H1-B Application Processing Analytics and Outcome Prediction

Yubing Bai
New York University
New York City, the USA
yb1059@nyu.edu

Minkang Yang
New York University
New York City, the USA
my1843@nyu.edu

***Abstract*—**

With growing number of applicants and limited number of H1-B visas, only part of H1-B applicants will be approved each year. Therefore, it is interesting to understand the selection preference and standard during the H1-B certification processing. In this paper, we investigate the main features of H1-B applicants who were certified and also those who failed using Hadoop and related big data tools. With the data on the official website of United States Department Of Labor, we analyze the H1-B data from 2 resources. The characteristics we choose comprise of the employers' job-related information and employees' information. Finally we provide a general view of H1-B visa certification situation, and depict what kinds of applicants are more likely to be certified.

Keywords—analytics, big data, H1-B, characteristics

I. INTRODUCTION

H1-B is a visa type that allows U.S. employers to temporarily employ foreign workers in specialty occupations. Since the acceptance rate of getting a HI-B visa is dropping down and the application fees is quite high, it becomes highly important to know the chances of acceptance based on the past records.

Big data techniques such as Hadoop and Hive allow us to process data over a long time period. Hadoop is a framework that allows for distributed processing of large datasets across clusters of commodity computers using map-reduce model. It is a flexible way to clean, normalize and profile raw data from different archives. Hive is a big data analysis tool based on map-reduce framework. It is user-friendly, easy to understand and learn. We utilize Hive to organize, classify and count the data, and from the result we can draw our picture of H1-B visa situation.

II. MOTIVATION

The United States has historically been the top destination for international students owing to its quality higher education system, welcoming culture, and relatively open labor market. In the last year, the United States remains the country of choice for the largest number of international students, hosting about 1.1 million of the 4.6 million enrolled worldwide. Those students are challenged to have a stable identification when they start their career in the USA, that is , most of them are expected to get the H1-B visa. This paper provide a statistics on those who has been approved H1-B, which can helps H1-B applicants understand the chance they might be certified. Also, for those who are not ready for finding jobs, the result acts as a instruction of companies and occupations that are preferred in H1-B application processing.

III. RELATED WORK

Some researchers has been working on the features of H-1B workers across firms. Based on the data on United States Citizenship and Immigration Services (USCIS) through a Freedom of Information Act (FOIA), Anna Maria Mayda[1] analyzed the characteristics of population with H1B visa from 1997 to 2012 and display the visa approval trending. As a result, The researchers give some basic observation including the H-1B holders' occupation distribution, geographical distribution, ranking of most H1-B approved companied, the top-receiving industry, intensity of use of H-1B visas. The researcher found that the number of H-1B holders employed by the commercial services sector has soared since 2008, mainly due to the increased use of H-1B workers and the opposition to increasing the size of the industry. the concentration of H-1B staff has increased significantly. From 2008 to 2012, the proportion of new initial employment H-1B granted by the top 20 petitioning companies has increased by 150%.

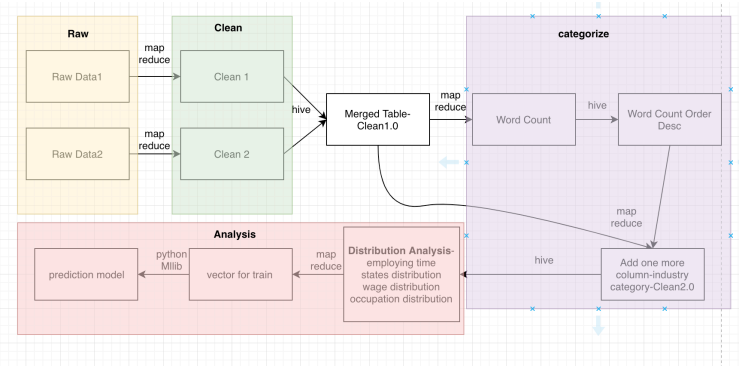
Also there exists some official surveys. According to the Characteristics of H-1B Specialty Occupation Workers [2] by U.S. Citizenship and Immigration Services from 2013 to 2014, we can see the general distribution of petitions, including the approval numbers, applicants' age, occupations, education background, salaries and so on. Such statistics are comprehensive but still need more updates for recent years due to the great possible changes in visa policy.

There are more relevant researches on extended dimensions considering the H1-B visa approval and its impact on the society. Giovanni Peri [3] used the inflow of foreign science, technology, engineering, and mathematics workers, made possibly by the H-1B visa program, to estimate the impact of STEM workers on the productivity of college and non-college-

educated American workers between 1990 and 2010. In terms of empirical analysis, they used variation in foreign-born STEM workers across US cities and time period to estimate their impact on native wages and employment. Their identification strategy relies on the H-1B supply-driven instrument. They closed their analysis by estimating the long-run effect of STEM on total factor productivity(TEP) and skill-biased productivity(SBP). Their results indicate that STEM workers spur economics growth by increasing productivity, especially that of college-educated workers.

IV. Design and Implementation

A. Design Details



The Design diagram above shows the overall process of the application. First, we download the raw data from official website [4][5]. Second, we run MapReduce to clean and normalize the raw data into cleaned data 1.0 with 12 columns. Third, we merged the data from two different sources together. Fourth, we use MapReduce word count and hive order function to extract the high frequency key word and categorize each record into one industry field, that's out data 2.0. Fifth, parse the data 2.0 into hive again and do distribution analysis, including employing time distribution, states distribution, wage distribution, and occupation distribution. Sixth, we transform each records into a vector and feed in into out certified/denied prediction model to predict a application's decision.

1. Data cleaning

We use Map Reduce to filter out those incomplete rows with some of the field blank or meaningless. We also filtered out those irrelevant rows and left 12 rows. Another thing worth mention is the separator. The final output of the reducer file is one string per row. We formatted the string this way: "column1{separator}column2{separator}...". At the beginning, we chose the "&" as operator, but when we try to create tables from two sources of data separated by "&", both of our two tables got some random column misplace all over the place. It turns out that some of the content of the column contains "&" themselves. We then found out that the original data actually contains all kinds of random symbols in the content. In the end, we first use regular expression to split the data into different parts and then replace the tab in each parts with space, then after cleaning each parts and filter out bad records, we concatenate all the parts together with '/t'.

2. Data normalization

1). Time format

There are 4 time fields, namely submission_date, decision_date, employer_start, and employer_end. They have different format. Submission_date is hyphen-separated, and the other three is forward-slash-separated and some of them have extra space within the date. In order for later time duration analyzing, we need to normalize all the time into the same format. We choose the year, month. day format(e.g 20180402) without any separator between year, month, and date. This format is convenient because it's purely digits. We could be parse this time format as double/int into hive without any extra effort.

2). Wage format

Unit of wage is given in the original data including year, month, week, bi-week, day and hour. Convert the wage rate to that with unit of year and correct them to two decimal places. Delete all the comma, quote and any other unexpected chars in the number.

3). Others

Names of jobs, occupations, enterprises and locations are string type. For the convenience of following word matching and counting, convert letters to lower case.

3. Data Categorizing

Now that both of data source is cleaned and normalized, we could merge them into the hive table, called it data 1.0. It has 12 fields, namely case_num string,case_status, submission_date, decision_date, employer_start, employer_end, employer_name, employer_state, employer_city, job_title, wage_rate, wage_unit. We would like to categorize each record into a industry field. We first extract the employer_name and job_title columns and run a MapReduce to get the over all word count of the employer and job related words. Then with the help of hive order by DESC function, we extract the high frequency key word. Then we first categorize these key word into 8 categories, namely administrative, architect and engineer, education, medical, computer, finance and business, entertainment, and other. With these high frequency keyword, we run another MapReduce on data 1.0, and add the category field for each record. This is our data 2.0.

4. Data Profiling

Check the maximum and minimum length, and maximum and minimum value for columns that we concerned with using Hadoop Map Reduce to make sure that the output from data cleaning is reasonable. If there is data that doesn't make sense, we need to step back and refractor out data cleaning and normalization part to filter out the bad data.

5. Distribution Analysis

Distribution analysis is completed with Hive. Import cleaned, normalized and well-categorized data and then use Hive commands to list, order and count. Five related aspects are considered in all in terms of the visa processing status and the condition of the applicants.

1). Processing duration analysis

Processing duration is calculated with case submission date and case decision date. According to the normalized data shown above, each data is a eight digit integer, so the duration is calculated by:

Processing duration = (floor(decision_date/10000) - floor(submission_date/10000))*365 + (floor(decision_date/100)%100 - floor(submission_date/100)%100)*30 + (decision_date%100 - submission_date%100).

Use SELECT clause to list records in an order of increasing processing time and group them by wether the case has been approved. The number of cases certified or denied for corresponding processing length is displayed in a table.

2). Employment duration analysis

Similar to the processing duration, employment is calculated with employment start date and employment end date. Data format is a eight digit integer, thus the duration is calculated by:

Employment duration = (floor(employer_end/10000) - floor(employer_start/10000))*365 + (floor(employer_end/100)%100 - floor(employer_start/100)%100)*30 + (employer_end%100 - employer_start%100).

Use SELECT clause to list records in an order of increasing employment duration and group them by wether the case has been approved. The number of cases certified or denied for corresponding employment period length is displayed in a table.

3). Working location distribution

There is no consecutive meaning between state names, so the whole records are split into two subtables, one of certified cases and another of denied cases, to separately show the region distribution of certified/denied applications.

Use SELECT clause to count the total number of cases in certain state and use JOIN clause to merge the certified results and denied results into one table.

4). Wage rate distribution

Wage per year is usually a at least 5 digit number which is a great number to categorize. Therefore, describe wage level by wage_rate/100, to decrease the number of wage rate category.

Use SELECT clause to get the relationship between the number of passed cases and wage_rate.

5). Occupation distribution

Six main occupation fields are examined, each of them are labelled from 0 to 6. Repeat the same procedure as working location analysis and get the occupation distribution of numbers of certified/denied cases.

6. H1-B application result prediction model

We characterize a H1-B application case by five dimensions, including wage rate, job location, processing duration, employing duration, occupation. After data normalization, wage rate is calculated by year in floating number; duration is a day-unit value calculated by end data minus start date; case_status, occupation and location is labelled with integer starting from 0.

We generate 67-dimension feature vectors row by row. Because the comparison of the value of those labels are meaningless, occupation and location are flattened as zero-initialized vectors of 8 dimensions and 55 dimensions respectively, and if the label number is n, then the nth record in that vector is assigned to 1. All the other fields are one dimension vector. Case status is set as 1 when the case is certified and 0 otherwise.

There are one million vectors in total, and randomly select 80 percents of them as training data and the rest of them as test data.

Build case decision forecast model by Logistic Regression Method and prediction result is assessed by log loss and AUC curve.

B. Description of Datasets

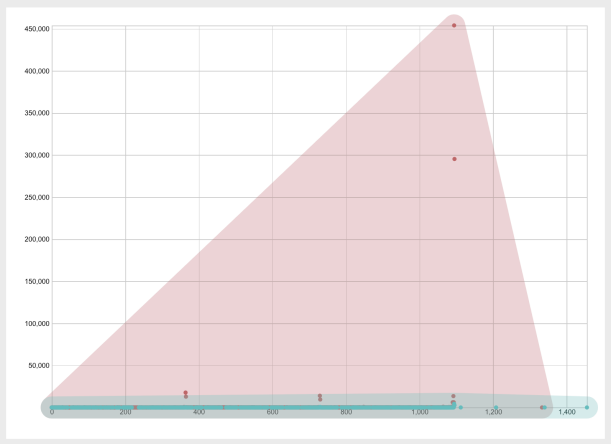
The first data source is the official website of United States Department Of Labor. The file is presented as .xlsx and the data generated directly according to what applicants filled in the form. The data contains hundreds of thousands of history application records in recent 2 years. Each records contains the following columns: case_number, case_status, case_submitted, decision_date, visa_class, employment_start_date, employment_end_date, employer_name, employer_business_dba, employer_address, employer_city, employer_state, employer_postal_code, employer_country, employer_province, employer_phone, employer_phone_ext, agent_representing_employer, agent_attorney_name, agent_attorney_city, agent_attorney_state, job_title, soc_code, soc_name, naics_code, total_workers, new_employment, continued_employment, change_previous_employment, new_concurrent_emp, change_employer, amended_petition, full_time_position, prevailing_wage, pw_unit_of_pay, pw_wage_level, pw_source, pw_source_year, pw_source_other, wage_rate_of_pay_from, wage_rate_of_pay_to, wage_unit_of_pay, h1b_dependent, willful_violator, support_h1b, labor_con_agree, public_disclosure_location, worksite_city, worksite_county, worksite_state, worksite_postal_code, original_cert_date.

The second data source is Foreign Labor Certification Data Center. The data is presented in text and already rows and columns and a document for the column field. One file could have hundreds of thousands of rows, and each row has totally 37 fields, e.g SUBMITTED_DATE, CASE_NO, NAME, ..., JOB_CODE, WAGE_SOURCE, OTHER_WAGE_SOURCE. Some fields are not so relevant to the possible insights we would derive from the data, while others could be useful.

V. RESULTS

So totally we have around 1.01 million records and after cleaning, we got around 919168 records left. We normalize , categorize, and analyze these data. We also use the records to train our prediction model to predict the decision of a given application. Here are some results we found:

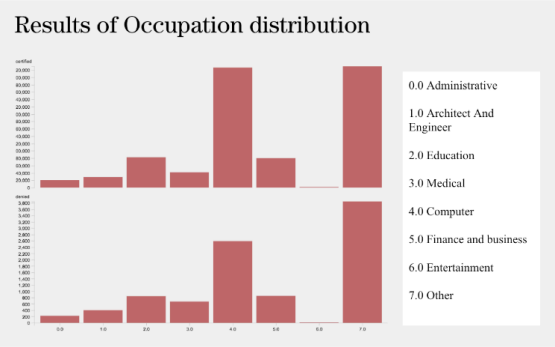
1. Employer time and application decision relation



X-axis: employer time Y-axis: applications counts.

From the diagram we could see that denied cases are spread all over the different range to employer time. While the certified data have a high concentration among the employer time around 2-3 years. The result infer that and job offer with employer time of 2-3 years would have a much higher opportunity to get certified for h1b.

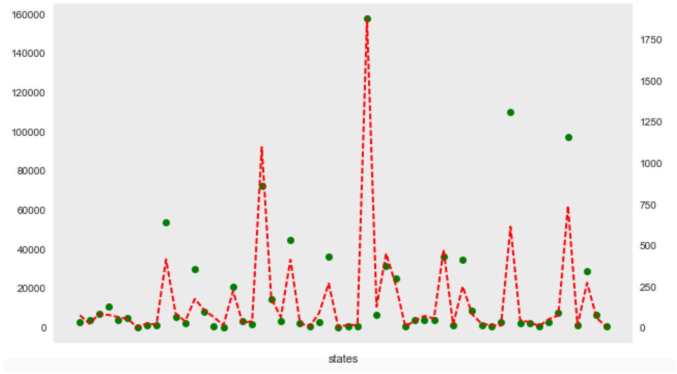
2. Industry distribution



X-axis: industry categories Y-axis: applications counts

From this diagram we could see that there is high concentration in the computer field for both denied and certified application. The denied and certified counts is almost half and half. And it does show that applications in the field of computer has higher certified rate.

3. States distribution



X-axis: states(cannot show state name due to size)

Y-axis: application counts

From the graph we could clearly see there is one state has much more application than other other states. The state is California.

4. Forecast results

We randomly select 80% of our data records as training data set and the rest of them as test set. The goodness of prediction model is evaluated by two dimensions: accuracy and AUC curve. Accuracy refers to the correctly predicted cases as a percent of the total number of cases, which shows the performance of our prediction accuracy. AUC curve shows the capability of distinguishing between two kind of cases. It is depicted by: TPR = f(FPR), where TPR is the true positive result as a percent of the number of positive cases and FPR is the number of false positive results as a percent of the number of negative cases.

Finally we get a Accuracy of 98% and a AUC of 0.67, which shows that our model is capable to distinguish passed and failed applications with a high accuracy. This in return proves that the factors we select do have an impact on certification decision.

VI. Future Work

Given more time, we would like to include more years of data into the analysis, we we could have a bigger over all picture of the trend of h1b visa.

VII. Conclusion

From the analytics we point out that there is a concentration of h1b holders in technical occupation, and there is a strong relationship between stable working status and possibility of success. Besides, CA has much more appliers than other other states.

A prediction model is built based on those observations and performs well in case test, which in return proves that the factors we select do have an impact on certification decision.

Acknowledgment

(This section can be used to thank the people/companies/ organizations who have made data available to you, for example. You can also list NYU HPC and any HPC people who were particularly helpful, for example.)

We would like to appreciate NYU HPC and DUMBO for their services. We also would like to appreciate The Office of Foreign Labor Certification and official website of United States Department Of Labor for giving us their data freely.

REFERENCE

(Add references for all of the papers/texts that you refer to in your paper. You will probably want to include the papers you read that were related to your project. You may have websites to reference, the Hadoop book, the MapReduce paper, the Pig Latin paper, etc. Some references are added below as an example. Links to original datasets and APIs used should be listed here and referenced in this paper where appropriate.)

1. Mayda, Anna Maria, et al. "New Data and Facts on H-1B Workers across Firms." The Role of Immigrants and Foreign Students in Science, Innovation, and Entrepreneurship. University of Chicago Press, 2018.
2. Citizenship, U. S. "Characteristics of H1B Specialty Occupation Workers." (2013).
3. Peri, Giovanni, Kevin Shih, and Chad Sparber. "STEM workers, H-1B visas, and productivity in US cities." Journal of Labor Economics 33.S1 (2015): S225-S255.
4. http://www.flcdatacenter.com/CaseH1B.aspx
5. https://www.foreignlaborcert.doleta.gov/performancedata.cfm