Analysis based on H1_B Visa petition data and approval prediction

ORIE4741: Chengnan Xu(cx223), Tian Yu(ty364)

1. Problem definition

In our project, we want to predict the outcome of H1B_visa applicants based on the dataset. H1_B visa, non-immigration visa category for international students who want to work in America, is highly desired. It required petition employees at least a bachelor's degree. As more and more employees go into America, based on Citizenship and Immigration Services(USCIS) grants 85,000 H1_B visas[1] every year, the visa outcome for the applicants is of importance.

The input to our algorithm is based on the dataset features and we will also do some feature engineering such as converted to 0 or 1 for the feature. The output would be the case status whether certified or denied.

2. About the dataset

2.1 Dataset overview

Our columns features were divided into 2 types: 6 nominal variables, 4 continuous features and Our dataset is from Kaggle listed under the name "H-1B Visa Petitions 2011-2016 dataset". There are 10 features and more than 3 million examples in our dataset. The columns in the dataset include case status, employer name, worksite coordinates, job title, prevailing wage, occupation code, and year filed. The dataset contains some missing data and the data values are heterogeneous including continuous values, discrete values, nominal data and text.

	CASE_STATUS	EMPLOYER_NAME	SOC_NAME	JOB_TITLE	FULL_TIME_POSITION	PREVAILING_WAGE	FILING_YEAR	WORKSITE	LONGITUDE	LATITUDE
1	CERTIFIED-WITHDRAWN	UNIVERSITY OF MICHIGAN	BIOCHEMISTS AND BIOPHYSICISTS	POSTDOCTORAL RESEARCH FELLOW	N	36067.0	2016.0	ANN ARBOR, MICHIGAN	-83.743038	42.280826
2	CERTIFIED-WITHDRAWN	GOODMAN NETWORKS, INC.	CHIEF EXECUTIVES	CHIEF OPERATING OFFICER	Y	242674.0	2016.0	PLANO, TEXAS	-96.698886	33.019843
3	CERTIFIED-WITHDRAWN	PORTS AMERICA GROUP, INC.	CHIEF EXECUTIVES	CHIEF PROCESS OFFICER	Y	193066.0	2016.0	JERSEY CITY, NEW JERSEY	-74.077642	40.728158
4	CERTIFIED-WITHDRAWN	GATES CORPORATION, A WHOLLY-OWNED SUBSIDIARY OF TOMKINS PLC	CHIEF EXECUTIVES	REGIONAL PRESIDEN, AMERICAS	Y	220314.0	2016.0	DENVER, COLORADO	-104.990251	39.739236
5	WITHDRAWN	PEABODY INVESTMENTS CORP.	CHIEF EXECUTIVES	PRESIDENT MONGOLIA AND INDIA	Y	157518.4	2016.0	ST. LOUIS, MISSOURI	-90.199404	38.627003

Table.1 First 5 points from the unprocessed dataset

2.2 Dataset Characteristics

We processed some of the existing features, created new features that we thought could be useful for prediction and discarded some features using the library Pandas. For example, in our training data, some original features like SOC NAME is not directly related to our CASE_STATUS, Therefore, we calculated the success rate and total number of applications to create new highly correlated features The detailed explanation is below.

CASE_STATUS: We only maintain 'CERTIFIED' and 'DENIED' in CASE_STATUS, with labeled as '1' and '0'.

FULL_TIME_POSITION: For the full time position column, there are two cases, {Y, N} for applicants who have full time position or not. We labeled “FULL TIME POSITION” = “Y” as 1 and “FULL TIME POSITION” = “N” as 0.

EMPLOYER_ACCEPTANCE: We created a feature for the ratio of H-1B applicants who were certified per employer.

EMPLOYER_ACCEPTANCE: Also categorized into six types for ratio of acceptance and converted the data into one-hot-k representation. We created a feature for the success rate per SOC type,converted to one-hot-k representation.

JOB_ACCEPTANCE: We created a feature for the success rate per Job type, converted to one-hot-k representation.

WORKSITE: For the format of {City, State}, we only included ”State” and converted the data into one-hot-k representation.

WAGE_CATEGRORY: we created a feature for the prevailing wage which is the average wage paid to employees with similar qualifications. Category the average wage into five kinds{very low, low, medium, high, very high} and then converted to one-hot-k representation.

FILING_YEAR: columns Year the applicants were filed, we converted data of different year into type int and applied one-hot-k representation.

2.3 Data Exploration

Our label CASE_STATUS has seven different categories. To get more information about the CASE_STATUS, we plot the histogram of case status versus number of petition of the visa petition.

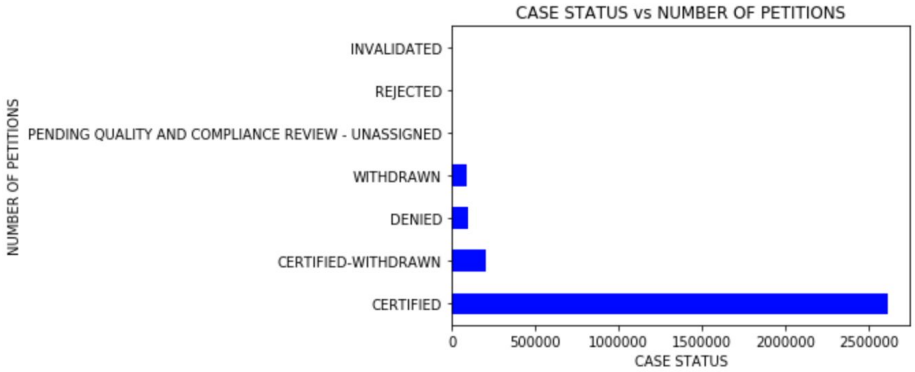

Fig. 1.1 the histogram of seven classes in CASE_STATUS

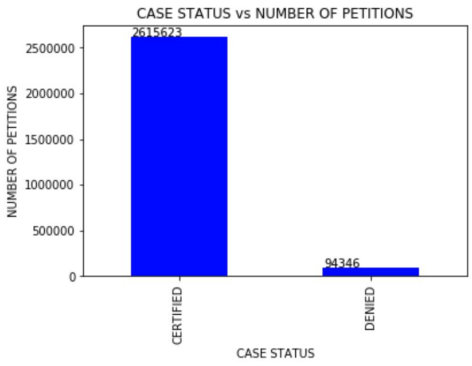

Fig.1.2 imbalanced classes in CASE_STATUS

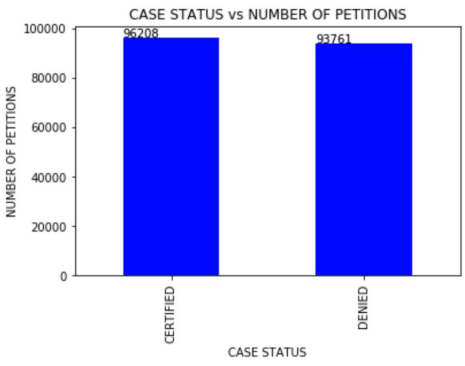

Fig.1.3 equal ratio after undersampling

From Fig.1, we observe that the total amount of some CASE_STATUS outcome classes are relevant quite low. We decided to discard the 'INVALIDATED,' 'REJECTED', 'PENDING QUALITY AND COMPLIANCE REVIEW - UNASSIGNED' class because of the trivial amount.

Then i compute the top employers sponsoring H1'B, we can see the top are indians' company which maybe they want to hire more technique people to increase companies' profits.

```
INFOSYS LIMITED                                    130241
TATA CONSULTANCY SERVICES LIMITED                  64358
WIPRO LIMITED                                      43679
DELOITTE CONSULTING LLP                            36667
ACCENTURE LLP                                      32983
IBM INDIA PRIVATE LIMITED                          28166
MICROSOFT CORPORATION                              22373
HCL AMERICA, INC.                                  22330
ERNST & YOUNG U.S. LLP                             18217
LARSEN & TOUBRO INFOTECH LIMITED                   16724
CAPGEMINI AMERICA INC                              16032
COGNIZANT TECHNOLOGY SOLUTIONS U.S. CORPORATION    15448
GOOGLE INC.                                        12545
IGATE TECHNOLOGIES INC.                            12196
IBM CORPORATION                                    10690
Name: EMPLOYER_NAME, dtype: int64
```

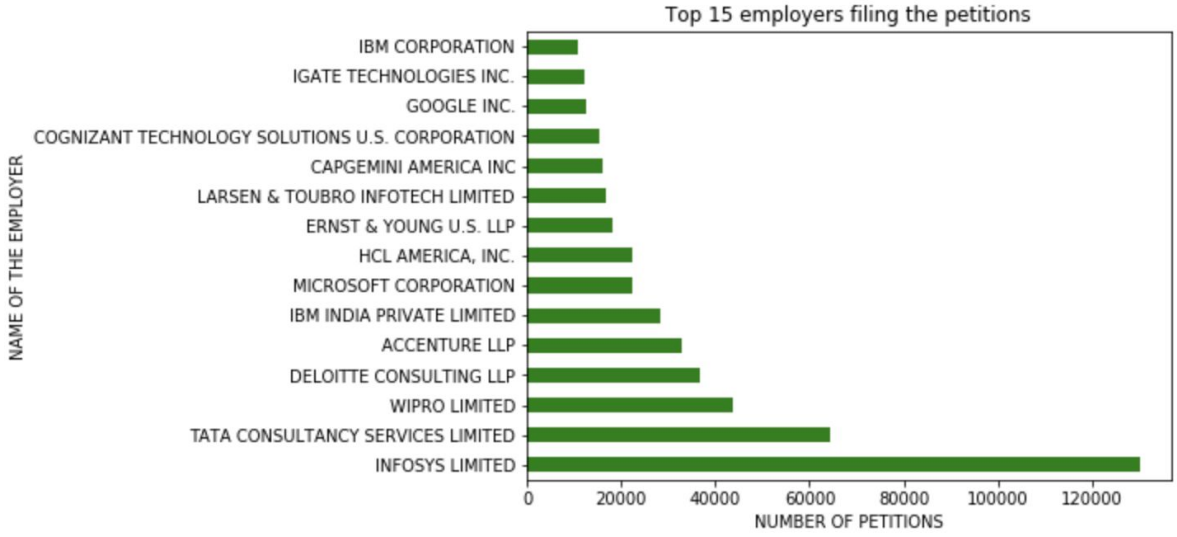

Figure 2 Top employers sponsoring H1'B

2.4 Cleaning the Data

As with any dataset, there were many missing values. By applying pandas.dataframe.isnull(), we statistically summarized the missing data in the dataset and we remove rows with null values for "EMPLOYER_NAME", "SOC_NAME", "JOB_TITLE", "FULL_TIME_POSITION", "PREVAILING_WAGE" by applying dataframe.dropna. The missing data is like Employer) name is missing about 20 numbers, Job title is about 0, full_time_position is about 1, the prevailing wage is about more than 50;

We also plot the histogram of CASE_STATUS vs number of petitions. From Fig.2.1, we can see the dataset is highly imbalanced, with only 3.5% of case status being declined. The imbalanced dataset may contribute to model's overfitting problem, the problem and solution will be elaborated in the preliminary models part.

After the pre-processing steps described above, we split the training and test sets 80:20. Training set had a total of 2.4 million examples. Due to the inherent bias in our dataset towards the "CERTIFIED" label(Fig. 2.1), the model learning from class imbalanced training dataset performed badly in the test datatest. The precision of "CERTIFIED" class is up to 95% while the precision of "DENIED" class is lower than 50%. The model is overfitting.

To address imbalanced classes problem, undersampling the majority class is a good choice, especially when we have millions of rows in our project dataset. Undersampling can be defined as removing some observations of the majority class and it can only be applied after splitting test and train set. After resampling, we have an equal ratio of 'CERTIFIED' data points to 'DENIED' data points(Fig. 2.2) , but a smaller quantity of data to train machine learning models on.

2.5 Feature selection

Since the preprocessed dataset in one-hot-encoding format has too many features. We need to eliminate unimportant feature. We employ the RFE(Recursive Feature Elimination) function in sklearn. The operating principle of the RFE is as follows. First, the estimator is trained on the initial set of features and the importance of each feature is obtained. Then, the least important features are pruned from the current set of features. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached. By employing RFE, we finally use the features with key words of 'EMPLOYER', 'FILING', 'FULL', 'JOB', 'SOC', 'WAGE', and 'WORKSITE'.

3. Models

3.1 Logistic regression

Logistic regression is a statistical machine learning algorithm that classifies the data by considering outcome variables on extreme ends and tries makes a logarithmic line that distinguishes between them.

The logistic regression is suitable for the classification scenario. It has the low calculation cost and very efficient in terms of time and memory requirements. However, this model has relatively low classification accuracy and not suitable for the dataset with missing data or large feature space.

In logistic regression, we mainly pay attention to the hyperparameter of C and class_weight. A smaller of C values specify stronger regularization. And class_weight of the 'auto' modeclass_weight can over/undersample by given weights.

The confusion matrix of the logistic regression and result indicators are as follows.

```
[[14722  4033]
 [ 1907 29293]]
              precision    recall  f1-score   support

           0       0.89      0.78      0.83     18755
           1       0.88      0.94      0.91     31200

    accuracy                           0.88     49955
   macro avg       0.88      0.86      0.87     49955
weighted avg       0.88      0.88      0.88     49955
```

Figure.3 Confusion matrix and results indicators of logistic Regression

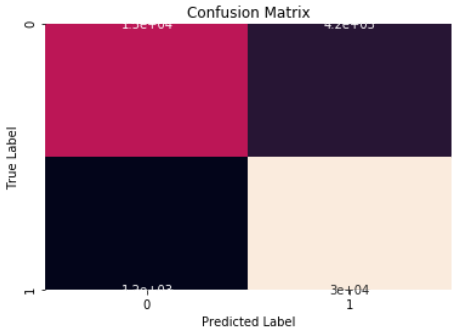

Figure 4 Heat map of the confusion matrix of logistic Regression

3.2 Random forest classifier

Random forest is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classification or mean prediction of the individual trees. There are several advantages[2] of this method:

- Small number of data loss does not change the regression model
- The high generalization ability
- Avoid the probability of overfitting
- Do not need to choose the hyperparameters specifically

For the random forest model, there are also some parameters that we need to consider:

- Number of features: the quantity of features in the model
- N estimators: the performance of the model
- Min sample leaf: the ability of data capturing
- Criterion: the validation criterion

While we implement the random forest model, we take the following values for our regression model to find the optimal prediction result. For the estimators, with the higher number of estimators can result in better performance in regression and prediction, but simultaneously, it can also bring the fact that the processing time will be much longer.

Parameters	Features	Estimators	min_sample_leaf	Criterion
Values	6	90	2	mse

Table. 2 The optimal parameter of the random forest classifier

The confusion matrix of the logistic regression and result indicators are as follows.

```
[[14722  4033]
 [ 1907 29293]]
              precision    recall  f1-score   support

           0       0.89      0.78      0.83     18755
           1       0.88      0.94      0.91     31200

    accuracy                           0.88     49955
   macro avg       0.88      0.86      0.87     49955
weighted avg       0.88      0.88      0.88     49955
```

Figure. 5 Confusion matrix and results indicators of random forest classifier

3.3 Support Vector classifier

The Support Vector Machine can also be used as a classification method, maintaining all the main features that characterize the algorithm which is the maximal margin. There are several advantages[3] that the SVR algorithm has, for instance, SVR model tends to have strong generalization ability and can reduce the probability of overfitting. There are several kernels in this model that needed to consider, they are as follows:

- Linear Kernel
- Polynomial Kernel
- Radial-basis Function(RBF) Kernel
- Sigmoid Kernel

In the support vector regression model, the RBF kernel is also known as the Gaussian Kernel, after we compared those four different kernels, we found out that the RBF kernel is the best fit for our model. And the formula for this kernel function is:

In this kernel function, we need to focus on two different hyperparameters, which is:

- C: Cost parameter of error item
- Gamma: The parameter in the kernel function

In this model, in a gesture to avoid overfitting, we used the Cross Validation with k-ford equals 5 to find the optimal hyperparameters in the Support Vector Regression model. And the optimal hyperparameters we got based on the input of our training data. I firstly use default value to get the accuracy. Since the dataset is too large, then I use 2000 data and used in test data, get the accuracy about 85%. Then I use heat map to find the optimal parameter based on different C and gamma and choose kernel as ‘RBF’

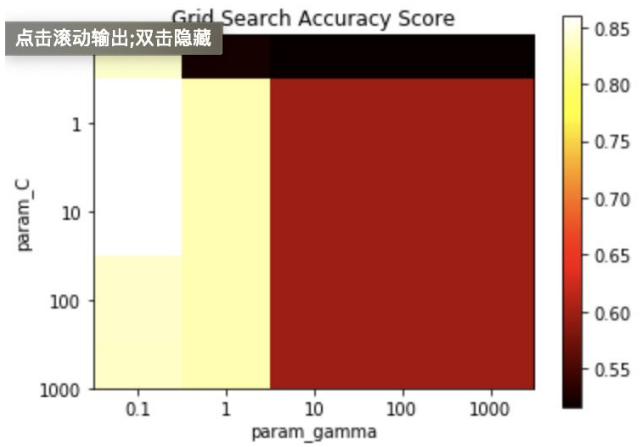

Figure. 6 Heat map for different parameter(C, gamma)

We could see that the C equals to 10 and gamma equals to 0.1, getting the accuracy about 87.3%.

3.4 Gaussian Naive Bayes classifier

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem.

The confusion matrix of the logistic regression and result indicators are as follows.

```
[[ 7777 10978]
 [ 2608 28592]]
              precision    recall  f1-score   support

           0       0.75      0.41      0.53     18755
           1       0.72      0.92      0.81     31200

    accuracy                           0.73     49955
   macro avg       0.74      0.67      0.67     49955
weighted avg       0.73      0.73      0.71     49955
```

Figure.7 Confusion matrix and results indicators of gaussian naive bayes

3.5 Model effectiveness comparison

From the bar plot, we can observe the logistic regression has a better performance on the H-1B Visa Petitions 2011-2016 dataset with prediction accuracy of 89.1%

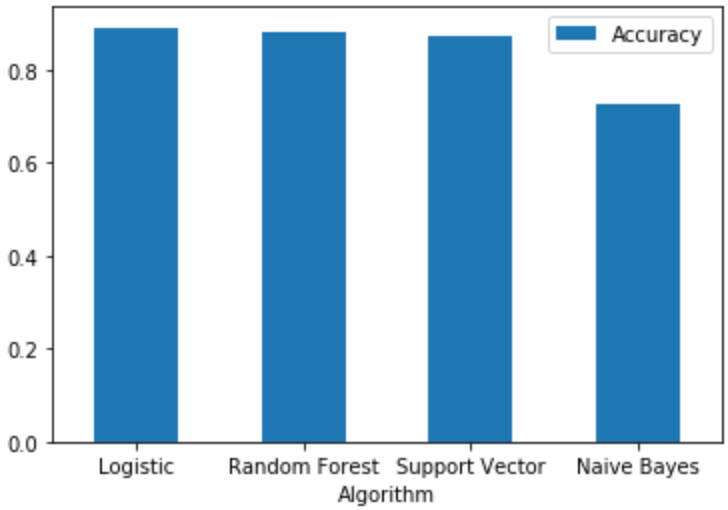

Figure.8 Prediction accuracy comparison

4. Conclusion

Using techniques[4] we learned in class(K-fold cross validation, logistic regression, support vector machines) and also some other techniques learned outside(random forest), we were able to find models that can predict whether one will experience denied or approved for the H1B application given the dataset we used.

We do feel confident that our results can be used in production to see whether one will be denied based on the dataset information. Our project results can be used by government or non-profit agencies to help predict the people such that non-native students in america have the opportunities to work here and start preparing to be higher probabilities to be approved such like working in some information science related job or find the bigger company like google. In the future, there is potential to add more features to help increase the accuracy.

5. Reference

[1] https://webpages.uncc.edu/sshinde5/

[2] K.-Luckyson, S. Saha, and S-R.Dey, “Predicting the direction of stock market prices using random forest” *Applied Mathematical Finance*, 2016.

[3] C.Y.Yeh, C-W. Huang, and Shie-J.Lee, “A multiple-kernel support vector regression approach for a stock market price forecasting,” *Science Direct.*, 2010.

[4] Andrew NG, “Machine Learning, Stanford University course”.