H1_B Visa petition data analysis and approval prediction

Chengnan Xu(cx223), Tian Yu(ty364)

1 Problem definition

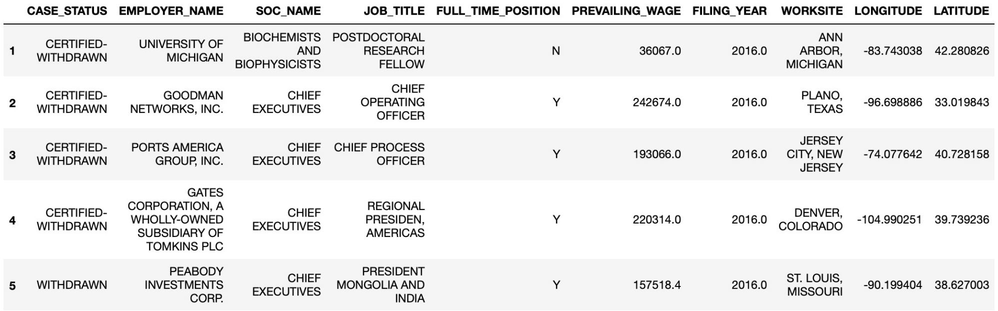
In our project, we want to predict the outcome of H1B_visa applicants based on the dataset. H1_B visa, non-immigration visa category for international students who want to work in America, is highly desired. It required petition employees at least a bachelor's degree. As more and more employees go into America, based on Citizenship and Immigration Services(USCIS) grants 85,000 H1_B visas every year, the visa outcome for the applicants 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 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 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1	CERTIFIED-WITHDRAWN	UNIVERSITY OF MICHIGAN	BIOCHEMISTS AND BIOPHYSICISTS	POSTDOCTORAL RESEARCH FELLOW	N	36067.0	2016.0	ANN ARBOR, MICHIGAN	-83.743038	42.280826
2	CERTIFIED-WITHDRAWN	GOODMAN NETWORKS, INC.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3	CERTIFIED-WITHDRAWN	PORTS AMERICA GROUP, INC.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4	CERTIFIED-WITHDRAWN	GATES CORPORATION, A WHOLLY-OWNED SUBSIDIARY OF TOMKINS PLC	CHIEF EXECUTIVES	REGIONAL PRESIDEN, AMERICAS	Y	220314.0	2016.0	DENVER, COLORADO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5	WITHDRAWN	PEABODY INVESTMENTS CORP.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

Fig. 1: First 5 points from the unprocessed dataset

2.2 Feature engineering

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.

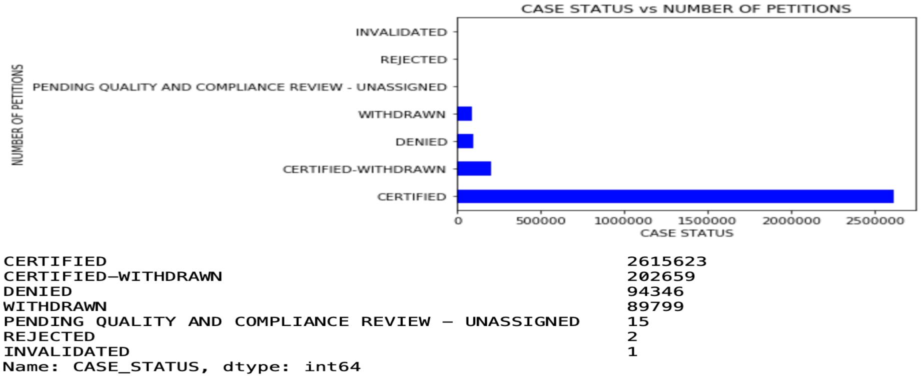
WORKSITE: For the format of { City, State} , we only included ” State” and converted the data into one-hot-k representation.

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 Exploratory data analysis

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.



```
CERTIFIED                                             2615623
CERTIFIED-WITHDRAWN                                   202659
DENIED                                                94346
WITHDRAWN                                             89799
PENDING QUALITY AND COMPLIANCE REVIEW - UNASSIGNED    15
REJECTED                                              2
INVALIDATED                                           1
Name: CASE_STATUS, dtype: int64
```

Fig. 2 the histogram of seven classes in CASE_STATUS

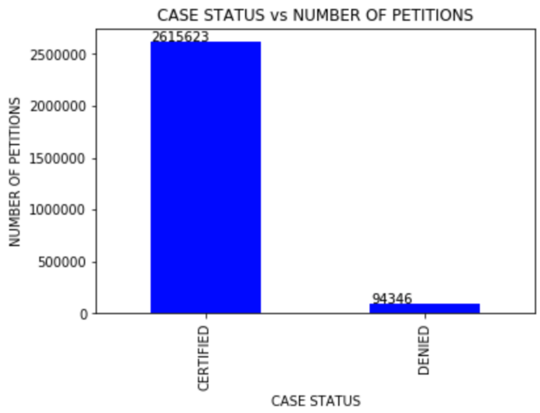


Fig.3 imbalanced classes in CASE_STATUS

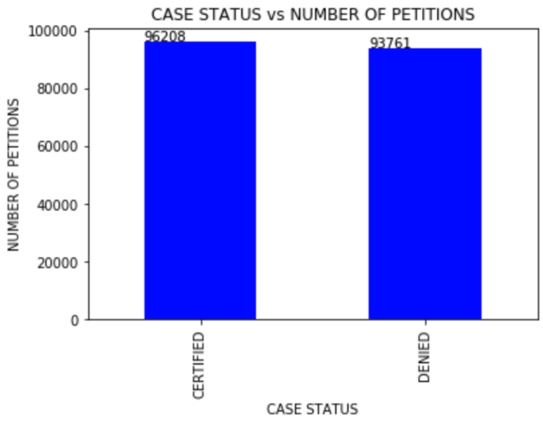


Fig.4 equal ratio after undersampling

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.

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 apply dataframe.dropna

We also plot the histogram of CASE_STATUS vs number of petitions. From Fig.2, 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.

3 Preliminary Models

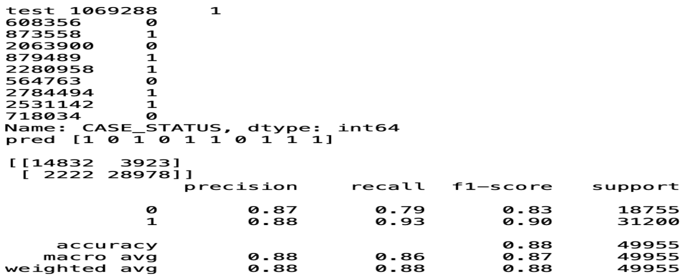
3.1 Overfitting and underfitting

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), the model learning from 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. 3) , but a smaller quantity of data to train machine learning models on.

3.2 Decision tree and model effectiveness

In order to test our model performance, we calculate the confusion matrix by the help function in the sklearn library. We also used the metric of precision, recall, and f1-score to evaluate the model performance on “CERTIFIED” and “DENIED” class.



```
test 1069288     1
608356      0
873558      1
2063900     0
879489      1
2280958     1
564763      0
2784494     1
2531142     1
718034      0
Name: CASE_STATUS, dtype: int64
pred [1 0 1 0 1 1 0 1 1 1]

 [[14832  3923]
  [ 2222 28978]]
              precision    recall  f1-score   support

           0       0.87      0.79      0.83     18755
           1       0.88      0.93      0.90     31200

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

Fig.5 the metrics of the decision tree performance

The overall accuracy is about 88 percent.

4 Next steps

Moving forward we will fit alternative models to the dataset such as SVM, random forest, as well as other models. We also want to determine the best prediction model for the data, also apply more techniques we learned in class such as ridge regression or LASSO to help predict better. Besides, we will use like crose validation or included k-fold to help predict well. Comparing with each methods’ performance, we will choose the most optimal one for H1B_Visa approval prediction for applicants.