H-1B Labor Condition Application Case Status Prediction

1. Introduction

The H-1B is a visa in the United States under the immigration and Nationality Act, section which allows U.S. employers to employ foreign workers in specialty occupations. As talents from all over the world flow into the USA, the process of receiving the H-1B visa becomes more and more competitive. The first step to apply for an H-1B visa is Labor Condition Application. The LCA is a form that requires U.S. Department of Labor (DOL) approval in order for an employer to file an H-1B petition for a temporary professional worker. Due to the scarcity and uncertainty of H-1B release, people who intend to work in the USA always pay much attention on it and are eager to figure out the standards behind. The US Department of Labor does require some documents for LCA but we do not know what factors really affect the certification rate of LCA. In this project, we attempt to predict the status of the LCA petition based on the visa-application metadata. The intent of this study is to discover how LCA petition status is influenced by attributes of applicants. The classification models designed in this report could be utilized by H-1B aspirants, employers, and third-party agents who provide relevant service to increase the likelihood of certification and have a better understanding of LCA, before and after filing the petition.

2.1 Data retrieval and cleaning

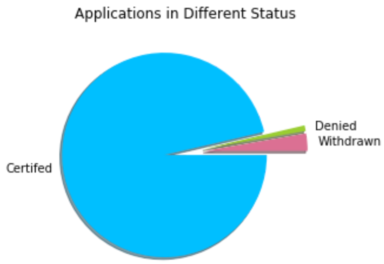
The data was downloaded from the website of the Office of Foreign Labor Certification: https://www.foreignlaborcert.doleta.gov/pdf/PerformanceData/2019/H-1B_Disclosure_Data_FY2019.xlsx.
We have 64617 observations in total with 52 variables. The variable descriptions can be seen in https://www.foreignlaborcert.doleta.gov/pdf/PerformanceData/2019/H-1B_FY19_Record_Layout.pdf
The raw data is messy, so we do the following cleaning and manipulation work:

1. Calculate annual salary.
2. Convert all date variable to datetime type.
3. When training classification model, consider only ‘CERTIFIED’ and ‘DENIED’ cases.
4. Create dummies for categorical variables.

2.2 Exploratory analysis

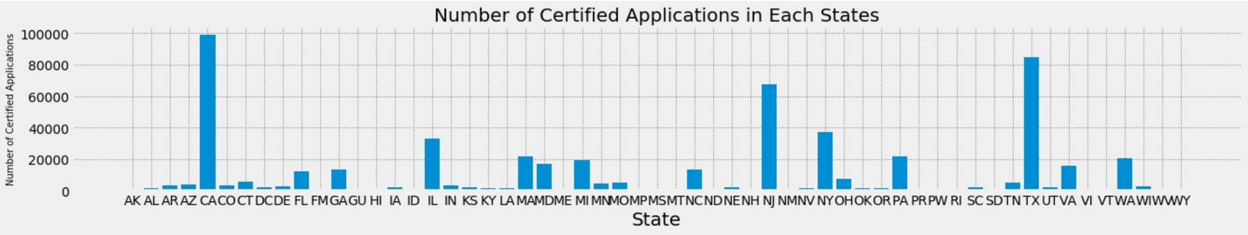
2.2.1 Visualize the proportion of various application statuses

As shown in the graph below, our outcome variables are quite imbalanced. The size of certified group outweighs those of the denied and withdrawn groups in a large scale. This shortcoming may affect the correctness of our prediction, as the prediction will always lean to certified. We will solve this later.



2.2.2 Visualize certified applications in different states

As shown, some particular states have a lot more applicants than others, and they are California, Texas, New York and New Jersey, places which are technology-intensive and capital-intensive.



2.3 Prediction model

2.3.1 oversampling

From the exploratory analysis, it is clearly seen that the number of certified cases exceeds the number of denied cases greatly (note that this is not the certification of H1B visa, it is the certification of LCA). If we use the imbalanced data to train classification model, especially logistic regression and tree models, the minor group in our project, the denied group, are likely to be ignored and the prediction result would tend to fall in certified group.

To solve this problem, we use synthetic minority oversampling techniques on our dataset. The basic idea of SMOTE is to synthetically create denied samples: for each existing denied observation in initial dataset, we find one of its denied neighbors and create a new denied sample using the combination of it and its neighbor.

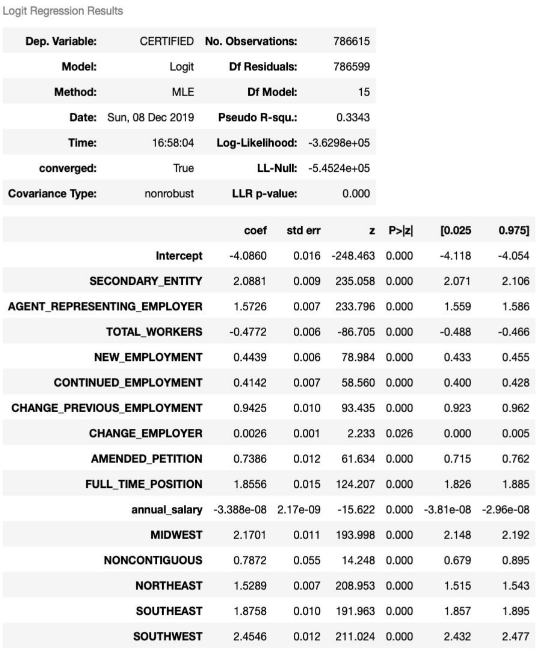
After SMOTE, we have 1123736 observations with half denied and half certified. We divide them into 70% training samples and 30% testing samples.

2.3.2 Logistic Regression

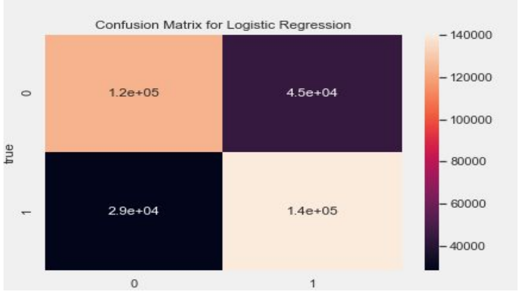
We start from logistic regression mainly because our desired dependent variable is binary. Besides, it is not only a relatively simple method but also can help us select features. Now let’s analyze the summary of model to find out significant variables. Logistic Regression Generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of certified or denied. At the beginning, we incorporated all potential features, then we selected all significant features. The final logistic regression is on the next page.

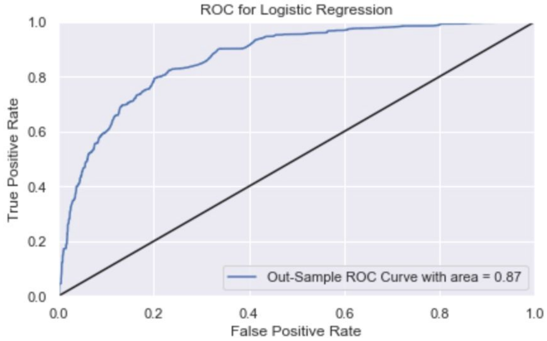
According to the results below, it is easy to find by the coefficient and p-value that except for “change_employer”, all other factors have statistically significant impact on the outcome variables.

We also display ROC curve and confusion matrix here. ROC Curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. Therefore, the closer the curve is to the left top corner, the better the performance is. The confusion matrix displays the prediction result compared with actual value. The lighter the color of the top-left and bottom-right parts, the better prediction the model does.



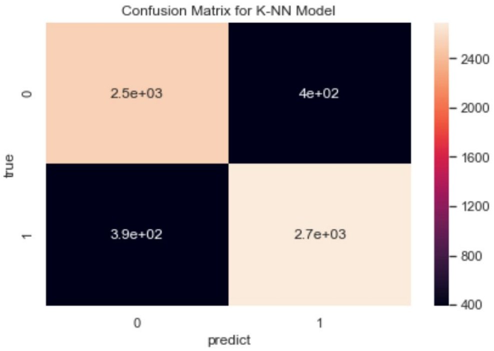
```
Logit Regression Results

        Dep. Variable:        CERTIFIED   No. Observations:      786615
               Model:             Logit       Df Residuals:      786599
              Method:               MLE           Df Model:          15
                Date:  Sun, 08 Dec 2019     Pseudo R-squ.:       0.3343
                Time:          16:58:04     Log-Likelihood: -3.6298e+05
           converged:              True            LL-Null: -5.4524e+05
     Covariance Type:         nonrobust        LLR p-value:       0.000

                                    coef   std err         z  P>|z|     [0.025     0.975]
                     Intercept   -4.0860     0.016  -248.463  0.000     -4.118     -4.054
              SECONDARY_ENTITY    2.0881     0.009   235.058  0.000      2.071      2.106
   AGENT_REPRESENTING_EMPLOYER    1.5726     0.007   233.796  0.000      1.559      1.586
                 TOTAL_WORKERS   -0.4772     0.006   -86.705  0.000     -0.488     -0.466
                NEW_EMPLOYMENT    0.4439     0.006    78.984  0.000      0.433      0.455
          CONTINUED_EMPLOYMENT    0.4142     0.007    58.560  0.000      0.400      0.428
    CHANGE_PREVIOUS_EMPLOYMENT    0.9425     0.010    93.435  0.000      0.923      0.962
               CHANGE_EMPLOYER    0.0026     0.001     2.233  0.026      0.000      0.005
              AMENDED_PETITION    0.7386     0.012    61.634  0.000      0.715      0.762
            FULL_TIME_POSITION    1.8556     0.015   124.207  0.000      1.826      1.885
                 annual_salary -3.388e-08  2.17e-09   -15.622  0.000  -3.81e-08  -2.96e-08
                       MIDWEST    2.1701     0.011   193.998  0.000      2.148      2.192
                NONCONTIGUOUS     0.7872     0.055    14.248  0.000      0.679      0.895
                    NORTHEAST     1.5289     0.007   208.953  0.000      1.515      1.543
                    SOUTHEAST     1.8758     0.010   191.963  0.000      1.857      1.895
                    SOUTHWEST     2.4546     0.012   211.024  0.000      2.432      2.477
```

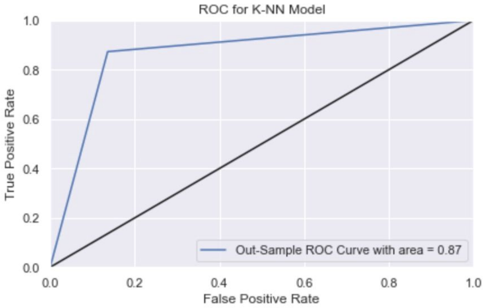




2.3.3 KNN

KNN is one of the most basic methods when doing classification. Computing time for KNN is highest among all the other methods for this dataset. So we choose a subsample to do this classification. K we choose here is 10 by muti-loops checking.

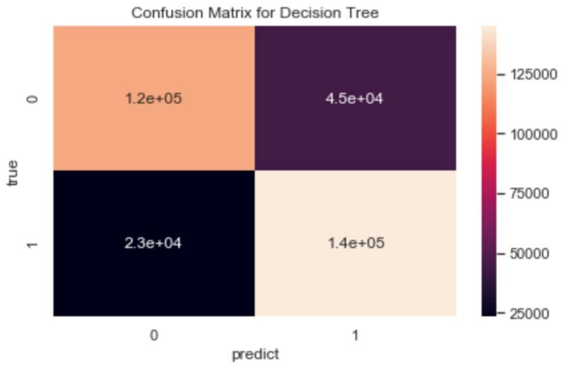


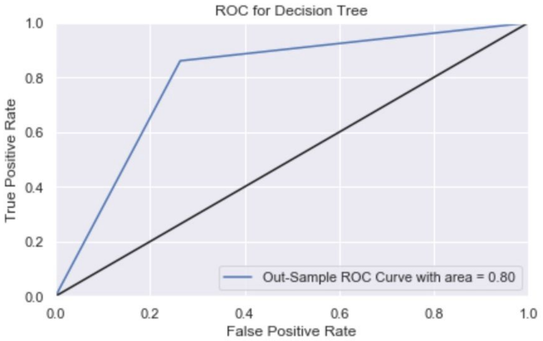


The result is good in terms of confusion matrix and ROC curve. However, the performance of KNN could be overestimated due to oversampling. When using SMOTE to do oversampling, we manufactured observations based on nearest neighbors. As a result, if we use k-nearest neighbors, the testing results would be affected heavily by manufactured samples.

2.3.4 Decision Tree

Decision tree works by splitting the dataset recursively, which means that the subsets that arise from a split are further split until a predetermined termination criterion is reached. At each step, the split is made based on the independent variable that results in the largest possible reduction in heterogeneity of the predicted variable. It creates binary decision trees and decision rules that has more sensitivity and specificity. (max_depth=5)



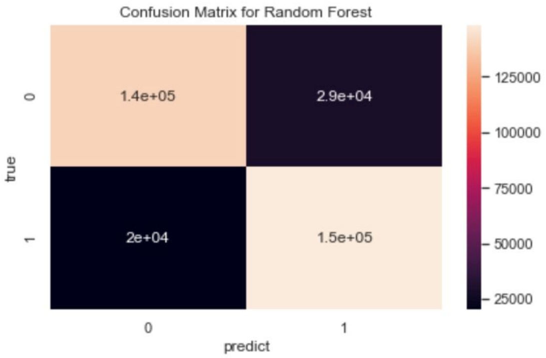


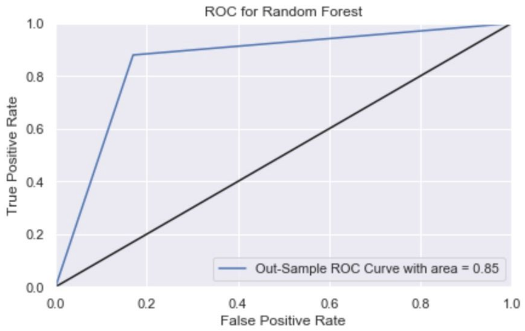
Based on the decision tree, we find that people fulfilling the following criteria are predicted to receive a certified decision:

1. secondary entity, not continued employment, not agent representing employer, not Southwest
2. secondary entity, not continued employment, not agent representing employer, Southwest, full-time position
3. secondary entity, not continued employment, agent representing employer
4. secondary entity, not amended position, full-time position
5. secondary entity, amended position, total workers more than 2

2.3.5 Random Forests

Random forests are an ensemble learning method by constructing a multitude of decision trees by selecting samples with replacements and features randomly. (n_estimators=1000, max_depth=5, max_features=5)





2.3.5 Summary of Classification

Below is a summary of classification results:

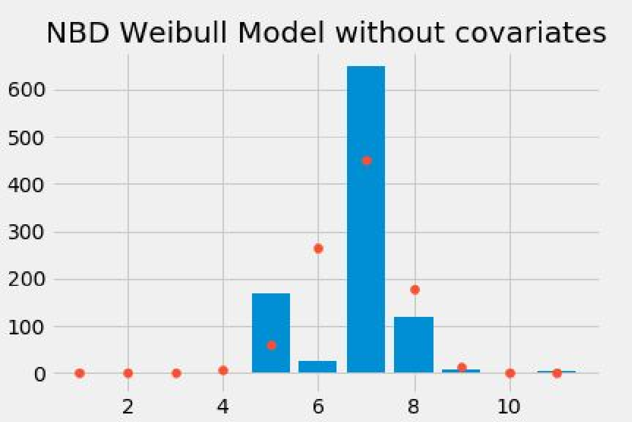
Model	**Accuracy**	**AUC**
Logistic Regression	0.78286	0.86887339
K-NN	0.869	0.86887396
Decision Regression	0.79857	0.7986
Random Forests	0.854897796	0.8549

It is clearly seen in the table that both Random Forests and K-NN has a good performance. However, as we have mentioned, the good performance of K-NN could result from Oversampling method we use to solve data imbalance. Therefore, we should conclude that random forests have the best performance.

2.4 Timing model —— Weibull-Gamma Model

Since waiting for the result of H-1B applications is an anxious process for all applicants, the time until a decision comes out is also an eye-catching topic. Here we use 3 models, Weibull model, Weibull-Gamma model and Weibull model with covariates. Below is the distribution of estimated waiting time and actual waiting time. Below is the result of parameters estimated:

Continued Employ	Secondary Entity	Agent presenting	Total worker	Change previous	Change employ	Amended petition	Full time position
0.681	-0.124	-0.189	-0.419	0.019	-0.024	-0.036	-1.31



3. Conclusion and Recommendations

In conclusion, we use 4 different models to help predict LCA petition certification rate. Among which, random forests have the best performance. According to our model, secondary entity, not continued employment applicants are more likely to get certified. An interesting thing is that salary seems not to be an important features in our model, which means applicants do not have to consider too much on salary for LCA.

Besides, we use timing model, Weibull model to predict waiting time. Based on this model, we can calculate hazard for an applicant. More specifically, if an application hasn't been decided, we can calculate the probability of getting processed in the next day. This can be displayed on the website, which could give applicants an expectation about how many more days they have to wait.

Link of Data：https://drive.google.com/open?id=1Ze4xggxU5jdQzaEoSaC5aRytWUOeStrD

Link of Video：https://drive.google.com/open?id=1u3n0UtKcW5K0S6O1XS37oUvMTvrU47Ck

Work by: Jiaxi Zhou(jz3150), Jiaqi Zhang(jz3155),Xiaoyu Chen(xc2525),Jiajie Li(jl5536)