Machine Learning -2
Coded Project

Business Report

DSBA – Course

Created by – Rishabh Gupta

Foreword

Context –

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.

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.

Contents

List of Tables

List of Figures

Objective

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.

Data Analysis of problem statement –

The dataset contains data corresponding to polished and unpolished stones.

Sheet name – *EasyVisa.csv*

Data Dictionary –

1. **case_id:** ID of each visa application
2. **continent:** Information of continent the employee
3. **education_of_employee**: Information of education of the employee
4. **has_job_experience:** Does the employee has any job experience? Y= Yes; N = No
5. **requires_job_training:** Does the employee require any job training? Y = Yes; N = No
6. **no_of_employees:** Number of employees in the employer's company
7. **yr_of_estab:** Year in which the employer's company was established
8. **region_of_employment**: Information of foreign worker's intended region of employment in the US.
9. **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.
10. **unit_of_wage:** Unit of prevailing wage. Values include Hourly, Weekly, Monthly, and Yearly.

11. **full_time_position:** Is the position of work full-time? Y = Full-Time Position; N = Part-Time Position
12. **case_status**: Flag indicating if the Visa was certified or denied

Categorization of variables –

Continuous Variables - no_of_employees, yr_of_estab, prevailing_wage

Categorical Variables - case_id, continent, education_of_employee, has_job_experience , requires_job_training, region_of_employment, unit_of_wage, full_time_position, case_status

Data Overview –

To start the data analysis, we need to import the necessary libraries, specify the working directory, and load the dataset. Then, we will view the first five rows using head () function to get a preliminary understanding of the data. The Dataset has 25480 number of rows with 12 columns.

	case_id	continent	education_of_employee	has_job_experience	requires_job_training	no_of_employees	yr_of_estab	region_of_employment	prevailing_wage	unit_of_v
0	EZYV01	Asia	High School	N	N	14513	2007	West	592.20290	
1	EZYV02	Asia	Master's	Y	N	2412	2002	Northeast	83425.65000	
2	EZYV03	Asia	Bachelor's	N	Y	44444	2008	West	122996.86000	
3	EZYV04	Asia	Bachelor's	N	N	98	1897	West	83434.03000	
4	EZYV05	Africa	Master's	Y	N	1082	2005	South	149907.39000	

TABLE 1 - TOP 5 ROWS OF DATASET

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25480 entries, 0 to 25479
Data columns (total 12 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   case_id                25480 non-null  object
 1   continent              25480 non-null  object
 2   education_of_employee  25480 non-null  object
 3   has_job_experience     25480 non-null  object
 4   requires_job_training  25480 non-null  object
 5   no_of_employees        25480 non-null  int64
 6   yr_of_estab            25480 non-null  int64
 7   region_of_employment   25480 non-null  object
 8   prevailing_wage        25480 non-null  float64
 9   unit_of_wage           25480 non-null  object
 10  full_time_position     25480 non-null  object
 11  case_status            25480 non-null  object
dtypes: float64(1), int64(2), object(9)
memory usage: 2.3+ MB
```

TABLE 2 - BASIC INFO. OF DATASET

Regarding the datatype of the columns in the dataset, we can confirm that-

• We can observe that there around 9 object datatypes columns and 3 numerical datatypes

• There are no missing values in the dataset

Missing value treatment and Analysis-

- On analysis, we can observe there are no null values in the dataset.
- Also, there are no duplicate entries.
- We have dropped column ‘case_id’ as it will have no effect on predictive analysis.

Statistical Summary –

Using Describe () function, we can analyses the summary statistics of the dataset –

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
case_id	25480	25480	EZYV01	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
continent	25480	6	Asia	16861	NaN	NaN	NaN	NaN	NaN	NaN	NaN
education_of_employee	25480	4	Bachelor's	10234	NaN	NaN	NaN	NaN	NaN	NaN	NaN
has_job_experience	25480	2	Y	14802	NaN	NaN	NaN	NaN	NaN	NaN	NaN
requires_job_training	25480	2	N	22525	NaN	NaN	NaN	NaN	NaN	NaN	NaN
no_of_employees	25480.00000	NaN	NaN	NaN	5667.04321	22877.92885	-26.00000	1022.00000	2109.00000	3504.00000	602069.00000
yr_of_estab	25480.00000	NaN	NaN	NaN	1979.40993	42.36693	1800.00000	1976.00000	1997.00000	2005.00000	2016.00000
region_of_employment	25480	5	Northeast	7195	NaN	NaN	NaN	NaN	NaN	NaN	NaN
prevailing_wage	25480.00000	NaN	NaN	NaN	74455.81459	52815.94233	2.13670	34015.48000	70308.21000	107735.51250	319210.27000
unit_of_wage	25480	4	Year	22962	NaN	NaN	NaN	NaN	NaN	NaN	NaN
full_time_position	25480	2	Y	22773	NaN	NaN	NaN	NaN	NaN	NaN	NaN
case_status	25480	2	Certified	17018	NaN	NaN	NaN	NaN	NaN	NaN	NaN

TABLE 3 - STATISTICAL SUMMARY OF DATASET

Observations-

Categorical Variables:

- Continent: 6 categories, with Asia being the most frequent.
- Education of Employee: 4 categories, with Bachelor's most common.
- Has Job Experience: Binary (Y/N), with Y (Yes) most frequent.
- Requires Job Training: Binary (Y/N), with N (No) most common.
- Region of Employment: 5 regions, with Northeast most common.
- Unit of Wage: 4 units, with Year most frequent.
- Full-Time Position: Binary (Y/N), with Y (Full-Time) most common.
- Case Status: Binary (Certified/Denied), with Certified most frequent.

Numerical Variables:

- No. of Employees: Mean = 5,667; Max = 602,069.
- Yr. of Estab: Ranges from 1800 to 2016.
- Prevailing Wage: Range from 2.13 to 319,210.

Other Insights:

- Asia is the most common continent of origin for applicants.
- Most applicants hold a Bachelor's degree.
- Many applicants have job experience and do not require job training.
- The Northeast region is the most popular for employment.
- The majority of applications are for full-time positions.
- Certified visas are more common than denied ones, indicating a high approval rate.

Exploratory Data Analysis

Lets investigate the dataset –

- We are checking the negative values in the number of employees, so that we can handle them accordingly.

  Here we got to know that we have 33 cases like this and hence, we will take the absolute value of those values.(Plz refer code).

- Now we will check the unique values in categorical columns.

```
continent
Asia             16861
Europe            3732
North America     3292
South America      852
Africa             551
Oceania            192
Name: count, dtype: int64
--------------------------
education_of_employee
Bachelor's     10234
Master's        9634
High School     3420
Doctorate       2192
Name: count, dtype: int64
--------------------------
has_job_experience
Y    14802
N    10678
Name: count, dtype: int64
--------------------------
requires_job_training
N    22525
Y     2955
Name: count, dtype: int64
--------------------------
region_of_employment
Northeast    7195
South        7017
West         6586
Midwest      4307
Island        375
Name: count, dtype: int64
```

```
unit_of_wage
Year     22962
Hour      2157
Week       272
Month       89
Name: count, dtype: int64
--------------------------
full_time_position
Y    22773
N     2707
Name: count, dtype: int64
--------------------------
case_status
Certified    17018
Denied        8462
Name: count, dtype: int64
```

TABLE 4 – CATEGORICAL COLUMNS

- **Continent:**

  The majority of applicants come from Asia.

- **Education Level of Employees:**
  Most applicants hold a Bachelor's degree, with a notable number also having a Master's degree.
- **Job Experience:**
  A significant portion of applicants possess job experience.
- **Job Training Requirement:**
  The vast majority of applicants do not require job training.
- **Region of Employment:**
  The Northeast and South regions have the highest number of applicants.
- **Wage Unit:**
  Most wages are reported on an annual basis.
- **Employment Type:**
  The majority of positions are full-time.
- **Case Status:**
  There are more certified cases compared to denied ones.

Numerical Variables - A histogram for visualizing the distribution of numerical –

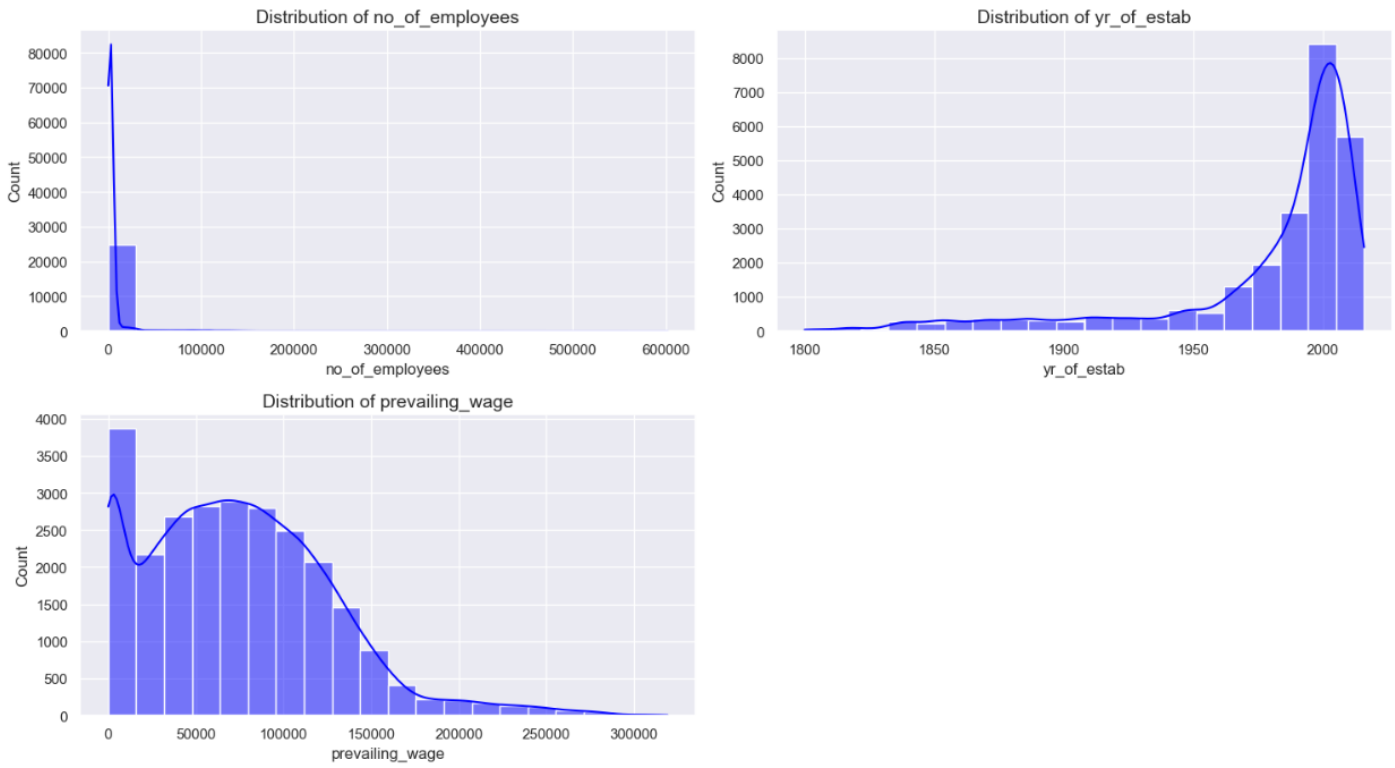


Figure 1 - Histogram plot for Numerical variables

- Distribution of no_of_employees:
    - The distribution is highly skewed to the right, indicating that most companies have a relatively small number of employees.
    - There are a few companies with a very large number of employees, which might be outliers or represent large corporations.
- Distribution of yr_of_estab:
    - The distribution is also skewed to the right, but less so than the number of employees.
    - This suggests that a majority of companies were established in recent years, with a smaller number of older companies.
- Distribution of prevailing_wage:
    - The distribution of prevailing wage is also right-skewed, indicating that most positions have a lower wage, with a few positions having significantly higher wages.

The dataset appears to be skewed towards smaller companies with lower prevailing wages.

Further, lets boxplots for numerical variables for deeper analysis –

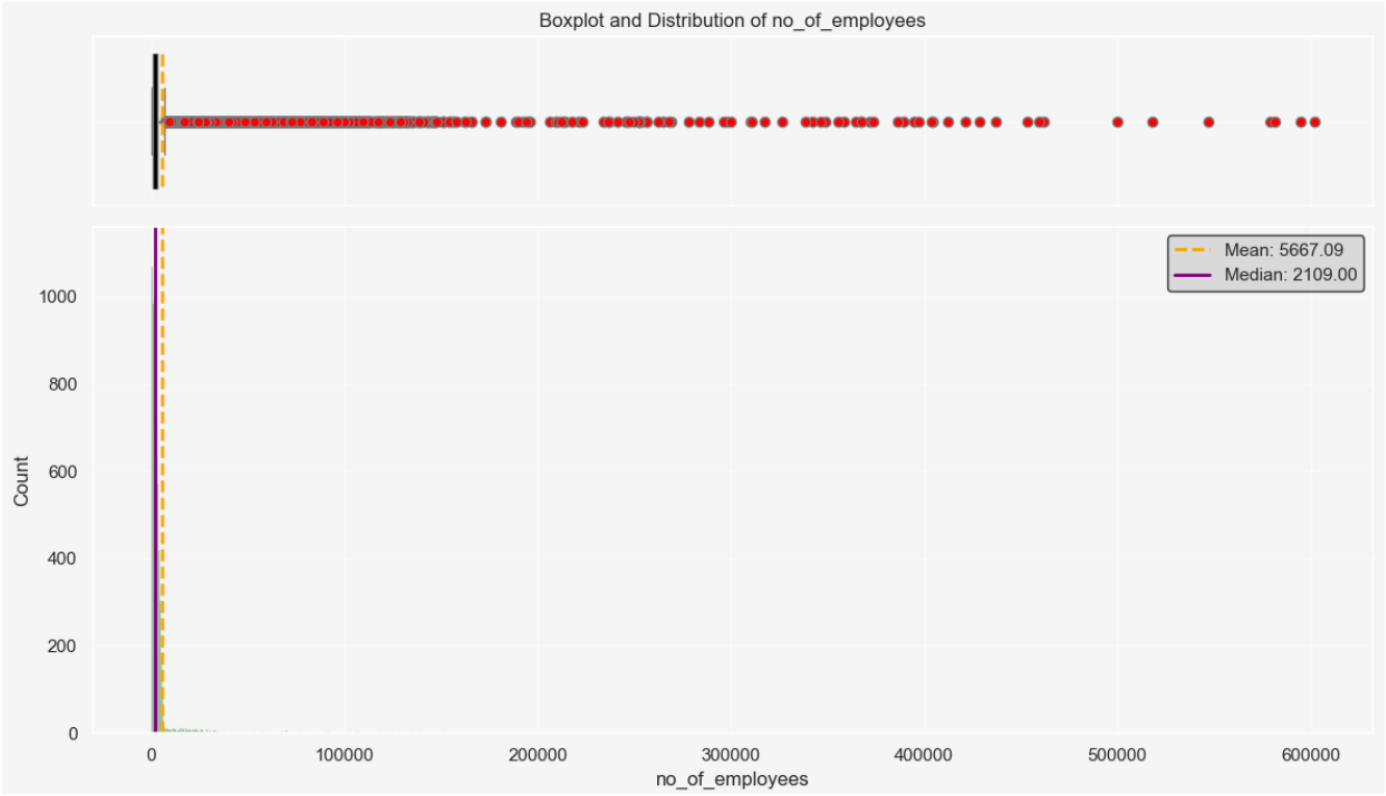


Figure 2 - Boxplot for no_of_employee

Observations –

- Boxplot:
  - Median: Half the companies have fewer than 2109 employees.
  - Mean: Average is higher due to the influence of outliers.
  - Right-skewed: Long right whisker indicates outliers.
- Histogram:
  - Confirms right-skewed distribution.
  - Peak around lower employee counts.
  - Long tail towards higher counts.

- Mix of small and large companies.
- Outliers likely represent large corporations.
- Mean is influenced by outliers; median might be a better measure of central tendency.

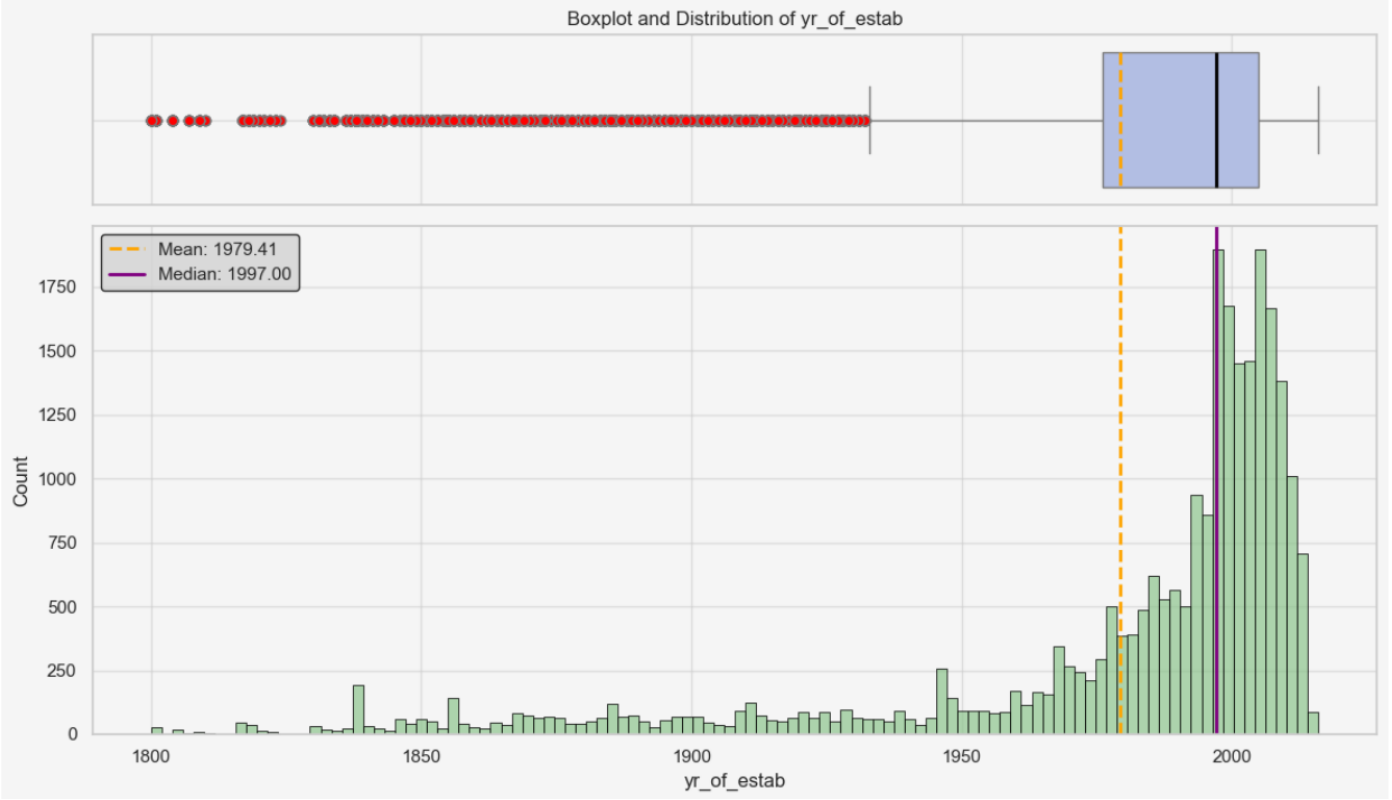


Figure 3 - Boxplot for yr_of_estab

- Boxplot:
  - Median: Most companies were established around 1997.
  - Mean: The average year of establishment is slightly earlier at 1979.41.
  - Outliers: There are a few older companies that are considered outliers.

- Histogram:
  - Confirms the right-skewed distribution.
  - A peak around the recent years, indicating a higher frequency of newer companies.
  - A long tail towards the left, representing older companies.
- The dataset likely contains a mix of old and new companies.
- The right-skewed distribution suggests that the industry has seen significant growth in recent years.

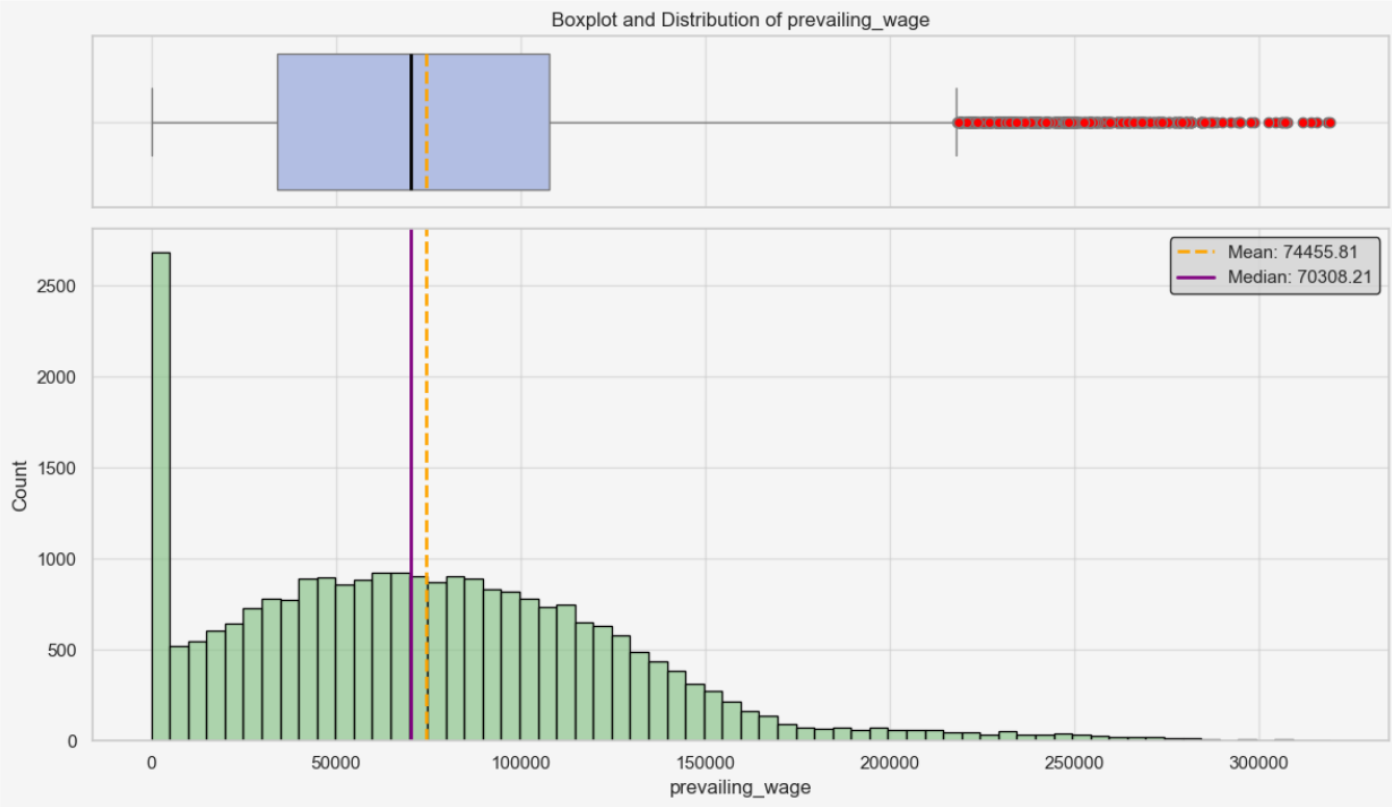


Figure 4 - Boxplot for prevailing_wage

- Boxplot:
  - Median: Half of the positions have a prevailing wage below $70,308.21.
  - Mean: The average wage is slightly higher at $74,455.81, likely influenced by a few high-wage positions.
  - Outliers: There are some high-wage outliers, suggesting a few positions with exceptionally high salaries.

- Histogram:
    - Confirms the right-skewed distribution.
    - A peak around the lower wage range.
    - A long tail towards higher wages.

- The dataset likely contains a mix of low and high-paying positions.
- The right-skewness suggests that most positions have lower wages, with a few high-paying exceptions.

Further for deeper understanding, lets barplot for Categorical variables-

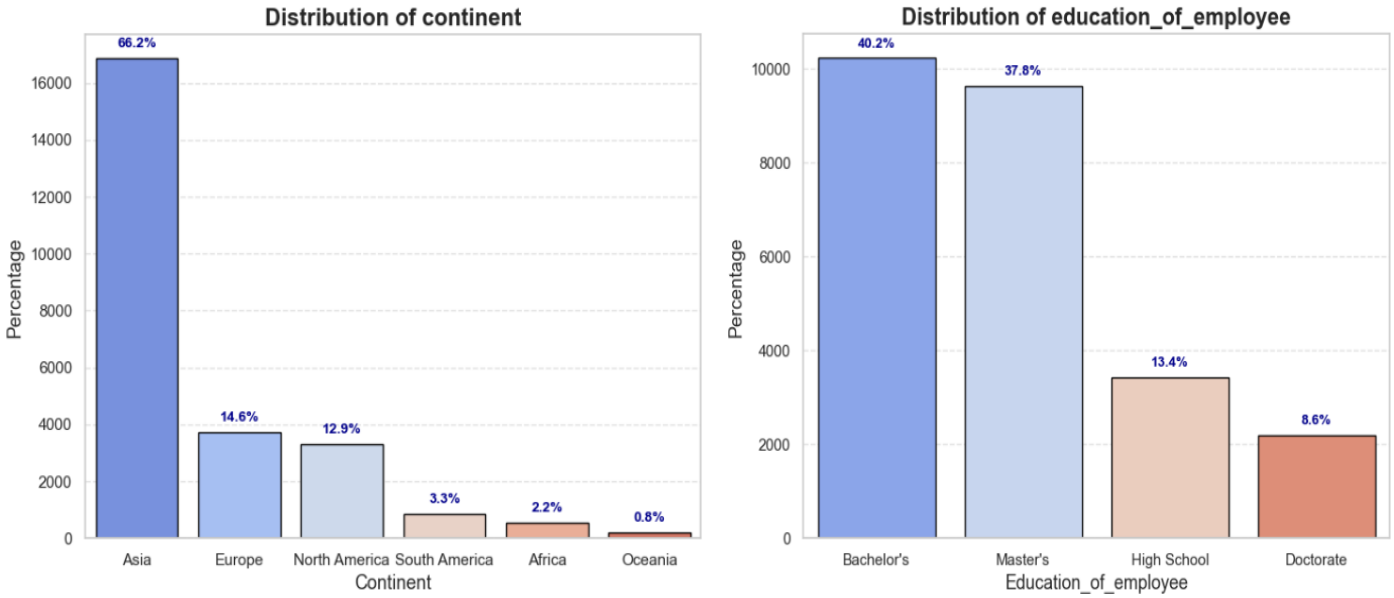


Figure 5 - Barplot for Content and Education variables

- Asia: The bar for Asia is significantly taller than the others, indicating a dominant presence.
- Europe and North America: These continents have the next highest representation, though still significantly lower than Asia.
- South America, Africa, and Oceania: These continents have the smallest representation in the data.
- Bachelor's and Master's: These two categories have the highest representation, indicating a preference for higher education qualifications in the workforce.
- The organization may prioritize higher education qualifications for its workforce.

- High School and Doctorate: These categories have a much smaller proportion of employees, suggesting that while some positions may require a high school diploma, a significant number of roles require higher education.

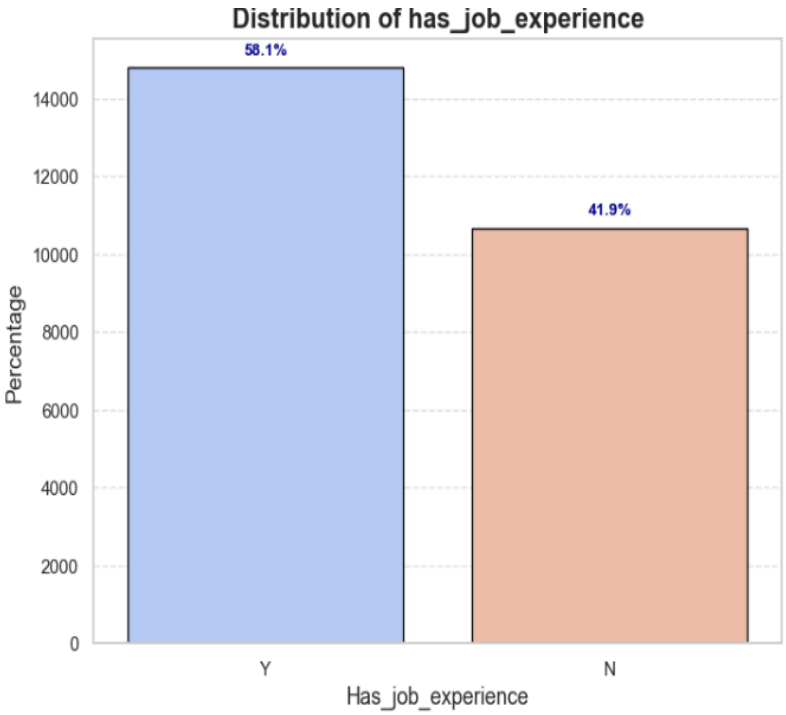


Figure 6 - Barplot for Job exp

- The distribution is almost evenly split between individuals with and without prior job experience.
- There is a slight majority of individuals who have prior job experience.
- With Job Experience: Approximately 58.1% of the individuals have prior job experience.
- Without Job Experience: Approximately 41.9% of the individuals do not have prior job experience.

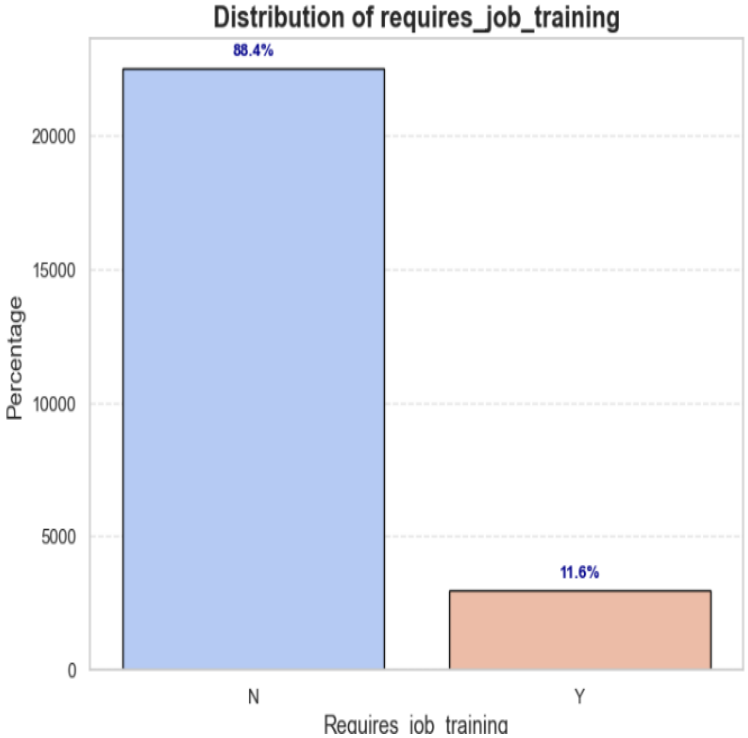
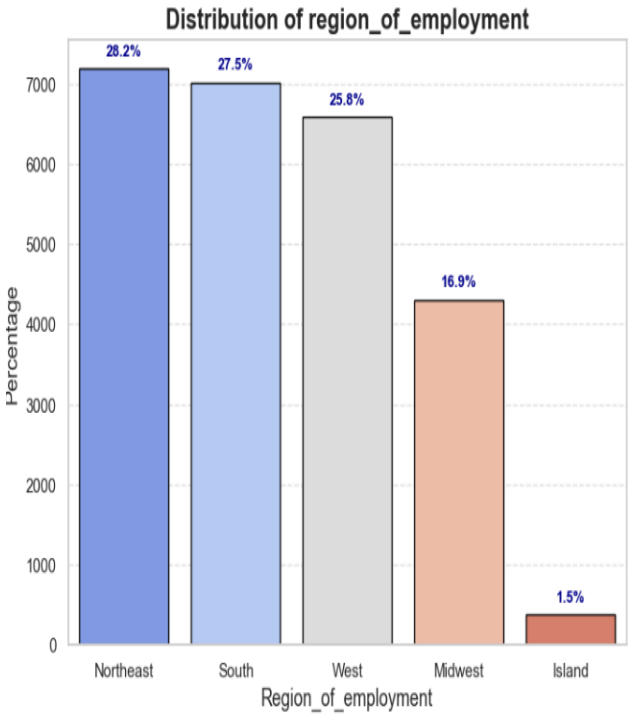


Figure 7 - Barplot for Job training

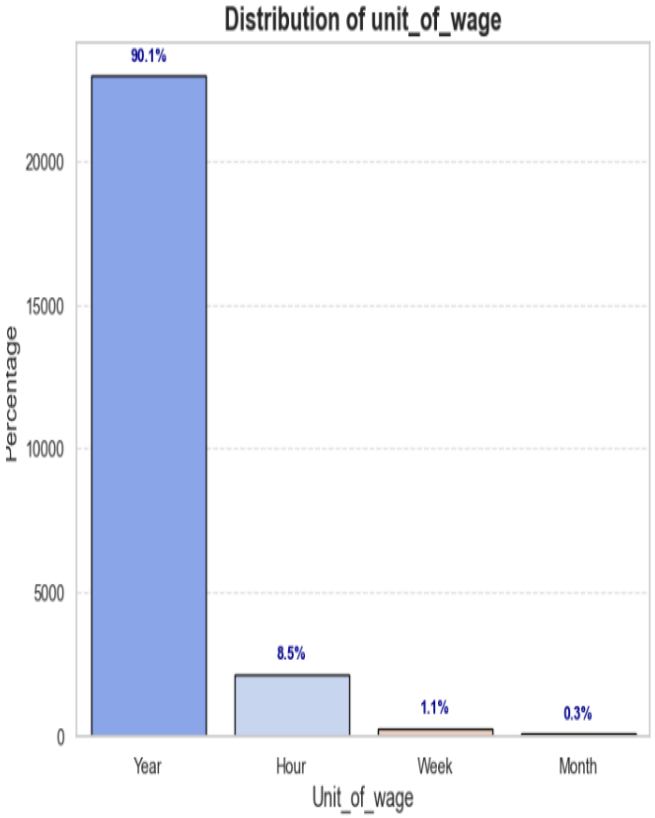
- The distribution is heavily skewed towards jobs that do not require job training.
- A small percentage of jobs require job training.
- No Job Training Required: Approximately 88.4% of the jobs do not require any formal job training.
- Job Training Required: Only 11.6% of the jobs require some form of job training.
- The dataset may represent a field or industry where many jobs have low barriers to entry, requiring minimal or no formal training.



The Northeast and South regions have slightly higher representation compared to the West, Midwest, and Island regions.

- Northeast: Approximately 28.2% of the jobs are located in the Northeast region.
- South: Around 27.5% of the jobs are located in the South region.
- West: Approximately 25.8% of the jobs are located in the West region.
- Midwest: Around 16.9% of the jobs are located in the Midwest region.
- Island: Only 1.5% of the jobs are located in the Island region.

Figure 8 - Barplot for Region of Employment

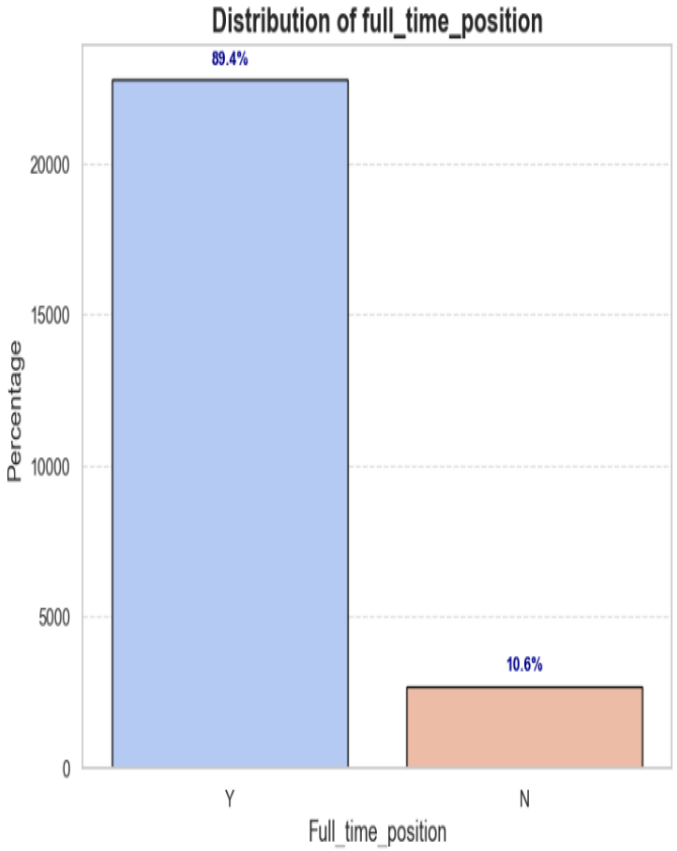


The distribution is heavily skewed towards the "Year" unit, with a significant majority of the data using this unit.

- The other units, "Hour," "Week," and "Month," have a much smaller representation.

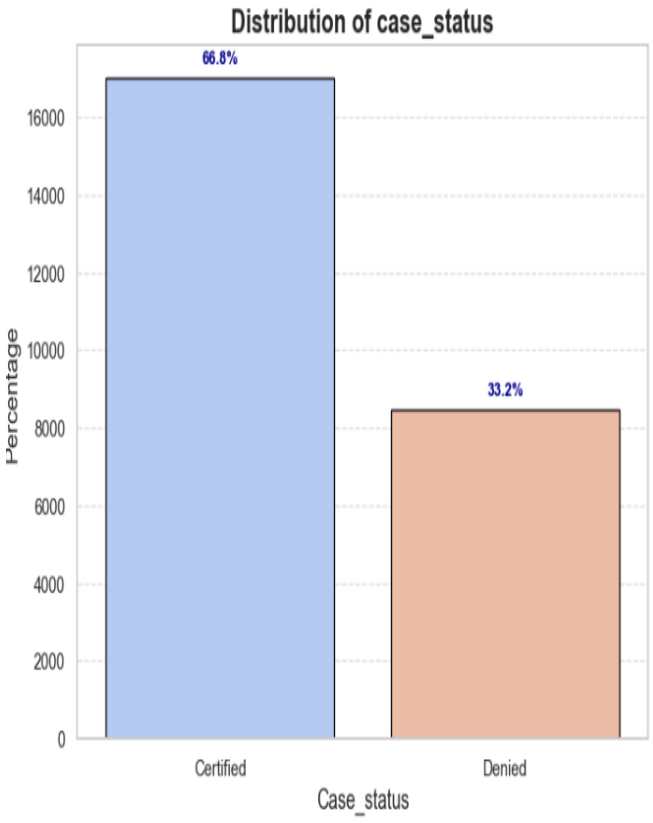
- Year: Approximately 90.1% of the wage units are based on a yearly basis.
- Hour: Around 8.5% of the wage units are based on an hourly basis.
- Week: Approximately 1.1% of the wage units are based on a weekly basis.
- Month: Only 0.3% of the wage units are based on a monthly basis.

Figure 9 - Barplot for Unit for Wage



- The distribution is heavily skewed towards full-time positions.
- A small percentage of the positions are not full-time.
- Full-Time: Approximately 69.4% of the positions are full-time.
- Part-Time/Contract: Around 10.6% of the positions are not full-time.

Figure 10 - Barplot for Full time position



- The distribution is skewed towards "Certified" cases.
- A smaller proportion of cases are "Denied".

- Certified: Approximately 66.8% of the cases have been certified.
- Denied: Around 33.2% of the cases have been denied.

Figure 11 - Barplot for Case studies

Examining the distribution of the prevailing wage with respect to the target variable (case_status) and analyzing outliers –

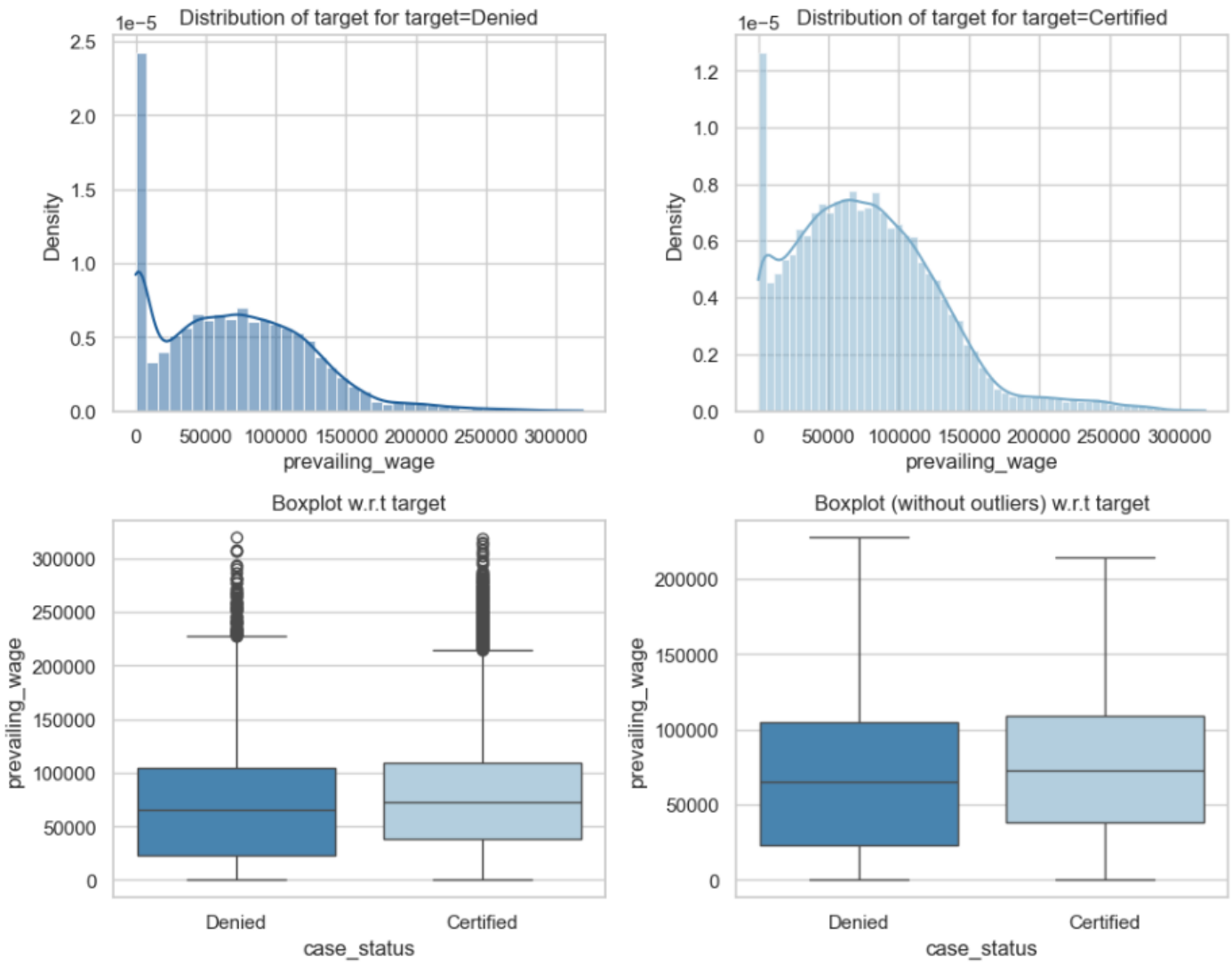


Figure 12 – Plot for prevailing_wage distribution with outliers

Observations:

Right-skewed distribution: Both the "Denied" and "Certified" cases exhibit a right-skewed distribution for prevailing wage. This indicates that a majority of cases have lower prevailing wages, with a few cases having significantly higher wages.

Denied Cases:

- Density Plot: The density plot shows a peak around lower wage values, with a long tail extending towards higher wages. This suggests that a significant proportion of denied

cases have lower prevailing wages, and there are a few cases with exceptionally high wages.

- Boxplot: The boxplot shows a wide range of prevailing wages, with a few outliers on the higher end. The median wage for denied cases is lower compared to certified cases.

Certified Cases:

- Density Plot: The density plot also shows a right-skewed distribution, with a peak around lower wage values. However, the peak is slightly higher and more pronounced compared to denied cases, indicating a higher concentration of cases with moderate wages.

- Boxplot: The boxplot shows a similar pattern to denied cases, with a wide range of wages and a few outliers. However, the median wage for certified cases is higher compared to denied cases.

- Prevailing Wage as a Factor: Prevailing wage appears to be a factor influencing case outcomes. Cases with higher prevailing wages might have a higher likelihood of being certified, possibly due to factors like job complexity, required skills, or market demand.
- Outliers: The presence of outliers in both categories suggests that there might be a few cases with exceptionally high wages, which could be due to specific job roles, industries, or geographical locations.

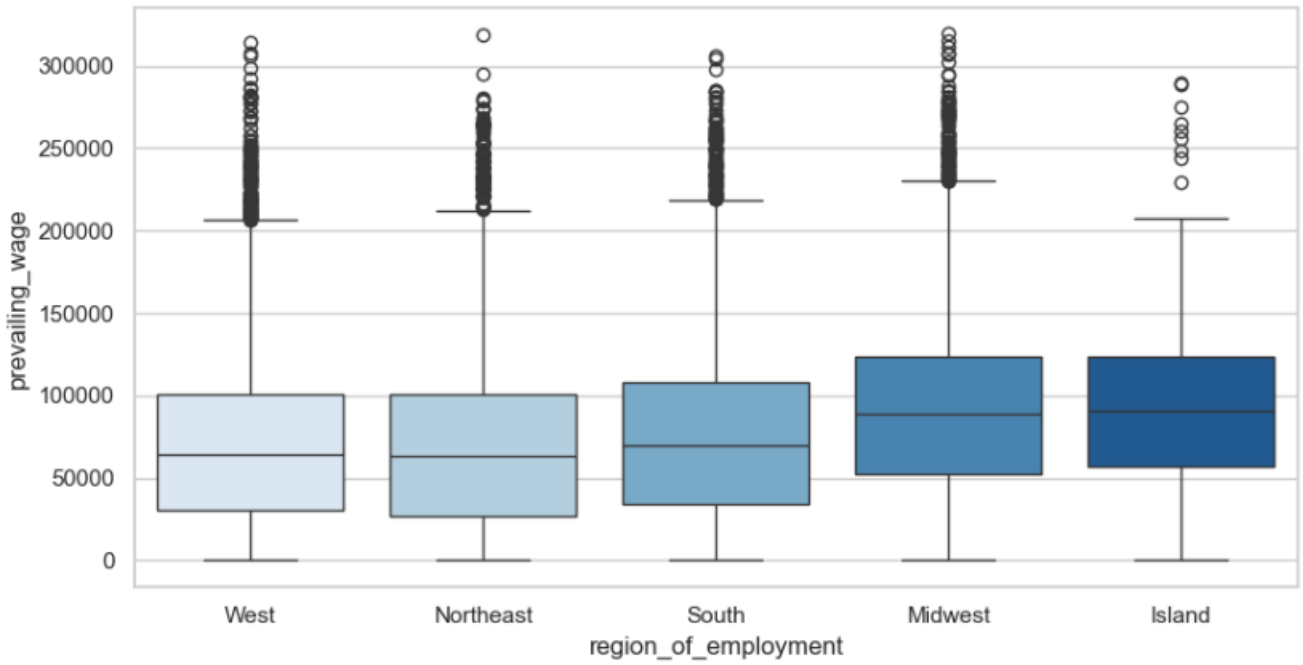


Figure 13 – Boxplot for prevailing wage distribution across different regions of employment

- Right-Skewed Distribution: The distribution of prevailing wages is skewed to the right in all regions. This indicates that a majority of jobs have lower wages, with a few high-paying positions.

- Regional Variations:
  - Northeast and Midwest: These regions have a higher median wage compared to other regions.
  - West and South: These regions have a lower median wage compared to the Northeast and Midwest.
  - Island: The Island region has the lowest median wage and a relatively narrower range of wages.

- The presence of outliers in all regions suggests that there are a few positions with exceptionally high wages, which could be due to specific job roles, industries, or geographical locations.

Here's crosstab distribution for Education of employee vs region of their employment –

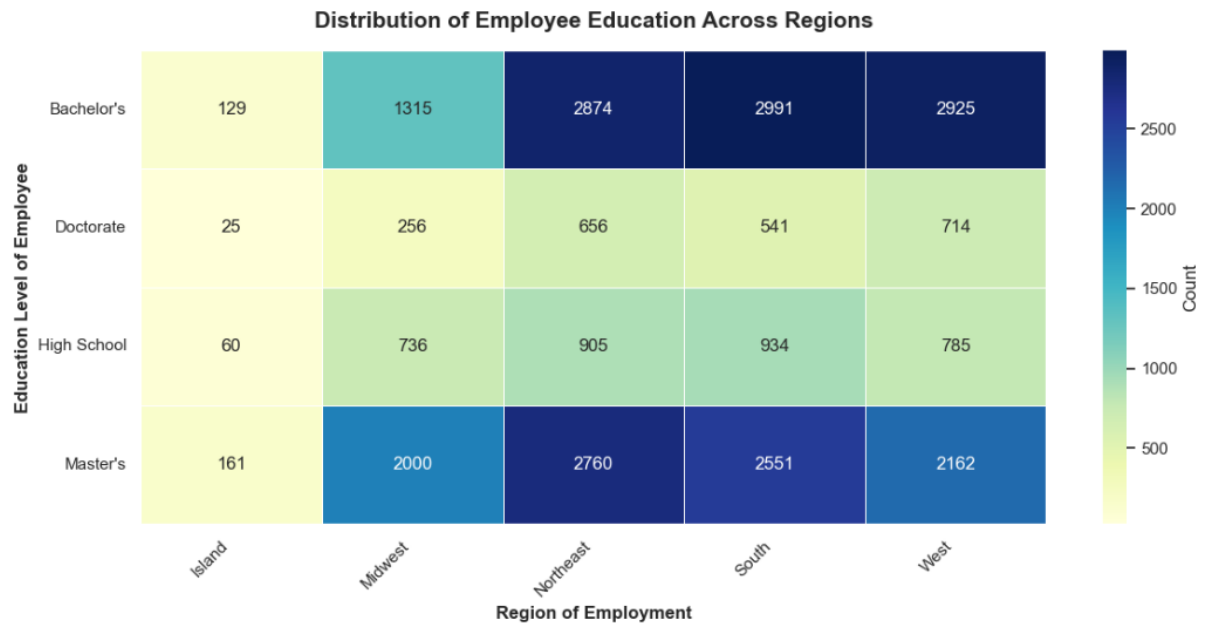


Figure 14 – CrossTab for Education vs Region of employment

- Region-wise Distribution:

  - Northeast and South: These regions have a higher concentration of employees, particularly those with Master's and Bachelor's degrees.
  - Midwest and West: These regions also have a significant number of employees, with a similar distribution across education levels.

- Island: This region has the lowest number of employees across all education levels.

- Education Level Distribution:
  - Master's and Bachelor's: These two levels have the highest number of employees across all regions, indicating a preference for higher education qualifications.
  - High School and Doctorate: These levels have a lower number of employees, suggesting that while some positions may require a high school diploma, a significant number of roles require higher education.

Here's stacked barplot for Case_studies and Region of employment –

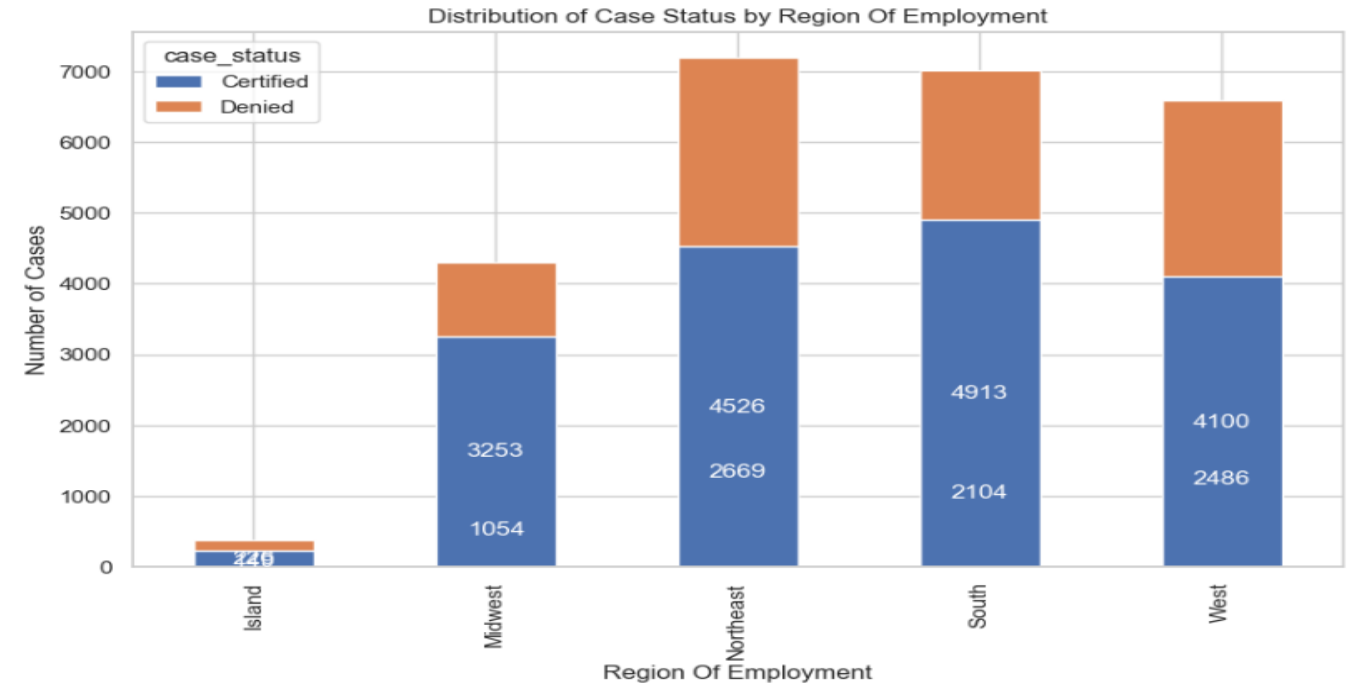


Figure 15 – Barplot for case_studies vs Region of employ

- Northeast and South: These regions have a higher number of both certified and denied cases compared to other regions. This suggests a higher volume of cases processed in these regions.
- Midwest and West: These regions have a moderate number of cases, with a relatively balanced distribution between certified and denied cases.
- Island: This region has the lowest number of cases, with a majority of them being certified.

Possible Interpretations:

- The higher number of cases in the Northeast and South regions could be attributed to factors such as a higher population density, a larger number of businesses, or a higher concentration of industries.
- The lower number of cases in the Island region might be due to various factors, including geographical constraints, industry concentration, or specific labor market dynamics.
- The relatively balanced distribution of certified and denied cases in some regions might indicate a consistent review process and decision-making criteria.

Here's stacked barplot for Case_studies and Region of employment –

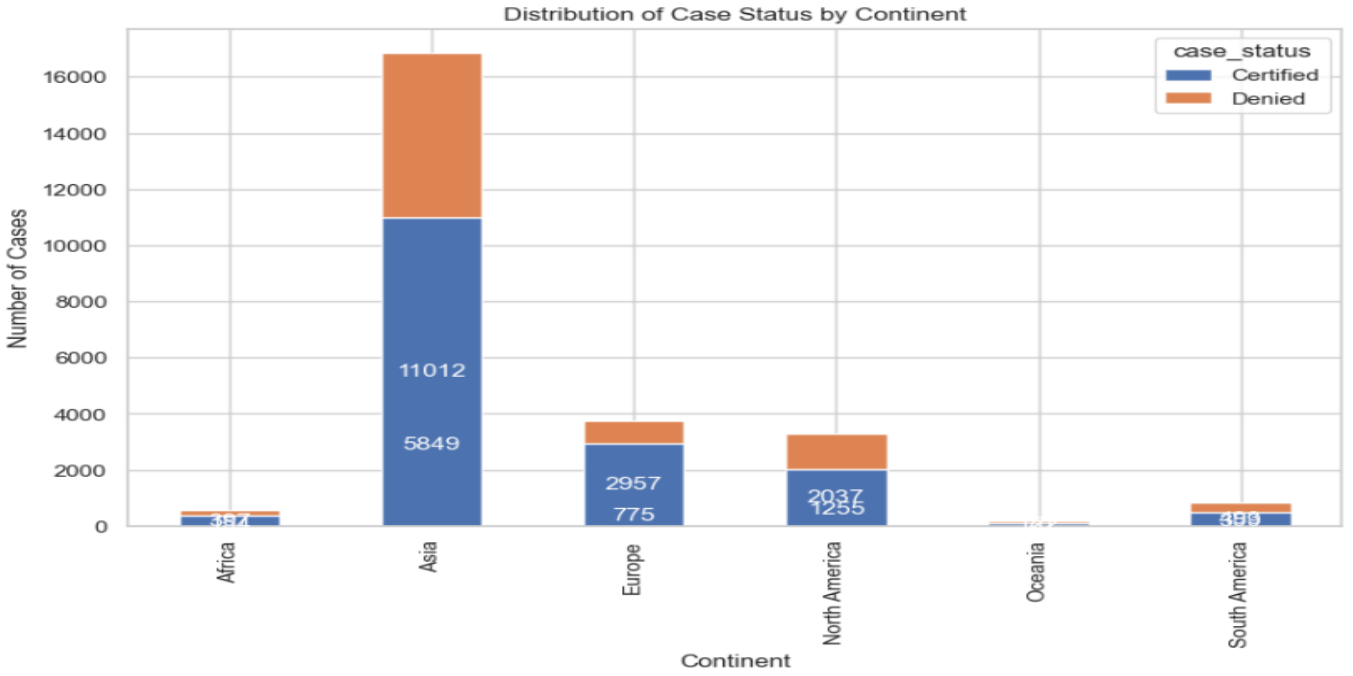


Figure 16 – Barplot for Case_studies vs Continent

- Asia: This continent has the highest number of both certified and denied cases, indicating a high volume of cases processed in this region.
- North America and Europe: These continents have a significant number of cases, with a relatively balanced distribution between certified and denied cases.
- Africa, Oceania, and South America: These continents have a lower number of cases, with a majority of them being certified.
- The higher number of cases in Asia might be attributed to factors such as a large population, a growing economy, and a high demand for skilled labor.

- The lower number of cases in certain continents could be due to various factors, such as economic conditions, immigration policies, or data availability.

Here's stacked barplot for Case_studies and Region of employment –

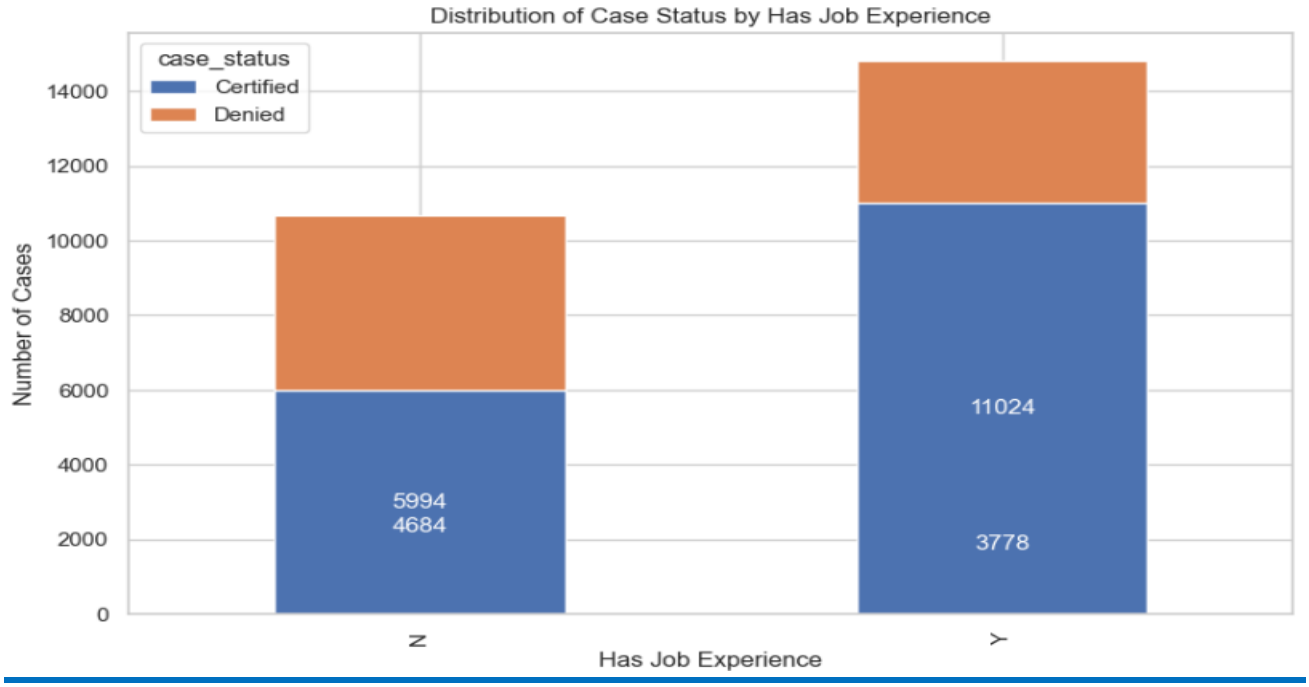


Figure 17 – Barplot for Case_studies vs Job_expreience

- Individuals with Job Experience: This group has a higher number of both certified and denied cases compared to those without job experience.
- Individuals without Job Experience: This group has a lower number of cases, with a majority of them being denied.
- Experience as a Factor: Having prior job experience might increase the likelihood of a case being certified, possibly due to factors like relevant skills, work experience, or employer references.
- Skill Gap: Individuals without job experience might face challenges in meeting the specific requirements of certain job roles, leading to a higher denial rate.

Here's stacked barplot for Case Status by Job Experience and Training Requirement–

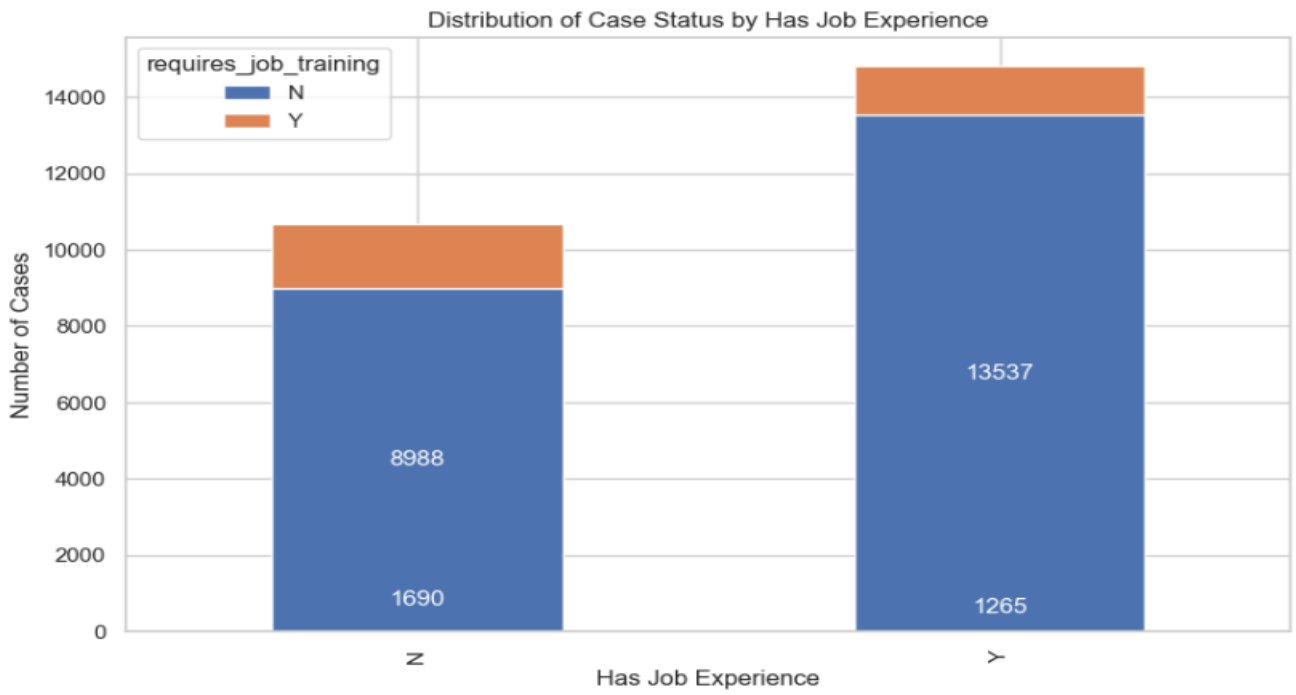


Figure 18 – Barplot for Case Status by Job Experience and Training Requirement

- Individuals with Job Experience:
  - A significant portion of individuals with job experience do not require job training.
  - The number of certified cases is higher than denied cases for this group.

- Individuals without Job Experience:
  - A majority of individuals without job experience require job training.
  - The number of denied cases is higher than certified cases for this group.

- Job Experience and Training: Individuals with prior job experience are more likely to be certified, possibly due to their existing skills and knowledge. Those without prior experience may require additional training to meet the job requirements.

- Training Requirement: Jobs that require training might have stricter eligibility criteria or specific skill sets, leading to a higher rejection rate for individuals without the necessary qualifications.

Here's stacked barplot for Case Status w.r.t unit by wage -

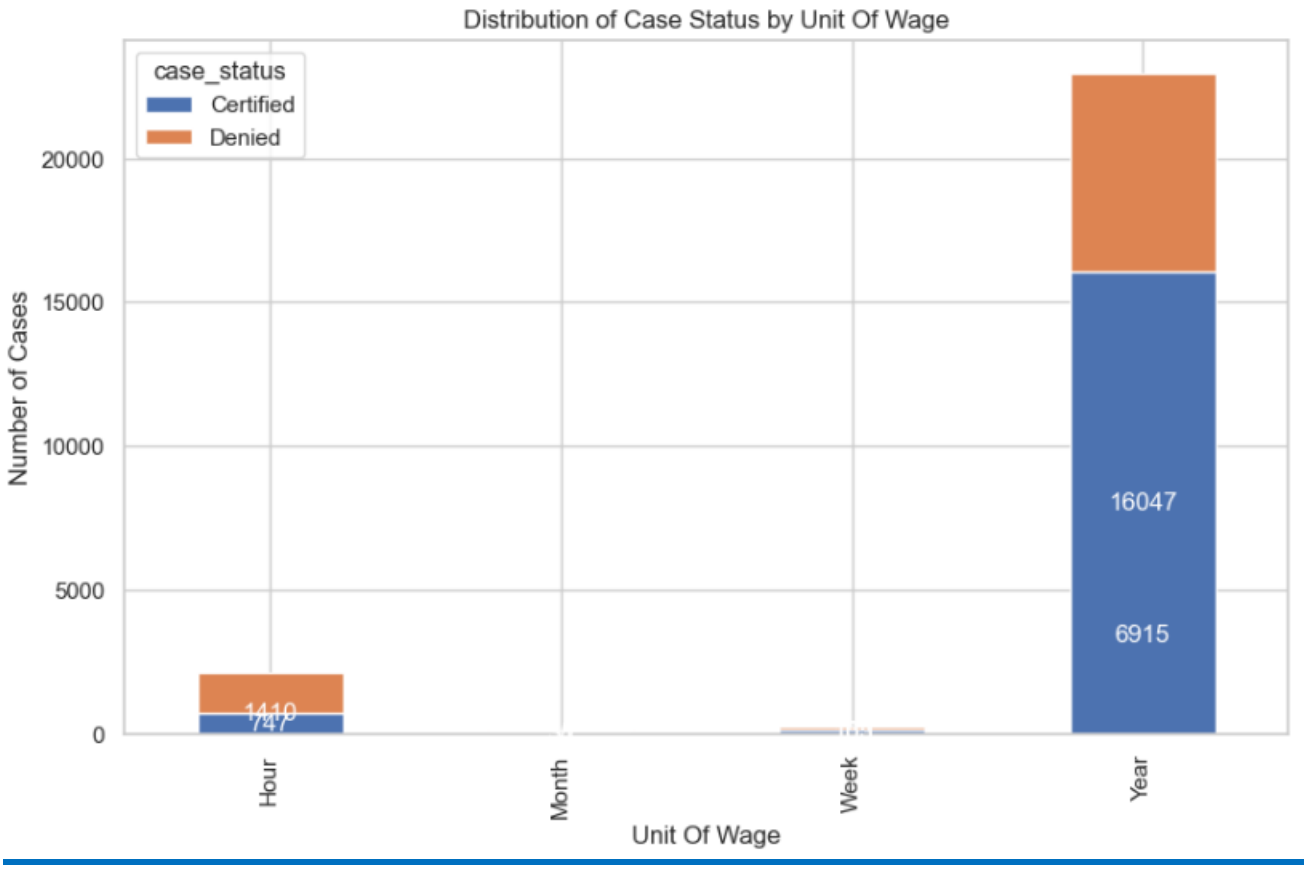


Figure 19 – Barplot for Case Status w.r.t unit by wage

- Year: This unit of wage has the highest number of both certified and denied cases, indicating a significant proportion of cases with annual salary information.

- Hour, Month, and Week: These units have a much lower number of cases, suggesting that most cases involve annual salaries rather than hourly, weekly, or monthly wages.

- Data Availability: The prevalence of annual salary data might be due to data collection practices or industry standards.

- Job Types: The use of hourly, weekly, or monthly wages might be more common in specific industries or for certain types of jobs, such as part-time or temporary positions.

Data Preprocessing

We want to predict effectiveness of Visa approvals and potential candidates. Before we proceed to build a model, we'll have to encode categorical features.

We'll split the data into train and test to be able to evaluate the model that we build on the train data.

- Missing value treatment (if needed) – there are no missing values (refer page 8)
- Feature engineering (if needed)
- Outlier detection and treatment (if needed)
- Preparing data for modeling

Outlier Treatment –

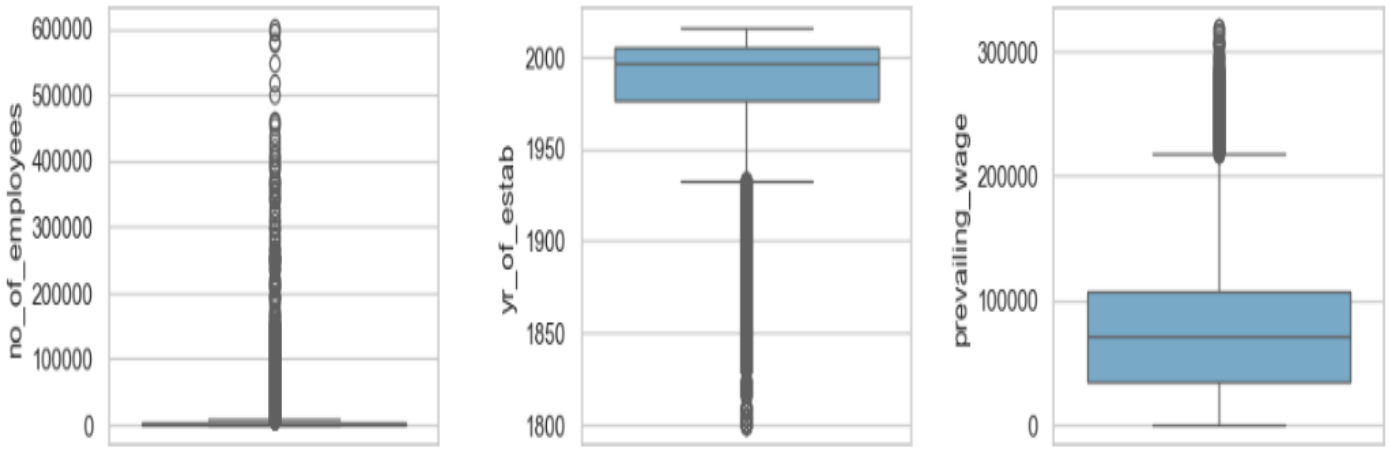


Figure 20 – Outlier summary

Observations:

- no_of_employees:
  - The distribution is heavily right-skewed, with a long tail towards higher values.
  - This indicates that most companies have a relatively small number of employees, while a few companies have a significantly larger number of employees.

- yr_of_estab:
  - The distribution is concentrated around the recent years, with a few older companies.
  - The boxplot shows a range of establishment years, with some outliers representing very old companies.

- prevailing_wage:
  - The distribution is also right-skewed, with a long tail towards higher values.
  - This indicates that most positions have lower wages, while a few positions have significantly higher wages.

- The right-skewed distribution of no_of_employees and prevailing_wage suggests that the dataset might contain a mix of small and large companies, as well as low and high-paying positions.

- There are no alarming outliers but we will keep for evaluation.

Model Building

- Our goal is to predict which visas will be certified.
- Before building the model, we need to encode the categorical features.
- We will split the data into training and testing sets to evaluate the performance of the model built on the training data.

```
Shape of Training set :  (17836, 21)
Shape of test set :  (7644, 21)
Percentage of classes in training set:
case_status
1   0.66792
0   0.33208
Name: proportion, dtype: float64
Percentage of classes in test set:
case_status
1   0.66784
0   0.33216
Name: proportion, dtype: float64
```

TABLE 5 – SPLIT DATA IN TRAINING AND TEST SETS

Observations -

- Training Set:
  - Shape: (17836, 21) - This means the training set has 17,836 rows (samples) and 21 columns (features).

  Class Distribution:

  - Class 1 (likely "Certified"): 66.792%
  - Class 0 (likely "Denied"): 33.208%

- Test Set:
  - Shape: (7644, 21) - This means the test set has 7,644 rows (samples) and 21 columns (features).

  Class Distribution:

  - Class 1 (likely "Certified"): 66.784%

    - Class 0 (likely "Denied"): 33.216%

- Both the training and test sets have a similar number of features (21).
- The class distribution is almost identical in both sets, with a slight majority of cases belonging to class 1.
- This indicates that the dataset is reasonably balanced, with a similar proportion of cases in both classes.

Model Evaluation Criterion

The model can make incorrect predictions in two key scenarios:

1. The model predicts that the visa application will be certified, but in reality, it should be denied.
2. The model predicts that the visa application will be denied, but in reality, it should be certified.

Which scenario is more critical?

Both scenarios are significant because:

- Certifying a visa application that should be denied could result in an unqualified individual taking the job, depriving U.S. citizens of potential employment opportunities.
- Denying a visa application that should be certified would cause the U.S. to miss out on valuable talent that could contribute to the economy.

How can we minimize these errors?

We can use the F1 Score as the evaluation metric for the model. A higher F1 Score indicates a better balance between minimizing False Negatives and False Positives.

Additionally, we will use balanced class weights to ensure the model gives equal attention to both classes.

Decision Tree - Model Building and Hyperparameter Tuning

We are going to create, configure, and train a decision tree classifier using the scikit-learn library

```
DecisionTreeClassifier
DecisionTreeClassifier(random_state=1)
```

Figure 21 – Decision Tree

Now let's check model performance on training set –

	Accuracy	Recall	Precision	F1
0	1.00000	1.00000	1.00000	1.00000

Figure 22 – Model performance on training set

- High Accuracy: The model correctly classified 100% of the cases.
- Perfect Precision and Recall: The model correctly identified all positive and negative cases.
- Excellent F1-Score: The harmonic mean of precision and recall is 1.0, further confirming the model's strong performance.

Now let's check model performance on test set –

	Accuracy	Recall	Precision	F1
0	0.66484	0.74280	0.75223	0.74749

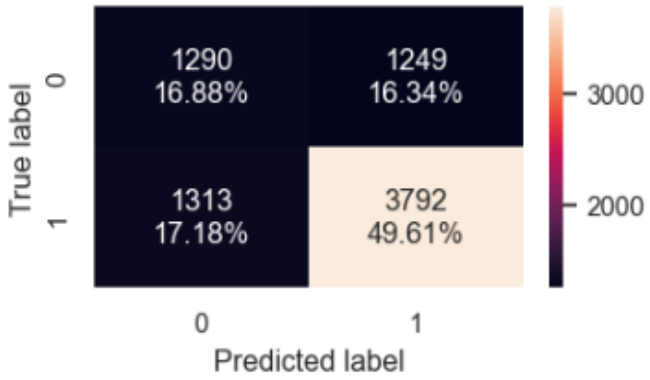


Figure 23 – Model performance on test set

- Confusion Matrix Analysis:
  - True Positives (Bottom Right): The model correctly predicted 3,792 positive cases (49.61% of total predictions).
  - False Positives (Top Right): The model incorrectly predicted 1,249 cases as positive when they were actually negative (16.34%).
  - False Negatives (Bottom Left): There were 1,313 cases incorrectly predicted as negative when they were actually positive (17.18%).
  - True Negatives (Top Left): The model correctly predicted 1,290 negative cases (16.88%).

- Performance Metrics:
  - Accuracy (66.48%): The proportion of total predictions (both true positives and true negatives) that were correct.
  - Recall (74.28%): The model's ability to correctly identify positive cases, indicating how well it minimizes false negatives.
  - Precision (75.22%): Indicates the correctness of positive predictions, showing how well it avoids false positives.
  - F1 Score (0.74749): Represents the harmonic mean of precision and recall, balancing both aspects of model performance.

Overall the **Decision Tree Classifier demonstrates signs of overfitting**, achieving perfect scores on the training data but showing significantly lower performance on the test data.

The **feature importance plot** provides insights into the relative importance of different features in predicting the target variable (likely "case_status").

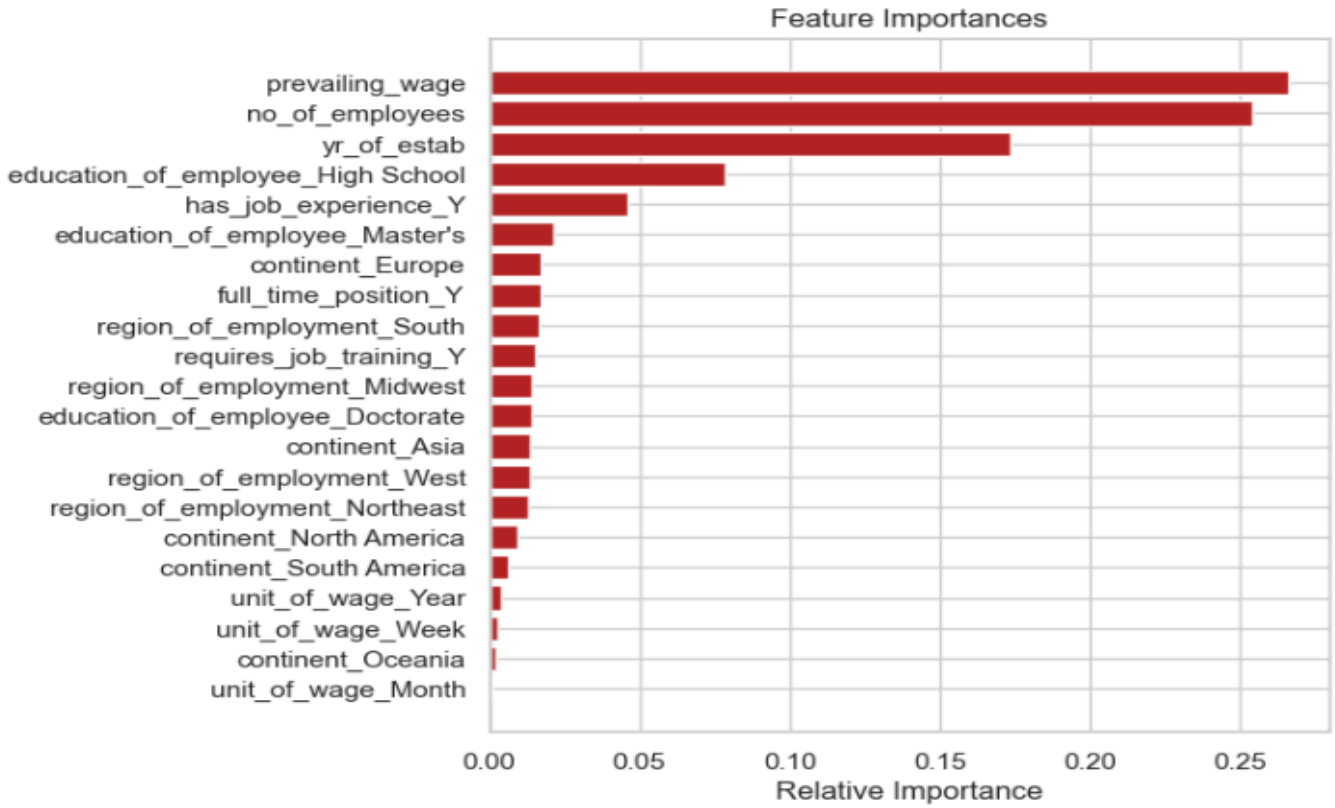


Figure 24 – Feature importance plot

Key Features:

- Prevailing Wage: This feature has the highest importance, indicating that it plays a significant role in determining the case status.
- Number of Employees: This feature also has a high importance, suggesting that the size of the company might be a relevant factor.
- Year of Establishment: The year of establishment seems to be a moderately important factor.
- Education of Employee: Features related to the employee's education level, such as "High School" and "Master's," have moderate importance.

Hyperparameter Tuning - Decision Tree

```
DecisionTreeClassifier
DecisionTreeClassifier(class_weight='balanced', max_depth=10, max_leaf_nodes=2,
                       min_impurity_decrease=0.0001, min_samples_leaf=3,
                       random_state=1)
```

Figure 25 – Hyper metric tuning

Now let's plot and analyze Confusion matrix for **train data** on tuned estimator –

	Accuracy	Recall	Precision	F1
0	0.71255	0.93192	0.72007	0.81241

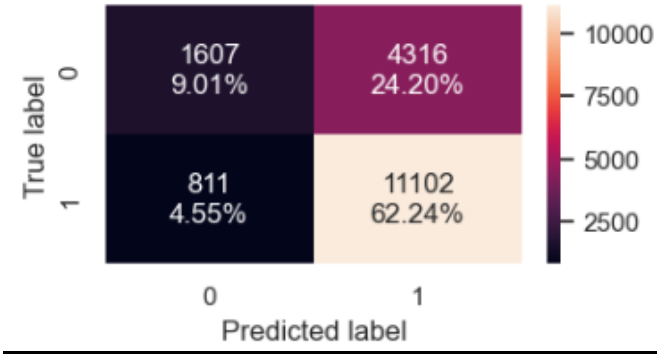


Figure 26 – Confusion matrix for train data on tuned estimator

- The confusion matrix provides a detailed breakdown of the model's predictions.
  - True Positives (TP): 11102 - The model correctly predicted 11102 positive cases.
  - True Negatives (TN): 1607 - The model correctly predicted 1607 negative cases.
  - False Positives (FP): 4316 - The model incorrectly predicted 4316 negative cases as positive.
  - False Negatives (FN): 811 - The model incorrectly predicted 811 positive cases as negative.

- Performance Metrics:
  - Accuracy: 0.71255 - The model correctly predicted 71.25% of the cases.
  - Recall: 0.93192 - The model correctly identified 93.19% of the positive cases.
  - Precision: 0.72007 - 72.01% of the positive predictions made by the model were correct.

- F1-Score: 0.81241 - The harmonic mean of precision and recall, indicating a balance between precision and recall.

Interpretation:

- The model exhibits reasonable performance overall. However, there is room for improvement, particularly in terms of precision.
- High Recall: The model is good at identifying positive cases, but it tends to overpredict positive cases, leading to a lower precision.
- Moderate Accuracy: The model correctly classifies a majority of cases, but there are still a significant number of misclassifications.

Now let's plot and analyze Confusion matrix for **test data** on tuned estimator -

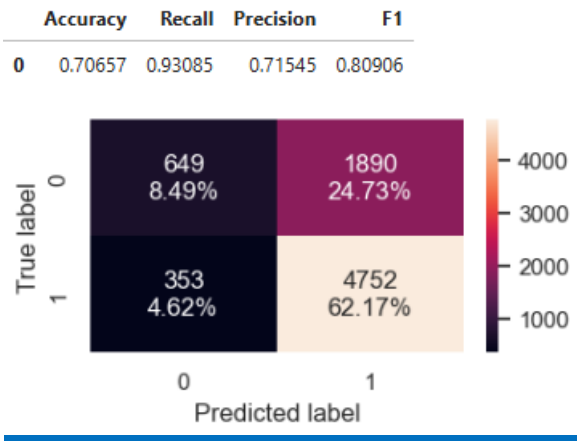


Figure 27 – Confusion matrix for test data on tuned estimator

- The confusion matrix provides a detailed breakdown of the model's predictions.
  - True Positives (TP): 4752 - The model correctly predicted 4752 positive cases.
  - True Negatives (TN): 649 - The model correctly predicted 649 negative cases.
  - False Positives (FP): 1890 - The model incorrectly predicted 1890 negative cases as positive.
  - False Negatives (FN): 353 - The model incorrectly predicted 353 positive cases as negative.
- Performance Metrics:
  - Accuracy: 0.70657 - The model correctly predicted 70.66% of the cases.
  - Recall: 0.93085 - The model correctly identified 93.08% of the positive cases.
  - Precision: 0.71545 - 71.55% of the positive predictions made by the model were correct.
  - F1-Score: 0.80906 - The harmonic mean of precision and recall, indicating a balance between precision and recall.
-

- Interpretation:
  - The model exhibits reasonable performance overall. However, there is room for improvement, particularly in terms of precision.
  - High Recall: The model is good at identifying positive cases, but it tends to overpredict positive cases, leading to a lower precision.
  - Moderate Accuracy: The model correctly classifies a majority of cases, but there are still a significant number of misclassifications.

After tuning, the Decision Tree demonstrates improved generalization, with training and test performance metrics aligning more closely. This indicates a better balance between model complexity and predictive capability.

The metrics for the test data, particularly recall, remain high, highlighting the model's effectiveness in identifying the positive class.

Bagging - Model Building and Hyperparameter Tuning

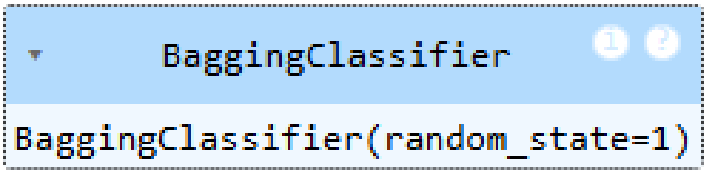


Figure 28 – Bagging Classifier

Now let's plot and analyze Confusion matrix for **train data** w.r.t bagging classifier -

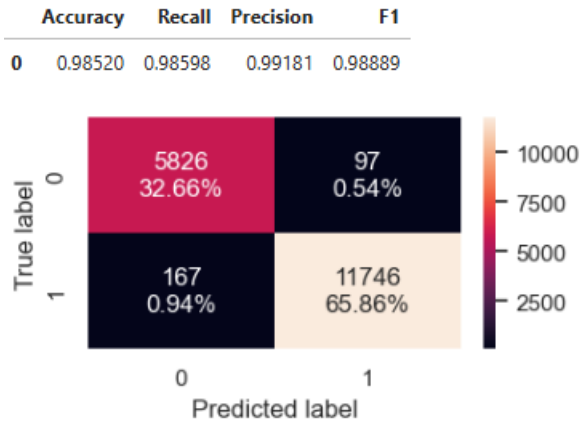


Figure 29 – Confusion matrix for train data (bagging classifier)

- The confusion matrix provides a detailed breakdown of the model's predictions.
    - True Positives (TP): 11746 - The model correctly predicted 11746 positive cases.
    - True Negatives (TN): 5826 - The model correctly predicted 5826 negative cases.
    - False Positives (FP): 97 - The model incorrectly predicted 97 negative cases as positive.
    - False Negatives (FN): 167 - The model incorrectly predicted 167 positive cases as negative.
- Performance Metrics:
    - Accuracy: 0.98520 - The model correctly predicted 98.52% of the cases.
    - Recall: 0.98598 - The model correctly identified 98.59% of the positive cases.
    - Precision: 0.99181 - 99.18% of the positive predictions made by the model were correct.
    - F1-Score: 0.98889 - The harmonic mean of precision and recall, indicating a balance between precision and recall.

Interpretation:

The model exhibits excellent performance across all metrics. This suggests that the model is highly accurate in classifying cases into their correct categories. There are very few false positives and false negatives, indicating that the model is both sensitive and specific.

Now let's plot and analyze Confusion matrix for **test data** w.r.t bagging classifier –

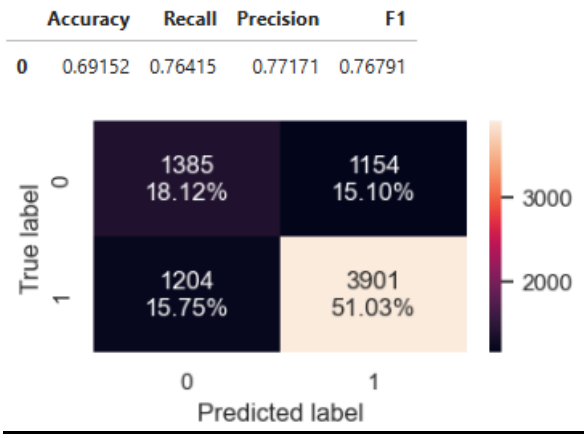


Figure 30 – Confusion matrix for test data (bagging classifier)

- The confusion matrix provides a detailed breakdown of the model's predictions.
    - True Positives (TP): 3901 - The model correctly predicted 3901 positive cases.
    - True Negatives (TN): 1385 - The model correctly predicted 1385 negative cases.
    - False Positives (FP): 1154 - The model incorrectly predicted 1154 negative cases as positive.
    - False Negatives (FN): 1204 - The model incorrectly predicted 1204 positive cases as negative.

- Performance Metrics:
  - Accuracy: 0.69152 - The model correctly predicted 69.15% of the cases.
  - Recall: 0.76415 - The model correctly identified 76.41% of the positive cases.
  - Precision: 0.77171 - 77.17% of the positive predictions made by the model were correct.
  - F1-Score: 0.76791 - The harmonic mean of precision and recall, indicating a balance between precision and recall.

Interpretation:

- The model exhibits reasonable performance overall. However, there is room for improvement, particularly in terms of accuracy and precision.
- The model correctly classifies a majority of cases, but there are still a significant number of misclassifications.
- Good Recall: The model is relatively good at identifying positive cases, but there are some false negatives.

The Bagging Classifier demonstrates strong performance on the training data, indicating potential overfitting, as shown by a notable drop in metrics when evaluated on the test data.
Although recall remains high, the decline in test accuracy suggests that the model may struggle to generalize effectively to new data and could benefit from further tuning to mitigate overfitting.

Hyperparameter Tuning - Bagging Classifier

Hyperparameter tuning for a Bagging Classifier focuses on finding the best set of hyperparameters to boost its performance, minimize overfitting, and improve its ability to generalize to new data.

Bagging, or Bootstrap Aggregating, works by training multiple base models (often decision trees) on different subsets of the training data. The predictions from these models are then combined to create a more reliable and robust final model.

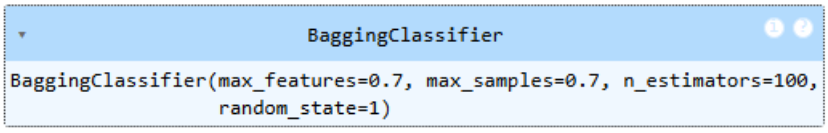


Figure 31 – Bagging classifier(hypertuned)

Now let's plot and analyze Confusion matrix for **train data** w.r.t bagging classifier(hypertuned) –

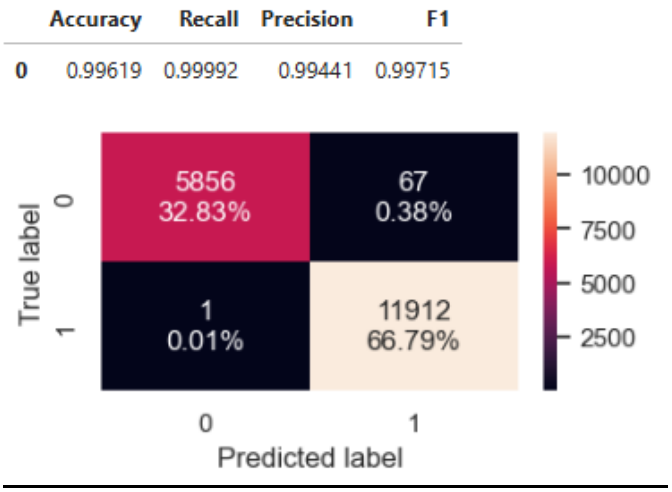


Figure 32 – Confusion matrix for train data - Bagging classifier(hypertuned)

- The confusion matrix provides a detailed breakdown of the model's predictions.
    - True Positives (TP): 11912 - The model correctly predicted 11912 positive cases.
    - True Negatives (TN): 5856 - The model correctly predicted 5856 negative cases.
    - False Positives (FP): 67 - The model incorrectly predicted 67 negative cases as positive.
    - False Negatives (FN): 1 - The model incorrectly predicted 1 positive case as negative.

- Performance Metrics:
    - Accuracy: 0.99619 - The model correctly predicted 99.62% of the cases.
    - Recall: 0.99992 - The model correctly identified 99.99% of the positive cases.
    - Precision: 0.99441 - 99.44% of the positive predictions made by the model were correct.
    - F1-Score: 0.99715 - The harmonic mean of precision and recall, indicating a balance between precision and recall.

The model exhibits excellent performance across all metrics. This suggests that the model is highly accurate in classifying cases into their correct categories. There are very few false positives and false negatives, indicating that the model is both sensitive and specific.

Now let's plot and analyze Confusion matrix for **test data** w.r.t bagging classifier(hypertuned) –

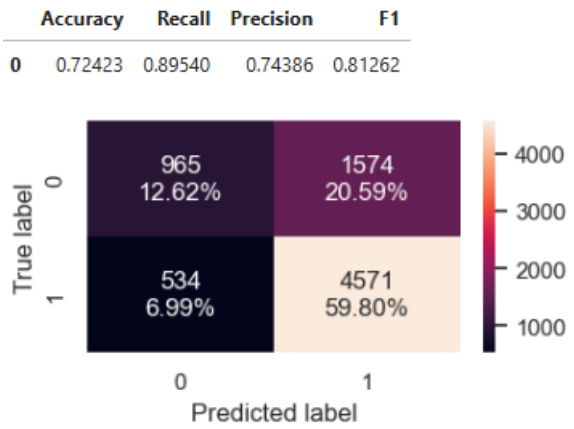


Figure 33 – Confusion matrix for test data - Bagging classifier(hypertuned)

- The confusion matrix provides a detailed breakdown of the model's predictions.
  - True Positives (TP): 4571 - The model correctly predicted 4571 positive cases.
  - True Negatives (TN): 965 - The model correctly predicted 965 negative cases.
  - False Positives (FP): 1574 - The model incorrectly predicted 1574 negative cases as positive.
  - False Negatives (FN): 534 - The model incorrectly predicted 534 positive cases as negative.
- Performance Metrics:
  - Accuracy: 0.72423 - The model correctly predicted 72.42% of the cases.
  - Recall: 0.89540 - The model correctly identified 89.54% of the positive cases.
  - Precision: 0.74386 - 74.39% of the positive predictions made by the model were correct.
  - F1-Score: 0.81262 - The harmonic mean of precision and recall, indicating a balance between precision and recall.

Interpretation:

- Due to hyperparameter tuning, the model exhibits reasonable performance overall. However, there is room for improvement, particularly in terms of precision.
- Good Recall: The model is relatively good at identifying positive cases, but it tends to overpredict positive cases, leading to a lower precision.
- Moderate Accuracy: The model correctly classifies a majority of cases, but there are still a significant number of misclassifications.

Random Forest – Model building

It works by building multiple decision trees during training and combines their outputs to make a more accurate and stable prediction. The idea behind Random Forest is to reduce overfitting and increase generalization by creating a 'forest' of trees and aggregating their predictions.

Let's plot and analyze Confusion matrix for **train data** (Random forest)–

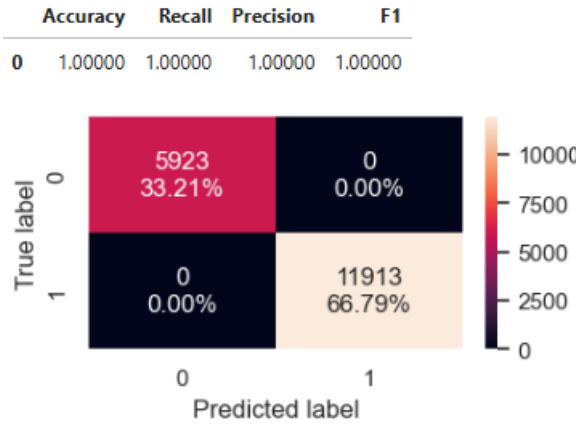


Figure 34 – Confusion matrix for train data – Random forest

- High Accuracy: The model correctly classified 100% of the cases.
- Perfect Precision and Recall: The model correctly identified all positive and negative cases.
- Excellent F1-Score: The harmonic mean of precision and recall is 1.0, further confirming the model's strong performance.

Reasons for Perfect Performance:

- Data Quality: High-quality data with clear distinctions between classes.
- Model Selection: Appropriate choice of model for the given problem.
- Hyperparameter Tuning: Optimal model configuration.

- Balanced Dataset: Equal representation of both classes in the training data.

Let's plot and analyze Confusion matrix for **test data** (Random forest)–

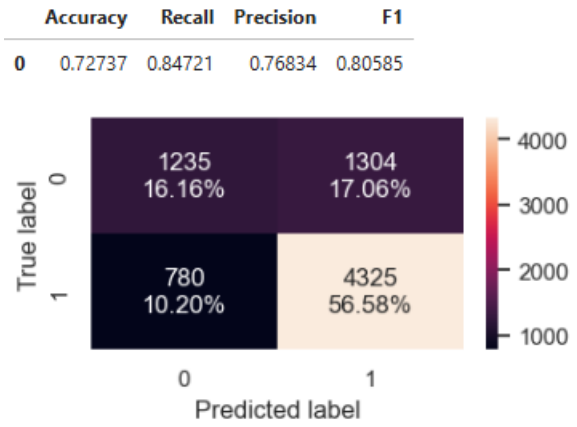


Figure 35 – Confusion matrix for test data – Random forest

- The confusion matrix provides a detailed breakdown of the model's predictions.
  - True Positives (TP): 4325 - The model correctly predicted 4325 positive cases.
  - True Negatives (TN): 1235 - The model correctly predicted 1235 negative cases.
  - False Positives (FP): 1304 - The model incorrectly predicted 1304 negative cases as positive.
  - False Negatives (FN): 780 - The model incorrectly predicted 780 positive cases as negative.

- Performance Metrics:
  - Accuracy: 0.72737 - The model correctly predicted 72.74% of the cases.
  - Recall: 0.84721 - The model correctly identified 84.72% of the positive cases.
  - Precision: 0.76834 - 76.83% of the positive predictions made by the model were correct.
  - F1-Score: 0.80585 - The harmonic mean of precision and recall, indicating a balance between precision and recall.

The Random Forest Classifier achieves perfect training scores, a strong indication of overfitting, as it has learned the training data exactly. However, there is room for improvement, particularly in terms of precision.

- Good Recall: The model is relatively good at identifying positive cases, but it tends to overpredict positive cases, leading to a lower precision.
- Moderate Accuracy: The model correctly classifies a majority of cases, but there are still a significant number of misclassifications.

Hyperparameter Tuning – Random Forest

```
RandomForestClassifier(max_depth=10, min_samples_split=7, n_estimators=80,
                       oob_score=True, random_state=1)
```

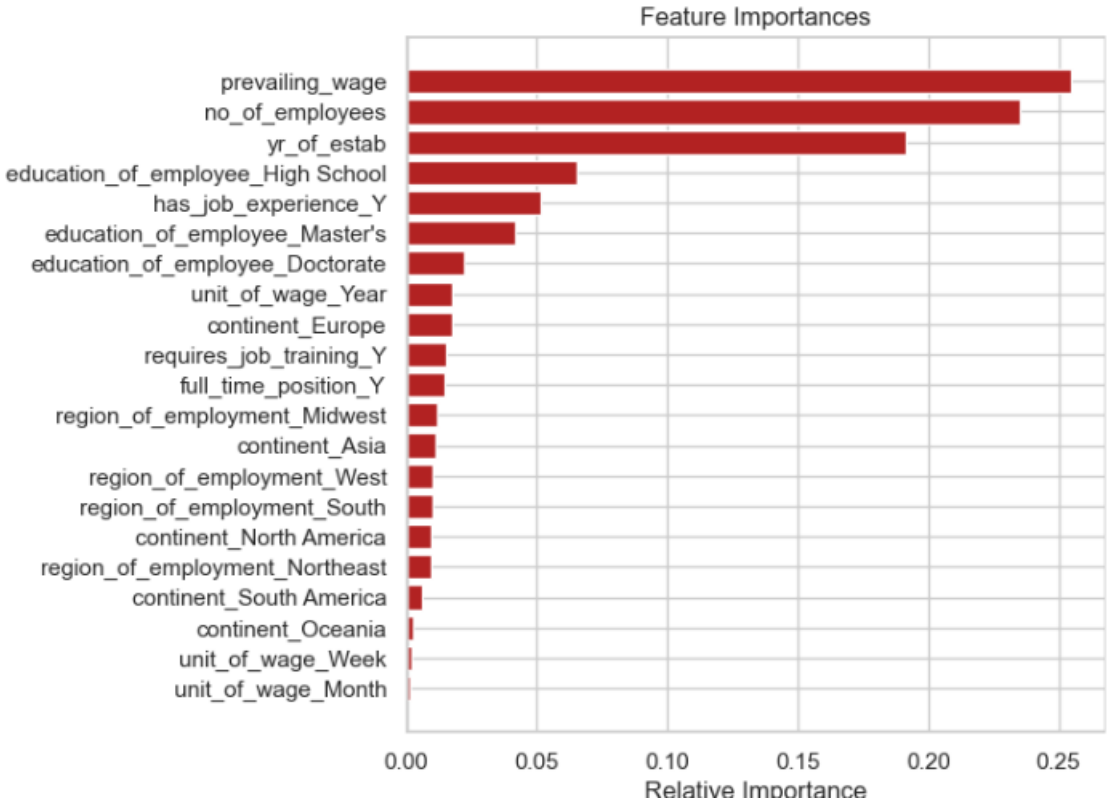


Figure 36 – Feature importance – random forest (Hypertuned)

Now let's plot and analyze Confusion matrix for **train data** -Random Forest (hypertuned) –

	Accuracy	Recall	Precision	F1
0	0.77377	0.90649	0.78710	0.84259

Figure 37 – Confusion matrix for train data – Random forest(hypertuned)

- Confusion Matrix:
    - True Positives (TP): 10799 - The model correctly predicted 10799 positive cases.
    - True Negatives (TN): 3002 - The model correctly predicted 3002 negative cases.
    - False Positives (FP): 2921 - The model incorrectly predicted 2921 negative cases as positive.
    - False Negatives (FN): 1114 - The model incorrectly predicted 1114 positive cases as negative.
- Performance Metrics:

    - Accuracy: 0.77377 - The model correctly predicted 77.38% of the cases.
    - Recall: 0.90649 - The model correctly identified 90.65% of the positive cases.
    - Precision: 0.78710 - 78.71% of the positive predictions made by the model were correct.
    - F1-Score: 0.84259 - The harmonic mean of precision and recall, indicating a balance between precision and recall.

Interpretation:

- The model exhibits reasonable performance overall. However, there is room for improvement, particularly in terms of precision.
- Good Recall: The model is relatively good at identifying positive cases, but it tends to overpredict positive cases, leading to a lower precision.
- Moderate Accuracy: The model correctly classifies a majority of cases, but there are still a significant number of misclassifications.

Now let's plot and analyze Confusion matrix for **test data** -Random Forest (hypertuned) –

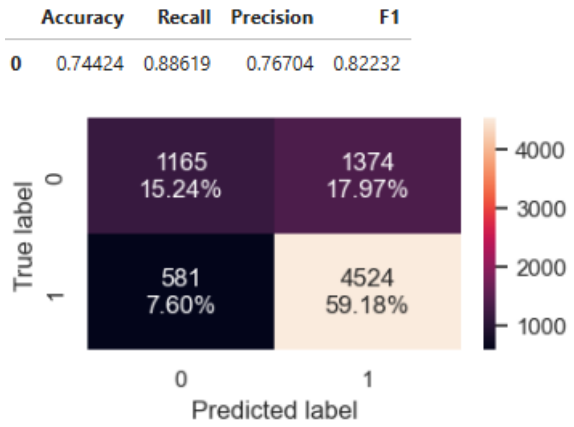


Figure 38 – Confusion matrix for test data – Random forest(hypertuned)

- Confusion Matrix:
    - True Positives (TP): 4524 - The model correctly predicted 4524 positive cases.
    - True Negatives (TN): 1165 - The model correctly predicted 1165 negative cases.

    - False Positives (FP): 1374 - The model incorrectly predicted 1374 negative cases as positive.
    - False Negatives (FN): 581 - The model incorrectly predicted 581 positive cases as negative.

- Performance Metrics:
    - Accuracy: 0.74424 - The model correctly predicted 74.42% of the cases.
    - Recall: 0.88619 - The model correctly identified 88.62% of the positive cases.
    - Precision: 0.76704 - 76.70% of the positive predictions made by the model were correct.
    - F1-Score: 0.82232 - The harmonic mean of precision and recall, indicating a balance between precision and recall.

Interpretation:

- The model exhibits reasonable performance overall. However, there is room for improvement, particularly in terms of precision.
- Good Recall: The model is relatively good at identifying positive cases, but it tends to overpredict positive cases, leading to a lower precision.
- Moderate Accuracy: The model correctly classifies a majority of cases, but there are still a significant number of misclassifications.

Following hyperparameter tuning, the Random Forest Classifier exhibits reduced overfitting, with training metrics no longer achieving perfect scores.
The test metrics are relatively aligned with the training metrics, reflecting better model generalization and showcasing a strong recall and a solid F1 score.

Boosting - Model Building

Boosting is a powerful ensemble technique used in machine learning that focuses on converting a set of weak learners (typically simple models with limited predictive power) into a strong learner.

AdaBoost Classifier

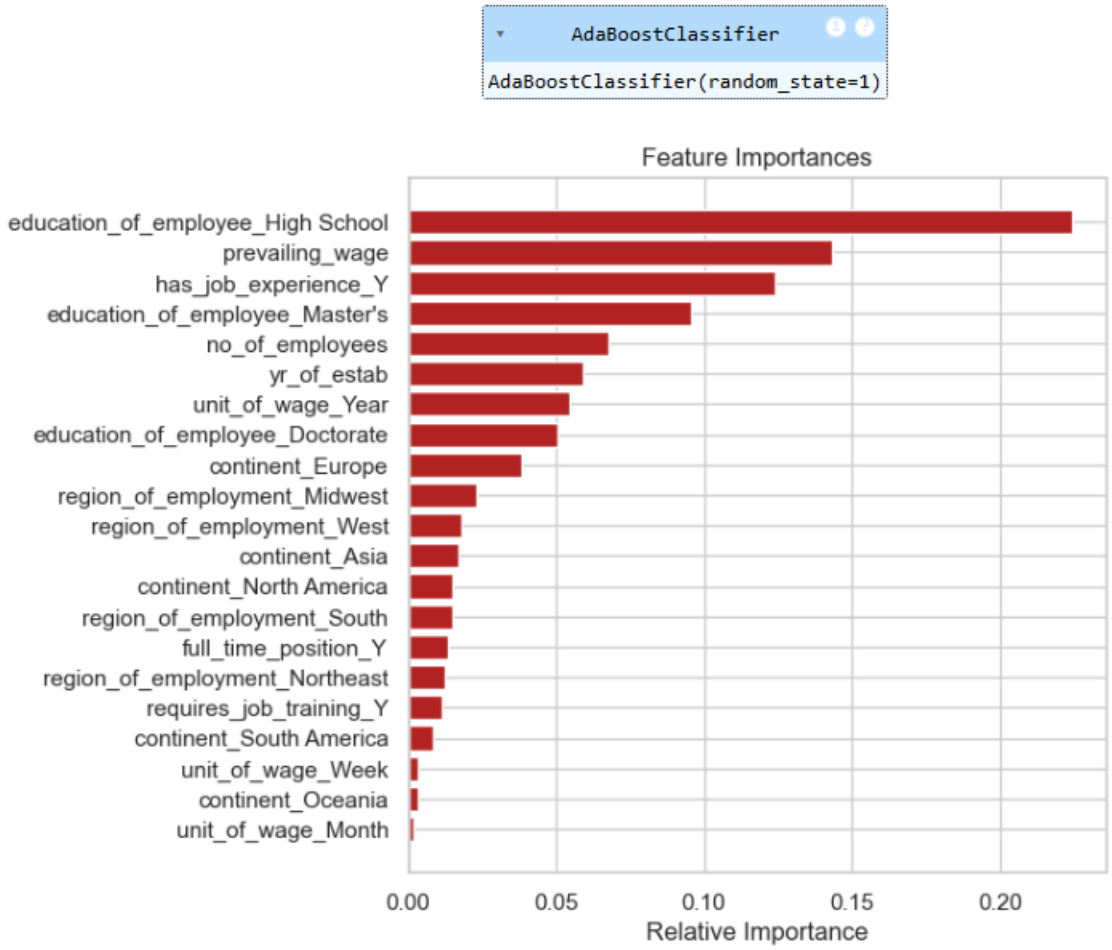


Figure 39 – AdaBoost classifier

Now let's plot and analyze Confusion matrix for **train data** – AdaBoost :

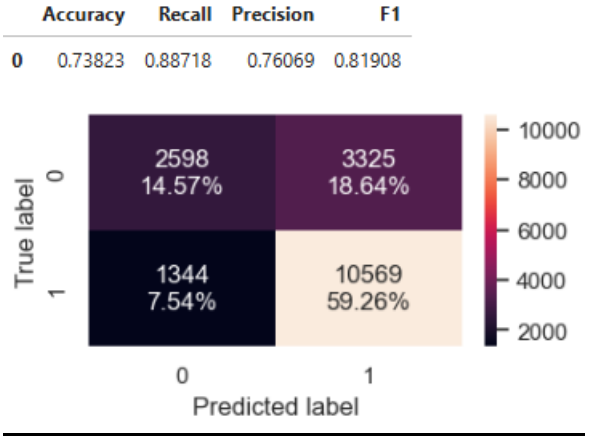


Figure 40 – Confusion matrix for train data – AdaBoost

- Confusion Matrix:
    - True Positives (TP): 10569 - The model correctly predicted 10569 positive cases.
    - True Negatives (TN): 2598 - The model correctly predicted 2598 negative cases.
    - False Positives (FP): 3325 - The model incorrectly predicted 3325 negative cases as positive.
    - False Negatives (FN): 1344 - The model incorrectly predicted 1344 positive cases as negative.

- Performance Metrics:

    - Accuracy: 0.73823 - The model correctly predicted 73.82% of the cases.
    - Recall: 0.88718 - The model correctly identified 88.72% of the positive cases.
    - Precision: 0.76069 - 76.07% of the positive predictions made by the model were correct.
    - F1-Score: 0.81908 - The harmonic mean of precision and recall, indicating a balance between precision and recall.

Interpretation:

- The model exhibits reasonable performance overall. However, there is room for improvement, particularly in terms of precision.
- Good Recall: The model is relatively good at identifying positive cases, but it tends to overpredict positive cases, leading to a lower precision.
- Moderate Accuracy: The model correctly classifies a majority of cases, but there are still a significant number of misclassifications.

Now let's plot and analyze Confusion matrix for **test data** – AdaBoost :

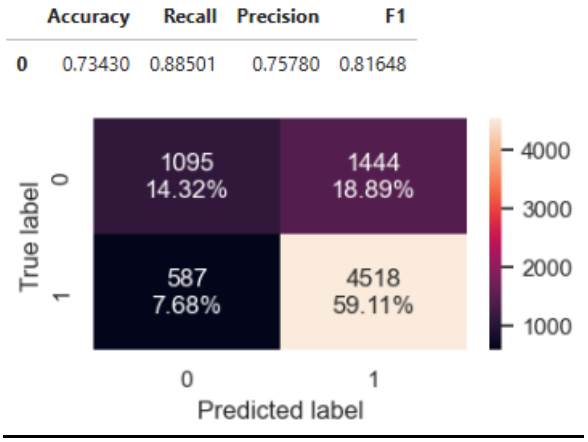


Figure 41 – Confusion matrix for test data – Adaboost

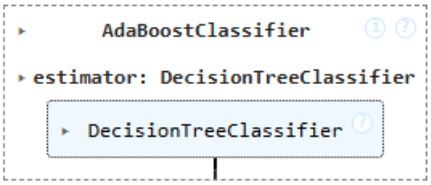
- Confusion Matrix:
  - True Positives (TP): 4518 - The model correctly predicted 4518 positive cases.
  - True Negatives (TN): 1095 - The model correctly predicted 1095 negative cases.
  - False Positives (FP): 1444 - The model incorrectly predicted 1444 negative cases as positive.
  - False Negatives (FN): 587 - The model incorrectly predicted 587 positive cases as negative.

- Performance Metrics:
  - Accuracy: 0.73430 - The model correctly predicted 73.43% of the cases.
  - Recall: 0.88501 - The model correctly identified 88.50% of the positive cases.
  - Precision: 0.75780 - 75.78% of the positive predictions made by the model were correct.
  - F1-Score: 0.81648 - The harmonic mean of precision and recall, indicating a balance between precision and recall.

Interpretation:

- The model exhibits reasonable performance overall. However, there is room for improvement, particularly in terms of precision.
- Good Recall: The model is relatively good at identifying positive cases, but it tends to overpredict positive cases, leading to a lower precision.
- Moderate Accuracy: The model correctly classifies a majority of cases, but there are still a significant number of misclassifications.

Hyperparameter Tuning - AdaBoost Classifier



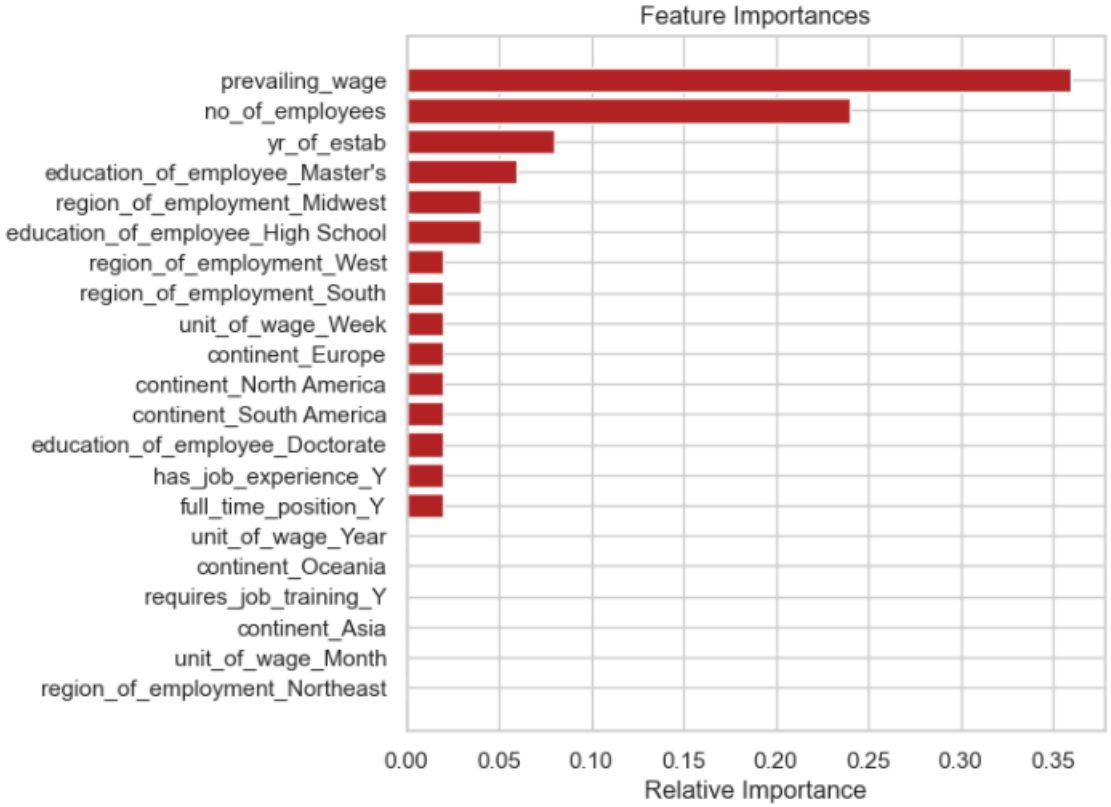


Figure 42 – AdaBoost classifier(hyperparameter tuning) – important features

Now let's plot and analyze Confusion matrix for **train data** – AdaBoost(hypertuned) :

	Accuracy	Recall	Precision	F1
0	0.75443	0.88391	0.77844	0.82783

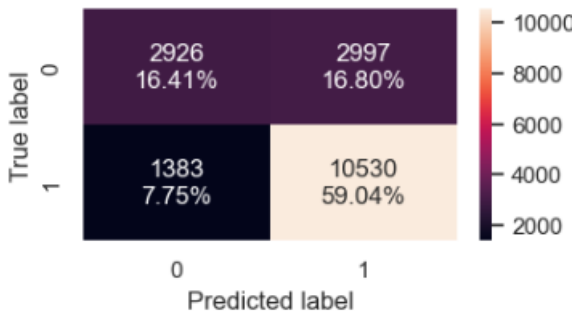


Figure 43 – Confusion matrix for train data – Adaboost(hypertuned)

- Confusion Matrix:

    - True Positives (TP): 10530 - The model correctly predicted 10530 positive cases.
    - True Negatives (TN): 2926 - The model correctly predicted 2926 negative cases.
    - False Positives (FP): 2997 - The model incorrectly predicted 2997 negative cases as positive.
    - False Negatives (FN): 1383 - The model incorrectly predicted 1383 positive cases as negative.

- Performance Metrics:

    - Accuracy: 0.75443 - The model correctly predicted 75.44% of the cases.
    - Recall: 0.88391 - The model correctly identified 88.39% of the positive cases.
    - Precision: 0.77844 - 77.84% of the positive predictions made by the model were correct.
    - F1-Score: 0.82783 - The harmonic mean of precision and recall, indicating a balance between precision and recall.

Interpretation:

- The model exhibits reasonable performance overall. However, there is room for improvement, particularly in terms of precision.
- Good Recall: The model is relatively good at identifying positive cases, but it tends to overpredict positive cases, leading to a lower precision.
- Moderate Accuracy: The model correctly classifies a majority of cases, but there are still a significant number of misclassifications.

Now let's plot and analyze Confusion matrix for **test data** – AdaBoost (hypertuned) :

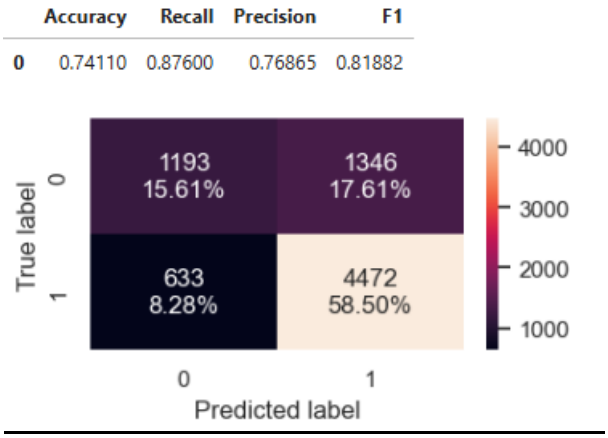


Figure 44 – Confusion matrix for test data – Adaboost(hypertuned)

- Confusion Matrix:

    - True Positives (TP): 4472 - The model correctly predicted 4472 positive cases.
    - True Negatives (TN): 1193 - The model correctly predicted 1193 negative cases.
    - False Positives (FP): 1346 - The model incorrectly predicted 1346 negative cases as positive.
    - False Negatives (FN): 633 - The model incorrectly predicted 633 positive cases as negative.

- Performance Metrics:

    - Accuracy: 0.74110 - The model correctly predicted 74.11% of the cases.
    - Recall: 0.87600 - The model correctly identified 87.60% of the positive cases.
    - Precision: 0.76865 - 76.87% of the positive predictions made by the model were correct.
    - F1-Score: 0.81882 - The harmonic mean of precision and recall, indicating a balance between precision and recall

Interpretation:

- The model exhibits reasonable performance overall. However, there is room for improvement, particularly in terms of precision.
- Good Recall: The model is relatively good at identifying positive cases, but it tends to overpredict positive cases, leading to a lower precision.
- Moderate Accuracy: The model correctly classifies a majority of cases, but there are still a significant number of misclassifications.

The hyperparameter tuning of the AdaBoost Classifier led to a modest increase in training accuracy, while preserving high recall and achieving a strong F1 score. The metrics on the test data remained stable after tuning, suggesting that the model is generalizing effectively to unseen data.

Gradient Boosting

The Gradient Boosting Classifier is an ensemble learning algorithm that builds a model in a stage-wise manner, optimizing for a loss function by adding new models that correct the errors of the previous ones. It is particularly effective for regression and classification tasks due to its ability to handle both linear and non-linear relationships in the data.

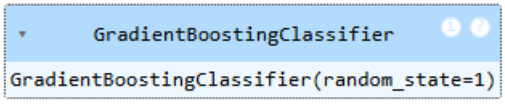


Figure 45 – Gradient booster

Now let's plot and analyze Confusion matrix for **train data** – Gradient booster -

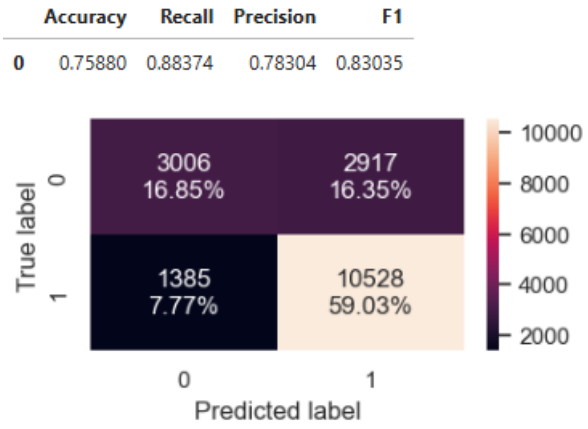


Figure 46 – Confusion matrix for Training data - Gradient booster

- Confusion Matrix:
  - True Positives (TP): 4472 - The model correctly predicted 4472 positive cases.
  - True Negatives (TN): 1193 - The model correctly predicted 1193 negative cases.
  - False Positives (FP): 1346 - The model incorrectly predicted 1346 negative cases as positive.
  - False Negatives (FN): 633 - The model incorrectly predicted 633 positive cases as negative.
- Performance Metrics:
  - Accuracy: 0.74110 - The model correctly predicted 74.11% of the cases.
  - Recall: 0.87600 - The model correctly identified 87.60% of the positive cases.
  - Precision: 0.76865 - 76.87% of the positive predictions made by the model were correct.
  - F1-Score: 0.81882 - The harmonic mean of precision and recall, indicating a balance between precision and recall.

Interpretation:

- The model exhibits reasonable performance overall. However, there is room for improvement, particularly in terms of precision.
- Good Recall: The model is relatively good at identifying positive cases, but it tends to overpredict positive cases, leading to a lower precision.
- Moderate Accuracy: The model correctly classifies a majority of cases, but there are still a significant number of misclassifications.

Now let's plot and analyze Confusion matrix for **test data** – Gradient booster -

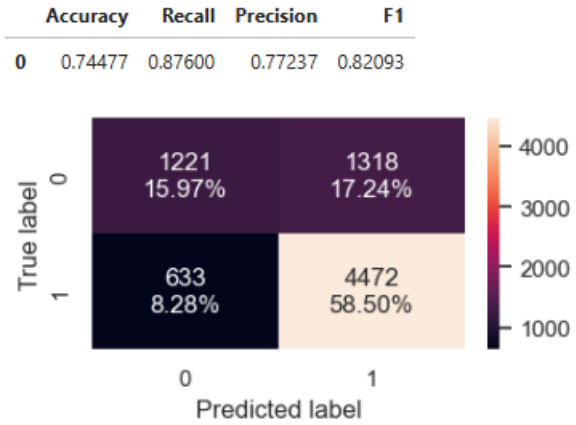


Figure 47 – Confusion matrix for Test data - Gradient booster

- Confusion Matrix:
  - True Positives (TP): 10528 - The model correctly predicted 10528 positive cases.
  - True Negatives (TN): 3006 - The model correctly predicted 3006 negative cases.
  - False Positives (FP): 2917 - The model incorrectly predicted 2917 negative cases as positive.
  - False Negatives (FN): 1385 - The model incorrectly predicted 1385 positive cases as negative.
- Performance Metrics:
  - Accuracy: 0.75880 - The model correctly predicted 75.88% of the cases.
  - Recall: 0.88374 - The model correctly identified 88.37% of the positive cases.
  - Precision: 0.78304 - 78.30% of the positive predictions made by the model were correct.
  - F1-Score: 0.83035 - The harmonic mean of precision and recall, indicating a balance between precision and recall.

Interpretation:

- The model exhibits reasonable performance overall. However, there is room for improvement, particularly in terms of precision.
- Good Recall: The model is relatively good at identifying positive cases, but it tends to overpredict positive cases, leading to a lower precision.
- Moderate Accuracy: The model correctly classifies a majority of cases, but there are still a significant number of misclassifications.

Hyperparameter Tuning - Gradient Boosting Classifier

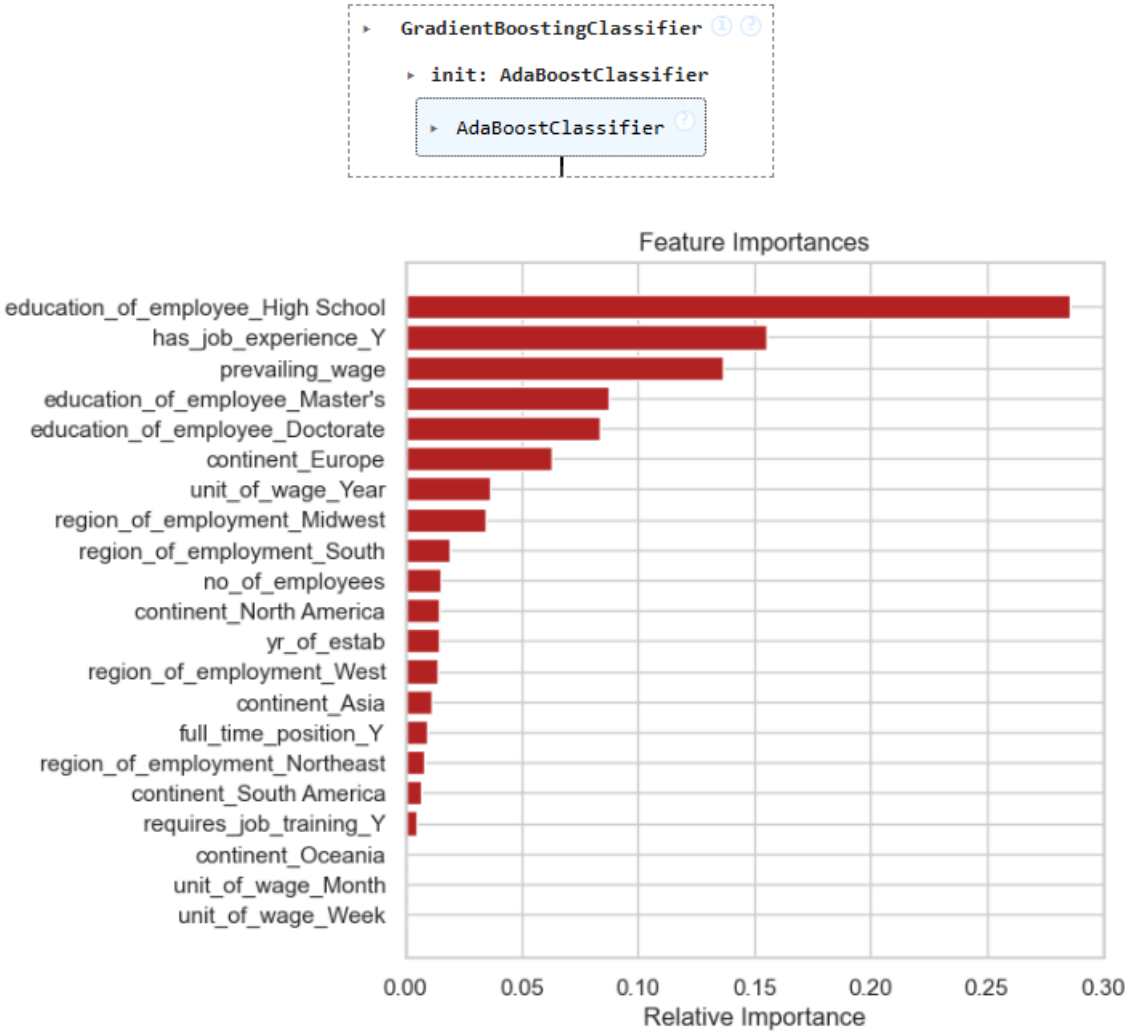


Figure 48 – Gradient booster(hypertuned) and important features

Now let's plot and analyze Confusion matrix for **train data** – Gradient booster(hypertuned) :

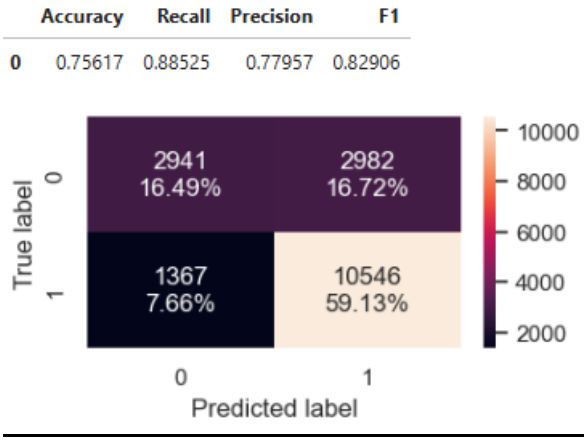


Figure 49 – Confusion matrix for Train data - Gradient booster(hypertuned)

Confusion Matrix:

- **True Positives (TP):** 10546 - The model correctly predicted 10546 positive cases.
- **True Negatives (TN):** 2941 - The model correctly predicted 2941 negative cases.
- **False Positives (FP):** 2982 - The model incorrectly predicted 2982 negative cases as positive.
- **False Negatives (FN):** 1367 - The model incorrectly predicted 1367 positive cases as negative.

Performance Metrics:

- **Accuracy:** 0.75617 - The model correctly predicted 75.62% of the cases.
- **Recall:** 0.88525 - The model correctly identified 88.53% of the positive cases.
- **Precision:** 0.77957 - 77.96% of the positive predictions made by the model were correct.
- **F1-Score:** 0.82906 - The harmonic mean of precision and recall, indicating a balance between precision and recall.

Interpretation:

The model exhibits reasonable performance overall. However, there is room for improvement, particularly in terms of precision.

- **Good Recall:** The model is relatively good at identifying positive cases, but it tends to overpredict positive cases, leading to a lower precision.
- **Moderate Accuracy:** The model correctly classifies a majority of cases, but there are still a significant number of misclassifications.

Now let's plot and analyze Confusion matrix for **test data** – Gradient booster(hypertuned) :

	Accuracy	Recall	Precision	F1
0	0.74372	0.87875	0.77000	0.82078

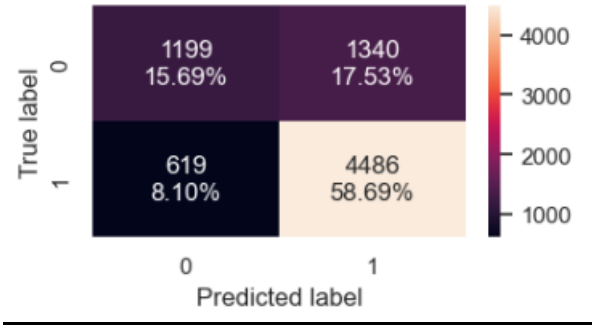


Figure 50 – Confusion matrix for Test data - Gradient booster(hypertuned)

Confusion Matrix:

- **True Positives (TP):** 4486 - The model correctly predicted 4486 positive cases.
- **True Negatives (TN):** 1199 - The model correctly predicted 1199 negative cases.
- **False Positives (FP):** 1340 - The model incorrectly predicted 1340 negative cases as positive.
- **False Negatives (FN):** 619 - The model incorrectly predicted 619 positive cases as negative.

Performance Metrics:

- **Accuracy:** 0.74372 - The model correctly predicted 74.37% of the cases.
- **Recall:** 0.87875 - The model correctly identified 87.88% of the positive cases.
- **Precision:** 0.77000 - 77.00% of the positive predictions made by the model were correct.
- **F1-Score:** 0.82078 - The harmonic mean of precision and recall, indicating a balance between precision and recall.

Interpretation:

The model exhibits reasonable performance overall. However, there is room for improvement, particularly in terms of precision.

After tuning, the Gradient Boosting Classifier shows stable and closely aligned metrics between the training and test data, indicating good generalization.

The high recall and consistent F1 scores across both datasets demonstrate the model's ability to accurately identify most positive cases while maintaining a balance between precision and recall.

XGBoost Classifier

The **XGBoost Classifier** (Extreme Gradient Boosting) is an optimized, highly efficient implementation of gradient boosting that is designed to be faster and more accurate than other gradient boosting models. It is widely used in machine learning competitions and real-world applications due to its performance and flexibility.

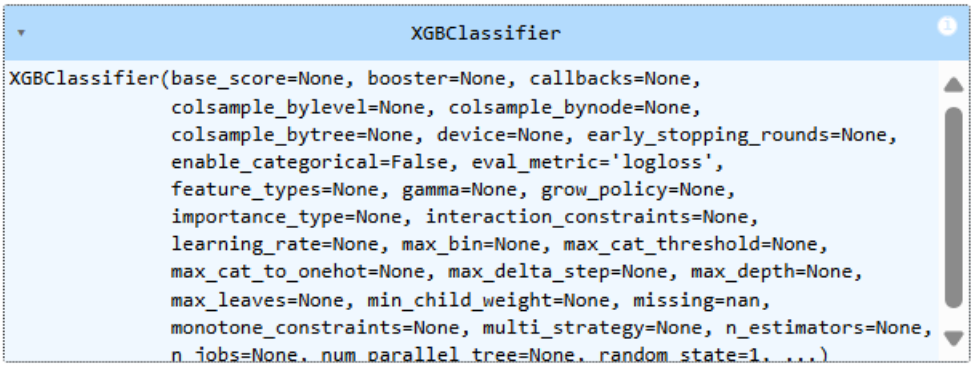


Figure 51 – XGBoost Classifier

Now let's plot and analyze Confusion matrix for **train data** – XGBoost

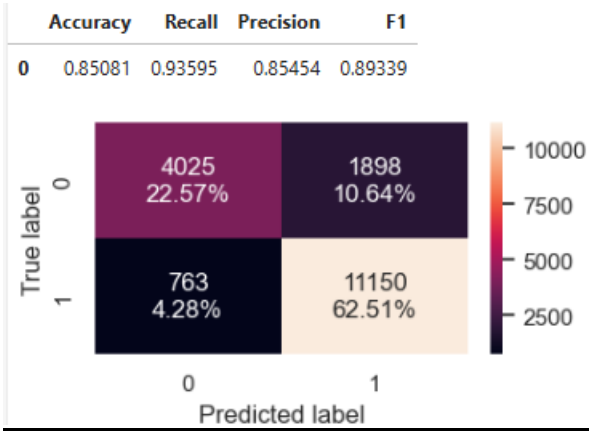


Figure 52 – Confusion matrix for Train data - XGBoost

Confusion Matrix:

- **True Positives (TP):** 11150 - The model correctly predicted 11150 positive cases.
- **True Negatives (TN):** 4025 - The model correctly predicted 4025 negative cases.
- **False Positives (FP):** 1898 - The model incorrectly predicted 1898 negative cases as positive.
- **False Negatives (FN):** 763 - The model incorrectly predicted 763 positive cases as negative.

Performance Metrics:

- **Accuracy:** 0.85081 - The model correctly predicted 85.08% of the cases.
- **Recall:** 0.93595 - The model correctly identified 93.59% of the positive cases.
- **Precision:** 0.85454 - 85.45% of the positive predictions made by the model were correct.
- **F1-Score:** 0.89339 - The harmonic mean of precision and recall, indicating a balance between precision and recall.

Interpretation:

The model exhibits strong performance overall. It has a high accuracy, recall, precision, and F1-score. This suggests that the model is effective in classifying cases correctly.

Now let's plot and analyze Confusion matrix for **test data** – XGBoost

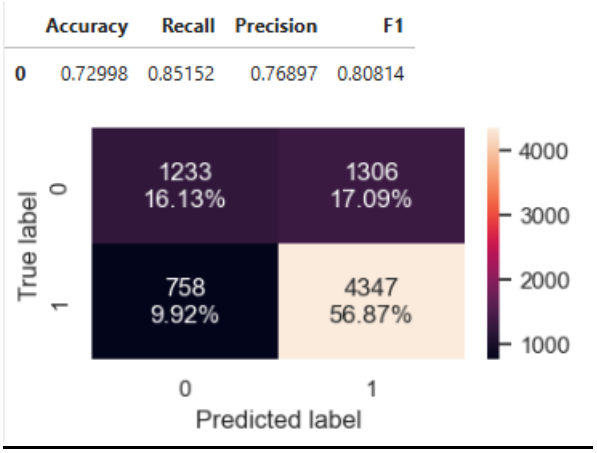


Figure 53 – Confusion matrix for Train data – XGBoost

Confusion Matrix:

- **True Positives (TP):** 4347 - The model correctly predicted 4347 positive cases.
- **True Negatives (TN):** 1233 - The model correctly predicted 1233 negative cases.
- **False Positives (FP):** 1306 - The model incorrectly predicted 1306 negative cases as positive.
- **False Negatives (FN):** 758 - The model incorrectly predicted 758 positive cases as negative.

Performance Metrics:

- **Accuracy:** 0.72998 - The model correctly predicted 72.998% of the cases.
- **Recall:** 0.85152 - The model correctly identified 85.15% of the positive cases.
- **Precision:** 0.76897 - 76.897% of the positive predictions made by the model were correct.

- **F1-Score:** 0.80814 - The harmonic mean of precision and recall, indicating a balance between precision and recall.

Interpretation:

The model exhibits reasonable performance overall. However, there is room for improvement, particularly in terms of precision.

- **Good Recall:** The model is relatively good at identifying positive cases, but it tends to overpredict positive cases, leading to a lower precision.
- **Moderate Accuracy:** The model correctly classifies a majority of cases, but there are still a significant number of misclassifications.

Hyperparameter Tuning - XGBoost Classifier

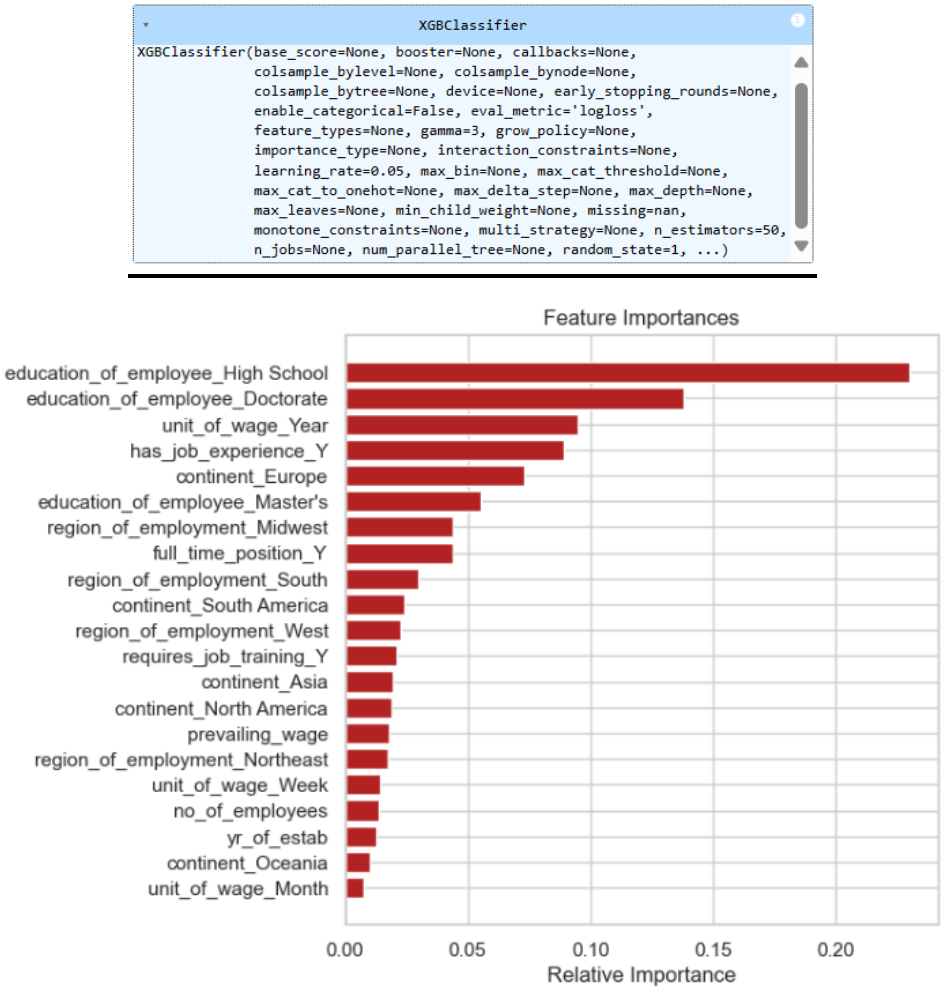


Figure 54 – XGBoost Classifier(hypertuned) and important features

Now let's plot and analyze Confusion matrix for **train data** – XGBoost(hypertuned)

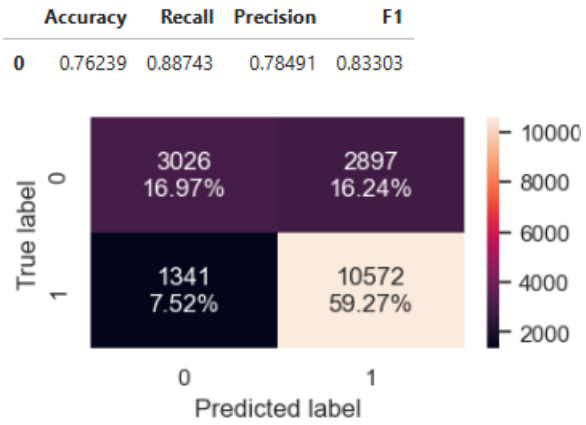


Figure 55 – Confusion matrix for Train data - XGBoost(hypertuned)

Confusion Matrix:

- **True Positives (TP):** 10572 - The model correctly predicted 10572 positive cases.
- **True Negatives (TN):** 3026 - The model correctly predicted 3026 negative cases.
- **False Positives (FP):** 2897 - The model incorrectly predicted 2897 negative cases as positive.
- **False Negatives (FN):** 1341 - The model incorrectly predicted 1341 positive cases as negative.

Performance Metrics:

- **Accuracy:** 0.76239 - The model correctly predicted 76.24% of the cases.
- **Recall:** 0.88743 - The model correctly identified 88.74% of the positive cases.
- **Precision:** 0.78491 - 78.49% of the positive predictions made by the model were correct.
- **F1-Score:** 0.83303 - The harmonic mean of precision and recall, indicating a balance between precision and recall.

Interpretation:

The model exhibits reasonable performance overall. However, there is room for improvement, particularly in terms of precision.

- **Good Recall:** The model is relatively good at identifying positive cases, but it tends to overpredict positive cases, leading to a lower precision.
- **Moderate Accuracy:** The model correctly classifies a majority of cases, but there are still a significant number of misclassifications.

Now let's plot and analyze Confusion matrix for **test data** – XGBoost(hypertuned)

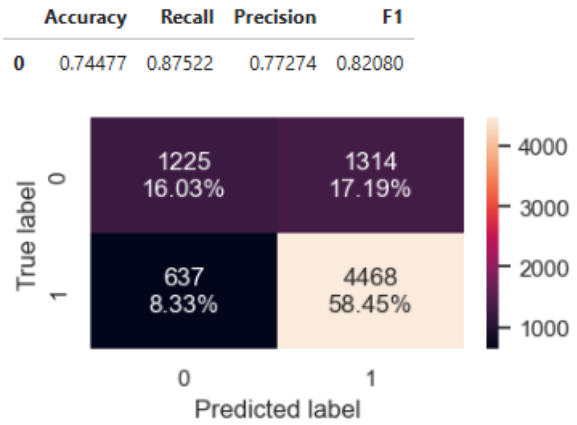


Figure 56 – Confusion matrix for Train data - XGBoost(hypertuned)

Confusion Matrix:

- **True Positives (TP):** 4468 - The model correctly predicted 4468 positive cases.
- **True Negatives (TN):** 1225 - The model correctly predicted 1225 negative cases.
- **False Positives (FP):** 1314 - The model incorrectly predicted 1314 negative cases as positive.
- **False Negatives (FN):** 637 - The model incorrectly predicted 637 positive cases as negative.

Performance Metrics:

- **Accuracy:** 0.74477 - The model correctly predicted 74.48% of the cases.
- **Recall:** 0.87522 - The model correctly identified 87.52% of the positive cases.
- **Precision:** 0.77274 - 77.27% of the positive predictions made by the model were correct.
- **F1-Score:** 0.82080 - The harmonic mean of precision and recall, indicating a balance between precision and recall.

Interpretation:

- **After hyperparameter tuning, the performance of the XGBoost Classifier has improved on both the training and test datasets, showing higher accuracy, recall, precision, and F1-score compared to the untuned model.**
- **Good Recall:** The model is relatively good at identifying positive cases, but it tends to overpredict positive cases, leading to a lower precision.

Stacking Classifier

A **Stacking Classifier** is an ensemble learning technique that combines multiple classification models to improve predictive performance. It works by training a set of base models (also known as level-0 models) and then using another model (the meta-model or level-1 model) to combine the predictions of these base models.

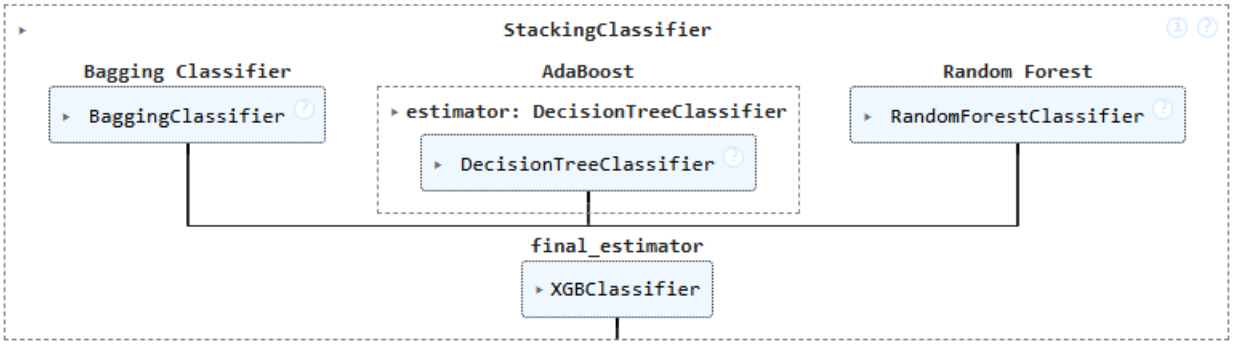


Figure 57 – Stacking classifier

Now let's plot and analyze Confusion matrix for **train data** – stacking classifier

	Accuracy	Recall	Precision	F1
0	0.74507	0.86435	0.77842	0.81914

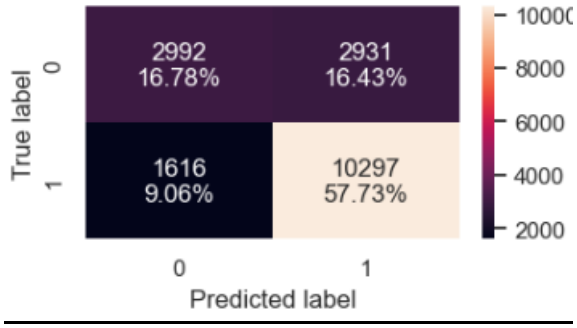


Figure 58 – Confusion matrix for Train data – stacking classifier

Confusion Matrix:

- **True Positives (TP):** 10297 - The model correctly predicted 10297 positive cases.
- **True Negatives (TN):** 2992 - The model correctly predicted 2992 negative cases.
- **False Positives (FP):** 2931 - The model incorrectly predicted 2931 negative cases as positive.
- **False Negatives (FN):** 1616 - The model incorrectly predicted 1616 positive cases as negative.

Performance Metrics:

- **Accuracy:** 0.74507 - The model correctly predicted 74.51% of the cases.
- **Recall:** 0.86435 - The model correctly identified 86.44% of the positive cases.
- **Precision:** 0.77842 - 77.84% of the positive predictions made by the model were correct.
- **F1-Score:** 0.81914 - The harmonic mean of precision and recall, indicating a balance between precision and recall.

Interpretation:

The model exhibits reasonable performance overall. However, there is room for improvement, particularly in terms of precision.

- **Good Recall:** The model is relatively good at identifying positive cases, but it tends to overpredict positive cases, leading to a lower precision.
- **Moderate Accuracy:** The model correctly classifies a majority of cases, but there are still a significant number of misclassifications.

Now let's plot and analyze Confusion matrix for **test data** – stacking classifier -

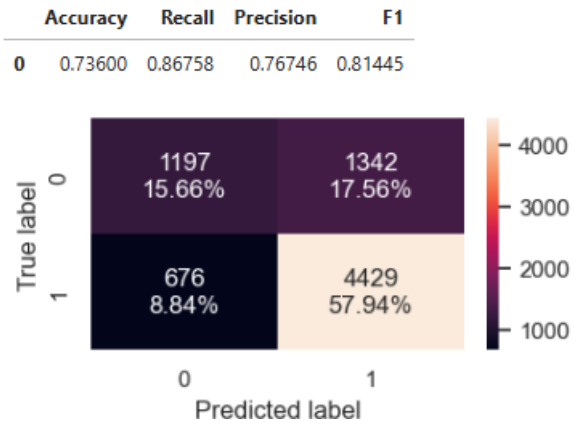


Figure 59 – Confusion matrix for test data – stacking classifier

Confusion Matrix:

- **True Positives (TP):** 4429 - The model correctly predicted 4429 positive cases.
- **True Negatives (TN):** 1197 - The model correctly predicted 1197 negative cases.
- **False Positives (FP):** 1342 - The model incorrectly predicted 1342 negative cases as positive.
- **False Negatives (FN):** 676 - The model incorrectly predicted 676 positive cases as negative.

Performance Metrics:

- **Accuracy:** 0.73600 - The model correctly predicted 73.60% of the cases.
- **Recall:** 0.86758 - The model correctly identified 86.76% of the positive cases.
- **Precision:** 0.76746 - 76.75% of the positive predictions made by the model were correct.
- **F1-Score:** 0.81445 - The harmonic mean of precision and recall, indicating a balance between precision and recall.

Interpretation:

The model exhibits reasonable performance overall. However, there is room for improvement, particularly in terms of precision.

- **Good Recall:** The model is relatively good at identifying positive cases, but it tends to overpredict positive cases, leading to a lower precision.
- **Moderate Accuracy:** The model correctly classifies a majority of cases, but there are still a significant number of misclassifications.

Stacking Classifier – Hyperparameter tuning(bagging)

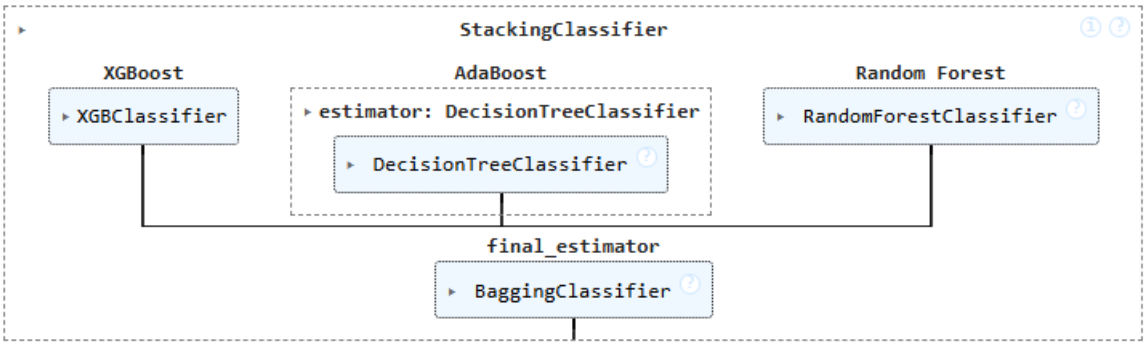


Figure 60 – Stacking classifier(hypertuned)

Now let's plot and analyze Confusion matrix for **train data** – stacking classifier (hypertuned)

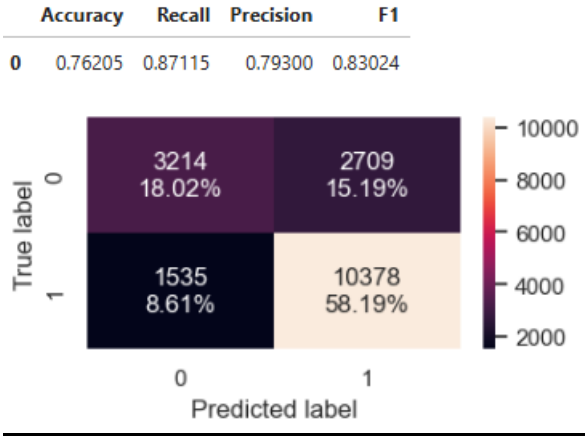


Figure 61 – Confusion matrix for Train data – stacking classifier(hypertuned)

Confusion Matrix:

- **True Positives (TP):** 10378 - The model correctly predicted 10378 positive cases.
- **True Negatives (TN):** 3214 - The model correctly predicted 3214 negative cases.
- **False Positives (FP):** 2709 - The model incorrectly predicted 2709 negative cases as positive.
- **False Negatives (FN):** 1535 - The model incorrectly predicted 1535 positive cases as negative.

Performance Metrics:

- **Accuracy:** 0.76205 - The model correctly predicted 76.21% of the cases.
- **Recall:** 0.87115 - The model correctly identified 87.12% of the positive cases.
- **Precision:** 0.79300 - 79.30% of the positive predictions made by the model were correct.
- **F1-Score:** 0.83024 - The harmonic mean of precision and recall, indicating a balance between precision and recall.

Interpretation:

The model exhibits reasonable performance overall. However, there is room for improvement, particularly in terms of precision.

- **Good Recall:** The model is relatively good at identifying positive cases, but it tends to overpredict positive cases, leading to a lower precision.
- **Moderate Accuracy:** The model correctly classifies a majority of cases, but there are still a significant number of misclassifications.

Now let's plot and analyze Confusion matrix for **test data** – stacking classifier (hypertuned)

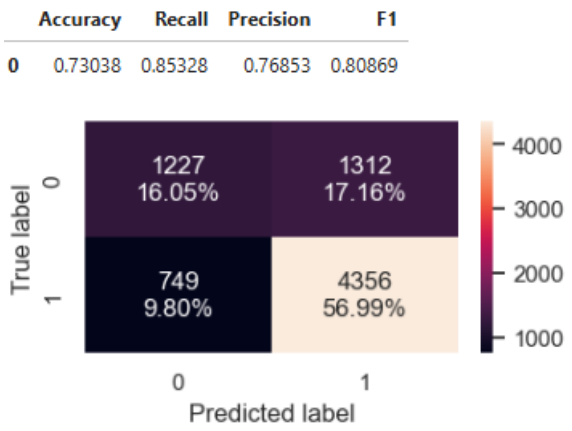


Figure 62 – Confusion matrix for test data – stacking classifier(hypertuned)

Confusion Matrix:

- **True Positives (TP):** 4356 - The model correctly predicted 4356 positive cases.
- **True Negatives (TN):** 1227 - The model correctly predicted 1227 negative cases.
- **False Positives (FP):** 1312 - The model incorrectly predicted 1312 negative cases as positive.
- **False Negatives (FN):** 749 - The model incorrectly predicted 749 positive cases as negative.

Performance Metrics:

- **Accuracy:** 0.73038 - The model correctly predicted 73.04% of the cases.
- **Recall:** 0.85328 - The model correctly identified 85.33% of the positive cases.
- **Precision:** 0.76853 - 76.85% of the positive predictions made by the model were correct.
- **F1-Score:** 0.80869 - The harmonic mean of precision and recall, indicating a balance between precision and recall.

Interpretation:

The model exhibits reasonable performance overall, **better than the first stacking.**

Good Recall: The model is relatively good at identifying positive cases, but it tends to overpredict positive cases, leading to a lower precision.

Model Performance Comparison and Final Model Selection

Training performance Comparison –

Training performance comparison:

	Decision Tree	Tuned Decision Tree	Bagging Classifier	Tuned Bagging Classifier	Random Forest	Tuned Random Forest	Adaboost Classifier	Tuned Adaboost Classifier	Gradient Boost Classifier	Tuned Gradient Boost Classifier	XGBoost Classifier	XGBoost Classifier Tuned	Stacking Classifier	Stacking Classifier2
Accuracy	1.00000	0.71255	0.98520	0.99619	1.00000	0.77377	0.73823	0.75443	0.75880	0.75617	0.85081	0.76239	0.74507	0.76205
Recall	1.00000	0.93192	0.98598	0.99992	1.00000	0.90649	0.88718	0.88391	0.88374	0.88525	0.93595	0.88743	0.86435	0.87115
Precision	1.00000	0.72007	0.99181	0.99441	1.00000	0.78710	0.76069	0.77844	0.78304	0.77957	0.85454	0.78491	0.77842	0.79300
F1	1.00000	0.81241	0.98889	0.99715	1.00000	0.84259	0.81908	0.82783	0.83035	0.82906	0.89339	0.83303	0.81914	0.83024

TABLE 6 – TRAINING DATA FINAL PERFORMANCE COMPARISON

Test performance Comparison –

Testing performance comparison:

	Decision Tree	Tuned Decision Tree	Bagging Classifier	Tuned Bagging Classifier	Random Forest	Tuned Random Forest	Adaboost Classifier	Tuned Adaboost Classifier	Gradient Boost Classifier	Tuned Gradient Boost Classifier	XGBoost Classifier	XGBoost Classifier Tuned	Stacking Classifier	Stacking Classifier2
Accuracy	0.66484	0.70657	0.69152	0.72423	0.72737	0.74424	0.73430	0.74110	0.74477	0.74372	0.72998	0.74477	0.73600	0.73038
Recall	0.74280	0.93085	0.76415	0.89540	0.84721	0.88619	0.88501	0.87600	0.87600	0.87875	0.85152	0.87522	0.86758	0.85328
Precision	0.75223	0.71545	0.77171	0.74386	0.76834	0.76704	0.75780	0.76865	0.77237	0.77000	0.76897	0.77274	0.76746	0.76853
F1	0.74749	0.80906	0.76791	0.81262	0.80585	0.82232	0.81648	0.81882	0.82093	0.82078	0.80814	0.82080	0.81445	0.80869

TABLE 7 – TEST DATA FINAL PERFORMANCE COMPARISON

While the **Decision Tree** and **Random Forest models** tend to overfit, their tuned versions offer a more balanced solution. The Gradient Boosting Classifier and XGBoost models, particularly when tuned, demonstrate consistent and dependable performance.

Lets have plot between Gradient boost and Tuned Random forest w.r.t to Important features for final selection –

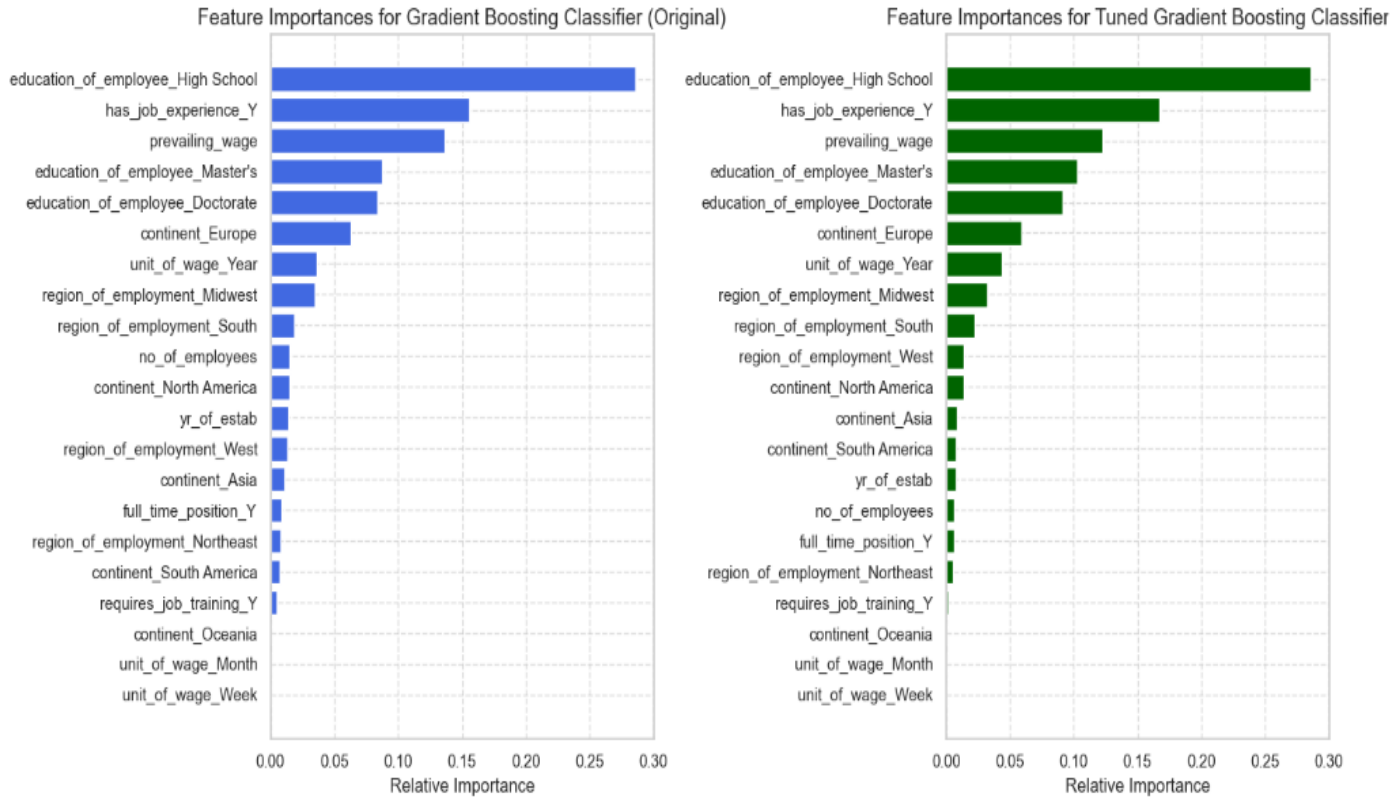


Figure 63 – Important features Gradient boost vs Tuned Random forest

Both the original and tuned Gradient Boosting Classifiers exhibit similar trends in feature importance. However, the **tuned model** seems to have refined the importance rankings, potentially leading to better performance.

Hence, Tuned Random Forest model is performing very well.

Actionable Insights & Recommendations –

Profile of Applicants Likely to Have Visa Approved when -

- Education Level: Higher education is strongly preferred. At a minimum, applicants should have a Bachelor's degree, with Master's and Doctoral degrees being especially desirable.
- Job Experience: Applicants should have some relevant job experience.
- Prevailing Wage: The median prevailing wage for employees whose visa applications are approved is around $72,000.

Also, other important factors can be -

- Unit of Wage: Applicants with an annual wage structure are more likely to be approved.
- Continent: Applicants from Europe, Africa, and Asia generally have a higher likelihood of visa certification.
- Region of Employment: Visa approvals are more frequent for applications to work in the Midwest. Furthermore, regional demands for specific qualifications are as follows:
  - High School: The South region has the highest demand, followed by the Northeast.
  - Bachelor's Degree: The South region leads in demand, with the West region following.
  - Master's Degree: The Northeast has the greatest demand, followed by the South.
  - Doctorate: The West region has the highest demand, with the Northeast in second

Profile of Applicants Likely to Have Visa Denied:

- Education Level: Applicants without any degree or those who have only completed high school.
- Job Experience: Applicants lacking job experience are more likely to be denied.
- Prevailing Wage: The median prevailing wage for employees whose visa applications are denied is around $65,000.

Also, other important factors can be -

- Unit of Wage: Applicants with an hourly wage structure are more likely to face visa denial.

- Continent: While nationality or ethnicity should not ideally influence visa decisions, applicants from South America, North America, and Oceania have historically had a higher rate of visa denials.

Insights:

1. Education and Experience Matter: Feature importance analysis reveals that education level and job experience are key predictors of visa approval. This suggests that applicants with higher education and relevant experience are more likely to receive approval, supporting policy objectives aimed at attracting skilled talent.
2. Wage as an Indicator: Prevailing wage, particularly for hourly positions, also plays a significant role in the likelihood of visa approval. Higher wages are associated with a greater chance of approval, potentially reflecting the demand for specialized skills in higher-paying roles.
3. Model Performance: The tuned Gradient Boosting Classifier demonstrates strong performance across multiple metrics (accuracy, recall, precision, F1 score), indicating it is a reliable model for predicting visa approvals. The model's balanced performance ensures both precision and fairness, making it a valuable tool.
4. Interestingly, attributes such as whether the job opportunity is full-time or part-time, whether an employee requires additional job training, the annual prevailing wage for the occupation in the US, the year the employer was established, or the number of employees in the organization are not significant factors and do not have a substantial impact on whether a case is certified or denied.

Recommendations:

1. Prioritize High-Impact Applications: With limited resources, the OFLC could prioritize reviewing applications from individuals with higher education, relevant experience, and higher prevailing wages, as these factors are most likely to result in approval.
2. Model Deployment: Deploy the tuned Gradient Boosting Classifier as a decision support tool to assist visa officers in making more informed, data-driven decisions. It’s important, however, to ensure that the final decision always incorporates human judgment and adheres to legal and ethical standards.
3. Continuous Monitoring and Improvement: Regularly evaluate and update the model using new data to ensure its continued accuracy and relevance. This includes revisiting feature importance to adapt to changes in the labor market and immigration policies.
4. Transparency and Fairness: Ensure that the model’s use remains transparent, avoiding any bias against certain applicant groups. Regular audits for fairness and bias should be conducted to maintain the integrity of the visa approval process and uphold public trust.