Final Report

Team Member: Min Dai, Jie Lu, Ying Di

INTRODUCTION

Our project is to predict the case status of LCA (Labor Condition Application) of H-1B application submitted by the employer to hire non-immigrant workers under the H-1B visa program.

We are interested in this project because H-1B is very important for all the non-immigrants who want to work in America. We think employers, employees and government can benefit from the ML model.

- Employer can know the process of getting their LCA approved.
- Employee can use our tool to figure out their changes of having their H-1B application approved and they could also use our models to figure out what portions of their applications are lowering their chances of succeeding.
- There are a lot of applications each year, government can use this model to reduce their work.

DATASET

We used a 97MB dataset from kaggle. The dataset is published by U.S. Department of Labor. It contains administrative data from employer's Labor Condition Applications, and the H1B certification decisions made by Department's Office of Foreign Labor Certification, Employment and Training Administration. The dataset contains data that determination was issued from 10/1/2016 to 6/30/2017.

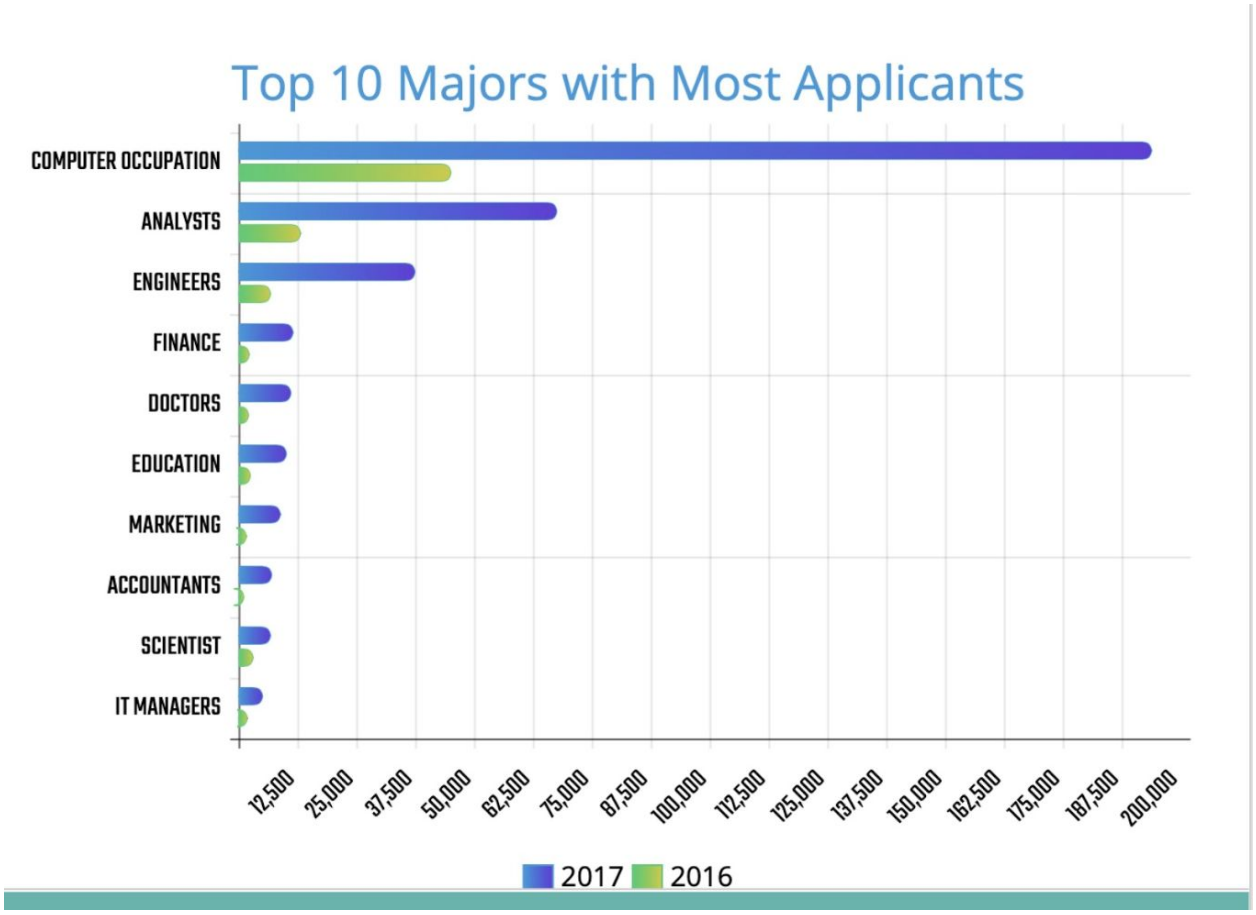
DATA EXPLORATION/UNDERSTANDING

Number of applicants in different state

```
df['WORKSITE_STATE'].value_counts()
```

```
CA    88573
TX    50262
NY    39925
```

Conclusion: California has most applicants because we guess California has a lot of non-immigrants since it’s Silicon Valley.



Conclusion: 50% of the applicants from Computer Occupation. Seems like computer is a hot job for non-immigrants.

The top 10 majors with highest deny rate

CASE_STATUS	CERTIFIED	DENIED	Total	Perc
REPORTERS AND CORRESPONDENTS	41	55	96	57.29%
REAL ESTATE	59	73	132	55.30%
PUBLIC RELATIONS	337	254	591	42.98%
MULTIMEDIA ARTISTS AND ANIMATORS	301	96	397	24.18%
MECHANICS	79	25	104	24.04%
HEALTHCARE	935	294	1229	23.92%
COUNSELORS	289	90	379	23.75%
SALES AND RELATED WORKERS	191	49	240	20.42%
RELIGIOUS WORKERS	47	10	57	17.54%
SCIENTIST	7499	1355	8854	15.30%

Conclusion: It seems like these job has highest deny rate. If people work is belonged to any of these, they will probably be denied.

Mean and median wages for certified and denied cases:

	PREVAILING_WAGE		WAGE_RATE_OF_PAY_FROM		wage_diff	
	mean	median	mean	median	mean	median
CASE_STATUS						
CERTIFIED	71861.934456	68952.0	81622.842755	75000.0	9760.908299	3008.00
DENIED	64101.767578	63794.0	69480.145537	65958.0	5378.377959	0.99

(Note: wage_diff = actual_wage - prevailing_wage)

Conclusion: PREVAILING_WAGE is the standard wage for this job, WAGE_RATE_OF_PAY_FROM is the employer's proposed wage. The chart shows the mean, median and difference of certified and denied wage. If applicants' WAGE_RAGE_OF_PAY_FROM is lower than PREVAILING_WAGE, they most likely be denied. If applicants' real wage is lower than the mean or median wage of PREVAILING_WAGE or WAGE_RATE_PAY_FROM, they also likely to be denied.

DATA PREPROCESSING

The dataset has some missing values, we dropped the cases with missing values.
We also dropped some duplicate features and unnecessary features for predicting.
Except for int and float, there are also object data type in the dataset, which is not allowed in the training model. To fix this issue, we used Label Encoder(for CASE_STATUS) and get_dummies()(for other features) to convert data in object type into numbers.

MODELING

We applied three classification models in the training process: Decision Tree Model, Random Forest Model, and Naive Bayes Model.
When we trained the model, we seperated the dataset into a train set (containing 60% of total data) and a test set (containing 40% of total data). We set stratify=y in the train model function, so the proportion of values in the sample produced will be the same as the proportion of original dataset.

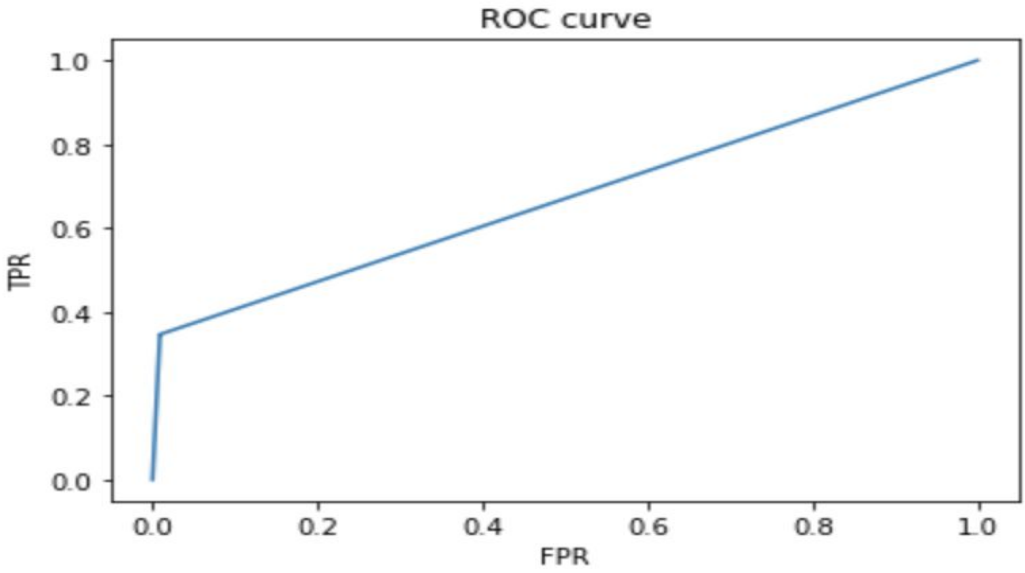
EVALUATION

As we can notice from the evaluation result and ROC curve for the 3 models we used, all models have a very high accuracy score(>98%). However, when we look at roc auc score, the 3 models are not as good as we expected. Decision Tree model got the best score among the 3. Naive

Bayes got only 0.5 which means this model makes random guesses on our dataset.

In terms of precision-recall, we need to compare the scores based on our use case. Our goal is to predict h1b result. For government, they want to use our model to save time and increase efficiency. So government will care more about the certified cases. They do not want to let any unqualified applicant get approval. Therefore, decision tree model is the best one for government as it has the lowest false negative value. But for applicants, they hope that their case would not be denied if it's qualified. So they want the lowest false positive. At this time, random forest would be the best choice for them.

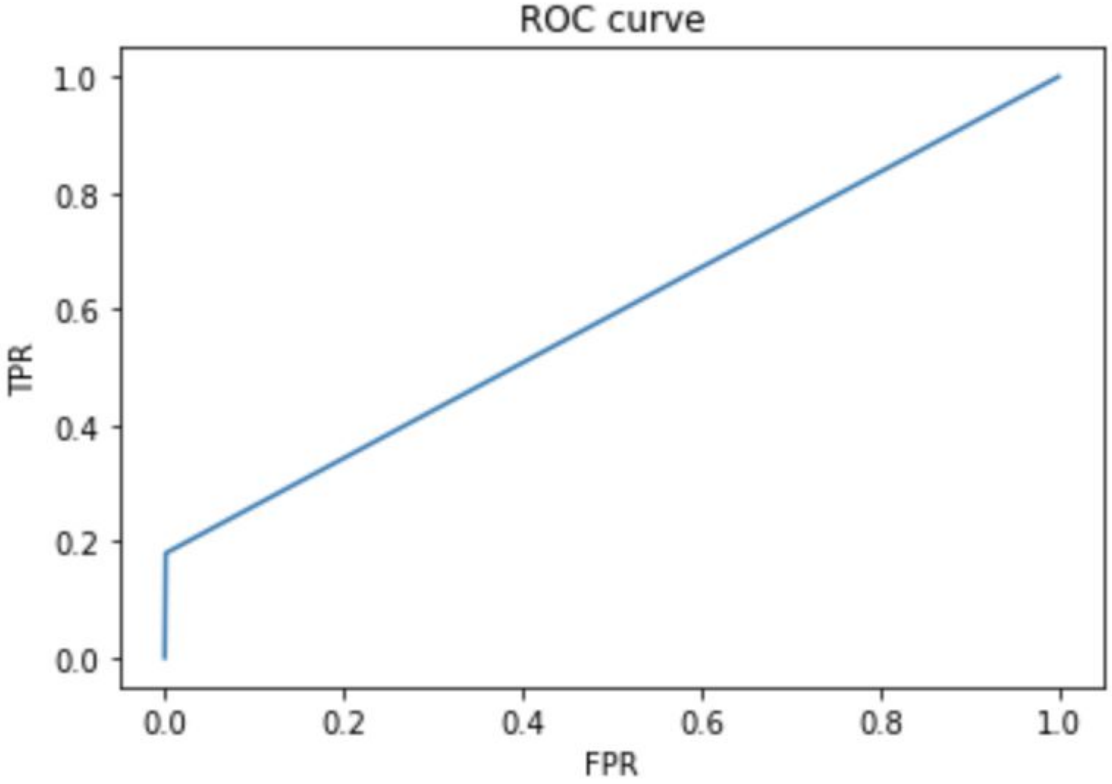
```
Decision Tree
model score:
0.9825397080832301
roc auc score:
0.6687283615874008
              precision    recall  f1-score   support

           0       0.99      0.99      0.99    183947
           1       0.34      0.35      0.34      2476

   micro avg       0.98      0.98      0.98    186423
   macro avg       0.67      0.67      0.67    186423
weighted avg       0.98      0.98      0.98    186423

confusion matrix
[[182313   1634]
 [  1621    855]]
```



Random Forest
model score:
0.9880701415597861
roc auc score:
0.5893380741132036

	precision	recall	f1-score	support
0	0.99	1.00	0.99	183947
1	0.70	0.18	0.29	2476
micro avg	0.99	0.99	0.99	186423
macro avg	0.84	0.59	0.64	186423
weighted avg	0.99	0.99	0.98	186423

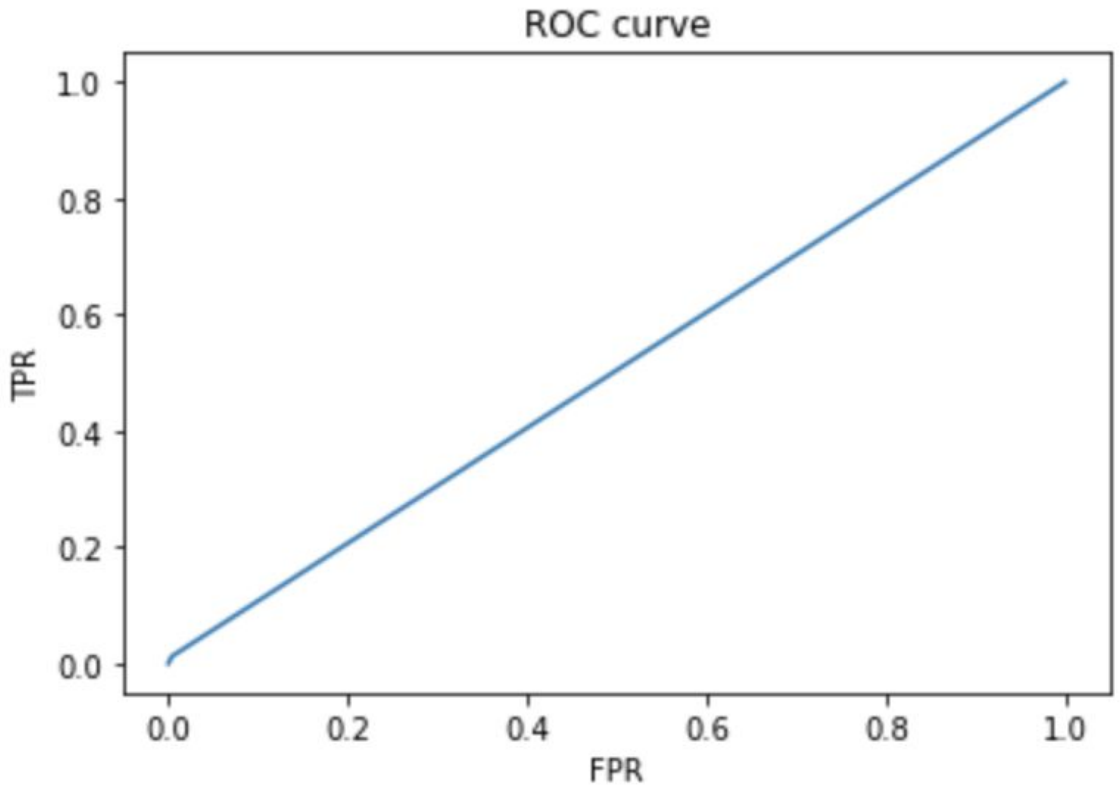
confusion matrix
[[183754 193]
[2031 445]]


ROC curve
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0

Naive Bayes
model score:
0.9830332094215842
roc auc score:
0.504109227377806

	precision	recall	f1-score	support
0	0.99	1.00	0.99	183947
1	0.04	0.01	0.02	2476
micro avg	0.98	0.98	0.98	186423
macro avg	0.51	0.50	0.51	186423
weighted avg	0.97	0.98	0.98	186423

confusion matrix
[[183230 717]
[2446 30]]



We also used K-fold cross validation model to evaluate the accuracy of the models: (K=5)
For each model, K-fold gave out 5 test scores, we calculated the mean test score.

```
Cross-validated scores for Decision Tree: [0.98247716 0.9817083  0.98219107 0.98329966 0.98294174]
average Cross-validated score for Decision Tree: 0.9825235858949861

Cross-validated scores for Random Forest: [0.98773401 0.98794858 0.98818102 0.98789493 0.98778743]
average Cross-validated score for Random Forest: 0.9879091951559476

Cross-validated scores for Naive Bayes: [0.98388971 0.9825308  0.98281689 0.98301357 0.98290598]
average Cross-validated score for Naive Bayes: 0.9830313905840932
```

From the results, we can see that all the models have high K-fold evaluation score. Among them, Random Forest score has slightly better score than the other models. Generally, all three models have high accuracy.

SERVICE

POST /predict — Entrypoint to our prediction function

Parameters

Parameter	Value	Description	Parameter Type	Data Type
model_select	Classification_Tree	**Model to use for predict**	formData	string
case_submitted_day	1	**Case submitted day**	formData	integer
case_submitted_month	1	**Case submitted month**	formData	integer
case_submitted_year	(required)	**Case submitted year**	formData	integer
soc_name	ACCOUNTANTS	**Occupation**	formData	string
naics_code	(required)	**Industry code associated with the employer For examples: 541511, 611110**	formData	integer
total_workers	(required)	**Total number of foreign workers requested by the employer**	formData	integer
full_time_position	Y	**If full time position, Y or N**	formData	string
prevailing_wage	(required)	**Prevailing wage for the job being requested for temporary labor condition.**	formData	double
pw_unit_of_pay	Bi-Weekly	**The pw unit**	formData	string
pw_source	CBA	**The pw source**	formData	string
pw_source_year	(required)	**Year the prevailing wage source was issued.**	formData	integer
wage_rate_of_pay_from	(required)	**Employer's proposed wage rate.**	formData	double
wage_unit_of_pay	Bi-Weekly	**The wage unit**	formData	string
h1b_dependent	Y	**If h1b dependent, Y or N**	formData	string
willful_violator	Y	**If employer is a willful violator, Y or N**	formData	string
worksite_state	AK	**The worksite state**	formData	string

Response Messages

HTTP Status Code	Reason	Response Model	Headers
200	Classification performed successfully		
500	Unable to perform the classification		

Try it out!

The above pictures show our service. We have health and predict function. Health is used to test if the service works or not. Predict is used to predict case status.
For predict function, we have model_select, which allows user to choose from classification tree, random forest and naive bayes to predict the case status. Then user input the necessary information such as case submitted day, month, year. Finally, clicking the 'Try it out' button to predict the case status.
We made dropdown button for most of the available input so that user can easily choose it without typing. It also can avoid invalid input.

CONCLUSION

We think this service can be used in 2 ways.
First, Employers can add this service in their internal website for H-1B application. When they collected all the information of the employee, they can directly input this information to check the if this case will be certified or not. Employee should also have access to this service.
Second, government can also add this service to their internal website. Officer who deal with the LCA can use this first or later to evaluate their work.

Final thoughts: If we had a long time to improve the project, what we will do next are:

1. Train as many models as we can, evaluate these models so we can find the best one.
2. Do more analysis of these features to find the deep relations between the specific features and result.
3. Improve the UI to make it more interactive. We can prompt corresponding error if the input is invalid.