Visa Approval Prediction using ML

Project 5 – Ensemble Techniques and Model Tuning

June 04, 2025

Submitted By:
Alex Kyeremateng Botwe

Contents / Agenda

List of Tables

List of Figures

List of Figures

Executive Summary

- In FY 2016 alone, Office of Foreign Labor Certification (OFLC) processed 775,979 employer applications covering close to 1.7 million positions for temporary and permanent labor certifications indicating a 9% increase in the overall number of processed applications from the previous year. The current manual evaluation process is increasingly unsustainable. To streamline this process:
  - A scalable, data-driven solution was required to streamline application review and highlight high-probability approvals.
  - Implementing automated checks for completeness and objective eligibility criteria at the intake stage to significantly reduce the volume of applications requiring extensive manual review was necessary.
- EasyVisa therefore developed a machine learning model that predicts the likelihood of visa certification for foreign labor applications. The goal is to support OFLC in automating and prioritizing visa cases, reducing manual workload, and improving decision accuracy, especially given the growing volume of applications yearly.

- **Business Insights:**

  a. **Job Experience & Wage Influence Approval:**

     - Applicants with documented job experience and competitive wages have a higher chance of approval.

  b. **Education Matters, But Not Linearly:**

     - Even High School education has strong predictive power suggesting that roles may not require advanced degrees but demand experience.

  c. **Regional & Continental Differences Exist:**

     - Applications from Europe and regions like the Midwest are more likely to succeed, potentially due to economic needs or program alignment.

  d. **Company Profile Plays a Role:**

     - Older and larger companies may receive slightly more favorable consideration

Selected Model: Tuned AdaBoost Classifier

Dataset	Accuracy	Recall	Precision	F1 Score
Training	0.7527	0.8853	0.7760	0.8271
Validation	0.7529	0.8831	0.7773	0.8268
Test	0.7398	0.8810	0.7651	0.8190

- The AdaBoost Classifier model exhibits consistent performance across all datasets, suggesting strong generalization and minimal overfitting

- Education level, job experience, and wage structure are the most influential predictors of visa approval.

- **Business Recommendations:**

  a. **Recommendation for OFLC:**

  - Integrate the model as a decision support system to flag high-confidence certifications
  - Focus manual effort on low-confidence or borderline applications

  b. **Recommendation for Employers:**

  - Clearly highlight job experience, education, and competitive wage offers in submissions
  - Consider offering salaries in annual formats to align with system expectations

  c. **Recommendation for Policymakers:**

  - Investigate geographic or educational feature biases
  - Explore how training and wage norms affect applicant fairness across industries

- **Conclusions:**

  - The Office of Foreign Labor Certification (OFLC) can significantly facilitate the process of visa approvals and recommend a suitable profile for the applicants for whom the visa should be certified or denied through predictive modeling.

  - The AdaBoost model delivers a robust, interpretable, and scalable solution to streamline these visa certifications. With an F1 score of approximately 0.82 on the test set, it provides reliable classification while maintaining sensitivity to both approval and rejection cases.

  - This model can serve as a cornerstone of an intelligent visa processing pipeline, improving efficiency without compromising compliance

Business Problem Overview and Solution Approach

- **Business Problem Overview:**

U.S. businesses face a growing challenge in finding and attracting qualified talent, both domestically and internationally, to remain competitive. The Office of Foreign Labor Certification (OFLC), under the Immigration and Nationality Act (INA), helps address workforce shortages by allowing foreign workers to enter the U.S. on a temporary or permanent basis.

The OFLC reviews employer applications and grants labor certifications if no qualified U.S. workers are available for a given position at the prevailing wage. However, the increasing volume of applications reaching nearly 776,000 employer applications for close to 1.7 million positions in FY 2016 is making the manual review process tedious and unsustainable.

EasyVisa, a data-driven solutions firm, has been tasked with developing a classification model to:

- Automate visa approval facilitation by developing a model to predict the likelihood of visa approval for applicants.

- Provide certification/denial recommendations based on the model's predictions and identify the key influencing factors and recommend whether a visa application should be certified or denied.

Solution Approach

To predict the likelihood of visa certification for foreign labor applications, **a machine learning model** was built using the following steps:

1. **Data Analysis**

    - We analyzed the dataset provided by OFLC which contained information about various attributes of the application and the final case status (certified/denied). Some other common features included in the data were wage unit, wage offered, region of employment, educational level, job experience and continent of employee's origin

2. **Data Cleaning**

    - Handle missing values: Checked and Impute missing data

    - Handle outliers: Identified and addressed extreme values that could skew the model

3. **Exploratory Data Analysis (EDA)**

- Analyze the distribution of individual features.
- Explore relationships between features and the target variable using stacked bar plots for categorical features vs. targe and box plots for numerical features vs. target
- Identify highly correlated features to avoid multicollinearity issues in some models using heatmap
- Obtain visual insights into the characteristics of the data before modeling, using distribution plot

4. **Data Preprocessing**

- Converted and encoded categorical variables to derive new features
- Outlier detection, Feature engineering and Data preparation for modeling were performed

5. Model Training and Selection

Given the nature of the classification problem (binary classification), and assuming class imbalance, the following models were considered:

i. **Baseline**: Random Forest

ii. **Boosting Models:** Gradient Boosting, AdaBoost and XGBoost

Handling Imbalanced Data: Visa approvals are likely to be much more frequent than denials, leading to an imbalanced dataset. To address this, we used techniques like:

- Oversample the train data,
- Undersample the train data.
- Ensemble Methods specifically designed for imbalance

Model Training: Train all the selected models on the training data

Model Selection: The validation set was used for model selection and hyperparameter tuning

6. **Model Evaluation & Validation**

   - Evaluated performance on train, test and validation sets using: Accuracy, Recall, Precision and F1
   - Plotted confusion matrix for further evaluation
   - Model Performance Improvement was done using Hyperparameter Tuning on 4 models namely Random Forest, AdaBoost Classifier, Gradient Boosting Classifier and XGBoost Classifier

7. **Important Features:** Identified the most influential features that significantly impact the case status.

EDA – Univariate Analysis

- **Education of Employee**

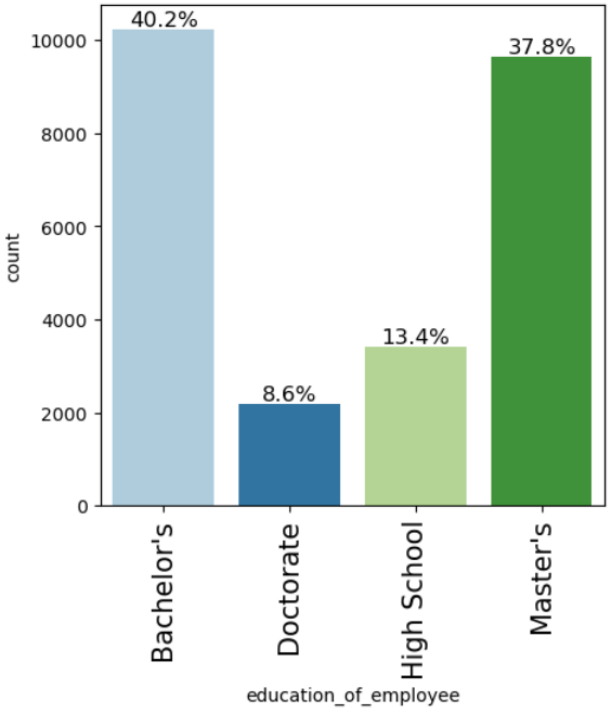


Fig 1: Univariate Analysis Education of Employee

Observations:

- The largest group of employees (40.2%) holds a Bachelor's degree
- Employees with a Master's degree constitute the second largest group at 37.8%
- High School graduates make up 13.4% of the population, while those with a Doctorate degree are the least represented group at 8.6%
- The vast majority of individuals (approximately 78%) possess either a Bachelor's or Master's degree, indicating a strong presence of highly educated individuals in the dataset.

Link to Appendix slide on data background check

- **Region of Employment**

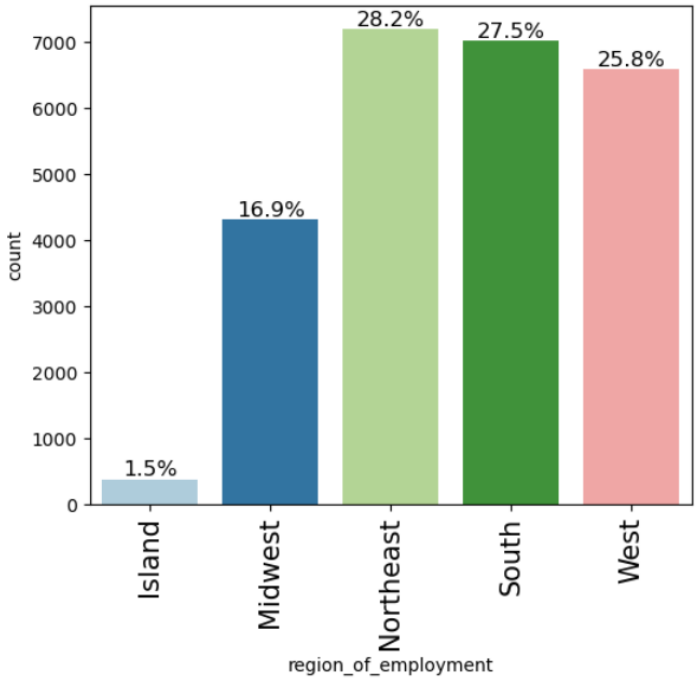


Fig 2: Univariate Analysis Region of Employment

Observations:

- The Northeast Region accounts for the largest proportion of employment opportunities at 28.2%
- The South and West Regions also have a significant share, with 27.5% and 25.8% respectively, indicating a relatively even distribution of opportunities across these three major regions
- The Midwest Region represents 16.9% of the employment opportunities
- The Island has a very small share, accounting for only 1.5% of the employment opportunities
- The majority of employment opportunities with human resource shortages are concentrated in the Northeast, South, and West regions of the United States

Link to Appendix slide on data background check

- **Job Experience**

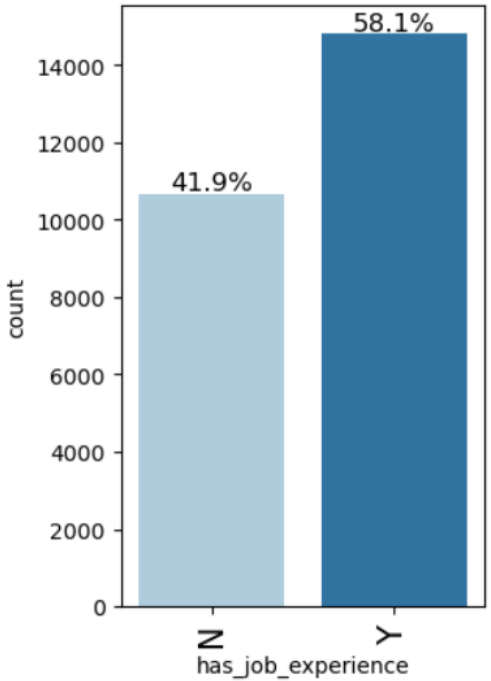


Fig 3: Univariate Analysis Job Experience

Observations:

- The larger proportion of applicants (58.1%) have prior job experience
- A substantial number of applicants (41.9%) do not have prior job experience

Link to Appendix slide on data background check

- **Number of Employees**

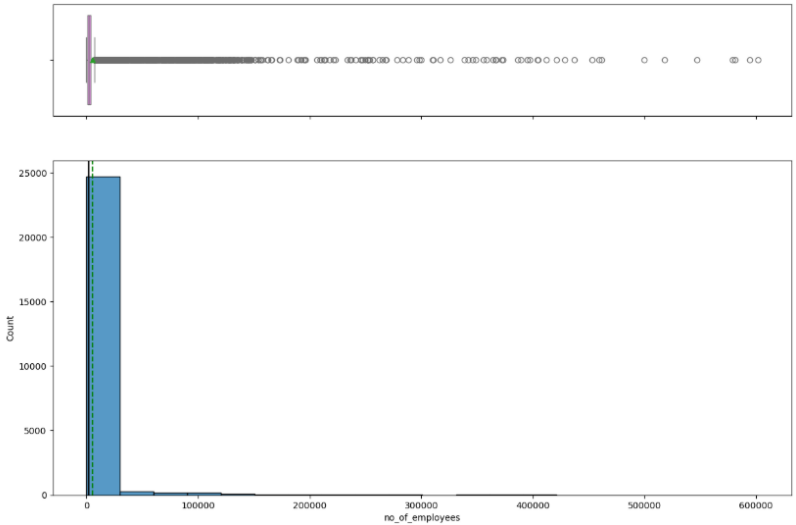


Fig 4: Univariate Analysis Number of Employees

Observations:

- The no of employees is highly skewed to the right. A vast majority of the companies or employers have a relatively small number of employees
- The histogram shows a very tall bar at the far left, indicating that most companies have a very low number of employees, likely in the tens or hundreds
- Both the box plot and the histogram clearly show the presence of many outliers. These outliers represent companies with a very high number of employees (hundreds of thousands), extending far to the right of the main distribution
- This suggests that most employers in this dataset are small to medium-sized, but there's a significant presence of very large companies which are outliers in terms of employee count.

Link to Appendix slide on data background check

- **Year of Establishment**

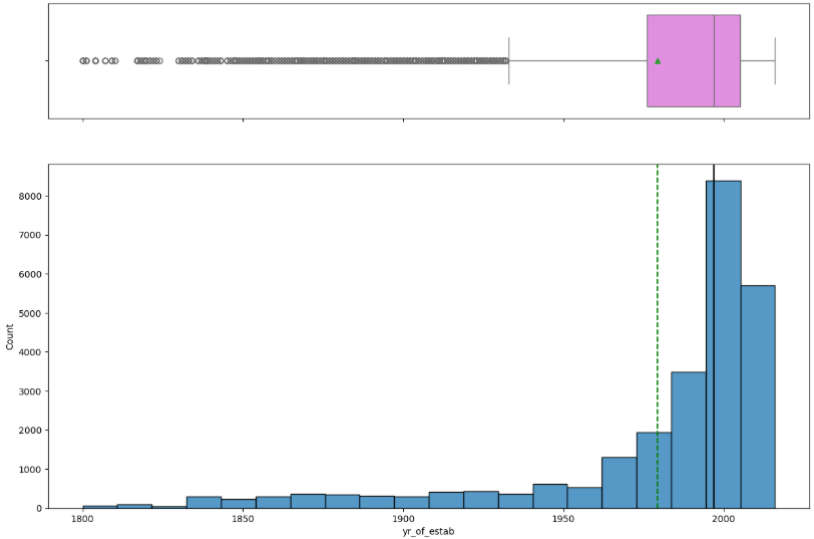


Fig 5: Univariate Analysis Year of Establishment

Observations:

- The vast majority of companies were founded more recently, particularly in the late 1990s and 2000s. There's a very large peak around the year 2000 and slightly before
- There's a clear increasing trend in the number of establishments as we move towards more recent years
- There are also a number of older establishments, with some dating back to the early 19th century (1800s) but very few
- The box plot shows a left-skewed distribution suggesting that the older establishments are considered outliers in the context of the main cluster of newer establishments
- There is therefore a clear indication that a strong trend of companies were established more recently, especially around the turn of the 21st century

Link to Appendix slide on data background check

- **Full Time Position**

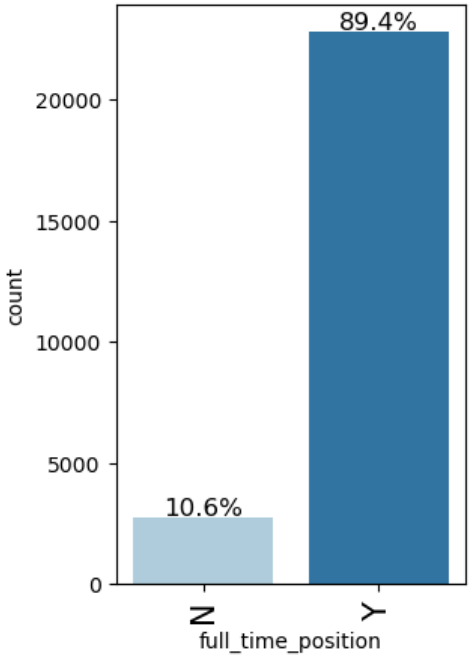


Fig 6: Univariate Analysis Full Time Position

Observations:

- A dominant proportion of the positions (89.4%) are full-time
- Only a small percentage (10.6%) of the positions are not full-time, which could include part-time, temporary, or other non-full-time arrangements.
- The data strongly suggests that the vast majority of the job opportunities in this dataset are for full-time employment

Link to Appendix slide on data background check

- **Case Status**

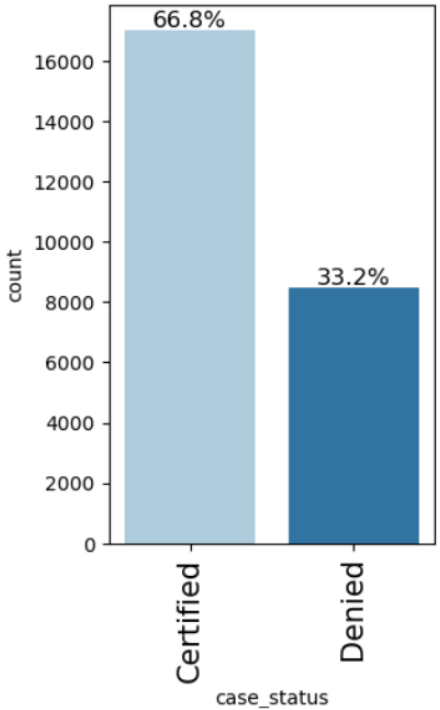


Fig 7: Univariate Analysis Case Status

Observations:

- The largest proportion of visa applications, 66.8%, are "Certified" (approved)
- A substantial portion of applications, 33.2%, are "Denied."
- The dataset is imbalanced, with significantly more certified cases than denied cases. This imbalance should be considered when building a classification model, as it can affect model performance if not addressed

Link to Appendix slide on data background check

EDA – Bivariate Analysis

- **Correlation Check**

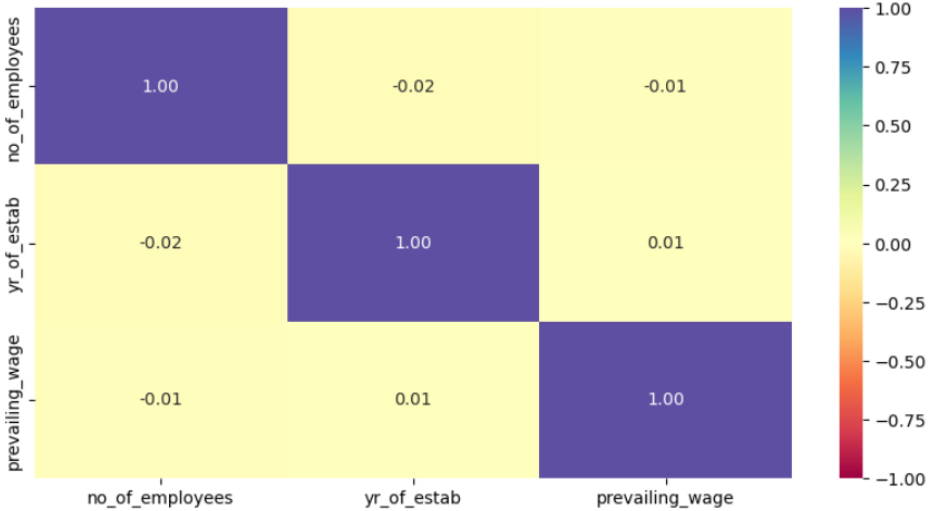


Fig 8: Correlation Plot

Observations:

- **'no_of_employees' and 'yr_of_estab':** The correlation of -0.02, suggest that almost no linear relationship. This means that the number of employees does not tend to systematically increase or decrease as the year of establishment changes

- **'no_of_employees' and 'prevailing_wage':** The correlation of -0.01, again indicates a negligible linear relationship. This implies that the number of employees in a company does not have a linear association with the prevailing wage

- **'yr_of_estab' and 'prevailing_wage':** The correlation is 0.01, showing virtually no linear connection. This means that the year a company was established does not linearly relate to the prevailing wage it offers

The three numerical features are therefore largely independent of each other in a linear sense

Link to Appendix slide on data background check

Education of Employee vs Case Study

case_status	Certified	Denied	All
education_of_employee			
All	17018	8462	25480
Bachelor's	6367	3867	10234
High School	1164	2256	3420
Master's	7575	2059	9634
Doctorate	1912	280	2192

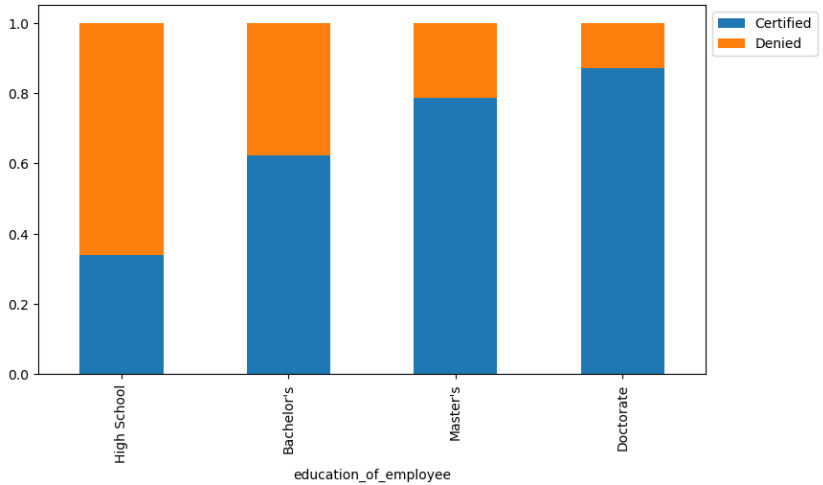


Fig 9: Education of Employee vs Case Study

Observations:

- A significantly larger proportion of High School applications are Denied compared to Certified. This confirms the table's observation that High School applicants have a lower approval rate
- For Bachelor's degree holders, the proportion of Certified cases are higher than Denied, but there's still a notable proportion of denials
- Master's degree holders show a much higher proportion of Certified cases compared to Denied, indicating a better approval rate than Bachelor's
- Doctorate holders have the highest proportion of Certified cases, with a very small Denied cases. This suggests that applicants with a Doctorate degree have the highest visa approval rate among all education levels.

Link to Appendix slide on data background check

- **Continent vs Case Study**

case_status	Certified	Denied	All
continent			
All	17018	8462	25480
Asia	11012	5849	16861
North America	2037	1255	3292
Europe	2957	775	3732
South America	493	359	852
Africa	397	154	551
Oceania	122	70	192

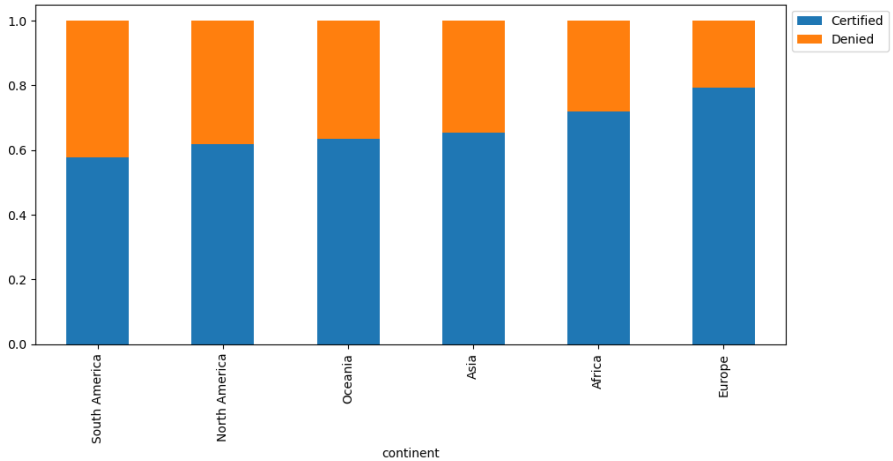


Fig 10: Continent vs Case Study

Observations:

- South America and North America have lower certification rates compared to other continents
- Asia and Oceania have relatively moderate certification rates
- Africa and Europe exhibit a higher proportion of certified visa applications. However, Europe appears to have the highest certification rate among all continents
- This suggests that the applicant's continent of origin might be a relevant factor influencing the case status

Link to Appendix slide on data background check

- **Job Experience vs Case Study**

case_status	Certified	Denied	All
has_job_experience			
All	17018	8462	25480
N	5994	4684	10678
Y	11024	3778	14802

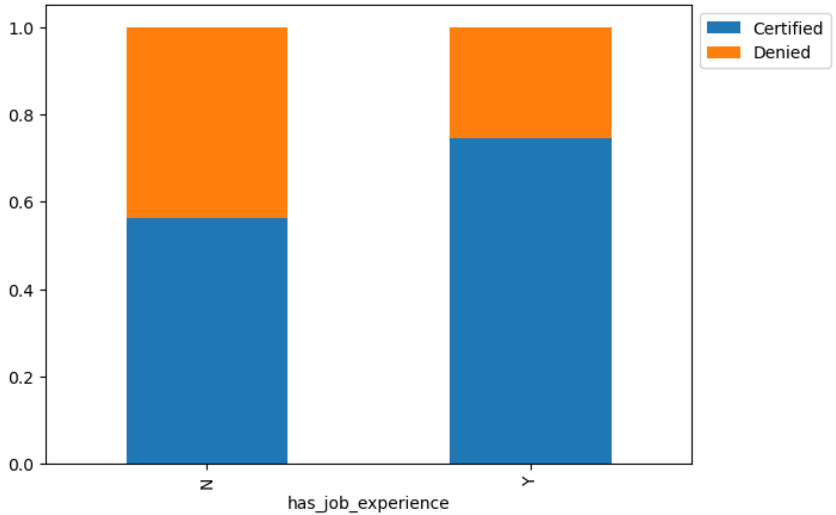


Fig 11: Job Experience vs Case Study

Observations:

- More applications were submitted by individuals with job experience (14,802) than those without job experience (10,678)
- Out of 25,480 applications, 17,018 (approximately 66.8%) were certified and 8,462 (approximately 33.2%) were denied
- Applications of the individual with job experience have a significantly higher certification rate compared to those without job experience
- This suggest that having job experience appears to be a strong positive predictor for receiving labor certification, significantly increasing the likelihood of an application being certified and reducing the chance of denial

Link to Appendix slide on data background check

- **Region of Employment vs Prevailing Wage**

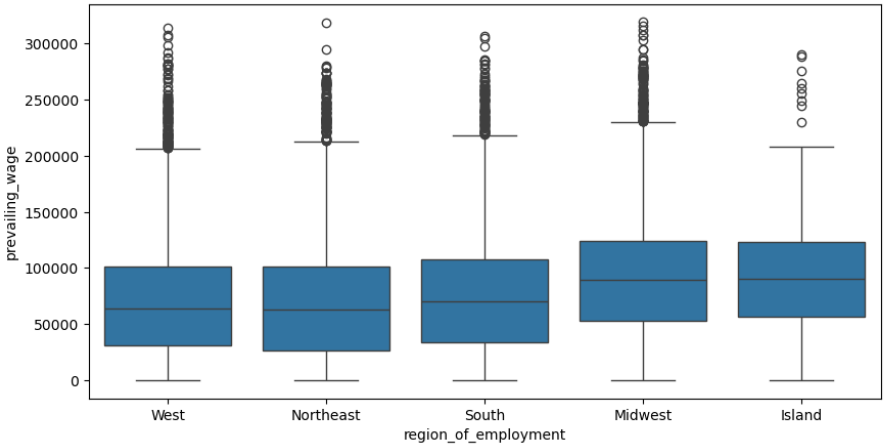


Fig 12: Region of Employment vs Prevailing Wage

Observations:

- All regions appear to be positively skewed, meaning the tail of the distribution extends further to the higher wage values
- The prevailing wages span a very wide range across all regions, from near $0 to over $300,000, with numerous outliers at the higher end
- There are noticeable differences in the median prevailing wages across regions
- The prevailing wage is not similar across all regions of the US
- While all regions have a wide range of wages and many high outliers, the central tendency and spread of prevailing wages vary geographically

Link to Appendix slide on data background check

- **Distribution Plot of Prevailing Wage vs Case Study**

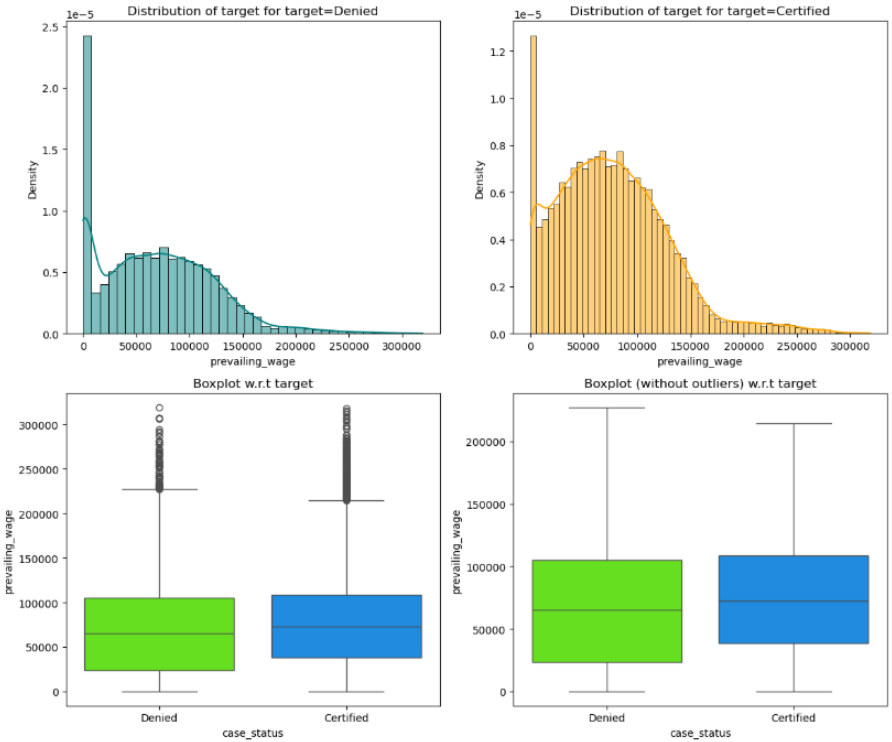


Fig 13: Prevailing Wage vs Case Study

Observations:

- The distribution of prevailing wages for Denied cases is more concentrated at lower wage ranges, showing a peak in density roughly between $0 and $50,000
- The distribution for Certified cases , while also peaking at lower wages, is generally shifted towards higher wage values and has a more substantial presence in the $50,000 to $100,000 range.
- The median prevailing wage for Certified applications is higher than for Denied applications
- Both Certified and Denied categories have numerous high-wage outliers, suggesting that even at very high prevailing wages, applications can be both denied or certified
- The prevailing wage therefore appears to be a significant factor influencing the outcome of labor certification applications

Link to Appendix slide on data background check

- **Distribution Plot of No. of_Employees vs Case Study**

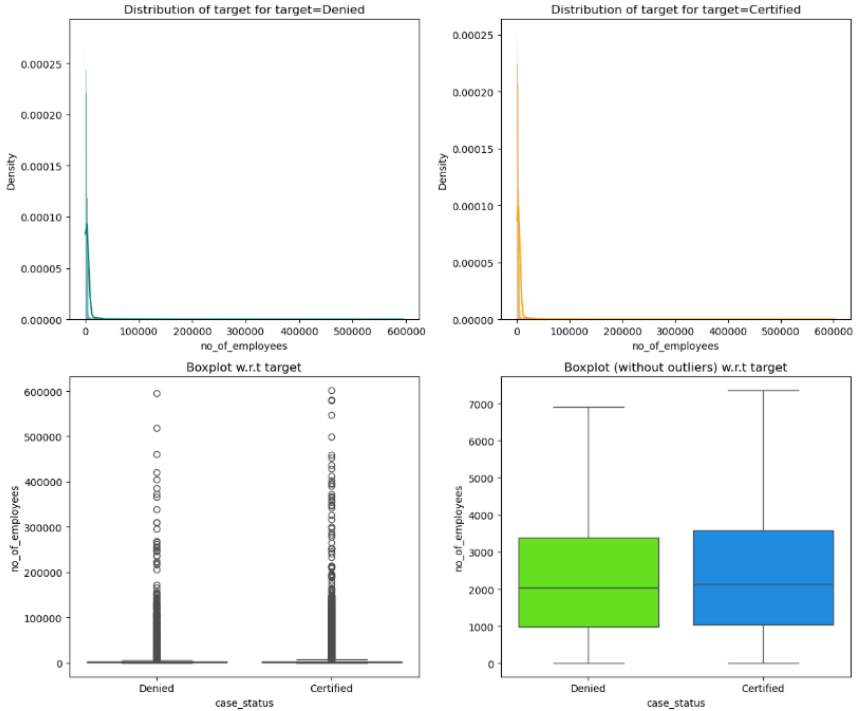


Fig 14: No. of_Employees vs Case Study

Observations:

- The distribution for denied cases is extremely skewed towards companies with a very small number of employees
- The highest density is at or very close to 0 employees, indicating that most denied applications come from very small businesses
- The distribution for certified cases is also heavily skewed towards companies with a small number of employees, with the highest density near 0
- There are a massive number of outliers for both denied and certified cases. This suggests that while most applications are from small companies, very large companies also apply and can have their applications denied or certified.
- The number of employees do not appear to be a strong distinguishing factor between denied and certified labor certification applications

Link to Appendix slide on data background check

- **Year of Establishment vs Case Study**

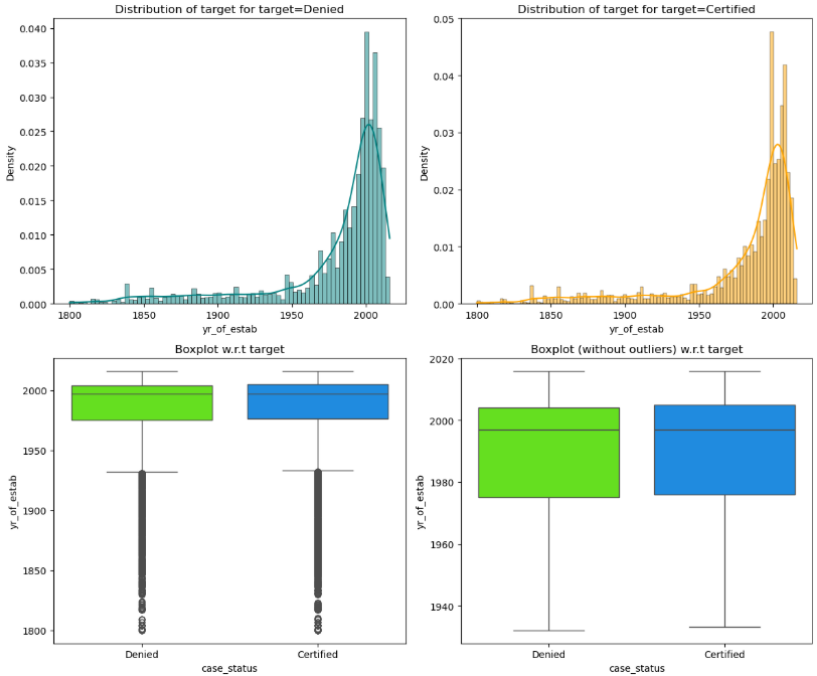


Fig 15: Year of Establishment vs Case Study

Observations:

- The distribution for denied cases is heavily skewed towards more recently established companies between 1990s and early 2000s
- The distribution for certified cases shows a very similar pattern to denied cases: a strong skew towards more recent establishments, with the highest density also in the late 1990s and early 2000s.
- Both denied and certified applications overwhelmingly come from companies established more recently, particularly in the late 1990s and early 2000s.
- Based on these plots, the year of establishment of the employer's company does not appear to be a significant distinguishing factor between denied and certified labor certification applications.

Link to Appendix slide on data background check

- **Unit of Wage vs Case Study**

case_status	Certified	Denied	All
unit_of_wage			
All	17018	8462	25480
Year	16047	6915	22962
Hour	747	1410	2157
Week	169	103	272
Month	55	34	89

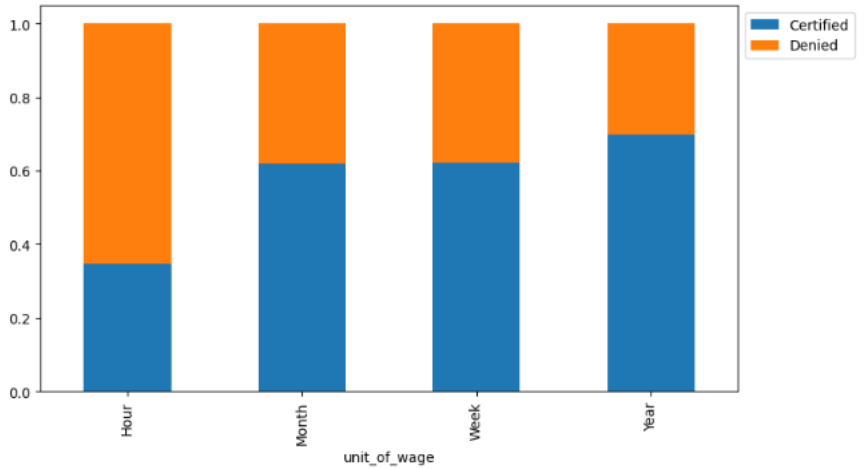


Fig 16: Unit of Wage vs Case Study

Observations:

- The vast majority of labor certification applications (22,962 out of 25,480 total applications) specify wages on a yearly basis
- The likelihood of certification varies significantly depending on the unit of wage.
- Hourly wages have the lowest certification rate, with a majority of applications being denied (approximately 65% denied)
- Monthly and Weekly wages have similar certification rates, with about 62% of applications being certified
- Yearly wages have the highest certification rate, with approximately 73% of applications being certified
- This suggests that the "unit_of_wage" is a relevant factor in the case status, with yearly wage applications having a considerably higher success rate

Link to Appendix slide on data background check

- **Full Time Position vs Case Study**

case_status	Certified	Denied	All
full_time_position			
All	17018	8462	25480
Y	15163	7610	22773
N	1855	852	2707

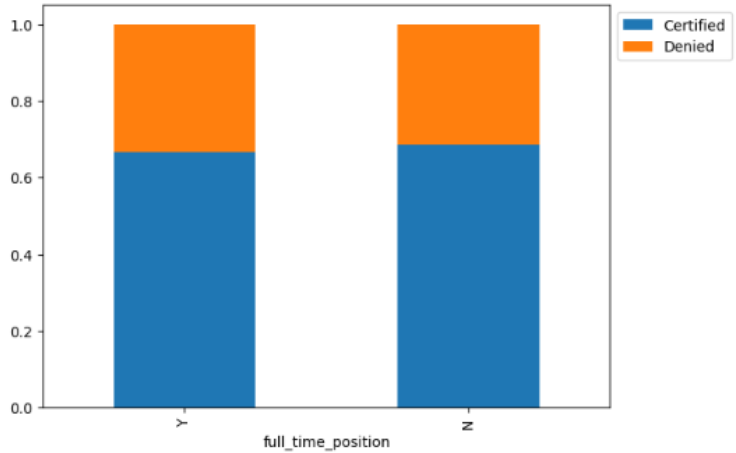


Fig 17: Full Time Position vs Case Study

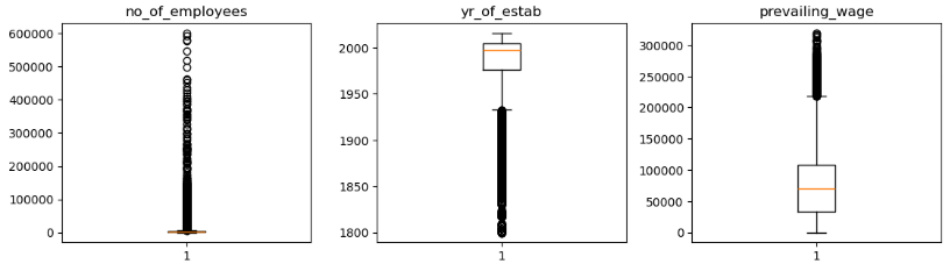
Observations:

- The vast majority of labor certification applications (22,773 out of 25,480 total applications) are for full-time positions.
- The certification rates for full-time and non-full-time positions are quite similar.
- The difference in certification rates between full-time and non-full-time positions is minimal, suggesting that whether a position is full-time or not does not significantly impact the likelihood of certification

Link to Appendix slide on data background check

Data Preprocessing

- Outlier Check



- **no_of_employees:** There is a severe positive skew and a massive number of high-value outliers which signifies that the vast majority of employers applying for labor certifications have a very small number of employees

- **yr_of_estab:** The majority of companies in the dataset were established relatively recently, primarily within the last few decades of the 20th century and the beginning of the 21st century. The numerous data points below the lower whisker represent companies established much earlier

- **prevailing_wage:** This plot clearly shows a significant number of high-value outliers in the prevailing_wage variable. This means that while most job positions fall within a broad "typical" wage range, there are a substantial number of positions that offer exceptionally high wages

Data Preparation for Modeling

We defined Dependent and independent variables:

- **Dependent:** case_status which is the primary outcome of the prediction
- **Independent Variables:** continent, education_of_employee, has_job_experience, requires_job_training, no_of_employees, yr_of_estab, region_of_employment, prevailing_wage, unit_of_wage, full_time_position

- We encoded categorical features
- Intercept was added to data and dummies were created for independent features
- We split the data into train, validation and test to be able to evaluate the model that we build on the train data.
  - First, we split data into training and temporary set
  - Then we split the temporary set into train and validation

```
Shape of Training set :  (15288, 21)
Shape of the Validation set:  (5096, 21)
Shape of test set :  (5096, 21)
```

The class ratios are almost the same across all three data set indicating that there will be no bias in the model

```
Percentage of classes in training set:
case_status
1    0.667844
0    0.332156
Name: proportion, dtype: float64
```

```
Percentage of classes in validation set:
case_status
1    0.667975
0    0.332025
Name: proportion, dtype: float64
```

```
Percentage of classes in test set:
case_status
1    0.667975
0    0.332025
Name: proportion, dtype: float64
```

Model Evaluation Criterion

i. Model predicts that the visa application will get certified but in reality, the visa application should get denied.

ii. Model predicts that the visa application will not get certified but in reality, the visa application should get certified.

iii. Both the cases are important as:

- If a visa is certified when it had to be denied a wrong employee will get the job position while US citizens will miss the opportunity to work on that position

- If a visa is denied when it had to be certified the U.S. will lose a suitable human resource that can contribute to the economy

In order to reduce losses, F1 Score can be used as the metric for evaluation of the model. The greater the F1 score the higher the chances of minimizing False Negatives and False Positives.
We use balanced class weights so that model focuses equally on both classes.

Model Performance Summary

- **Overview of final ML model and its parameters**

Objective: The Tuned AdaBoost Classifier has demonstrated strong capabilities in learning from historical visa application data to accurately classify new applications. Its consistent performance across different datasets indicates its potential as a reliable tool for initial assessment and prioritization, thereby reducing the manual burden on OFLC personnel

Model Type: AdaBoost Classifier

Key Model Parameters: The key model parameters were **base_estimator, learning_rate, n_estimators**

Key Performance Metrics: The model's effectiveness was evaluated using a comprehensive set of metrics, including: **Accuracy, Recall, Precision, and F1 Score**, across its training, validation, and test datasets.

- **Summary of most important features used by the ML model for prediction**

  - **Experience and wage level** are central to predicting visa outcomes, aligning with labor certification priorities of proving no local worker is available at a fair wage

  - **Educational qualifications** matter, but surprisingly, less formal education (High School) also plays a strong predictive role, perhaps due to demand in less specialized job sectors

  - **Geographical factors** (region of employment and continent of origin) subtly influence predictions, hinting at systemic labor trends or socio-political factors

- **Summary of Key Performance Metrics for Training, Validation and Test data on all Models**

Metric	Tuned Random Forest (Training)	Tuned Adaboost Classifier (Training)	Tuned Gradient Boost Classifier (Training)	XGBoost Classifier Tuned (Training)	Tuned Random Forest (Validation)	Tuned Adaboost Classifier (Validation)	Tuned Gradient Boost Classifier (Validation)	XGBoost Classifier Tuned (Validation)	Tuned Adaboost Classifier (Test)
Accuracy	0.776622	0.752747	0.753663	0.762493	0.749804	0.752943	0.753728	0.725667	0.739796
Recall	0.911166	0.885309	0.879432	0.97953	0.890423	0.883079	0.874853	0.946687	0.881022
Precision	0.787656	0.776013	0.779833	0.745064	0.770659	0.777347	0.782243	0.725901	0.765051
F1 Score	0.844921	0.827066	0.826643	0.846359	0.826223	0.826846	0.82596	0.821775	0.818951

- All models show some degree of overfitting, as indicated by the performance drop from training to validation data
- OFLC's concern is ensuring qualified candidates are not missed, so Recall and F1 are especially important.
- XGBoost has the highest Validation Recall (0.9466) but lower precision, whereas AdaBoost with a Validation Recall (0.8831) balances recall and precision

- The Tuned Adaboost Classifier was selected as the best model, as its performance on the test set is solid, indicating good generalization capabilities to unseen data.

- The F1 score, being a harmonic mean of precision and recall, offers a balanced view, and here, Adaboost's test F1 score of 0.8189 is quite good.

APPENDIX

Data Background and Contents

- **Data Overview**

The dataset consists of 25480 rows and 12 columns, representing data information about different attributes of employee and the employer to help in developing a classification model to facilitate the process of visa approvals and recommend a suitable profile for the applicants for whom the visa should be certified or denied

case_id	continent	education_of_employee	has_job_experience	requires_job_training	no_of_employees	yr_of_estab	region_of_employment	prevailing_wage
EZYV01	Asia	High School	N	N	14513	2007	West	592.2029
EZYV02	Asia	Master's	Y	N	2412	2002	Northeast	83425.6500
EZYV03	Asia	Bachelor's	N	Y	44444	2008	West	122996.8600
EZYV04	Asia	Bachelor's	N	N	98	1897	West	83434.0300
EZYV05	Africa	Master's	Y	N	1082	2005	South	149907.3900

Table 1: Top 5 rows of the Dataset

- **Data Background**

The dataset EasyVisa was used in the preparation of machine learning-based solution for predicting the likelihood of visa certification for foreign labor applications. The goal is to help OFLC provide certification/denial recommendations based on the model's predictions and identify the key influencing factors and recommend whether a visa application should be certified or denied and also to facilitate the process of visa approvals

- **Data Contents**

```
 0   case_id                25480 non-null  object
 1   continent              25480 non-null  object
 2   education_of_employee  25480 non-null  object
 3   has_job_experience     25480 non-null  object
 4   requires_job_training  25480 non-null  object
 5   no_of_employees        25480 non-null  int64
 6   yr_of_estab            25480 non-null  int64
 7   region_of_employment   25480 non-null  object
 8   prevailing_wage        25480 non-null  float64
 9   unit_of_wage           25480 non-null  object
 10  full_time_position     25480 non-null  object
 11  case_status            25480 non-null  object
dtypes: float64(1), int64(2), object(9)
memory usage: 2.3+ MB
```

Table 2: Information on the Data Set

There are three datatypes namely: float64(1), int64(2) and object(9) with 3 numerical and 9 categorical (strings). The target variable is the case_status, which is of object type.

Data Description

- case_id: ID of each visa application
- continent: Information of continent the employee
- education_of_employee: Information of education of the employee
- has_job_experience: Does the employee has any job experience? Y= Yes; N = No
- requires_job_training: Does the employee require any job training? Y = Yes; N = No
- no_of_employees: Number of employees in the employer's company
- yr_of_estab: Year in which the employer's company was established
- region_of_employment: Information of foreign worker's intended region of employment in the US.
- prevailing_wage: Average wage paid to similarly employed workers in a specific occupation in the area of intended employment. The purpose of the prevailing wage is to ensure that the foreign worker is not underpaid compared to other workers offering the same or similar service in the same area of employment.
- unit_of_wage: Unit of prevailing wage. Values include Hourly, Weekly, Monthly, and Yearly.
- full_time_position: Is the position of work full-time? Y = Full Time Position; N = Part Time Position
- case_status: Flag indicating if the Visa was certified or denied

Statistical Summary

	no_of_employees	yr_of_estab	prevailing_wage
count	25480.000000	25480.000000	25480.000000
mean	5667.043210	1979.409929	74455.814592
std	22877.928848	42.366929	52815.942327
min	-26.000000	1800.000000	2.136700
25%	1022.000000	1976.000000	34015.480000
50%	2109.000000	1997.000000	70308.210000
75%	3504.000000	2005.000000	107735.512500
max	602069.000000	2016.000000	319210.270000

Table 3: Statistical Summary of the Dataset

- There are no missing values in the dataset

Observations

- **Consistent Data Volume:** All three variables have the same count of 25,480 indicating that these specific columns are either completely populated or that any missing values were treated.
- **Presence of Outliers and Skewness:** no_of_employees and prevailing_wage both show a significant difference between their mean and median, with the mean being higher than the median indicating that these distributions are right-skewed
- **Potential Data Quality Issues:** The min value for no_of_employees is -26.00. This is a clear data quality issue that needs to be addressed through cleaning
- **Varied Ages of Companies**: The yr_of_estab variable shows a wide range, from 1800 to 2016. This indicates that the dataset includes companies that have been established for a very long time, as well as relatively new ones.

Model Building

- **Original Data**

Model	Training Performance	Validation Performance
Bagging	0.9892	0.7737
Random Forest	1.0	0.805
Gradient Boosting	0.8291	0.8266
AdaBoost	0.8194	0.8166
XGBoost	0.8963	0.8079

- **Gradient Boosting:** This stands out as the best performer in terms of generalization, exhibiting the least overfitting and achieving the highest F1 validation score among all models (0.8266). This suggests it is the most robust model for predicting on unseen data
- **AdaBoost:** This a close second in terms of generalization and stability, with very low overfitting
- **Random Forest and XGBoost:** These show good validation performance, but with more evidence of overfitting compared to Gradient Boosting and AdaBoost
- **Bagging:** This model shows the most significant overfitting, making its high training score less reliable for real-world predictions.

For the purpose of predicting visa approvals, the Gradient Boosting model appears to be the most suitable choice due to its excellent balance of high performance and strong generalization ability. However, further fine-tuning of the best-performing models (Gradient Boosting, AdaBoost, XGBoost) could yield even better results.

Oversampled Data

Phase	Label	Count
Before Oversampling	Certified	10,210
	Denied	5,078
After Oversampling	Certified	10,210
	Denied	10,210

Dataset	Shape
train_X (features)	(20,420, 21)
train_y (labels)	(20,420,)

- **Class Imbalance (Before Oversampling):**Initially, there was a significant class imbalance: "Certified" had 10,210 samples, while "Denied" had only 5,078. Such imbalance could cause a classifier to be biased toward predicting the majority class ("Certified").
- **Class Balance (After Oversampling):** Oversampling was used to increase the number of "Denied" samples to match the number of "Certified" samples. The new training set has equal class distribution, with 10,210 samples for each label, reducing bias and improving model generalization for the minority class.
- **Dataset Size and Shape:** The training feature set (train_X) now contains 20,420 samples with 21 features. The label vector (train_y) also has 20,420 entries, aligning correctly with train_X.

Oversampling effectively corrected the class imbalance by synthetically increasing the minority class. This allows the model to learn better representations of both classes, particularly the previously underrepresented "Denied" label

Performance Results

Model	Training Performance	Validation Performance
Bagging	0.9875	0.7665
Random Forest	1.0000	0.7965
Gradient Boosting	0.8072	0.8173
AdaBoost	0.7981	0.8118
XGBoost	0.8709	0.8129

- **Gradient Boosting:** This remains a top contender. Despite a tiny drop in validation score, its training-validation gap is still minimal, indicating exceptional generalization
- **AdaBoost:** This also maintains very strong and stable performance, making it a reliable option
- **XGBoost:** Shows a positive impact from oversampling, improving its validation performance and reducing overfitting, making it a very strong candidate
- **Random Forest:** This still perfectly fits the training data, indicating its powerful capacity and continued overfitting
- **Bagging:** This model continues to show severe overfitting, achieving near-perfect training performance but a significantly lower validation score

In terms of overall balanced performance and generalization, Gradient Boosting and XGBoost (now with improved generalization post-oversampling) appear to be the most promising models.

Undersampled Data

Phase	Label	Count
Before Undersampling	Certified	10,210
	Denied	5,078
After Undersampling	Certified	5,078
	Denied	5,078

Dataset	Shape
train_X (features)	(10,156, 21)
train_y (labels)	(10,156,)

- **Class Imbalance (Before Undersampling):**The dataset originally had a class imbalance, with "Certified" (10,210) being roughly twice as common as "Denied" (5,078). This could lead to biased models favoring the majority class.
- **Class Balance (After Undersampling):** To correct the imbalance, undersampling was performed on the majority class ("Certified") to match the number of "Denied" samples.
- Both classes now have equal representation (5,078), promoting balanced learning and reducing model bias.
- **Dataset Shape After Undersampling:** The training dataset (train_X) has 10,156 rows (sum of 5,078 + 5,078) and 21 features. The labels (train_y) correctly match the number of training instances.

Undersampling successfully addressed the class imbalance, resulting in a balanced training set that is more suitable for training unbiased classification models. However, care should be taken as undersampling may discard potentially useful data from the majority class.

Performance Results

Model	Training Performance	Validation Performance
Bagging	0.9804	0.7057
Random Forest	1.0000	0.7417
Gradient Boosting	0.7281	0.7766
AdaBoost	0.7051	0.7619
XGBoost	0.8720	0.7459

- **Gradient Boosting:** The training performance has dropped considerably, and its validation performance has also decreased significantly compared to both original and oversampled approaches
- **AdaBoost:** This also shows a marked decrease in both training and validation performance after undersampling
- **XGBoost:** Training performance is consistent with the oversampled scenario, but its validation performance has dropped notably, making it the lowest performer on the validation set among the boosting models in this undersampled scenario
- **Random Forest:** This still perfectly fits the training data but suffers a substantial drop in validation performance
- **Bagging:** This model continues to show severe overfitting. Crucially, its validation performance has deteriorated significantly after undersampling

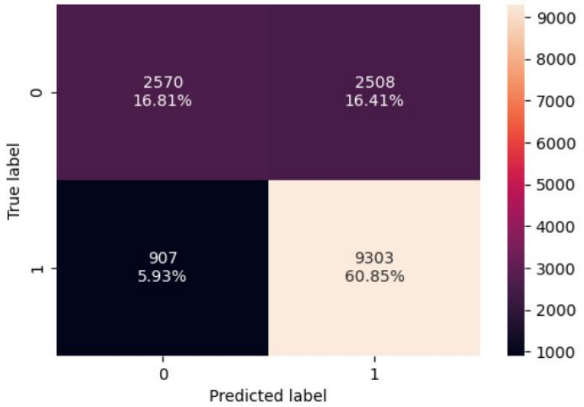
The models trained on the oversampled data (especially Gradient Boosting and XGBoost from that set) would be the recommended choice for deployment..

Model Performance Evaluation and Improvement

- **Hyperparameter Tuning - Random Forest**

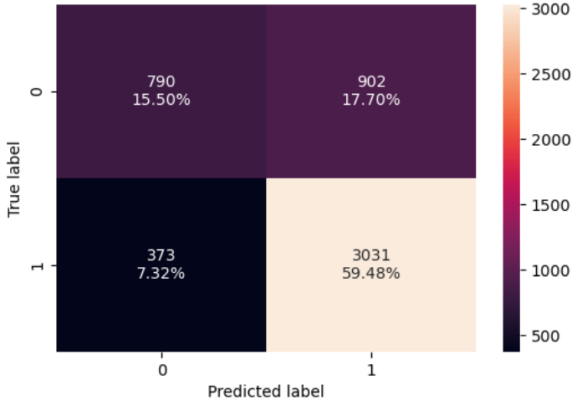
```
RandomForestClassifier
RandomForestClassifier(max_depth=10, min_samples_split=5, n_estimators=30,
                       oob_score=True, random_state=1)
```

Model Performance on the Training Set



	Accuracy	Recall	Precision	F1
0	0.776622	0.911166	0.787656	0.844921

Model Performance on the Validation Set



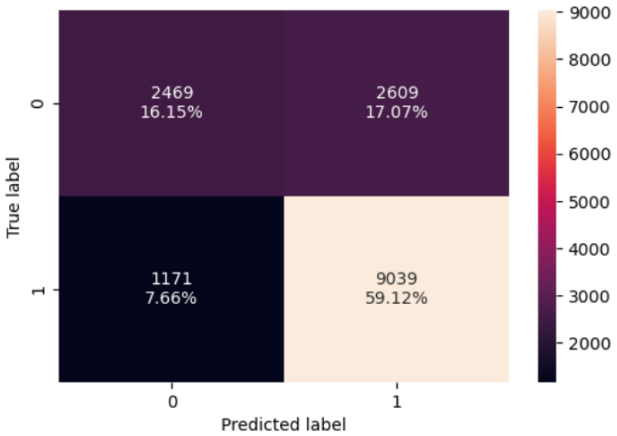
	Accuracy	Recall	Precision	F1
0	0.749804	0.890423	0.770659	0.826223

The model demonstrates strong performance in identifying the majority class 'Certified' cases, achieving high Recall and good Precision for this class. However, it shows a clear weakness in correctly identifying the minority class 'Denied' cases, as indicated by the relatively high number of False Positives

Hyperparameter Tuning - AdaBoost Classifier

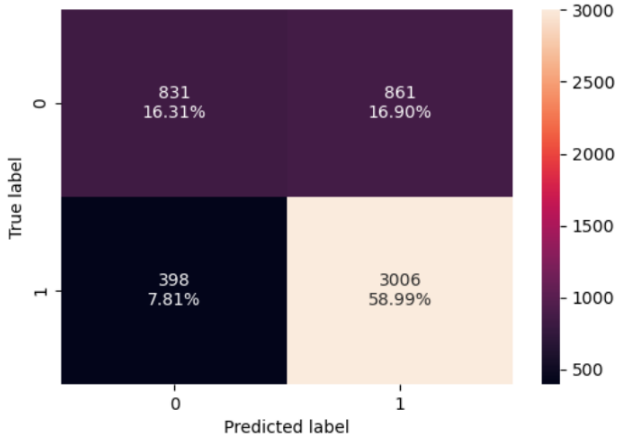
```
AdaBoostClassifier
estimator: DecisionTreeClassifier
DecisionTreeClassifier(max_depth=3, random_state=1)
DecisionTreeClassifier
DecisionTreeClassifier(max_depth=3, random_state=1)
```

Model Performance on the Training Set



	Accuracy	Recall	Precision	F1
0	0.752747	0.885309	0.776013	0.827066

Model Performance on the Validation Set



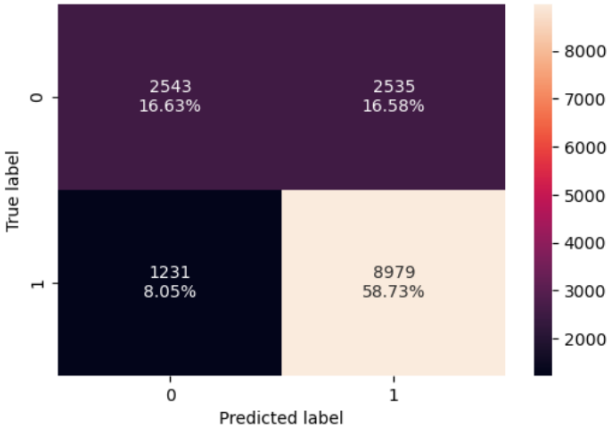
	Accuracy	Recall	Precision	F1
0	0.752943	0.883079	0.777347	0.826846

The model is generally effective at classifying the majority class. The consistent performance across different evaluation set sizes suggests a stable model. However, the high number of False Positives indicates that improving the model's ability to correctly classify the minority class ('Denied' cases) should be a priority for further optimization.

Hyperparameter Tuning - Gradient Boosting Classifier

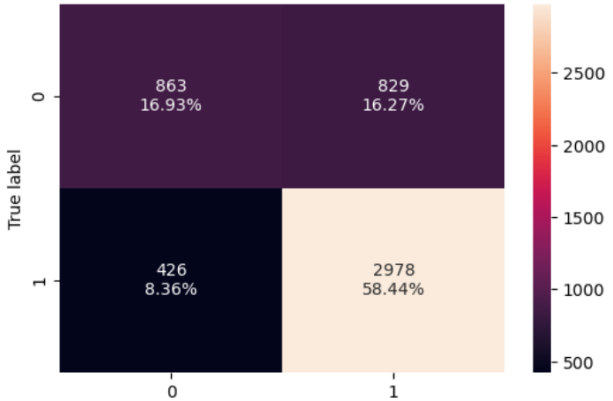
```
GradientBoostingClassifier(init=AdaBoostClassifier(random_state=1),
                           learning_rate=0.05, max_features=0.8, random_state=1,
                           subsample=0.7)
```

Model Performance on the Training Set



	Accuracy	Recall	Precision	F1
0	0.753663	0.879432	0.779833	0.826643

Model Performance on the Validation Set



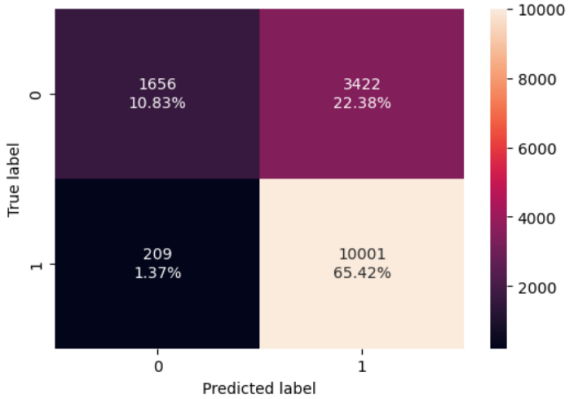
	Accuracy	Recall	Precision	F1
0	0.753728	0.874853	0.782243	0.82596

This model demonstrates consistent and strong performance, particularly for the majority class. The primary challenge remains the accurate identification of the minority class, as indicated by the persistent presence of a substantial number of False Positives

Hyperparameter Tuning - XGBoost Classifier

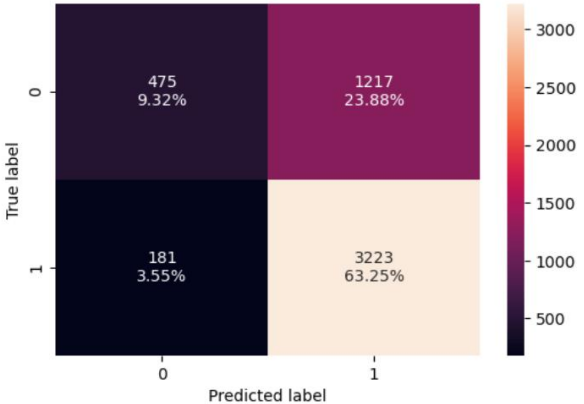
```
XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=None, device=None, early_stopping_rounds=None,
              enable_categorical=False, eval_metric='logloss',
              feature_types=None, feature_weights=None, gamma=1,
              grow_policy=None, importance_type=None,
              interaction_constraints=None, learning_rate=0.1, max_bin=None,
              max_cat_threshold=None, max_cat_to_onehot=None,
              max_delta_step=None, max_depth=None, max_leaves=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              multi_strategy=None, n_estimators=100, n_jobs=None,
```

Model Performance on the Training Set



	Accuracy	Recall	Precision	F1
0	0.762493	0.97953	0.745064	0.846359

Model Performance on the Validation Set



	Accuracy	Recall	Precision	F1
0	0.725667	0.946827	0.725901	0.821775

This model demonstrates a strong bias towards the majority class ('Certified'). It is excellent at finding most of the positive cases (high Recall) but comes at the cost of poorly identifying the minority class ('Denied'), resulting in a high number of False Positives.

Model Comparison and Final Model Selection

Training Performance Comparison:

	Tuned Random Forest	Tuned Adaboost Classifier	Tuned Gradient Boost Classifier	XGBoost Classifier Tuned
Accuracy	0.776622	0.752747	0.753663	0.762493
Recall	0.911166	0.885309	0.879432	0.979530
Precision	0.787656	0.776013	0.779833	0.745064
F1	0.844921	0.827066	0.826643	0.846359

Validation Performance Comparison:

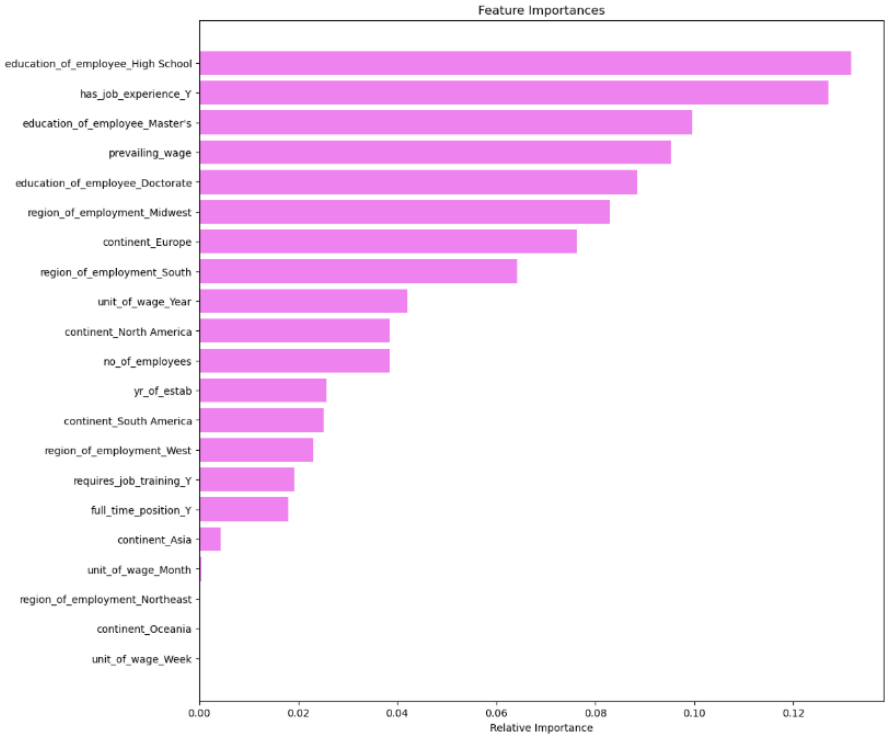
	Tuned Random Forest	Tuned Adaboost Classifier	Tuned Gradient Boost Classifier	XGBoost Classifier Tuned
Accuracy	0.749804	0.752943	0.753728	0.725667
Recall	0.890423	0.883079	0.874853	0.946827
Precision	0.770659	0.777347	0.782243	0.725901
F1	0.826223	0.826846	0.825960	0.821775

Test Performance Metrics on the Best Model:

	Tuned Adaboost Classifier
Accuracy	0.739796
Recall	0.881022
Precision	0.765051
F1	0.818951

- The Tuned AdaBoost Classifier has an outstanding generalization ability and stability and shows Minimal Overfitting, Consistent Performance and Good Balance of Metrics
- This combination of strong generalized performance and high stability makes the Tuned AdaBoost Classifier the most reliable choice for predicting visa case status for OFLC

- **Important features of the Final Model**



Top Most Important Features:

1. **education_of_employee_High School**: This is the most important feature suggesting that this education level is a strong predictor of the outcome of a visa approval or denial
2. **has_job_experience_Y:** The presence of job experience is the second most important feature. This is intuitively logical, as prior work experience would naturally influence a visa application's outcome
3. **education_of_employee_Master's:** Having a Master's degree is the third most important feature. This, along with high school education, highlights the significant influence of educational attainment levels on the model's predictions.

Final Model Summary

Dataset	Accuracy	Recall	Precision	F1 Score
Training	0.7527	0.8853	0.7760	0.8271
Validation	0.7529	0.8831	0.7773	0.8268
Test	0.7398	0.8810	0.7651	0.8190

Overall Accuracy on Test Set is 73.98%:
- This indicates that approximately 74% of all visa applications were correctly classified by the model on unseen data

Recall on Test Set is 88.10%:
- This means the model correctly identifies 88.10% of all actually Certified visa applications

Precision on Test Set is 76.51%:
- This means there is a roughly 23.5% chance that an application predicted as 'Certified' might actually be 'Denied, which is considered as a trade-off

F1 Score on Test Set is 81.90%:
- This signifies a robust and effective balance between correctly identifying positive cases and minimizing false alarms

Best Model

The Tuned AdaBoost Classifier achieved the highest F1 score of 0.8190 on the test set and therefore classified as the best performing model for visa case status classification. Its high recall ensures that a significant majority of actual positive cases are identified, while the strong F1 score confirms a good overall balance between precision and recall. This model can handle new applications and offer a valuable data-driven solution to facilitate the visa approval process

References

Great Learning. (n.d.) *Supervised Learning - Classification*. **Great Learning.** https://olympus.mygreatlearning.com/courses/124980/modules/items/7078869?pb_id=18483

Great Learning

Happy Learning !