IOSR Journal of Engineering (IOSR JEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 49-54

Data Preprocessing and Analysis for H-1b Visa Petitions

Tanvi Kulkarni[#1], Gauri Nayak[#2], Rachana Devasthali[#3], Nikita Dhuri[#4], Saishree Godbole[#5] Razia Sardinha[#6], Derrick Nazareth[#7]
Information Technology Department, Padre Conceicao College ofEngineering, Verna, Goa, India
#1tanvinurag@gmail.com
#2nayaquegauri33@gmail.com
#3rachna.devasthali@gmail.com
#4nikitaa24@gmail.com
#5saishreegodbole@gmail.com

Abstract: *H-1B Visa in United States allows employers to employ foreign skilled workers in specialty occupations. This paper addresses the approach to analyze the effect of different attributes in determining the success of the petition and study the different trends in the scenario. We also aim to predict the outcome of a given petition based on various attributes such as employer name, soc_name, wages, filing year and the worksite. As the H-1B visa category is one of the highly coveted ones, this approach can be used by both the individual and the employer in between applying for the visa and getting the final decision to be informed of the outcome before it occurs.*
Keywords: *H-1B Visa, Data analysis, Data visualizations, Data preprocessing, Data mining.*

I. Introduction

H-1B is a type of non-immigrant visa in the United States that allows US companies to employ foreign skilled workers in various specialized fields.This visa requires the applicant to have a job offer from an employer in the US before they can file an application to the US immigration service (USCIS). USCIS grants 85,000 H-1B visas every year, even though the number of applicants far exceed that number. The duration of stay is three years, extendable to six years. The selection process is claimed to be based on a lottery, hence how the attributes of the applicants affect the final outcome is unclear. Therefore we have made an attempt to explore the dataset of H-1B visa applications in order to determine the contributing factors to the application outcome.

In our work, we focus on

1) Data preprocessing of two datasets, one for the period of 2011 to 2016 and other from 2017 to 2018.
2) Analysis of the two datasets to determine the factors which affect the outcome of H-1B VISA petition.

The next section covers literature survey. The description of the datasets used is made in section 3. The topic of data preprocessing is described in section 4. The analysis phase is covered in section 5. The section 6 explains findings which are the outcome of analysis phase while section 7 describes the conclusions made. Finally the section 8 briefs about further work and closing section includes references made.

II. Literature Survey

Although this area currently doesn't seem to be well-studied, we were able to find some reports that were helpful. Most papers only focused on prediction, hence leaving an exploratory data analysis of H-1B visa applications datasets as a unique feature of our project.

We found three relevant reports on similar studies that used the same dataset as we did.

First report conducted a detailed data analysis and visualization for H-1B application distribution based on different input features such as location, salary, year and job type [5]. Although they had a prediction algorithm based on K-means clustering and decision trees, they provided prediction accuracies for only a small subset of job types instead of an average one. Overall, this report gave us good insight on the distribution of our data.

Second report, that was drafted by UCSD students used AdaBoost, Random Forest, Logistic Regression and Naive Bayes to predict the case status, however there was no analysis done in this paper. [4].

The third report, drafted by students of Stanford University, used the SGD Classifier from the scikit-learn library [9] to implement Naive Bayes, Logistic Regression and soft-margin linear SVM. They preferred stochastic gradient descent over batch gradient descent due to the size of the dataset. They also used neural networks with l2 regularization, Naïve Bayes as well as Logistic Regression for prediction and presented an

excellent comparison amongst all the methods used based on training and test accuracy. Their findings showed that Neural Network with l2 regularization outperformed all the other models with 98% training accuracy and 82% test accuracy on the balanced test data.[9]

III. Datasets Used

We have used two datasets for this project. One of it is taken from kaggle[1]. This dataset has about 3 million data points, 7 features and 1 class label. the data spans over the years 2011-2016. Another dataset is taken from United States Department of Labor's (USDL) Office of Foreign Labor Certification (OFLC)[2]. This dataset contains 50000 data points and data spanning from 2017-2018.

For the analysis we have considered the attributes common to both the data sets. These attributes include EMPLOYER_NAME, CASE_STATUS, YEAR_OF APPLICATION, WORKSITE, WAGES AND SOC_NAME.

IV. Data Preprocessing

Data preprocessing has a significant impact on the outcome of analysis phase and on performance of a ML algorithm. Data preprocessing is a data mining technique that involves transforming raw data into an understandable format.

Why we use Data Preprocessing ?

Real-world data is often **Incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data,**Inconsistent**:containing discrepancies in codes or names and **Noisy**: containing errors or outliers. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.

Moreover, in real world data, the representation of data often uses too many features,but only a few of them may be related to the target concept.

Preprocessing Done

A. For 3 million rows dataset:

We split the WORKSITE attribute consisting city and state values into two separate attributes WORKSITE_STATE & WORKSITE_CITY. WORKSITE_CITY was discarded.

We dropped the attributes such as DECISION_MONTH, DECISION_YEAR,CASE_SUBMITTED_DAY, DECISION_DAY, CASE_SUBMITTED_MONTH, WAGE_UNIT_OF_PAY, H-1B_DEPENDENT, WILLFUL_VIOLATOR, WORKSITE_POSTAL_CODE whichwere not common to both the datasets used.

Also, we noticed that the JOB_TITLE feature represents highly redundant information with the SOC_NAME feature, therefore we discarded JOB_TITLE.

We then deleted the rows that had VISA_CLASS attribute value such as 'E3 Australian', 'H1B1 CHILE', 'H1B1 SINGAPORE' since we needed to analyze H-1B VISA.

After successful deletion of the rows as mentioned above we dropped VISA_CLASS Attribute.

Rows with values such as 'hour','month','bi-weekly','week' for the attribute PW_UNIT_OF_PAY were deleted. Only rows with ' annum' value were retained. After dropping rows the column PW_UNIT_OF_PAY was dropped.

For the attribute CASE STATUS: We excluded the cases 'WITHDRAWN', since 'WITHDRAWN' decisions are either made by the petitioning employer or the applicant, therefore not predictive of USCIS's future behavior. We also converted 'CERTIFIED-WITHDRAWN' label to 'CERTIFIED' since these petitions were certified by USCIS to begin with, but later withdrawn by the employer or the employee. Also we converted 'REJECTED' to 'DENIED'. We removed the entries that had the values 'INVALIDATED' and 'PENDING QUALITY AND COMPLIANCE REVIEW-UNASSIGNED' for the case status.

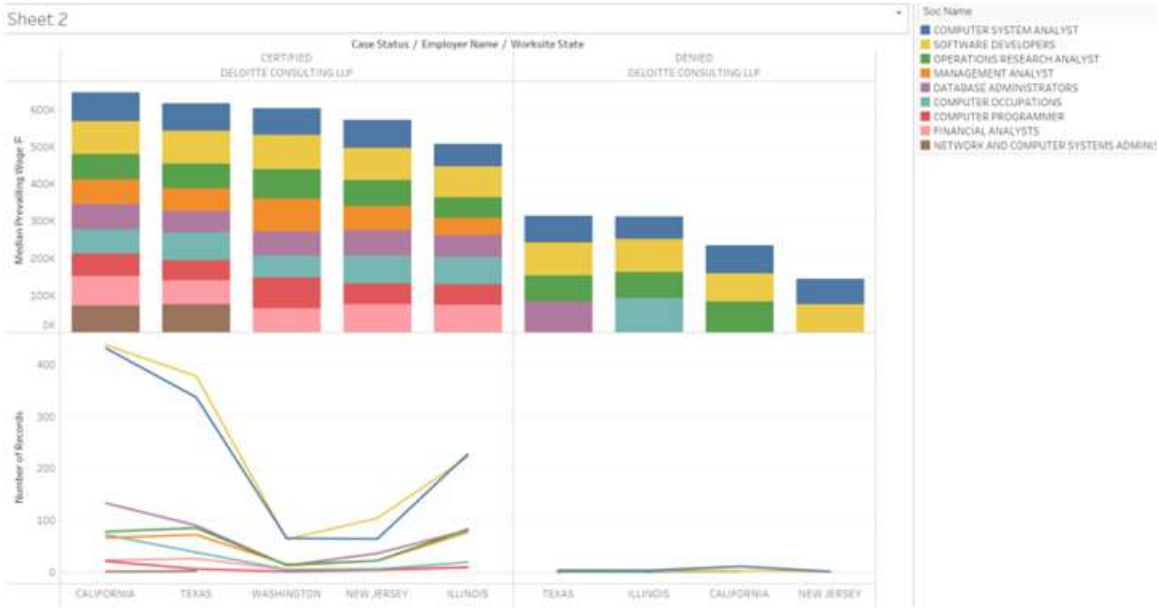
For the attributes SOC_NAME and EMPLOYER_NAME: Since entries that were in uppercase were considereddifferent from those entries that were bearing the same value in lowercase,we first converted all the attribute values under SOC_NAME & EMPLOYER_NAME into Uppercase. We also removed the entries that contained corrupted data bearing unnecessary numbers and characters that didn't make any sense. We replaced numerous spelling mistakes in both these attributes.

For more plots visit

https://drive.google.com/file/d/0B3nulL6mfs3GNFdManVRRXk5eng0QjZmV2VJdlZQM1NHU0k0/view?usp=sharing

B. For 50000 rows dataset

We split the WORKSITE attribute consisting city and state values into two separate attributes WORKSITE_STATE & WORKSITE_CITY. WORKSITE_CITY was discarded.



Wedropped the attributes such as DECISION_DATE, EMPLOYMENT_START_DATE, EMPLOYMENT_END_DATE, EMPLOYER_ADDRESS, EMPLOYER_CITY, EMPLOYER_STATE, EMPLOYER_POSTAL_CODE, EMPLOYER_COUNTRY, which were not common between the 2 datasets.
We then deleted the rows that had VISA_CLASS attribute values such as 'E-3 Australian', 'H-1B1 CHILE', 'H-1B1 SINGAPORE' present in the second dataset.
Later VISA_CLASS attribute was dropped.
Rows with value for PW_UNIT_OF_PAY attribute as per week, per hour ,bi-weekly, per month were removed. Then PW_UNIT_OF_PAY was dropped.
CASE STATUS: We excluded the cases 'WITHDRAWN', since 'WITHDRAWN' decisions are either made by the petitioning employer or the applicant, therefore not predictive of USCIS's future behavior. We also converted 'CERTIFIED-WITHDRAWN' label to 'CERTIFIED' since these petitions were certified by USCIS to begin with, but later withdrawn by the employer or the employee.
YEAR: The values of this attribute were in the float point form. We converted the same into integer format.
SOC_NAME and EMPLOYER_NAME: We first converted all the values under this attribute into Uppercase as some entries that were already in uppercase were treated as different from the entries bearing the same value but in lowercase. We also removed the entries that contained corrupted data bearing unnecessary numbers and characters that didn't make any sense. We replaced numerous spelling mistakes in these attribute values.

V. Analysis

In the analysis phase, we explored two datasets of H-1B visa applications and their outcomes in order to determine the contributing factors to the application outcome.
We tried answering some of the interesting questions like:

1) What makes an applicant more likely to be certified?
2) Which employers have higher success rates of visa application acceptance?
3) Salary variations with respect to SOC_NAMES.
4) Number of applications per year.
5) Most prominent work states across USA.

We used various data visualizations for uncovering the patterns in data and presenting them in a simplified manner. Some of the basic visualizations used were plots like bar graphs, line graphs and area graphs that were implemented using **Seaborn**, a python library. We used **Tableau Public**, an excellent open source visualization software to implement complex plots such as Box-Whisker plots, Violin plots, Treemaps, Multigraphs, Bubblecharts, Scatter plots. Some other visualizations such as Coxcomb Plots were drawn using an online tool called **Vizzlo** and for implementing plots like Chloropleth maps we used **Plotly**, another python library.

For more plots visit
https://drive.google.com/file/d/0B3nulL6mfs3GNFdManVRRXk5eng0QjZmV2VJdlZQM1NHU0k0/view?usp=sharing

VI. Findings

Following are the briefed findings of our analysis:

i. There has been a steady increase in the number of H-1B visa applications from 2011 onward
ii. Indian companies such as Infosys and TCS are dominating the chart with the highest number of applications.
iii. In general comparison with other states, California unsurprisingly has the highest no. of certified applications overall, from 2011 to 2016 (184000), followed by Texas, New York and Illinois.
iv. Amongst all the datasets Washington pays the highest wages (84000), immediately followed by California (79000) from 2011-2016
v. For Computer Programmers, Washington provides the highest wages above 150K, over a period of 2011 to 2016.
vi. Among all the states, California has an increasing trend of certified applications for Computer Programmers, with more than 10K applications certified in 2015.
vii. California has the highest number of applications for statisticians in 2018.
viii. In New Jersey, statisticians have the highest median wages above 200K, followed by California.
ix. For Statisticians, California saw a hike in both, the no. of applications, as well as the prevailing wages in 2016 over a period of 2011 to 2016.
x. California proves to be the hotbed for Software Developers in terms of number of applications getting certified (above 30K) and provides highest wages above 200K.
xi. The denied Software Developers have lower wages compared to their certified counterparts, ranging from wages greater than 50,000 USD in 2011 to wages greater than 100,000 in 2016.
xii. During 2011 to 2016 Infosys has the highest no. of certified petitions, followed by TCS and Deloitte.
xiii. In 2018, TCS and Deloitte are biggest players having most certified petitions, followed by Infosys and Cognizant.

A. Analysis of top 3 employers during the period 2011 to 2016

1) INFOSYS:

2011:Computer System Analysts,Computer Occupations, Computer Programmers, Management Analyst had higher no. of certified applications in California. Computer Systems Analyst, Computer Programmerswere highly paid in New York.Computer Occupations were highly paid in California. Management Analysts were highly paid in Texas.
2012: Computer System Analysts,Computer Occupations,Computer Programmers, Management Analyst had higher no. of certified applications in California. Computer System Analysts were highly paid in California,New York and Texas equally.
2013: Computer System Analysts were highly paid in California,New York and Texas equally. Computer Programmers, Management Analysts were highly paid in California,New York.
2014:Computer System Analysts were highly paid in California and Texas. there was a decrease in no. of certified records in Texas, California,New York.
2015:Overall there was a drastic increase in the no. of certified applications compared to 2014.
2016:Overall there was a decrease in the no. of certified applications compared to 2015.

2) DELOITTE

In 2011: Computer System Analyst were certified the most specially in California & Texas. Software Developers were highly paid in California,New York,Texas. Computer Systems Analyst, Management Analysts were highly paid in New York &Texas.
2012:Software Developers were certified the most in California,New York & Texas. Software Developers & Computer Systems Analyst were highly paid in New York. Management Analysts were highly paid in California.
2013:There was a reduction in the no. of certified records for Software Developers in California,New York,Texas. Whereas there was growth in the no. of certified records for Computer System Analysts and Management Analysts.
2014 & 2015:There was a increase in no. of certified records for Software Developers, Computer System Analyst, Management Analyst in California,New York,Texas. Computer System Analyst gained higher wages in New York and Texas during 2014 and in California for 2015.

2016:There was a Reduction in the no. of certified records for Computer Occupations,Computer Programmers, Computer System Analyst.but increase in the no. of certified records for Management Analysts.

3) WIPRO:

Computer Programmers, Computer System Analysts are highly paid and have higher no. of certified applications in California followed by New York and Texas.
2011:Computer Programmers were certified most in California and were paid highest in Texas.
2012: Computer Occupations were certified the most in California followed by New York. Computer Programmers were highly paid in California.There was an overall increase in the no.of certified applications.
2013:In California there was a increase and in New York there was decrease in number of certified applications.ComputerProgrammers were highly paid California.
2014: In California there was a increase and in New York there was a decrease in no. of applications.Computer System Analysts were highly paid in New York and Texas.
2015: No. of certified applications decreased slightly for Computer Programmers.Computer Programmers were highly paid Texas.for Computer System Analyst no. of certified records increased and were highly paid in California.
2016: Overall the no. of certified records reduced compared to 2015. Computer Programmers gained maximum wages in New York.Management Analysts emerged as a new entry in top 5 SOC_NAME for Wipro and got highest wages in New York.

B. Analysis of top 3 employers in 2018

1) *INFOSYS:* no. of certified applications were the highest for California followed by Texas for Computer System Analyst.Computer System Analysts were highest paid in all the top 5 SOC_NAMEs. Management Analysts were second highest and network & computer administrator were third highest for Texas, California, New Jersey, Illinois. Network and computer administrator were the second highest for only Washington.

2) *DELOITTE:* no. of certified applications were the highest for California followed by Texas for Computer System Analyst and Software Developers. Computer System Analyst were highly paid for Texas,California,Washington,New Jersey,Illinois, followed by Software Developers and operation research analysts and Management Analysts. No. of data administrators certified were highest in California followed by Texas.

3) *TCS:* Computer Occupations were highly paid,followed by Computer Programmers being second highest and Computer System Analyst being third highest in Texas, California, New Jersey,Illinois. Computer System Analyst being second highest for Washington. no.of certified applications were the highest for Computer Programmers in California followed by Computer Occupations and Computer System Analysts.

C. Analysis with respect to topmost analyst positions

Top 5 analysts with respect to no.of certified records for the year 2018 are Computer System Analyst, Management Analysts, operating research analysts, market research analysts & financial analysts. Computer System Analysts being the highest paid by cognizant, Deloitte, IBM, Infosys, TCS. Followed by Management Analysts and operation research analysts. No. of certified records is maximum for Computer System Analyst in Infosys in California followed by Texas. For Cognizant, California has the highest certified no. of applications.

VII. Conclusion

For the years 2011-2016 the topmost employers based on the number of certified petitions were Infosys, TCS, Deloitte, Wipro, IBM respectively. While the topmost SOC_NAMES are Computer System Analyst, Software Developers, Computer Programmers, Computer Occupations, Management Analysts. The topmost worksites during the same period were California, Texas, New York, New Jersey, Illinois.

However for the year 2018, the above rankings were shuffled as follows: Deloitte, TCS, Infosys, Cognizant, IBM respectively. The topmost SOC_NAMES were Software Developers, Computer System Analysts, Computer Occupations, Computer Programmers, Operation research analyst respectively. The topmost worksites were California, Texas, New York, New Jersey, Maryland respectively.

Rreferences

[1]. "H-1B Fiscal Year (FY) 2018 Cap Season," USCIS. [Online]. Available:https://www.uscis.gov/working-united-states/temporary-workers/h-1b-specialty-occupationsand-fashion-models/h-1b-fiscal-year-fy-2018-cap-season. [Accessed: 20-Oct-2017].
[2]. "High-skilled visa applications hit record high," CNNMoney. [Online].Available:http://money.cnn.com/2016/04/12/technology/h1b-cap-visa-fy-2017/index.html. [Accessed: 20-Oct-2017].
[3]. "Using Text Analysis To Predict H-1B Wages," The Official Blog of BigML.com, 01-Oct2013. [Online]. Available: https://blog.bigml.com/2013/10/01/using-text-analysisto-predict-H-1B-wages/.[Accessed:20-Oct-2017].

[4]. "Predicting Case Status of H-1B Visa Petitions." [Online]. Available:https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a054.pdf.
[5]. "H-1B Visa Data Analysis and Prediction by using K-means Clustering and Decision Tree Algorithms." [Online]. Available: https://github.com/Jinglin-LI/H1B-VisaPrediction-by-Machine-Learning-Algorithm/blob/master/H1B\%20Prediction\%20Research\%20Report.pdf.
[6]. H-1B Visa Petitions 2011-2016 — Kaggle. [Online].Available:https://www.kaggle.com/nsharan/hvisa/data.[Accessed:20-Oct-2017].
[7]. Fabian Pedregosa, Ga¨elVaroquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Edouard Duchesnay. Scikit-learn: Machine Learning in Python, Journal of Machine Learn- ´ ing Research, 12, 2825-2830 (2011).
[8]. Fernando P´erez and Brian E. Granger. IPython: A System for Interactive Scientific Computing, Computing in Science & Engineering, 9, 21-29 (2007).
[9]. BelizGunel, OnurCezmiMutlu. *Predicting the Outcome of H-1B Visa Applications*, CS229 Term Project Final Report, Stanford University.
[10]. Darshit A. Pandya. Predicting filed H-1B Visa petition's status, International Research Journal Of Engineering and technology (IRJET), Volume:05, Issue: August,2018.
[11]. https://www.foreignlaborcert.doleta.gov/pdf/performancedata/2018/h-1b_disclosure_data_fy2018-q3.xlsx *(link to the dataset of H-1B applications for the year 2018)*
[12]. S. B. Kotsiantis, D. Kanellopoulos, P. E. Pintelas. Data Preprocessing for Supervised Learning, World Academy of Science, Engineering and Technology, Volume:1 No.12, 2007.