EasyVisa Data Analysis

Python Foundations : PGP-DSBA

April 5th ,2023

Contents / Agenda

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

Business Problem Overview and Solution Approach

Problem Definition: The problem is to analyze a dataset containing information about visa applications processed by the Office of Foreign Labor Certification (OFLC) in the United States and develop a machine learning-based solution to facilitate the process of visa approvals and recommend suitable profiles for visa certification or denial based on significant drivers influencing the case status.

Solution Approach / Methodology: The solution approach involves the following steps: Data Cleaning and Preparation:

- The first step is to clean and prepare the data for analysis. This includes handling missing values, removing duplicates, and transforming the data into a format suitable for machine learning algorithms.
- Exploratory Data Analysis (EDA): The next step is to perform exploratory data analysis to gain insights into the data. This involves visualizing the data and identifying trends, patterns, and outliers that could affect the analysis.

Business Problem Overview and Solution Approach

- Feature Engineering: After EDA, feature engineering is performed to create new features that may help improve the performance of the machine learning model. This includes transforming categorical variables into numerical variables, creating interaction terms, and scaling the data.
- Model Development: Once the data is prepared and features engineered, several machine learning algorithms such as logistic regression, decision trees, and random forests are trained and tested to develop the classification model. The model is evaluated using various metrics such as accuracy, precision, recall, and F1 score.
- Model Selection and Optimization: Based on the evaluation metrics, the best-performing model is selected and optimized using hyperparameter tuning and cross-validation techniques.
- Model Deployment: The final step is to deploy the model and use it to predict the visa status of new visa applications.
- Recommendations: Based on the analysis, actionable insights and recommendations can be made to improve the visa approval process and reduce the likelihood of visa application denials.

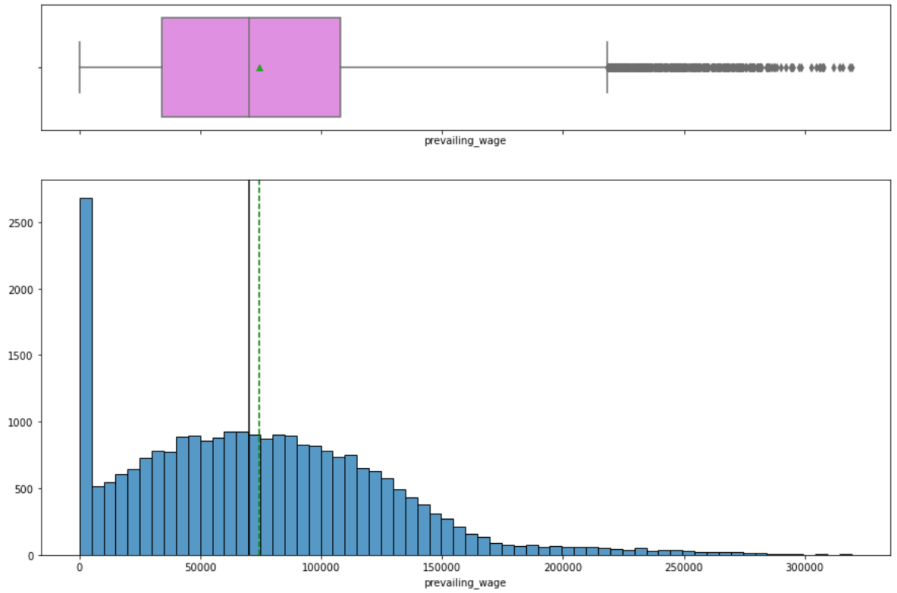
Data Overview

- *The dataset has 25,480 rows and 12 columns. There are nine (9) object types, which means they have text values, one float value, and two integer values.*
- *There are no missing or duplicated values in the dataset.*
- *The average prevailing wage for employees in comparable roles is 74,456.*
- *The maximum number of employees is 602,069.*
- *The average year of company formation was 1979.*

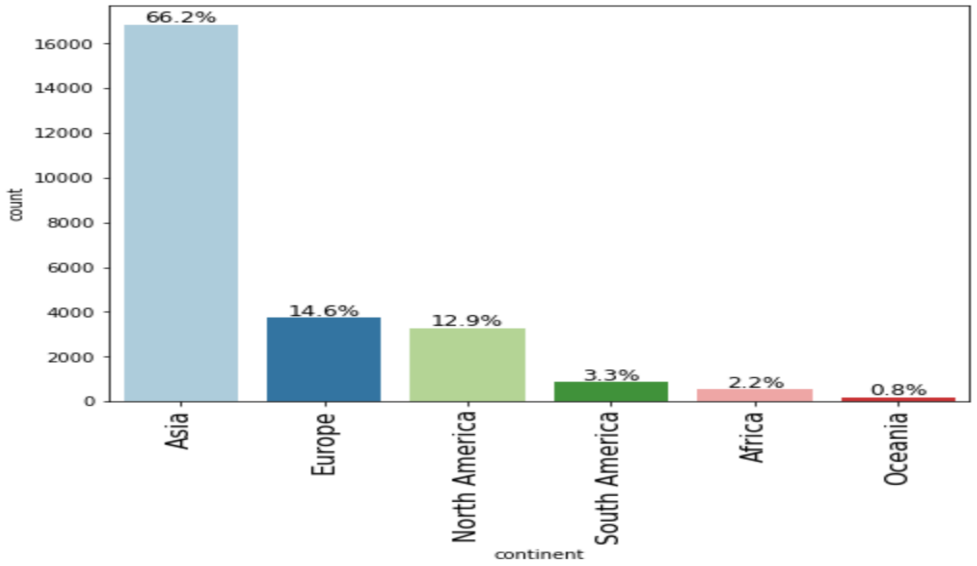
Link to Appendix slide on data background check

EDA Results

- The distribution for prevailing wage is rightly skewed
- The boxplot shows that this variable has many outliers to the right.

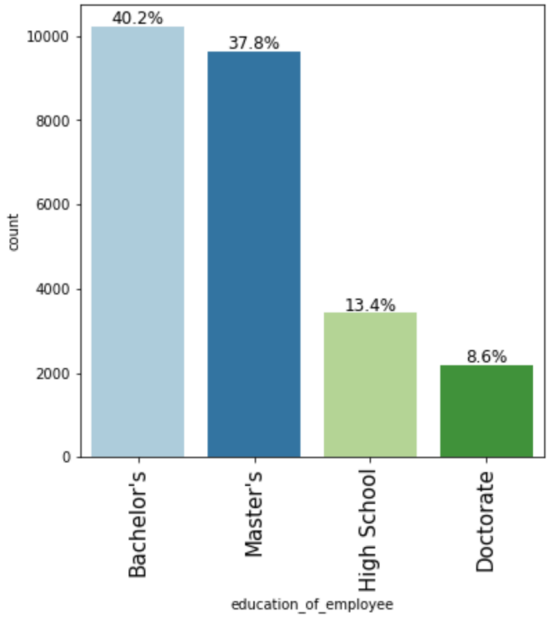


- Asia had the most applicants, with 66.2%.
- Europe and North America are second and third, respectively, with 0.8%, followed by South America, Africa, and Oceania.

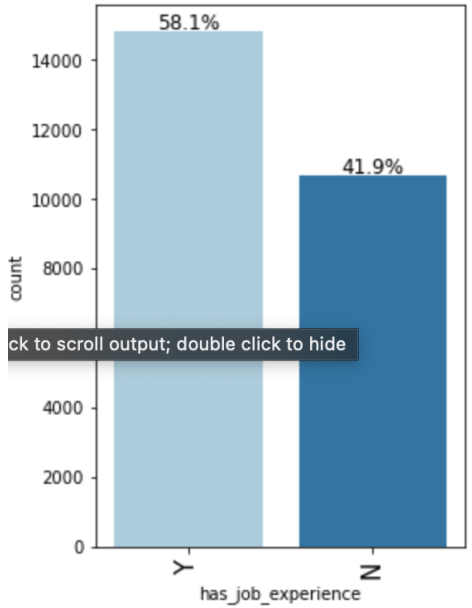


EDA Results

- According to the graph, 40.2% of applicants have a bachelor's degree, 37.8% have a master's degree, 13.4% have only a high school diploma, and 8.6% have a doctorate degree.

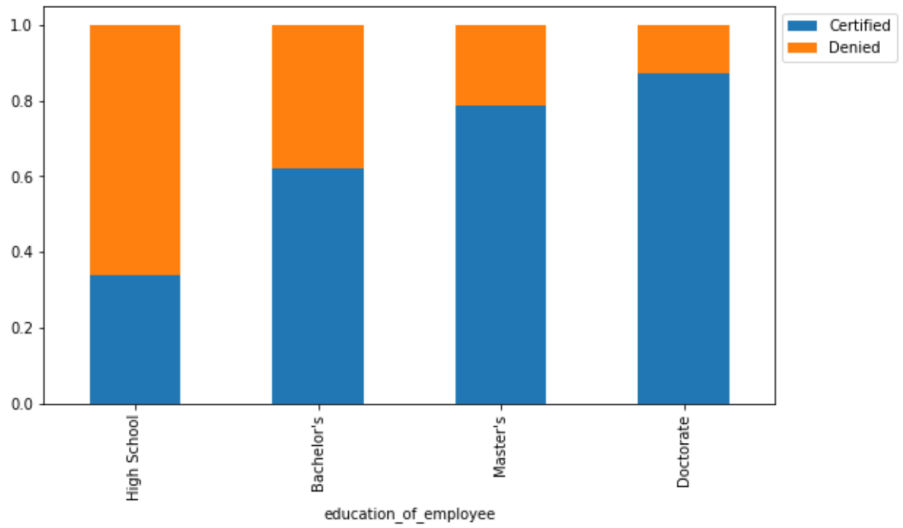


• More than half of the applicants (58.1%) have work experience; 41.9% have no work experience.

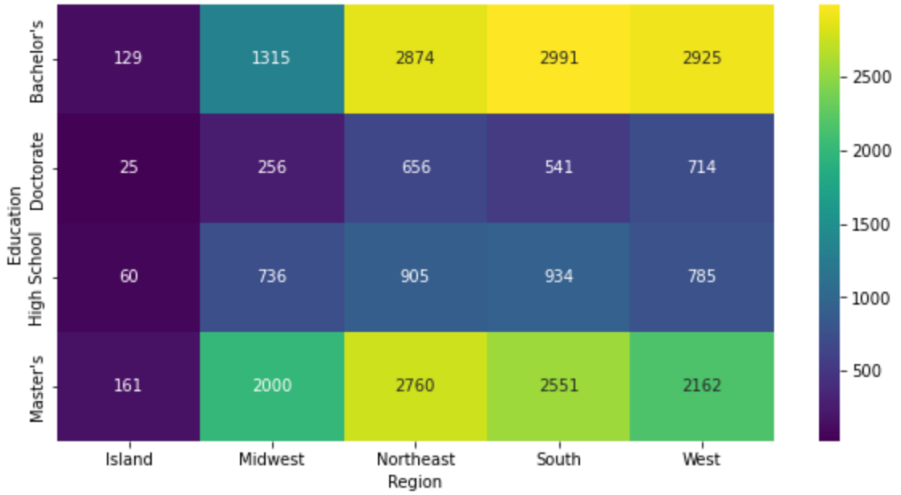


EDA Results

- The distribution below shows that education level has an effect on visa certification.
- Candidates with only a high school diploma have a very low rate of visa acceptance.
- The higher the employee's education, the better the chances, with a doctorate degree having the highest rate of visa approval.



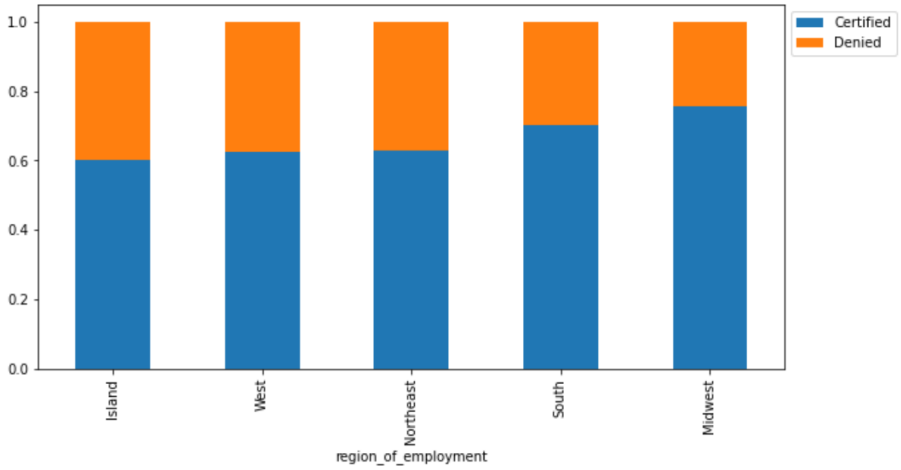
- The heat map chart shows that different regions have different educational background requirements.
- The island accepts the fewest applicants, which may be due to population size.

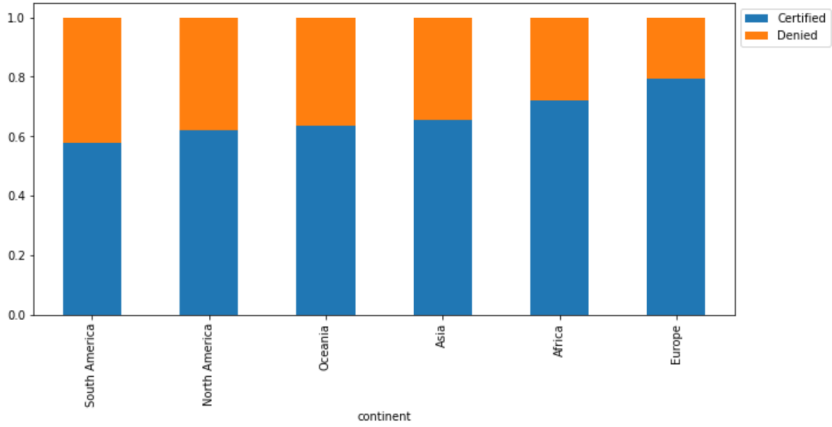


EDA Results

- The island has the highest denial rate of any region, at around 40%.
- The west and northeast have denial rates of 38% and 37%, respectively.
- The Midwest has the lowest denial rate of 25%.

- Europe and Africa have the highest visa approval rates on the continent, with 79% and 72%, respectively.
- Asia, Oceania, and North America have approval rates of 65%, 64%, and 62%, respectively.
- With 58% approval, South America has the lowest approval rate.

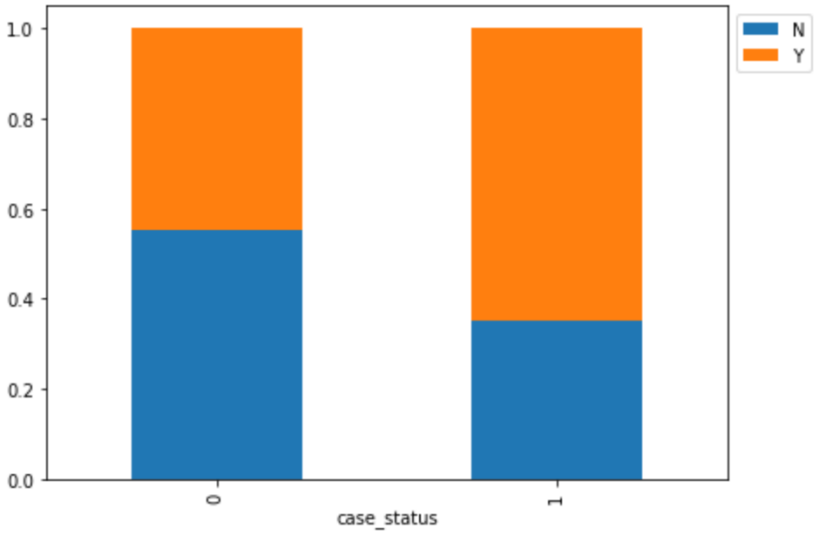


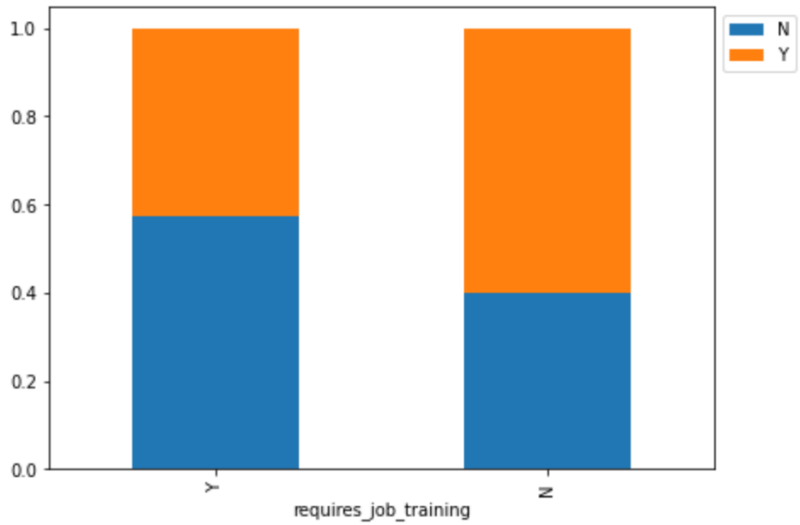


EDA Results

- Despite their job experience, 26% of applicants were denied a visa.
- 44% of applicants with no prior work experience were denied a visa.
- This distribution demonstrates that having work experience has an impact on visa approval.

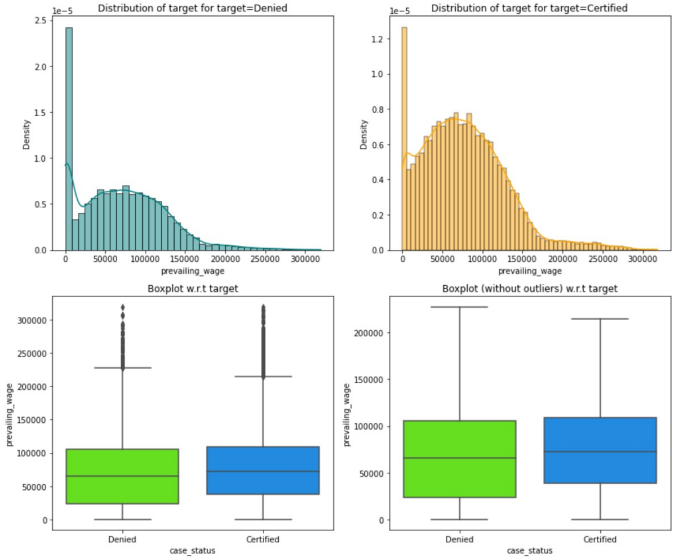
- Yes, very few applicants with previous work experience needed job training.
- This distribution shows that even with job experience, only 9% require job training.



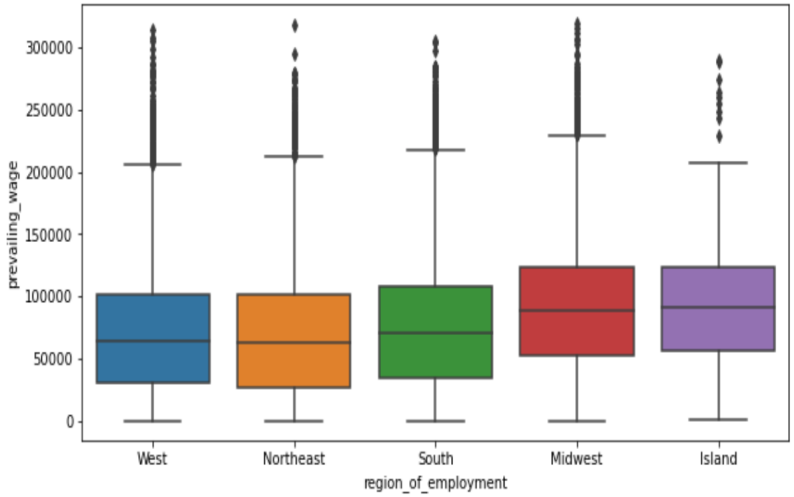


EDA Results

- There are outliers in both distribution boxplots. • This distribution demonstrates that the prevailing wage is slightly similar across industries.
- The distribution demonstrates that case status varies slightly with prevailing wages across all regions.

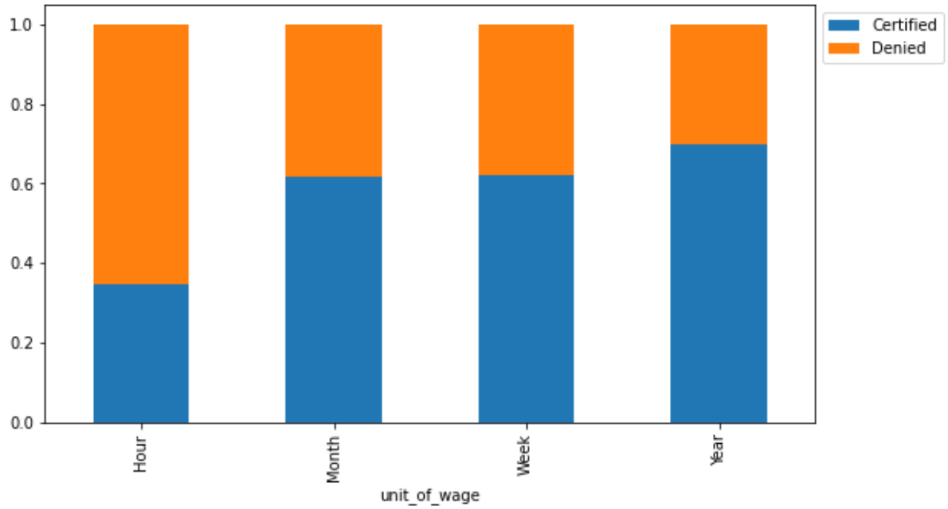


- The prevailing wage is slightly similar across this distribution.



EDA Results

- The various units have an effect on visa certification, with the hourly unit having the lowest certified rate.
- The yearly unit, as well as the monthly and weekly units, has the highest certified rate.



Data Preprocessing

- The data set contains no duplicates or missing values.
- Outliers in the box plot for prevailing wage and year of establishment were discovered.
- Data preparation will be used to forecast which visas will be approved.

Model Performance Summary

- **Create a predictive model that can predict which visas will be approved.**
- **The F1 score will be used as a performance evaluation metric.**

i. If a visa is certified when it should have been denied, the wrong employee will be hired, while US citizens will miss out on the opportunity to work on that position.

ii. If a visa is denied when it is required to be certified, the United States will lose a qualified human resource who can contribute to the economy.

iii. The higher the F1 score, the better the chances of minimizing False Positives and False Positives.

- **The most important predictors of getting a visa certified are employee education, job experience, and prevailing wage.**

Model Performance Summary

Model	Train accuracy	Test Accuracy	Train recall	Test recall	Train precision	Test precision	Train F1	Test F1
Decision Tree	1.0	0.64	1.0	0.72	1.0	0.74	1.0	0.73
Tuned Decision Tree	0.71	0.70	0.93	0.93	0.72	0.71	0.81	0.80
Bagging Classifier	0.98	0.68	0.98	0.75	0.99	0.76	0.99	0.75
Tuned Bagging Classifier	0.98	0.72	0.99	0.88	0.97	0.74	0.98	0.81
Random Forest	1.0	0.70	1.0	0.80	1.0	0.76	1.0	0.78
Tuned Random Forest	0.77	0.74	0.89	0.87	0.79	0.77	0.84	0.81

Model Performance Summary

Adaboost Classifier	0.73	0.73	0.89	0.88	0.75	0.75	0.82	0.81
Tuned Adaboost Classifier	0.71	0.71	0.78	0.78	0.79	0.79	0.78	0.78
Gradient Boost Classifier	0.75	0.74	0.88	0.87	0.78	0.77	0.83	0.82
Tuned Gradient Boost Classifier	0.76	0.74	0.88	0.87	0.78	0.77	0.83	0.81
XGBoost Classifier	0.82	0.73	0.92	0.85	0.83	0.76	0.87	0.80
XGBoost Classifier Tuned	0.76	0.74	0.88	0.87	0.79	0.77	0.83	0.82
Stacking Classifier	0.77	0.74	0.88	0.86	0.79	0.77	0.83	0.81

Model Building And Model Improvement

Building steps

• First, the data must be separated into training and testing sets, and we use the training data model to compute predictions over the testing data.

• First, the data must be separated into training and testing sets, and we use the training data model to compute predictions over the testing data.

• Compute predictions and evaluate the model to ensure performance.

• Finally, if you are not satisfied with the performance results, you can tune the hyperparameters.

DECISION TREE

Model performance

The decision tree is overfitting the training data.

Model improvement after hyper parameter tuning

The recall is still overfitting, the decision and f1 test score improved after hyper parameter tuning

BAGGING CLASSIFIER

Model performance

- The bagging classifier is overfitting the training data
- The test f1 score decrease from initial model

Model improvement

The training data is still overfitting after hyperparameter tuning, although there is an increase in test f1 score

Model Building And Model Improvement

ADABOOST

Model performance

The Adaboost model shows no improvement from the initial model and precision decrease

Model improvement

There was a decrease in both recall and f1 test score after hyper parameter tuning

GRADIENT BOOST

Model performance

The gradient boost shows the highest f1 test score and a balance across all ranges which gives the best model to predict if a visa will be certified

Model improvement

No improvement shown except a decrease in f1 test score after hyper parameter tuning

STACKING CLASSIFIER

Model performance

The stacking classifier shows improvement from all ranges except a decrease in test f1 score from previous model

Executive Summary (Insights and recommendations)

Based on the analysis of the visa application dataset, the following actionable insights and recommendations can be made:

- The prevailing wage is the most significant driver of visa approval, followed by the intended region of employment and the number of employees in the employer's company. Employers should ensure that the prevailing wage paid to foreign workers is at or above the average wage paid to similarly employed workers in the area of intended employment.

- Employers should prioritize applicants who have relevant job experience and do not require job training. This can increase the likelihood of visa approval and reduce the time and cost involved in providing job training to foreign workers.

- Full-time positions are more likely to be approved for visa certification compared to part-time positions. Employers should consider offering full-time positions to foreign workers to increase the chances of visa approval.

Executive Summary (Insights and recommendations)

Based on the analysis of the visa application dataset, the following actionable insights and recommendations can be made:

- The most important factor in determining whether a visa will be granted is the prevailing wage, followed by the desired region of employment and the size of the employer's business. Companies must make sure that the prevailing wage provided to foreign employees is equal to or higher than the mean wage for similarly employed workers in the region of intended employment.

- Companies should give preference to candidates with appropriate work experience who do not need on-the-job training. This can save the time and expense associated with job training for foreign workers while also increasing the likelihood that a visa will be approved.

- Compared to part-time jobs, full-time positions have a higher likelihood of having their visa certification approved. Businesses should think about hiring foreign workers full-time to improve their chances of getting their visas approved.

Executive Summary (Insights and recommendations)

- The number of employees in the employer's organization and the anticipated region of employment are the next two most important factors that influence visa approval. Companies must ensure that the prevailing wage provided to foreign workers is equal to or higher than the average wage paid to similarly employed workers in the planned employment region.

- Candidates with relevant work experience and no need for on-the-job training should be given preference by employers. This may enhance the likelihood that a visa will be granted and decrease the time and expense needed to educate foreign personnel for their jobs.

- Part-time jobs are less likely to get their visa certification accepted than full-time jobs. To improve the likelihood that a visa will be approved, employers ought to think about hiring foreign workers full-time.

- The data revealed the following insights:

I. It shows that 66.8% of visas were certified and 33.2% were denied;

II. employee education (high school), job experience, and prevailing wage all had a significant influence on whether visas should be certified or denied.

Great Learning

Happy Learning !