Enhancing Visa Approval Predictions: Insights and Recommendations

Ensemble Techniques - Pam Lozano

January 19, 2024

Contents / Agenda

- Data Information
- Business Problem Overview and Solution Approach
- Data Background and Contents
- EDA Results
- Data Preprocessing
- Model Evaluation
- Decision Tree - Model Building and Hyperparameter Tuning
- Bagging - Model Building and Hyperparameter Tuning
- Boosting - Model Building and Hyperparameter Tuning
- Stacking Classifier
- Model Performance Comparison and Final Model Selection
- Executive Summary

Objective

Develop a machine learning classification model to assist EasyVisa in shortlisting visa candidates with higher approval chances, recommending profiles based on influential factors, and streamlining the visa approval process.

Key Focus Areas:

1. Conduct exploratory data analysis (EDA) to understand feature distributions and identify patterns impacting visa approval.
2. Preprocess the data by handling missing values, encoding categorical variables, and scaling numerical features for effective model training.
3. Build, fine-tune, and compare multiple classification models, including Decision Tree, Random Forest, and boosting algorithms, to select the best-performing model for visa approval prediction.

Data Information

The data contains the different attributes of employee and the employer.

.

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 have any job experience? Y=Yes; N=No
requires_job_training	Does the employee require any job experience? 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
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 booking was canceled or not.

Business Problem Overview

- **High Demand for Talent:** The U.S. faces a competitive challenge in meeting high demand for skilled human resources, both domestically and internationally.
- **Compliance Complexity:** Businesses, under the Immigration and Nationality Act (INA), grapple with compliance complexities when hiring foreign workers to address workforce shortages.
- **OFLC's Approval Challenge:** The Office of Foreign Labor Certification (OFLC) deals with a laborious visa approval process, reviewing a substantial number of applications due to an annual increase in applicants.
- **Data-Driven Solution Need:** The surge in applicants prompts the need for a data-driven solution, leading OFLC to enlist EasyVisa for the development of a machine learning model to streamline visa approvals.
- **OFLC's Efficiency Goal:** OFLC aims to enhance the efficiency and accuracy of visa approvals by utilizing machine learning to recommend candidate profiles based on influential factors.
- **EasyVisa's Role:** Analyze data and build a classification model that not only facilitates visa approvals but also aids in strategically selecting candidates with optimal approval chances.

The ultimate aim is to leverage machine learning and data-driven insights to revolutionize the visa approval process, enhancing efficiency, accuracy, and strategic decision-making for EasyVisa and the Office of Foreign Labor Certification (OFLC), ultimately contributing to the seamless integration of skilled human resources into the U.S. workforce.

Solution Approach

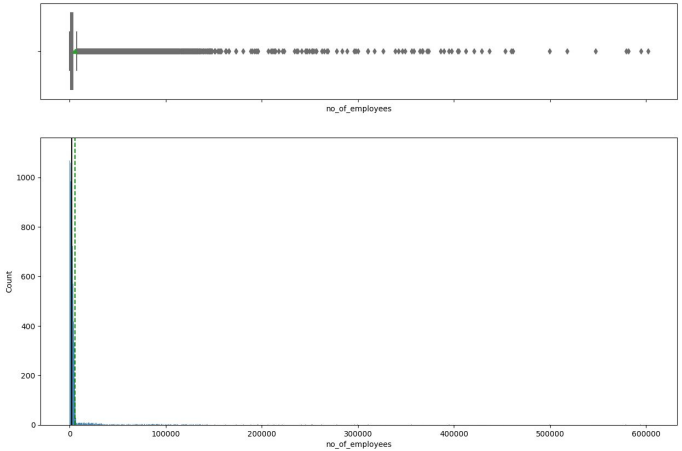
- Conduct thorough exploratory data analysis (EDA) to understand feature distributions and identify patterns influencing visa approval.
- Preprocess the dataset by handling missing values, encoding categorical variables, and scaling numerical features for effective model training.
- Implement multiple classification models, including Decision Tree, Random Forest, and boosting algorithms (e.g., XGBoost, AdaBoost), fine-tuning their hyperparameters for optimal performance.
- Explore the development of a stacking classifier, combining the strengths of multiple models to enhance overall predictive accuracy.
- Compare and evaluate the performance of different models using appropriate metrics such as accuracy, precision, recall, F1 score, and ROC-AUC.
- Integrate the selected model into EasyVisa's processes, providing a data-driven solution for shortlisting visa candidates with higher approval chances and contributing to a streamlined visa approval process for the Office of Foreign Labor Certification (OFLC).

Data Background and Contents

- Our dataset comprises 25,480 rows and 12 columns, providing a comprehensive foundation for our analysis.
- There are a mix of data types in our dataset: integers, objects, and floats, contributing to the diversity of information available for our analysis.
- There are no missing values in our dataset.
- There are no duplicate values in our dataset.
- The average number of employees is 5,667 which is a central point of our dataset.
- The median number of employees is 2,109, and the maximum number of employees is 602,069.

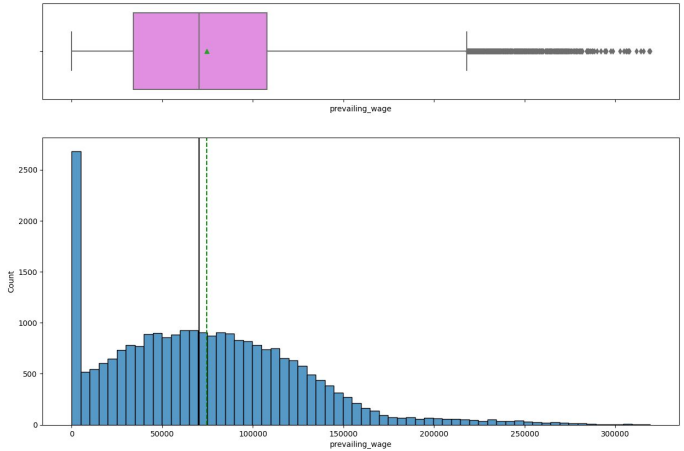
Exploratory Data Analysis

Univariate Analysis - Number of Employees



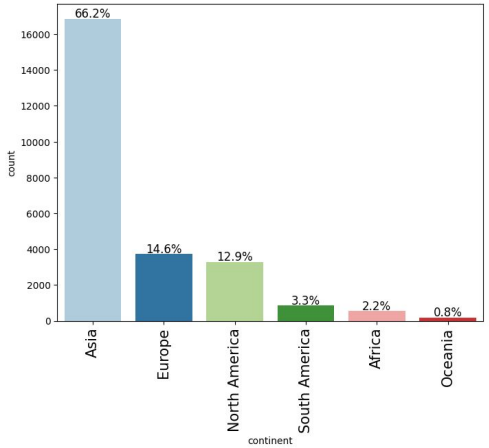
- Outliers are evident throughout the distribution, suggesting potential extreme values.

Univariate Analysis - Prevailing Wage



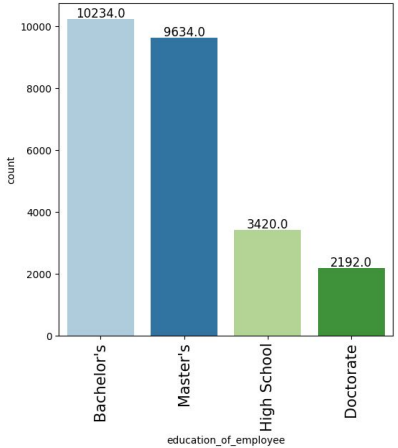
- Outliers are evident on the right of the distribution, suggesting potential extreme values.
- The distribution is right-skewed, with a median of approximately 70,000 in prevailing wage.

Univariate Analysis - Continent



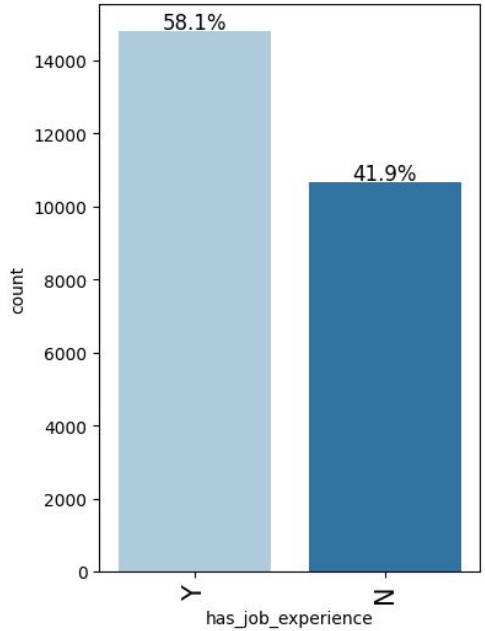
- The majority of employees are from Asia at 66.2%
- The continent with the least amount of employees come from Oceania at .8%.

Univariate Analysis - Education of Employee



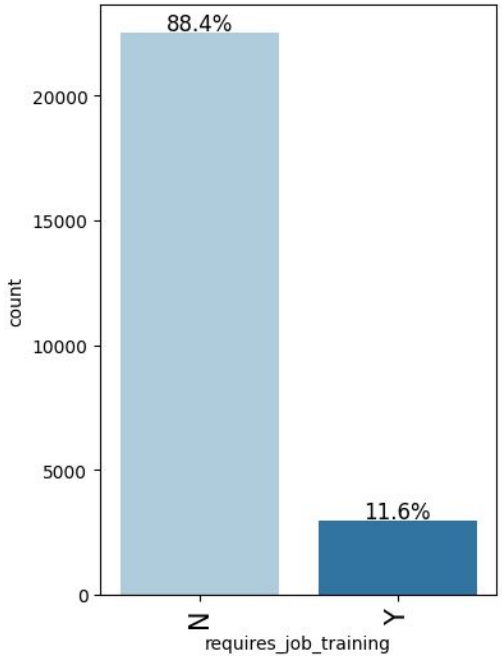
- The majority of employees have at least a bachelor’s degree.
- 2,192 employees have a doctorate.

Univariate Analysis - Job Experience



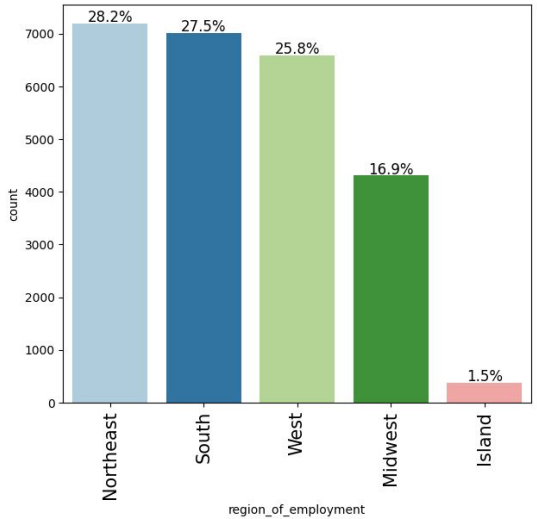
- The majority of employees already have job experience at 58.1%

Univariate Analysis - Job Training



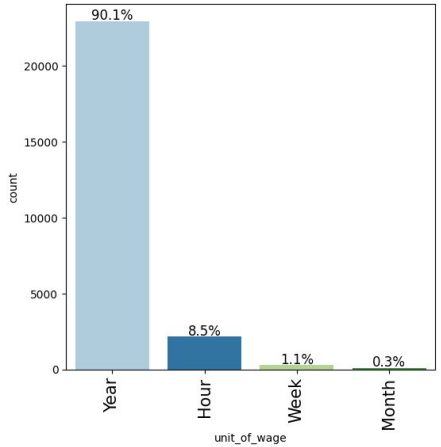
- The majority of employees at 88.4% do not require any job training.

Univariate Analysis - Region of Employment



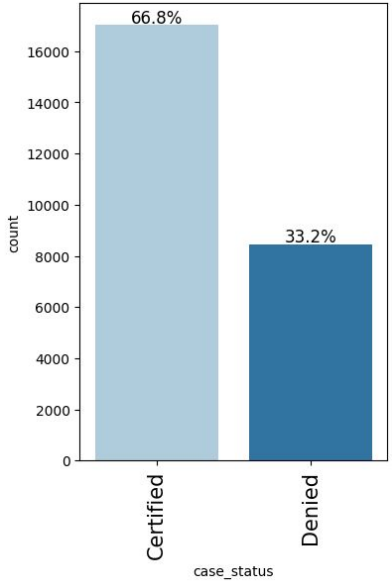
- The most common region of employment comes from the Northeast at 28.2%.
- The least common region of employment comes from the Island of 1.5%.

Univariate Analysis - Unit of Wage



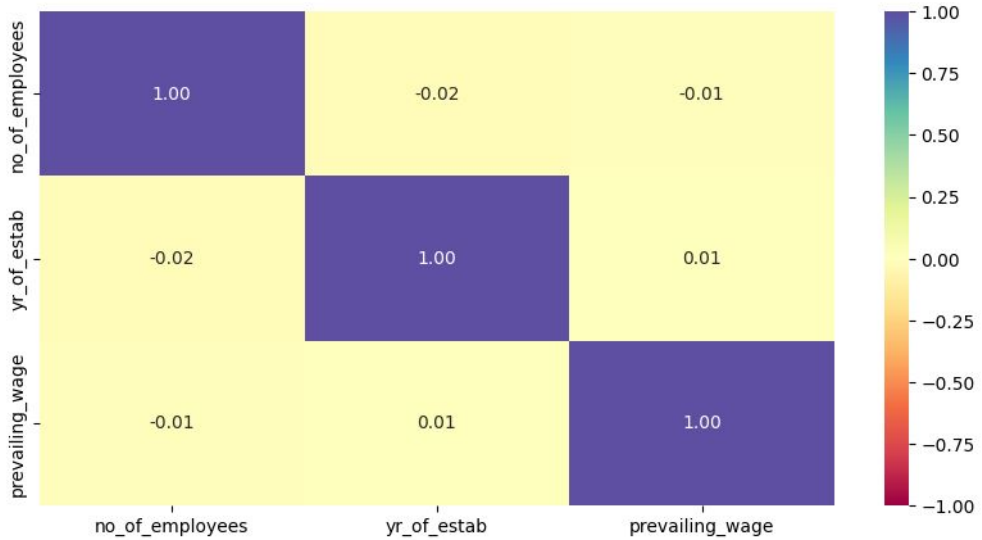
- The majority of employees are yearly salaried employees with 90.1%.

Univariate Analysis - Case Status



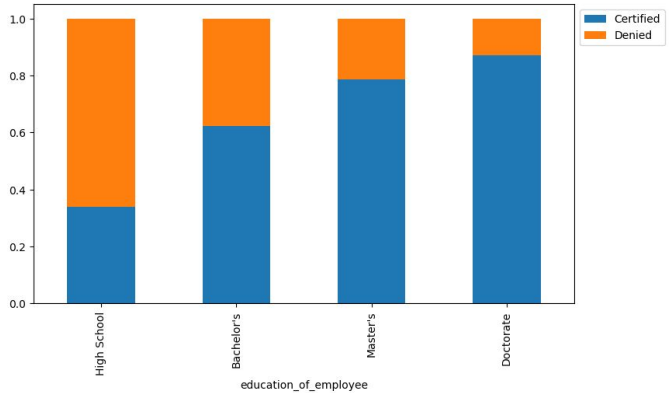
- Most employees are certified at 66.8%.

Bivariate Analysis



- There isn’t a strong correlation between number of employees, year established, and prevailing wage.
- A negative correlation exists between number of employees and prevailing wage, and year established.

Bivariate Analysis - Education vs Case Status

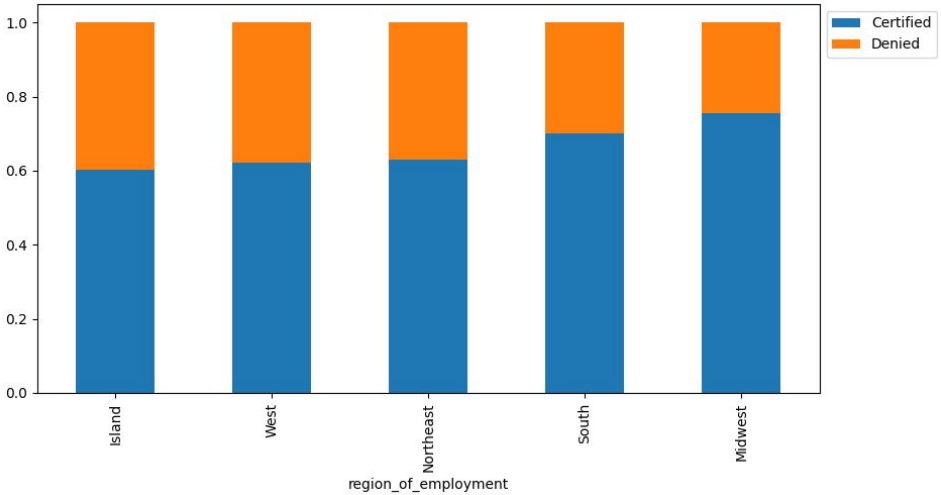


- The higher level of education, the more employees that get certified.

Bivariate Analysis - Education vs Region of Employment

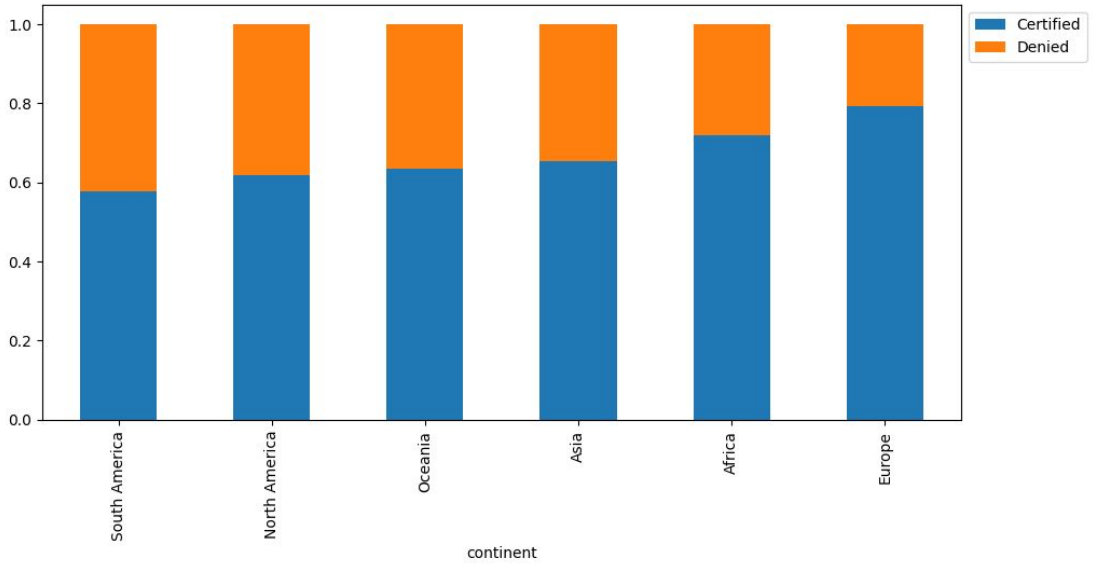
- Employees from the South Region contain more employees with a high school and a bachelor's degree.
- The Northeast Region has more of the employees with a higher level of education ranging from Master's and Doctorate.

Bivariate Analysis - Region of Employment vs Case Status



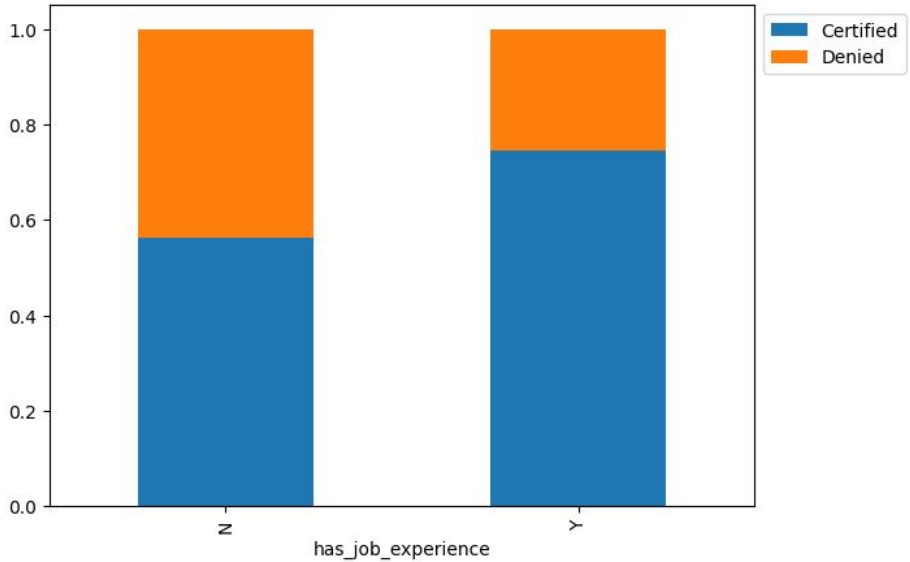
- Employees that come from the Midwest have more employees that get certified.
- The area that has the least amount of employees to get certified is from the Island Region.

Bivariate Analysis - Continent vs Case Status



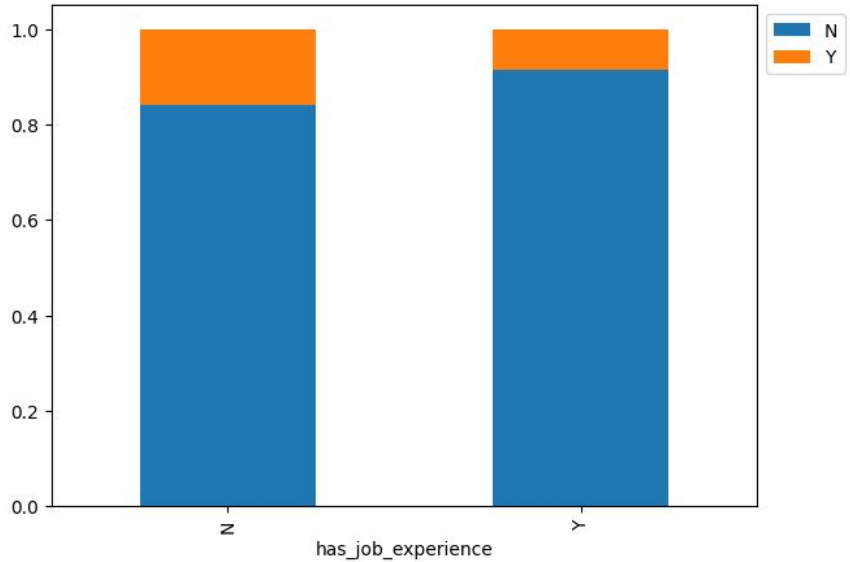
- The employees that are more certified come from Europe.
- The continent that has the least amount of certified employees are from South America.

Bivariate Analysis - Job Experience vs Case Status



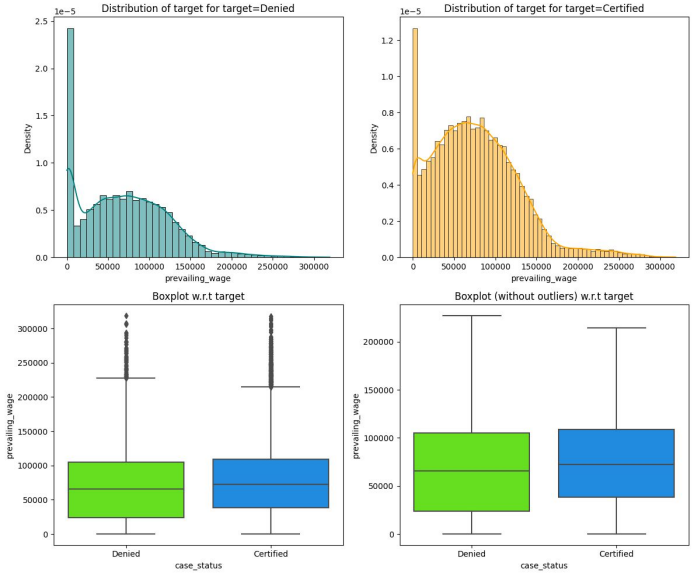
- Employees that have prior job experience are more likely to get certified.

Bivariate Analysis - Job Experience vs Requires Job Training



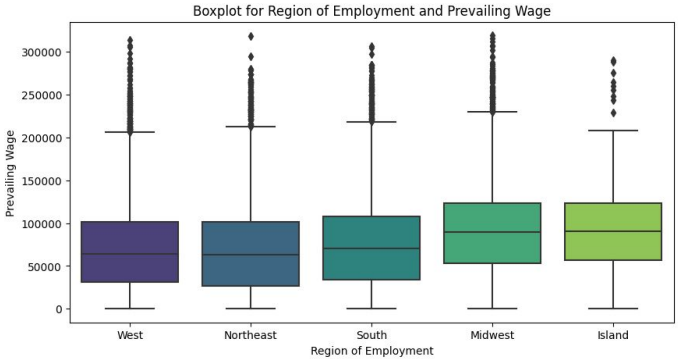
- Employees with more job experience requires job training.

Bivariate Analysis - Prevailing Wage vs Case Status



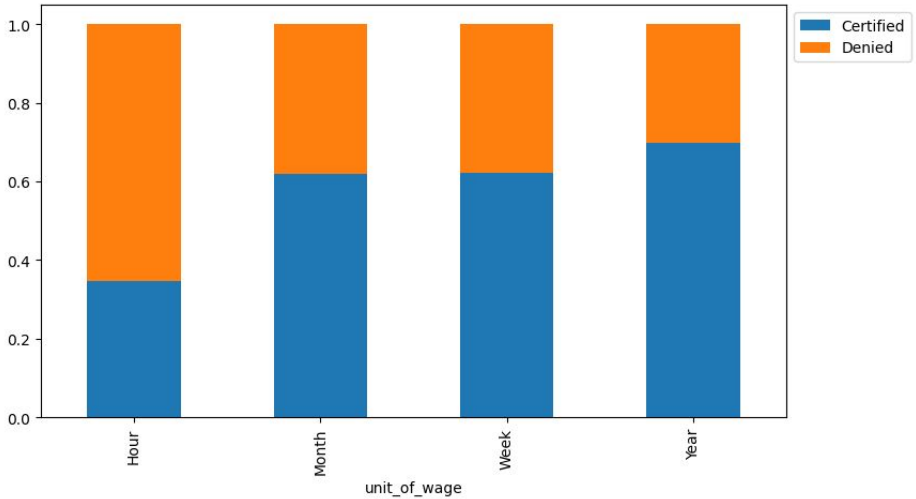
- A positive correlation exists between prevailing wage and case status.

Bivariate Analysis - Region of Employment vs Prevailing Wage



- The Midwest and Island region show a higher prevailing wage than other regions.

Bivariate Analysis - Unit of Wage vs Case Status



- Salaried employees have more likelihood to be certified while hourly employees will have more tendency to be denied.

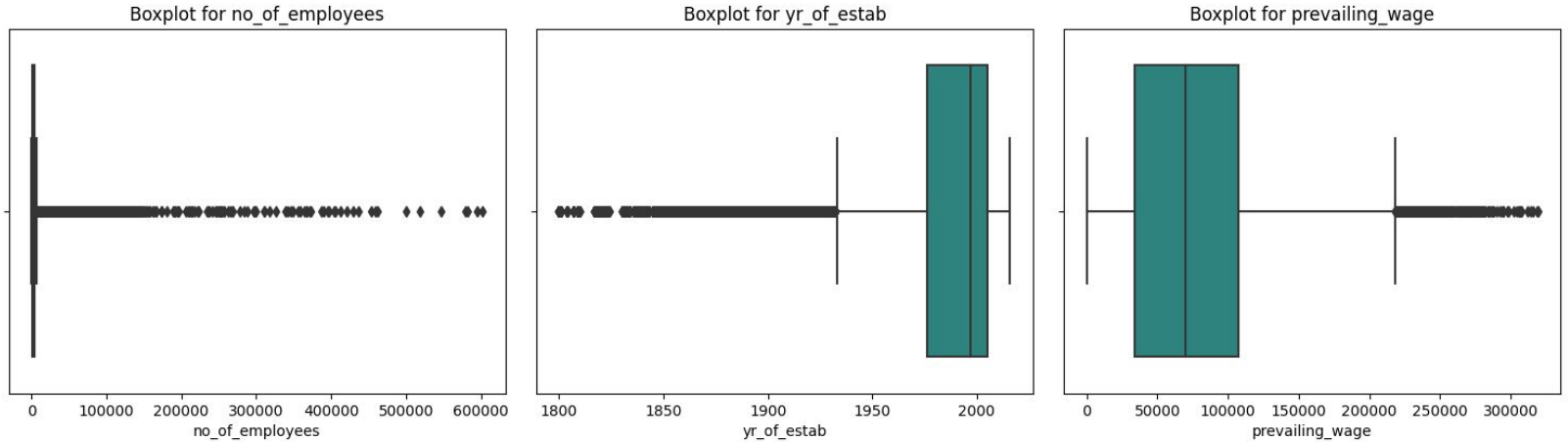
Data Preprocessing

Data Preprocessing

We want to predict which visa will be certified. To achieve this:

- We'll encode categorical features.
- The data will be split into train and test sets for model evaluation.
- Various models will be built using the train data, and we'll assess the best performing model.

Data Preprocessing



- Several outliers have been identified in the data, but no treatment will be applied.

Model Evaluation Criteron

Model Evaluation Criteron

First, we will create functions to calculate different metrics and a confusion matrix so that we don't have to use the same code repeatedly for each model.

- The model_performance_classification_sklearn function will be used to check the model performance of models.
- The confusion_matrix_sklearn function will be used to plot the confusion matrix.

Decision Tree - Model Building and Hyperparameter Tuning

Decision Tree - Pre - Pruning

Training Performance

	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Testing Performance

	Accuracy	Recall	Precision	F1
0	0.660387	0.739275	0.748958	0.744085

- The decision tree model exhibited perfect performance on the training data but demonstrated strong generalization on the test set, achieving a balanced trade-off between accuracy, precision, recall, and F1 score.

Decision Tree - Post Pruning

Training

	Accuracy	Recall	Precision	F1
0	0.712548	0.931923	0.720067	0.812411

Testing

	Accuracy	Recall	Precision	F1
0	0.706567	0.930852	0.715447	0.809058

- Post-pruning, the decision tree model achieves strong and consistent performance on both the training and testing sets, showcasing that it is well-balanced.

Bagging - Model Building and Hyperparameter Tuning

Bagging Classifier - Pre - Pruning

Training Performance

	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Testing Performance

	Accuracy	Recall	Precision	F1
0	0.723967	0.834868	0.770845	0.80158

- Pre-pruning the training set achieved a flawless performance, and demonstrated trong demonstrated a strong generalization on the testing set, showcasing robust predictive capabilities.

Bagging Classifier - Post Pruning

Training

	Accuracy	Recall	Precision	F1
0	0.995234	0.999496	0.993409	0.996443

Testing

	Accuracy	Recall	Precision	F1
0	0.731293	0.874241	0.75966	0.812933

- The training set shows a potential risk of overfitting while the testing set has a strong performance suggesting a balanced model that avoids significant underfitting.

Random Forest - Model Building and Hyperparameter Tuning

Random Forest - Pre - Pruning

Training Performance

	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Testing Performance

	Accuracy	Recall	Precision	F1
0	0.719911	0.835651	0.766164	0.7994

- The training set achieved perfect performance, and demonstrated strong generalization on the testing set, showcasing its consistent and reliable predictive prowess.

Random Forest- Post Pruning

Training

	Accuracy	Recall	Precision	F1
0	0.77209	0.900865	0.78819	0.840769

Testing

	Accuracy	Recall	Precision	F1
0	0.74359	0.881685	0.768482	0.821201

- The post-pruned random forest model performs well on both the training and testing sets, suggesting a balanced fit without notable overfitting or underfitting concerns.

AdaBoost - Model Building and Hyperparameter Tuning

AdaBoost - Pre - Pruning

Training Performance

	Accuracy	Recall	Precision	F1
0	0.738058	0.886259	0.760937	0.81883

Testing Performance

	Accuracy	Recall	Precision	F1
0	0.734301	0.883252	0.75858	0.816182

- A consistent performance exists between the training and testing sets, indicating a well-balanced model without prominent signs of overfitting or underfitting.

AdaBoost - Post Pruning

Training

	Accuracy	Recall	Precision	F1
0	0.75527	0.88777	0.777418	0.828938

Testing

	Accuracy	Recall	Precision	F1
0	0.742282	0.880705	0.767628	0.820288

- AdaBoost post-pruning demonstrates consistent and balanced performance across the training and testing sets, suggesting effective generalization without clear signs of overfitting or underfitting.

Gradient Boosting - Model Building and Hyperparameter Tuning

Gradient Boosting - Pre-Pruning

Training Performance

	Accuracy	Recall	Precision	F1
0	0.757849	0.883657	0.782095	0.82978

Testing Performance

	Accuracy	Recall	Precision	F1
0	0.745814	0.878355	0.772305	0.821923

- Gradient boosting pre-pruning shows consistent and comparable performance on both training and testing sets, indicating effective model generalization without signs of overfitting or underfitting.

Gradient Boosting - Post Pruning

Training

	Accuracy	Recall	Precision	F1
0	0.753756	0.885671	0.776894	0.827724

Testing

	Accuracy	Recall	Precision	F1
0	0.745029	0.881489	0.770021	0.821993

- Gradient boosting post-pruning exhibits stable and balanced performance across both training and testing sets, reflecting effective generalization without noticeable signs of overfitting or underfitting.

XGBoost - Model Building and Hyperparameter Tuning

XGBoost - Pre - Pruning

Training Performance

	Accuracy	Recall	Precision	F1
0	0.840884	0.930664	0.8464	0.886534

Testing Performance

	Accuracy	Recall	Precision	F1
0	0.726583	0.850735	0.765826	0.80605

- The XGBoost model, pre-pruning, exhibits high accuracy, recall, and precision on the training set, translating into robust predictive capabilities. However, on the testing set, there is a slight decrease in performance, suggesting potential overfitting.

XGBoost - Post Pruning

Training

	Accuracy	Recall	Precision	F1
0	0.758971	0.889532	0.780339	0.831365

Testing

	Accuracy	Recall	Precision	F1
0	0.74516	0.879138	0.771267	0.821677

- The XGBoost model, post-pruning, demonstrates consistent performance on both the training and testing sets, indicating reliable generalization without significant overfitting or underfitting.

Stacking Classifier

Stacking Classifier

Training Performance

	Accuracy	Recall	Precision	F1
0	0.766764	0.873248	0.796982	0.833373

Testing Performance

	Accuracy	Recall	Precision	F1
0	0.748038	0.86288	0.782277	0.820604

- The Stacking model performs well on both training and testing sets, showing consistent and effective predictive capabilities.

Model Performance Comparison and Final Model Selection

Checking Model Performance

- Decision Tree:
  - Achieves perfect performance on the training set but shows a drop in metrics on the testing set, indicating potential overfitting.
- Bagging (Bootstrap Aggregating):
  - Pre-pruning demonstrates exceptional performance on the training set with good generalization to the testing set.
  - Post-pruning maintains high accuracy and recall on both sets, suggesting a well-balanced model.
- Boosting:
  - AdaBoost pre-pruning yields good performance on both sets, avoiding overfitting.
  - Post-pruning maintains high accuracy and recall, demonstrating improved precision.
- Gradient Boosting:
  - Pre-pruning achieves high accuracy and recall on both sets, indicating robust predictive capabilities.
  - Post-pruning maintains strong performance, showcasing consistent generalization.
- XGBoost:
  - Pre-pruning demonstrates good performance on both sets, showcasing balanced metrics.
  - Tuned XGBoost enhances precision on the testing set, indicating improved model sensitivity.
- Stacking:
  - Demonstrates robust performance on both sets, showcasing the potential benefits of combining diverse models.

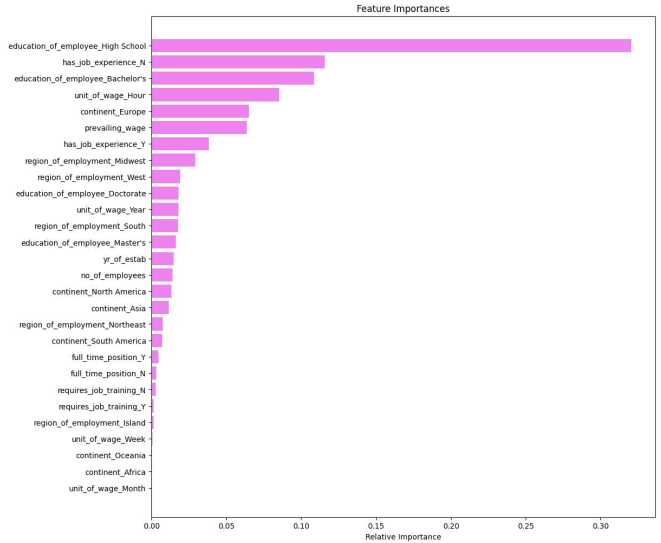
This comparative analysis highlights the strengths and weaknesses of each model, aiding in informed model selection for specific use cases.

Model Performance Summary

- Overfitting Observation:
  - The majority of the models are exhibiting signs of overfitting on the training data, as indicated by a higher F1-score on the training set compared to the testing set.
- Best Test F1-Score:
  - The Bagging Classifier achieves the highest F1-score on the test data, showcasing strong predictive capabilities. However, it's essential to note that it also demonstrates overfitting tendencies on the training data.
- Tuned Random Forest Performance:
  - The Tuned Random Forest model follows closely with the second-highest test F1-score. It shows a more generalized performance, striking a balance between predictive power and avoiding overfitting on the training data.

In summary, while the Bagging Classifier performs well on the test set, the Tuned Random Forest offers a promising alternative with a more balanced trade-off between training and testing performance. Further exploration and fine-tuning could enhance the model's generalization capabilities.

Important Features - Post Pruning



- High School Degree, Job Experience, and Bachelor’s Degree are the most important features for the final model

Executive Summary

Executive Summary - Conclusions

- **Education and Experience Impact:** Higher education and job experience positively influence visa approval, emphasizing the importance of qualified candidates.
- **Training Requirement Consideration:** Applicants requiring job training may face additional scrutiny, and assessing this factor is essential in predicting visa outcomes.
- **Company Characteristics Matter:** Employer attributes, such as company size and establishment year, play a role in visa decisions, indicating the need to consider these factors during the evaluation process.
- **Regional Dynamics Influence:** The region of intended employment is a significant factor affecting visa outcomes, highlighting the importance of understanding regional variations and preferences.
- **Competitive Prevailing Wages:** Offering competitive prevailing wages is crucial, as it positively correlates with higher chances of visa approval and reflects fair compensation practices.
- **Preference for Full-Time Positions:** Full-time positions have a higher likelihood of visa approval compared to part-time positions, suggesting a preference for stability and commitment in employment.

These conclusions provide valuable insights for EasyVisa to enhance its decision-making process and tailor its recommendations to increase the efficiency of visa approvals.

Executive Summary - Recommendations

- **Feature Importance Analysis:**
  - Identify key drivers influencing visa approval.
  - Focus on variables like education, job experience, training, company attributes, and prevailing wage.
- **Model Development and Integration:**
  - Build a robust machine learning model for predicting visa outcomes.
  - Integrate the model into the approval process to automate initial screening.
- **Streamlining and Efficiency:**
  - Streamline the visa approval process by prioritizing high-risk and high-potential applications.
  - Reduce manual workload and enhance efficiency with automated screening.
- **Continuous Improvement and Compliance:**
  - Monitor and update the model regularly to adapt to changes.
  - Ensure compliance with legal and ethical standards in immigration and employment practices.

These recommendations aim to enhance the accuracy, efficiency, and compliance of EasyVisa's services, providing actionable insights for visa approval decisions.