PGP-DSBa project report

ML-2 – Coded Project

**BY**

**Ishaan Shakti Jayaraman**

**PGPDSBA.O.JULY24.A**

Contents

[LIST OF FIGURES 3](#_Toc185188639)

[INTRODUCTION 4](#_Toc185188640)

[1.1 Objective 4](#_Toc185188641)

[EXPLORATORY DATA ANALYSIS 5](#_Toc185188642)

[2.1 Problem Definition 5](#_Toc185188643)

[2.2 Data Contents 5](#_Toc185188644)

[2.3 Data Dictionary 5](#_Toc185188645)

[2.4 Statistical Summary 7](#_Toc185188646)

[2.5 Univariate Analysis 8](#_Toc185188647)

[2.5.1 No of Employees 8](#_Toc185188648)

[2.5.2 Year of Establishment 9](#_Toc185188649)

[2.5.3 Prevailing Wage 10](#_Toc185188650)

[2.5.4 Continent 11](#_Toc185188651)

[2.5.5 Education of Employee 12](#_Toc185188652)

[2.5.6 Job Experience 13](#_Toc185188653)

[2.5.7 Job Training 14](#_Toc185188654)

[2.5.8 Region of Employment 15](#_Toc185188655)

[2.5.9 Unit of Wage 16](#_Toc185188656)

[2.5.10 Full Time Position 17](#_Toc185188657)

[2.5.11 Case Status 18](#_Toc185188658)

[2.6 Bivariate Analysis 19](#_Toc185188659)

[2.6.1 Heatmap 19](#_Toc185188660)

[2.6.2 Education of employee vs Case Status 20](#_Toc185188661)

[2.6.3 Continent vs Case Status 21](#_Toc185188662)

[2.6.4 Continent vs Education of Employee 22](#_Toc185188663)

[2.6.5 Job Experience vs Case Status 23](#_Toc185188664)

[2.6.6 Unit of Wage vs Case Status 24](#_Toc185188665)

[2.6.7 Full Time Position vs Case Status 25](#_Toc185188666)

[DATA PREPROCESSING 26](#_Toc185188667)

[3.1 Missing Value Check 26](#_Toc185188668)

[3.2 Duplicate Value Check 26](#_Toc185188669)

[3.3 Outlier Check 26](#_Toc185188670)

[3.4 Feature Engineering 26](#_Toc185188671)

[3.5 Data Preparation 26](#_Toc185188672)

[MODEL BUILDING – ORIGINAL DATA 27](#_Toc185188673)

[4.1 Model Evaluation Criterion 27](#_Toc185188674)

[4.2 Building the Models 27](#_Toc185188675)

[MODEL BUILDING – OVERSAMPLED DATA 29](#_Toc185188676)

[5.1 Building the Models 29](#_Toc185188677)

[MODEL BUILDING – UNDERSAMPLED DATA 31](#_Toc185188678)

[6.1 Building the Models 31](#_Toc185188679)

[MODEL PERFORMANCE IMPROVEMENT USING HYPERPARAMETER TUNING 33](#_Toc185188680)

[6.1 Gradient Boost Machine Model with Original Data 33](#_Toc185188681)

[6.2 Adaboost Model with Original Data 34](#_Toc185188682)

[6.3 XGBoost Model with Oversampled Data 34](#_Toc185188683)

[MODEL PERFORMANCE COMPARISON & FINAL MODEL SELECTION 35](#_Toc185188684)

[7.1 Performance of the Best Model on the Test Set 36](#_Toc185188685)

[7.2 Feature Importances 36](#_Toc185188686)

[ACTIONABLE INSIGHTS & RECOMMENDATIONS 37](#_Toc185188687)

# LIST OF FIGURES

[Figure 1 - Statistical Summary of the Dataset 7](#_Toc185188493)

[Figure 2 - Histogram Boxplot of No of Employees 8](#_Toc185188494)

[Figure 3 - Histogram Boxplot of Year of Establishment 9](#_Toc185188495)

[Figure 4 - Histogram Boxplot of Prevailing Wages 10](#_Toc185188496)

[Figure 5 - Labeled Barplot of Continent 11](#_Toc185188497)

[Figure 6 - Labeled barplot of Education of Employee 12](#_Toc185188498)

[Figure 7 - Labeled Barplot of Job Experience 13](#_Toc185188499)

[Figure 8 - Labeled Barplot of Job Training 14](#_Toc185188500)

[Figure 9 - Labeled Barplot of Region of Employment 15](#_Toc185188501)

[Figure 10 - Labeled Barplot of Unit of Wage 16](#_Toc185188502)

[Figure 11 - Labeled Barplot of Full Time Position 17](#_Toc185188503)

[Figure 12 - Labeled Barplot of Case Status 18](#_Toc185188504)

[Figure 13 - Heatmap of all Numerical Variables 19](#_Toc185188505)

[Figure 14 - Stacked Barplot of Education of Employee vs Case Status 20](#_Toc185188506)

[Figure 15 - Stacked Barplot of Continent vs Case Status 21](#_Toc185188507)

[Figure 16 - Stacked Barplot of Continent vs Education of Employee 22](#_Toc185188508)

[Figure 17 - Stacked Barplot of Job Experience vs Case Status 23](#_Toc185188509)

[Figure 18 - Stacked Barplot of Unit of Wage vs Case Status 24](#_Toc185188510)

[Figure 19 - Stacked Barplot of Fulltime Position vs Case Status 25](#_Toc185188511)

[Figure 20 - Boxplot of all Numerical Variables 26](#_Toc185188512)

[Figure 21 Training & Validation Performance on Original Data 28](#_Toc185188513)

[Figure 22 - Training & Validation Difference on Original Data 28](#_Toc185188514)

[Figure 23 - Training & Validation Performance on Oversampled Data 30](#_Toc185188515)

[Figure 24 - Training & Validation Difference on Oversampled Data 30](#_Toc185188516)

[Figure 25 - Training & Validation Performance on Undersampled Data 32](#_Toc185188517)

[Figure 26 - Training & Validation Difference on Undersampled Data 32](#_Toc185188518)

[Figure 27 - Training & Validation Performance on Tuned GBM Model 33](#_Toc185188519)

[Figure 28 - Training & Validation Performance on Tuned Adaboost Model 34](#_Toc185188520)

[Figure 29 - Training & Validation Performance on Tuned XGBoost Model 34](#_Toc185188521)

[Figure 30 - Training & Validation Model Comparison 35](#_Toc185188522)

[Figure 31 - Performance of GBM model on Unseen Test Data 36](#_Toc185188523)

[Figure 32 - Feature Importances of the GBM Model 36](#_Toc185188524)

# INTRODUCTION

Business communities in the United States are facing high demand for human resources, but one of the constant challenges is identifying and attracting the right talent, which is perhaps the most important element in remaining competitive. Companies in the United States look for hard-working, talented, and qualified individuals both locally as well as abroad.

The Immigration and Nationality Act (INA) of the US permits foreign workers to come to the United States to work on either a temporary or permanent basis. The act also protects US workers against adverse impacts on their wages or working conditions by ensuring US employers' compliance with statutory requirements when they hire foreign workers to fill workforce shortages. The immigration programs are administered by the Office of Foreign Labor Certification (OFLC).

OFLC processes job certification applications for employers seeking to bring foreign workers into the United States and grants certifications in those cases where employers 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.

## Objective

In FY 2016, the OFLC processed 775,979 employer applications for 1,699,957 positions for temporary and permanent labor certifications. This was a nine percent increase in the overall number of processed applications from the previous year. The process of reviewing every case is becoming a tedious task as the number of applicants is increasing every year.

The increasing number of applicants every year calls for a Machine Learning based solution that can help in shortlisting the candidates having higher chances of VISA approval. OFLC has hired the firm EasyVisa for data-driven solutions. You as a data scientist at EasyVisa have to analyze the data provided and, with the help of a classification model:

1. Facilitate the process of visa approvals.
2. Recommend a suitable profile for the applicants for whom the visa should be certified or denied based on the drivers that significantly influence the case status.

# EXPLORATORY DATA ANALYSIS

## 2.1 Problem Definition

Business communities in the United States face significant challenges in identifying and attracting the right talent, crucial for competitiveness. The Immigration and Nationality Act allows foreign workers to fill workforce shortages while protecting U.S. workers' conditions. However, the increasing volume of applications for labor certifications has made the manual review process by the Office of Foreign Labor Certification (OFLC) tedious and inefficient.

A machine learning-based solution is needed to assist in shortlisting candidates with higher chances of visa approval, streamlining the process, and improving the accuracy of recommendations for visa certifications or denials.

## 2.2 Data Contents

The dataset (EasyVisa.csv) contains an individual visa application processed by the Office of Foreign Labor Certification

* There are 25480 observations in the dataset.
* There are 12 columns in the dataset containing a detailed view of each visa application.
* There are 9 categorical variables in the dataset.
* There are 3 numerical variables (2 integer types and 1 float type) in the dataset.
* “case\_status” consists of entirely unique values, This column will be dropped from the dataset.
* There are no null values.
* There are no duplicate values.

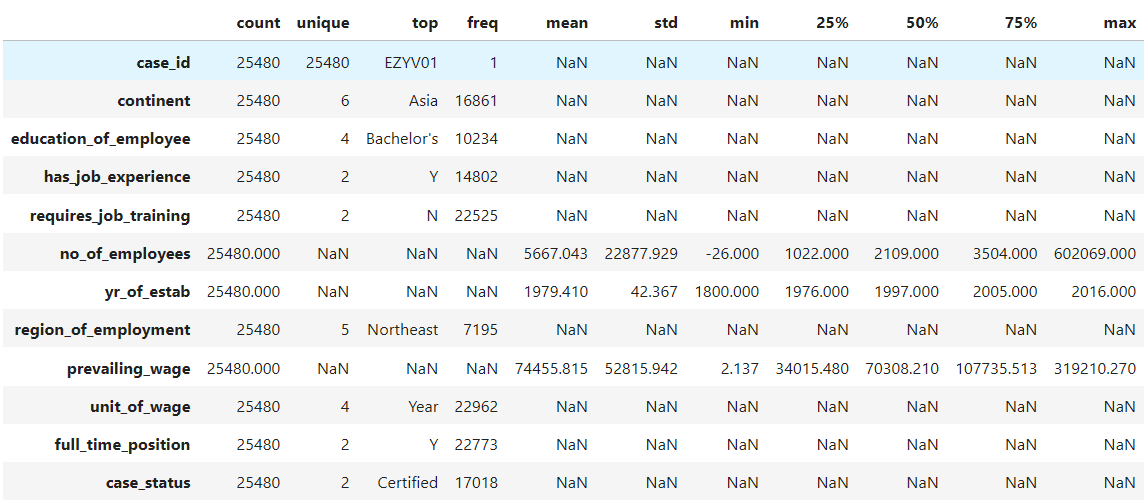
## 2.3 Data Dictionary

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

* 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

## 2.4 Statistical Summary

Figure 1 - Statistical Summary of the Dataset



**Insights:**

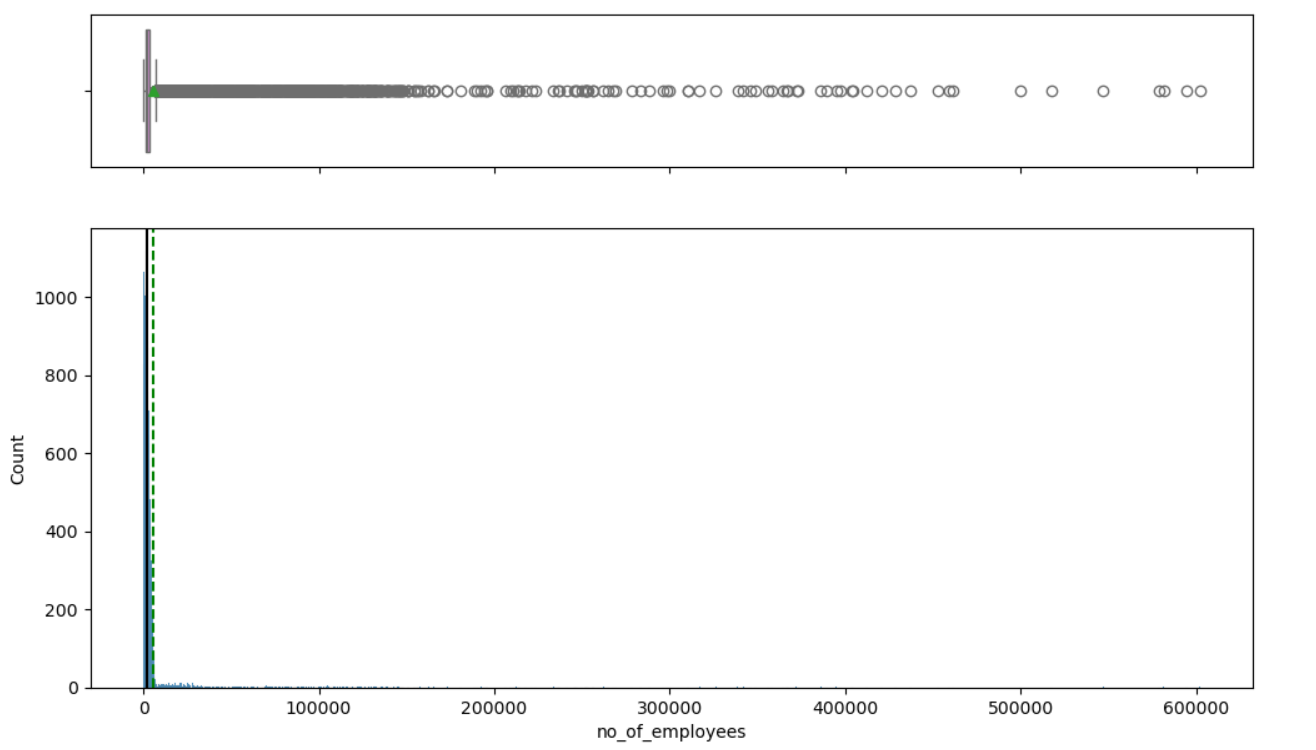
* There are 6 unique continents, with Asia being the most common with 16,861 observations.
* 4 education levels of employees with Bachelor's being the most frequent with 10,234 observations.
* Observations of employees with job experience is more common than those without experience.
* Observations of employees that don’t require job training is more common than those that require job training.
* There is an average of 5,667 employees per company with a high variation with -26 being the lowest value and 602,069 being the highest number.
* -26 of employees is not a possible number, this is most likely a mistake. Negative values will be turned into absolute values during pre-processing.
* On average the year of establishment per company is 1979.
* The average prevailing wage is 74,455.82, with a minimum value of 2.14 and a maximum value of 319,210.27.
* Majority of the observations have their visas certified, a total of 17,018.

## 2.5 Univariate Analysis

Univariate analysis is the simplest form of statistical analysis that involves examining the distribution and characteristics of a single variable. It focuses on summarizing and visualizing the data for one variable to understand its central tendency, spread, and distribution.

### 2.5.1 No of Employees

Figure 2 - Histogram Boxplot of No of Employees

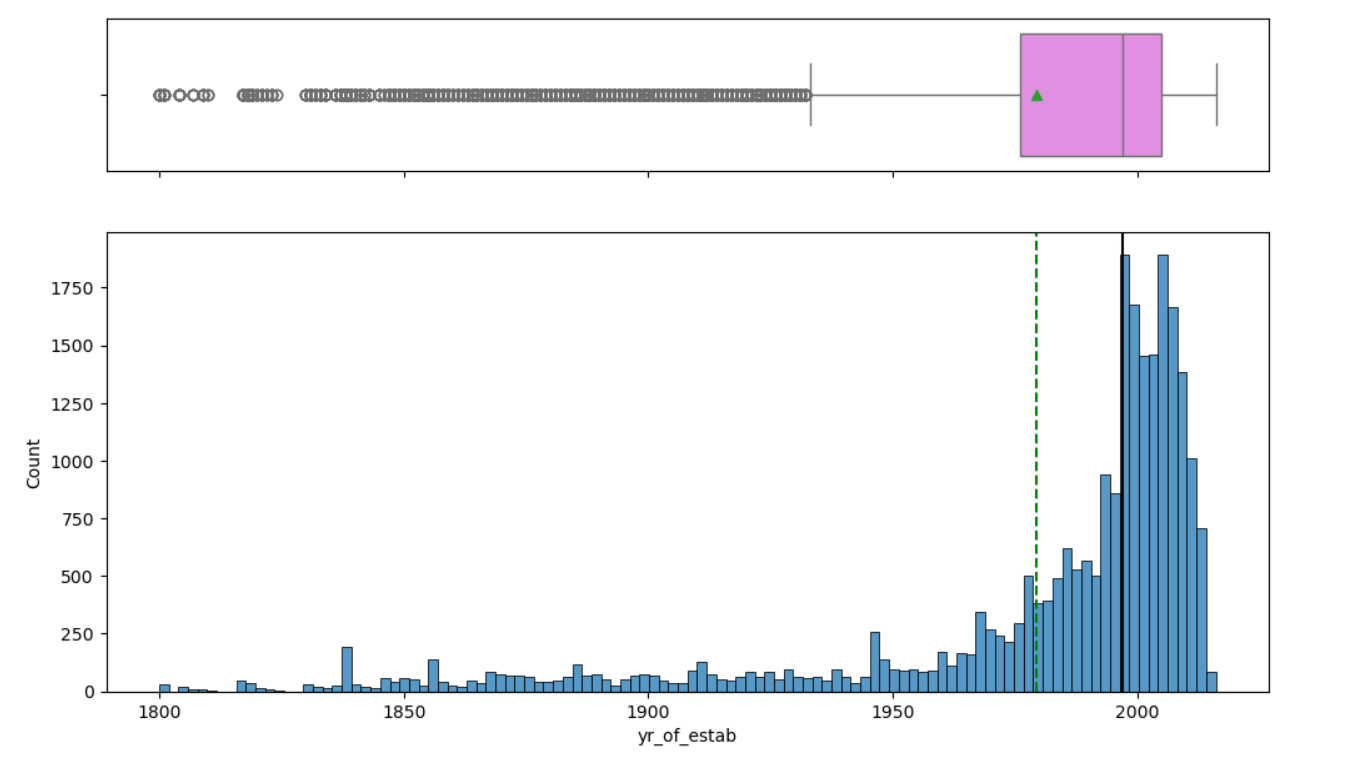


**Insights:**

* The distribution is heavily skewed to the right with many outliers.
* Due to the large variation in the data, the mean looks closer to 0.

### 2.5.2 Year of Establishment

Figure 3 - Histogram Boxplot of Year of Establishment

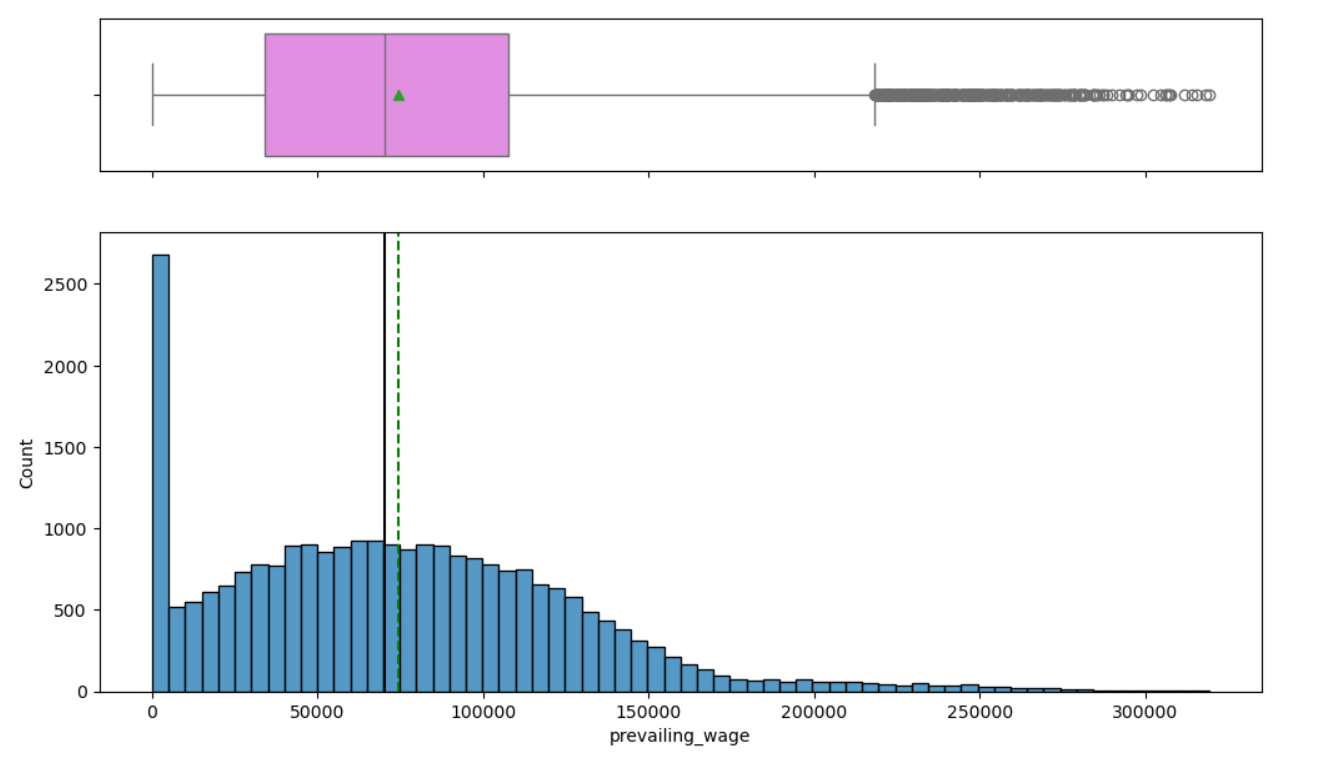


**Insights:**

* The distribution of the chart is left skewed.
* There are outliers present on the left side of the distribution.
* The mean year of establishment is around 1979 close to the 25th Percentile of the data.

### 2.5.3 Prevailing Wage

Figure 4 - Histogram Boxplot of Prevailing Wages

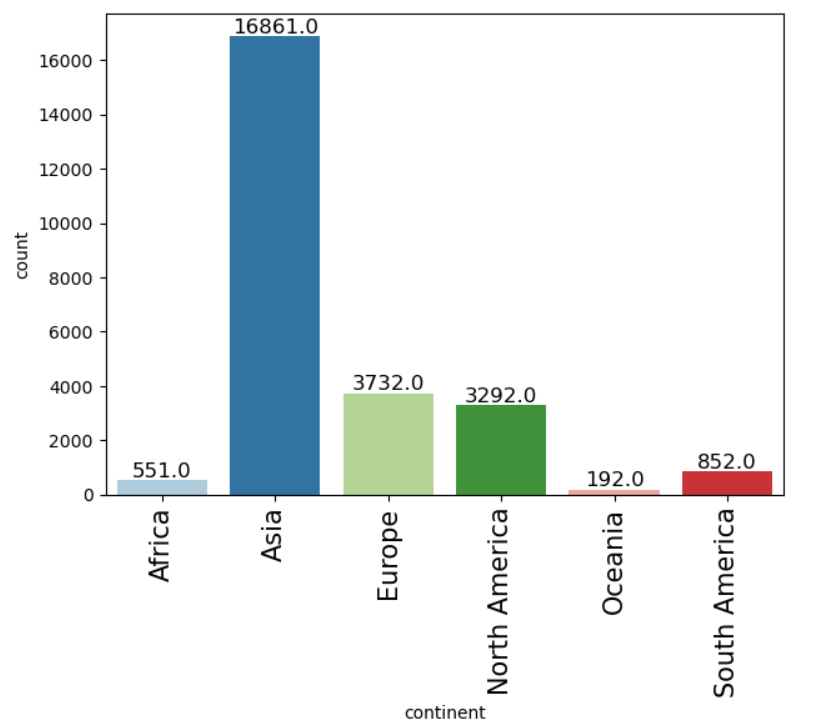


**Insights:**

* The distribution is slightly right skewed.
* The mean and median are relatively close to each other.

### 2.5.4 Continent

Figure 5 - Labeled Barplot of Continent

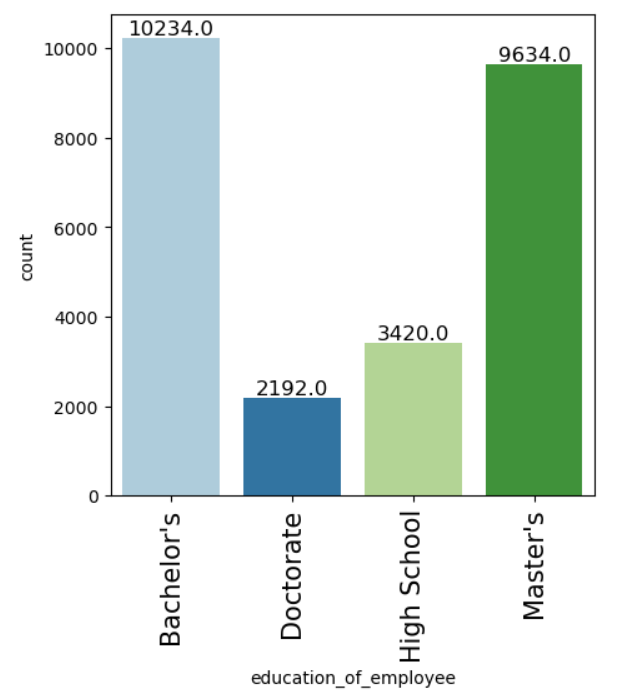
****

**Insights:**

* Majority of the applicant cases are from Asia.
* Oceania has the lowest number of applicant cases.

### 2.5.5 Education of Employee

Figure 6 - Labeled barplot of Education of Employee

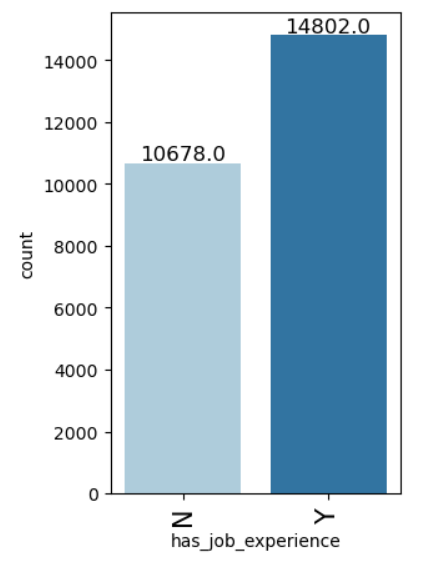


**Insights:**

* The number of employees holding bachelor’s degree is the highest followed by master’s degree.
* Doctorate degree holders are the least.

### 2.5.6 Job Experience

Figure 7 - Labeled Barplot of Job Experience

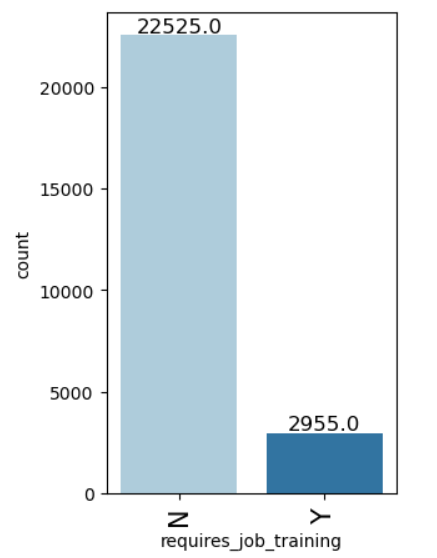


**Insights:**

* Majority of the employees have job experience approximately 58% of the dataset and the rest do not have job experience.

### 2.5.7 Job Training

Figure 8 - Labeled Barplot of Job Training

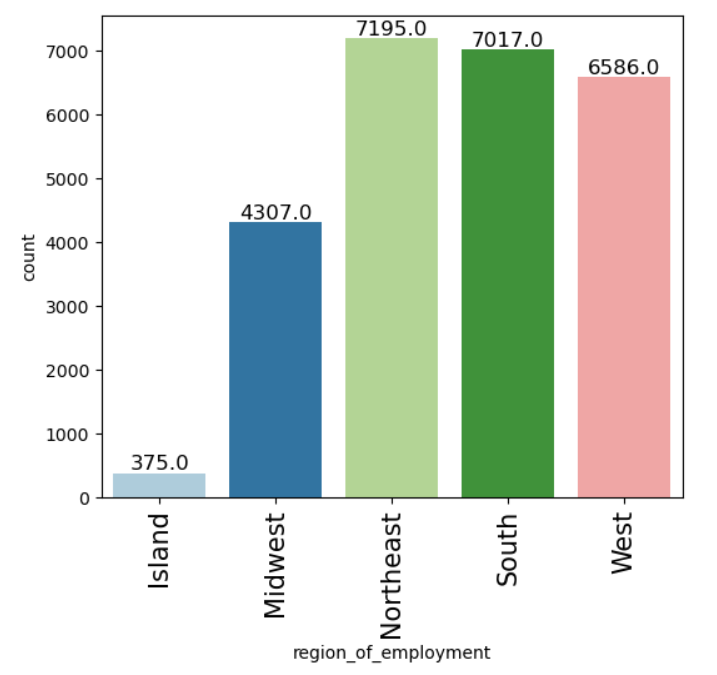


**Insight:**

* Huge majority of the employees don’t require job training approximately 88% of the data, The rest require job training.

### 2.5.8 Region of Employment

Figure 9 - Labeled Barplot of Region of Employment

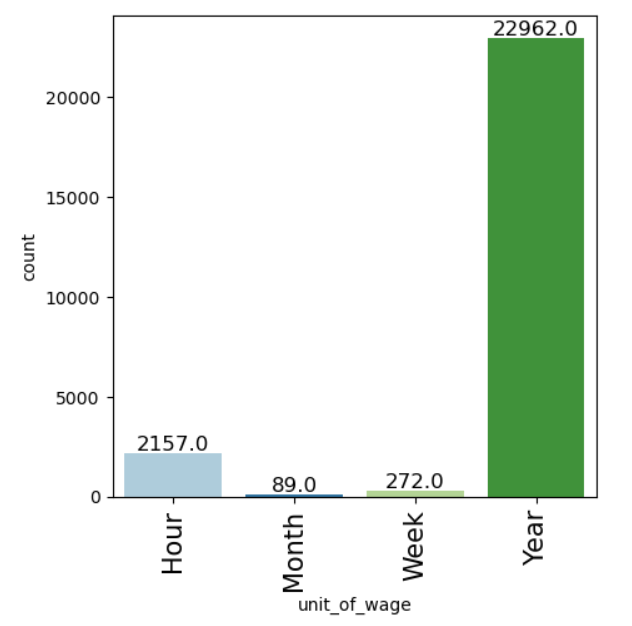


**Insights:**

* The intended region of employment for majority of the employees is the Northeast, followed by South and West.
* Island region is the least intended region for employees.

### 2.5.9 Unit of Wage

Figure 10 - Labeled Barplot of Unit of Wage

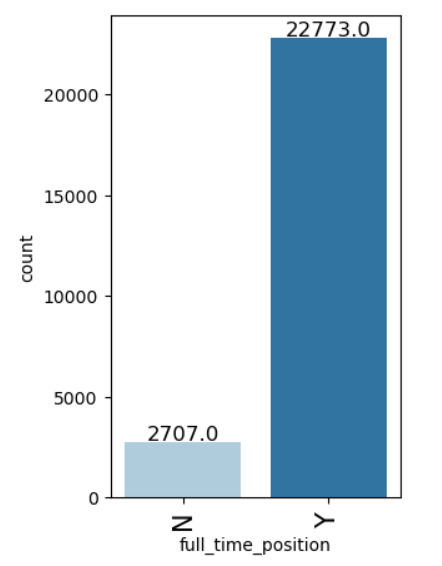


**Insights:**

* Most common unit of prevailing wage is Yearly pay.
* Monthly unit of wage is the least observed in the dataset

### 2.5.10 Full Time Position

Figure 11 - Labeled Barplot of Full Time Position

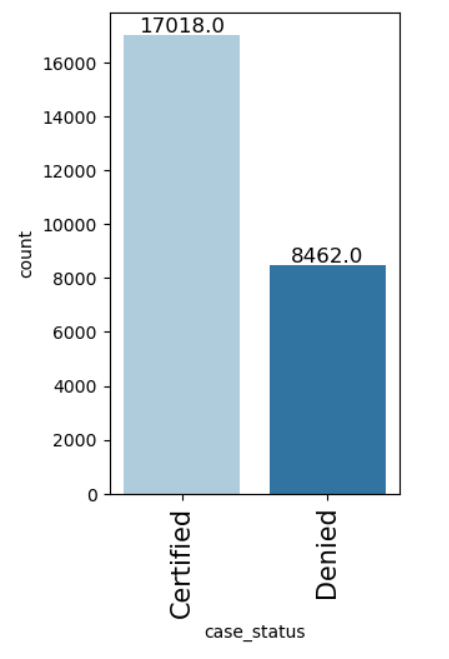


**Insights:**

* Majority of applicant employees’ position of work is full time, approximately 90% of the observations.
* 10% of the observations position of work is part time.

### 2.5.11 Case Status

Figure 12 - Labeled Barplot of Case Status



Insights:

* Majority of the applicants had their visas certified, approximately 67% of the total observations.
* 33% of the observed applicants were denied visas.

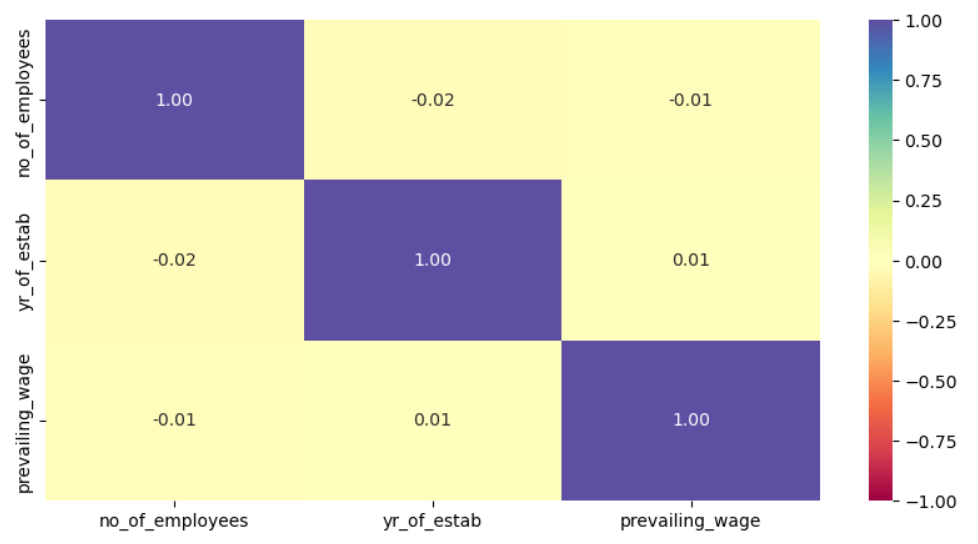
## 2.6 Bivariate Analysis

Bivariate analysis examines the relationship between two variables to understand how they interact or correlate with each other. It can identify patterns, trends, and dependencies between the variables, such as whether changes in one variable correspond to changes in another.

Utilising various charts such as heatmap etc. will help us determine if a relationship exists and provide insights into cause-and-effect dynamics or associations between variables.

### 2.6.1 Heatmap

Figure 13 - Heatmap of all Numerical Variables

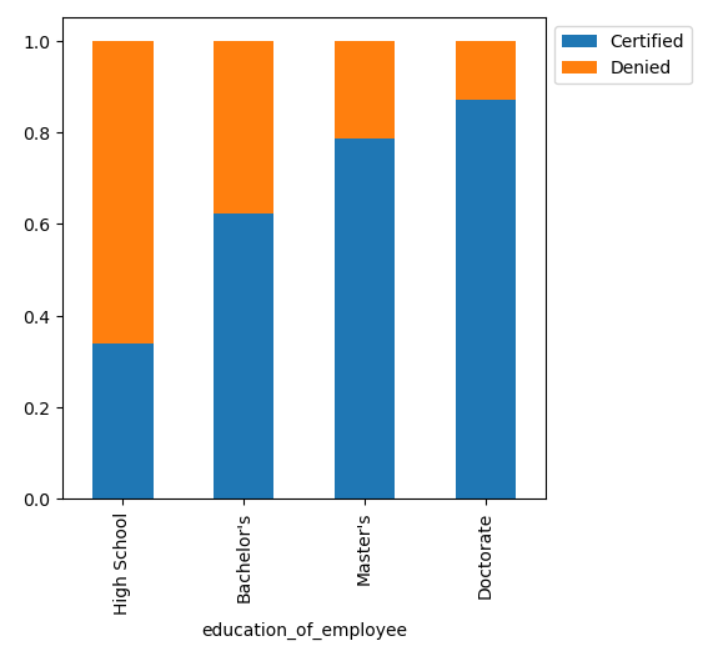


**Insights:**

* There is weak to negative correlation between the 3 numerical variables
* Majority of the dataset consists of categorical variables so any relationships need to be observed by analyzing categorical variables with each other.

### 2.6.2 Education of employee vs Case Status

Figure 14 - Stacked Barplot of Education of Employee vs Case Status

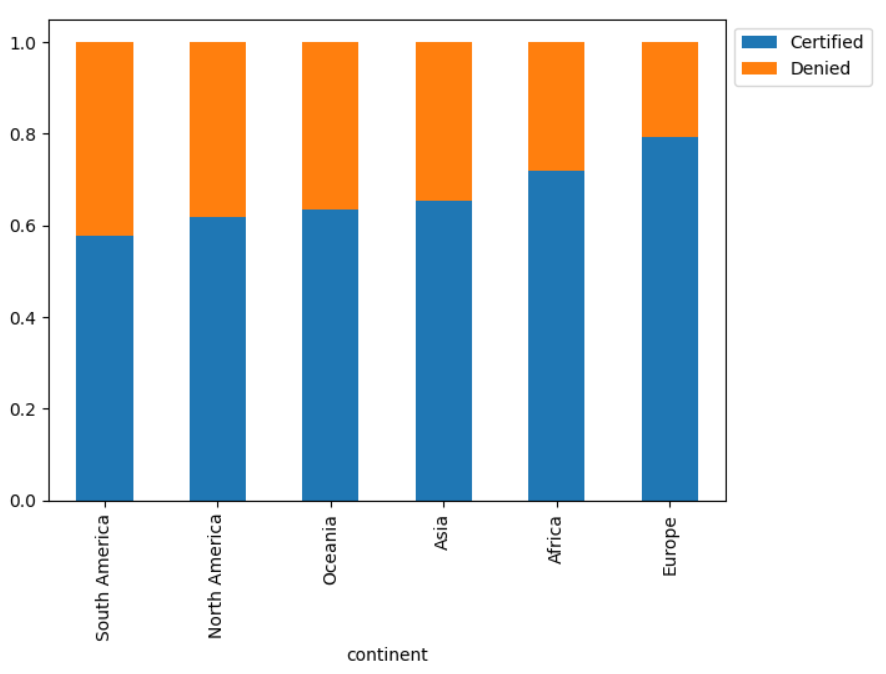


**Insights:**

* Employees with doctorate degrees have a high chance of getting visas certified, approximately 87% of doctorate degree holders have their visas certified.
* Employees with only high school degree have the least chance of getting visas certified, approximately 34% of them get certified the rest are denied visas.

### 2.6.3 Continent vs Case Status

Figure 15 - Stacked Barplot of Continent vs Case Status

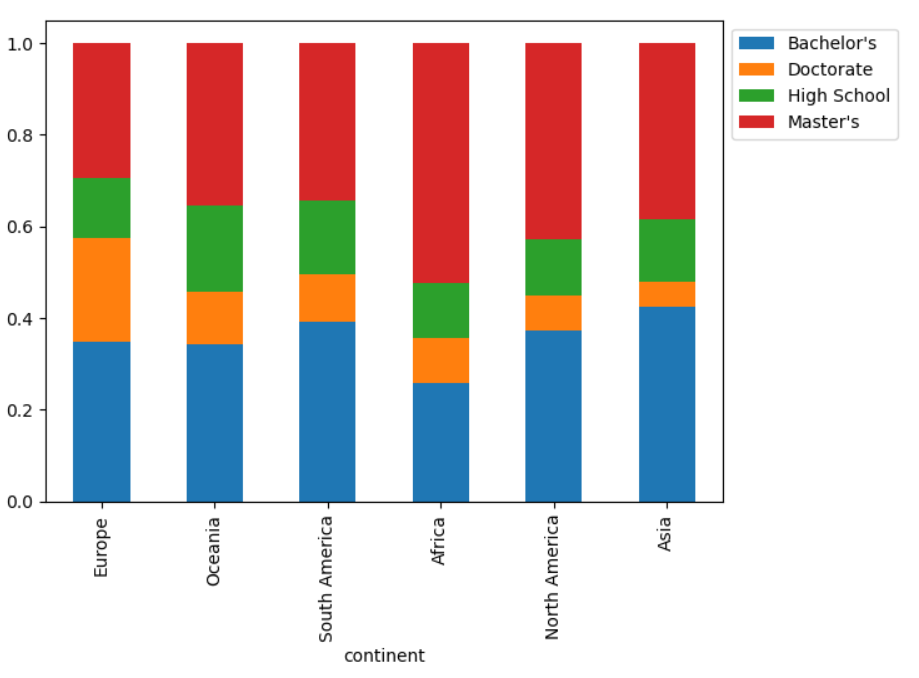


**Insights:**

* Majority of the certified visas are from Europe, approximately 78% get their visas approved from Europe.
* Africa is the second highest region that has higher chance of getting visas certified.
* North America, Oceania and Asia have similar chances of having visas certified, approximately around 60%
* South America has the least amount of certified visas.

### 2.6.4 Continent vs Education of Employee

Figure 16 - Stacked Barplot of Continent vs Education of Employee

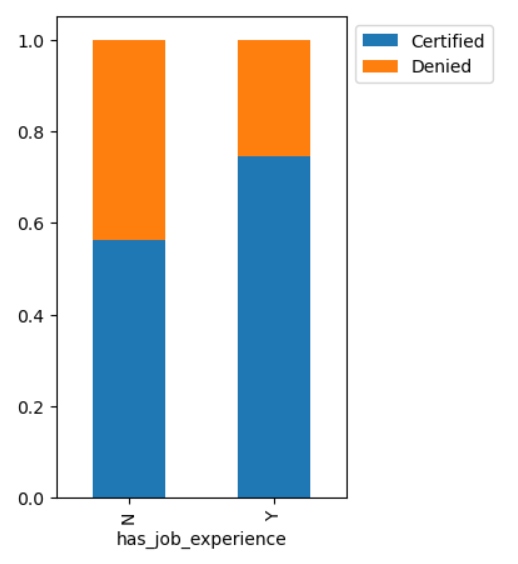


**Insights:**

* Majority of Master’s Degree holders are from Africa followed by North America and Asia. South America and Oceania have similar percent of Master Degree holders approximately around 35%. Europe has the least amount of Master’s Degree Holders.
* Asia has the highest number of applicants with Bachelor’s Degree followed by South America. Europe, Oceania and North America have a similar percentage of Bachelor’s Degree holders, approximately around 35%. Africa has the least number of applicants with Bachelor’s Degree
* Europe has the largest number of doctorate holders, approximately around 20% of total applicants from Europe compared to other regions. The smallest percentage of Doctorate holders are from Asia.
* Oceania has the highest percentage of applicants with High school degree while Africa has the least percentage.

### 2.6.5 Job Experience vs Case Status

Figure 17 - Stacked Barplot of Job Experience vs Case Status

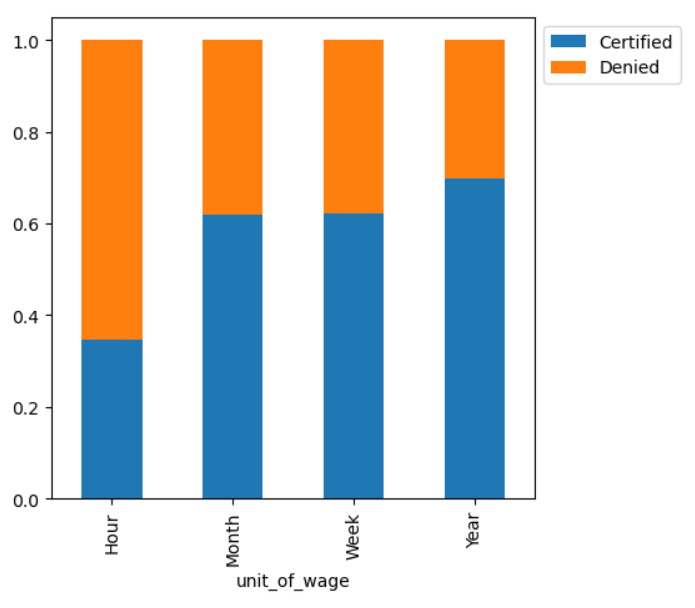


**Insights:**

* Around 40% of the applicants with no job experience have had their visas denied.
* Around 75% of the applicants with prior job experience have their visas certified, This implies having prior work experience plays a role in whether a person will have their visas certified or denied.

### 2.6.6 Unit of Wage vs Case Status

Figure 18 - Stacked Barplot of Unit of Wage vs Case Status

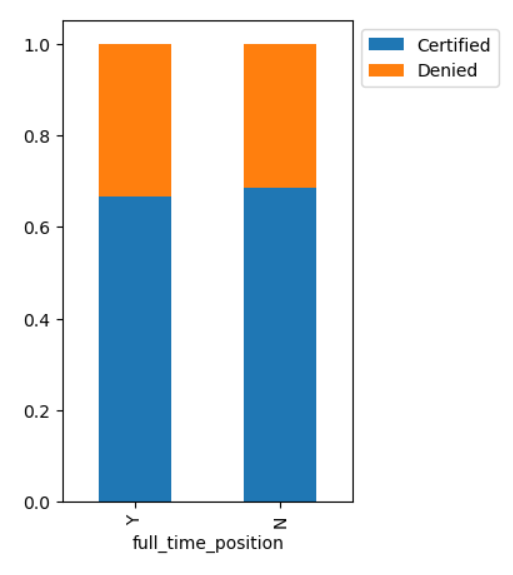


**Insights:**

* Highest number of certifications are for applicants with Yearly pay unit of wage\
* Hourly pay unit of wage has the most number of denied applicants.
* Applicants with Monthly and Weekly pay unit of wage have approximately 40% chance of being denied.

### 2.6.7 Full Time Position vs Case Status

Figure 19 - Stacked Barplot of Fulltime Position vs Case Status



**Insights:**

* Working full time or part time has similar percentages of being denied visa, approximately 35% of the applicants are denied visas.

# DATA PREPROCESSING

## 3.1 Missing Value Check

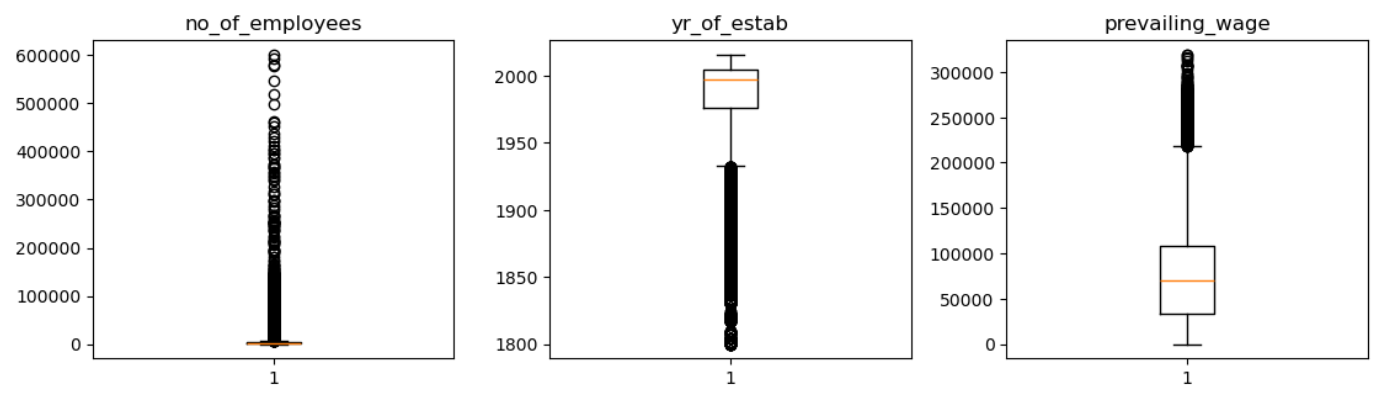
There are no missing values in the dataset therefore no treatment is required.

## 3.2 Duplicate Value Check

There are no duplicate values in the dataset therefore no treatment is required.

## 3.3 Outlier Check

Figure 20 - Boxplot of all Numerical Variables



There are outliers present in the dataset; these will be considered as genuine values as they can contribute to providing insights so these outliers will not be needing treatment.

## 3.4 Feature Engineering

Feature engineering involves transforming or creating the variables from the raw data to improve the performance of the machine learning model. For this data the values of “certified” and “denied” have been encoded as “1” and “0” respectively.

## 3.5 Data Preparation

The negative values in “no\_of\_employees” column have been turned into absolute values. i.e the negative values are transformed into positive values. All categorical variables will be encoded and then the data will be split into training set, testing set and validation set. The data is first split in 80:20 split creating a temporary data set and then the data is split further into a 75:25 split of training and validation set.

# MODEL BUILDING – ORIGINAL DATA

## 4.1 Model Evaluation Criterion

The purpose of the model is to create a machine learning based solution that can help in shortlisting the candidates having higher chances of VISA approval.

The model can make wrong predictions such as

1. Predicting an applicant that should be certified but is denied.
2. Predicting an applicant that should be denied but is certified.

It is important to reduce both the false positives and false negatives as they both can lead to negative impacts such as allowing unqualified people to take up a position they are not fit for and by denying visa to a qualified person which can reduce productivity in companies not being able to fill positions faster. The F1 score will be used as a metric for evaluation as the greater the score, the greater the chances are to minimizing false positives and false negatives.

## 4.2 Building the Models

Different models will be built to ensure that the meaningful insights from the data are extracted which will lead to better prediction of the result. The following models are built with priority to the F1 score and the most suitable will be tuned to get the best result.

1. Bagging Model
2. Random Forest Model
3. Gradient Boosting Machine Model
4. AdaBoost Model
5. Decision Tree Model
6. XGBoost Model

Figure 21 Training & Validation Performance on Original Data

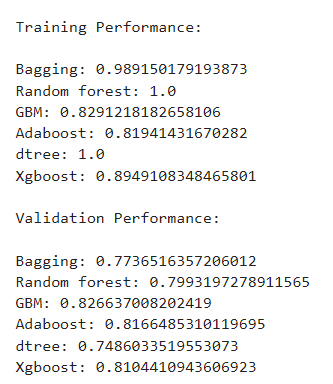
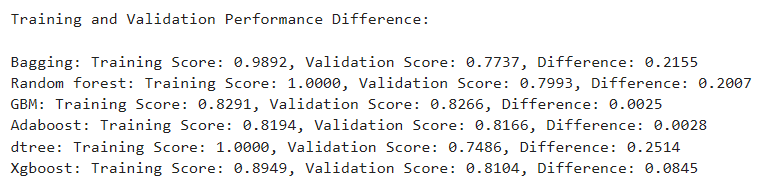


Figure 22 - Training & Validation Difference on Original Data



* GBM, Adaboost are the most suitable for further tuning due to their strong validation performance and minimal overfitting
* XGBoost has a very slightly higher difference but is showing decent performance
* Bagging, Random Forest and Decision Tree models are showing signs of overfitting.
* GBM, Adaboost and XGboost can be considered for tuning.

# MODEL BUILDING – OVERSAMPLED DATA

Model building with oversampled data refers to the process of training machine learning models using a dataset where the class distribution has been adjusted to address class imbalance through oversampling. Class imbalance occurs when one class has significantly fewer instances than the other class, which can lead to biased models that favor the majority class.

## 5.1 Building the Models

After Oversampling, The counts of Certified and Denied status are equal at 10,211.The size of the training dataset has been increased to 20,422 observations. The following models are built with priority to the F1 score and the most suitable will be tuned to get the best result.

1. Bagging Model
2. Random Forest Model
3. Gradient Boosting Machine Model
4. AdaBoost Model
5. Decision Tree Model
6. XGBoost Model

Figure 23 - Training & Validation Performance on Oversampled Data

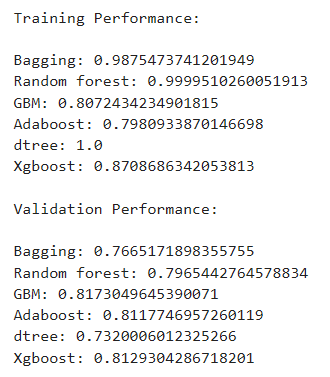
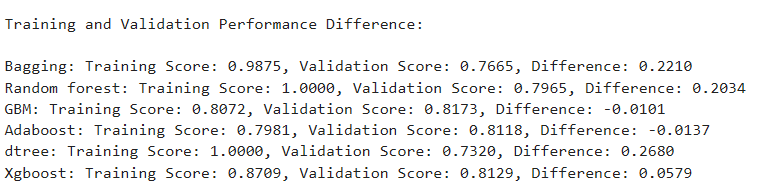


Figure 24 - Training & Validation Difference on Oversampled Data



* Gradient Boost Machine Model and Adaboost Model performs the best overall with minimal differences between training and validation performance.
* Random Forest and Bagging and Decision Tree Models show overfitting, with large performance differences between training and validation scores.
* XGBoost performs well but not as well as Gradient boost and Adaboost models

# MODEL BUILDING – UNDERSAMPLED DATA

Model building with undersampled data is a technique used to address class imbalance in a dataset by reducing the number of instances in the majority class to balance the class distribution. This approach can help improve the performance of machine learning models, especially when the dataset has a disproportionate number of instances in one class compared to the other.

## 6.1 Building the Models

After Undersampling, The counts of Certified and Denied status are equal at 5,077.The size of the training dataset has been decreased to 10,154 observations. The following models are built with priority to the F1 score and the most suitable will be tuned to get the best result.

1. Bagging Model
2. Random Forest Model
3. Gradient Boosting Machine Model
4. AdaBoost Model
5. Decision Tree Model
6. XGBoost Model

Figure 25 - Training & Validation Performance on Undersampled Data

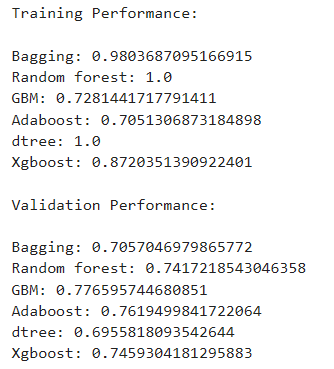
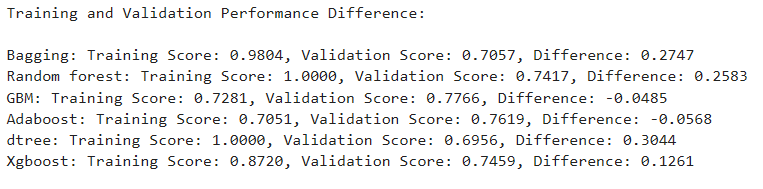


Figure 26 - Training & Validation Difference on Undersampled Data



* Gradient Boost Machine Model and Adaboost Model performs the best as it has the smallest training-validation performance gap.
* Random Forest, Bagging and Decision Tree Model shows a large performance difference, indicating possible overfitting on the training data.
* XGBoost performs well on training data but shows a moderate difference on the validation data.

# MODEL PERFORMANCE IMPROVEMENT USING HYPERPARAMETER TUNING

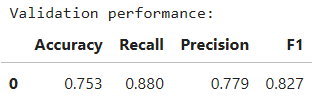
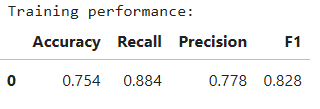
Hyperparameter tuning is the process of selecting the best set of hyperparameters to improve a machine learning model's performance. It involves methods like Grid Search, which tests all combinations of hyperparameters, and Random Search, which samples randomly. We will be using Random Search method with F1 score as the metric of interest for tuning the models.

The three models which will be tuned are

1. Gradient Boost Machine Model with Original Data
2. Adaboost Model with Original Data
3. XGBoost Model with Oversampled Data

## 6.1 Gradient Boost Machine Model with Original Data

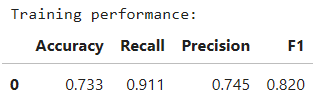
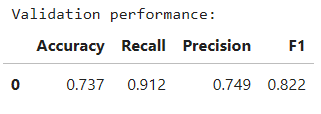
Figure 27 - Training & Validation Performance on Tuned GBM Model



After hypertuning the gradient boost machine model, we arrive at the above results. The model has strong training performance and validation performance at 0.828 and 0.827 respectively with F1 score as the primary metric.

## 6.2 Adaboost Model with Original Data

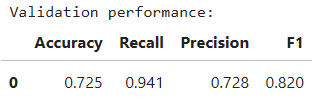
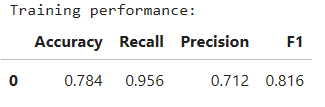
Figure 28 - Training & Validation Performance on Tuned Adaboost Model



After hypertuning the adaboost model, we arrive at the above results, The model has good performance on training and validation data with an F1 score of 0.822 and 0.820 respectively showing minimal difference between the F1 scores indicating that it is stable and well-tuned.

## 6.3 XGBoost Model with Oversampled Data

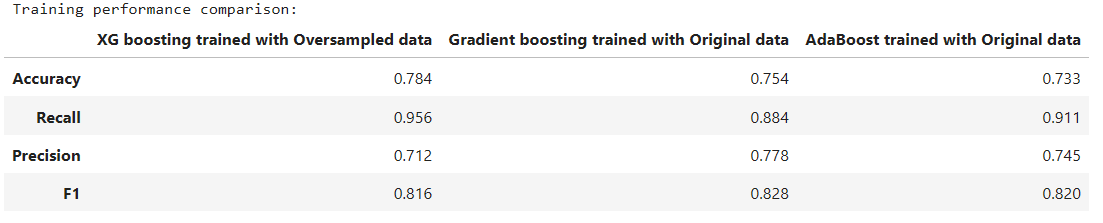
Figure 29 - Training & Validation Performance on Tuned XGBoost Model

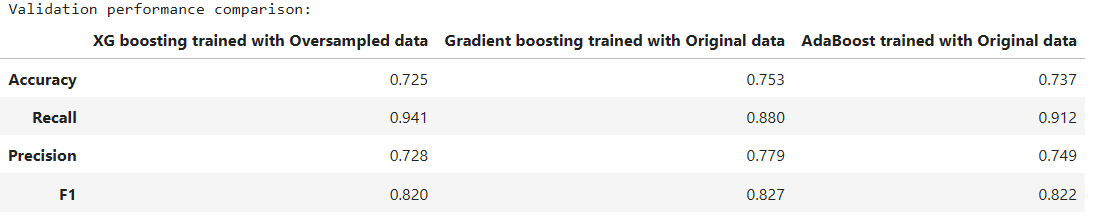


After hypertuning the xgboost model, we arrive at the above results, The model shows a high score of 0.816 on training set and a score of 0.820 on the validation set as F1 score is the primary metric. The model performs well.

# MODEL PERFORMANCE COMPARISON & FINAL MODEL SELECTION

Figure 30 - Training & Validation Model Comparison





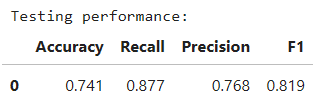
Comparing all the models we can infer the following

* XGBoost model provides the best accuracy but lower precision which leads to an increase of false positives.
* Gradient Boosting Machine model has the best balance in terms of F1 score (both training and validation), making it a strong candidate for overall performance.
* Adaboost model shows excellent recall, but its precision and F1 score are slightly lower than Gradient Boosting Machine Model.

After comparing the various tuned models, the most suitable model is the Gradient Boost Machine model which is tuned on the original data. This model shows a higher training and validation score along with high stability and does not suffer from overfitting or underfitting.

## 7.1 Performance of the Best Model on the Test Set

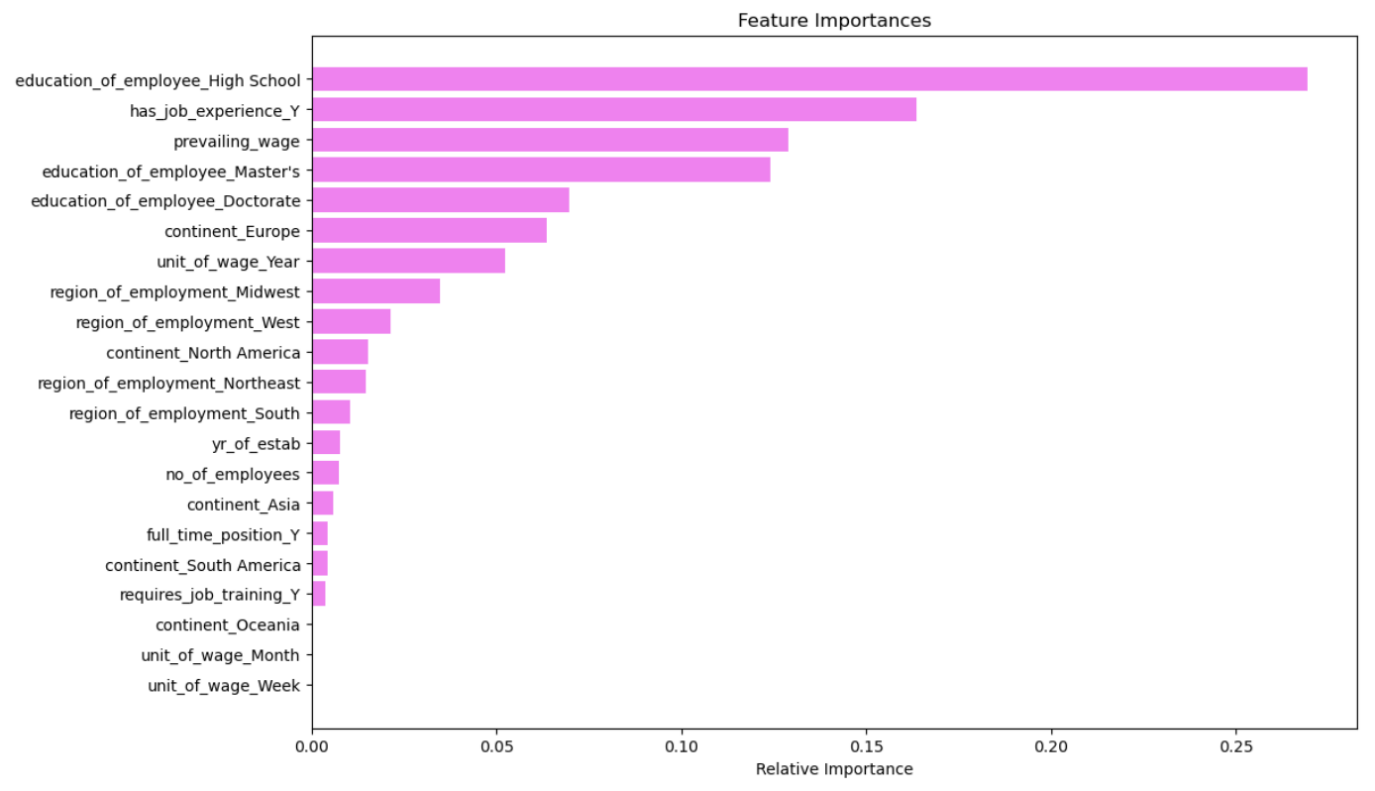
Figure 31 - Performance of GBM model on Unseen Test Data



The performance of the Gradient Boosting Machine model on the testing set shows it performing well with minimal variations in the score compared to the training and validation data. This confirms that the model is the most suitable as we have a strong prediction rate of around 82%.

## 7.2 Feature Importances

Figure 32 - Feature Importances of the GBM Model



The feature importances show that education of employee, job experience and prevailing wages are the most important factors when predicting if a visa should be certified or denied. If a person has a maximum education level of high school, the chances of denial are very high implying that masters or bachelors are required at minimum.

The least important factors are job training and monthly and weekly unit of wage.

# ACTIONABLE INSIGHTS & RECOMMENDATIONS

* Applicants with at least a Bachelor's degree are more likely to have their visa certified, with Master's and Doctorate degrees having the highest chances.
* Applicants with prior job experience are more likely to have their visa certified compared to those with no experience. It is necessary to prioritize applicants with prior job experience for shortlisting.
* Applicants with a high prevailing wage have a higher likelihood of visa certification. Those applicants with higher wages can be shortlisted in advance to streamline the process.
* Applicants from Europe, Africa, and Asia have higher chances of visa approval, while those from North and South America, and Oceania are more likely to face visa denials.
* Visa certification has a higher chance of being granted for roles in the Midwest and South regions of the US.
* High wages for salaried positions strongly influence visa approval.
* Applicants with higher education levels (minimum bachelor’s degree), relevant job experience, and competitive wages should be prioritized during screening as they are most likely to get their visa certified.
* Ensure that the applicant's educational background is clearly documented and matches the requirements of the job position to expedite the review process.
* Require detailed job experience documentation to assess the candidate’s suitability for the role, especially for positions that demand specific skills or certifications.
* Shortlist applications where the wage unit is yearly rather than hourly, as hourly wage jobs are more likely to be denied.
* Applicants from Europe, Africa and Asia who have a minimum of Bachelor’s Degree can be shortlisted in advance to streamline the process as they have the highest chances of getting approval.