Classification of Work Visa Approval

1. Abstract

H1-B Visa is the most sought-after non-immigrant visa that allows foreign workers to work in United States in specialty occupation. In 2019, more than 1 million applicants applied to get an H-1B visa [1] including new applications, renewals and transfer of H1-B to another company. There were more than 180,000 new applicants for H1-B [2], however, only 80,000 applications were picked up in the lottery process for taking it further to USCIS for approval.

The uncertainty in getting an H1-B visa creates employment and legal status uncertainties for a job application and high legal and visa processing fees [3] for the organization over the period of employment. We plan to use the anonymized dataset for 2019 that United Status Department of Labor publishes publicly [4] and apply data science techniques to improve predictability of approval.

2. Introduction

United States Department of Labor publishes a dataset each year for different categories of Visa such as L visa, H visa as well as PERM applications. The dataset includes employer details such as Name, Address, if it is H1-B dependent as well as employment details such as Job Title and proposed wage.

We start with null hypothesis that there is a correlation between the profile of an employer and the outcome of a visa application. We plan to model that in this paper and predict whether an application will get rejected or approved. However, we limit the dataset to H1-B visas which is most popular work visa for an immigrant, but it can be extended to other visa types as well.

3. Related Work

Though the dataset is published publicly, there hasn't been much work in analyzing this to best of our knowledge. We also saw new features being added each year and were not available in past. For instance, 'master's exception' and 'other information about universities', more details on companies like whether they are 'H1-B dependent' or 'Willful Violator' of H1-B were added in 2017 dataset, 'attorney information' was added in 2015 and so on. We believe this data could be helpful in modelling employer and application profiles to have better predictions than last year.

4. Dataset and Features

Parsing the 2019 excel data into, we found 589,414 cases for H1-B applications. The dataset contains features [5] that gives information about employer and visa applicant. Out of all the features, due to the contextual important, we selected [7] below:

CASE_STATUS: Excluding the Withdrawn and Certified-Withdrawn, Certified decision is considered as 1 outcome in the resulting dataset and Denied as 0. This is used to model the outcome.

VISA_CLASS: Only H1-B visa class is being modeled in this paper which contributes to the majority of datapoints, we exclude the records for other work visas such as E-3 Australian, H-1B1 Chile and H-1B1 Singapore.

EMPLOYER_NAME: The employer name submitting the visa application. We believe employer name is one of the important features to profile the visa application. As per NY times, some companies are manipulating the visa process by flooding the system [6].

SECONDARY_ENTITY: Whether the applicant will be places in a secondary location. This feature is assumed to be helpful since majority of Consultancy companies that are believed to outsource software services which have reputation to flood the visa processing [6].

AGENT_REPRESENTING_EMPLOYER: If another firm is representing the employer and its application. We plan to model this feature to see if an agency has high rejection rates as compared to another.

JOB_TITLE, SOC_NAME: Job title and SOC name have details about the position, occupation field and seniority of the applicant.

SOC_CODE, NAICS_CODE: They are standard categories of a job.

CONTINUED_EMPLOYMENT: If this is a re-new visa application

CHANGE_PREVIOUS_EMPLOYMENT: If an application will continue without changes in job duties

NEW_CONCURRENT_EMPLOYMENT: If the applicant will have an additional employer

CHANGE_EMPLOYER: If applicant will get the visa with a new employer

AMENDED_PETITION: If an applicant will work with the same employer with changes in duties FULL_TIME_POSITION: If this application is for full-time position

H-1B_DEPENDENT: If an employer is categorized to be H1-B dependent.

SUPPORT_H1B: If this application will be used in the future to file for H1-B petitions

WILLFUL_VIOLATOR: If an employer has violated H1-B rules in the past.

WAGE_RATE_OF_PAY_FROM: Employer's proposed wage rate

WAGE_UNIT_OF_PAY: Paycheck frequency.

TOTAL_WORKER: Total amount of workers in the company filing the application.

Here is a record in dataset:

CASE_STATUS	EMPLOYER_NAME	JOB_TITLE	SOC_NAME	PW_UNIT_OF_PAY	WAGE_RATE_OF_PAY_FROM	WAGE_RATE_OF_PAY_TO	WAGE_UNIT_OF_PAY
CERTIFIED	THE UNIVERSITY OF TEXAS HEALTH SCIENCE CENTER AT HOUSTON	ASSISTANT PROFESSOR - RESEARCH	BIOLOGICAL SCIENTISTS, ALL OTHER	Year	$85,000.00	$85,000.00	Year

5. Methods

We clean the original dataset [8] to get a common format to run logistic regression on the non-textual features and textual features as well. We observe the dataset and see contribution of individual features on the predicted decision.

For Logistic regression [19], we select the features [9] `SECONDARY_ENTITY`, `AGENT_REPRESENTING_EMPLOYER`, `TOTAL_WORKERS`, `NEW_EMPLOYMENT`, `CONTINUED_EMPLOYMENT`, `CHANGE_PREVIOUS_EMPLOYMENT`, `NEW_CONCURRENT_EMPLOYMENT`, `CHANGE_EMPLOYER`, `AMENDED_PETITION`,`FULL_TIME_POSITION`, `H1B_DEPENDENT` and `WILLFUL_VIOLATOR`. Probability of outcome (y) given by:

$$\mathcal{P}_{\boldsymbol{w}}(y = \pm 1|\boldsymbol{x}) \equiv \frac{1}{1 + e^{-y\boldsymbol{w}^T\boldsymbol{x}}},$$

We also plan to run Multinomial Naïve Bayes [20] to see how likely it is to get a reject based on occupation and company name in the visa application [10]. Initial features include `JOB_TITLE`, `EMPLOYER_NAME`, `SOC_CODE`, `SOC_NAME` and `NAICS_CODE`. Probability for Multinomial Naïve Bayes is given by:

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y), \qquad \hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

where N_{yi} is the is the number of times feature *i* appears in a sample of class *y*, N_y is the total count and α accounts for features that are not present to prevent zero probabilities.

We calculate wage information and structure in unique buckets based on its distribution [11] from features `WAGE_RATE_OF_PAY_FROM`, `WAGE_RATE_OF_PAY_TO`, `WAGE_UNIT_OF_PAY` and `PREVAILING_WAGE` and evaluate its relevance in logistic as well as Naïve Bayes models.

The random forest implementation used is from sklean.ensemble [18] which averages the result from each decision tree in the random forest. Each decision tree is internally generated to maximize the entropy given by

$$H(S) = \sum_{c \in C} -p(c) \log_2 p(c)$$

C ={yes,no}

6. Experiments, Results and Discussions

6.1 Logistic Regression

We process the data to make it in a format where logistic regression can run [7]. We also fix the missing data wherever it doesn't change the meaning [9]. We run logistic regression on the dataset [12] by separating training to validation set at a 30:70 ratio.

This gives a high accuracy of 99.175% but the confusion matrix looks like:

	True Negative	True Positive
Predicted Negative	48	3446
Predicted Positive	49	419979

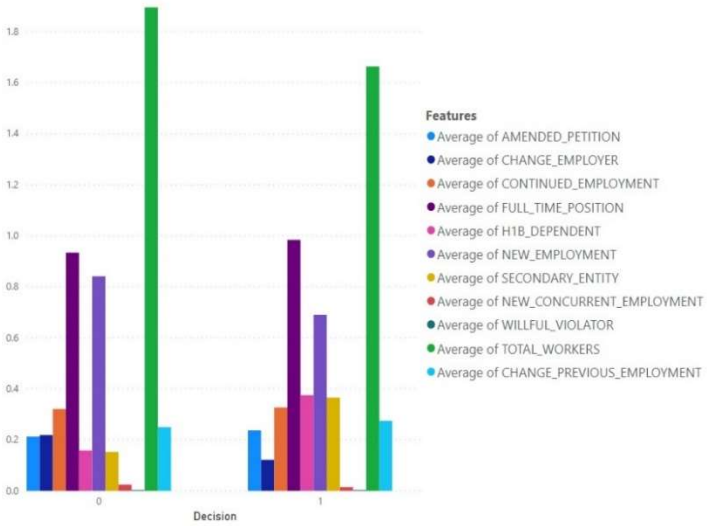

Fig: All logistic features average between accepted and rejected

This is because of the approval ratings are high compared to rejection-rate for year 2019. Logistic Regression doesn't improve the outcome. Features are distribution across Approved and Denied case of Visa suggests they are less relevant. So far, Visa decision seems having weak co-relation with features that profile employer and applicants.

When logistic regression is run on the wage information [11] the accuracy returned is 99.14% but is predicts a low negative outcome. After looking at the distribution of the wages, we implement a custom logic to capture all of them in comparable buckets.

There is been discrepancy in the data for previous years, for instance, most of the features that we processed logistic on is available in 2014 dataset [14] except wage information.

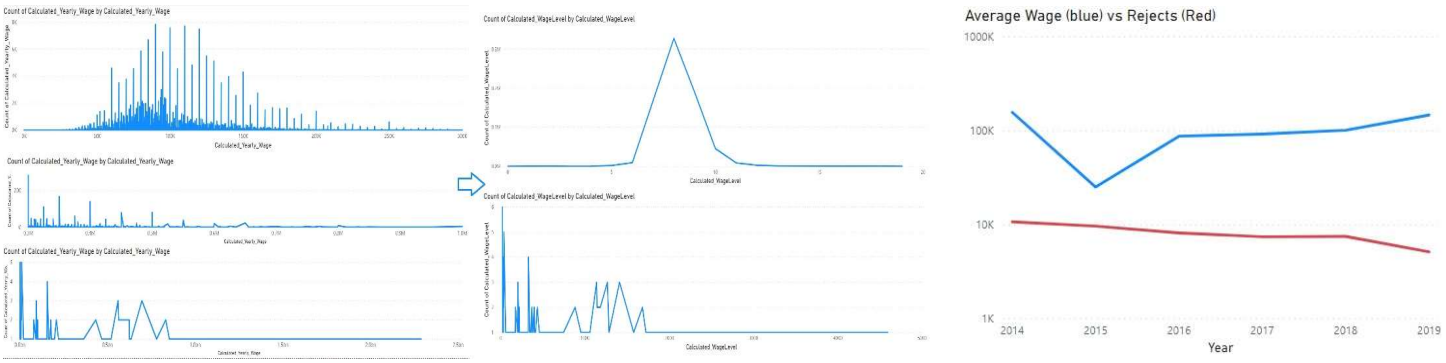

Fig: Wage distribution and normalization to wage levels.

Though there are some differences in wage features as well, for instance PREVAILING_WAGE that was not present earlier years, we can capture a common baseline to predict the wage and update the logic [15].

Logistic results in similar accuracy of more than 98.5% when it is run through the wage features in the new dataset [16] but with better accuracy on predicting true negatives. From normalization we could see an increase in rejection at the wage level 5-7 decrease from 7 to 14 as compared to approvals. We believe a high dimensional kernel can fit this pattern of data.

6.2 Naïve Bayes

To capture the true negative results in our model, the data is processed from a multiple text feature to build a dictionary of words that matter [12].

When running through Naïve Bayes model, with the addition of each feature there is improvement in accuracy and confusion matrix. The processed dataset [13] is run using features `JOB_TITLE, EMPLOYER_NAME, SOC_CODE, SOC_NAME` and `NAICS_CODE,` and this gives an accuracy of 98.49% which is slightly lower than logistic regression, but it captures the true negatives with confusion matrix:

	True Negative	True Positive
Predicted Negative	1466	3679
Predicted Positive	6660	577609

We processed above features with the new wage bucket information into the dictionary for naïve Bayes for years 2014-2019 [17] since they all retained the textual features of `JOB_TITLE, EMPLOYER_NAME, SOC_CODE, SOC_NAME` and `NAICS_CODE`. We saw accuracy of 94.37% and confusion matrix as:

	True Negative	True Positive
Predicted Negative	14554	34311
Predicted Positive	150135	3075885

6.3 Random Forest Model

For Random forest model, the model was generated using a 50% mix of approved and rejected cases. For this, entire rejected cases and randomly sampled equal count of approved cases was chosen. Although it has higher coverage of matching predicted negative with true negative, the accuracy of 64.21% is lower. Below chart shows the accuracy and confusion matrix overview:

	True Negative	True Positive
Predicted Negative	1831	1010
Predicted Positive	180609	323916

7. Conclusion and Future Work

We found the logistic model to be predicting high accuracy but hides the true negatives as it tries to fit the data. So, we believe that visa outcome is not as dependent on employer and job profile as we presumed in our null hypothesis, it has an element of random behavior in the decision. We captured individual company names, job titles and job categories to see if they are useful measures in modelling the accuracy. The result was a drop in total accuracy but higher level of predicting true negatives. We also evaluated random forest and logistic with modelling more features than were used in Naïve Bayes.

In future, the wage can be made to fit using kernel trick with a higher dimension to evaluate how good it fits the data to predict the outcome. Neural network and boosting can also be used for stronger learning from logistic, Naïve Bayes, SVM and Random Forest to predict the outcome.