

Machine Learning - 2 : Easy Visa

Jesmi George

Contents

1 Business Problem	5
1.1 Context	5
1.2 Objective	5
1.3 Data Overview	5
1.4 Statistical Summary of DataSet	5
1.5 Data Dictionary	6
2 Exploratory Data Analysis	7
2.1 Data Cleaning Steps (Before EDA)	7
2.2 Univariate Analysis	7
2.3 Bivariate Analysis	13
2.4 Insights from EDA	19
3 Data-Preprocessing	20
4 Model Building	21
4.1 Model Evaluation Criterion	21
4.2 Model Building on Original Training Data	21
4.3 Model Building on SMOTE Oversampled Data	23
4.4 Model Building on Random Undersampled Data	24
4.5 Model Performance Improvement using Hyperparameter Tuning	26
5 Model Performance Comparison and Final Model Selection	28
5.1 Overview	28
5.2 Training Performance Comparison	28
5.3 Validation Performance Comparison	28
5.4 Key Observations	29
5.5 Final Model Recommendation	30
6 Actionable Insights and Recommendations	30

List of Figures

1	Distribution of Applicant Continent	8
2	Distribution of Education Levels	8
3	Job Experience Distribution	9
4	Training Requirement Distribution	9
5	Distribution of Number of Employees	10
6	Distribution of Prevailing Wage	10
7	Distribution of Company Age	11
8	Distribution of Employment Regions	11
9	Unit of Wage Distribution	12
10	Full-Time vs Part-Time Positions	12
11	Visa Case Status Distribution	13
12	Correlation Heatmap	13
13	Pairplot with Case Status	14
14	Case Status vs Job Experience	14
15	Case Status vs Job Training Requirement	15
16	Case Status vs Education Level	15
17	Case Status vs Region	16
18	Case Status vs Unit of Wage	16
19	Case Status vs Full-Time Position	17
20	Prevailing Wage Distribution by Case Status	17
21	Company Age Distribution by Case Status	18
22	Employee Count Distribution by Case Status	19
23	Performance comparison on Original Training Set	22
24	Performance comparison on Testing Set	22
25	Performance Comparison on SMOTE Oversampled Training Set	23
26	Performance Comparison on Testing Set after SMOTE Training	23
27	Performance Comparison on Undersampled Training Set	25
28	Performance Comparison on Testing Set after RUS Training	25
29	Performance of Tuned Gradient Model on Test Set	26
30	Performance of Tuned AdaBoost Model on Test Set	27
31	Confusion Matrices of Tuned XGBoost Model on Test Set	27
32	Training Performance Comparison across Models	28
33	Validation Set Performance Comparison across Models	29
34	Visualization of the Feature Importance of Tuned AdaBoost Model	29
35	Performance of Tuned AdaBoost Model	30

List of Tables

1	Data Description	6
---	----------------------------	---

1 BUSINESS PROBLEM

1.1 CONTEXT

Businesses in the U.S. face high demand for skilled talent, often seeking workers both locally and abroad. The Immigration and Nationality Act (INA) allows foreign workers to work in the U.S. while protecting local workers' wages and conditions. The Office of Foreign Labor Certification (OFLC) oversees this process, approving applications only when employers show a shortage of qualified U.S. workers at fair wages.

1.2 OBJECTIVE

In FY 2016, the OFLC processed nearly 776K applications for 1.7M positions—a 9% increase from the previous year, making manual review increasingly difficult. EasyVisa aims to use Machine Learning to streamline visa shortlisting by predicting which applications are likely to be certified. Analyze the data and build a classification model that aids visa approval decisions by identifying key factors influencing case outcomes.

Key Questions to be answered :

1. What factors most influence whether a visa application is certified or denied?
2. Facilitate the process of visa approvals.

1.3 DATA OVERVIEW

- The dataset has 25480 rows and 12 columns.
- No duplicate records were found
- There are no null values in any columns
- The dataset contains 1 float column, 2 integer columns, and 9 categorical (object) columns.

1.4 STATISTICAL SUMMARY OF DATASET

- Most visa applicants are from Asia and hold a Bachelor's degree, typically with job experience and no training required.
- Company employee count has invalid values (e.g., -26), indicating data errors.
- The Northeast region has the highest applicant concentration.
- Company establishment years range from 1800 to 2016, with even newly formed companies having applicants.

- Prevailing wages range from \$2.13 to \$319,210, with unrealistic low values indicating incorrect records.
- Yearly wage is the most common wage unit reported.
- Most applicants apply for full-time roles, and the majority of these are certified.

1.5 DATA DICTIONARY

The data contains the different attributes of the employee and the employer. The detailed data dictionary is given below.

Field Name	Description
case_id	ID of each visa application
continent	Continent information of the employee
education_of_employee	Education details of the employee
has_job_experience	Does the employee have 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	Intended region of employment for the foreign worker in the US
prevailing_wage	Average wage paid to similarly employed workers in the intended area of employment
unit_of_wage	Unit of prevailing wage (Hourly, Weekly, Monthly, Yearly)
full_time_position	Is the job position full-time? (Y = Full-Time; N = Part-Time)
case_status	Indicates whether the visa was certified or denied

TABLE 1: DATA DESCRIPTION

2 EXPLORATORY DATA ANALYSIS

2.1 DATA CLEANING STEPS (BEFORE EDA)

The following data cleaning operations were performed to ensure the dataset was reliable and suitable for exploratory analysis:

- **Handling Invalid Employee Counts:** Negative values in no_of_employees were considered invalid and treated as missing. A total of 33 records contained missing values for this feature, which are addressed during later preprocessing.
- **Removing Non-Predictive Identifier:** The column case_id was dropped since it is a unique identifier for each application and does not contribute any predictive value to the model.
- **Dropping Raw Establishment Year:** The variable yr_of_estab was removed because the raw year of establishment is not inherently meaningful and behaves like a random numeric value without context.
- **Creating Company Age Feature:** A new variable, age_of_company, was created using the formula:

$$\text{Company Age} = 2016 - \text{yr_of_estab}$$

Assuming FY 2016 as the reference year (as stated in the objective), this feature better reflects company stability and potential influence on visa approval outcomes.

2.2 UNIVARIATE ANALYSIS

Continent of Applicant:

- About 66% of applicants are from Asia.
- Europe (14.6%) and North America (12.9%) follow, while other continents contribute less than 4%.

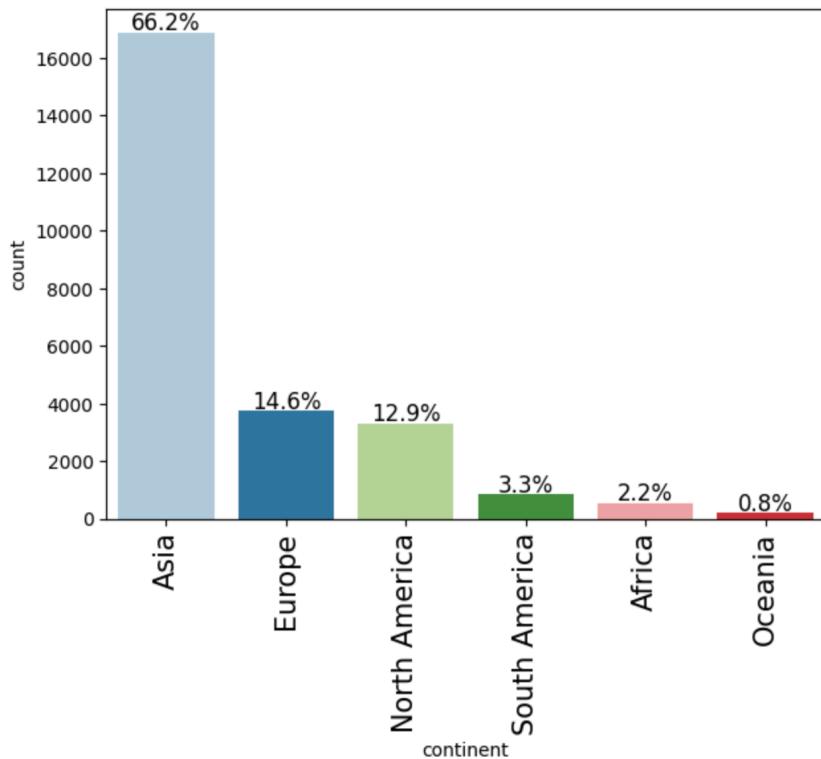


FIGURE 1: DISTRIBUTION OF APPLICANT CONTINENT

Education of Employee:

- Bachelor's and Master's degree holders form over 78% of all applicants.
- Doctorate applicants account for 8.6%, while High School graduates represent 13.4%.

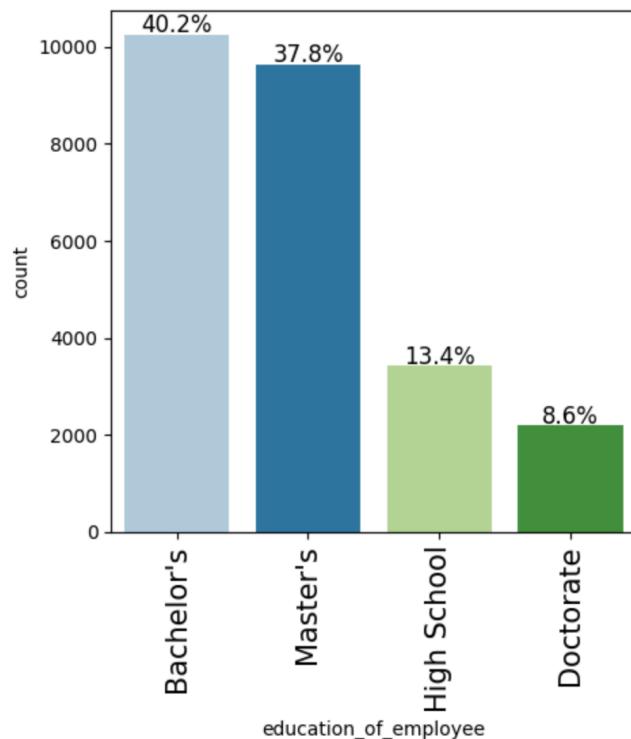


FIGURE 2: DISTRIBUTION OF EDUCATION LEVELS

Job Experience:

- Around 58% of applicants report having prior job experience.

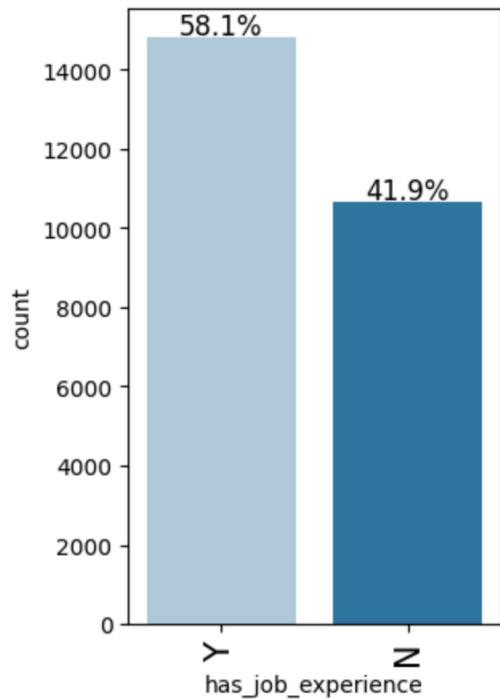


FIGURE 3: JOB EXPERIENCE DISTRIBUTION

Job Training Requirement:

- Nearly 88.4% of applicants do not require job training.
- This aligns with the high education levels of most applicants.

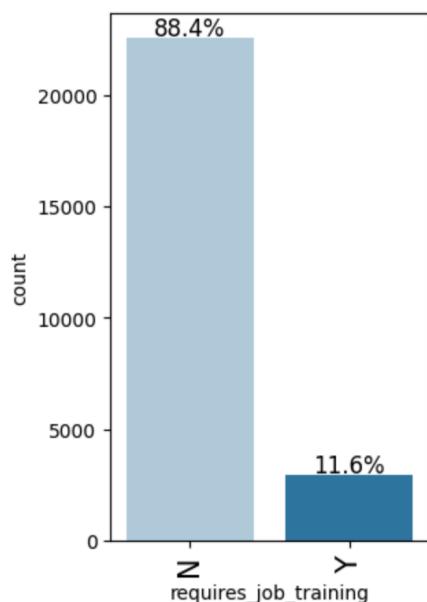


FIGURE 4: TRAINING REQUIREMENT DISTRIBUTION

Number of Employees in Company:

- Highly right-skewed distribution with large outliers.
- Most companies have under 1,000 employees, but some exceed 600,000.

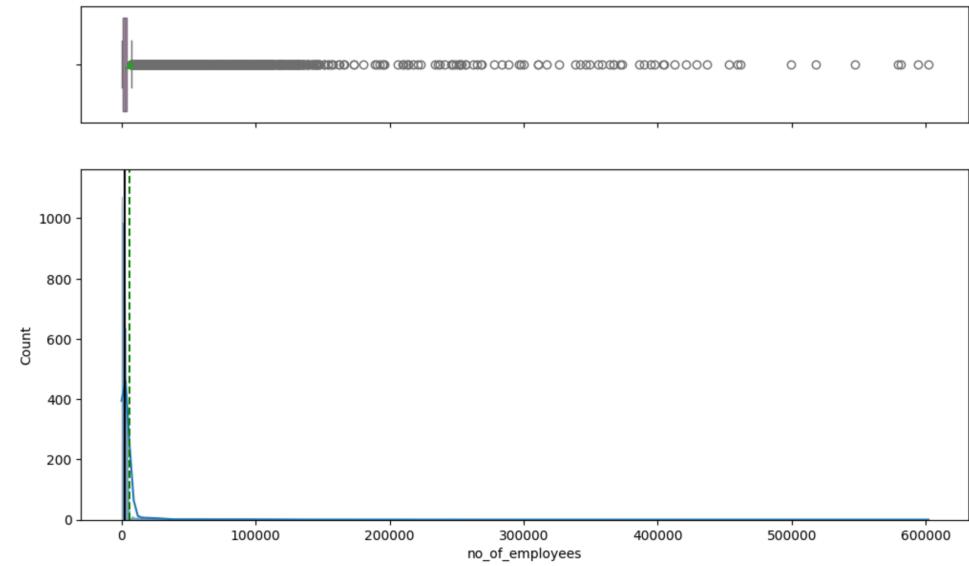


FIGURE 5: DISTRIBUTION OF NUMBER OF EMPLOYEES

Prevailing Wage:

- Wage distribution is right-skewed; most earn under \$150,000.
- Significant high-wage outliers exist.
- Very low wage values indicate erroneous entries.

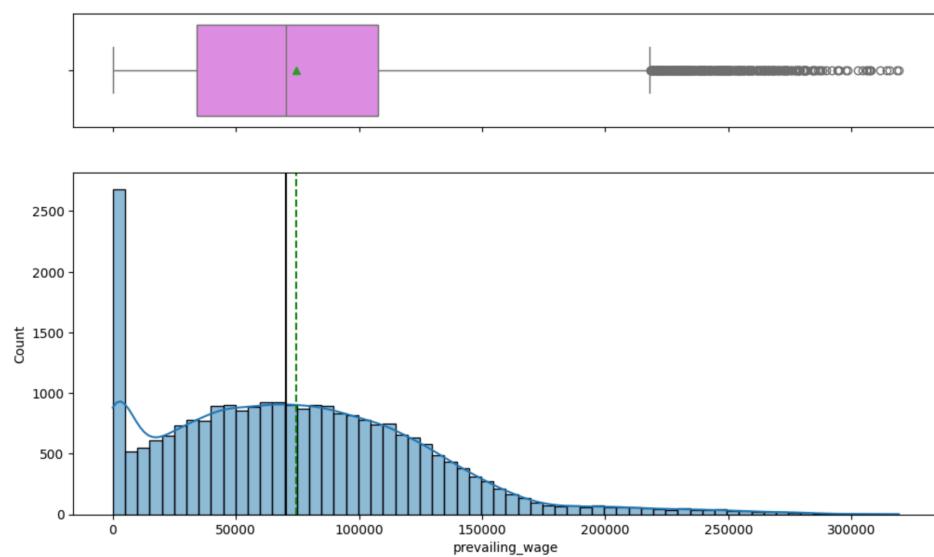


FIGURE 6: DISTRIBUTION OF PREVAILING WAGE

Age of Company:

- Right-skewed distribution with most companies under 50 years old.
- Few companies exceed 200 years in age.

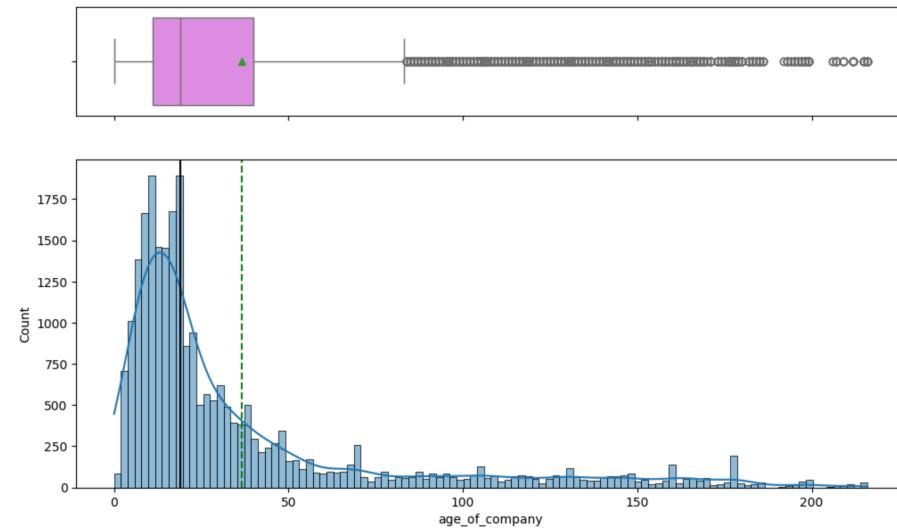


FIGURE 7: DISTRIBUTION OF COMPANY AGE

Region of Employment:

- Northeast and South regions account for about 28% each.
- West contributes 26% and Midwest 17%.
- Island region has less than 2%.

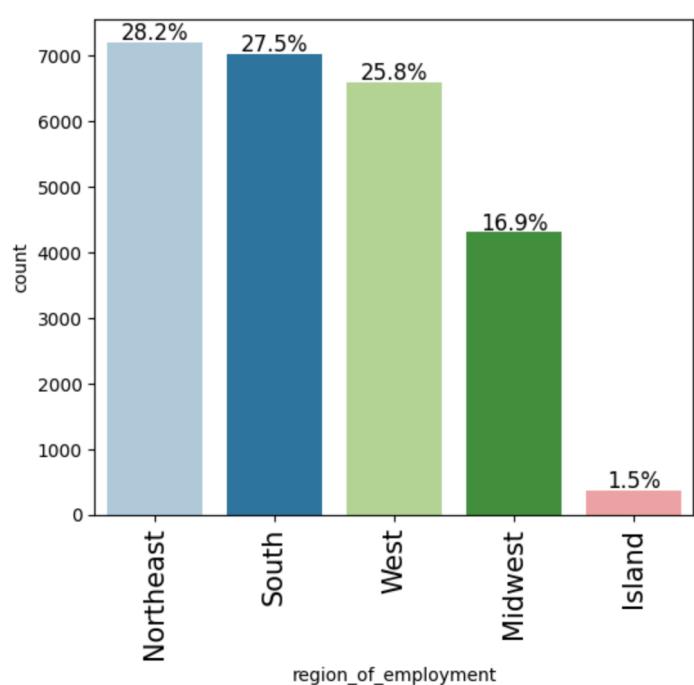


FIGURE 8: DISTRIBUTION OF EMPLOYMENT REGIONS

Unit of Wage:

- Nearly 90% wages are reported annually.
- Hourly wages form 8.5%, with weekly and monthly below 2%.

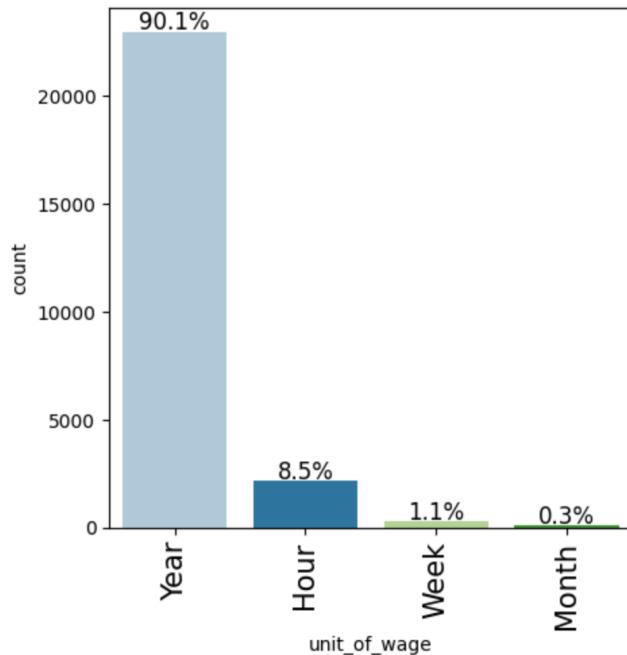


FIGURE 9: UNIT OF WAGE DISTRIBUTION

Full-Time Position:

- About 89.4% of applicants apply for full-time positions.

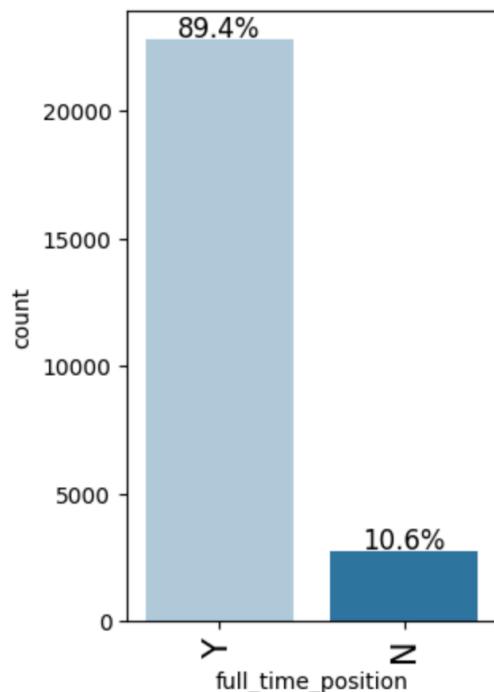


FIGURE 10: FULL-TIME VS PART-TIME POSITIONS

Case Status (Target Variable):

- Around 67% of applications are certified.
- Indicates a notable class imbalance.

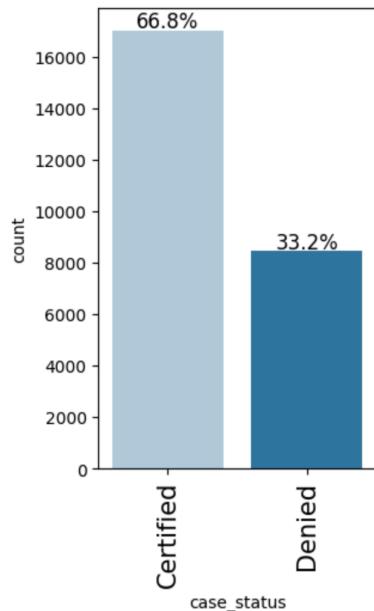


FIGURE 11: VISA CASE STATUS DISTRIBUTION

2.3 BIVARIATE ANALYSIS

Correlation Heatmap:

- No strong correlation between company age and number of employees.

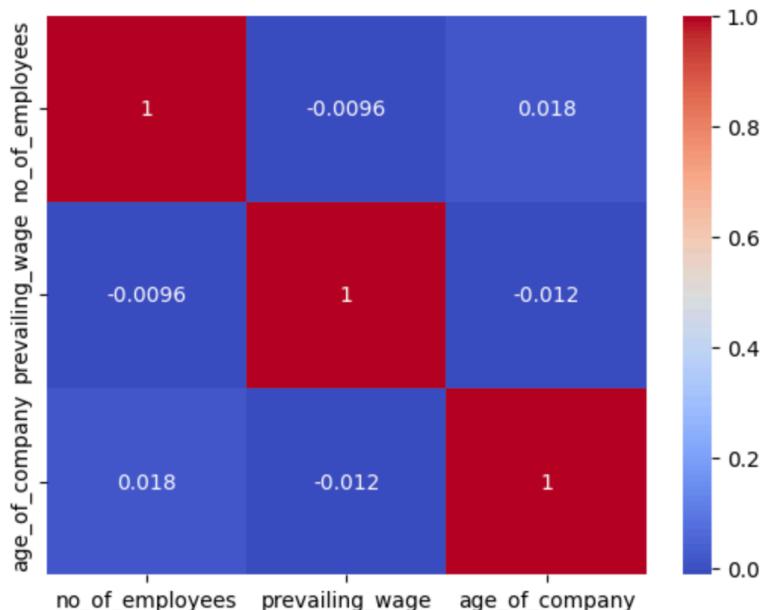


FIGURE 12: CORRELATION HEATMAP

Pairplot Analysis:

- No clear relationship between company size or age and visa approval.

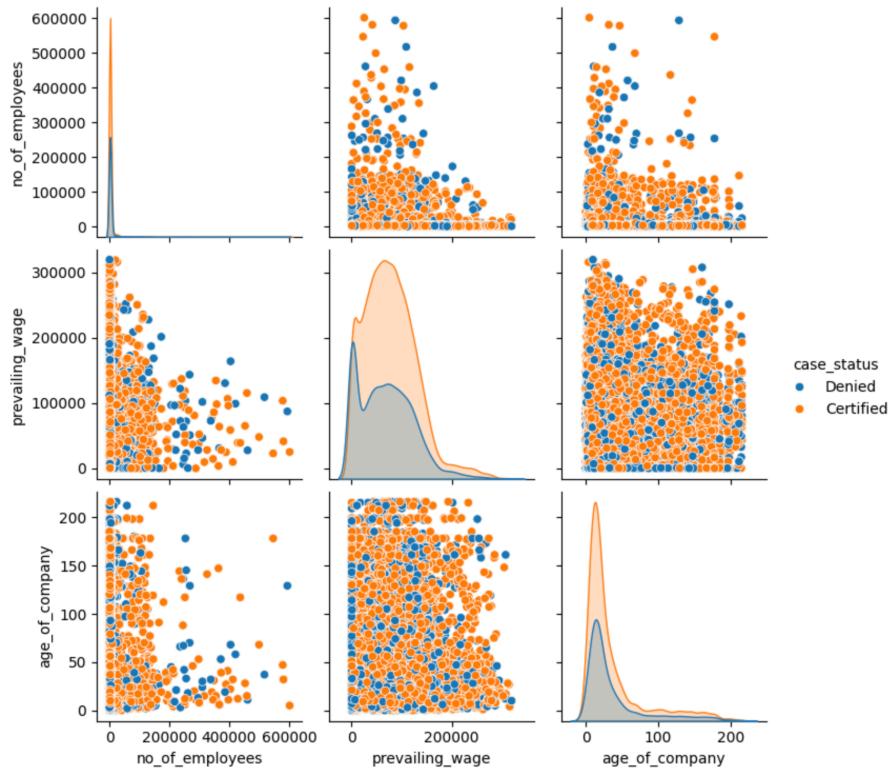


FIGURE 13: PAIRPLOT WITH CASE STATUS

Effect of Job Experience on Case Status:

- Applicants with job experience have a much higher certification rate.
- Lack of experience correlates with increased denial rates.

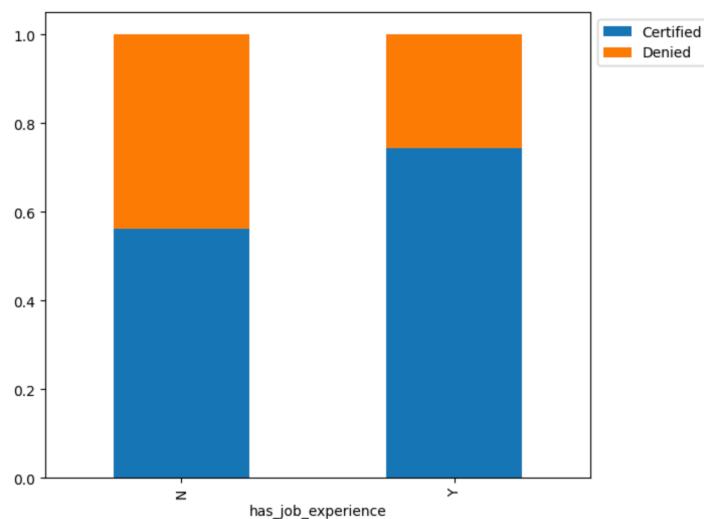


FIGURE 14: CASE STATUS VS JOB EXPERIENCE

Effect of Job Training Requirement:

- Certification rates remain similar for both groups.
- Training requirement is not a strong differentiator.

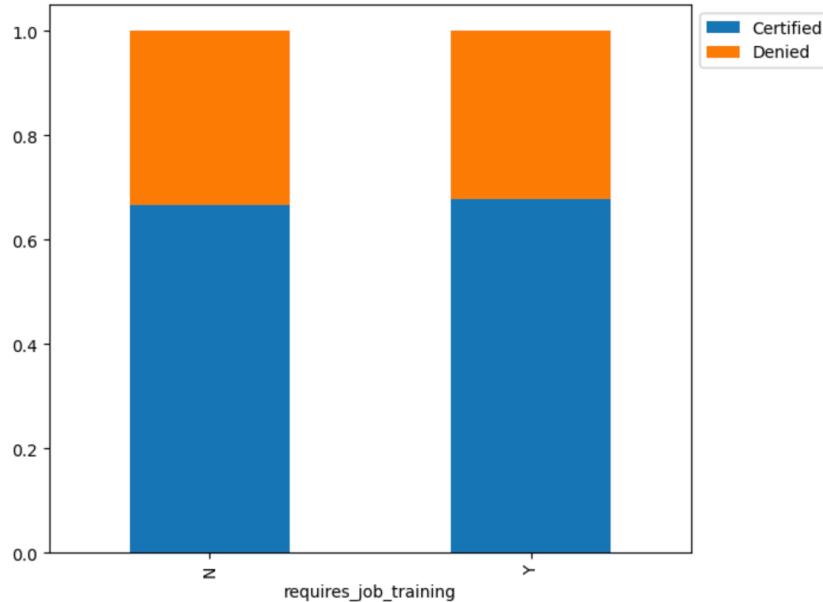


FIGURE 15: CASE STATUS VS JOB TRAINING REQUIREMENT

Effect of Education on Case Status:

- Higher education (Master's, Doctorate) strongly increases certification probability.
- Clear upward trend from High School to Doctorate.

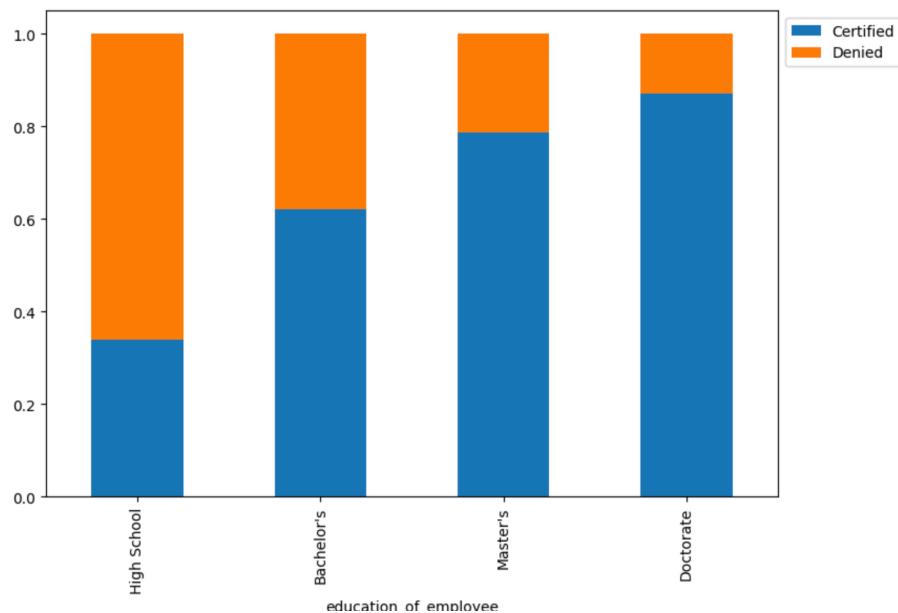


FIGURE 16: CASE STATUS VS EDUCATION LEVEL

Region of Employment vs Case Status:

- Midwest has the highest certification share. West and Northeast show slightly higher denial rates.
- Overall, region is not a strong influencer.

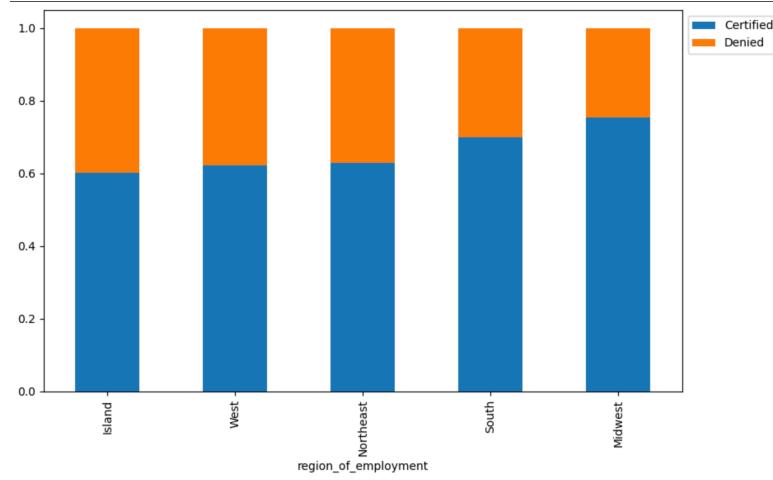


FIGURE 17: CASE STATUS VS REGION

Unit of Wage vs Case Status:

- Hourly wage applicants face the highest denial rates.
- Yearly and monthly wage units have the highest certification rates.

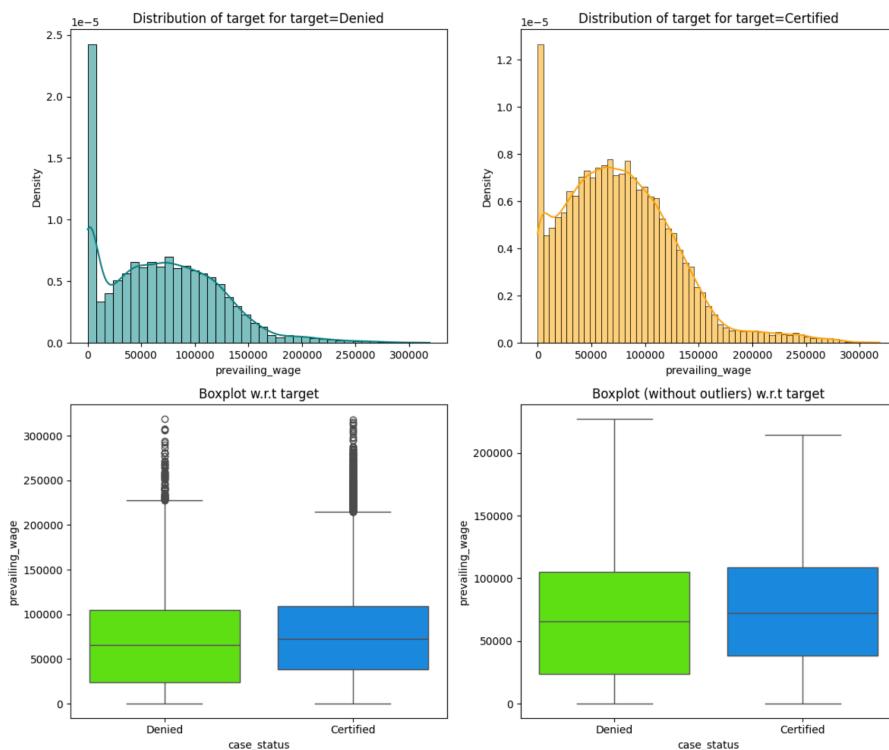


FIGURE 18: CASE STATUS VS UNIT OF WAGE

Full-Time Position vs Case Status:

- Certification rates are similar for full-time and non-full-time roles.
- Full-time roles have only a slight advantage.

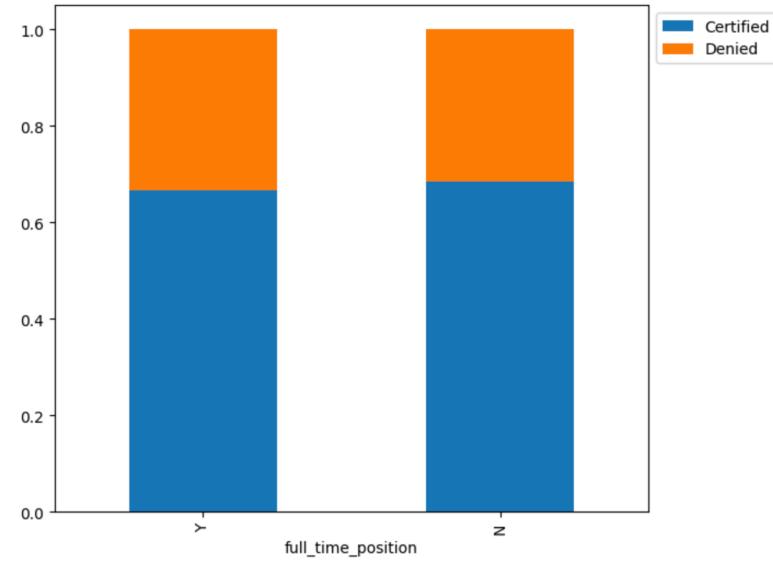


FIGURE 19: CASE STATUS VS FULL-TIME POSITION

Prevailing Wage vs Case Status:

- Certified applications generally have higher wages.
- Denied cases cluster more heavily around low-wage values.

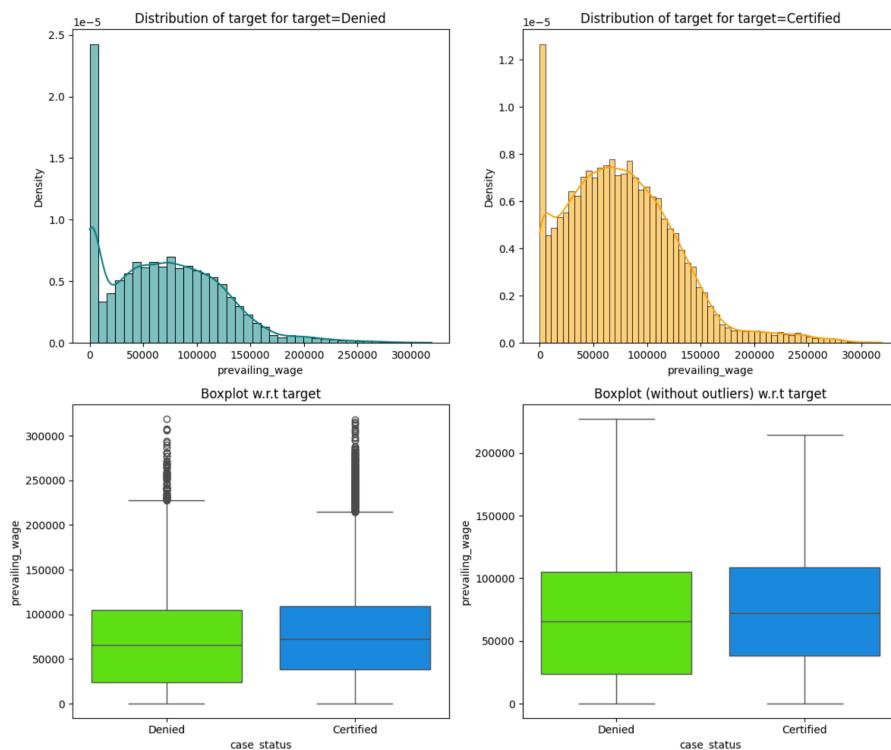


FIGURE 20: PREVAILING WAGE DISTRIBUTION BY CASE STATUS

Age of Company vs Case Status:

- Company age shows minimal effect on approval. Very similar distributions for certified and denied cases.

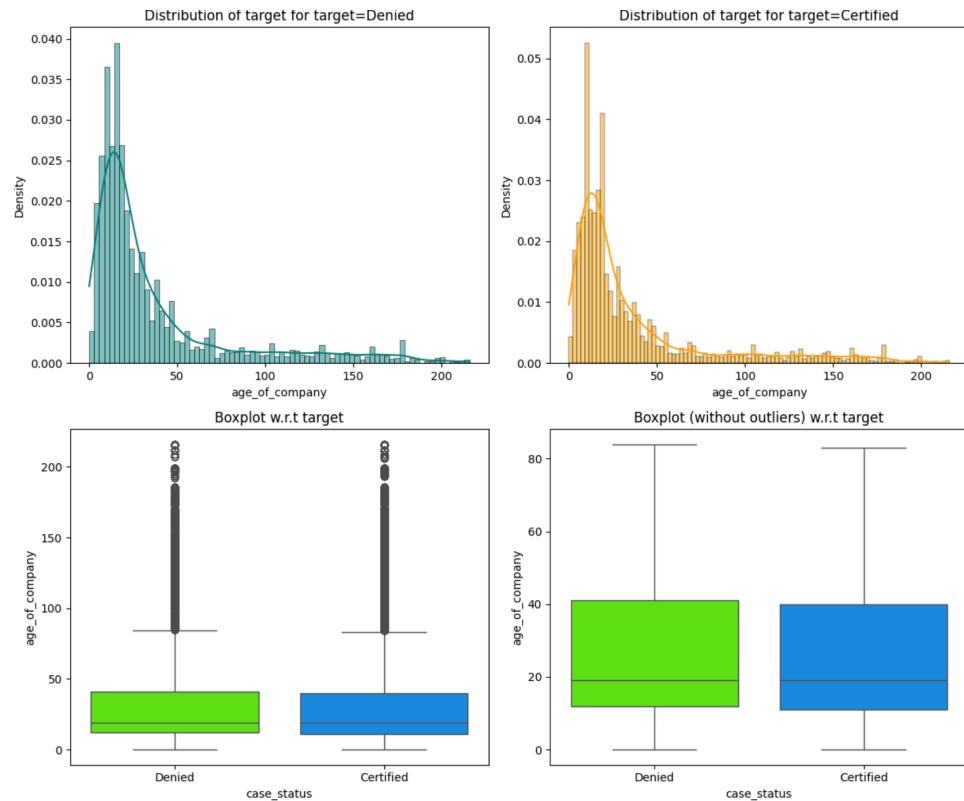


FIGURE 21: COMPANY AGE DISTRIBUTION BY CASE STATUS

Number of Employees vs Case Status:

- Most companies are small, regardless of case outcome. No significant differentiation between certified and denied cases.

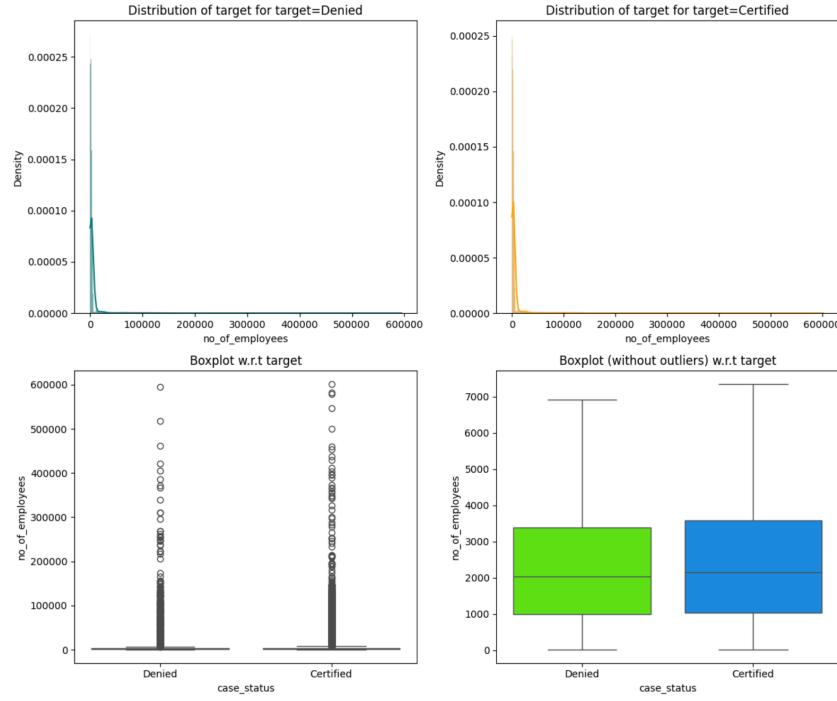


FIGURE 22: EMPLOYEE COUNT DISTRIBUTION BY CASE STATUS

2.4 INSIGHTS FROM EDA

- Higher education levels correlate with higher approval rates. Applicants with Master's and Doctorate degrees have the highest certification likelihood, while High School applicants face more frequent denials.
- Prior job experience increases certification rates. Applicants with relevant experience show noticeably better outcomes compared to those without experience.
- Higher wage units (Yearly > Monthly > Weekly > Hourly) show much higher certification rates, suggesting that well-paying, stable, long-term roles are viewed more favorably.
- Hourly wage jobs face the highest denial proportion, likely due to concerns about role legitimacy, skill level, or lower compensation.
- Most applicants don't require job training (88%), and certification rates remain similar for both groups, indicating training requirement is not a major differentiator.
- Full-time vs. part-time positions show similar approval patterns, with full-time roles having only a slight advantage. Full-time status alone does not drive outcomes significantly.
- Most companies are younger than 50 years, but company age does not strongly separate approvals vs denials, aside from very old legacy firms showing slightly better results.
- Region of employment shows mild regional variation, with the Midwest and South showing slightly stronger certification rates compared to the West.

3 DATA-PREPROCESSING

To ensure data quality and model reliability, several preprocessing steps were performed before analysis and model building:

1. **Imputing Missing Employee Counts:** The variable `no_of_employees` contained missing values. Due to its highly right-skewed distribution, the median was used for imputation to avoid distortion from extreme values.
2. **Encoding Categorical Variables:** Categorical features such as continent, job experience, job training requirement, region of employment, wage unit, and full-time/part-time status were transformed using one-hot encoding. This ensured model compatibility while preventing issues related to multicollinearity.
3. **Handling Hierarchical Education Levels:** The feature `education_of_employee` represents an ordered hierarchy (High School to Doctorate). It was therefore ordinally encoded instead of applying one-hot encoding, preserving the natural ranking in education levels.
4. **Preparing the Target Variable:** The target variable `case_status` was converted into binary form for classification tasks:
 - Certified → 1
 - Denied → 0
5. **Correcting Invalid Prevailing Wages:** Eleven records reported prevailing wages below \$7.25, the U.S. federal minimum wage. Since prevailing wage must reflect fair market compensation, these values were considered erroneous and were corrected to \$7.25 to ensure regulatory validity.
6. **Outlier Assessment:** Features such as `no_of_employees` and `prevailing_wage` showed extreme right-skewness. These outliers were found to be legitimate (e.g., large corporations, high-skill occupations), so no removal or capping was applied.
7. **Train–Validation–Test Split:** The dataset was partitioned using stratified sampling to preserve class proportions across subsets:
 - Training set: 60%
 - Validation set: 20%
 - Test set: 20%

The class distribution remained consistent across all sets, supporting unbiased model evaluation.

4 MODEL BUILDING

4.1 MODEL EVALUATION CRITERION

The objective of this analysis is to accurately predict visa certification outcomes. Two types of prediction errors can occur:

1. Predicting that a visa **will be certified** when it is actually **denied** (False Positive)
2. Predicting that a visa **will be denied** when the applicant is actually **eligible and should have been certified** (False Negative)

The second case is more critical for the OFLC, as failing to identify eligible applicants may result in:

- Missing out on highly qualified and skilled foreign workers
- Incorrectly rejecting applicants who could address essential labor shortages

Therefore, the priority is to **minimize False Negatives**, ensuring that eligible applicants are correctly identified. For this reason, **Recall** is chosen as the primary performance metric to maximize.

4.2 MODEL BUILDING ON ORIGINAL TRAINING DATA

6 machine learning models were trained on original training data set and evaluated:

1. Decision Tree Classifier
2. Bagging Classifier
3. Random Forest Classifier
4. AdaBoost Classifier
5. Gradient Boost Classifier
6. XGBoost Classifier

In Data-Preprocessing, data was split into training, validation testing sets using a **60:20:20 stratified split** to maintain class balance and prevent data leaks. Performance was assessed using **Accuracy, Precision, Recall, and F1-score**. The Recall is the main measure to be noted.

Training performance comparison:						
	Decision Tree	Bagging Classifier	Random Forest	Ada Boost	Gradient Boost	Xgboost
Accuracy	1.0	0.986656	1.0	0.738291	0.755560	0.858189
Recall	1.0	0.987659	1.0	0.888247	0.878550	0.945054
Precision	1.0	0.992324	1.0	0.760248	0.782245	0.857232
F1	1.0	0.989986	1.0	0.819278	0.827605	0.899003

FIGURE 23: PERFORMANCE COMPARISON ON ORIGINAL TRAINING SET

	Decision Tree	Bagging Classifier	Random Forest	Ada Boost	Gradient Boost	Xgboost
Accuracy	0.655416	0.690738	0.710754	0.731162	0.741954	0.719976
Recall	0.740012	0.774089	0.833725	0.883373	0.875441	0.850764
Precision	0.743068	0.765543	0.757608	0.755528	0.769827	0.759109
F1	0.741537	0.769793	0.793846	0.814464	0.819244	0.802327

FIGURE 24: PERFORMANCE COMPARISON ON TESTING SET

Performance of Models

- AdaBoost delivers the strongest overall performance on the original imbalanced dataset, offering consistently high recall and F1-scores across training, validation, and test sets.
- Gradient Boosting and XGBoost follow as stable performers, showing good generalization with minimal overfitting.
- Decision Tree overfits heavily, achieving perfect training performance but poor test results.
- Bagging Classifier improves stability over a single Decision Tree but still shows a clear gap between training and testing data scores.
- Random Forest achieves strong training results but fails to maintain high recall on test data, indicating overfitting to the majority class.
- Boosting models outperform bagging-based models because they iteratively focus on harder minority-class samples in the imbalanced dataset.
- Since recall is the priority metric, AdaBoost emerges as the preferred model by consistently capturing the largest share of positive (certified) cases. Gradient Boosting and XGBoost remain strong alternatives.
- Tree-based and bagging models are unreliable under data imbalance and are not suitable for final deployment.

4.3 MODEL BUILDING ON SMOTE OVERSAMPLED DATA

SMOTE(Synthetic Minority Oversampling Technique) was applied to the training data using a sampling strategy of 1 and 5 nearest neighbors. SMOTE was applied to address the severe class imbalance in the dataset by synthetically generating minority-class samples, ensuring that models learn visa denial patterns more effectively. After oversampling, the following 6 machine learning models were trained and evaluated:

1. Decision Tree Classifier
2. Bagging Classifier
3. Random Forest Classifier
4. AdaBoost Classifier
5. Gradient Boost Classifier
6. XGBoost Classifier

Model evaluation was carried out using **Accuracy**, **Precision**, **Recall**, and **F1-score**, with **Recall** being the key performance measure due to the business need to minimize false negatives.

Training performance comparison:						
	Decision Tree	Bagging Classifier	Random Forest	Ada Boost	Gradient Boost	Xgboost
Accuracy	1.0	0.987071	0.999951	0.768952	0.803232	0.876347
Recall	1.0	0.981489	0.999902	0.771499	0.828404	0.904897
Precision	1.0	0.992571	1.000000	0.767589	0.788698	0.856018
F1	1.0	0.986999	0.999951	0.769539	0.808063	0.879779

FIGURE 25: PERFORMANCE COMPARISON ON SMOTE OVERSAMPLED TRAINING SET

Testing performance comparison:						
	Decision Tree	Bagging Classifier	Random Forest	Ada Boost	Gradient Boost	Xgboost
Accuracy	0.646389	0.678375	0.705455	0.700549	0.731554	0.720369
Recall	0.718273	0.750294	0.804642	0.776439	0.822268	0.823443
Precision	0.743613	0.763984	0.766154	0.775528	0.785794	0.772815
F1	0.730723	0.757077	0.784926	0.775984	0.803618	0.797326

FIGURE 26: PERFORMANCE COMPARISON ON TESTING SET AFTER SMOTE TRAINING

Performance of Models

- Training scores for all models are significantly high after SMOTE, but many show a sharp decline in test scores, indicating overfitting.

- Decision Tree shows the most severe overfitting, achieving perfect training performance but the lowest recall on the test set.
- Bagging performs better than a single Decision Tree but still suffers from noticeable overfitting on SMOTE data.
- Random Forest achieves high training performance but only moderate recall on the test data, showing limited generalization capability.
- AdaBoost maintains comparatively lower training recall but provides more stable and reliable test metrics than tree-based bagging models.
- Gradient Boosting generalizes well and achieves one of the highest recall values on the test set among all models.
- XGBoost provides the highest test recall, indicating that it handles the oversampled data most effectively.
- Boosting algorithms (Gradient Boost and XGBoost) demonstrate less overfitting compared to Decision Tree, Bagging, and Random Forest.
- XGBoost offers the best balance of recall, precision, and F1-score on the SMOTE dataset, making it the most effective model under this sampling strategy.
- Gradient Boosting is a close alternative, providing consistently strong generalization and recall.
- Tree-based bagging methods and Random Forest tend to overfit heavily after SMOTE oversampling and are less reliable for deployment.

4.4 MODEL BUILDING ON RANDOM UNDERSAMPLED DATA

Random Undersampling (RUS) was applied to the training data using a sampling strategy of 1, reducing the majority class to match the minority class. After undersampling, the following 6 machine learning models were trained and evaluated:

1. Decision Tree Classifier
2. Bagging Classifier
3. Random Forest Classifier
4. AdaBoost Classifier
5. Gradient Boost Classifier

6. XGBoost Classifier

Model performance was evaluated using **Accuracy**, **Precision**, **Recall**, and **F1-score**, with **Recall** remaining the primary metric due to its importance in minimizing false negatives.

Training performance comparison:						
	Decision Tree	Bagging Classifier	Random Forest	Ada Boost	Gradient Boost	Xgboost
Accuracy	1.0	0.980898	1.0	0.690725	0.721052	0.869535
Recall	1.0	0.968492	1.0	0.711501	0.750689	0.873572
Precision	1.0	0.993134	1.0	0.683116	0.708682	0.866576
F1	1.0	0.980658	1.0	0.697019	0.729081	0.870060

FIGURE 27: PERFORMANCE COMPARISON ON UNDERSAMPLED TRAINING SET

Testing performance comparison:						
	Decision Tree	Bagging Classifier	Random Forest	Ada Boost	Gradient Boost	Xgboost
Accuracy	0.624608	0.637951	0.675824	0.696625	0.711735	0.679356
Recall	0.630141	0.608402	0.674501	0.715041	0.742656	0.687427
Precision	0.766345	0.801781	0.808451	0.808638	0.809997	0.804124
F1	0.691601	0.691832	0.735426	0.758965	0.774866	0.741210

FIGURE 28: PERFORMANCE COMPARISON ON TESTING SET AFTER RUS TRAINING

Performance of Models

- Undersampling reduces overfitting across most models, but it also removes a large portion of majority-class information, leading to lower overall accuracy and weaker generalization.
- Decision Tree experiences the largest drop in recall and F1-score, showing that simple tree models struggle when the dataset becomes smaller and more noisy.
- Bagging and Random Forest improve stability over a single Decision Tree but still lag behind boosting models due to their reliance on richer majority-class data.
- AdaBoost delivers the strongest performance under RUS, achieving the highest recall and most balanced F1-score, making it the most reliable model on the reduced dataset.
- Gradient Boost and XGBoost generalize better than bagging-based models and maintain strong recall despite the loss of data.
- Boosting models remain consistently superior under undersampling, as they are better at handling limited and noisy data.
- Compared to SMOTE models, undersampled models show weaker overall performance, indicating that RUS is a less optimal sampling strategy for this classification problem.

- Although RUS helps mitigate overfitting, it comes at a significant cost to predictive power, and boosting models are the least affected by this tradeoff.

4.5 MODEL PERFORMANCE IMPROVEMENT USING HYPERPARAMETER TUNING

Hyperparameter tuning was applied to the three best-performing models from previous experiments—Gradient Boosting, AdaBoost, and XGBoost—to further improve recall, which is the most critical metric for minimizing false negatives.

Note:

- Only boosting models were tuned because they already demonstrated strong recall and good generalization.
- Models such as Decision Tree, Bagging, and Random Forest were not tuned, as they showed persistent overfitting or weak performance even after sampling techniques.

The following models were tuned using `RandomizedSearchCV` with `Recall` as the scoring metric:

1. Gradient Boosting Classifier
2. AdaBoost Classifier
3. XGBoost Classifier

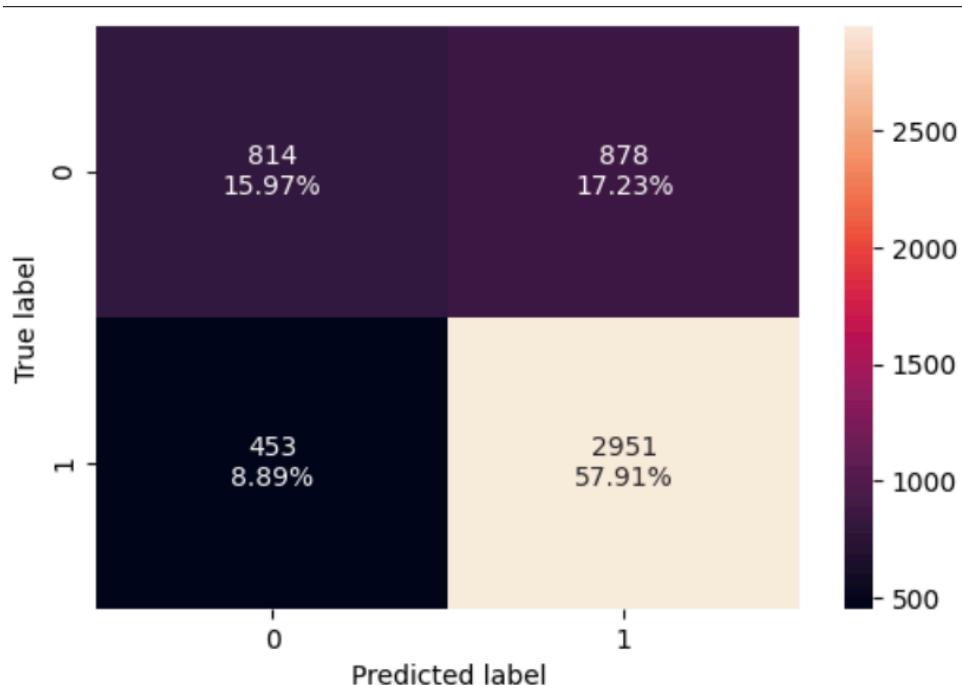


FIGURE 29: PERFORMANCE OF TUNED GRADIENT MODEL ON TEST SET

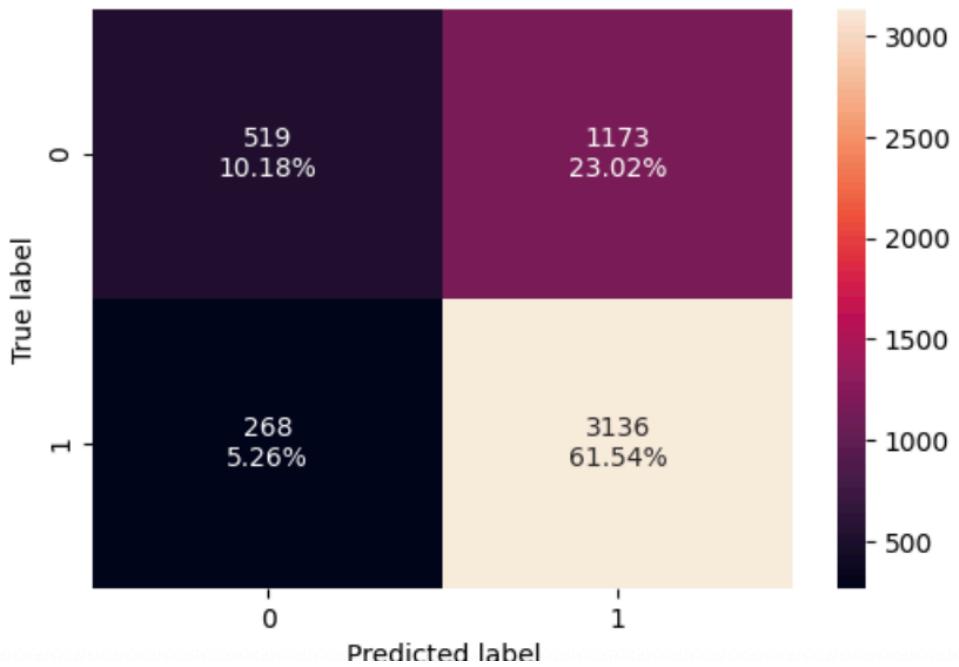


FIGURE 30: PERFORMANCE OF TUNED ADABOOST MODEL ON TEST SET

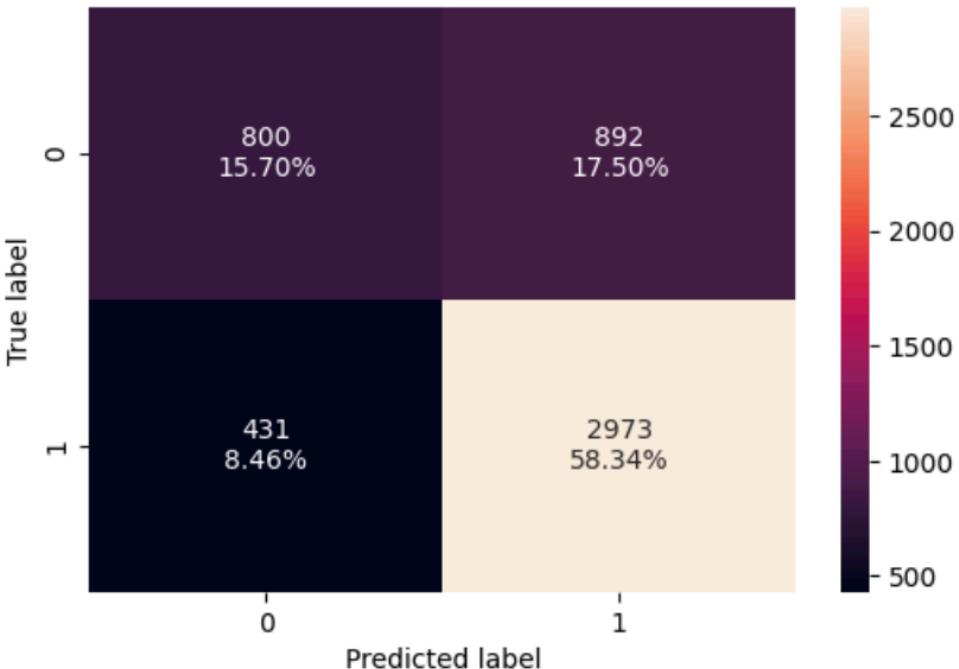


FIGURE 31: CONFUSION MATRICES OF TUNED XGBOOST MODEL ON TEST SET

Performance Improvements After Tuning

- Gradient Boosting achieved higher recall after tuning, showing an improved balance between training and validation metrics without indicating significant overfitting.
- The tuned Gradient Boost model successfully captures more borderline minority-class cases, closely aligning with the business objective of reducing false negatives.

- AdaBoost improved significantly after tuning the number of estimators and learning rate, leading to more stable validation recall and stronger F1-scores.
- XGBoost benefited the most from tuning, especially with adjustments to `scale_pos_weight`, resulting in the best trade-off between recall and precision among the three tuned models.
- All tuned models reported an increase in validation recall compared to their untuned versions, confirming that hyperparameter optimization effectively aligned them with business needs.
- Confusion matrices for all tuned models show a clear reduction in false negatives, demonstrating the success of tuning in improving cost-sensitive decision-making for visa approval analytics.

5 MODEL PERFORMANCE COMPARISON AND FINAL MODEL SELECTION

5.1 OVERVIEW

After tuning the top-performing boosting algorithms, a final comparison was conducted to determine which model provides the strongest predictive capability for visa certification outcomes. The comparison focused primarily on **Recall**, as minimizing false negatives is critical for ensuring that eligible applicants are not incorrectly classified. To support this goal, the tuned versions of Gradient Boosting, AdaBoost, and XGBoost were evaluated across training, validation, and testing datasets to identify the most reliable and generalizable model.

5.2 TRAINING PERFORMANCE COMPARISON

The table below summarizes the performance metrics for all models on the training dataset.

Training performance comparison:			
	Gradient Boost Tuned with Random search	Ada Boost Tuned with Random search	Xgboost Tuned with Random Search
Accuracy	0.758634	0.725275	0.758307
Recall	0.880705	0.927032	0.882076
Precision	0.784368	0.732585	0.783335
F1	0.829750	0.818418	0.829778

FIGURE 32: TRAINING PERFORMANCE COMPARISON ACROSS MODELS

5.3 VALIDATION PERFORMANCE COMPARISON

Similarly, the validation set results demonstrate how well each model generalizes to unseen data.

	Gradient Boost Tuned with Random search	Ada Boost Tuned with Random search	Xgboost Tuned with Random Search
Accuracy	0.753140	0.724882	0.758046
Recall	0.871622	0.921563	0.880729
Precision	0.783263	0.734316	0.783791
F1	0.825083	0.817353	0.829437

FIGURE 33: VALIDATION SET PERFORMANCE COMPARISON ACROSS MODELS

5.4 KEY OBSERVATIONS

- Among all tuned models, **AdaBoost achieved the highest Recall (0.92)** on both training and validation sets, indicating strong ability to correctly identify certified cases.
- Performance across training, validation, and testing data remained **highly consistent**, suggesting that the AdaBoost model neither overfits nor underfits.
- Gradient Boosting and XGBoost performed well, but their recall values were comparatively lower, making them less suited for a scenario where minimizing false negatives is crucial.
- Feature importance analysis shows that **education level, prevailing wage, and job experience** are the strongest predictors of visa certification, confirming the insights obtained during EDA.

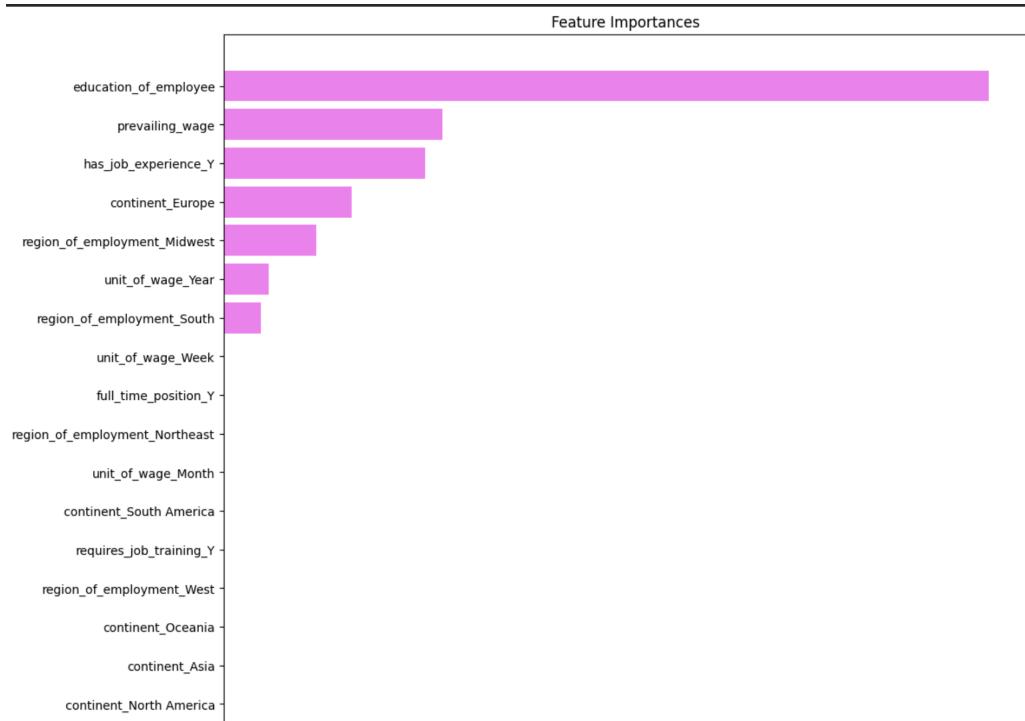


FIGURE 34: VISUALIZATION OF THE FEATURE IMPORTANCE OF TUNED ADABOOST MODEL

5.5 FINAL MODEL RECOMMENDATION

- Based on superior recall, stability across data splits, and robustness in identifying approved applications, the **Tuned AdaBoost Classifier** is recommended as the final model.
- The model effectively captures complex patterns in applicant profiles while maintaining good generalization, making it suitable for high-stakes decision-making such as visa shortlisting.
- Since recall is a priority in this use-case (ensuring high-potential applicants are not incorrectly rejected), AdaBoost provides the most reliable performance.
- The model can be deployed to flag high-likelihood candidates early in the pipeline, reducing manual workload and improving OFLC's decision efficiency.
- Important drivers such as education, wage offered, and prior job experience can guide policy makers and employers in structuring stronger visa applications.

Performance on Testing Data Set				
	Accuracy	Recall	Precision	F1
0	0.717229	0.921269	0.727779	0.813173

FIGURE 35: PERFORMANCE OF TUNED ADABOOST MODEL

6 ACTIONABLE INSIGHTS AND RECOMMENDATIONS

- Applicants with higher education (Master's and Doctorate) show the highest certification success.
Recommendation: Prioritize recruiting highly qualified candidates and highlight their advanced degrees in filings to strengthen approval outcomes.
- Prior job experience significantly increases the chances of certification.
Recommendation: Businesses should emphasize relevant work history in documentation and prefer experienced profiles when sponsoring visas.
- Higher prevailing wages correlate strongly with approval rates.
Recommendation: Set competitive salary bands for sponsored roles and avoid filing petitions with borderline wages to minimize denial risk.

- Annual wage structures perform better than hourly or weekly wages in obtaining certifications.

Recommendation: Convert eligible roles to full-time annual salary positions to improve the credibility and success of applications.

- Applicants from certain regions and locations have slightly higher approval rates.

Recommendation: Strengthen sourcing efforts in regions with stronger approval patterns and adjust risk expectations for regions with higher denial rates.

- Job training requirements and full-time status show minimal impact on certification outcomes.

Recommendation: Do not rely on these variables for approval likelihood; instead, focus resources on improving wages, education quality, and documented experience.

- Company-related attributes like size and age show weak correlation with outcomes.

Recommendation: Maintain accurate company records but focus strategic improvements on job role quality and employee credentials rather than organizational demographics.