OBTAINING AN H-1B VISA

Katrina Apiado
Zoe Atkins
Sharmaine Parani

TABLE OF CONTENTS

01

INTRODUCTION

Background, Problem Statement, Data Description

02

DATA

Description, Collection, Cleaning & Processing

03

MODELS & ANALYSIS

Decision Tree, Random Forest, AdaBoost, Analysis

04

CONSIDERATIONS

Ethical Considerations, Limitations

05

CONCLUSION

Final Thoughts, Future Recommendations

06

REFERENCES

Sources

EXPRESS MAIL

01

INTRODUCTION

6 PM ---
PARIS, FR
---JULY 17

WHAT IS AN H-1B VISA?

DESCRIPTION:

The H-1B visa is a nonimmigrant work visa that allows U.S. employers to hire foreign workers for specialty jobs that require a bachelor's degree or equivalent.

ELIGIBILITY:

- A job offer from a U.S. employer for a role that requires specialty knowledge
- Proof of a bachelor's degree or higher equivalent in that field
- Employer must show that there is a lack of qualified U.S. applicants for the role

V — 6 - PM — 17 JULY 098

PROBLEM STATEMENT

For nonimmigrant workers seeking to reside in the U.S. temporarily, applying to the H-1B program is a complex process. We wanted to see **if there were specific features that heavily affected whether an application would be certified or denied**.

With this project, we created and analyzed models that will **predict the likelihood of an international applicant obtaining an H-1B visa**, based on their descriptive features like starting wage, employer, etc.

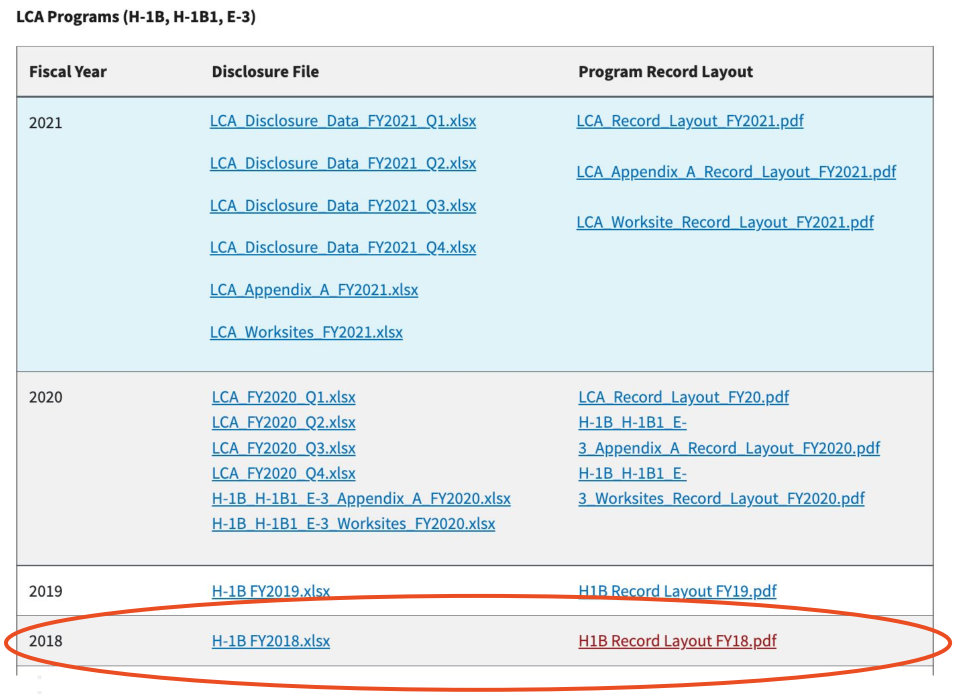
02 DATA

Collection, Cleaning, & Processing

FIRST DAY ISSUE

DATA COLLECTION AND PROCESSING

- Collected from the **U.S. Department of Labor**
  - Found under LCA programs
  - Fiscal Year "H-1B FY2018.xlsx"
- **Classification** problem



LCA Programs (H-1B, H-1B1, E-3)

Fiscal Year	Disclosure File	Program Record Layout
2021	LCA_Disclosure_Data_FY2021_Q1.xlsx LCA_Disclosure_Data_FY2021_Q2.xlsx LCA_Disclosure_Data_FY2021_Q3.xlsx LCA_Disclosure_Data_FY2021_Q4.xlsx LCA_Appendix_A_FY2021.xlsx LCA_Worksites_FY2021.xlsx	LCA_Record_Layout_FY2021.pdf LCA_Appendix_A_Record_Layout_FY2021.pdf LCA_Worksite_Record_Layout_FY2021.pdf
2020	LCA_FY2020_Q1.xlsx LCA_FY2020_Q2.xlsx LCA_FY2020_Q3.xlsx LCA_FY2020_Q4.xlsx H-1B_H-1B1_E-3_Appendix_A_FY2020.xlsx H-1B_H-1B1_E-3_Worksites_FY2020.xlsx	LCA_Record_Layout_FY20.pdf H-1B_H-1B1_E-3_Appendix_A_Record_Layout_FY2020.pdf H-1B_H-1B1_E-3_Worksites_Record_Layout_FY2020.pdf
2019	H-1B FY2019.xlsx	H1B Record Layout FY19.pdf
2018	H-1B FY2018.xlsx	H1B Record Layout FY18.pdf

DATA CLEANING

- **Drop Null Values**
  - Specific to which columns were significant by personal choice
- **Convert Object to Float/Int**
  - Label Encoding
    - States, Y/N
- **Region Assignment**
  - States → West, Midwest, Northeast, South
- **Column Split**
  - SOC_CODE delimiter = hyphen
  - first 2 = industry | last 4 = specific occupation in industry
- **Keep # of characters in string**
  - First 5 characters of POSTAL_CODES
  - First 3 characters of NAICS_CODE for industry
- **Extracting Month from CASE_SUBMITTED & EMPLOYMENT_START_DATE**

BOSTON
SEPT.1ST
N.Y.

U.S. DEPARTMENT OF HOMELAND SECURITY

U.S. Citizenship
and Immigration
Services

1

Dataset

573,512

Rows

27

Descriptive Features

SEPT.1ST
*** BELGIQUE ***
~ 1940 ~

DEPARTMENT OF LABOR
UNITED STATES OF AMERICA

DATA SUMMARY

The dataset used for training and testing has 25 input variables, with CASE_STATUS of either "**CERTIFIED**" or "**DENIED**" as the output.

- **CASE_STATUS** (output) = Status associated with the last significant event or decision. Valid values include “Certified” and “Denied.”
- **CASE_SUBMITTED_MONTH** = Month the application was submitted.
- **EMPLOYMENT_START_MONTH** = Beginning month of employment.
- **EMPLOYER_STATE** = State information of the Employer requesting temporary labor certification
- **EMPLOYER_POSTAL_CODE** = Postal code information of the Employer requesting temporary labor certification.
- **AGENT_REPRESENTING_EMPLOYER** = Y = Employer is represented by an Agent or Attorney; N = Employer is not represented by an Agent or Attorney.
- **TOTAL_WORKERS** = Total number of foreign workers requested by the Employer(s).
- **NEW_EMPLOYMENT** = Indicates requested worker(s) will begin employment for new employer, as defined by USCIS I-29.

DATA SUMMARY

The dataset used for training and testing has 25 input variables, with CASE_STATUS of either "CERTIFIED" or "DENIED" as the output.

- **CONTINUED_EMPLOYMENT** = Indicates requested worker(s) will be continuing employment with same employer, as defined by USCIS I-29.
- **CHANGE_PREVIOUS_EMPLOYMENT** = Indicates requested worker(s) will be continuing employment with same employer without material change to job duties, as defined by USCIS I-29.
- **NEW_CONCURRENT_EMP** = Indicates requested worker(s) will begin employment with additional employer, as defined by USCIS I-29.
- **CHANGE_EMPLOYER** = Indicates requested worker(s) will begin employment for new employer, using the same classification currently held, as defined by USCIS I-29.
- **AMENDED_PETITION** = Indicates requested worker(s) will be continuing employment with same employer with material change to job duties, as defined by USCIS I-29.
- **FULL_TIME_POSITION** = Y = Full Time Position; N = Part Time Position.
- **WAGE_RATE_OF_PAY_FROM** = Employer’s proposed wage rate.

DATA SUMMARY

The dataset used for training and testing has 25 input variables, with CASE_STATUS of either "CERTIFIED" or "DENIED" as the output.

- **WAGE_UNIT_OF_PAY** = Unit of pay. Valid values include “Hour", "Week", "Bi-Weekly", "Month", or "Year".
- **H1B_DEPENDENT** = Y = Employer is H-1B Dependent; N = Employer is not H-1B Dependent.
- **WILLFUL_VIOLATOR** = Y = Employer has been previously found to be a Willful Violator; N = Employer has not been considered a Willful Violator.
- **SUPPORT_H1B** = Y = Employer will use the temporary labor condition application only to support H-1B petitions or extensions of status of exempt H-1B worker(s); N = Employer will not use the temporary labor condition application to support H-1B petitions or extensions of status for exempt H-1B worker(s);
- **LABOR_CON_AGREE** = Y = Employer agrees to the responses to the Labor Condition Statements as in the subsection; N = Employer does not agree to the responses to the Labor Conditions Statements in the subsection.
- **WORKSITE_STATE** = State information of the foreign worker's intended area of employment.

DATA SUMMARY

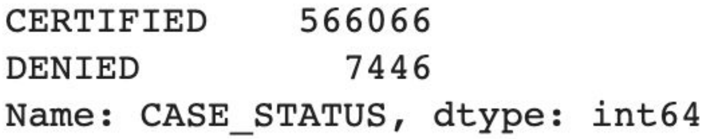
The dataset used for training and testing has 25 input variables, with CASE_STATUS of either "CERTIFIED" or "DENIED" as the output.

- **WORKSITE_POSTAL_CODE** = Zip Code information of the foreign worker's intended area of employment.
- **SOC_CODE2** = Occupation industry; Standard Occupational Classification
- **SOC_CODE4** = Specific role within industry; Standard Occupational Classification
- **NAICS_CODE3** = Subsector information; North American Industry Classification System
- **EMPLOYER_REGION** = Region information of the Employer requesting temporary labor certification
- **WORKSITE_REGION** = Region information of the foreign worker's intended area of employment.

98.7%

Of the outcome is 'CERTIFIED' for CASE_STATUS

Data is *highly* imbalanced → Stratified *k*-fold.



```
CERTIFIED    566066
DENIED         7446
Name: CASE_STATUS, dtype: int64
```

03 MODELS & ANALYSIS

Predictive Modeling for Decision Tree, Adaboost, Random Forest, Results

FIRST DAY ISSUE

OUR 3 CLASSIFICATION MODELS

DECISION TREE

Stratified K-Fold
Hyperparameter Tuning
Random Search CV
Classification Report
Confusion Matrix
Feature Importance
Predictive Performance

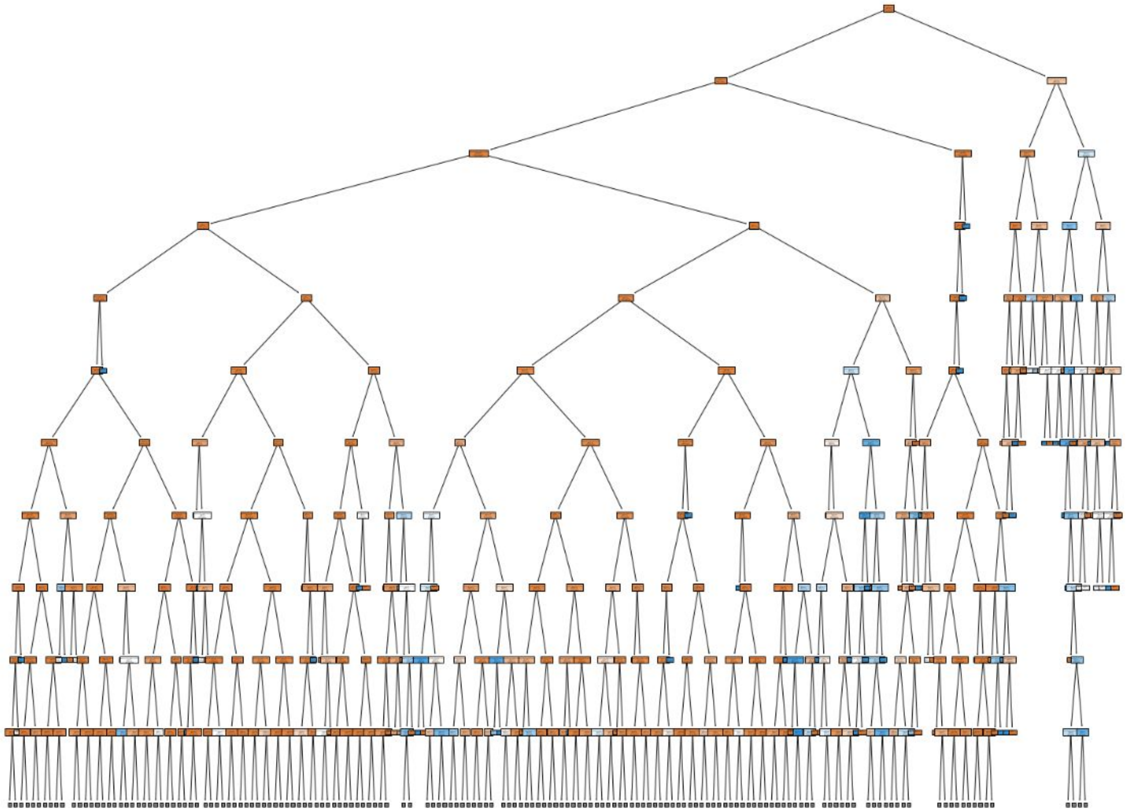
ADABOOST

Stratified K-Fold
Hyperparameter Tuning
Random Search CV
Classification Report
Confusion Matrix
Feature Importance
Predictive Performance

RANDOM FOREST

Stratified K-Fold
Hyperparameter Tuning
Random Search CV
Classification Report
Confusion Matrix
Feature Importance
Predictive Performance

DECISION TREE BEFORE HYPERPARAMETER TUNING

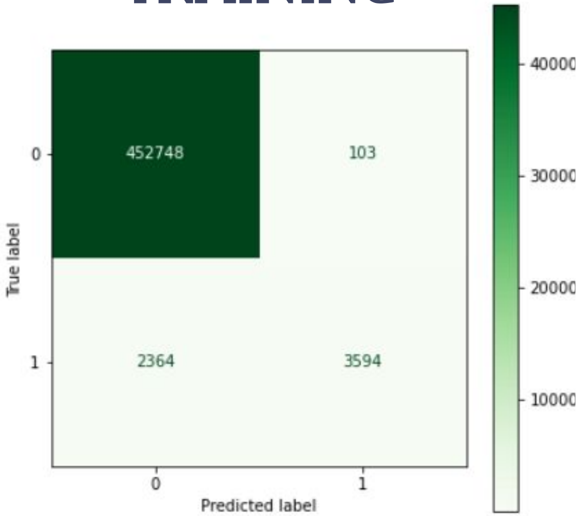


Initial Model:

- max_depth = 10
- feature_names = train_X.columns
- class_names = ['CERTIFIED','DENIED']

INITIAL CONFUSION MATRIX AND CLASSIFICATION REPORT

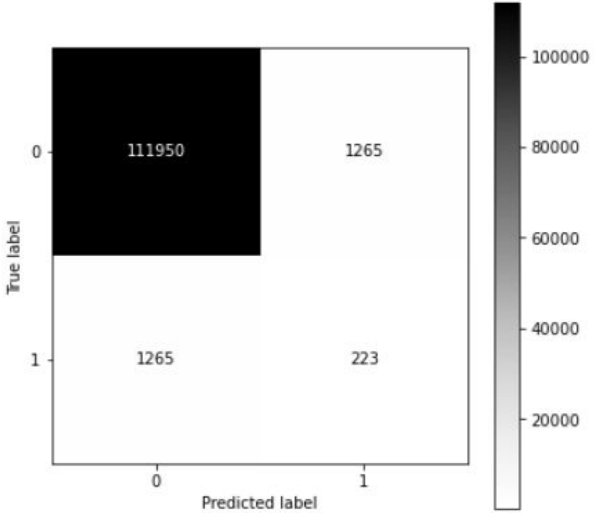
TRAINING



Macro: 0.87

Weighted: 0.98

TESTING



Macro: 0.57

Weighted: 0.98

Confusion Matrix:

- **Training set performance** (green) should be near perfect but there are flaws which are affected by imbalanced data
- **Testing set performance** (black) does worse, as there are more False Positives (FP) and False Negatives (FN)

Classification Report:

- Highly imbalanced # instances
- Macro Average F1-score = 0.57
- Overfitting data

HYPERPARAMETER TUNING

Initial Guess:

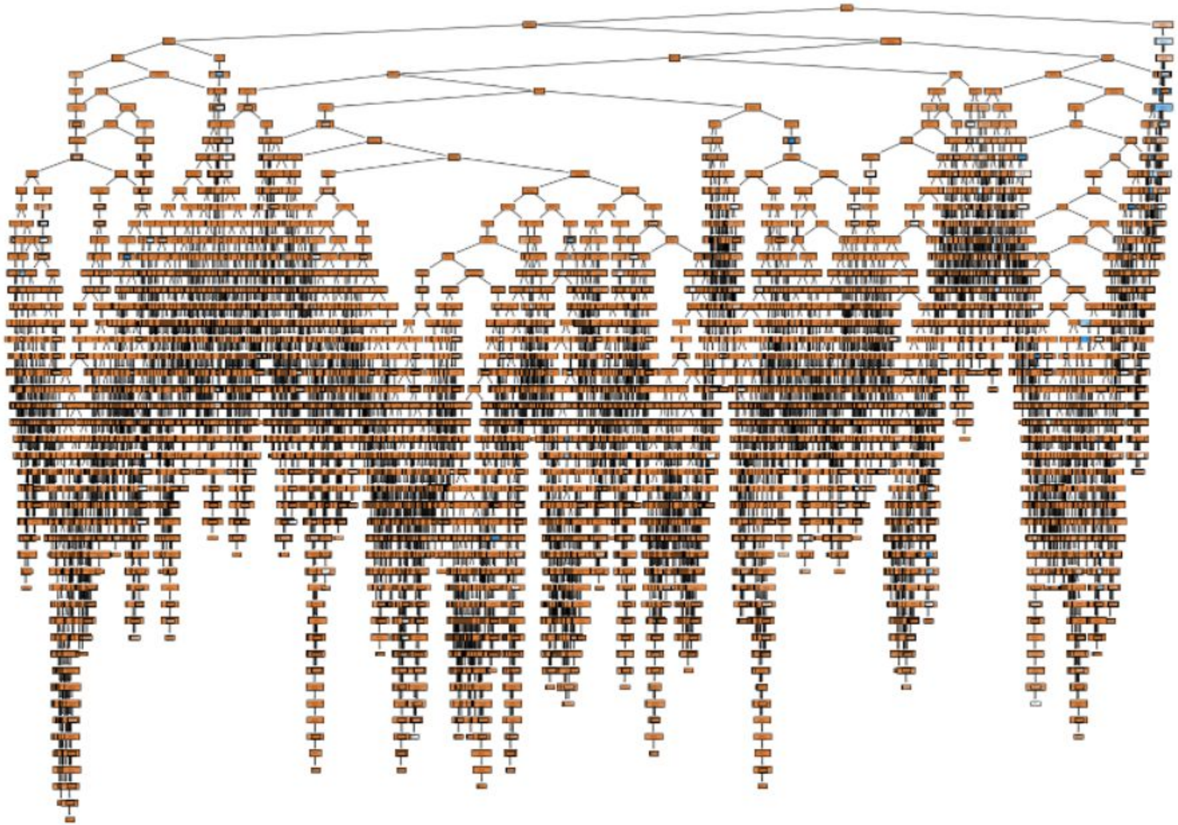
- max_depth: [10, 20, 30, 40] → 40
- min_samples_leaf: [10, 20, 30, 40, 100] → 10
- min_samples_split: [20, 40, 60] → 20
- macro f1 scoring

Adapted Hyperparameters:

- max_depth: [50, 100, 150] → 100
- min_samples_split: list(range(15, 24)) → 15
- min_samples_leaf: list(range(2, 10)) → 2

- Stopped hyperparameter tuning considering the time it took to fit the model in Randomized Search CV and little performance improvement

IMPROVED DECISION TREE AFTER HYPERPARAMETER TUNING



Stratified K-Fold CV:

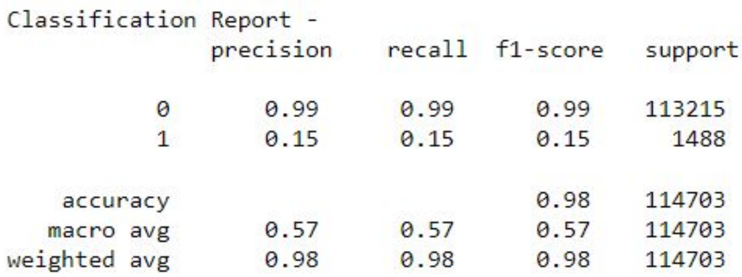
- 5 folds
- F1-macro scoring

Improved Model with new hyperparameters

- max_depth = 100
- min_samples_split = 15
- min_samples_leaf = 2

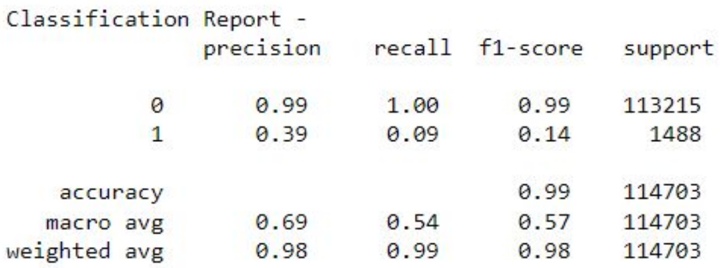
CLASSIFICATION REPORT: BEFORE AND AFTER HYPERPARAMETER TUNING

INITIAL



```
Classification Report -
              precision    recall  f1-score   support

           0       0.99      0.99      0.99    113215
           1       0.15      0.15      0.15      1488

    accuracy                           0.98    114703
   macro avg       0.57      0.57      0.57    114703
weighted avg       0.98      0.98      0.98    114703
```

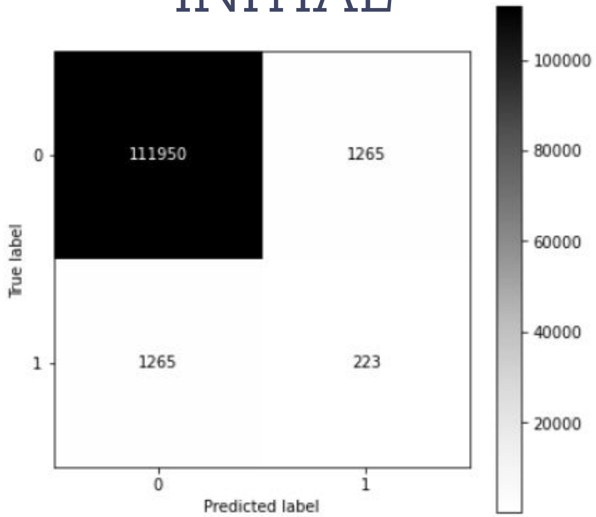
IMPROVED



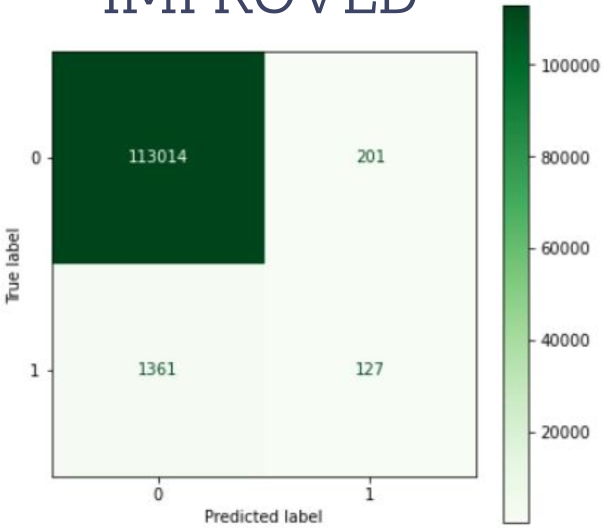
```
Classification Report -
              precision    recall  f1-score   support

           0       0.99      1.00      0.99    113215
           1       0.39      0.09      0.14      1488

    accuracy                           0.99    114703
   macro avg       0.69      0.54      0.57    114703
weighted avg       0.98      0.99      0.98    114703
```

CONFUSION MATRICES: BEFORE AND AFTER HYPERPARAMETER TUNING

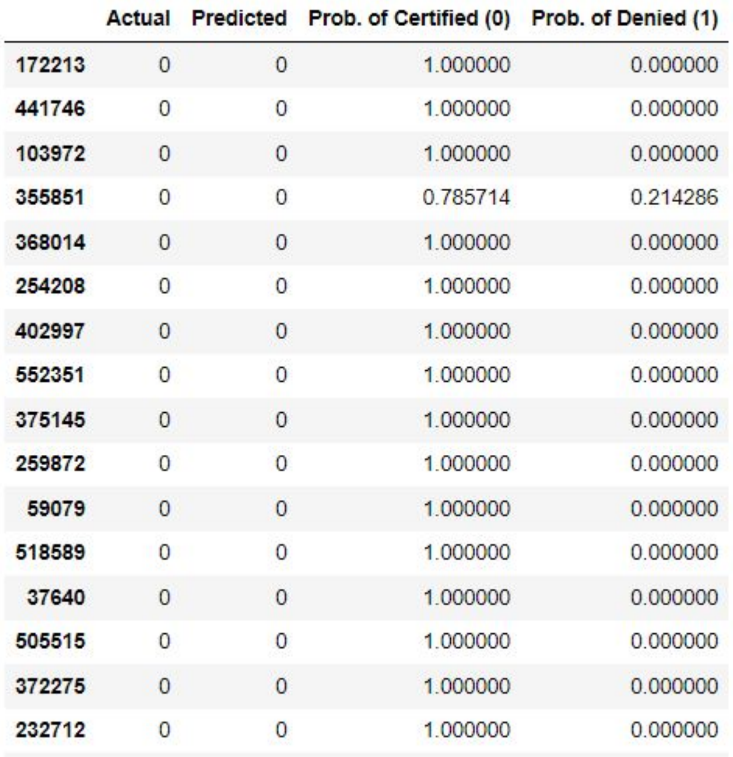
INITIAL



IMPROVED



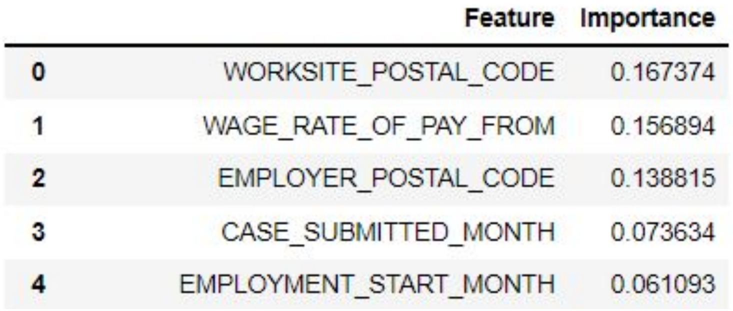
DECISION TREE: PREDICTIVE PERFORMANCE



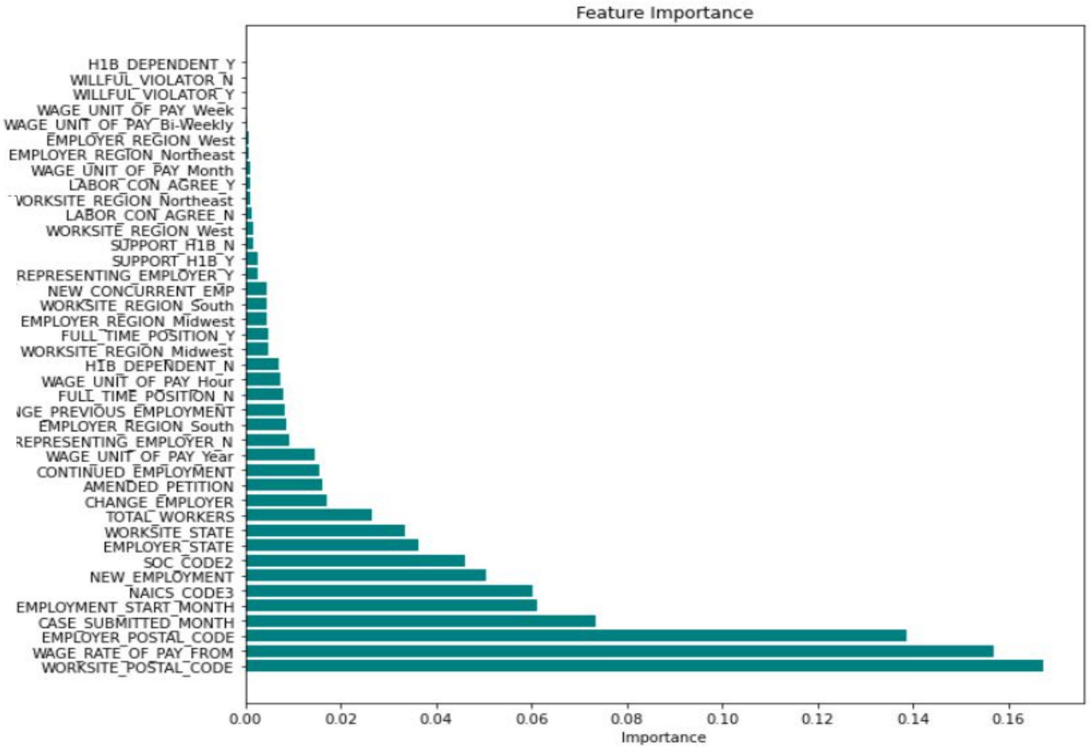
	Actual	Predicted	Prob. of Certified (0)	Prob. of Denied (1)
172213	0	0	1.000000	0.000000
441746	0	0	1.000000	0.000000
103972	0	0	1.000000	0.000000
355851	0	0	0.785714	0.214286
368014	0	0	1.000000	0.000000
254208	0	0	1.000000	0.000000
402997	0	0	1.000000	0.000000
552351	0	0	1.000000	0.000000
375145	0	0	1.000000	0.000000
259872	0	0	1.000000	0.000000
59079	0	0	1.000000	0.000000
518589	0	0	1.000000	0.000000
37640	0	0	1.000000	0.000000
505515	0	0	1.000000	0.000000
372275	0	0	1.000000	0.000000
232712	0	0	1.000000	0.000000

- Certified is 0, Denied is 1
- As seen from the confusion matrix, certified is not only more likely to be accurately predicted, but also more likely to be the result

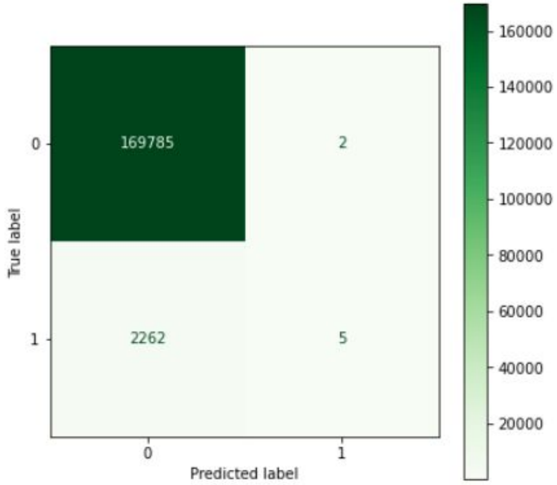
DECISION TREE: FEATURE IMPORTANCE

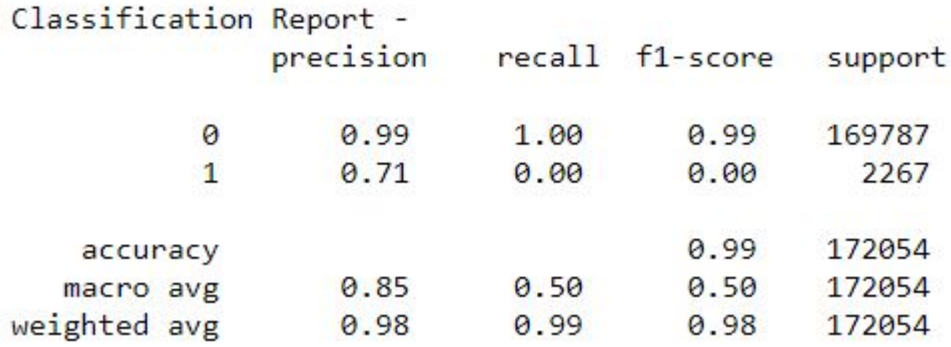


	Feature	Importance
0	WORKSITE_POSTAL_CODE	0.167374
1	WAGE_RATE_OF_PAY_FROM	0.156894
2	EMPLOYER_POSTAL_CODE	0.138815
3	CASE_SUBMITTED_MONTH	0.073634
4	EMPLOYMENT_START_MONTH	0.061093



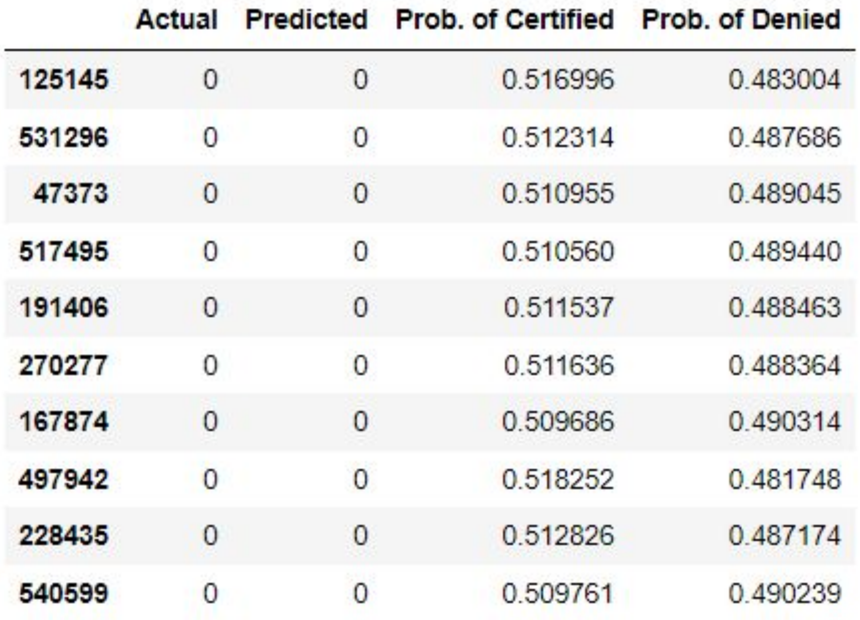
ADABOOST CLASSIFICATION REPORT AND CONFUSION MATRIX





```
Classification Report -
              precision    recall  f1-score   support

           0       0.99      1.00      0.99    169787
           1       0.71      0.00      0.00      2267

    accuracy                           0.99    172054
   macro avg       0.85      0.50      0.50    172054
weighted avg       0.98      0.99      0.98    172054
```

ADABOOST: PREDICTIVE PERFORMANCE



	Actual	Predicted	Prob. of Certified	Prob. of Denied
125145	0	0	0.516996	0.483004
531296	0	0	0.512314	0.487686
47373	0	0	0.510955	0.489045
517495	0	0	0.510560	0.489440
191406	0	0	0.511537	0.488463
270277	0	0	0.511636	0.488364
167874	0	0	0.509686	0.490314
497942	0	0	0.518252	0.481748
228435	0	0	0.512826	0.487174
540599	0	0	0.509761	0.490239

Certified is 0, Denied is 1

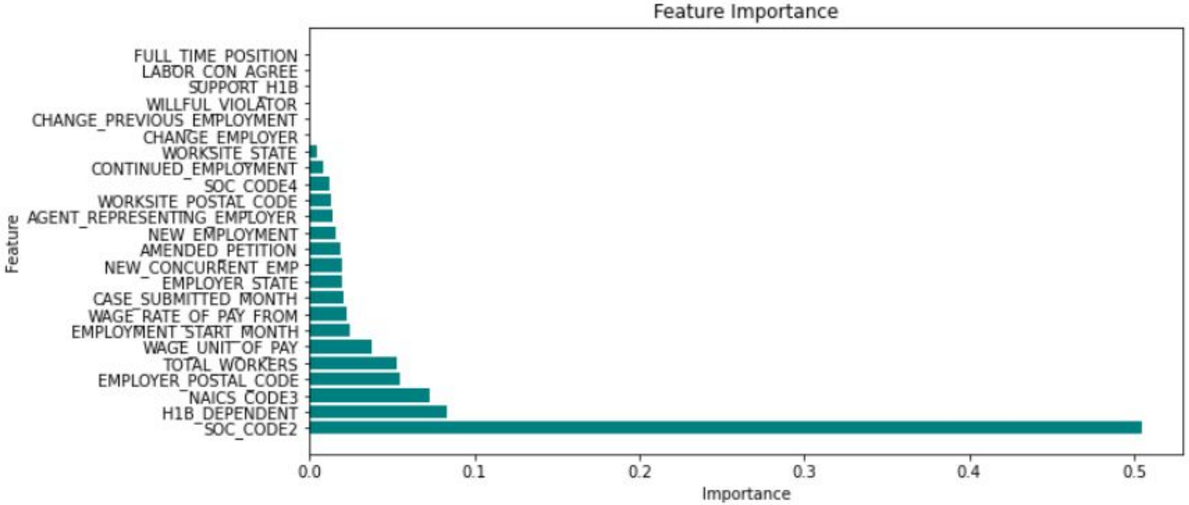
Stratified K-Fold CV:

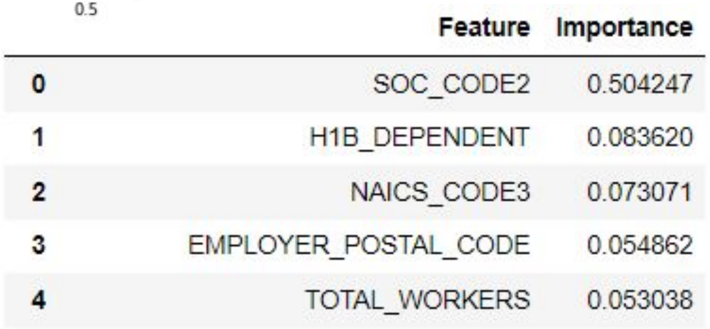
- 5 folds
- F1-macro scoring

Improved Model with new hyperparameters

- max_depth = 100
- min_samples_split = 15
- min_samples_leaf = 2

ADABOOST: FEATURE IMPORTANCE

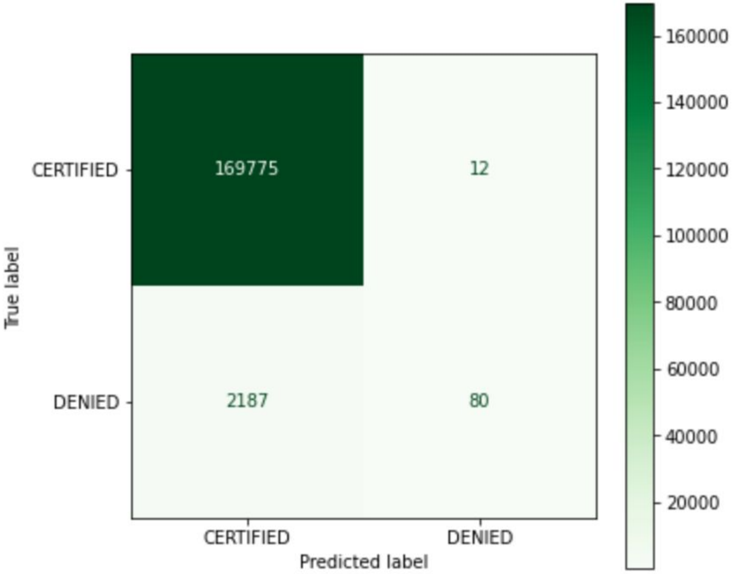


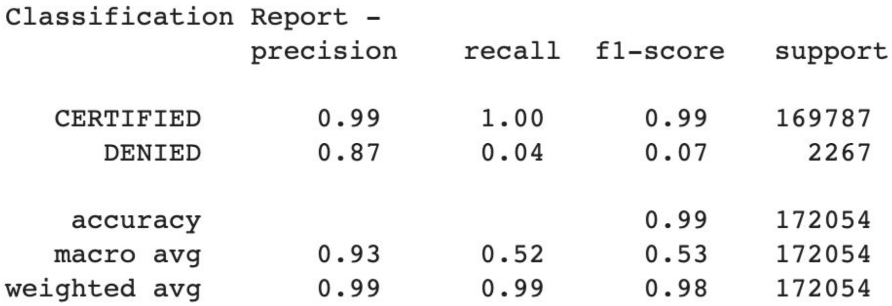


	Feature	Importance
0	SOC_CODE2	0.504247
1	H1B_DEPENDENT	0.083620
2	NAICS_CODE3	0.073071
3	EMPLOYER_POSTAL_CODE	0.054862
4	TOTAL_WORKERS	0.053038

RANDOM FOREST: CONFUSION & CLASSIFICATION

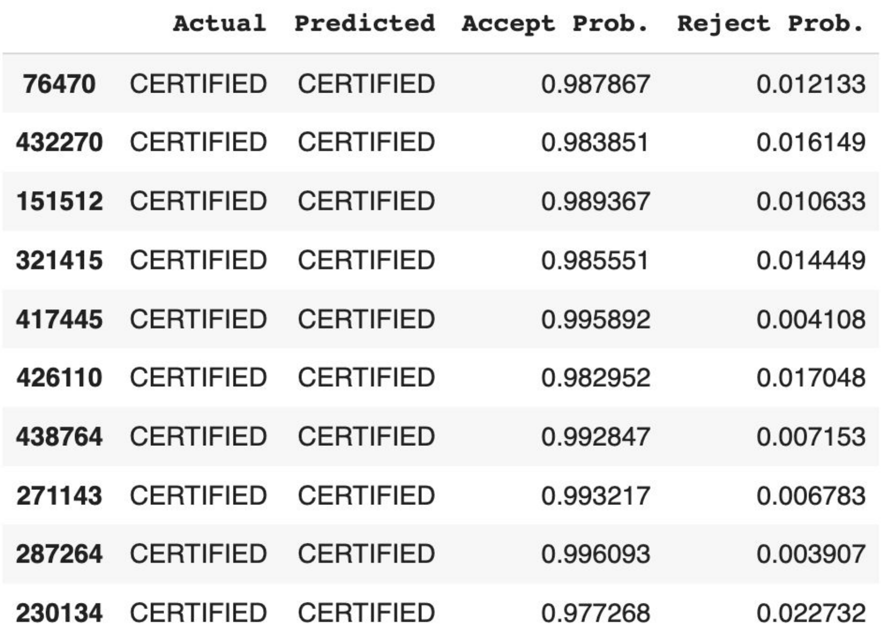
TESTING





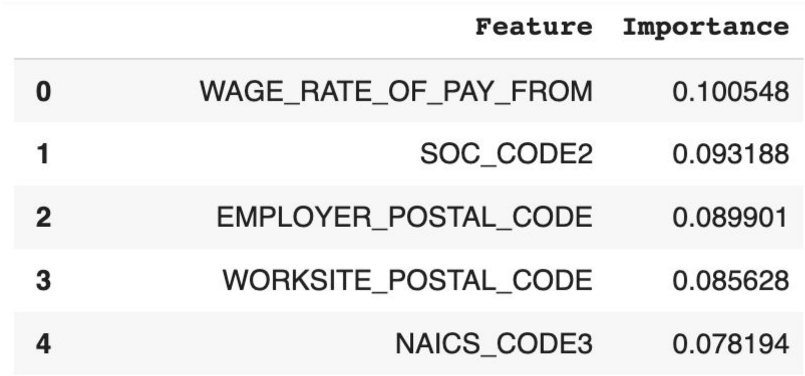
```
Classification Report -
             precision    recall  f1-score   support

   CERTIFIED       0.99      1.00      0.99    169787
      DENIED       0.87      0.04      0.07      2267

    accuracy                           0.99    172054
   macro avg       0.93      0.52      0.53    172054
weighted avg       0.99      0.99      0.98    172054
```

RANDOM FOREST: PREDICTIVE PERFORMANCE

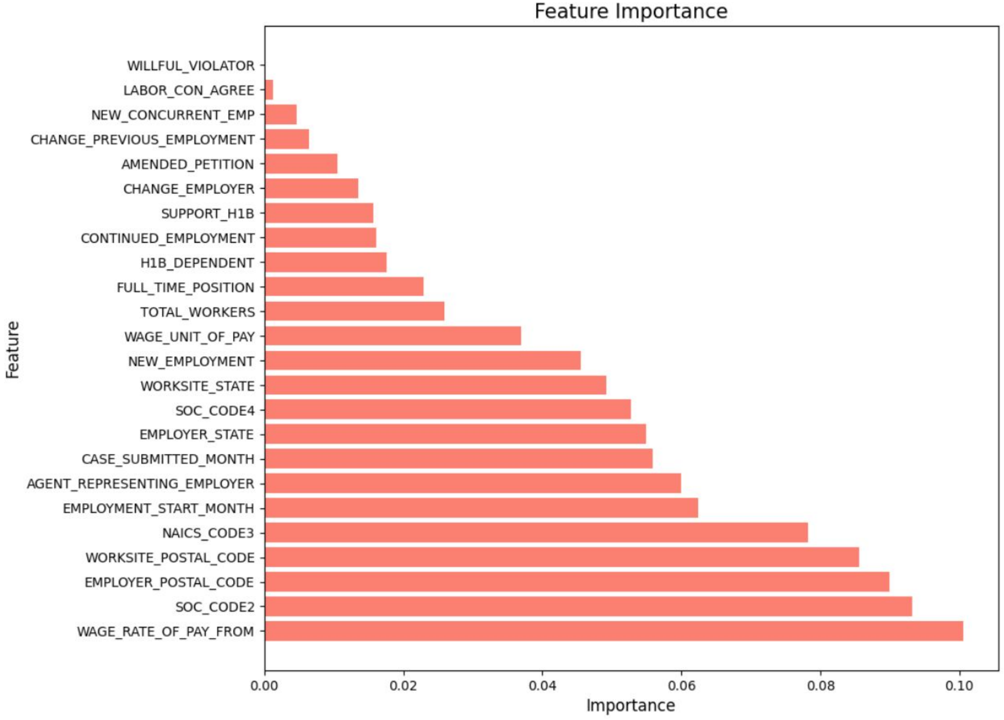


	Actual	Predicted	Accept Prob.	Reject Prob.
76470	CERTIFIED	CERTIFIED	0.987867	0.012133
432270	CERTIFIED	CERTIFIED	0.983851	0.016149
151512	CERTIFIED	CERTIFIED	0.989367	0.010633
321415	CERTIFIED	CERTIFIED	0.985551	0.014449
417445	CERTIFIED	CERTIFIED	0.995892	0.004108
426110	CERTIFIED	CERTIFIED	0.982952	0.017048
438764	CERTIFIED	CERTIFIED	0.992847	0.007153
271143	CERTIFIED	CERTIFIED	0.993217	0.006783
287264	CERTIFIED	CERTIFIED	0.996093	0.003907
230134	CERTIFIED	CERTIFIED	0.977268	0.022732

RANDOM FOREST: FEATURE IMPORTANCE



	Feature	Importance
0	WAGE_RATE_OF_PAY_FROM	0.100548
1	SOC_CODE2	0.093188
2	EMPLOYER_POSTAL_CODE	0.089901
3	WORKSITE_POSTAL_CODE	0.085628
4	NAICS_CODE3	0.078194



FINAL F1-SCORES

DECISION TREE

Macro Avg F1-score: **0.57**
Weighted avg F1-score:
0.98

ADABOOST

Macro Avg F1-score: **0.50**
Weighted avg F1-score:
0.98

RANDOM FOREST

Macro Avg F1-score: **0.53**
Weighted avg F1-score:
0.98

SUMMARY OF FEATURE IMPORTANCE BY MODEL

DECISION TREE

WORKSITE_POSTAL_CODE - **0.167**

WAGE_RATE_OF_PAY_FROM - **0.156**

EMPLOYER_POSTAL_CODE - **0.139**

CASE_SUBMITTED_MONTH - **0.0736**

EMPLOYMENT_START_MONTH - **0.0611**

RANDOM FOREST

WAGE_RATE_OF_PAY_FROM - **0.101**

SOC_CODE2 - **0.0932**

EMPLOYER_POSTAL_CODE - **0.0899**

WORKSITE_POSTAL_CODE - **0.0856**

NAICS_CODE3 - **0.0782**

ADABOOST

SOC_CODE2 - **0.504**

H1B_DEPENDENT - **0.0836**

NAICS_CODE3 - **0.0731**

EMPLOYER_POSTAL_CODE - **0.0549**

TOTAL_WORKERS - **0.0530**

WITH 'CERTIFIED' AS CASE_STATUS

Top 5 Applied Occupation Industries (SOC_CODE2)	
15	Computer & Mathematical
13	Business & Financial Operations
17	Architecture & Engineering
11	Management
29	Healthcare Practitioners and Technical

(according to the SOC Manual)

Top 5 Worksite Postal Codes (EMPLOYER_POSTAL_CODE)	
98052	Redmond, WA
94105	San Francisco, CA
94043	Mountain View, CA
19103	Philadelphia, PA
95054	Santa Clara, CA

04 CONSIDERATIONS

Ethical Considerations, Limitations, Future Extensions

FIRST DAY ISSUE

LIMITATIONS/ETHICAL CONSIDERATIONS

LIMITATIONS

- Within the scope of H-1B Visa (excluding classes of E-3 Australian, H-1B1 Chile, and H-1B1 Singapore)
- Highly imbalanced dataset
- Complex model with >500,000 rows, which was very time-consuming for analysis and our models often crashed before finishing running

ETHICAL CONSIDERATIONS

- Obtaining a visa can be crucial for nonimmigrants and thousands apply every year
- H-1B Visa is based on luck and should not have any biased results (as long as eligibility requirements are met)

05 CONCLUSION

Final Thoughts

FIRST DAY ISSUE

FUTURE EXTENSIONS

- Find a balanced dataset to increase accuracy of the model finding those "Denied" a visa
- Incorporating classes of E-3 Australian, H-1B1 Chile, and H-1B1 Singapore
- Analyze more fiscal years pre-COVID and post-COVID
- Use additional models and ensemble learning to make a more accurate model

CONCLUSION

As American citizens, we often overlook how we are born in the United States and immediately given citizenship. There are so many people who apply for H-1B visas everyday and it is up to a lottery system on whether they can stay in the country.

After analyzing the data, we could see that even the most important features did not have a large impact on whether the application was certified or denied. This is ultimately good news because it confirms that the process is randomized and not biased.

06

REFERENCES

6 PM ---

PARIS, FR

---JULY 17

REFERENCES

[1] Immigration Wait Times from Quotas have Doubled: Green Card Backlogs are Long, Growing, and Inequitable by David Bier :: SSRN

[2] H-1B Visa Lottery (How Does it Work?) | NNU Immigration

[3] Performance Data | U.S. Department of Labor

[4] H-1B Specialty Occupations

THANK YOU!

Dankie
ju faleminderit
faleminderit
شكرا
Grazias
Շնորհակալություն
Sağ ol
eskerrik asko
Дзякуй
তোমাকে ধন্যবাদ
hvala
trugéré
благодаря
Akeva
Chezu ba
gràcies
Salamat
zikomo
谢谢
hvala
děkuji
Tak
dank u

ಧನ್ಯವಾದಗಳು
សូមអរគុណអ្នក
Kamsahamnida
ຂໍຂອບໃຈທ່ານ
Lorem ipsum dolor
paldies
ačiū
ви благо[illegible]
misaotra
Terima kasih
Grazzi
Xie xie
Mauruuru
Dhanyawaadh
Welálin
баярлалаа
barka
Ahéhee'
Dhanyabaad
Thank you
miigwetch
manana
تشکر از شما
dziękuję

obrigado
ਤੁਹਾਡਾ ਧੰਨਵਾਦ
mulțumesc
спасибо
tapadh leibh
хвала
ďakujem
[illegible]
Waad ku mahadsan tahay
Gracias
Asante
Tack
Salamat
Rahmat
நன்றி
ధన్యవాదాలు
ขอบคุณ
tualumba
teşekkür ederim
Спасибі
آپ کا شکریہ
rahmat
cảm ơn bạn
Diolch yn fawr

Dankon
aitäh
takk fyri
salamat
kiitos
Merci
Grazas
დიდი მადლობა
Danke
σας ευχαριστώ
આભાર
Mèsi poutèt ou
Na gode
Mahalo
תודה
Dhanyawaad
köszönöm
þakka þér
Daalụ
terima kasih
Go raibh maith agat
grazie
ありがとう
matur nuwun

QUESTIONS?