H-1B Visa Applications Outcome Prediction

ECON481 Final Project Report

Ruobing Chen

1 Introduction

The H-1B is a non-immigrant working visa that allows foreign nationals who have graduated from a degree program in the United States to work in the United States for a specified period of time. This visa is often sought after by those in the fields of science, technology, engineering, and mathematics (STEM), as it can be extended for up to six years for those in these fields. However, the length of time an individual is permitted to stay on an H-1B visa can vary based on nationality. For example, citizens of Slovenia are only allowed to stay for a maximum of one year on an H-1B visa, while citizens of other countries may be allowed to stay for the full six-year maximum. The selection process for H-1B visas is random, with the United States Citizen and Immigration Services (USCIS) [1] using a computer-generated system to select the necessary number of visas to meet the annual quota. This lottery selection process can make it difficult to understand how the attributes of the applicants affect the outcome, as the selection is based on random chance rather than specific qualifications or characteristics of the applicants.

The H-1B visa program can be a useful way for foreign nationals with advanced degrees to work in the United States, particularly in STEM fields. However, the limited number of visas available and the random selection process can make it challenging for some applicants to obtain an H-1B visa. A prediction algorithm, as mentioned in the previous paragraph, could be a helpful resource for both H-1B applicants and their sponsoring employers in understanding the chances of obtaining an H-1B visa.

2 Exploratory Analysis

The data set in the 2018 and 2019 fiscal year for the algorithm was obtained from the Office of Foreign Labor Certification[2] by the United States of Labor with roughly 40,000 data points and 200 variables. Due to the time limitation, the project manually selected the 16 variables and the description of each variable is shown as below.

Variable Name	Variable Meaning	Dtype
CASE_NUMBER	Case ID	object
CASE_STATUS	Status	object
CASE_SUBMITTED	Submission Date	object
DECISION_DATE	Decision Date	object
FILLING_YEAR	Filling Year	int64
VISA_CLASS	Visa Class	object
JOB_TITLE	Job Title	object
SOC_CODE	Job Social Code	object
SOC_NAME	Job Social Name	object
FULL_TIME_POSITION	Full time position or not	object
EMPLOYER_NAME	Employer Name	object
AGENT_REPRESENTING_EMPLOYER	Have attorney or not	object
EMPLOYER_STATE	Employment State	object
EMPLOYER_CITY	Employment City	object
PREVAILING_WAGE	Wage	float64

2.1 Data Preprocessing

To obtain the best model, we first clean up the data with missing NaN values and remove H-1B1 Chile or H-1B1 Signgap data, hence the scope of the project is to focus on the H-1B visa application. In addition, all prevailing wage data is converted from type String o Float for further manipulation. The H-1B application has four outcomes: Rejected, Certified-Withdrawn, Withdrawn, and Certified. As shown in figure 1, the data is imbalanced with more outcomes on the Certified, and the objective of the project is to find the likelihood of eligibility. Therefore, the outcomes including Rejected and Certified-Withdrawn all treated as Denied. The target variable for the prediction will be only three outcomes: Certificated, Denied, and Withdrawn.

2.2 Data Visualization

1. H1-B Case Status: Figure 1 uses the bar plot to show that there are more data on the class certificates than other target outcomes. The significant difference demonstrates an imbalance of the data in the status with more outcomes on the certificated. This can be a problem when training a machine learning model in the future since it is not representative of the real-world problem and can lead to biased results.

2. Top 10 H-1B Visa Sponsor Employers: Figure 2 shows the top 10 employers who sponsor the H-1B for their employers. The outcome is not surprising since the top 3 employers in the plot are the biggest well-known recruiters in India with great outsourcing powers. Figure 3 indicates the sponsor with the top 10 highest certificated rates. Comparing two figures on the sponsor name, we can infer that the largest

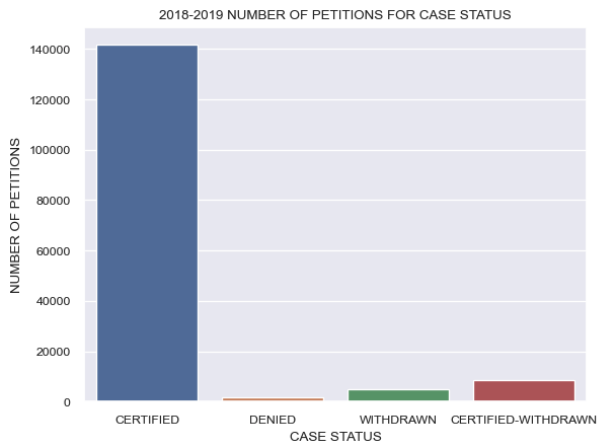

Figure 1: Number of petitions for case status

petition count may not necessarily come with the highest certification rate but there are other factors contributed to this.

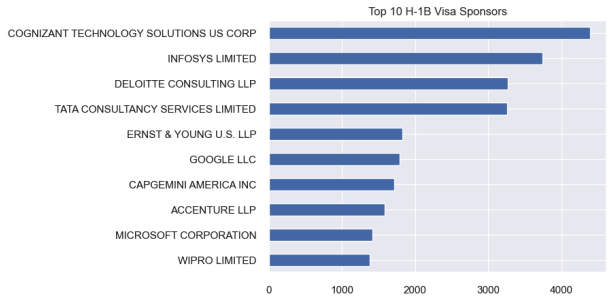

Figure 2: Top 10 H-1B visa sponsor

3. Wage Distribution: The histogram in Figure 4 shows the distribution of H-1B applicants' prevailing wage. The shape of the histogram is close to a bell-shaped curve but with a little skewed to the right. The mean prevailing wage is located in the center around 75,000, and the histogram shows the outliers located below 10,000. This histogram suggests that the majority of applicants in the group have wages that falls within a relatively narrow range, with few people's wages being significantly greater or less than the average wage.

2.3 Feature Engineering

1. Based on the above visualization, we need to solve the potential outliers issue on the prevailing wage data. Our approach will be first to find the wage in the 2 and 98 percentiles. Then we replace the min and max data with 2 percentiles and 98 percentiles wage.

2. Through the research on case status, we discovered the outcome Withdrawn are not depend on the decision of the applicant and petition, hence should not be served to predict the future outcome. Therefore, we choose to exclude the status Withdrawn and convert two target classes into binary variables where Certificated is 1 and Denied is 0.

	EMPLOYER_NAME	Acceptance_rate
2	TATA CONSULTANCY SERVICES LIMITED	0.998765
9	WIPRO LIMITED	0.997797
6	CAPGEMINI AMERICA INC	0.997661
10	AMAZON.COM SERVICES, INC.	0.995130
22	PRICEWATERHOUSECOOPERS ADVISORY SERVICES LLC	0.995098
12	IBM CORPORATION	0.994450
3	DELOITTE CONSULTING LLP	0.993458
7	ACCENTURE LLP	0.992322
18	JPMORGAN CHASE & CO.	0.991489
4	ERNST & YOUNG U.S. LLP	0.991223

Figure 3: Top 10 Highest Certification Rate H-1B visa sponsor

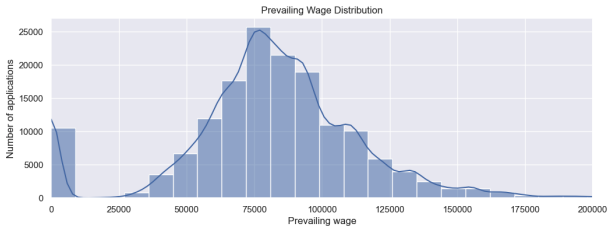

Figure 4: Top 10 H-1B visa sponsor

3. The descriptive statistics show that there are lots of unique values associated with job social code. To further use it in the classification model, we created a new feature name Occupation to cover the major occupation name and drop the column job title to avoid duplicated features.

4. The worksite information contains both City and State. We choose only to include State and convert to one-hot-k representation.

3 Methodology

In this study, we aimed to predict the outcome of H1B visa applications using machine learning techniques. To achieve this goal, we used three different models: logistics regression, decision tree, and random forest. We chose these models because they are widely used for classification tasks and have been shown to be effective in similar studies. In the following sections, we describe the data used in this study and the methodology we followed to train and evaluate the models. Our ultimate goal was to determine which model was the most accurate in predicting the outcome of H1B visa applications. Below is a brief description of the techniques based on the lecture material:

3.1 Logistic Regression

In logistic regression, the goal is to predict the probability that an example belongs to a certain class (e.g., "positive" or "negative") given its predictor variables. The probability that an example belongs to a certain class is calculated using the sigmoid function, which maps the output of the linear regression model to a value between 0 and 1. The sigmoid function is defined as follows:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{1}$$

where z is the output of the linear regression model:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{2}$$

In this equation, β_0 is the intercept term, and $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients that represent the influence of each predictor variable $x_1, x_2, \ldots, x_n$ on the probability of belonging to the "positive" class. These coefficients are learned during the training process using maximum likelihood estimation.

3.2 Decision Tree

A decision tree is a tree-like model used for classification and regression tasks. It works by dividing the training data into subsets based on the values of the attributes and using these splits to make predictions on new examples by following the branches of the tree.

Let T be a decision tree with root node r, n attributes $a_1, a_2, \ldots, a_n$, and a set of class labels C. To make a prediction on an example x with attribute values $x_1, x_2, \ldots, x_n$, we perform the following steps: Set the current node to r. While the current node is not a leaf node: If the value of the attribute a_v at the current node satisfies the condition specified by the node: Follow the left branch to the next node. Otherwise: Follow the right branch to the next node. Return the class label $c \in C$ associated with the current node as the prediction.

3.3 Random Forest

Random forests are a type of ensemble learning method for classification and regression tasks. They work by combining the predictions of multiple decision trees, which are trained on different subsets of the training data and using different subsets of the features. The final prediction is made by aggregating the predictions of the individual decision trees, typically using a simple majority vote or averaging the predicted probabilities.

4 Experiments Discussion

4.1 Results

Table 1: Experimental Results

	Balanced				Unbalanced			
	Tr. Acc.	Test Acc.	Prc.	Rcl.	Tr. Acc.	Test Acc.	Prc.	Rcl.
Logistics Regression	82.7%	82.7%	98%	85%	99.5%	99.46%	98%	99%
Logistics Regression L1	82.7%	82.8%	98%	68%	99.5%	99.46%	98%	99%
Logistics Regression L2	91.4%	91.6%	98%	85%	99%	99%	98%	99%
Decision Tree	97.1%	96.3%	99%	94%	98%	99%	99%	99%
Random Forest	97.7%	97.7%	98%	96%	99.4%	99.5%	98%	99%

Figure 5: Model Result

4.2 Model Discussion

In this study, we trained three models Logistics Regression, Decision Tree and Random Forest with the original data set by splitting it into 7:3 to create a training set and a test set. We then trained the logistic regression model on the training set and evaluated its performance on the test set. The results showed that the logistic regression model had a test F1 score of 0.99 a training F1 score of 0.99, and a precision of 0.98. The model also had a recall of 0.99. While the model seems to perform well on the test set with similar high training accuracy, the evaluation metrics show that the target variable Denied has an F-score of 0 where the F-score for Certificated is nearly 1.

Upon further investigation, the reason for this low f-score was due to imbalanced data. Specifically, there were significantly more instances of one class compared to the other, leading to skewed data distribution. This caused the model to heavily weigh one class, resulting in poor performance in the other class and a low f-score. To address this issue, the data were resampled to balance the distribution of classes by using the class-weight parameter that can be equal to "balanced" or a manual weight. It automatically adjusts the weight of each class based on the frequency of that class in the data. The rebalanced technique is applied to all models and helps to compensate for imbalanced data and improve the model's performance by 7 percent on the minority class. Further analysis is needed to identify potential ways to improve the performance of the minority class more effectively.

The table shows that the logistic regression model using the relative balance data indicates a lower F1 score. We sought to fine-tune our logistic regression model further using the L1 and L2 penalty. The L1 penalty, also known as the "lasso" penalty, adds a term to the cost function that is proportional to the absolute value of the coefficients. This encourages the model to assign small or zero coefficients to features that are not important for prediction. The L2 penalty, also known as the "ridge" penalty, adds a term to the cost function that is proportional to the square of the coefficients. This encourages the model to assign small coefficients to all features but does not encourage any coefficients

to be exactly zero. We found that using the L2 penalty resulted in the best performance for our logistic regression model. This is likely because the L2 penalty is less aggressive at assigning small coefficients to features, which can help prevent overfitting and improve the model's generalization.

In order to tune the Random Forest Model, the RandomizedSearchCV with 5-fold cross-validation is utilized to find the optimal hyperparameters including max_depth, min_samples_leaf, min_samples_split, and n_estimators. By setting the max depth to 8, the model has a better performance in capturing important patterns in the data. While the Decision Tree and Random Forest model seem to have similar performance, the Decision Tree is more effective at handling imbalanced data under this circumstance. The F-score for the minority class increased by 0.27 which indicates that the model is helpful in avoiding overlooking the minority class.

5 Conclusion & Future Work

In summary, it is possible to use machine learning models to predict the outcome of the H-1B Visa Application based on the applicant's information. Among the three models we tried using relatively balanced data, the results demonstrate both the random forest and decision tree models outperformed the logistic regression model. One potential reason is the random forest and decision tree models were able to handle better noise or outliers in the data, which may have impaired the performance of the logistic regression model. Another advantage of tree models is they are more generally robust to overfitting. The logistics models have relatively low recall which shows the model fails to predict a significant proportion of the actual positive cases as positive.

While the results of this study provide insights into the strengths and limitations of these models, there are several areas for future work that could help to improve their performance further. With more computational resources, the logistics regression with regularization needs to be fine-tuned with better parameters. It may be useful to explore more models such as Neural Networks and Support Vector Machines. Another direction for future work is to explore the use of resampling techniques to balance the classes in the data. This could include oversampling or undersampling the minority class, or using a combination of both approaches. By balancing the classes in the data, we may be able to improve the performance of the machine learning models and prevent them from being biased towards the majority class. Meanwhile, it is also crucial for the work to investigate the use of different feature engineering techniques to improve the quality of the input data. This may include techniques like feature selection, feature transformation, or feature extraction. By improving the input data, we may be able to achieve better results with any of the machine-learning models tested in this study.

References

[1] "The H-1B Visa, Explained." Boundless, www.boundless.com/immigration-resources/the-h-1b-visa-explained. Accessed 16 Dec. 2022.

[2] U.S. DEPARTMENT OF LABOR, H1-B Performance Data https://www.dol.gov/agencies/eta/foreign-labor/performance

[3] Predicting the Outcome of H-1B Visa Applications[online] from https://cs229.stanford.edu/proj2017/fin reports/5208701.pdf

[4] CASTRO , RAPHAEL. "Predicting Outcome for H-1B Eligibility in the US." Predicting Outcome for H-1B Eligibility in the US — Kaggle, 2017, www.kaggle.com/code/elraphabr/predicting-outcome-for-h-1b-eligibility-in-the-us.

Appendixes

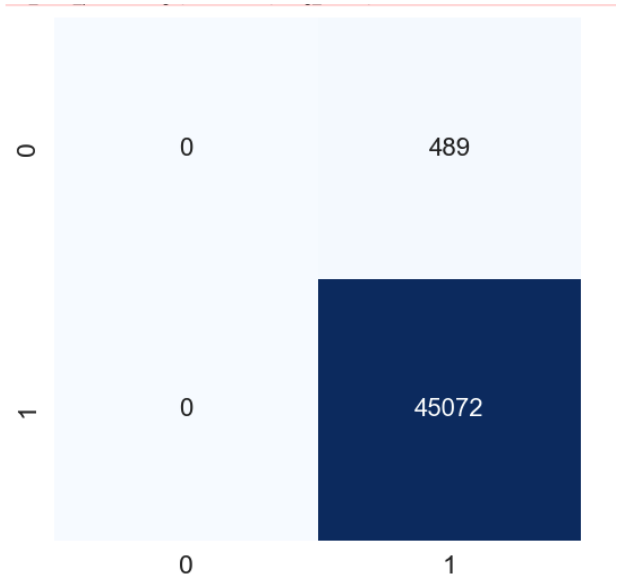

Figure 6: Imbalanced Data Confusion Matrix

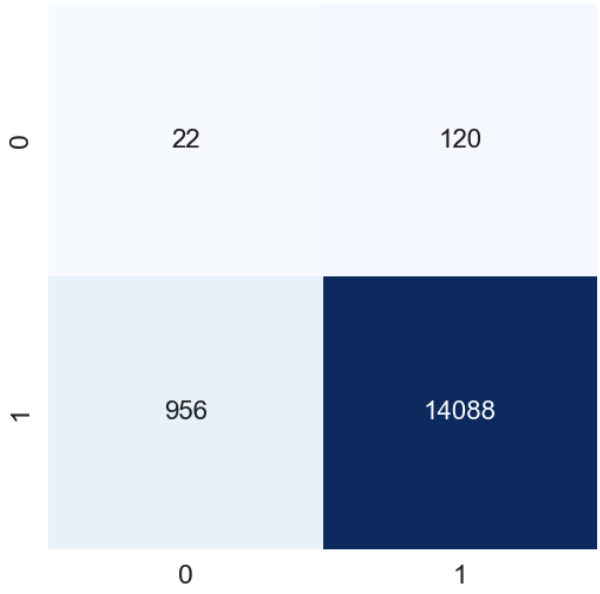

Figure 7: Balanced Data Confusion Matrix

```
              precision    recall  f1-score   support

           0       0.02      0.15      0.04       142
           1       0.99      0.94      0.96     15044

    accuracy                           0.93     15186
   macro avg       0.51      0.55      0.50     15186
weighted avg       0.98      0.93      0.95     15186

              precision    recall  f1-score   support

           0       0.16      0.98      0.27      1451
           1       1.00      0.94      0.97    135230

    accuracy                           0.94    136681
   macro avg       0.58      0.96      0.62    136681
weighted avg       0.99      0.94      0.96    136681

The f1 score for the training data: 0.9710614601638633
The f1 score for the testing data: 0.9632161903459593
```

Figure 8: Sample Classification Report