EASY VISA PROJECT

COURSE TITLE: ENSEMBLE TECHNIQUES

DATE: JUNE 22ND, 2022

CONTENTS

- Executive summary
- Business Problem Overview & solution approach
- Exploratory Data Analysis (EDA)
- Data preprocessing
- Model performance summary
- Appendix
- Business Insights and Recommendations

BUSINESS PROBLEM OVERVIEW & SOLUTION APPROACH

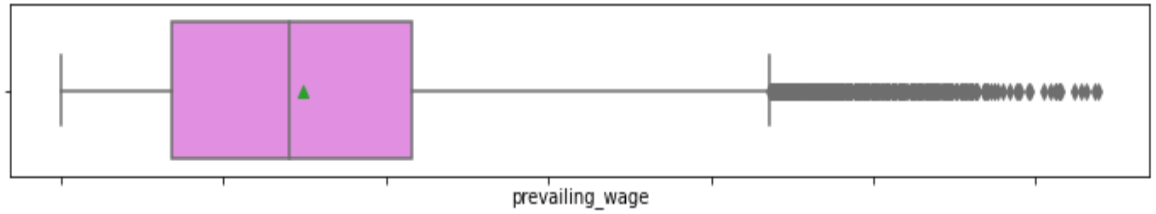
- There is a high demand for human resources in the business communities on the united states
- Companies are seeking out abroad for hardworking, talented and qualified individuals to fill the position
- There are agencies in the united states like The immigration and Nationality Act (INA), who gives the permit for foreign workers on either temporary or permanent basis, and also protect US workers against adverse impact on their wages or working conditions
- Secondly, The Office of Foreign Labor Certifications (OFLC) processes and grant job certification applications for employers , where they can demonstrate that there are not sufficient US workers available to perform the work at wages that meet or exceed the wage paid for the occupation in the area of intended employment.
- The process of reviewing every application is becoming a tedious task as the number of applicants is increasing every year, with a nine percent increase from the previous year
- To analyze the data with the help of classification model
- To facilitate the process of visa approvals and recommend a fitting profile for the applicants for whom the visa should be certified or denied based on the drivers that significantly influence the case status.

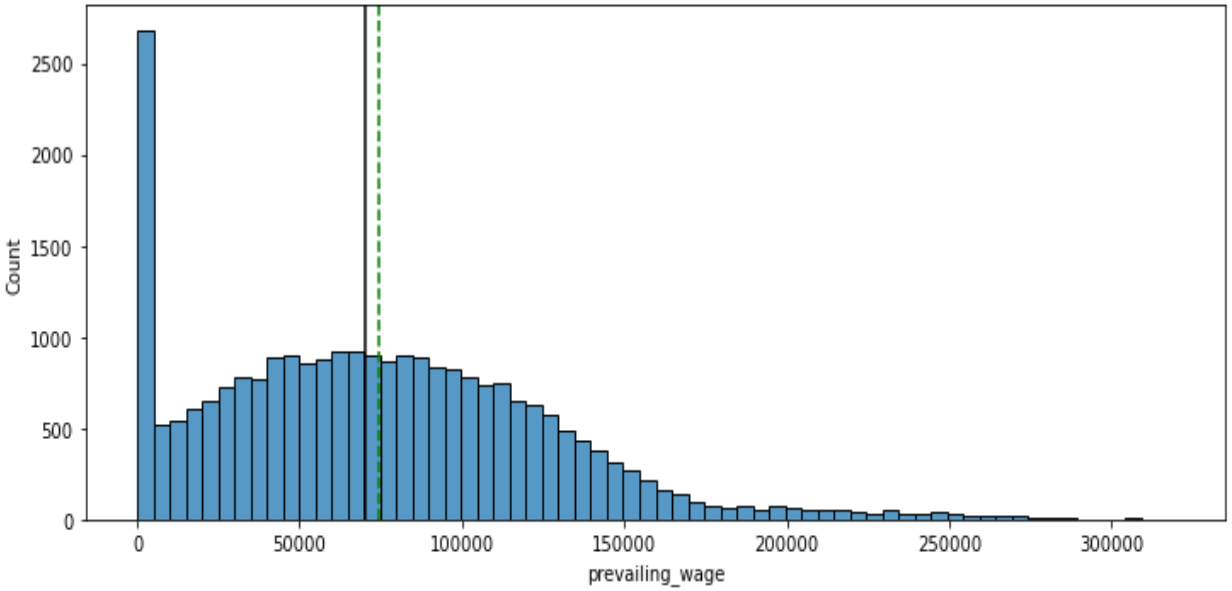
DATA OVERVIEW

- The data set consist of 25,480 rows and 12 columns
- we have nine (9) object type, which means they have text values and 1 float and 2 integer
- there are no missing or duplicated value in the dataset
- The average prevailing wage paid to employees in similar role is 74,456
- the maximum number of number of employees 602,069
- The average year of the company establishment was in 1979

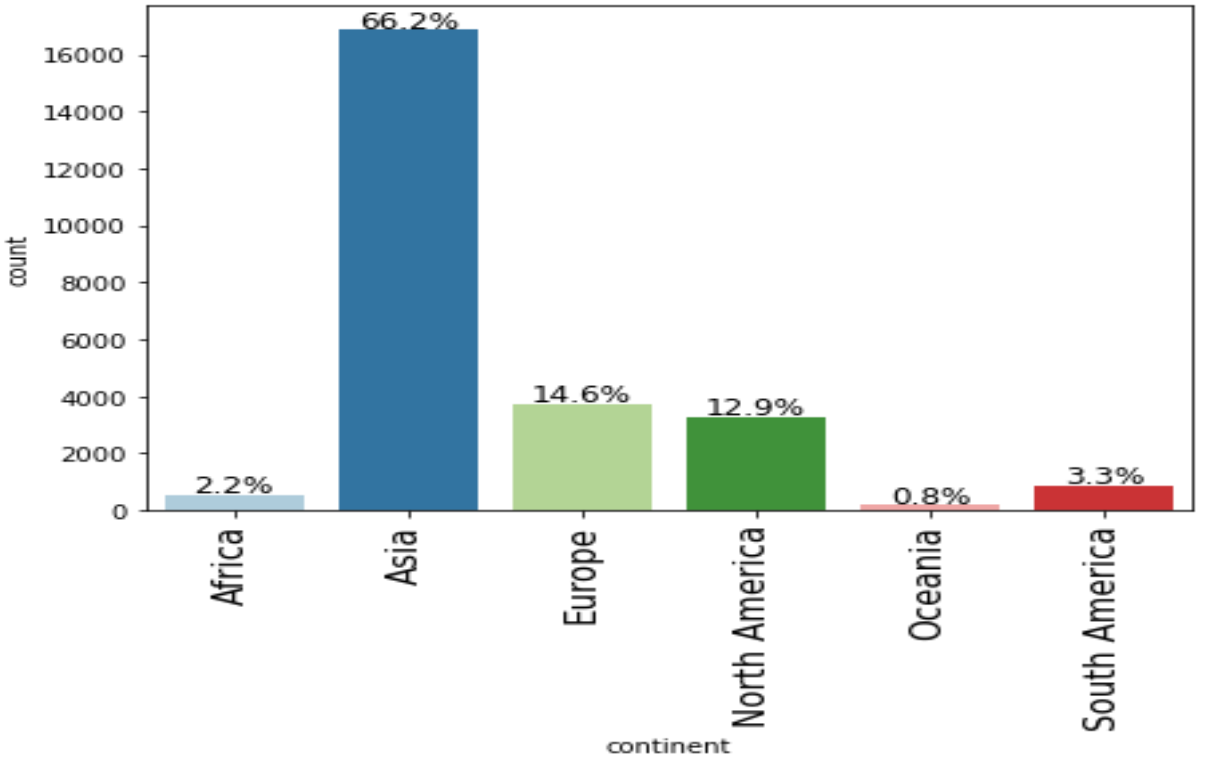
EXPLORATORY DATA ANALYSIS(EDA)

- The distribution for prevailing wage is rightly skewed
- The boxplot shows that there are a lot outliers to the right for this variable.
- 176 applicants had less than 100 as their prevailing wage



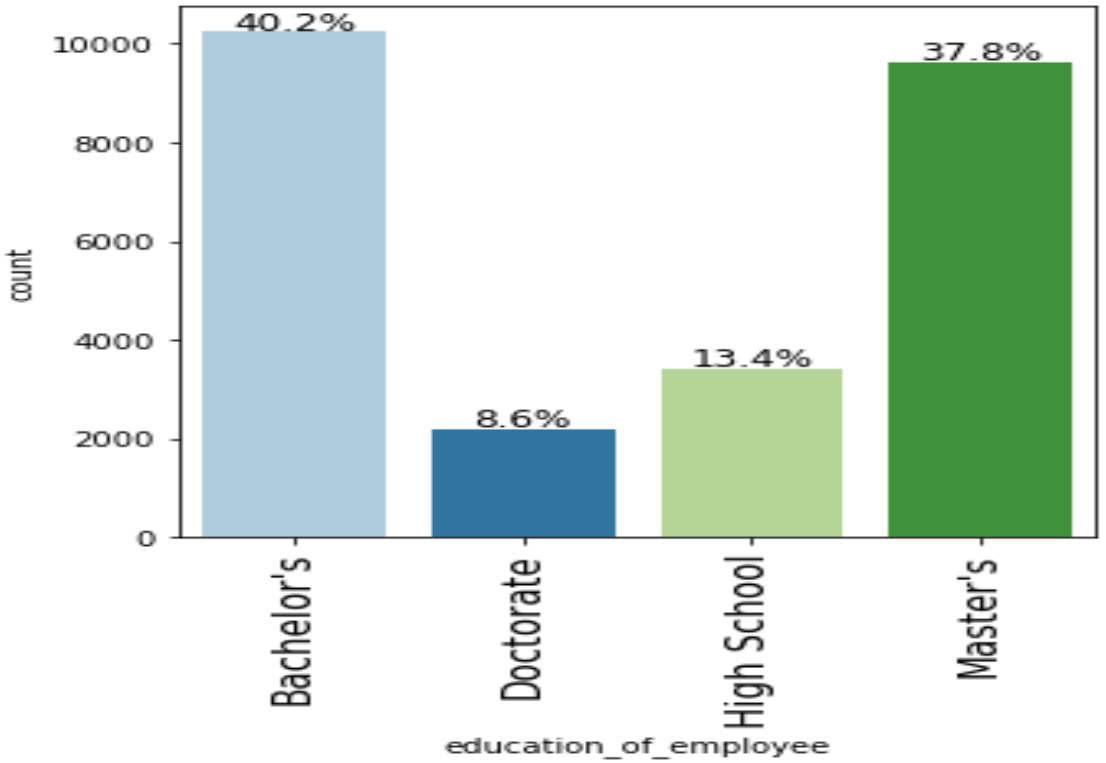


- Asia had the highest number of applicants at 66.2%
- We have Europe and north America coming in 2nd and 3rd respectively, followed by south America, Africa and Oceania at 0.8%

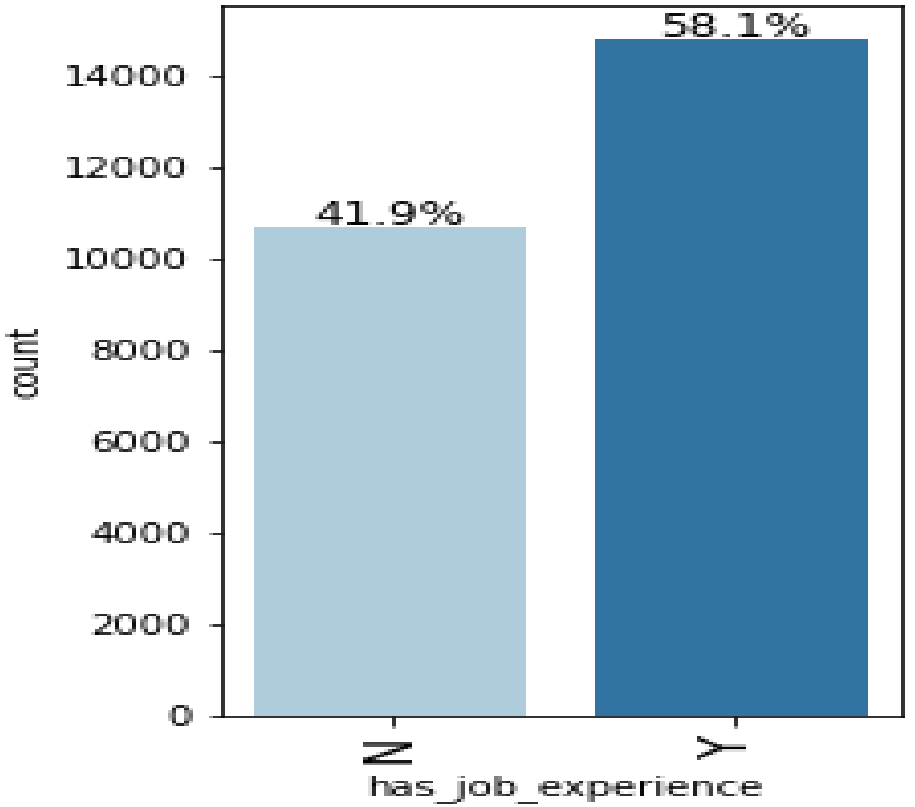


EXPLORATORY DATA ANALYSIS(EDA)

- The chart shows the highest of 40.2% applicants has bachelor's ,
- 37.8% has master's, 13.4% has just high school certificate
- And 8.6% has a doctorate degree

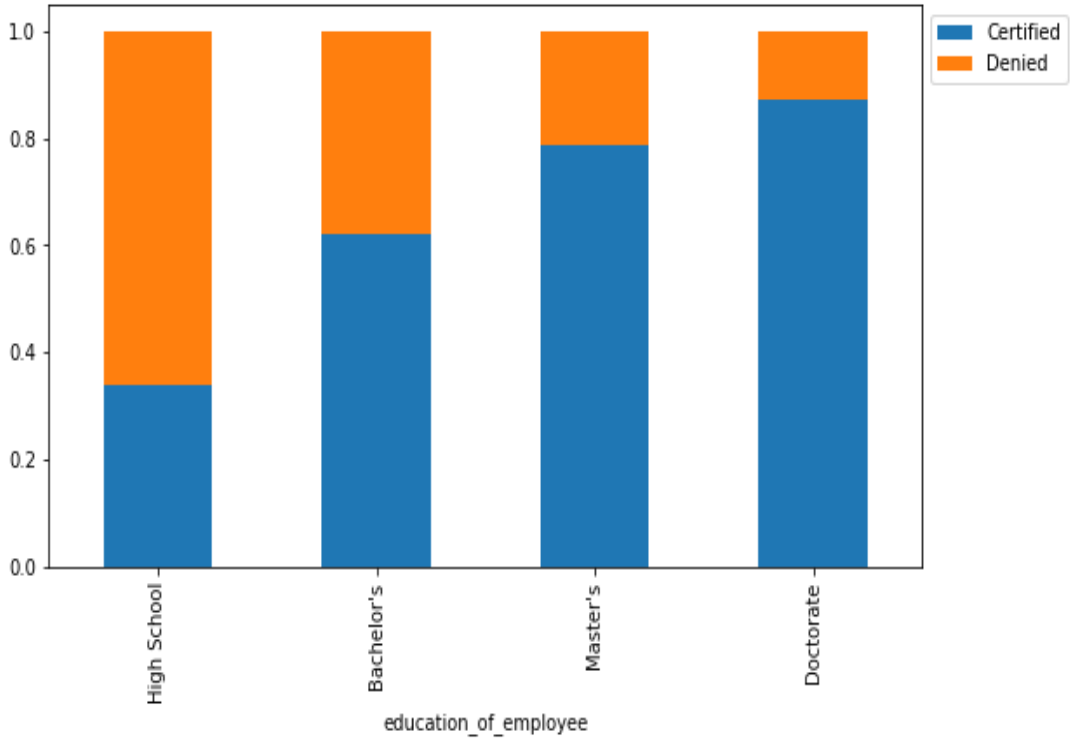


- More than half (58.1%) of the applicants has job experience
- 41.9% has no job experience

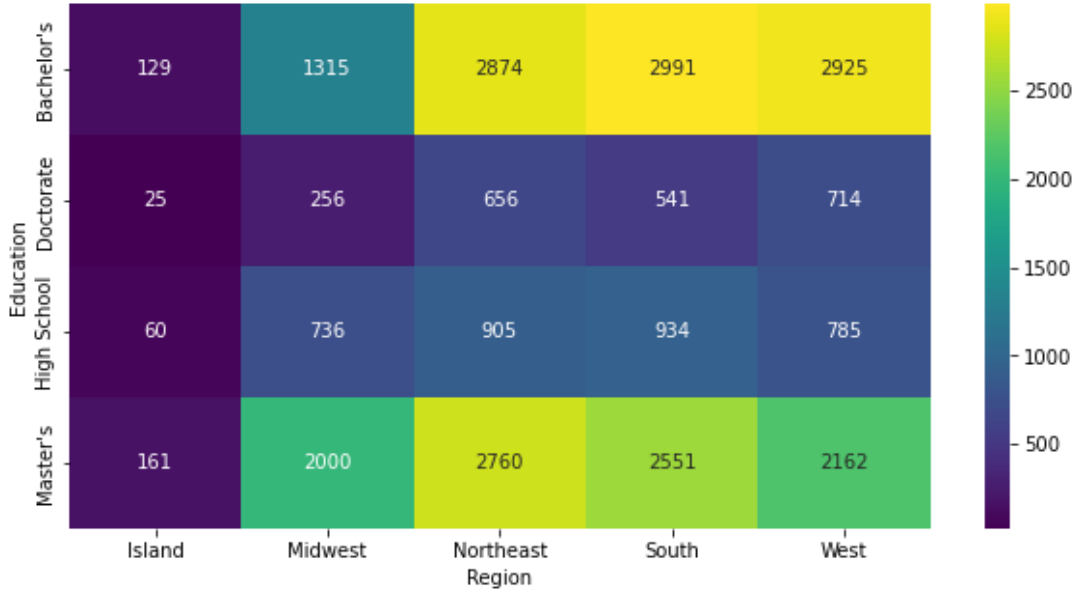


EXPLORATORY DATA ANALYSIS(EDA)

- The distribution below shows that the level of education has an impact on visa certification
- Candidates with just high school certificate has a very low rate on visa acceptance
- The higher the education of the employee the higher the chance with doctorate degree having the highest rate of visa approval

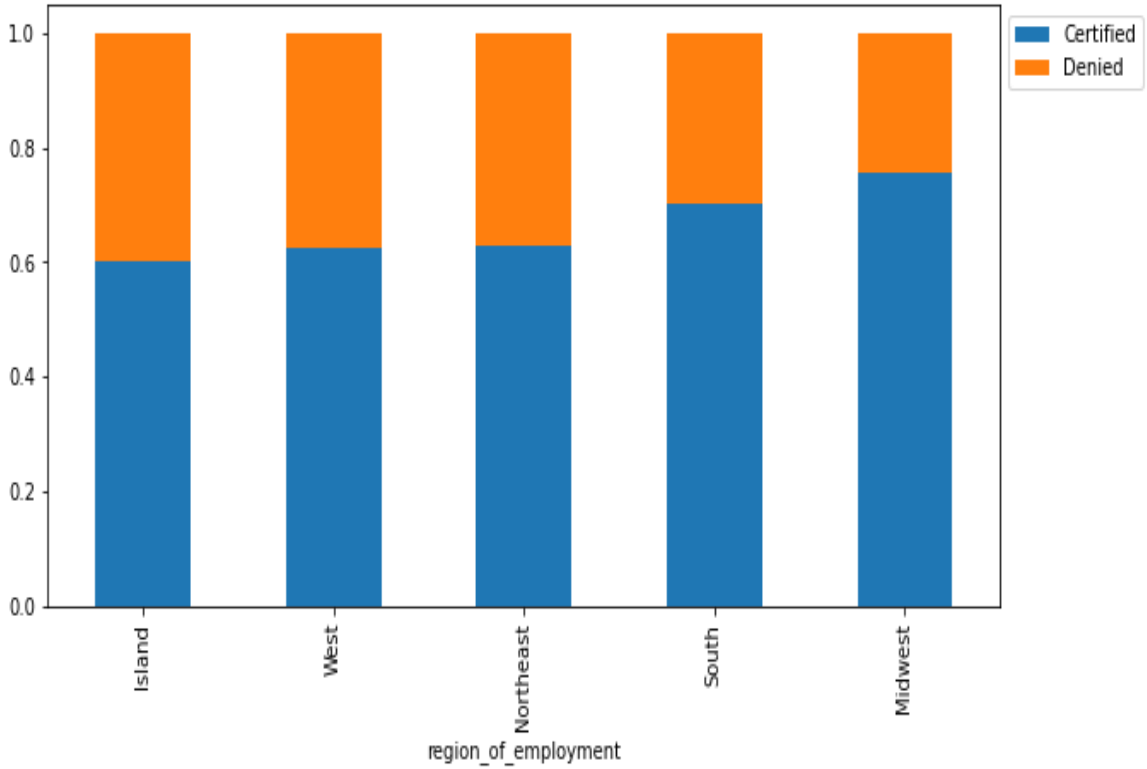


- The heat map chart shows that different regions have different requirement on educational background
- The island takes the lowest number of applicants which can be due to size of population

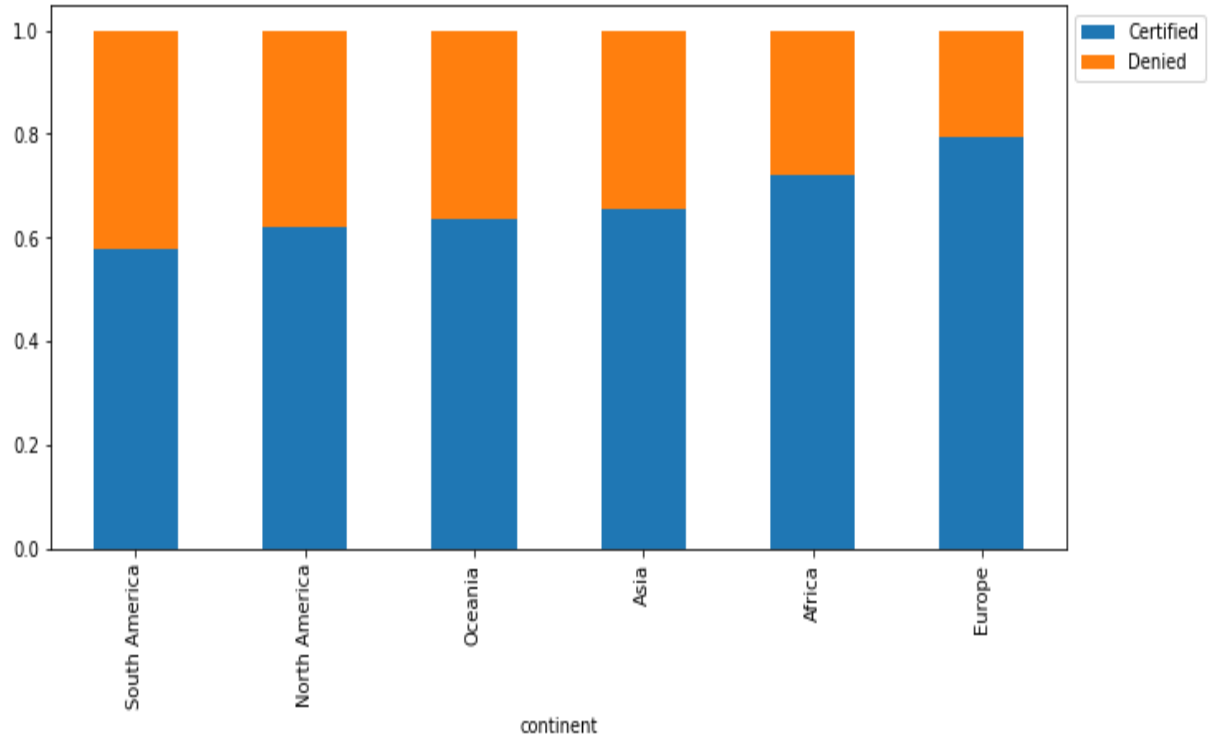


EXPLORATORY DATA ANALYSIS(EDA)

- The island has the highest denial rate across all regions with approximately 40%
- The west and northeast has 38% and 37% denial rate
- The Midwest has the lowest denial rate of 25%

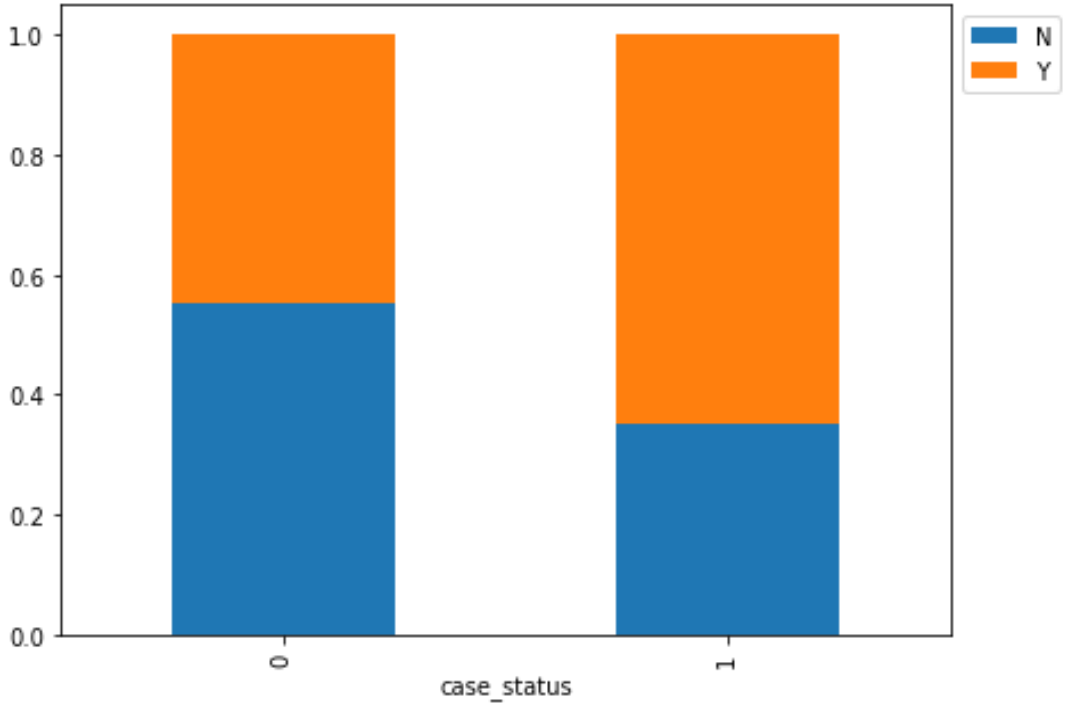


- Across the continent, Europe and Africa has the highest approval rate of visa with 79% and 72% respectively
- Asia, Oceania & North America has 65%, 64% & 62% approval rate respectively
- South America has the lowest approval rate with 58%

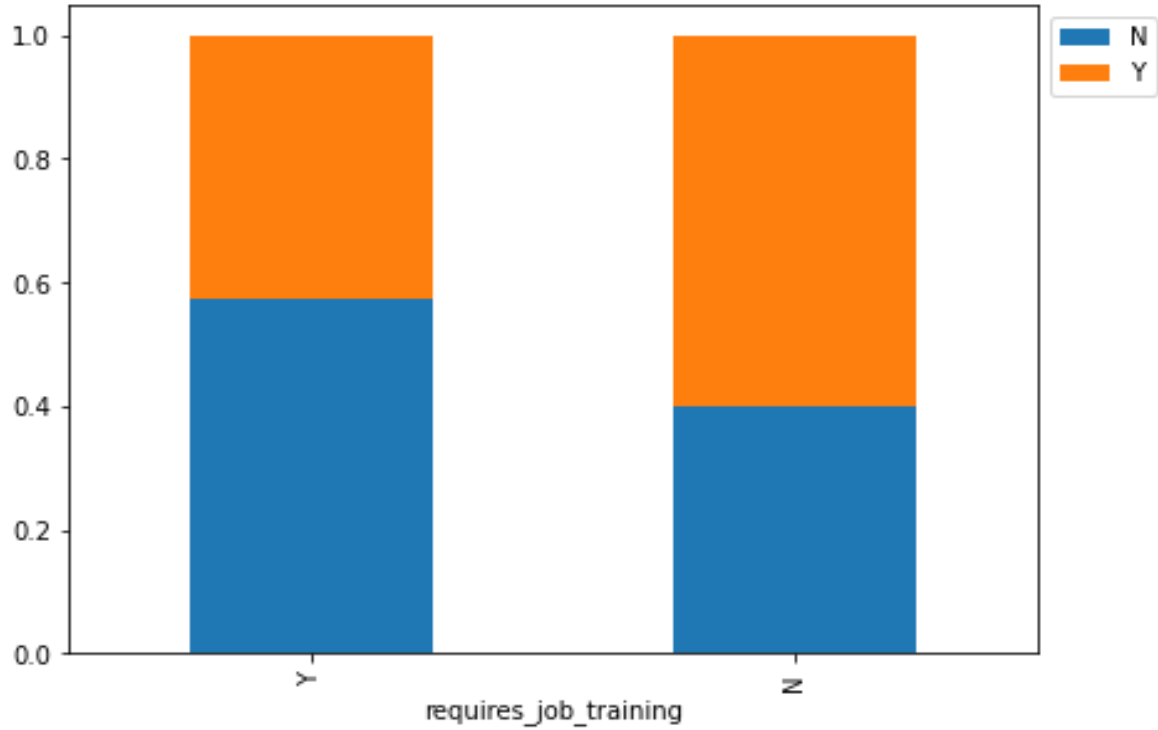


EXPLORATORY DATA ANALYSIS(EDA)

- 26% of applicants were denied visa application even with their job experience
- 44% of applicants who have no job experience were denied visa
- This distribution shows having job experience has an effect on visa certification

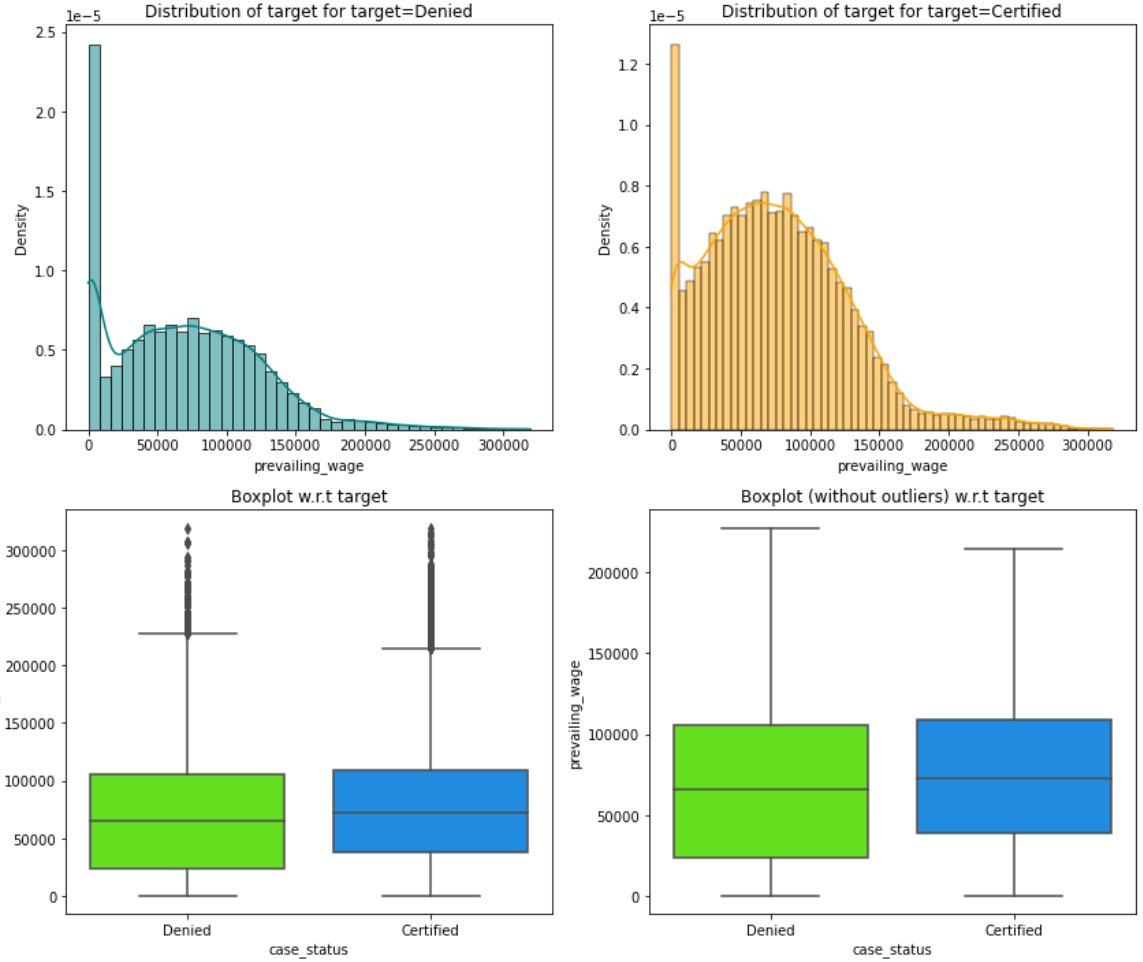


- Yes, very few applicants who have work experience required job training
- This distribution shows that just 9% requires job training even with job experience

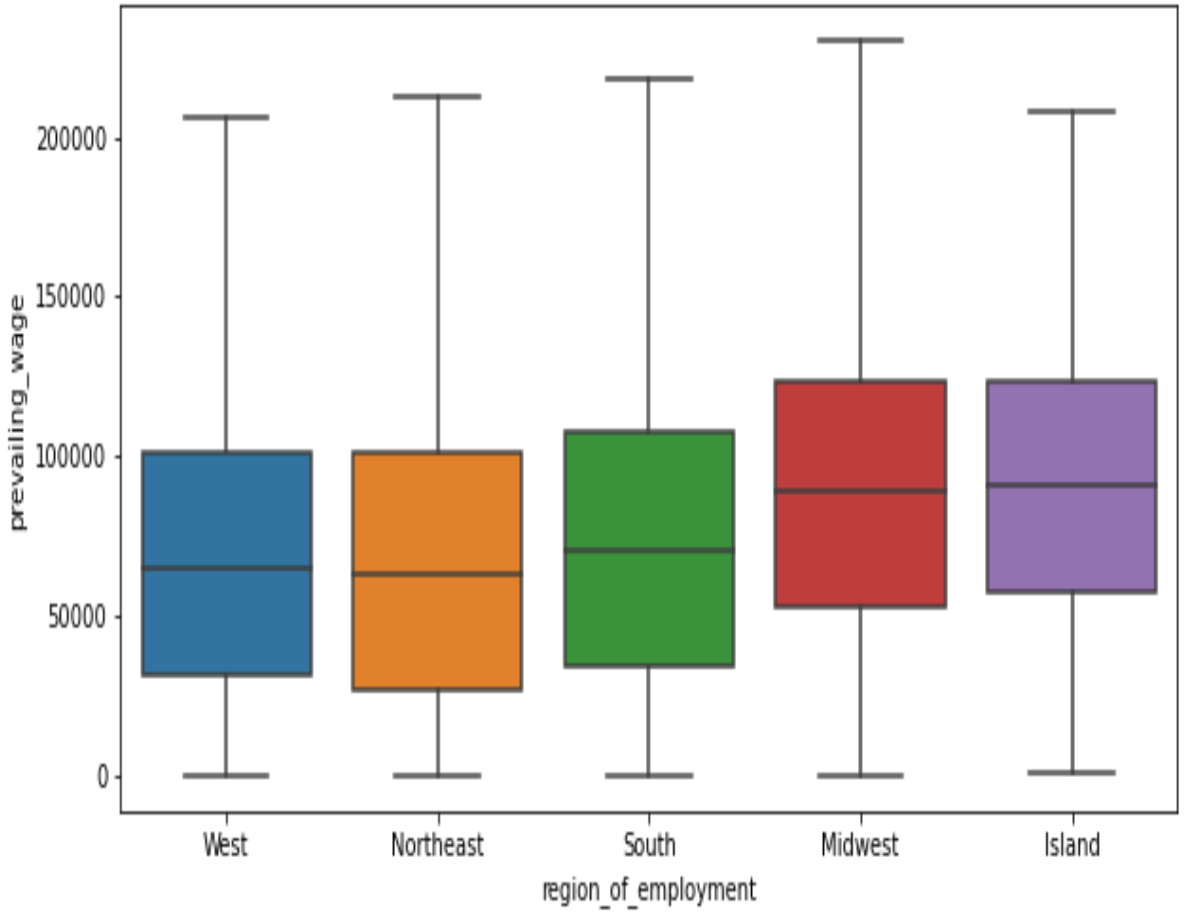


EXPLORATORY DATA ANALYSIS(EDA)

- There are outliers in the boxplots of both distribution
- The distribution shows the case status changes slightly with prevailing wages

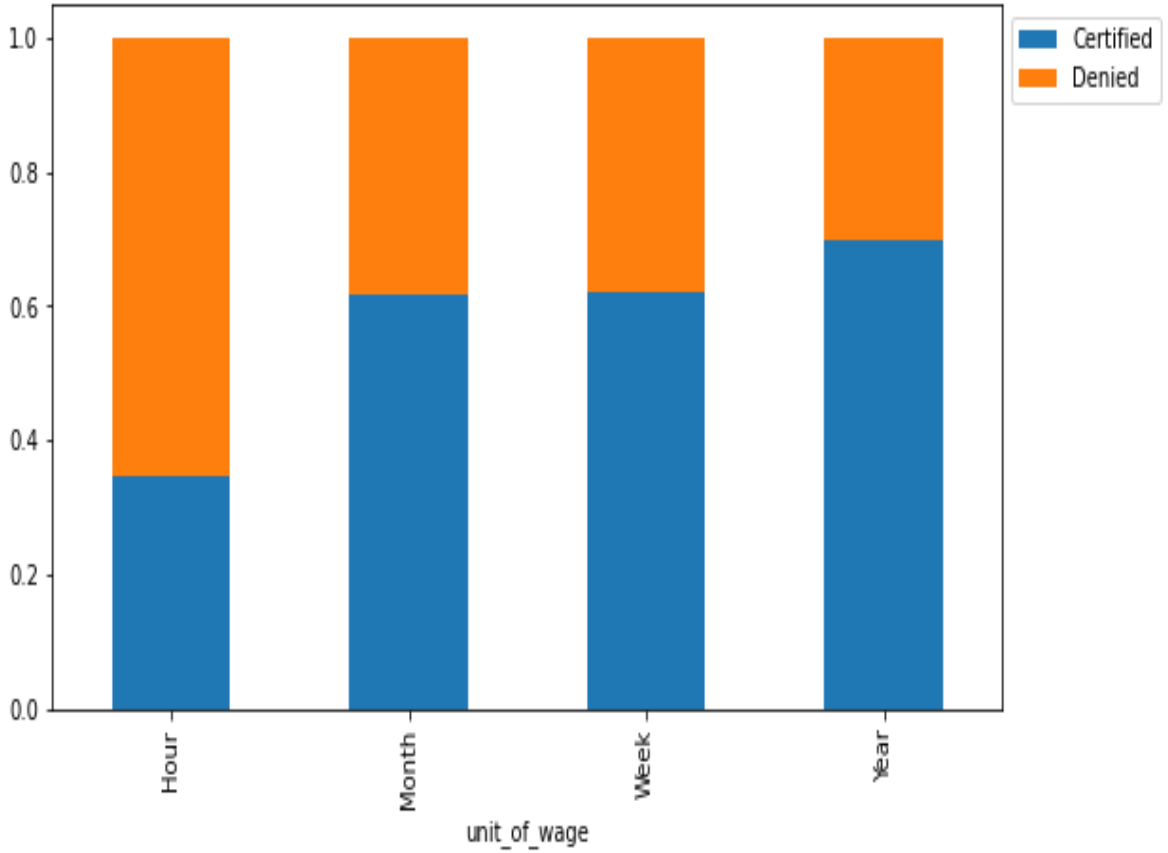


- This distribution shows prevailing wage is slightly similar across all regions



EXPLORATORY DATA ANALYSIS(EDA)

- The different units has an impact on visa been certified, with hourly unit having the lowest certified rate
- The yearly unit has the highest certified rate, along with the monthly and weekly



DATA PREPROCESSING

- There are no duplicates or missing value in the data set
- Outliers was found in the box plot for prevailing wage and year of establishment
- The data preparation will be used to predict which visa will be certified

MODEL PERFORMANCE SUMMARY

- Build a predictive model that can predict which visa will be certified
- F1 score will be used as performance metric of evaluation
  - if a visa is certified when it had to be denied a wrong employee will get the job position while US citizens will miss the opportunity to work on that position.
  - If a visa is denied when it had to be certified the U.S. will lose a suitable human resource that can contribute to the economy.
  - The greater the F1 score higher are the chances of minimizing False Negatives and False

- The most significant predictors of get a visa certified
  - Education of employee
  - Job experience
  - 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

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

Building steps

- Firstly data has to be split into the training and testing, we use the training data model to compute predictions over the testing data
- We build the classification models, after it is trained and tested, its stored in different variables
- Compute predictions and evaluate model to check the performance
- Finally, hyperparameter tuning if not satisfied with performance results to give options

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

RANDOM FOREST

- Model performance
  - The random forest is overfitting the training data
- Model improvement
  - There is an overall improvement on the random forest after hyper parameter tuning

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

MODEL BUILDING

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

BUSINESS INSIGHTS AND RECOMMENDATIONS

The following are the insights the data displayed

- it shows 66.8% visa were certified and 33.2% were denied
- Education of employee (high school), job experience and prevailing wage had a significant influence for whom visa should be certified or denied

The following will be recommended for EASY VISA to help in facilitating the visa process;

- The Information and service provided to visa applicants should be very precise and selective whereby the application should be denies if it doesn’t match the job requirements. This helps eliminates at first stage
- I would recommend the gradient boost model to help facilitate the visa process, this model is set to drive the factors that influence the approval or denial of a visa
- Europe and Africa has the highest acceptance visa rate, although this display almost the lowest number of applicants in the distribution, I would recommend a bench mark number of applicants from Asia which has approximately 66% applying