join the dataset as the previous paragraph described. Spark dataframe was used to analyse and manipulate data. Spark dataframe functions were heavily leveraged including groupBy, UDF, explode, and withcolumn, etc.

Salary and H1-B acceptance rate

Objective:

To verify if the higher salary you have, the higher H1-B acceptance chance you get, we compare the median of salary per year, state and employer with corresponding H1-B acceptance rate.

Scrapping:

We scrap salary data from the H1-B Visa Salary Database. We only considered job titles related to SOFTWARE ENGINEER, WEB DEVELOPER, DATA ENGINEER, DATA SCIENTIST and DATA ANALYST.

Cleaning:

First, we clean the job titles, we replace SOFTWARE ENGINEER II, JUNIOR SOFTWARE ENGINEER, …… to SOFTWARE ENGINEER, and for other job titles, so on and so forth.

Calculating:

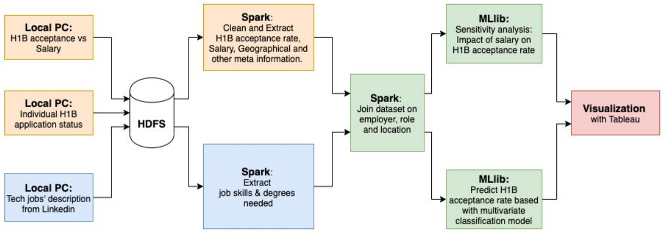
To verify if the higher salary you have, the higher H1-B acceptance chance you get, we calculate the median of salary per year, state and employer. First, we choose to use median rather than average because we want to avoid being affected by outliers. Second, we calculate median from year, state, and employer because year can represent economic situations, which might affect salary and h1-b acceptance rate, besides, same state and same employer are supposed to have similar salary ranges.

Analyzing:

We join the salary data, which has year, state, employer, median of salary with H1-B acceptance rate. We use mllib to build a linear regression model and get the linear regression equation, which has median of salary as independent variable, and H1-B acceptance rate as dependent variable. If we could find an equation with positive slope, and we can prove this result is significant by hypothesis testing , we can infer that when salary goes up, H1-B acceptance rate also becomes higher. Besides, we also use sensitivity analysis to see when the median of salary changes, how much the level of H1-B acceptance rate will be changed.

VI. Application Design

(Paste and explain your design diagram(s) here. Include a screenshot(s) of your visualization or UI.)


Local PC: H1B acceptance vs Salary
Local PC: Individual H1B application status
Local PC: Tech jobs' description from Linkedin
HDFS
Spark: Clean and Extract H1B acceptance rate, Salary, Geographical and other meta information.
Spark: Extract job skills & degrees needed
Spark: Join dataset on employer, role and location
MLlib: Sensitivity analysis: Impact of salary on H1B acceptance rate
MLlib: Predict H1B acceptance rate based with multivariate classification model
Visualization with Tableau

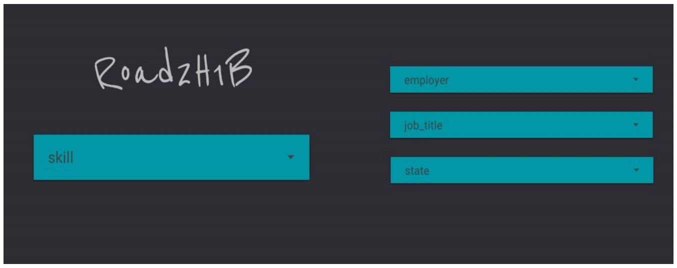


Figure 1 Query by meta information

We want to give our users the ability to picture their H1B acceptance rate with their existing jobs. Furthermore, we want them to know how learning different skill sets could influence the chance to secure a job in the tech industry. To do so, we created a website backed by a simple SQL database along with the regression model.

The webpage provides two different views for users to visualize what they searched for. In the first view, users can query the prediction results by selecting fields from the drop bar. This includes information such as employer, job title and geolocation as shown in Figure 1. As soon as users hit the search button, the confidence level of their H1B applications getting accepted will be prompted, with 0 being no chance and 1 being must accepted. This is done by running our pre-trained regression model on the back-end.

In the second view, the webpage provides users with the functionality to arbitrarily add more skills to their existing skill set and see if it can increase the confidence level. It is well-known that employees from big tech companies like Google and Facebook tend to be accepted with higher chances, and in fact, lots of the skill set these tech giants ask from candidates are in common. Having this view could potentially help users decide which skill set is more cost-effective to learn in the short time.

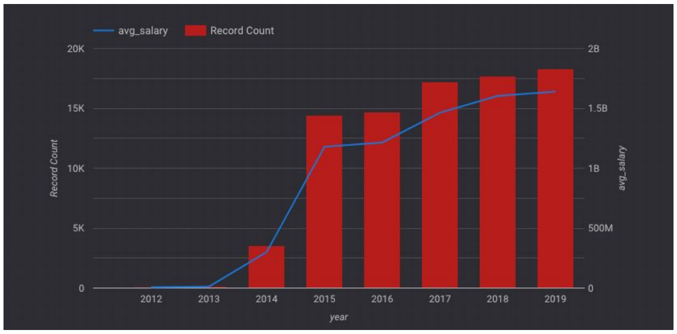


Figure 2. H1B & Salary trend from 2012 - 2019

Finally, asides from the searching results, the webpage also allows users to view the trend of H1B applications, as shown in Figure 2, and how they are correlated with different attributes.

Actuation or Remediation

(Describe the actuation or remediation response to the actionable insight. This is basically the action that can be initiated in response to the actionable insight produced by the analytic - the back-end of your application.)

VII. Analysis

(In this section, describe: Your experimental setup (tools, platforms), problems (with data, performance, tools, platforms, etc.). Describe what you learned. Discuss limitations of the application. Make recommendations for others, e.g. best practices.)

VIII. Conclusion

(One paragraph about the value, results, usefulness of your application.)

IX. Future Work

(Discuss possible future work for extending this project. Discuss how would you improve it, etc.)

Acknowledgment

(This section is optional. Use it to thank the people/companies/organizations who made data available to you, for example. You can list HPC people who were particularly helpful. List Amazon if you used an Amazon voucher. Cloudera for CDH.)

References

(Add references for all of the papers, texts, data sources.)

[1] KE, Barry; QIAO, Angela. Who Gets the Job and How are They Paid? Machine Learning Application on H-1B Case Data. *arXiv preprint arXiv:1904.10580*, 2019.

[2] GUNEL, Beliz; MUTLU, Onur Cezmi. Predicting the Outcome of H-1B Visa Applications.

[3] VEGESANA, Sharmila. Predictive analytics for classification of immigration visa applications: a discriminative machine learning approach. 2018.

[4] Sex Differences in Physician Salary in US Public Medical Schools, Anupam B. Jena, MD, PhD; Andrew R. Olenski, BS; Daniel M. Blumenthal, MD, MBA, September 2016

[5] H1-B employer data

[6] LinkedIn Job description data

[7] H1B Salary Database