AI-Driven Visa Approvals: Streamlining Foreign Labor Certification

EasyVisa: Post Graduate AI/ML Business Applications

April 22, 2025

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix

Executive Summary

- No of Employees - The presence of outliers exists for very large corporations that could have distinct operational practices, resource allocations, and economic impacts compared to smaller firms.
- Year Establish - the majority of establishments were founded between approximately 1950 and 2000, with the median year situated slightly above 2000
- Prevailing Wage - outliers represent industries or roles with premium compensation, highly specialized positions, and other exceptional circumstances that drive wages upward. Several data points extend well above this threshold, reaching as high as 300,000
- Applicants with Bachelor's degrees (40.2%), followed by those with Master's degrees (37.8%). In contrast, the proportions for Doctorate holders (8.6%) and employees with only High School diplomas (13.4%) are significantly lower
- The Northeast region leads with the highest percentage of employment at 28.2%, indicating a significant concentration of job opportunities in this area
- A substantial number of applicants entering the U.S. workforce under the Immigration and Nationality Act (INA) bring prior job experience
- The certification rate of 66.8% demonstrates the INA effectiveness in allowing skilled professionals to join the U.S. workforce
- Number of Employees vs. Year Established a -0.02 suggests that the size of an organization’s workforce is largely independent of its year of establishment
- Number of Employees vs. Prevailing Wage -0.01 his finding highlights that organizations, irrespective of their size, adhere to prevailing wage standards mandated by U.S. labor laws

Executive Summary

- Geographic location remains influential in the US, with European employment patterns factoring into the analysis
- Education Level (High School) – This feature holds the highest importance, indicating that an individual's education level plays a major role in classification outcomes
- The presence of various features with non-negligible importance indicates that the predictive task depends on a combination of factors rather than a single dominant variable.
- Prevailing Wage consistently remains a critical factor, aligning with the importance of compliance in visa applications.
- Yearly Wages applications associated with yearly wages dominate and enjoy the highest certification rates
- Weekly and Monthly Wages these categories exhibit moderate certification rates
- Year Established vs. Prevailing Wage correlation of 0.01 suggests that an organization’s age has little to no bearing on the wages it offers

Executive Summary

Actionable insights

- Majority of applicants hold a Bachelor degree 40% or Master degree 38%
- The Northeast region leads with the highest percentage of employment at 28.2%
- The South and West regions follow closely, with 28% and 26% of applicants finding employment there
- The Island region has the lowest percentage of applicants finding employment at just 2%
- Applicants with no job experience represents a significant portion but still less than those with experience
- Year establish and number of employees does not influence certification process. Which aligns with act's INA framework.
- Wage and Year established correlation has a weak positive 0.01 relation that suggest it has no bearing on the wages companies offers.

Executive Summary

Actionable insights

- Notable percentage of applicants without experience suggests opportunities for entry-level roles or training programs to integrate these individuals into the workforce effectively.
- The denial rate of 32% reflects the INA enforcement of stringent eligibility requirements, ensuring that only qualified candidates are approved to enter the workforce.
- The INA prioritizes other criteria, such as job qualifications, industry needs, and regional demands, to uphold equitable access to the U.S. workforce.
- Certification Rates by Unit of Wage Year - Certification Rate: 69.9%, Denial Rate: 30.1%
- The high certification rate aligns with the preference for stable, long-term employment contracts that fulfill INA requirements and not by Hour (747 Certified, 1410 Denied):
- Best Performer: GBM stands out as the most balanced algorithm, excelling in both training and validation performance, making it ideal for tasks requiring strong generalization.

Executive Summary

Actionable insights

- Cross-Validation Performance on the Training Dataset: Adaboost: With the highest score of 0.889 in cross-validation, Adaboost demonstrates exceptional performance in fitting the training data, suggesting its robustness in capturing patterns.
- Overfitting Risk: Algorithms like Bagging and Decision Tree exhibit overfitting tendencies, making them less suitable for datasets requiring robust validation performance.

Executive Summary

Recommendations for the Office of Foreign Labor Certification

- **To streamline the visa approval process**, we **propose establishing** a **robust, data-driven evaluation framework** that leverages advanced ensemble algorithms. **Our analysis** shows that **Gradient Boosting Machines (GBM)** demonstrate balanced performance on both training and validation sets, making them an excellent baseline for decision support. **In addition, models** like **AdaBoost** have **proven effective** in capturing complex patterns from the training data, though caution is warranted given the overfitting tendencies observed with methods such as Bagging and Decision Trees. **By incorporating rigorous cross-validation** and utilizing **ensemble methods** (e.g., **Random Forest as a meta-model** in a **stacking ensemble**), the **Office can improve consistency**, **reduce bias**, and ensure a more **reliable** and efficient process for **visa approvals.**

Executive Summary

Recommendations for the Office of Foreign Labor Certification

- For **recommending a suitable applicant profile**, it is essential to **integrate key drivers** that significantly influence certification outcomes. **Our findings indicate** that **prevailing wage is a critical factor**: **certified** applications exhibit **a higher median wage** (around $70,000) compared to denied applications (approximately $60,000). **Moreover,** the **analysis of wage units confirms** that **long-term**, **yearly wage structures are strongly associated** with **higher certification rates (69.9% certified vs. 30.1% denied**), emphasizing the importance of stable employment contracts over hourly arrangements. In addition, incorporating demographic and operational factors—such as the tendency for the majority of companies to have a small number of employees and the significant **regional employment patterns observed** in areas like the **Midwest and the southern region**—further refines the applicant profiling process.

Executive Summary

Recommendations for the Office of Foreign Labor Certification

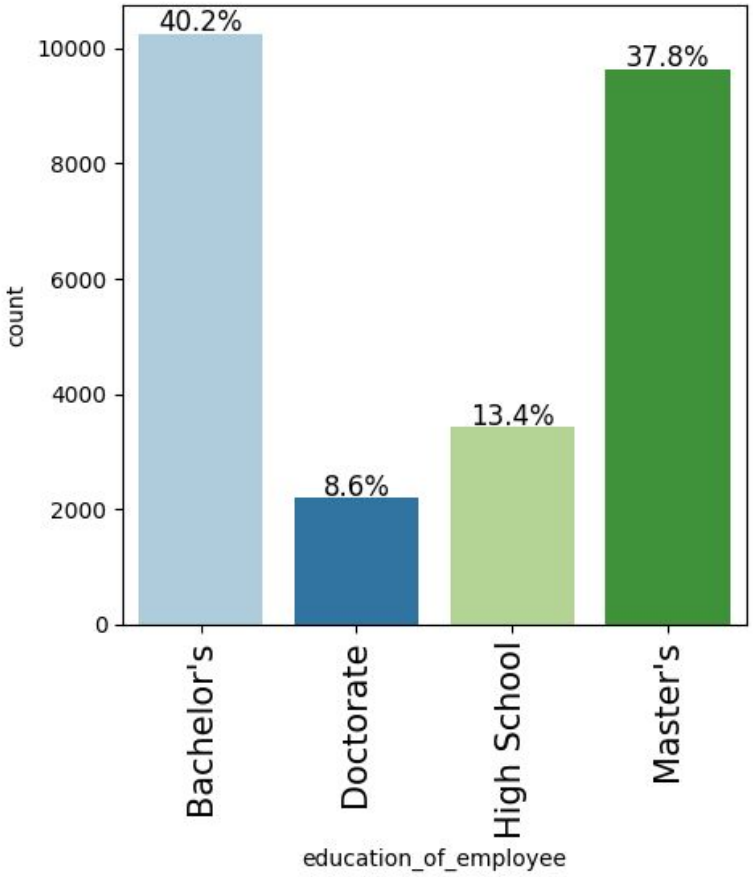
- **Integrating these insights into decision-making**, we recommend **developing a comprehensive, model-based decision-support system** that balances advanced algorithmic performance with practical applicant characteristics. This system should employ a validated ensemble approach—**leveraging GBM and AdaBoost outputs for strong predictive performance**—coupled **with feature engineering** that highlights **wage metrics, contract type, company size, and regional employment dynamics**. By aligning model-driven insights with operational criteria, the **Office of Foreign Labor Certification** can **achieve a dual objective**: **facilitating expedited visa approvals** while ensuring that **certification decisions** are **guided** by the most significant and **well-founded drivers** that predict case outcomes.

Business Problem Overview and Solution Approach

Under the **U.S. Immigration and Nationality Act (INA)**, foreign workers may obtain temporary or permanent visas only if U.S. employers demonstrate a shortage of qualified domestic candidates at competitive wages. This process is administered by the Office of Foreign Labor Certification (OFLC) to safeguard domestic labor markets. However, in FY 2016, the OFLC processed 775,979 employer applications for 1,699,957 positions—a **9% increase** from the previous year. **As a result**, the manual review process has become increasingly tedious and error-prone, straining administrative capacity and delaying visa approvals.

To address these challenges, it is recommended that OFLC **adopt a data-driven**, **machine learning solution** to automate the initial screening of visa applications. **By developing a robust classification model** that predicts visa approval outcomes based on key historical and contextual factors, the review process can be streamlined significantly. **This approach will reduce administrative burdens,** ensure more **consistent decision-making,** and **expedite overall processing times**—ultimately supporting the timely integration of qualified foreign talent into the U.S. workforce while maintaining strict **adherence to statutory requirements**.

EDA Results

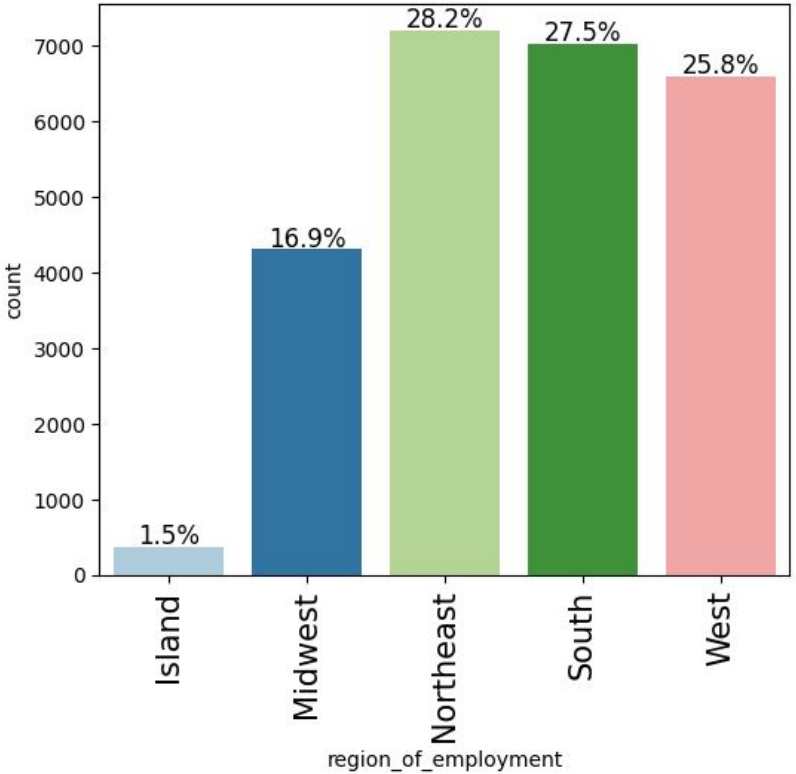


Education of Employee

This data provides insight into the educational composition of the workforce entering the United State. In the context of the Immigration and Nationality Act (INA), it underscores the importance of higher education qualifications among foreign workers seeking employment in the United States. **Bachelors** and **Masters** degrees combine is **78% of applicants**.

Link to Appendix slide on data background check

EDA Results

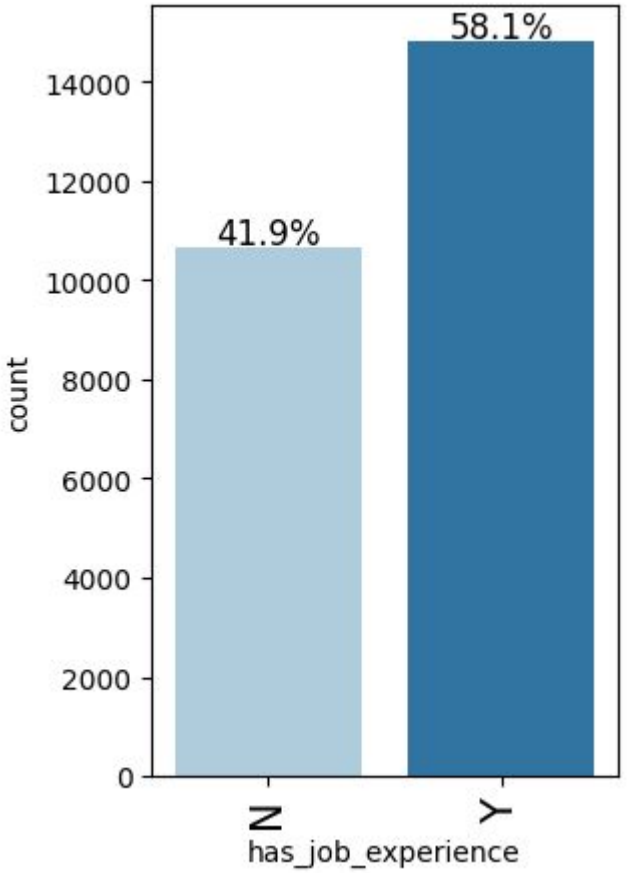


Region of Employment

This distribution reflects the varying **economic landscapes** and job markets across the United States. The dominance of the **Northeast**, **South,** and **West regions** aligns with their **larger** urban centers and **diverse industries**, which attract a significant portion of foreign workers. Meanwhile, the lower percentages in the **Midwest** and **Island** regions may indicate **fewer opportunities** or a **smaller demand** for foreign labor

Link to Appendix slide on data background check

EDA Results



Has Job Experience

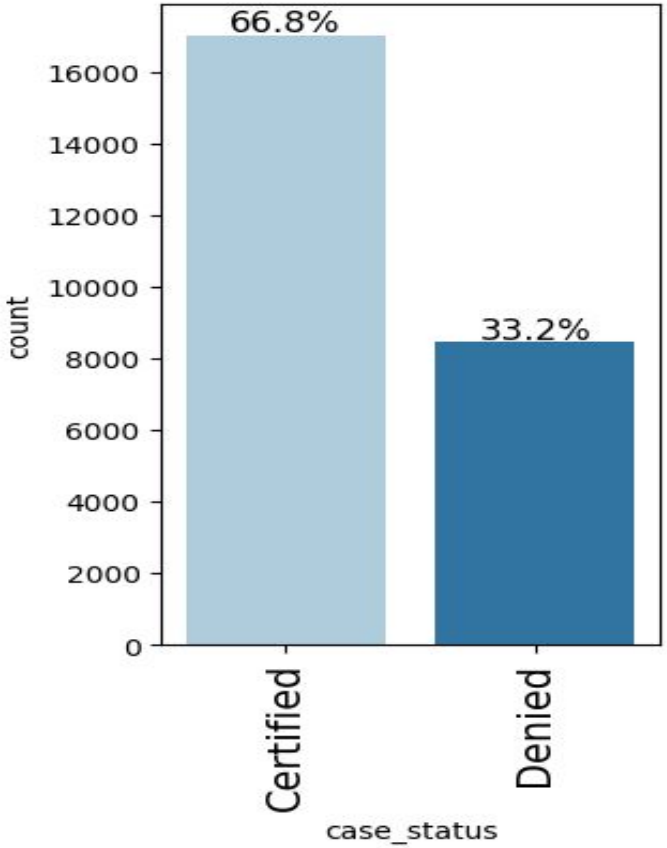
- A substantial number of applicants entering the U.S. workforce under the Immigration and Nationality Act (INA) bring prior job experience, which could make them more competitive in the job market.

58.1% of applicants have prior job experience **Y**, making them the majority.

41.9% of applicants lack job experience **N**, representing a significant portion but still less than those with experience.

Link to Appendix slide on data background check

EDA Results



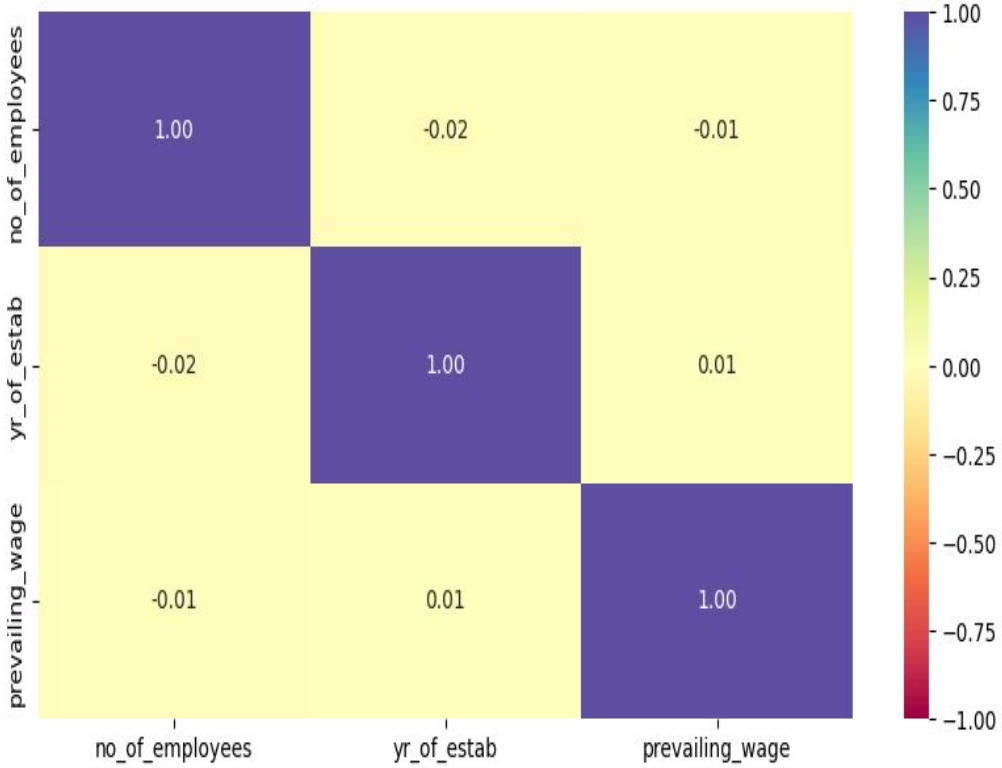
Case Status

As governed by the Immigration and Nationality Act (INA). The case status outcomes—**certified 66.8%** and **denied 33.2%**—highlight the rigorous selection process aimed at ensuring that applicants meet specific qualifications and standards set forth by U.S. immigration policies.

The **denial rate reflects** the act's enforcement of **stringent** eligibility requirements, ensuring that only qualified candidates are **approved** to enter the workforce

Link to Appendix slide on data background check

EDA Results



Correlationship

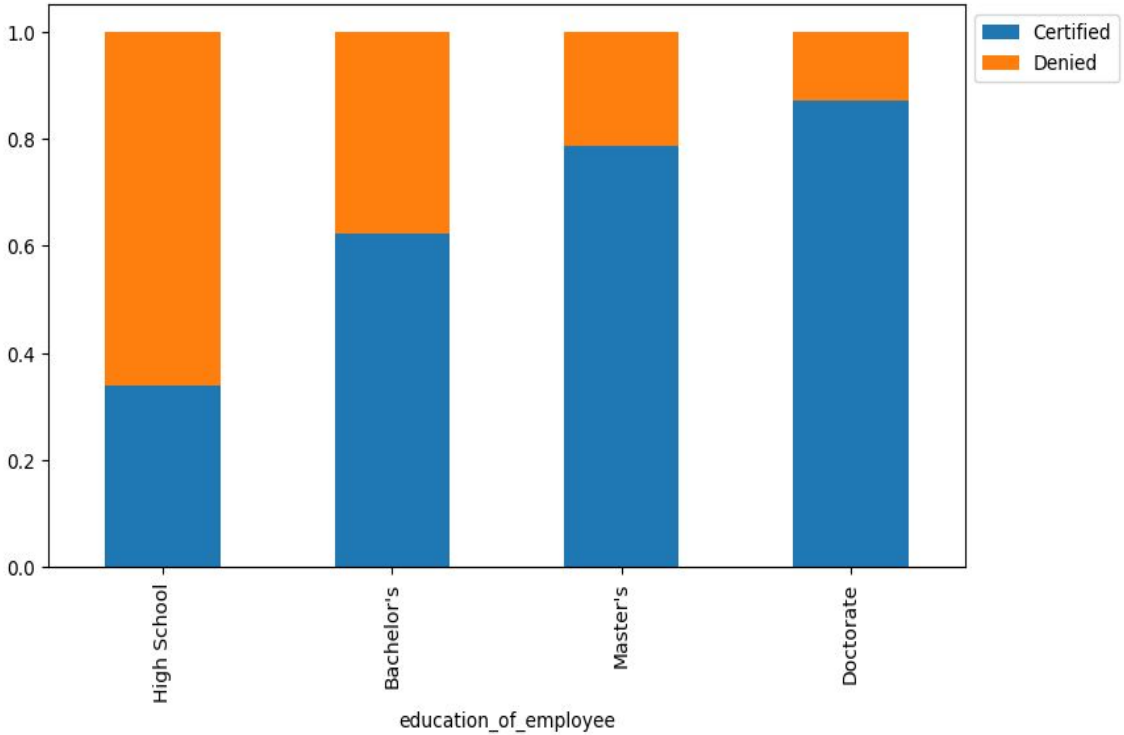
Number of Employees vs. Year Established The correlation of **-0.02** indicates a negligible negative relationship

Number of Employees vs. Prevailing Wage The correlation of **-0.01** reflects a minimal negative relationship between workforce size and prevailing wage

Year Established vs. Prevailing Wage 0.01 This indicates that both older and newer companies comply equally with prevailing wage requirements

Link to Appendix slide on data background check

EDA Results



Education of Employee Case Status

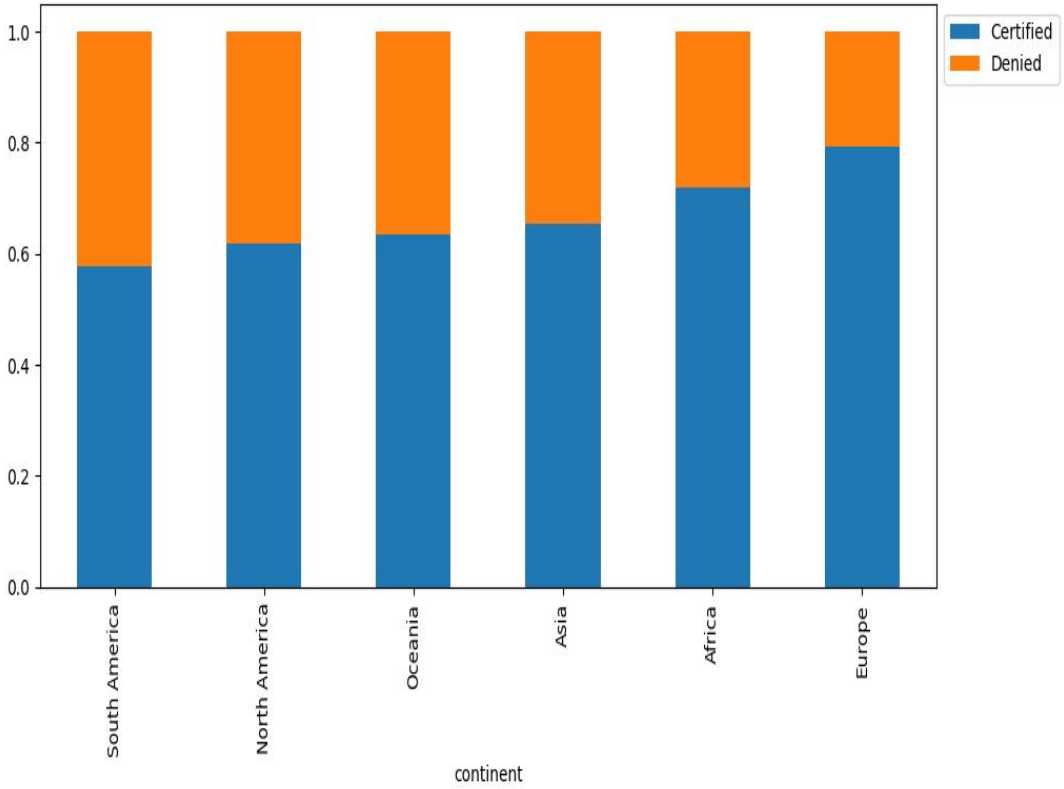
Bachelor's Degrees among them, 6,367 (62%) were certified, and 3,867 (38%) were denied.
This suggests a moderate certification rate for bachelor's degree holders, with a noticeable increase in opportunities compared to high school diploma holders

Master's Degrees Of the **9,634** applicants with a master's degree, 7,575 **(79%) were certified**, while 2,059 **(21%) were denied**.

Doctorate Degrees Out of 2,192 applicants with a doctorate, **1,912 (87%) were certified**, and only 280 **(13%) were denied**.

Link to Appendix slide on data background check

EDA Results



Case Status By Continent

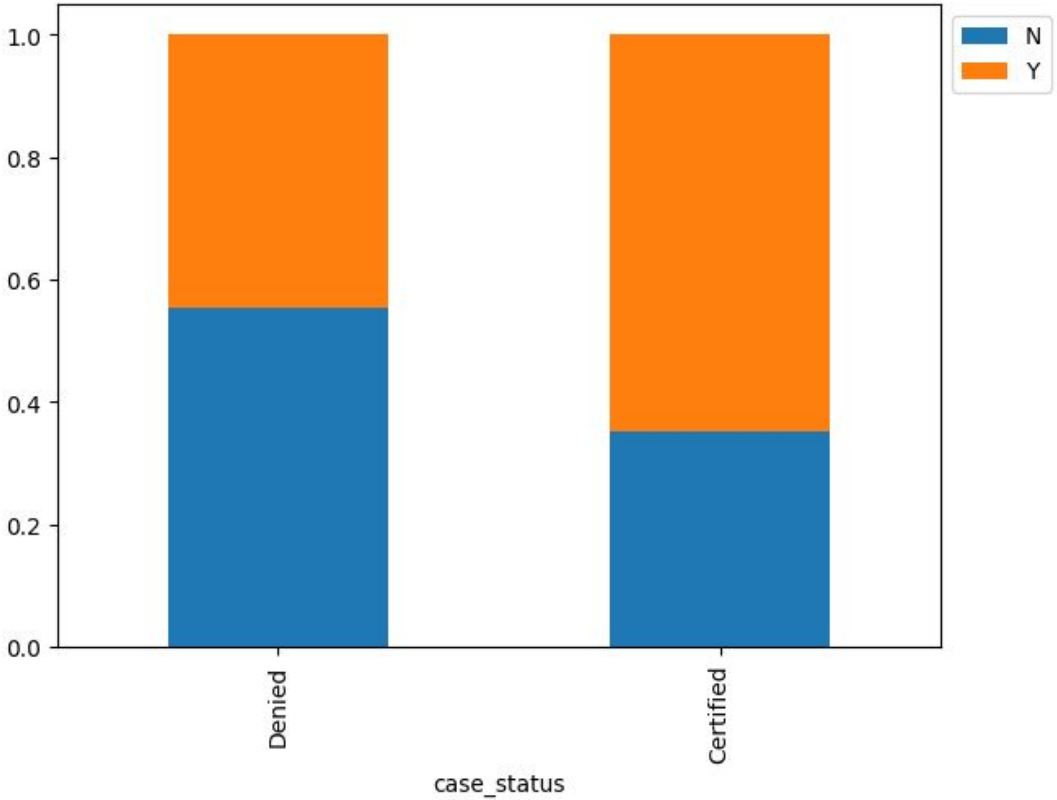
Asia has the **highest** number of applications, with a majority successfully **Certified: 11,012 (65.4%)**

Europe shows the **highest certification rate across all continents**, suggesting that applicants from this region often meet or exceed the INA's **Certified: 2,957 (79.2%)**

South America Denied: 359 (42.1%) indicates challenges faced by South American applicants

Link to Appendix slide on data background check

EDA Results



Case Status by Job Experience

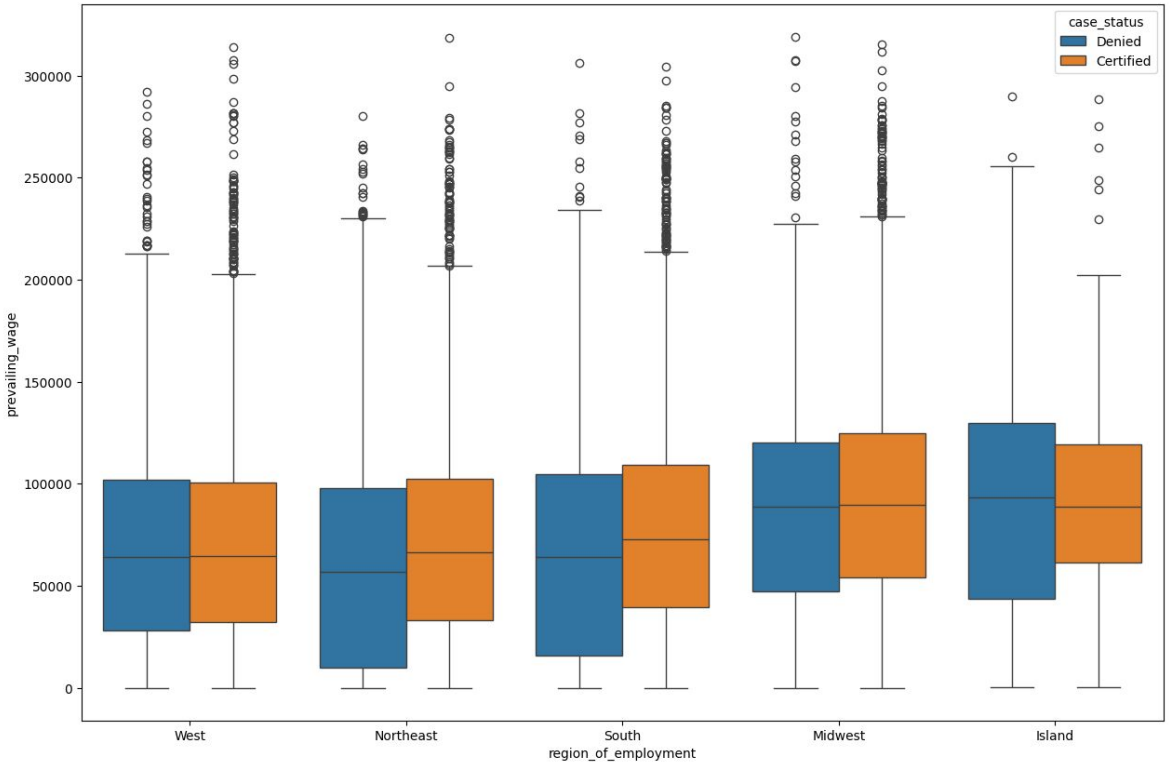
Out of 25,480 total cases, the majority (14,802, 58%) involve applicants with job experience (Y), while 10,678 (42%) involve applicants without job experience (N).

Certified Rate: 11,024 (**74.5%** of cases with job experience).

Denied Rate: 3,778 (**25.5%** of cases with job experience).

Link to Appendix slide on data background check

EDA Results



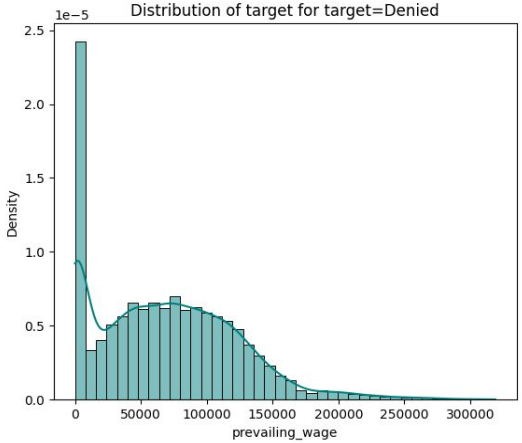
Prevailing Wage Across U.S. Regions

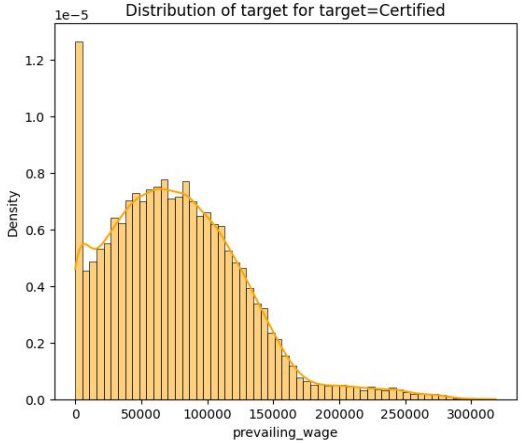
The **median prevailing wage is 50% relatively consistent across all regions,** indicating that wages offered to foreign workers under the Immigration and Nationality Act **(INA)** are **standardized to a degree**, regardless of location

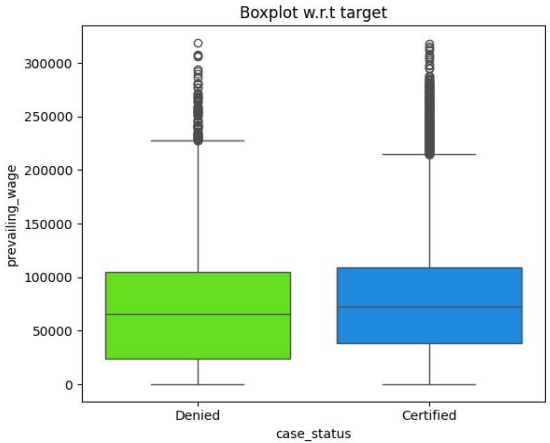
Wages for denied applications show a slightly lower distribution, which may indicate that these applications **did not meet** the **wage standards** or other criteria required for certification.

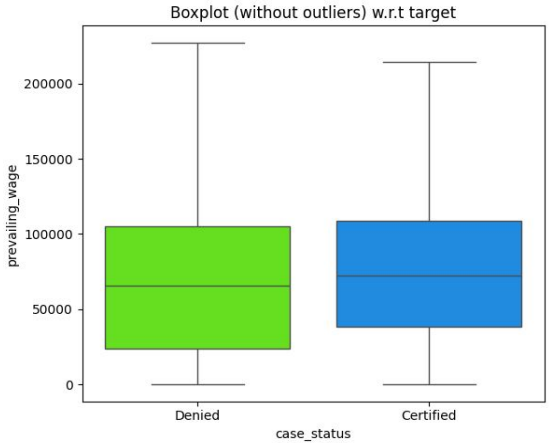
Link to Appendix slide on data background check

EDA Results









Distribution of Prevailing Wages

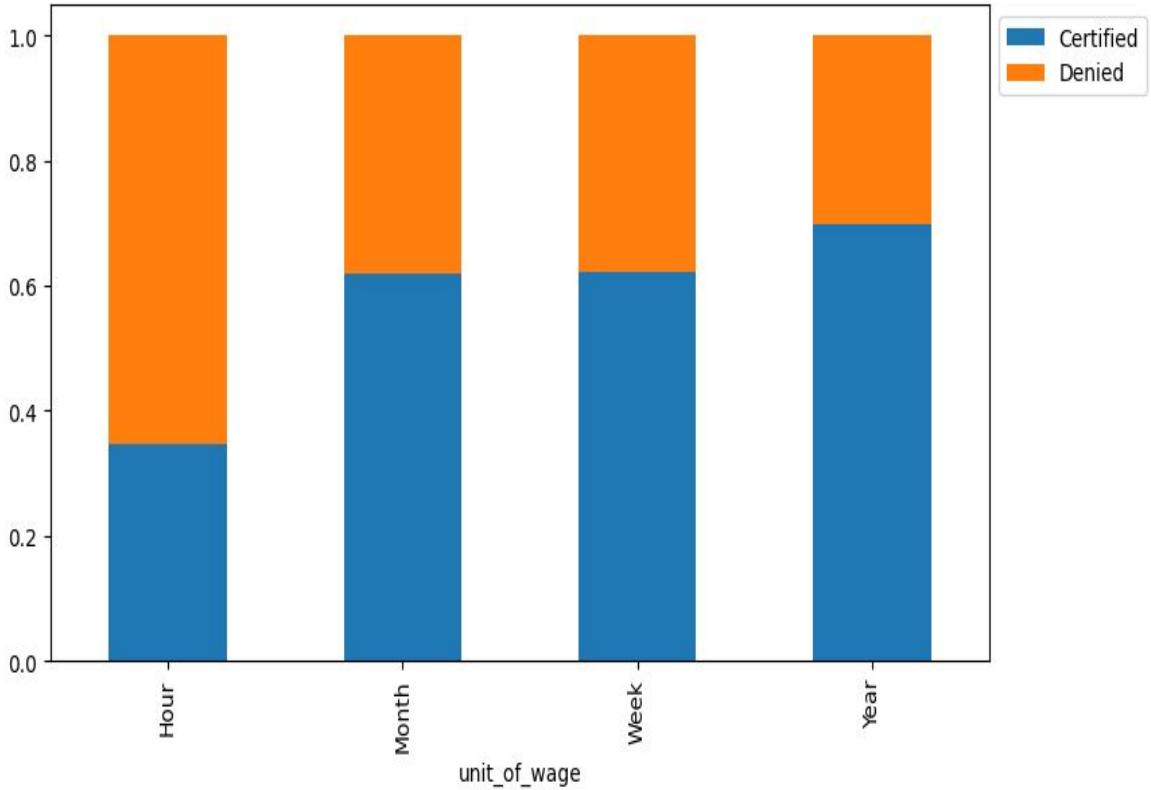
The **Denied Peak Density** density gradually declines as the prevailing wage increases. There are modest secondary **peaks** in the ranges of **50,000–60,000** and **90,000–100,000**.

Concentration at the Lower Wage Range for Both Denied and Certified distributions exhibit high density near the 0–10,000 range.

The **boxplots** indicate that **Certified cases** tend to command a **higher** prevailing **wage**. (**median around 70,000**) *compared to* ***Denied*** *cases* (***median around*** **60,000**)

Link to Appendix slide on data background check

EDA Results



Unit of Wage and Case Status

Certification Rates by Unit of Wage Year

- 16047 Certified
- 6915 Denied

Weekly wage-based applications exhibit a moderate certification rate, though the smaller sample size limits broader conclusions.

- Week (169 Certified, 103 Denied)
- Certification Rate: 62.1%
- Denial Rate: 37.9%

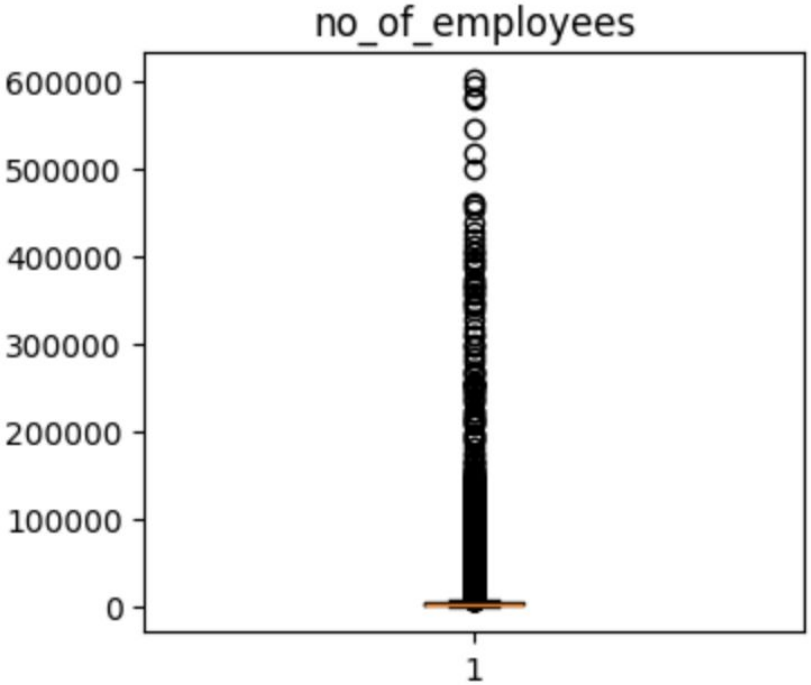
Link to Appendix slide on data background check

Data Preprocessing

- **Are there any duplicate entries in the dataset:**
  - No Duplicate entries exists
- **Missing value treatment:**
  - The dataset contains no missing values, as confirmed through a comprehensive verification process.
- **Outlier check**
  - Outliers does exists for **No of Employees, Year Established and Prevailing Wages**. After further analysis keeping the outliers in the dataset is the best option. The dataset is highly skewed and removing or correcting the outliers will impact the analysi

Data Preprocessing

- Outlier check No of Employees

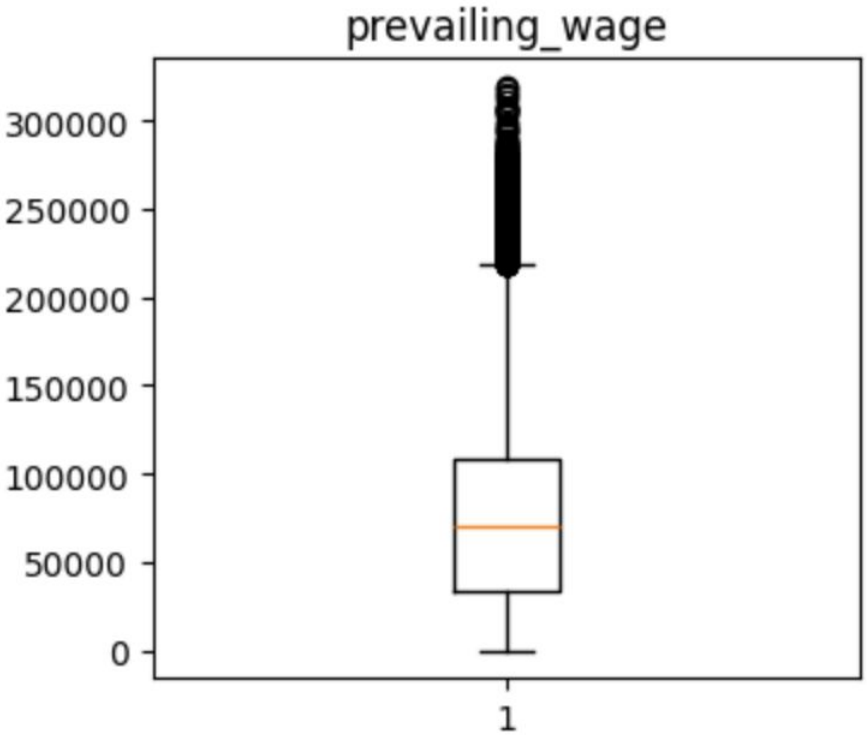


The **majority of companies** have a relatively **small number of employees**, as indicated by the **dense cluster** near the lower end of the scale. However, a significant number of **outliers**—represented by individual circles—**extend** all the way up to approximately **600,000 employees**.

This suggests that while typical companies remain small, a few organizations with extensive workforces markedly differentiate themselves from the majority.

Data Preprocessing

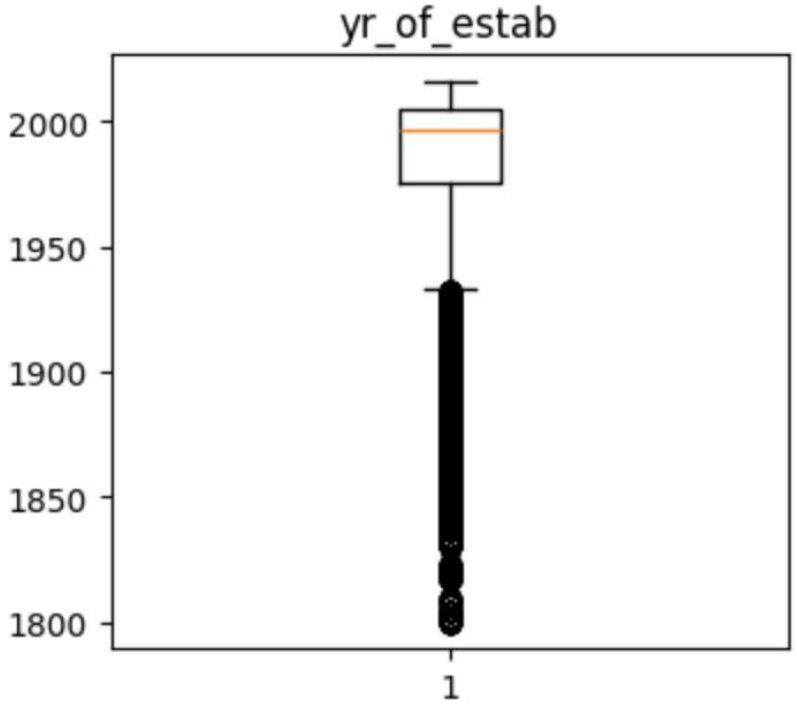
- Outlier check Prevailing Wage



Indicates that the **bulk** of the **wage data** lies **below** approximately **200,000**, which represents the upper bound of the interquartile range (IQR). However, **several data points** extend well above this threshold, **reaching as high as 300,000**. These points, plotted as outliers above the upper whisker, suggest the presence of **exceptionally high wages** relative to the majority of the dataset.

Data Preprocessing

- Outlier check Year Established



The **majority of companies** have a relatively **small number of employees**, as indicated by the **dense cluster** near the lower end of the scale. However, a significant number of **outliers**—represented by individual circles—**extend** all the way up to approximately **600,000 employees**.

This suggests that while typical companies remain **small**, a few organizations with extensive **workforces** markedly differentiate themselves from the **majority**.

Data Preprocessing

- Feature engineering

  The **No of Employee** column included **negative values**, which may skew data analysis. **To mitigate** this issue and maintain data integrity, replaced these negative values with their absolute values.

Continent	Negative Count
Africa	1
Asia	21
Europe	5
North America	5
South America	1
Africa	1

Data Preprocessing

Data preparation for modeling

- **Drop case id** from dataset. It is not a importance feature in the visa application approval process.
- **Case Status** updated Case Status to 1 to indicated a Certified Status, and 0 for Denied Status

One-hot encoding technique was applied to categorical features to convert to binary

To ensure effective model development and evaluation, the dataset was divided into three distinct subsets

- **Training Set:** Used for model learning and hyperparameters tuning to enhance predictive accuracy.
- **Validation Set:** Applied to refine the model and safeguard against data leakage, ensuring generalizability.
- **Test Set:** Reserved for inference on unseen data, providing an unbiased evaluation of final performance.

Model Building - Bagging - Decision Tree Classifier

Steps for Building Decision Tree Classifier with StratifiedKFold:

1. **Data Splitting Using StratifiedKFold**:
   - `StratifiedKFold` splits the data into **k folds** while maintaining the **class distribution** in each fold.
   - For each fold:
     - A subset of data (the validation set) is held out, and the remaining folds (the training set) are used to train the model.
     - The **DecisionTreeClassifier** is then trained exclusively on the current training set for that fold.

2. **Tree Construction on Training Fold**:
   Within each fold's training data, the DecisionTreeClassifier follows its standard process:

   - It selects the **best feature** to split the data based on the splitting criterion (e.g., Gini Impurity or Entropy).
   - Recursively splits the data until a stopping condition is met (e.g., maximum depth, minimum samples per leaf, or achieving pure leaf nodes).
   - As it trains, the model uses only the data available in that specific fold, ensuring that it does not "see" data from the validation set for that fold.

Model Building - Bagging - Decision Tree Classifier

3. **Evaluation on the Validation Fold**:

   - Once the tree is constructed, it is used to predict the labels for the validation set.
   - Performance metrics (e.g., accuracy, precision, recall) are calculated for the current fold.
   - This process is repeated for all k folds, with a new training and validation split for each iteration

4. **Aggregating Results**:

   - After all k folds are processed, the performance metrics from each fold are averaged to estimate the **DecisionTreeClassifier's** effectiveness.

Cross-Validation Performance on the Training Dataset

Decision Tree (dtree): At **0.739**, it has the lowest training performance, possibly due to its tendency to overfit and lack of ensemble mechanisms.

Validation Performance

Decision Tree (dtree): Scores **0.666**, showing the weakest generalization ability among the models.

Model Building - Bagging Classifier

Steps with Bagging Classifier and StratifiedKFold:

1. **Data Splitting Using StratifiedKFold:**

- `StratifiedKFold` splits the dataset into `k` folds, maintaining the proportional class distribution in each fold. For each fold:
  - A portion of the data is reserved as the validation set.
  - The remaining folds act as the training set for that iteration.

2. **Bagging Process with Training Fold:**

Even without explicitly setting `n_estimators`, the default behavior of the `BaggingClassifier` will still bootstrap the training data (i.e., sampling with replacement) to train multiple base estimators**. Here's how it works:**

- **Random Sampling**: Bootstrap samples are drawn from the training data using the specified `random_state`. This ensures the process is reproducible across different runs.
- **Base Estimators**: The BaggingClassifier uses the default base estimator which is 10 in this particular case, because we did not define the n_estimator for the model to train on the bootstrap samples. It generates multiple versions of the base model, each trained on a unique subset of the data.
- **Prediction Aggregation**: For classification tasks, the predictions from all base estimators are aggregated using majority voting. For regression tasks, predictions are averaged.

Model Building - Bagging Classifier

Steps with Bagging Classifier and StratifiedKFold:

3. **Prediction Aggregation:**

Once all base estimators are trained, predictions are combined:

- **Classification**: Majority voting is used to determine the final predicted class label.

4. **Validation with Held-Out Fold:**

- The trained Bagging Classifier is applied to the validation set (the held-out fold) to predict and assess performance metrics (e.g., accuracy, precision, etc.).
- This process is repeated for each fold in the StratifiedKFold, ensuring robust evaluation.

Cross-Validation performance on training dataset:

- **Bagging**: Scores **0.775**, showing moderate performance compared to the top algorithms.

Validation Performance:

- **Bagging**: At **0.691**, its performance drops significantly compared to cross-validation, indicating potential overfitting.

Model Building - Bagging - RandomForestClassifier

The **RandomForestClassifier** is an ensemble method built on the **Bagging technique**, where multiple decision trees are trained on different subsets of the data. Here's the process with how it works with StratifiedKFold

1. Splitting Data with StratifiedKFold:

- **StratifiedKFold** divides the dataset into k folds, maintaining the **class distribution** within each fold.
- For each fold:
  - One subset is held out as the validation set.
  - The remaining subsets are used as the training set.

2. Training RandomForest on Bootstrapped Training Data:

- For the current fold, RandomForestClassifier trains its ensemble of decision trees using the bootstrapped samples created from the training set.
- Each decision tree in the forest is independently trained on a unique subset, and features are randomly selected during splits.

3. Predicting on Validation Set:

- After the forest is trained on the training data, it predicts the target labels for the validation set (the held-out fold).
- The aggregated predictions from all trees (majority voting) provide the final output for the validation set.

Model Building - Bagging- RandomForestClassifier

4. Cross-Validation Process:

- The model is trained and evaluated on all k folds, and performance metrics (e.g., accuracy, precision, recall) are calculated for each fold.
- The average performance across folds gives a robust estimate of the model's generalization ability.

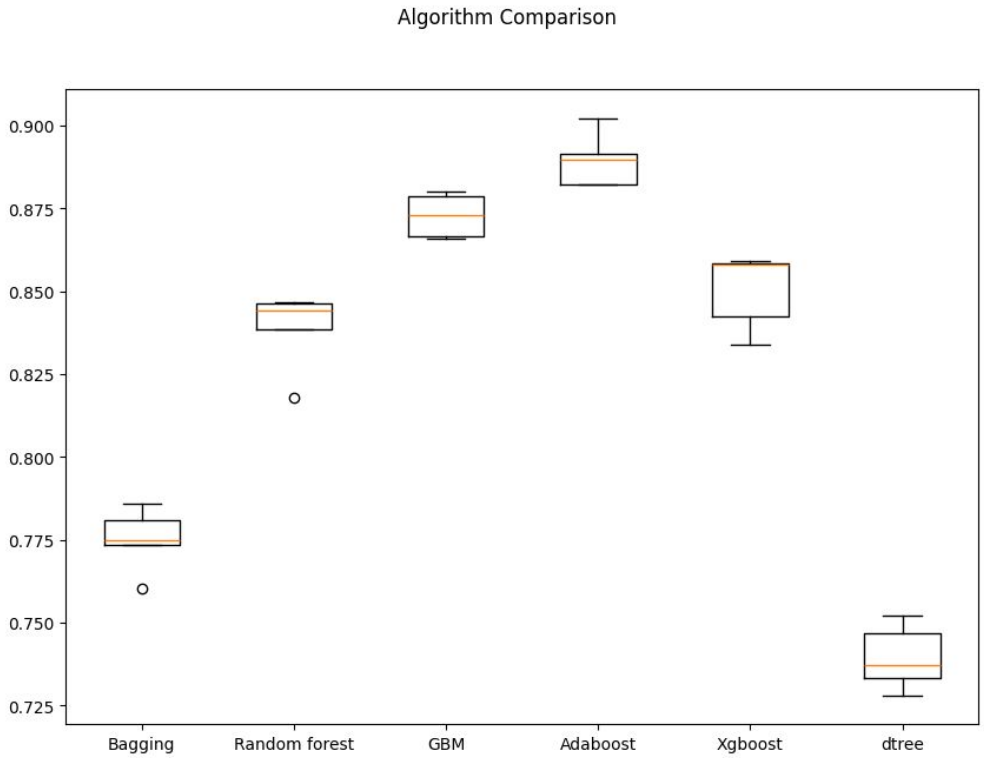
Cross-Validation performance on training dataset:

- Random Forest: Scores 0.838, highlighting its reliability in ensemble learning.

Validation Performance:

- Random Forest: Scores 0.718, exhibiting reliable predictions on unseen data.

Model Building - Algorithm Comparison on Training Set



Cross Validation Performance

Adaboost: With the highest score of 0.889 in cross-validation, Adaboost demonstrates exceptional performance in fitting the training data, suggesting its robustness in capturing patterns.

GBM (Gradient Boosting Machine): The second-best performer at 0.873, indicating a strong ability to model complex relationships.

Xgboost: Achieves 0.850, slightly below GBM but still indicative of high performance.

Random Forest: Scores 0.838, highlighting its reliability in ensemble learning.

Bagging: Scores 0.775, showing moderate performance compared to the top algorithms.

Decision Tree (dtree): At 0.739, it has the lowest training performance, possibly due to its tendency to overfit and lack of ensemble mechanisms

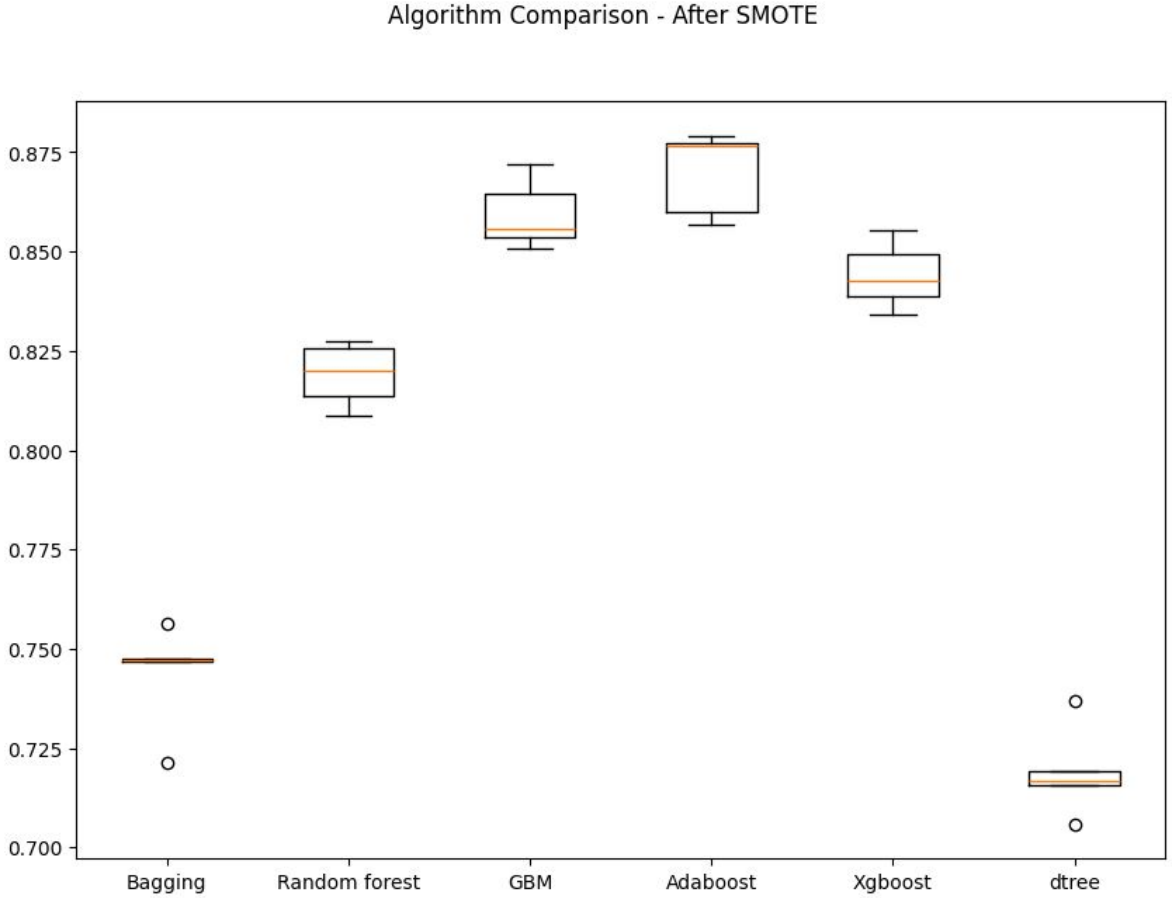
Key Takeaways: Best Performer: GBM stands out as the most balanced algorithm, excelling in both training and validation performance, making it ideal for tasks requiring strong generalization.

Model Improvement - Bagging

- Synthetic Minority OverSampling Technique (**SMOTE**) was **applied** to the training dataset to address bias and imbalance, resulting in improved model performance on the validation data.
- **Best Improvement: Adaboost** shows the highest **improvement (+0.1394)** after SMOTE is applied,

Algorithm	Before SMOTE	After SMOTE	Difference (Increase)
Bagging	0.6911	0.7486	+0.0575
Random Forest	0.7184	0.8187	+0.1003
GBM	0.7428	0.8529	+0.1101
Adaboost	0.7337	0.8731	+0.1394
XGBoost	0.7287	0.8411	+0.1124
Decision Tree	0.6659	0.7249	+0.0590

Model Improvement - Bagging - Algorithm Performance



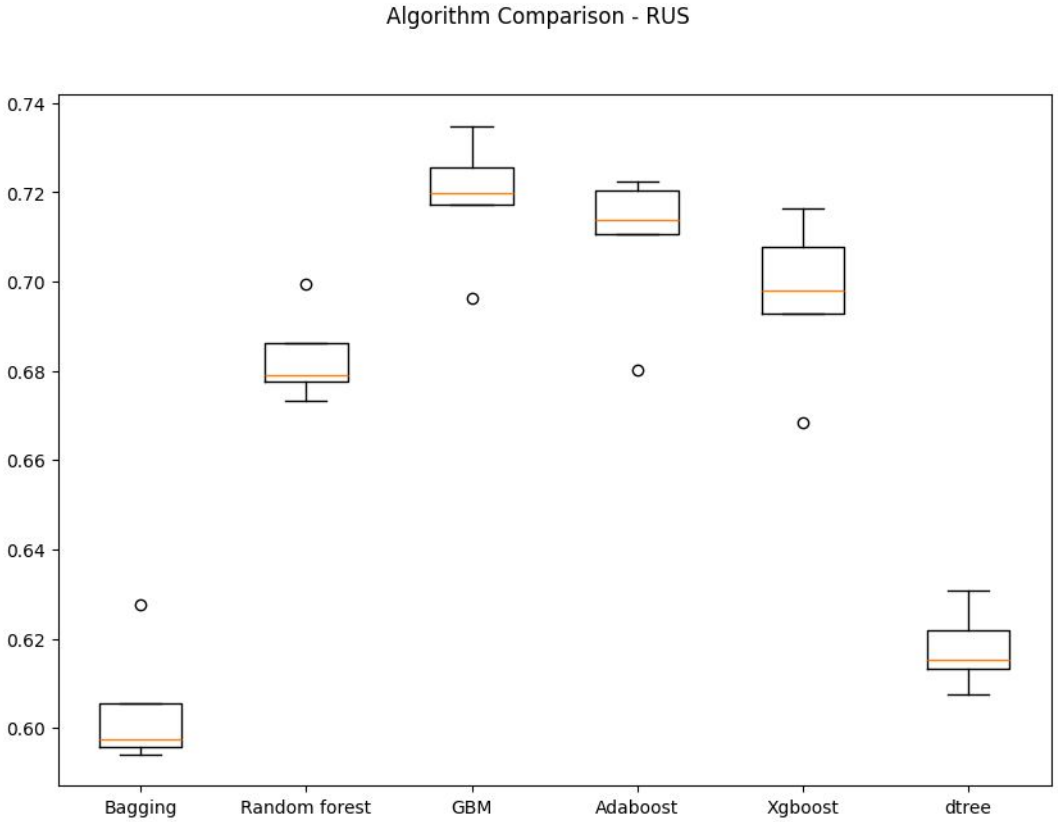
- **All models benefit from SMOTE**, as it addresses class imbalance by generating synthetic samples for the minority class, thereby improving the representation of all classes during training.
- The **improvement** is particularly pronounced for ensemble models like **Adaboost, GBM**, and **XGBoost**, which gain over **10% increase** in **performance.**

Model Improvement Bagging - RandomUnderSampling

- **Bagging benefits** the most from **RandomUnderSampling**, showing better validation performance due to reduced class imbalance.

Algorithm	Cross-Validation (Training)	Validation	Difference
Bagging	0.6041	0.6395	**+0.0354**
Random Forest	0.6831	0.6758	**-0.0073**
GBM	0.7187	0.7026	**-0.0161**
Adaboost	0.7094	0.6984	**-0.0110**
XGBoost	0.6966	0.6824	**-0.0142**
Decision Tree	0.6178	0.6162	**-0.0016**

Model Improvement Bagging - RandomUnderSampling



- **Slight Overfitting**: Ensemble methods like **Random Forest**, **GBM**, **Adaboost,** and **XGBoost** exhibit moderate drops in validation performance, suggesting sensitivity to undersampling while still maintaining stability
- **Consistent** but Limited: **Decision Tree** remains consistent but performs relatively poorly compared to more complex models

Model Improvement - Boosting - Hyperparameter Tuning

- The **AdaBoostClassifier's performance** after hyperparameter tuning demonstrates **noteworthy differences** between the **training set (**oversampled with SMOTE) and the validation set. Here's a detailed comparison:

Metric	Training Set	Validation Set	Difference
Accuracy	0.650718	0.70548	**+0.054762**
Recall	0.931923	0.930344	**-0.001579**
Precision	0.596465	0.714716	**+0.118251**
F1-Score	0.72738	0.808398	**+0.081018**

Model Improvement- AdaBoostClassifier Hyperparameter

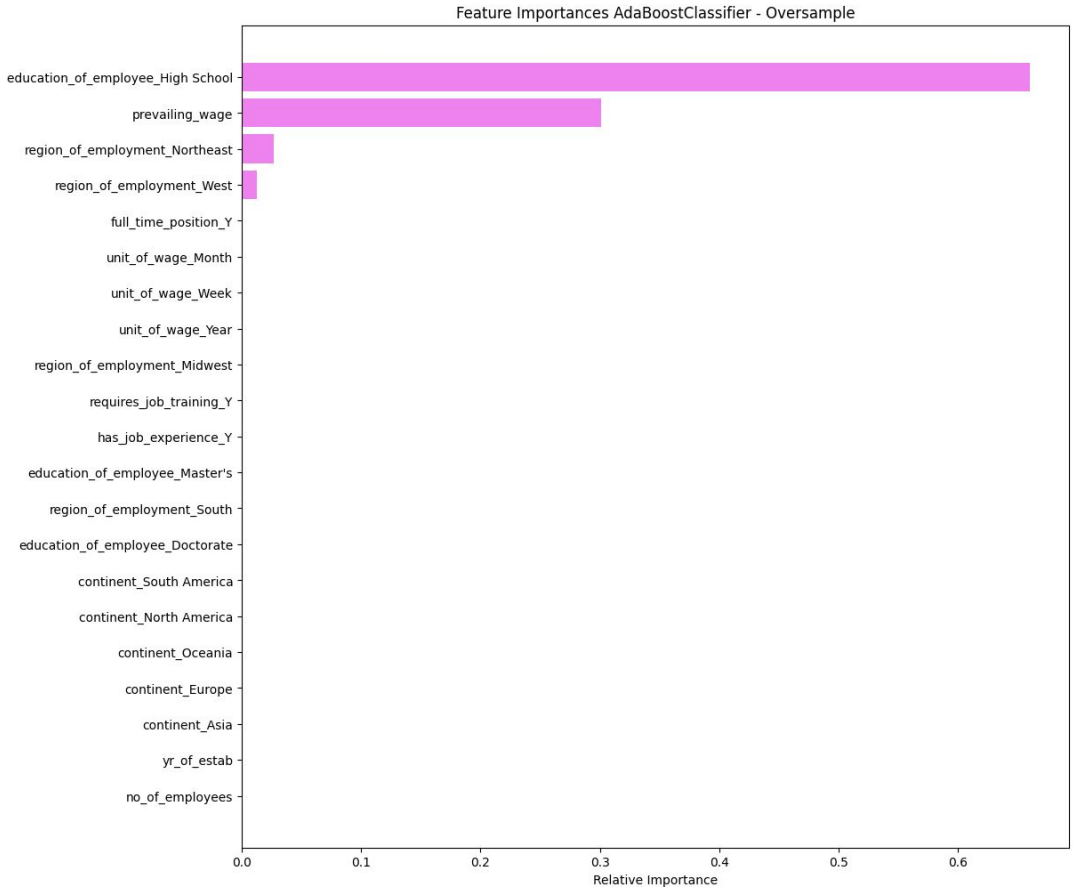
- **Accuracy validation** shows higher accuracy **(+5.48%)** compared to the training set, indicating better generalization and a well-tuned model that performs effectively on unseen data.

- **Recall remains high** and **consistent** between the **training and validation sets**, with a minor drop (-0.16%) in the validation set. This suggests the model maintains a strong ability to identify the positive class.

- **Precision increases** significantly on the validation set **(+11.83%)** compared to the training set, highlighting the model's ability to reduce false positives when applied to unseen data. This improvement is a direct benefit of hyperparameter tuning and the balanced dataset.

- **F1-Score a balance** between precision and recall, improves on the validation set by **+8.10%,** reflecting the model's enhanced performance at maintaining both sensitivity and specificity.

Model Improvement - AdaBoostClassifier Important Features



Education of Employee High School feature ranks as the most critical. It indicates that applicants with high school education have unique patterns impacting visa outcomes.

Prevailing Wage emerges as a key factor, underscoring its role in compliance with U.S. labor standards under the INA. Higher wage offerings might positively influence case outcomes.

Region of Employment Northeast is highly significant, suggesting potential regional dynamics such as economic factors or job demand influencing decisions.

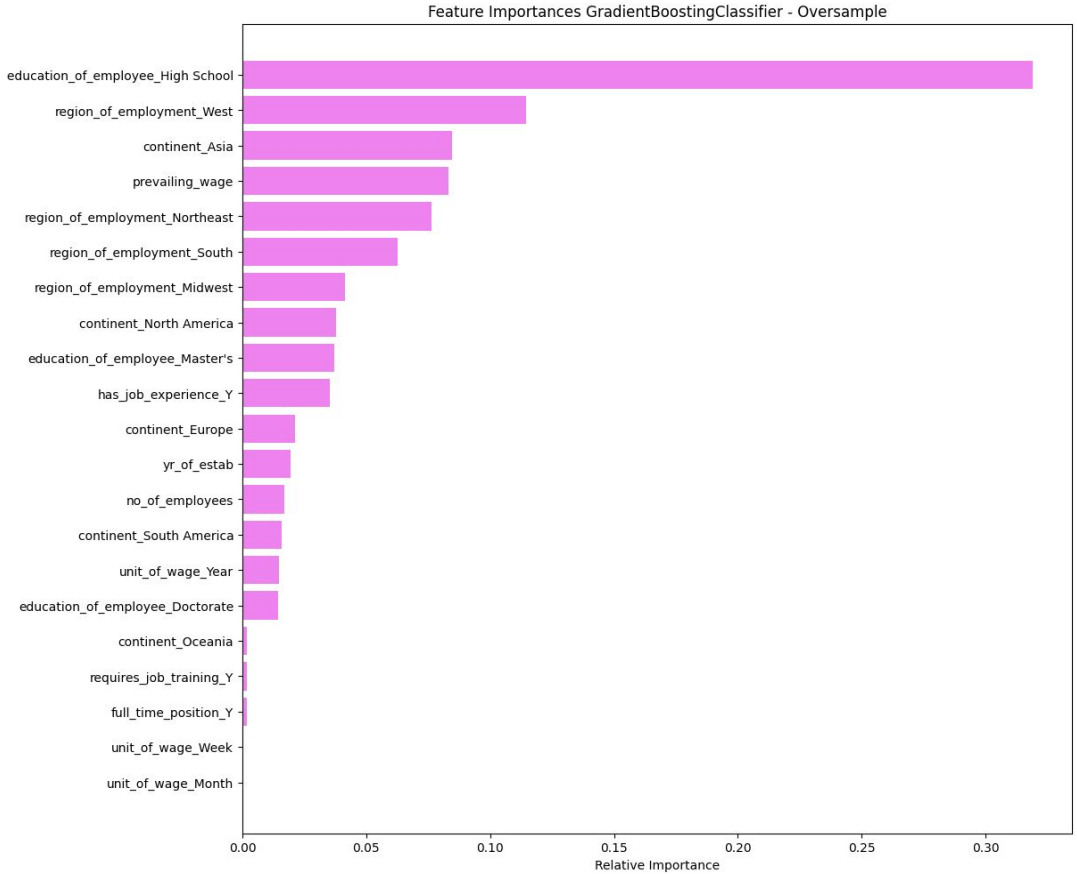
Similar to the Northeast, the **West region** shows high importance, likely tied to geographic-specific conditions affecting visa approvals.

Model Improvement GradientBoostClassifier

- The **GradientBoostClassifier's** performance on the validation set shows moderate drops in accuracy and recall compared to the training data.
- **Accuracy:** The training accuracy is 0.808151 compared to 0.734700 on the validation set—a decline of approximately 7.35 percentage points
- **Recall:** A minor decrease in recall (from 0.868967 to 0.849369, a drop of about 1.96 percentage points) indicates that the model retains most of its ability to capture true positives when applied to unseen data.
- **Precision:** Precision remains virtually unchanged with a negligible increase of about 0.024 percentage points. **F1-Score:** The overall F1-score decreases just slightly (from 0.819149 to 0.810468) by roughly 0.87 percentage points

Metric	Training	Validation	Difference (Val - Tr)
Accuracy	0.808151	0.734700	-0.073451
Recall	0.868967	0.849369	-0.019598
Precision	0.774734	0.774975	+0.000241
F1-Score	0.819149	0.810468	-0.008681

GradientBoostClassifier Feature Importance



Dominant Features is **Education of Employee High School** with a relative importance of approximately 0.30. This strong influence suggests that the educational background, particularly a high school education, is highly predictive in the model's decision-making process.

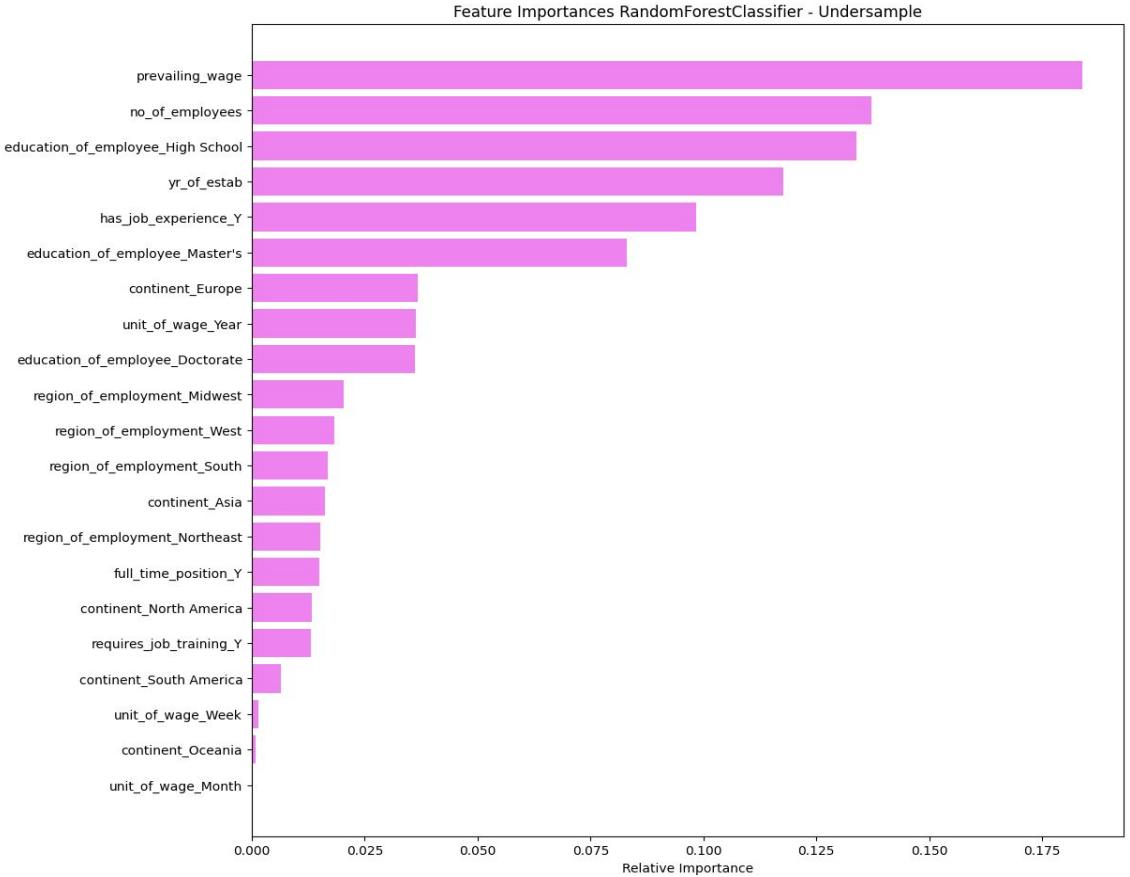
Secondary Influences following the top feature, **Region of Employment West** (≈0.15) and **Continent Asia** (≈0.10) are significant, indicating that geographic characteristics play an important role

Model Improvement - Bagging Hyper Parameter Tuning

- Performance metrics of the **RandomForestClassifier** on the undersampled dataset after hyperparameter tuning, along with the differences between the training and validation sets.
- **Accuracy & Recall: Both metrics decrease by approximately 9.7% and 9.2%**, respectively, on the validation set compared to the training set—an indication of overfitting or a generalization gap.
- **Precision** has a modest improvement **(+2.1% increase)** on the validation set. The overall **F1-Score drops** by about **3.9%** reflecting the **combined impact** of decreased recall and the slight gain in precision.

Metric	Training	Validation	Difference (Validation - Training)
Accuracy	0.801536	0.704463	-0.0971
Recall	0.807023	0.715281	-0.0917
Precision	0.798263	0.819247	+0.0210
F1-Score	0.802619	0.763742	-0.0389

RandomForestClassifier Feature Importance



RandomForestClassifier leverages a diverse set of predictors to drive its performance. The ensemble nature of the model, which involves random feature selection and aggregation of multiple decision trees, naturally results in many features contributing to the final prediction.

Model Improvement XGBClassifier HyperParameter Tuning

- The XGBoost classifier was designed with a structured approach to optimize its predictive performance. To ensure consistency in results, a fixed random state was applied, alongside an evaluation metric that measured model effectiveness
- A range of parameter options was considered, including adjustments to the number of estimators, learning rate, gamma value, subsampling ratio, and model depth. These variations were carefully tested to identify the most effective configuration.
- To refine the model selection process, a randomized search technique was utilized, incorporating a ten-fold validation approach. Nine of these folds were allocated for training, while the remaining fold was reserved for validation. This method enabled the identification of the best-performing model and optimal parameter settings. Once these selections were finalized, the model underwent training using an oversampled dataset to enhance its ability to recognize patterns effectively.

Model Improvement- XGBClassifier Hyper Parameter Tuning

Accuracy improved significantly in the validation set (0.66783 vs. 0.5), indicating better generalization compared to the oversampled training set.

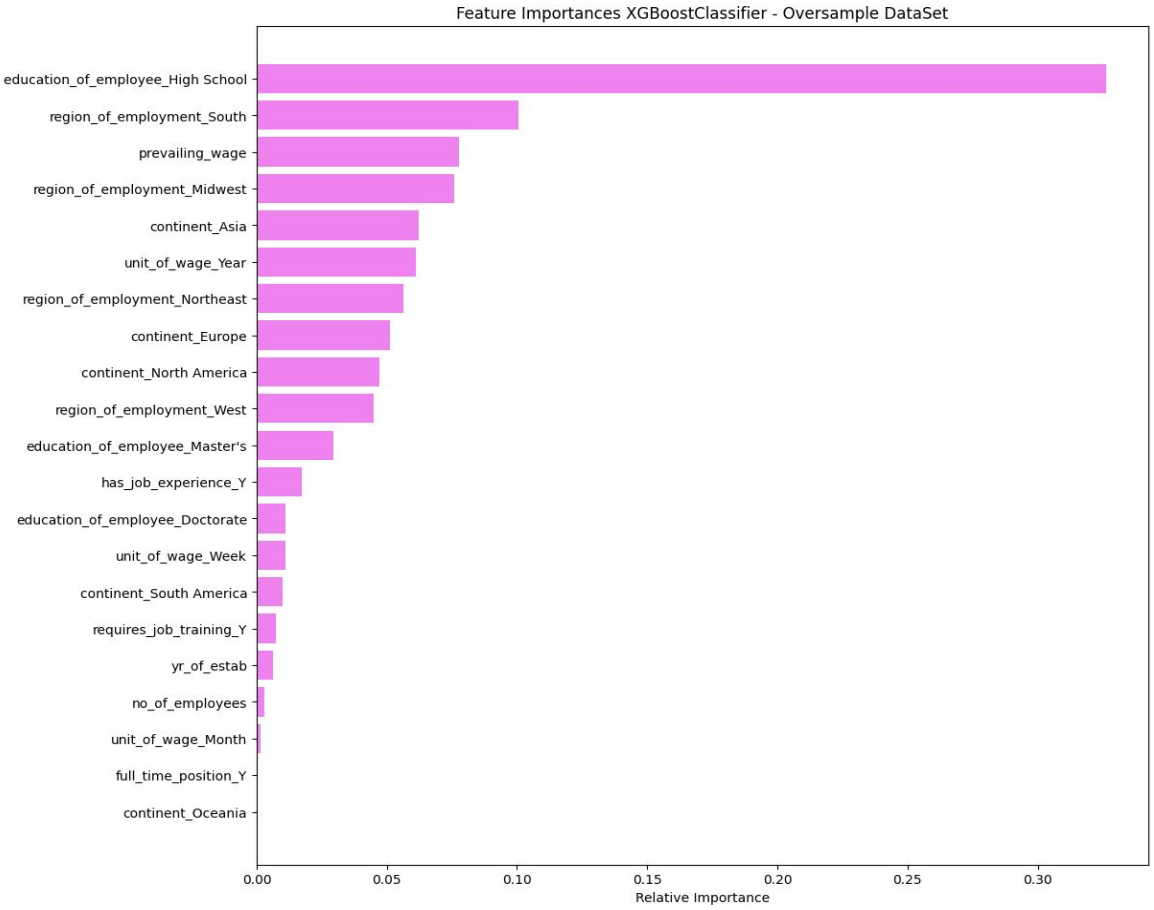
Recall is **consistent** at **1.0** across both datasets, suggesting that the model correctly captures all positive instances.

Precision has a **higher** in **the validation set** (0.66783 vs. 0.5), meaning fewer false positives, implying that the oversampling might have introduced noise affecting precision during training.

F1-Score increased in the validation set (0.80084 vs. 0.66667), confirming **better balance** between precision and recall.

Dataset	Accuracy	Recall	Precision
Training (Oversampled)	0.50000	1.00000	0.50000
Validation (Non-Oversampled)	0.66783	1.00000	0.66783

XGBClassifier Feature Importances



Key Insights

- **Education and geography** are dominant predictors, underscoring the impact of regional employment trends on classification results.
- **Financial aspects**, particularly prevailing wages and wage measurement units, directly influence predictions.
- **Continental employment trends** suggest broader geographic patterns beyond individual regions.

Model Performance Summary Training Data

Gradient Boosting (Oversampled) Analysis: This model demonstrates the **highest overall performance** with an accuracy of 80.8% and an F1 score of 81.9%

Recommendation: Given its strong generalization and balanced metrics, **Gradient Boosting** is the ideal candidate for deployment

Model	Accuracy	Recall	Precision	F1
Gradient Boosting (Oversampled Tuning)	**0.808151**	**0.868967**	**0.774734**	**0.819149**
XGBoost (Oversampled Tuning)	0.500000	1.000000	0.500000	0.666667
AdaBoost (Oversampled Tuning)	0.650718	0.931923	0.596465	0.727380
Random Forest (Undersampled Tuning)	**0.791744**	**0.803478**	**0.785054**	**0.794100**

Model Performance Summary Training Data

XGBoost (Oversampled):

- **Analysis:** While XGBoost achieved a perfect recall of 100%, its precision falls to 50%, resulting in a lower F1 score of 66.7% and an accuracy of only 50%. The high recall indicates that it is sensitive in detecting positive cases, but the low precision implies many false positives.
- **Recommendation:** XGBoost might be considered in applications where missing a true positive is extremely costly, but a review mechanism should be in place to filter out false positives. Tuning further—possibly focusing on increasing precision—may be necessary if this model is to be used in a balanced decision-making framework for **OFLC**

AdaBoost (Oversampled) and Random Forest (Undersampled):

- **Analysis:** AdaBoost shows a high recall (93.2%) but lower precision (59.6%) and overall F1 of 72.7%, indicating it captures many positives at the cost of increased false positives. **Random Forest**, **tuned on undersampled data**, demonstrates a balanced profile (accuracy 79.2%, recall 80.3%, precision 78.5%, F1 79.4%) but does not outperform the Gradient Boosting model.
- **Recommendation:** Both AdaBoost and Random Forest can be useful in scenarios that require diversified decision boundaries. Random Forest's balance suggests it is more reliable for general applications, while AdaBoost may serve well in specialized cases where maximizing recall is prioritized. Consider ensembling or further parameter tuning to optimize these models if necessary.

Model Performance Summary Validation Data

Overall Best Performer – Gradient Boosting tuned on oversampled data leads with the highest accuracy (73.47%) and the top F1 score (0.8105). Its balanced metrics—high recall (84.94%) coupled with solid precision (77.50%)—make it the top candidate for validation

Considerations for XGBoost and AdaBoost: XGBoost demonstrates perfect recall (1.0000), meaning it identifies all positive instances; however, this comes at the expense of precision (66.78%), leading to a slightly lower F1 score (0.8008) compared to Gradient Boosting

Model	Accuracy	Recall	Precision	F1
Gradient Boosting (Oversampled)	**0.734700**	**0.849369**	**0.774975**	**0.810468**
XGBoost (Oversampled)	0.667830	1.000000	0.667830	0.800837
AdaBoost (Oversampled)	0.705480	0.930344	0.714716	0.808398
Random Forest (Undersampled)	0.706643	0.720505	0.818497	0.766381

Model Performance Summary Validation Data

- **Random Forest Considerations:** Random Forest tuned on undersampled data achieves a solid precision (81.85%) but lags in recall (72.05%) and F1 (0.7664). This indicates it might miss a significant number of true positives, making it less ideal in applications where identifying all positive cases is crucial. Its performance suggests it is more conservative, which might be acceptable in contexts where false positives are particularly undesirable, but overall it does not match the balance seen in Gradient Boosting.
- **Overall, for validation performance,** Gradient Boosting emerges as the best model due to its strong accuracy, balanced precision and recall, and the highest F1 score, making it highly suitable for applications that require stable and reliable performance on unseen data.

Model Performance Summary Stacking

*A **StackingClassifier** was developed to **enhance overall predictive performance** by leveraging multiple base learners and a robust meta-model. The base models incorporated into the ensemble include **Logistic Regression**, **Support Vector Classifier (SVC)**, **XGBoostClassifier**, and a **Multi-Layer Perceptron Classifier (MLP)**. The ensemble's **final decision is driven by a RandomForestClassifier** serving as the meta-model. **Hyperparameter tuning** was performed using **RandomizedSearchCV** to systematically identify the optimal parameter combination, ensuring the **StackingClassifier** achieved the best possible balance between bias and variance.*

Trade-offs The approach achieves high accuracy; however, it requires significant computational resources to perform inference effectively.

Model	Accuracy
StackingClassifier	**0.8367**
Logistic Regression	0.7645
SVC	0.8931
XGB	0.8183
MLP	0.5372

APPENDIX

Data Background and Contents

- **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

Data Background and Contents

- What are the datatypes of the different columns in the dataset

case_id	Object
continent	Object
education_of_employee	Object
has_job_experience	Object
requires_job_training	Object
no_of_employees	Int64
yr_of_estab	Int64
region_of_employment	Object

Data Background and Contents

- What are the datatypes of the different columns in the dataset

prevailing_wage	float64
unit_of_wage	object
full_time_position	object
case_status	object

Great Learning

Happy Learning !