MACHINE LEARNING – 2

EASY VISA PROJECT STATEMENT

By: BENITA MERLIN.E

PGP-Data Science and Business Analytics.

BATCH: PGP DSBA. O. MAY24.A

Contents

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 PROJECT OVERVIEW

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.

1.2. OBJECTIVE

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

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

Facilitate the process of visa approvals. 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.

1.3 PROBLEM DEFINITION

- Build a Machine Learning model to predict the **visa case status** (Certified or Denied).
- Automate the process of shortlisting candidates likely to get visa certification.
- Provide recommendations to employers based on the key factors influencing visa approval.

2. DATA DESCRIPTION

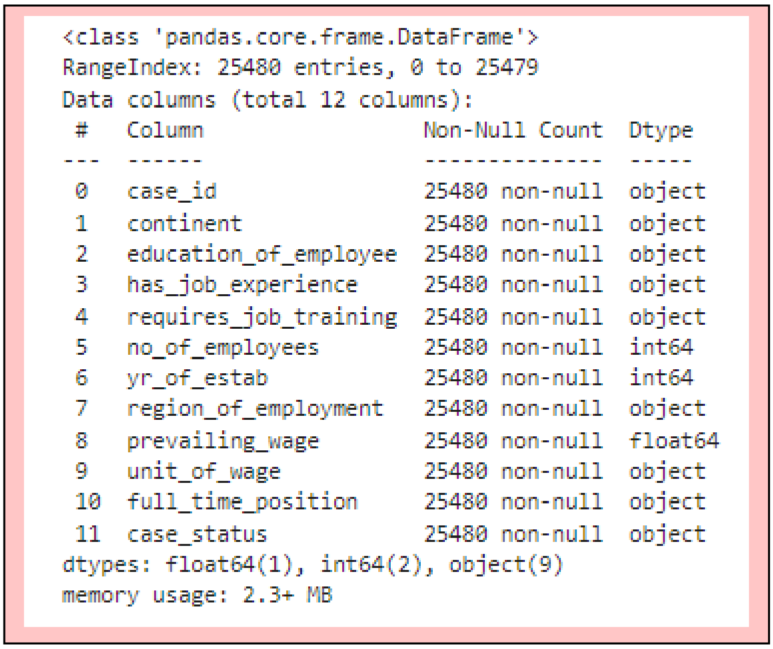
The data contains the different attributes of the employee and the employer. The detailed data dictionary is given below

2.1. DATA DICTIONARY

SI.NO	VARIABLE	DESCRIPTION
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

Table 1 Data Dictionary

2.2 DATA INFORMATION



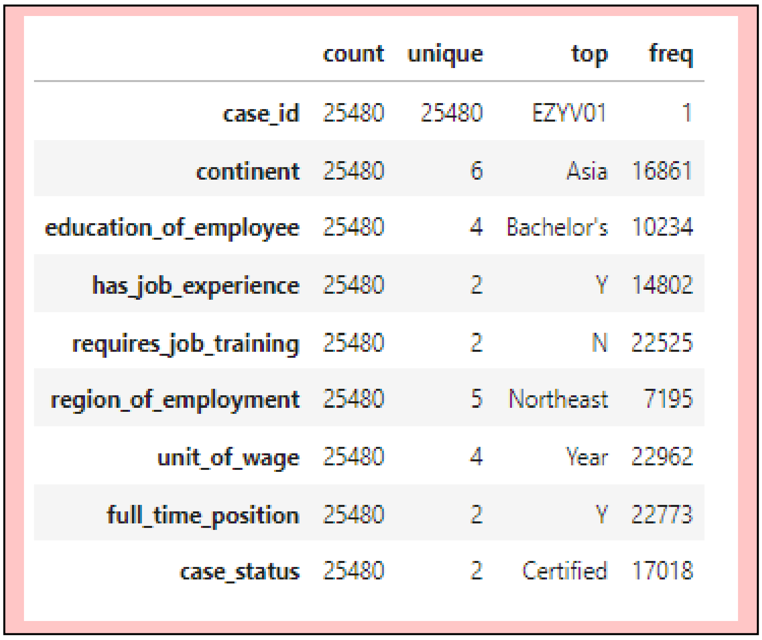
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25480 entries, 0 to 25479
Data columns (total 12 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   case_id                25480 non-null  object 
 1   continent              25480 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 Data Information

- 9 columns are categorical (object type), including variables like continent, case_status, and has_job_experience.
- 2 columns are numerical (no_of_employees, yr_of_estab).
- 1 column is continuous (prevailing_wage).
- The target variable for the models is case_status.
- case_id is a randomly assigned by INA for each case, which should not provide any material information for model building.
- No Missing Values: All columns have complete data, with no null values.
- No duplicate values.

3. STATISTICAL ANALYSIS

The statistical analysis provides a summary of the key metrics for the numerical columns in the dataset. This includes measures such as the count, mean, standard deviation, minimum, 25th percentile (Q1), median (50th percentile, Q2), 75th percentile (Q3), and maximum values for each column.



	count	unique	top	freq
case_id	25480	25480	EZYV01	1
continent	25480	6	Asia	16861
education_of_employee	25480	4	Bachelor's	10234
has_job_experience	25480	2	Y	14802
requires_job_training	25480	2	N	22525
region_of_employment	25480	5	Northeast	7195
unit_of_wage	25480	4	Year	22962
full_time_position	25480	2	Y	22773
case_status	25480	2	Certified	17018

Table 3 Statistical Summary

Observation:

- A significant majority of visa applicants come from Asia (66.2%).
- Bachelor’s degree holders form the largest group (40.2%) of applicants.
- Most applicants have prior job experience (58.1%) and do not require job training (88.4%).
- The Northeast region is the most common employment destination.
- Yearly wages are the dominant wage structure (90.1%).
- Most positions are full-time (89.3%).
- Visa certifications are more common than denials, with 66.8% of applications certified.

4. EXPLORATORY DATA ANALYSIS (EDA)

4.1. UNIVARIATE ANALYSIS

4.1.1. CASE STATUS

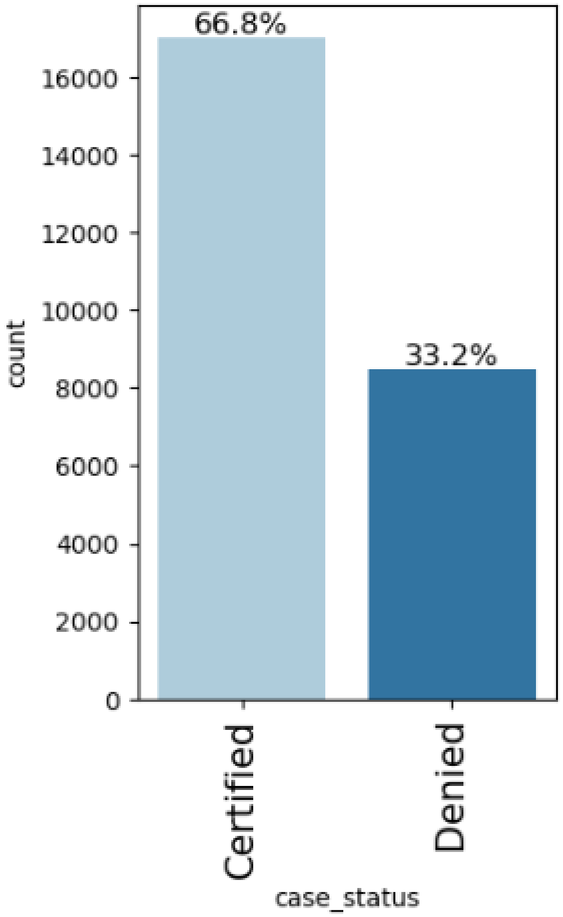


Figure 1 Bar plot of Case Status

OBSERVATION:

- Almost two-thirds of the visa applications are certified.

4.1.2. NO_OF_EMPLOYEES

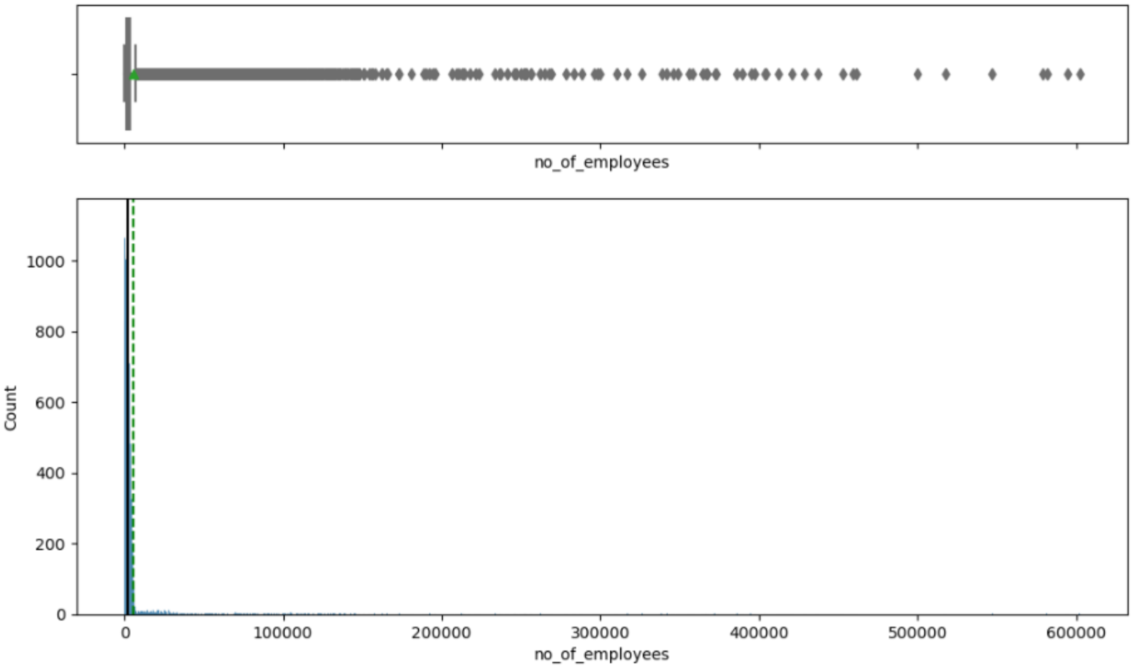


Figure 2 Distribution of No. of Employees Analysis

Top Plot (Boxplot):

- This is a boxplot illustrating the spread and outliers of the number of employees.
- The plot reveals a large number of extreme outliers, as indicated by many points far beyond the box.
- The bulk of the data is concentrated near the lower range, as shown by the compressed box and whiskers.
- A green line likely represents the median, which is close to the lower end of the data distribution.

Bottom Plot (Histogram):

- This is a histogram showing the frequency distribution of the number of employees.
- Most of the values are clustered near 0, with very few counts as the number of employees increases.
- The histogram confirms the boxplot's observation, highlighting a large skew towards smaller employee counts, with a few cases extending into very high numbers (long right tail).
- The green dashed line appears to represent a statistical measure such as the mean, which is also skewed due to the outliers.

Observation

- There is an extremely skewed distribution with most companies having a small number of employees, while a few have very large workforces.
- The large number of outliers suggests that while the majority of companies are small, there are some very large companies, creating a long tail in the data.
- The dataset likely contains many more small or medium-sized companies compared to large corporations.

4.1.3. YR_OF_ESTAB

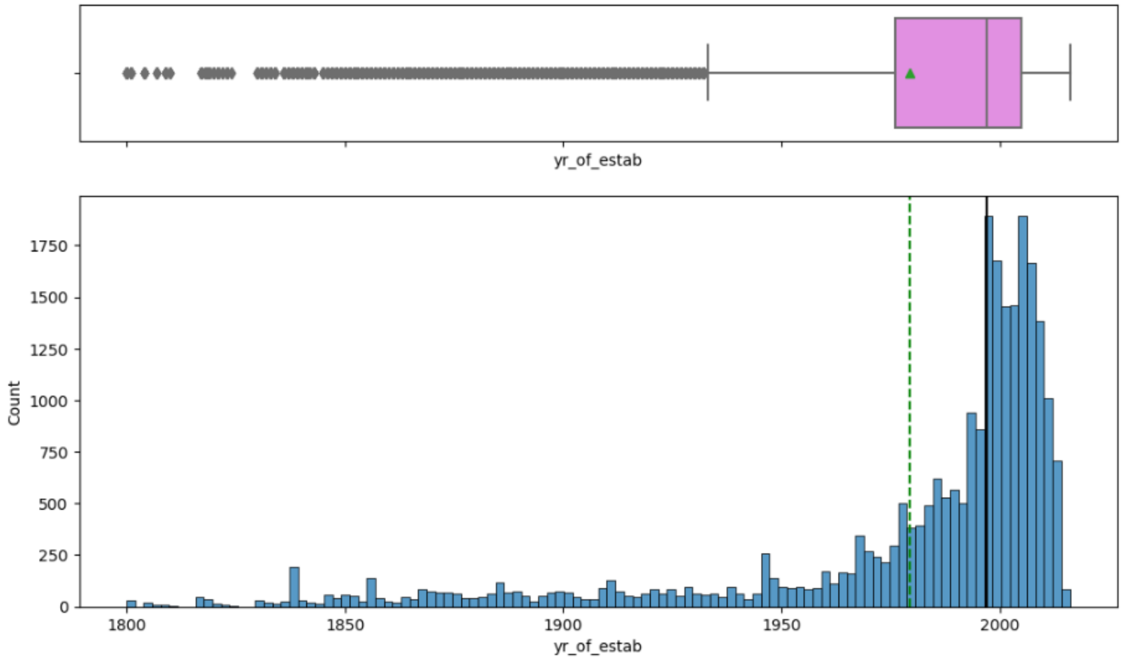


Figure 3 Distribution of yr_of_estab Analysis

Top Plot (Boxplot):

- The boxplot shows the spread of the establishment years, with a range of data points concentrated in the later years (1900 onwards).
- There are several outliers extending far to the left, representing entities established a long time ago (pre-1800s).
- The box itself represents the interquartile range (IQR), and the line within the box is the median, indicating that most entities were established in recent decades.
- A green triangle likely represents the mean, slightly skewed towards the more recent years.

Bottom Plot (Histogram):

- The histogram shows the frequency distribution of establishment years.

- There is a significant increase in the number of entities established from around the mid-1900s onwards, with a sharp rise in recent decades.
- The counts rise dramatically after the 1950s and peak around the late 1990s to early 2000s.
- A green dashed line indicates the mean, which aligns with the mid to late 20th century, matching the sharp increase seen in the histogram.

Observations:

- The data is right-skewed, indicating that most entities were established more recently, with fewer organizations being much older.
- Outliers are seen in the boxplot for organizations established much earlier, but these represent a small portion of the overall data.
- The majority of entities in this dataset were established after 1950, with a particularly large number in the last few decades.

4.1.4. PREVAILING_WAGE

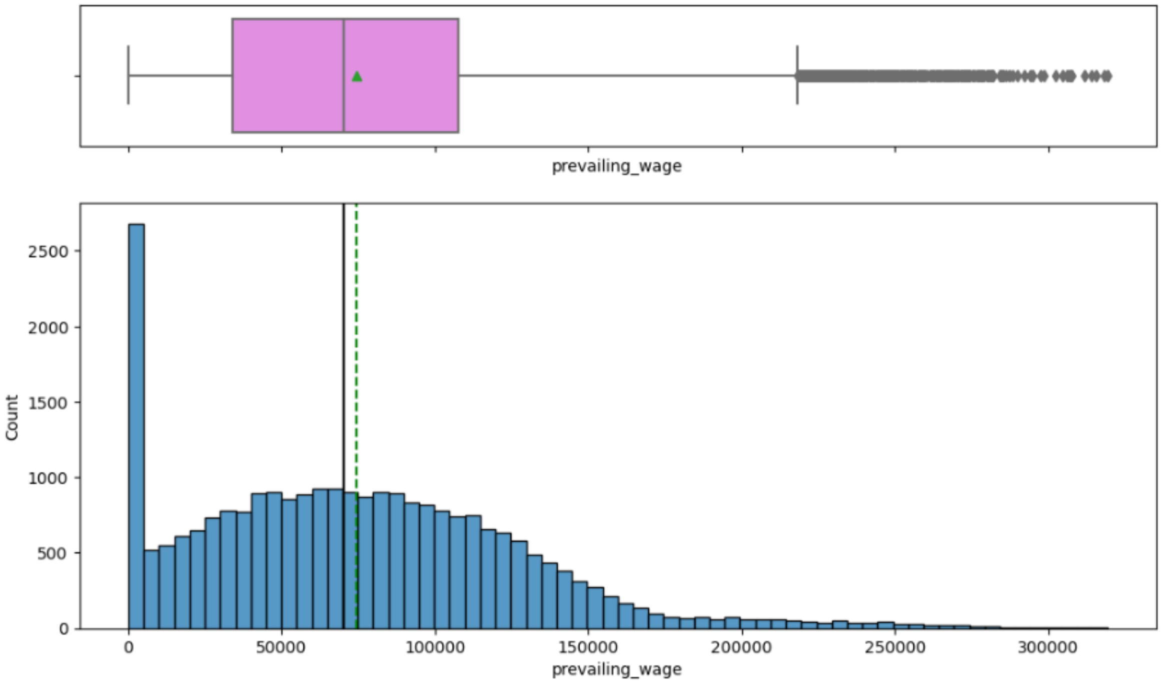


Figure 4 Distribution of Prevailing wage Analysis

Top Plot: Boxplot of prevailing_wage

Boxplot: This shows the five-number summary (minimum, first quartile, median, third quartile, maximum).

- The box represents the interquartile range (IQR), showing the spread of the middle 50% of the data.
- The green triangle represents the mean, while the line inside the box is the median.

- Whiskers extend to show the range of data within 1.5 times the IQR.
- There are outliers in the distribution (data points outside the whiskers, mostly on the higher end), indicating some extreme wage values.

Bottom Plot: Histogram of prevailing_wage

Histogram: The histogram represents the frequency distribution of the prevailing_wage.

- There is a large peak at zero, suggesting a significant number of entries with a prevailing_wage of zero.
- The distribution is right-skewed with a long tail extending towards higher wage values (e.g., beyond $100,000).
- The green dashed line likely represents the mean wage.

Key Observations:

- Outliers: The presence of outliers in the boxplot suggests that there are extreme values, which may need further investigation or treatment (e.g., capping or log transformation).
- Zero Wage: A significant number of data points have a prevailing_wage of zero, which could either be missing data or cases where wages weren't recorded.
- Skewness: The right skewness of the wage distribution indicates that most wages are clustered in the lower range, with a few very high wage values.

4.1.5. CONTINENT

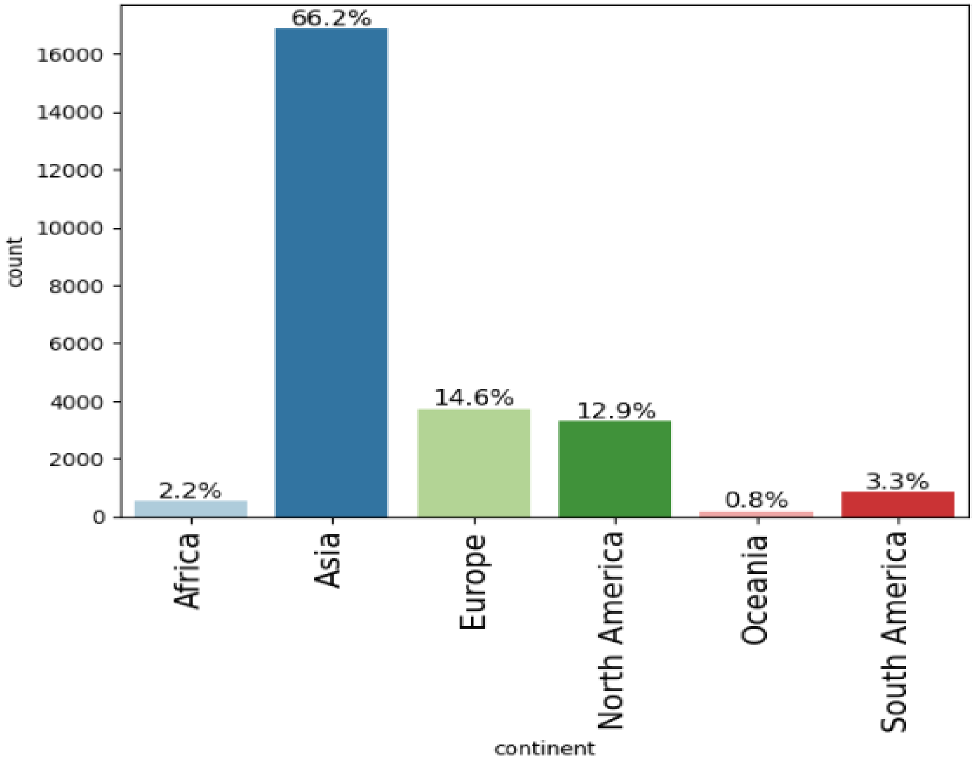


Figure 5 Bar plot of Continent

OBSERVATION:

- The majority (66%) of the visa applicants are from Asia, which makes sense given the high population of this continent.
- The lowest fraction (<1%) of the applicants are from Oceania, which also makes sense given its very low population.
- North America and Europe have close number of applicants (12.9% and 14.6%).

4.1.5. HAS_JOB_EXPERIENCE

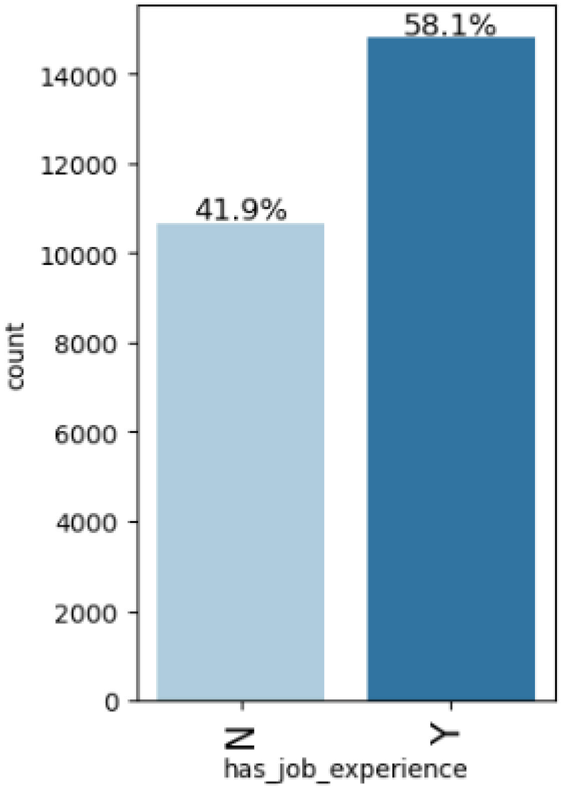


Figure 6 Bar plot of Has Job Experience

OBSERVATION:

- More than half (58%) of the applicants have job experience.

4.1.6. EDUCATION OF EMPLOYEE

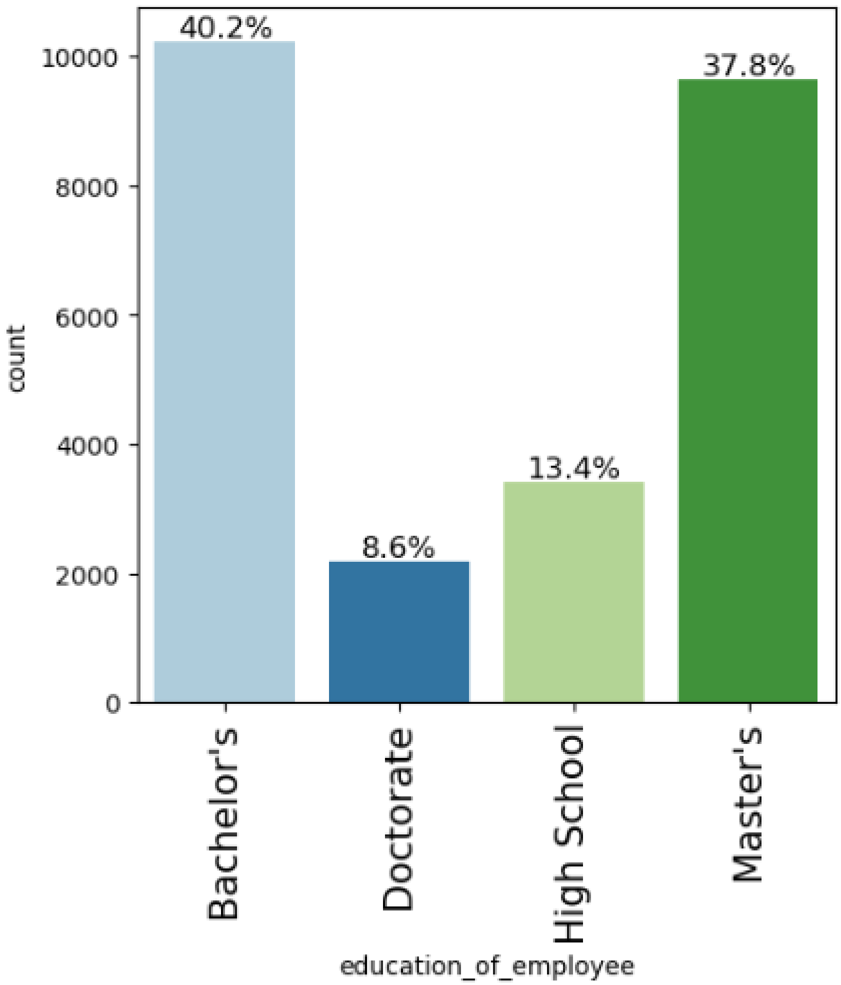


Figure 7 Bar plot of Education of Employee

OBSERVATION:

- The majority of the applicants have either bachelor's degrees (40.2%) or master's degrees (37.8%).
- Only 8.6% of the applicants have doctorate degrees.

4.1.7. REQUIRES_JOB_TRAINING

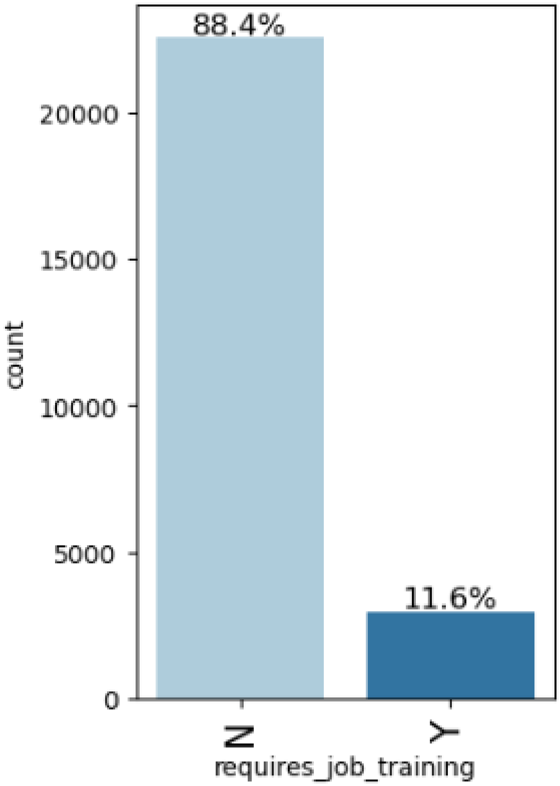


Figure 8 Bar plot of Requires Job Training

OBSERVATION:

- The vast majority (>88%) of the jobs do not require the applicants to receive training.

4.1.8. REGION_OF_EMPLOYMENT

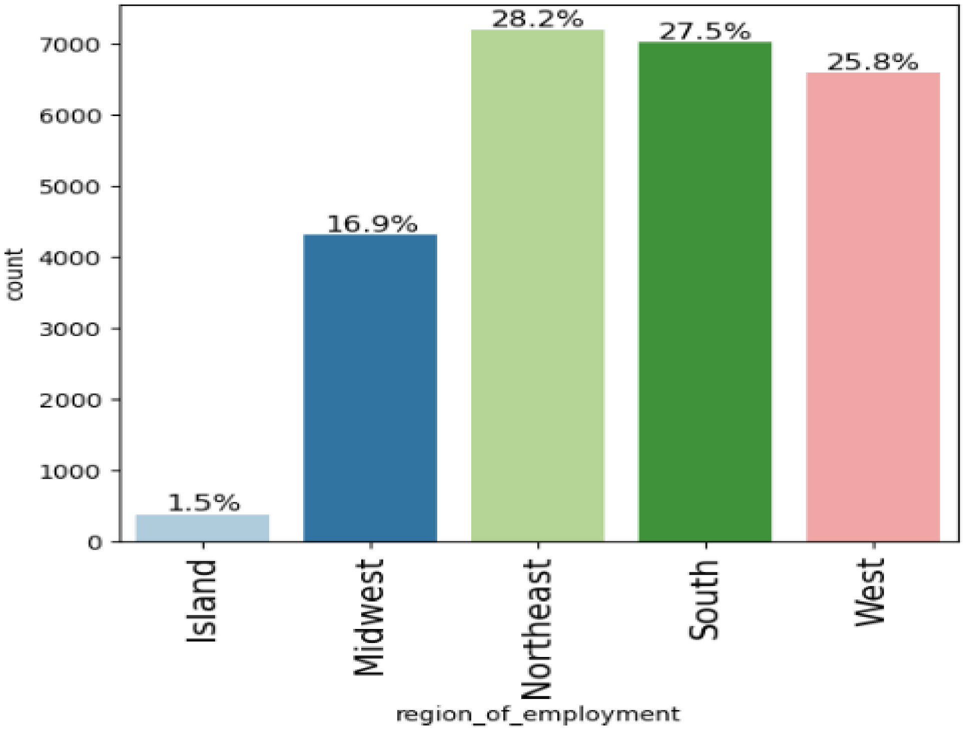


Figure 9 Bar plot of Region of Employment

OBSERVATION:

- Most of the applications are for employment in the Northeast, South, and West regions of the United States.
- This could be expected because the majority of the tech companies are in those regions and the populations of those regions are higher than the other regions of the United States.
- The Island region has the lowest number (1.5%) of work visa applicants.

4.1.9. FULL_TIME_POSITION

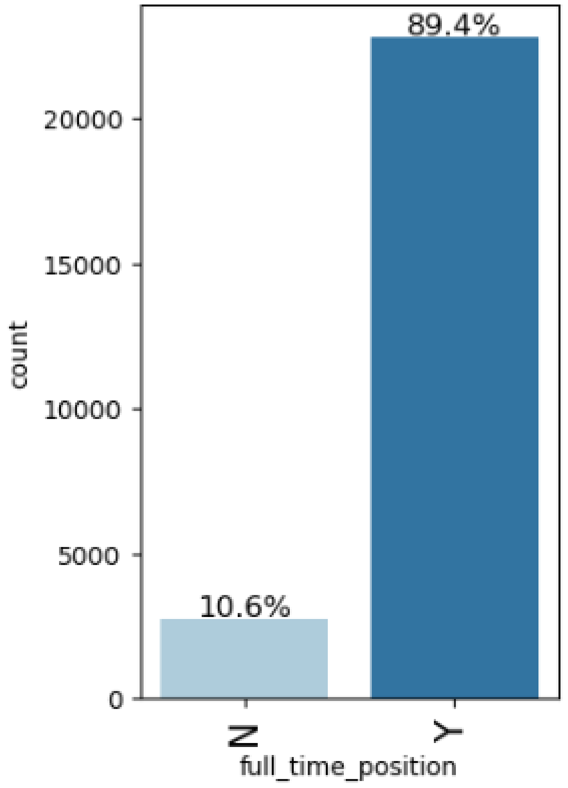


Figure 10 Bar plot of Full Time Position

OBSERVATION:

- More than 89% of the applications are related to full-time employment.

4.1.10. UNIT_OF_WAGE

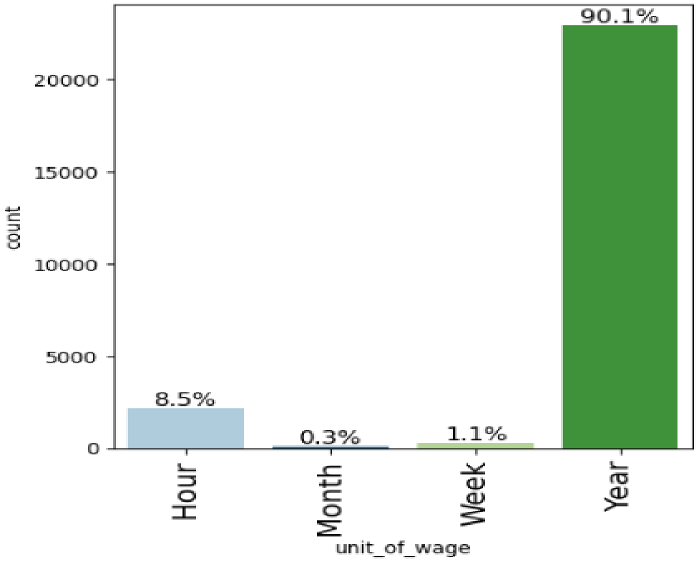
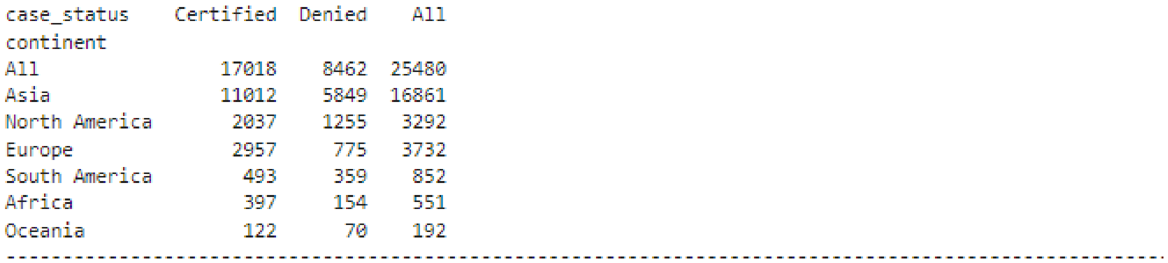


Figure 11 Bar plot of Unit of Wage

- The dominant majority (90%) of the applications are for the jobs whose prevailing wages are computed per year.

4.2. BIVARIATE ANALYSIS

4.2.1. CONTINENT VS CASE_STATUS



```
case_status    Certified  Denied    All
continent
All                17018    8462  25480
Asia               11012    5849  16861
North America       2037    1255   3292
Europe              2957     775   3732
South America        493     359    852
Africa               397     154    551
Oceania              122      70    192
------------------------------------------------------------------------------------------------------------------------
```

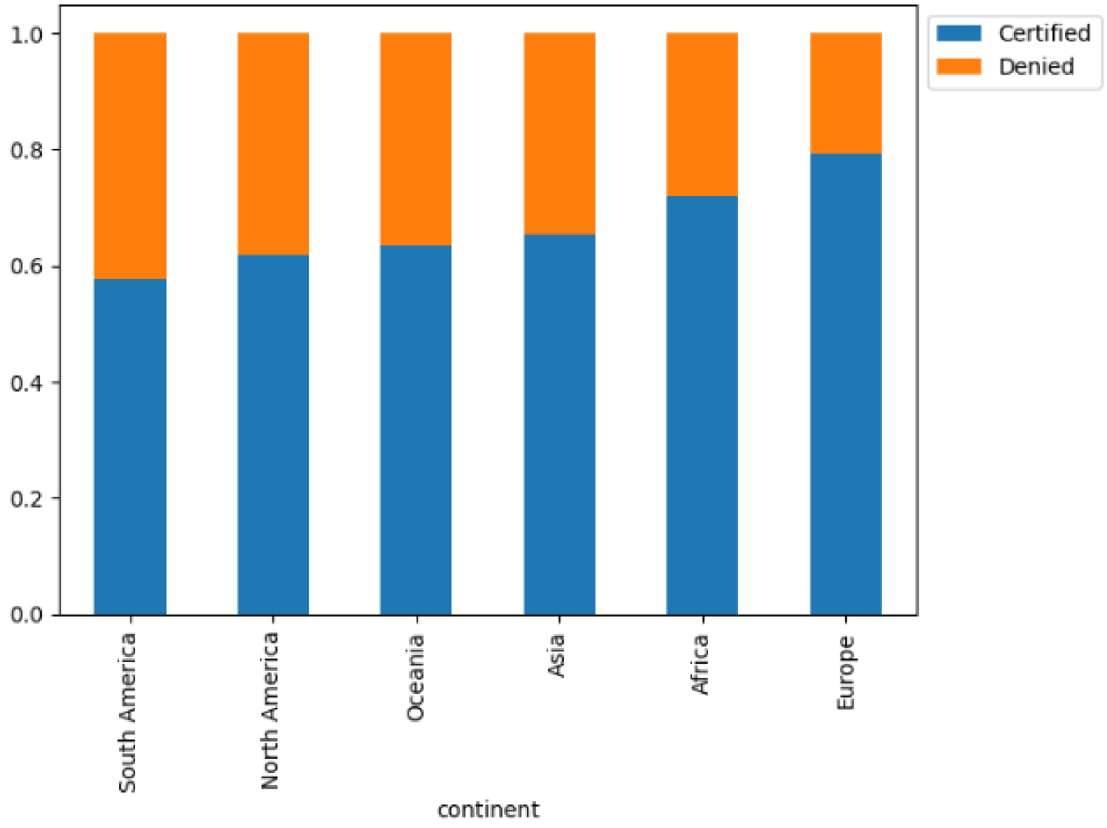
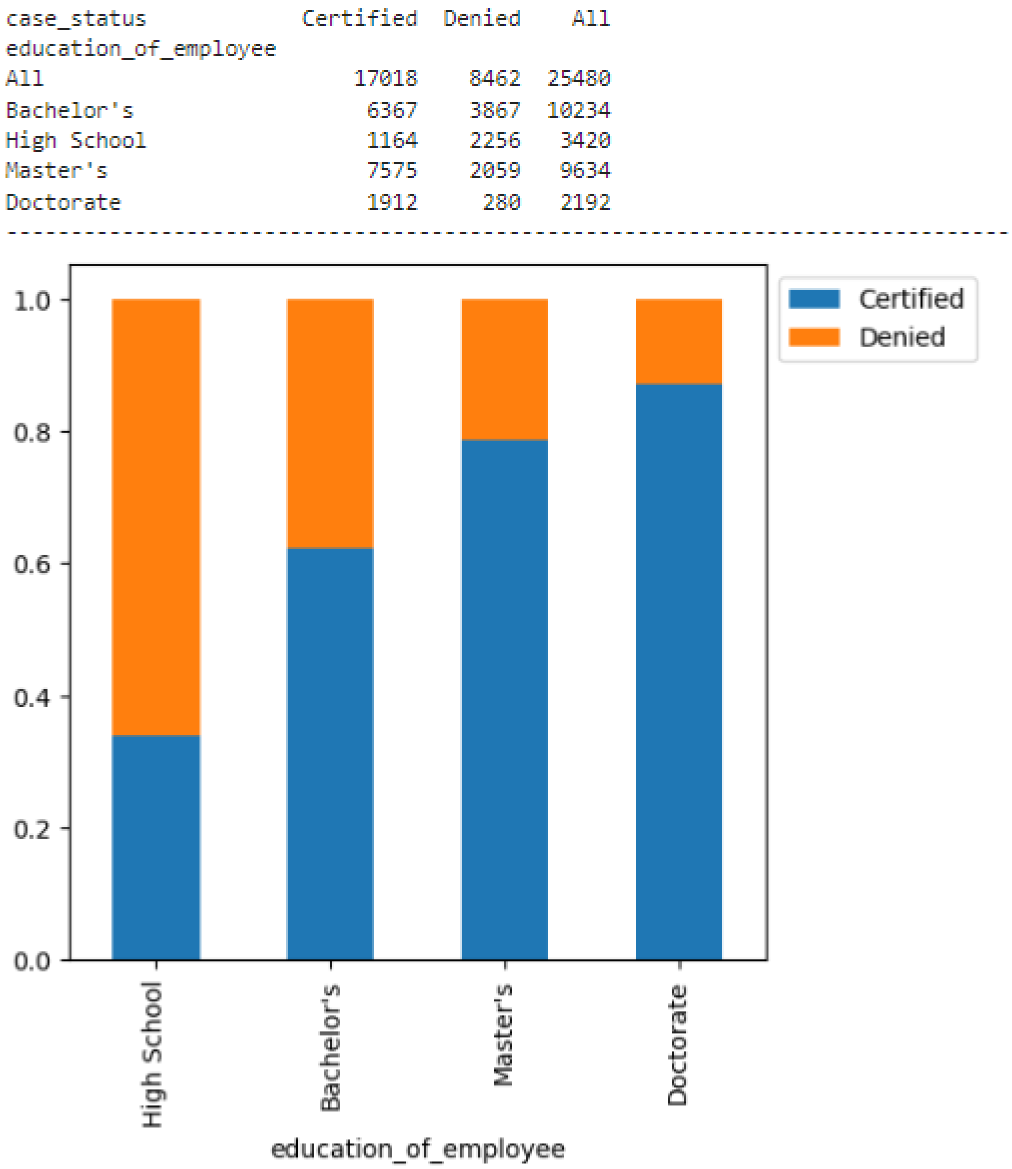


Figure 12 Continent vs Case status

OBSERVATION:

- Among different continents, Europe has the highest work visa certification rate (79%).
- The lowest work visa certification rate belongs to South America (58%).

4.2.2. EDUCATION OF EMPLOYEE VS CASE_STATUS



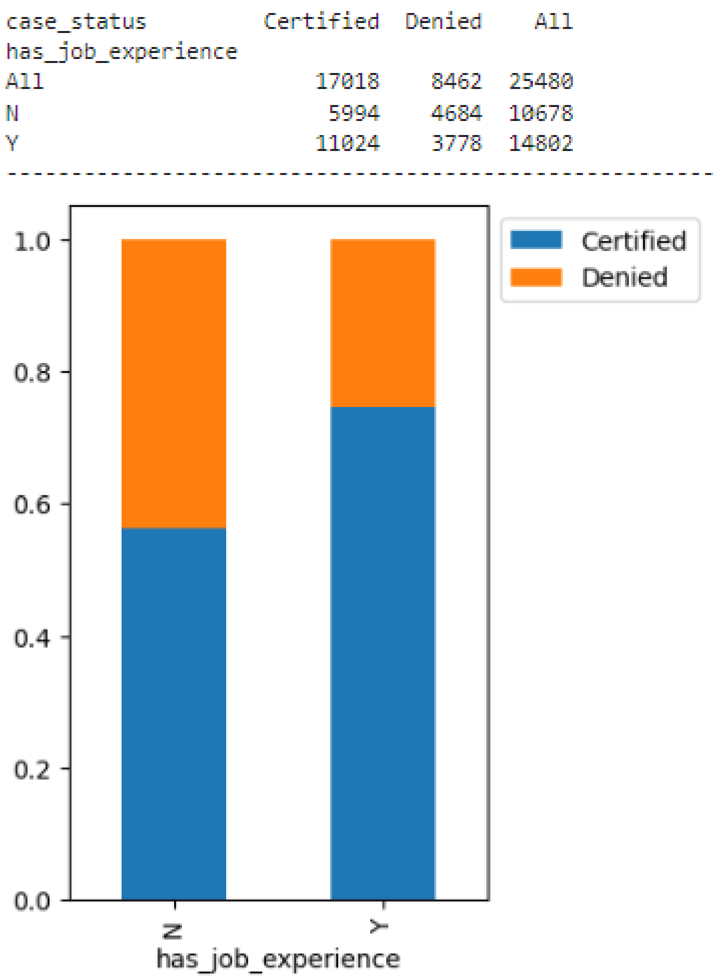
```
case_status           Certified  Denied    All
education_of_employee
All                       17018    8462  25480
Bachelor's                 6367    3867  10234
High School                1164    2256   3420
Master's                   7575    2059   9634
Doctorate                  1912     280   2192
------------------------------------------------------------------------------------------------------------------------
```

Figure 13 Education of Employee Vs Case Status

OBSERVATION:

- It is clear that the higher the education level of an applicants is, the more their chances of visa certification are.
- More specifically, while the visa certification likelihood of the applicants of a doctorate degree is 87%, this likelihood is only 34% for the applicants of high school education.

4.2.3. HAS_JOB_EXPERIENCE VS CASE_STATUS

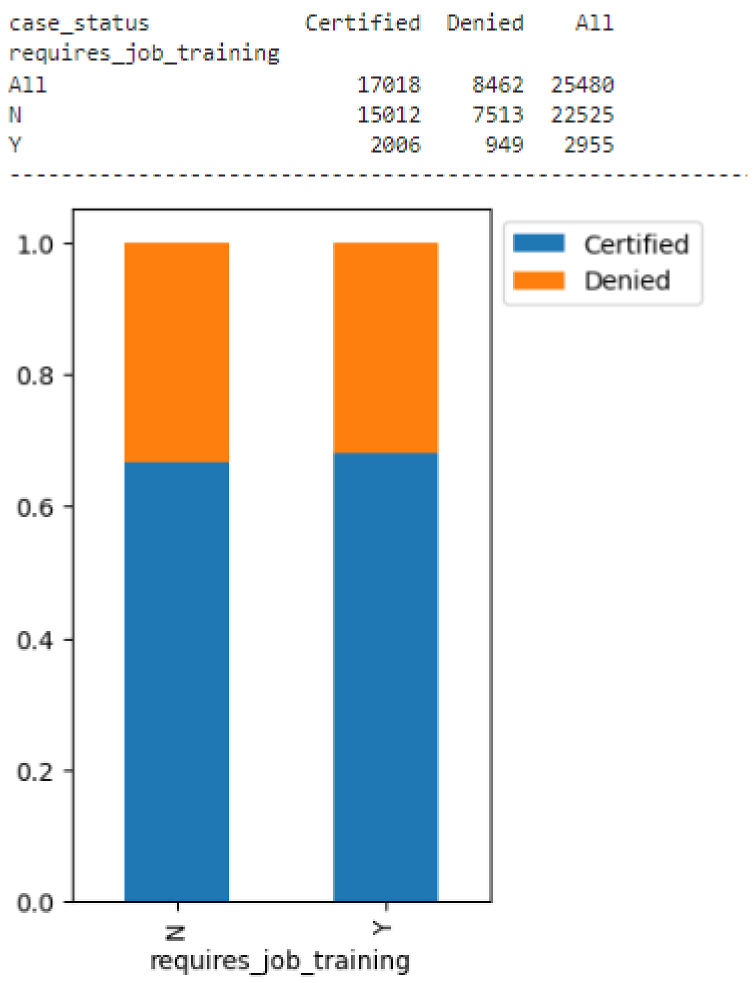


```
case_status         Certified  Denied    All
has_job_experience
All                     17018    8462  25480
N                        5994    4684  10678
Y                       11024    3778  14802
------------------------------------------------------------
```

Figure 14 Has Job Experience Vs Case Status

- Having job experience is found to have a positive effect on the visa certification likelihood.
- More specifically, about 74% of the experienced applicants are granted visas, while this percentages is only 56% for the inexperienced applicants.

4.2.4. REQUIRES_JOB_TRAINING VS CASE_STATUS



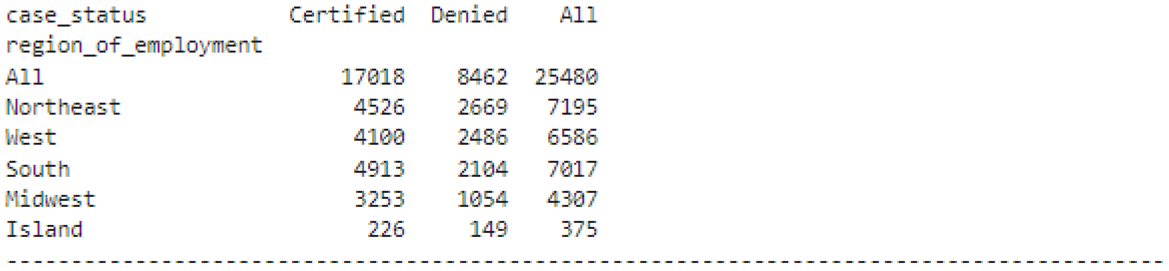
```
case_status           Certified  Denied    All
requires_job_training
All                       17018    8462  25480
N                         15012    7513  22525
Y                          2006     949   2955
------------------------------------------------------------
```

Figure 15 Requires Job Training Vs Case Status

OBSERVATION:

- The visa certification likelihood is found nearly unaffected by the job training requirement.

4.2.5. REGION_OF_EMPLOYMENT VS CASE_STATUS



```
case_status           Certified  Denied    All
region_of_employment
All                       17018    8462  25480
Northeast                  4526    2669   7195
West                       4100    2486   6586
South                      4913    2104   7017
Midwest                    3253    1054   4307
Island                      226     149    375
------------------------------------------------------------------------------------------
```

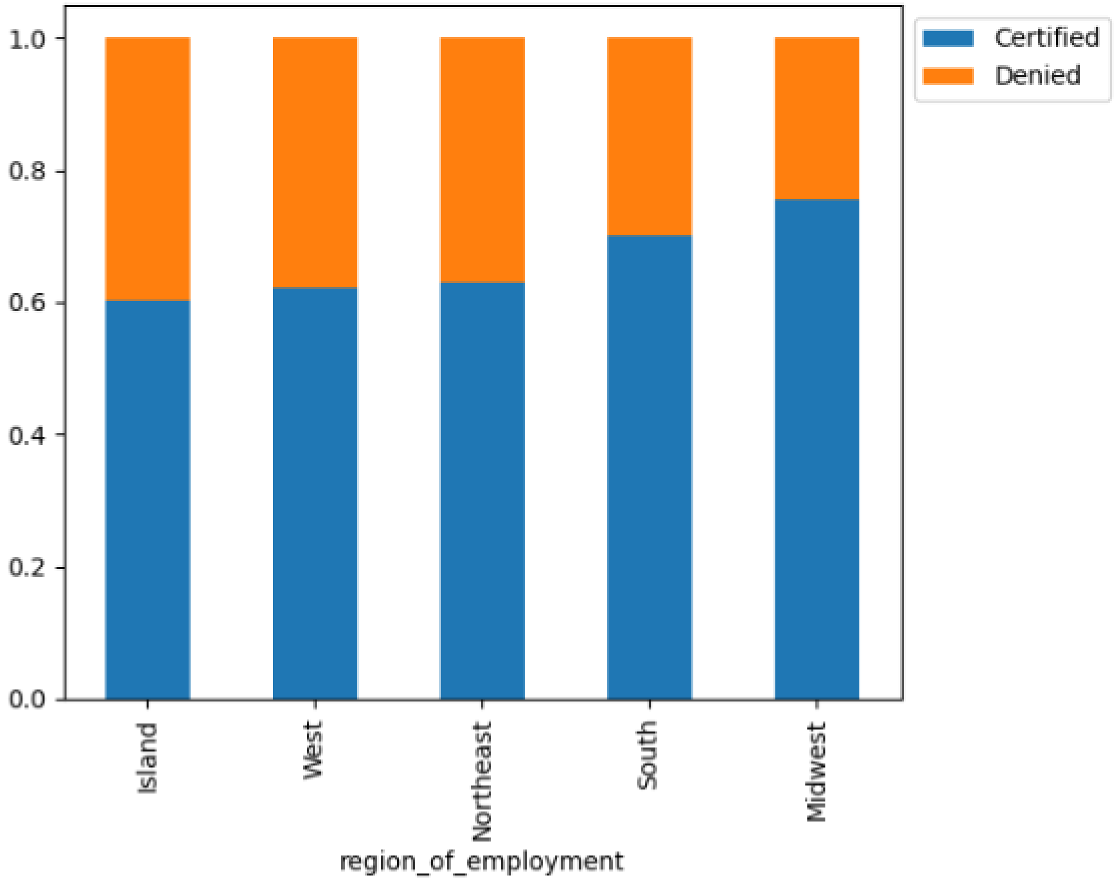
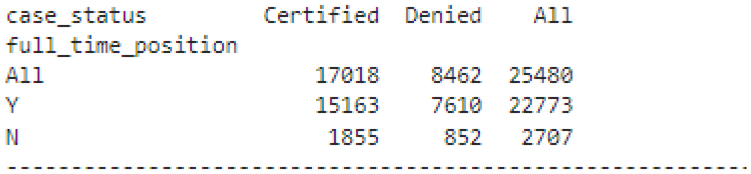


Figure 16 Region of Employment Vs Case Status

OBSERVATION:

- It appears that the visa applications filed by the employers within the Midwest region have the highest probability (~76%) of certification.
- The employers located in the Northeast, West, and Island regions have lower chances (60-63%) of visa certification.

4.2.6. FULL_TIME_POSITION VS CASE_STATUS



```
case_status        Certified  Denied    All
full_time_position
All                    17018    8462  25480
Y                      15163    7610  22773
N                       1855     852   2707
------------------------------------------------------------
```

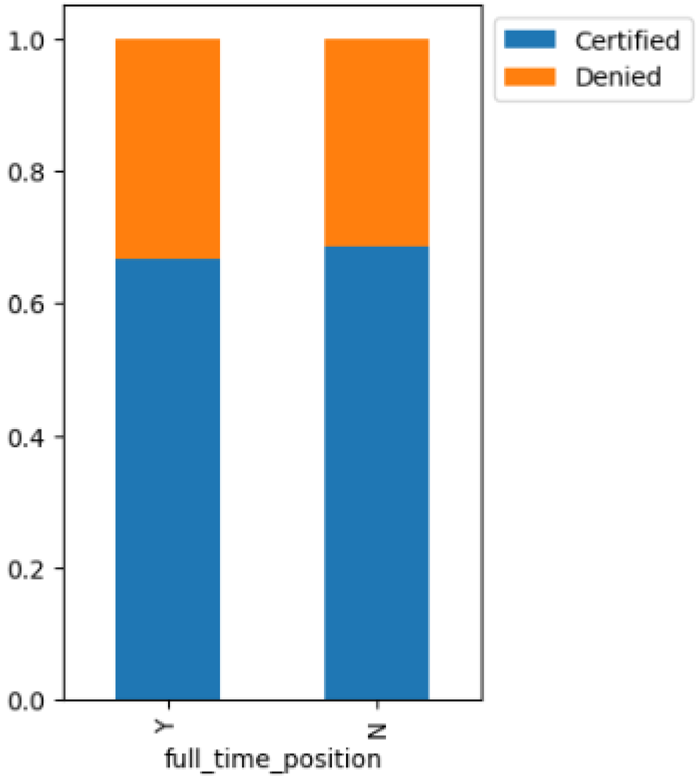
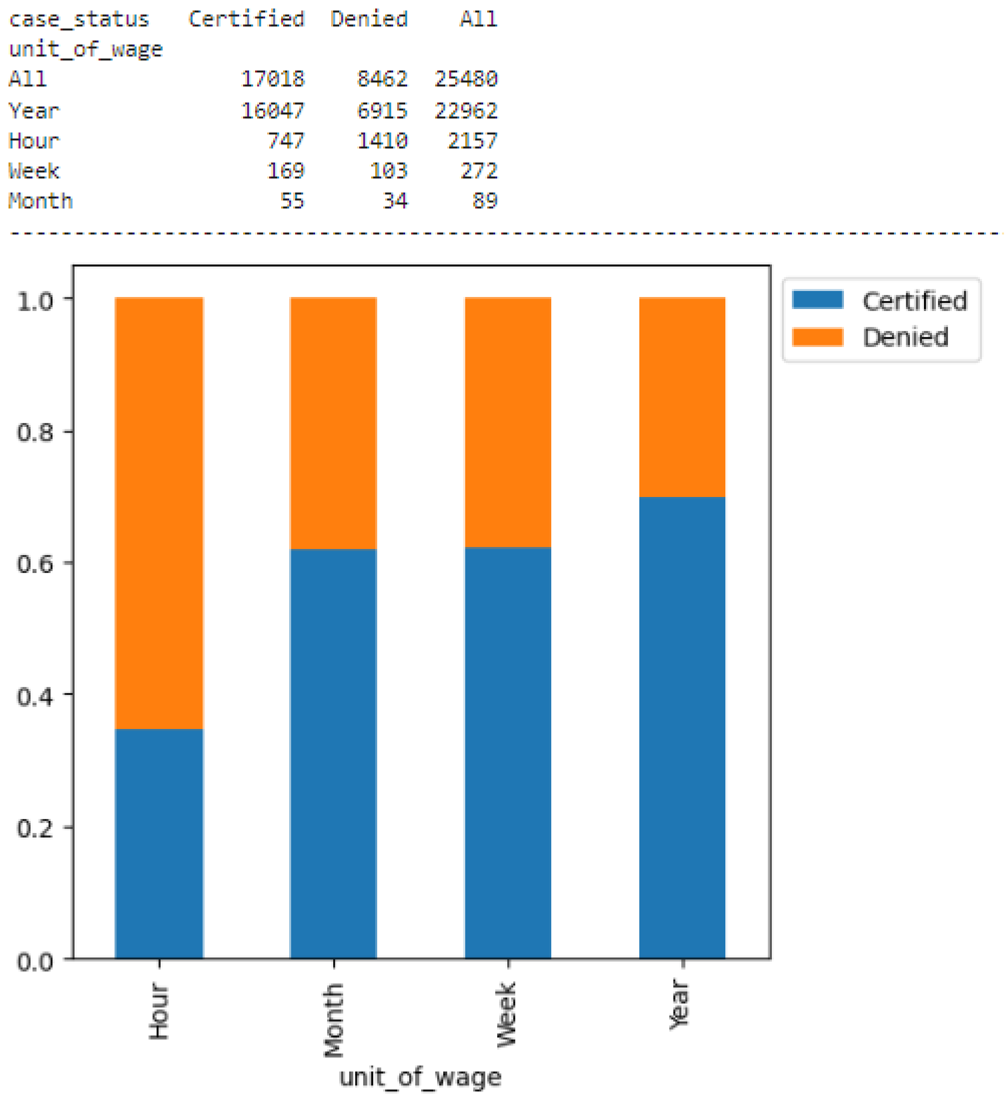


Figure 17 Full Time Position Vs Case Status

OBSERVATION:

- Visa certification seems to be unaffected by whether a position is full-time or part-time.

4.2.7. UNIT_OF_WAGE VS CASE_STATUS



```
case_status    Certified  Denied    All
unit_of_wage
All                17018    8462  25480
Year               16047    6915  22962
Hour                 747    1410   2157
Week                 169     103    272
Month                 55      34     89
------------------------------------------------------------------------------
```

Figure 18 Unit of Wage Vs Case Status

OBSERVATION:

- Those applicants whose wage unit is year are more likely than other applicants to be certified for a visa (~70% likelihood).
- The applicants who are paid by hour are the least likely to be certified for a visa (~35% likelihood).
- This could be predicted, because hourly jobs are usually less important for the growth of the United States and they could be done by normal American workers.

4.2.8. NO_OF_EMPLOYEES VS CASE_STATUS

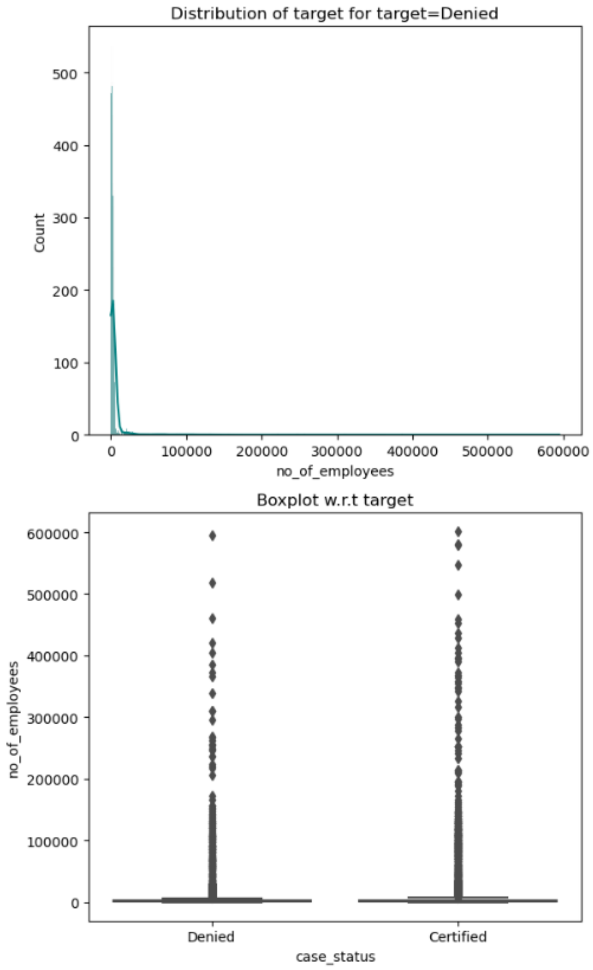
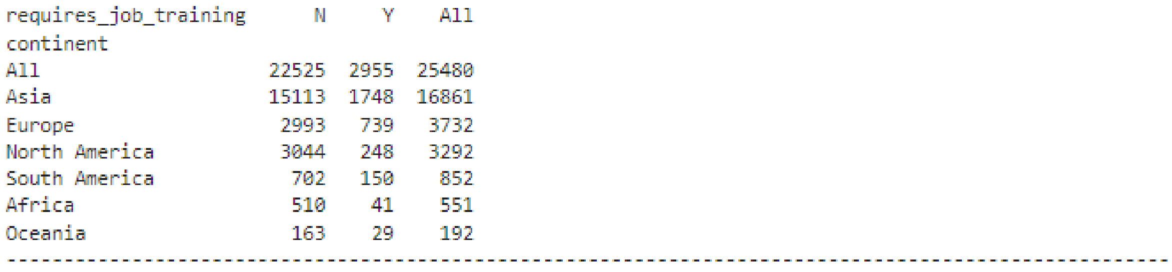


Figure 19 Number of Employees Vs Case Status

OBSERVATION:

- A very small difference is observed between the distributions of the employer's number of employees for those applications that are denied and those that are certified.
- As a result, it seems that the number of employees has insignificant effect on the likelihood of visa certification.

4.2.9. CONTINENT VS REQUIRES_JOB_TRAINING



```
requires_job_training      N     Y    All
continent
All                    22525  2955  25480
Asia                   15113  1748  16861
Europe                  2993   739   3732
North America           3044   248   3292
South America            702   150    852
Africa                   510    41    551
Oceania                  163    29    192
------------------------------------------------------------------------------------------
```

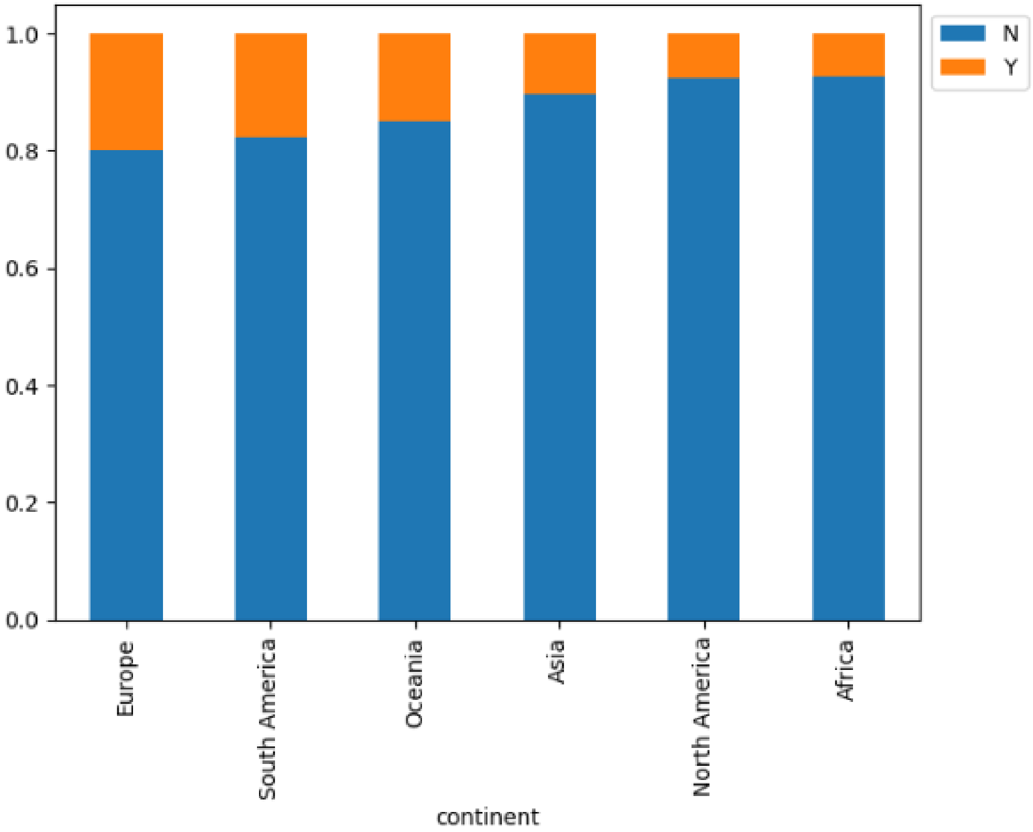


Figure 20 Continent Vs Requires Job Training

OBSERVATION:

- Among the applicants from different continents, a smaller ratio of those from Africa and North America need training than those from other continents.
- The highest ratio of the applicants who need training belongs to those from Europe.

4.3. MULTIVARIATE ANALYSIS

4.3.1. HEATMAP

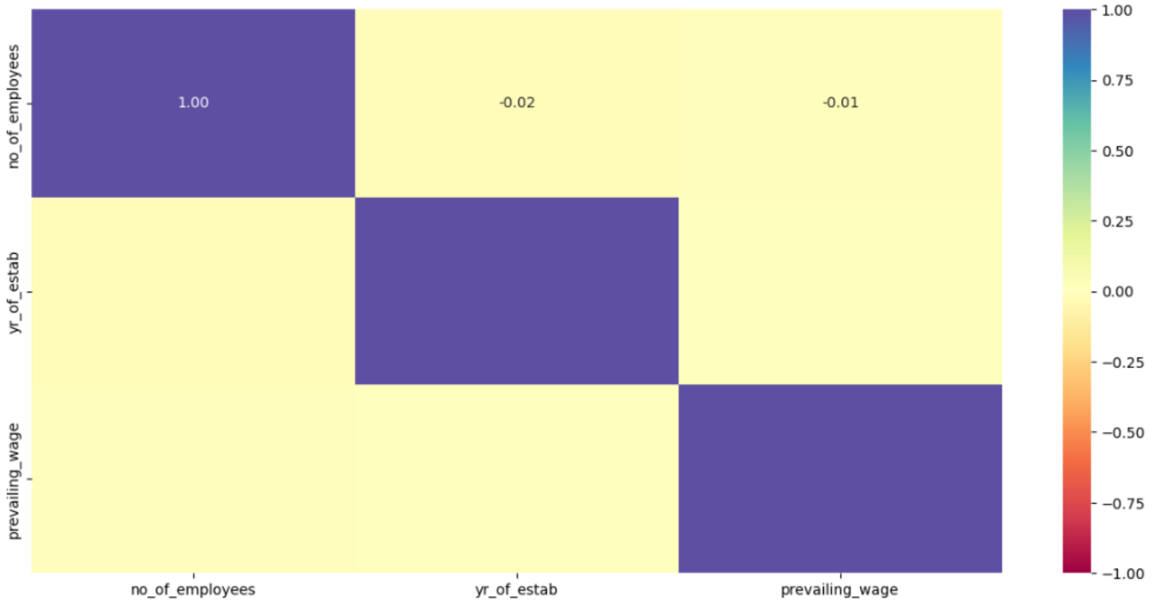


Figure 21 Heat Map

OBSERVATION:

- no_of_employees and yr_of_estab: Correlation is -0.02. This near-zero value suggests no significant linear relationship between the number of employees and the year the establishment was founded.
- no_of_employees and prevailing_wage: Correlation is -0.01. Again, this indicates a very weak and almost negligible negative relationship between the number of employees and the prevailing wage.
- yr_of_estab and prevailing_wage: Correlation is -0.01, meaning there is almost no linear relationship between the year the company was established and the prevailing wage.

4.3.2. PAIRPLOT

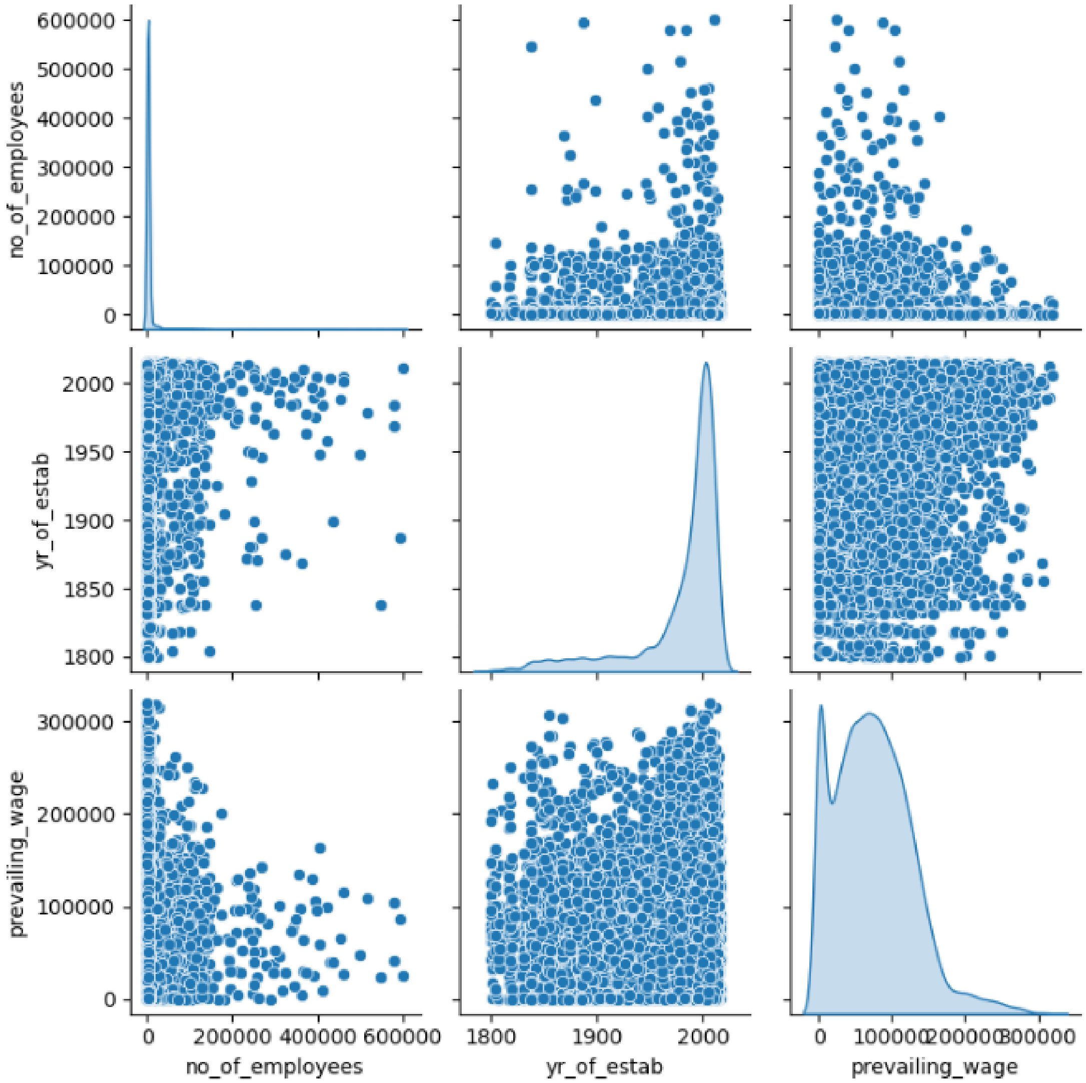


Figure 22 Pair Plot of Numeric Variables

OBSERVATION:

no_of_employees vs prevailing_wage:

- There's no strong linear relationship visible. Companies with a small number of employees dominate the data, and their wages seem to vary broadly.
- Some larger companies offer wages across the full range, but there's no clear upward trend.

yr_of_estab vs prevailing_wage:

- A slight positive relationship can be observed here. Newer companies (founded in more recent years) tend to offer higher wages compared to older companies.
- However, older companies still have a spread of wages, though concentrated on the lower side.

no_of_employees vs yr_of_estab:

- Most newer companies (post-2000) have a smaller number of employees, but some older companies show a larger workforce.
- There seems to be no strong relationship between the age of the company and the number of employees, indicating that the size of a company might not depend solely on its establishment year.

5. QUESTIONS

5.1. Does education play a role in Visa certification?

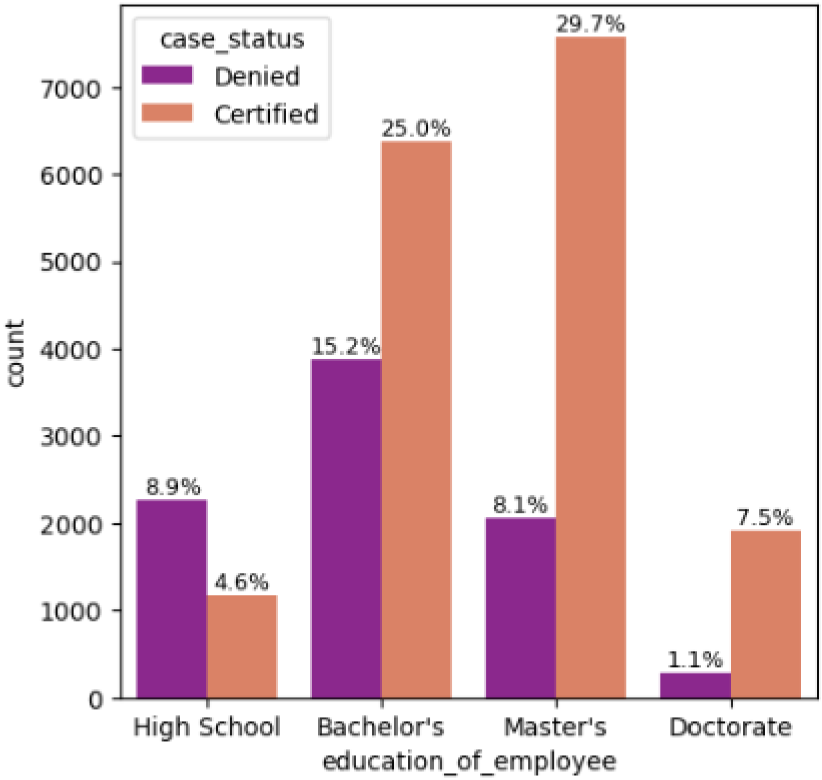


Figure 23 Education of Employees Vs Case Status

OBSERVATION:

- As the graph above shows, the ratio of applications being certified versus denied increases considerably as an applicant's highest level of education achieved increases.
- The ratio of an applicant with a high school diploma being approved versus denied is ~1:2, whereas the same ratio for an applicant with a doctorate is ~7:1.

5.2. How does the visa status vary across different continents?

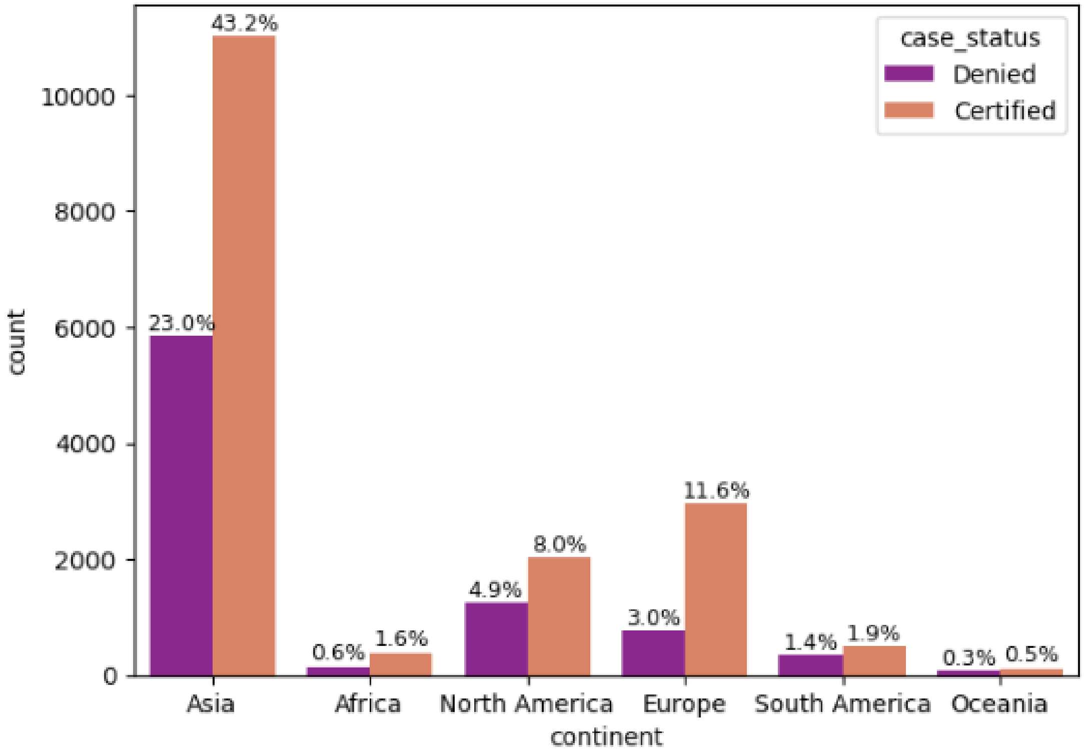


Figure 24 Continent Vs Case Status

OBSERVATION:

- Applicants from Asia comprise ~2/3 of all applications and these applicants have almost a 2:1 ratio of approvals to denials.
- An application from a European applicant has the best ratio of approvals to denials (~4:1).

5.3. In the United States, employees are paid at different intervals. Which pay unit is most likely to be certified for a visa?

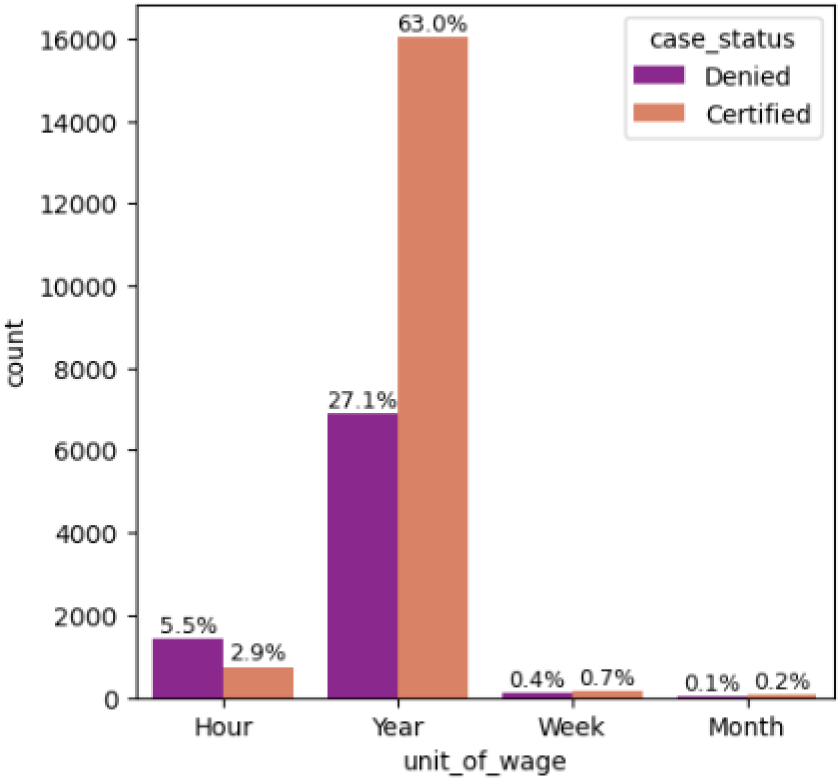


Figure 25 Unit of Wage Vs Case Status

OBSERVATION:

- Applicants who are applying to work in a job with an hourly rate have a ratio of approved versus denied applications of ~1:2. Additionally, these applicants comprise only ~8% of all applications, but comprise ~17% of all denials.
- Applicants from any other unit_of_wage category have a ratio of ~2:1, with applications for jobs with annual salaries showing nearly a 2.5:1 ratio of approvals to denials.

5.4. Does Region of employment influence visa status?

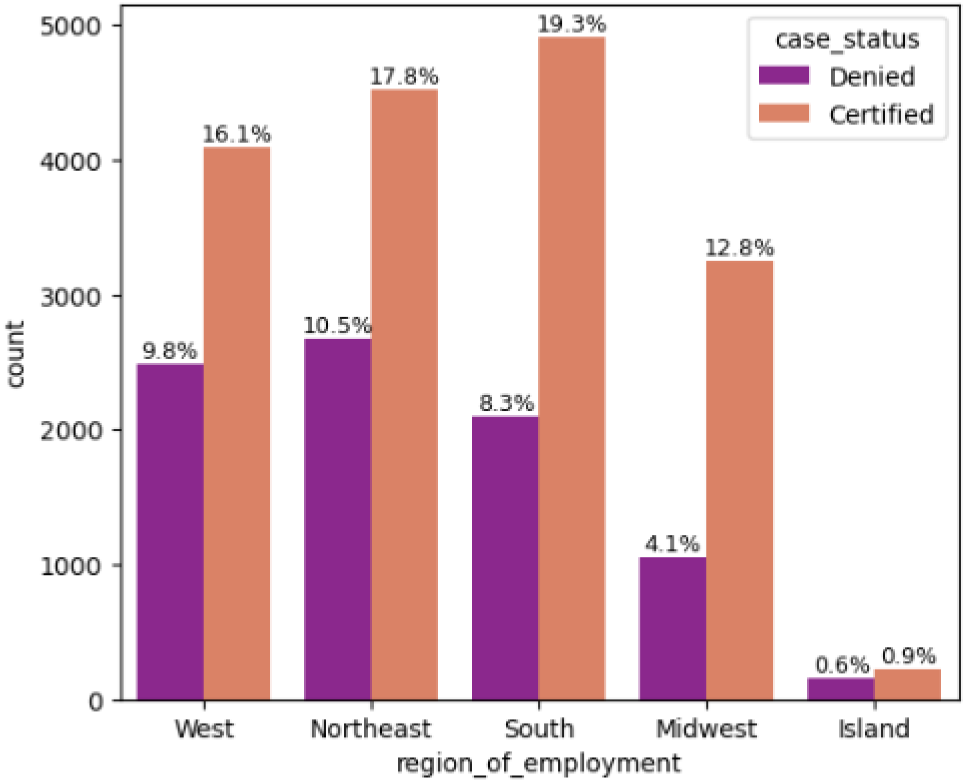


Figure 26 Region of Employment Vs Case Status

OBSERVATION:

- The majority of cases (16.1%) in the West region were certified, while 9.8% were denied.
- The Northeast also shows a higher proportion of certified cases (17.8%) compared to 10.5% being denied
- The South has the highest percentage of certified cases at 19.3%, with only 8.3% of cases denied.
- The Midwest has a lower total number of cases, with 12.8% certified and 4.1% denied.
- The Island region has the fewest cases overall, with a slight difference between certified (0.9%) and denied (0.6%).
- The South region has the highest certification rate, while the Midwest has the lowest certification percentage compared to other mainland regions.
- Islands account for the smallest share of employment cases, both certified and denied.

5.5. Experienced professionals might look abroad for opportunities to improve their lifestyles and career development. Does work experience influence visa status?

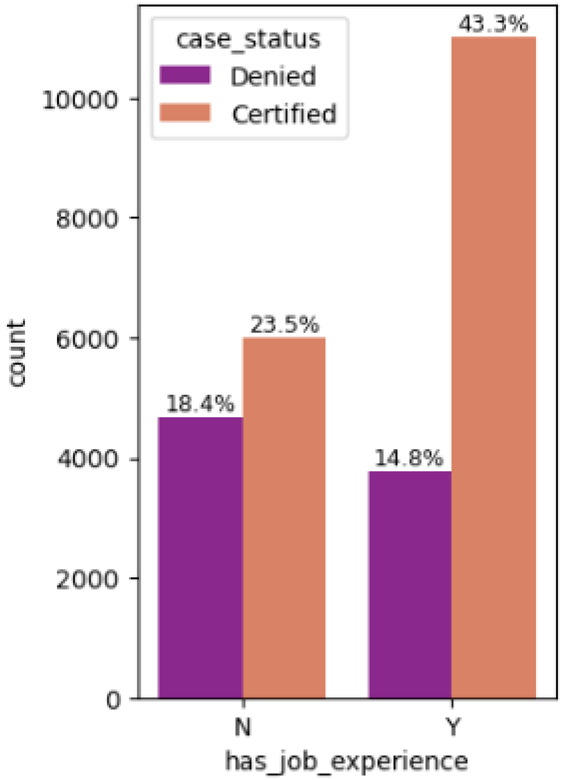


Figure 27 Has Job Experience Vs Case Status

OBSERVATION:

- Applicants with job experience have a ratio of approved to denied applications of ~3:1, whereas the same ratio for applicants without job experience have around a 5:4 ratio (i.e., approximately equivalent).

6. DATA PREPROCESSING

6.1. Treatment of Missing Values

Based on the initial evaluations, no values were missing in any of the columns. However, there were rows with unrealistic non-positive (<0) values of no_of_employees. To address this problem, these values are replaced with the median of no_of_employees.

- We will do missing value imputation after splitting the data into train, test and validation to **avoid data leakage**.

6.2. Feature Engineering

- The feature yr_of_estab is converted to yrs_snc_estab, containing the years since establishment.
- The columns yr_of_estab is dropped subsequently.

6.3. Detection and Treatment of Outliers

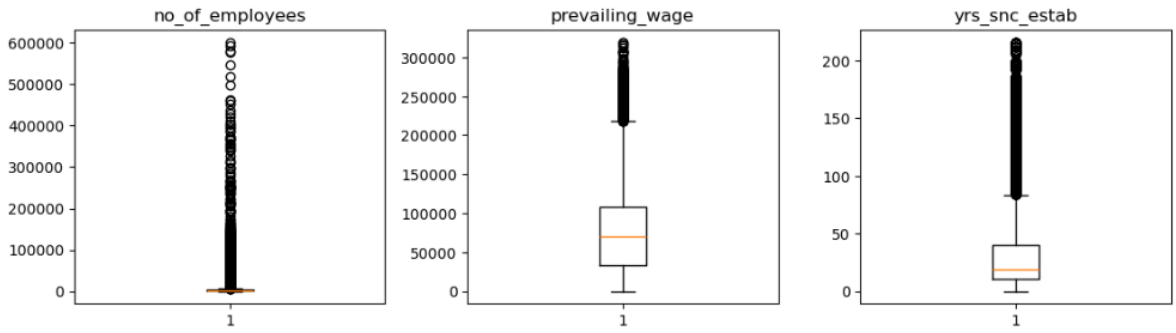


Figure 28 Detection of Outliers

Observation:

Since these values represent genuine outliers, we have chosen not to remove them from the analysis, as they reflect real-world booking behavior that could provide important insights for understanding customer patterns and trends

7. DATA PREPARATION FOR MODELING

- Encoding the values in the columns continent, has_job_experience, requires_job_training, full_time_position, region_of_employment, unit_of_wage and education_of_employee.
- Separation of independent and dependent variable.
- Splitting Data into Training, validation and Test Sets.
- Missing values are treated by imputing median.
- There is no duplicate values

- For the column case status, replaced the denied value with '*0*' and certified value with '1'.

- Created dummy variables for the columns mainly having datatype object and category.
- Let's inverse map the encoded values.

Reverse Mapping for Encoded Variables

- Inverse mapping returned original labels.

8. MODEL BUILDING

Model evaluation criterion

Model can make wrong predictions as:

1. Predicting an applicant visa should be certified but in reality, the applicant visa is denied.

2. Predicting an applicant visa should be denied but in reality, the applicant visa is certified.

Which case is more important?

- Both are important: If an applicant is approved when they would have been denied, an unqualified employee will get a job that should have been filled by a US citizen. If an applicant is denied when they should have been approved, U.S. companies will not be able to fill critical positions and the overall economy will not be as productive.

How to reduce this losses?

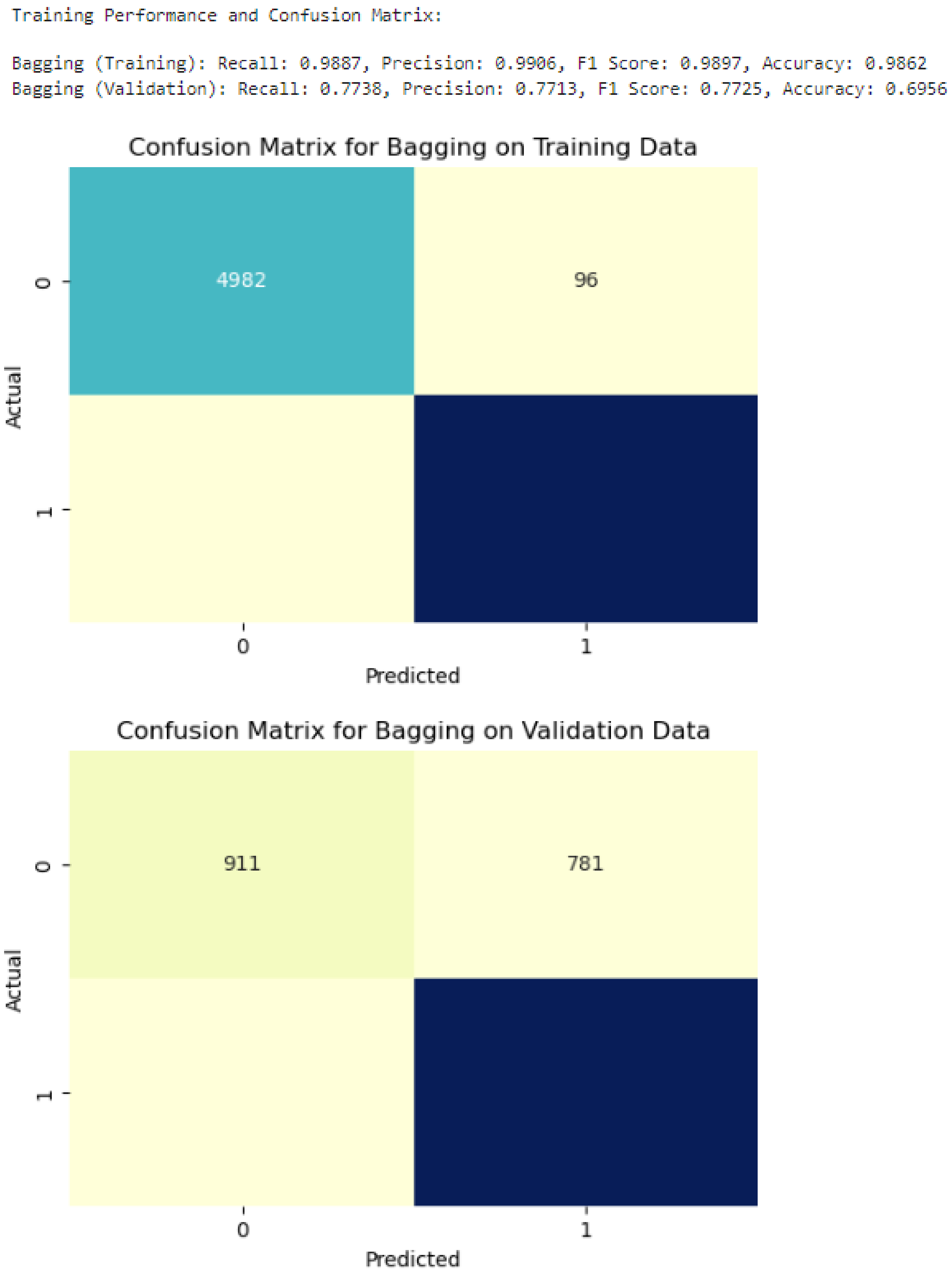
- As the process of reviewing each application is time and resource-intensive, this model should identify those candidates predicted to be approved, so agents can prioritize these applications.
- F1 Score can be used as the metric for evaluation of the model, as the greater the F1 score, the higher the chances of minimizing False Negatives and False Positives.
- We will use balanced class weights, where applicable, so that model focuses equally on both classes.

Let's define a function to output different metrics (including recall) on the train and test set and a function to show confusion matrix so that we do not have to use the same code repetitively while evaluating models.

8.1. MODEL BUILDING - ORIGINAL DATA

8.1.1 BAGGING CLASSIFIER

8.1.1.1 Training and Validation Performance



```
Training Performance and Confusion Matrix:

Bagging (Training): Recall: 0.9887, Precision: 0.9906, F1 Score: 0.9897, Accuracy: 0.9862
Bagging (Validation): Recall: 0.7738, Precision: 0.7713, F1 Score: 0.7725, Accuracy: 0.6956
```

Confusion Matrix for Bagging on Training Data

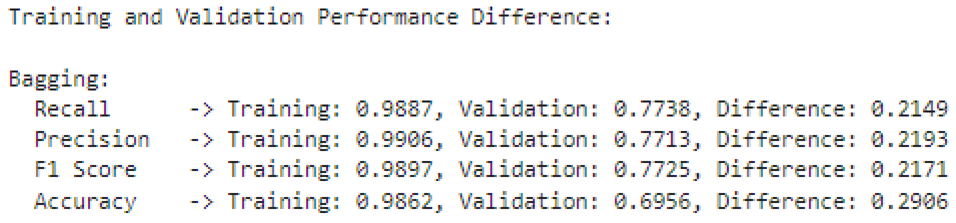
Confusion Matrix for Bagging on Validation Data

Figure 29 Training and Validation Performance of Bagging Classifier

Observations

- Shows strong performance on the training set with high recall (0.9887) and F1 score (0.9897), indicating it captures nearly all positive cases well.
- However, there's a notable drop in validation recall (0.7738) and accuracy (0.6956), suggesting some overfitting. This model may need regularization or tuning to improve generalization.

8.1.1.2 Training and Validation Performance Difference



```
Training and Validation Performance Difference:

Bagging:
  Recall      -> Training: 0.9887, Validation: 0.7738, Difference: 0.2149
  Precision   -> Training: 0.9906, Validation: 0.7713, Difference: 0.2193
  F1 Score    -> Training: 0.9897, Validation: 0.7725, Difference: 0.2171
  Accuracy    -> Training: 0.9862, Validation: 0.6956, Difference: 0.2906
```

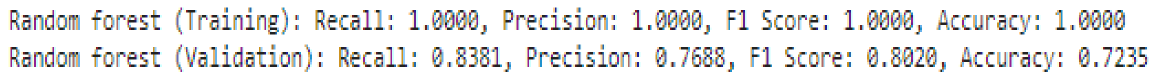
Table 4 Training and Validation Performance Difference of Bagging Classifier

Observation:

- There is a notable performance drop from training to validation, especially in **accuracy** (0.2906) and **precision** (0.2193), which implies some level of overfitting.
- The difference in **recall** (0.2149) and **F1 score** (0.2171) also indicates that while Bagging performs well in training, it struggles to generalize as effectively to new data.

8.1.2 RANDOM FOREST CLASSIFIER

8.1.2.1 Training and Validation Performance



```
Random forest (Training): Recall: 1.0000, Precision: 1.0000, F1 Score: 1.0000, Accuracy: 1.0000
Random forest (Validation): Recall: 0.8381, Precision: 0.7688, F1 Score: 0.8020, Accuracy: 0.7235
```

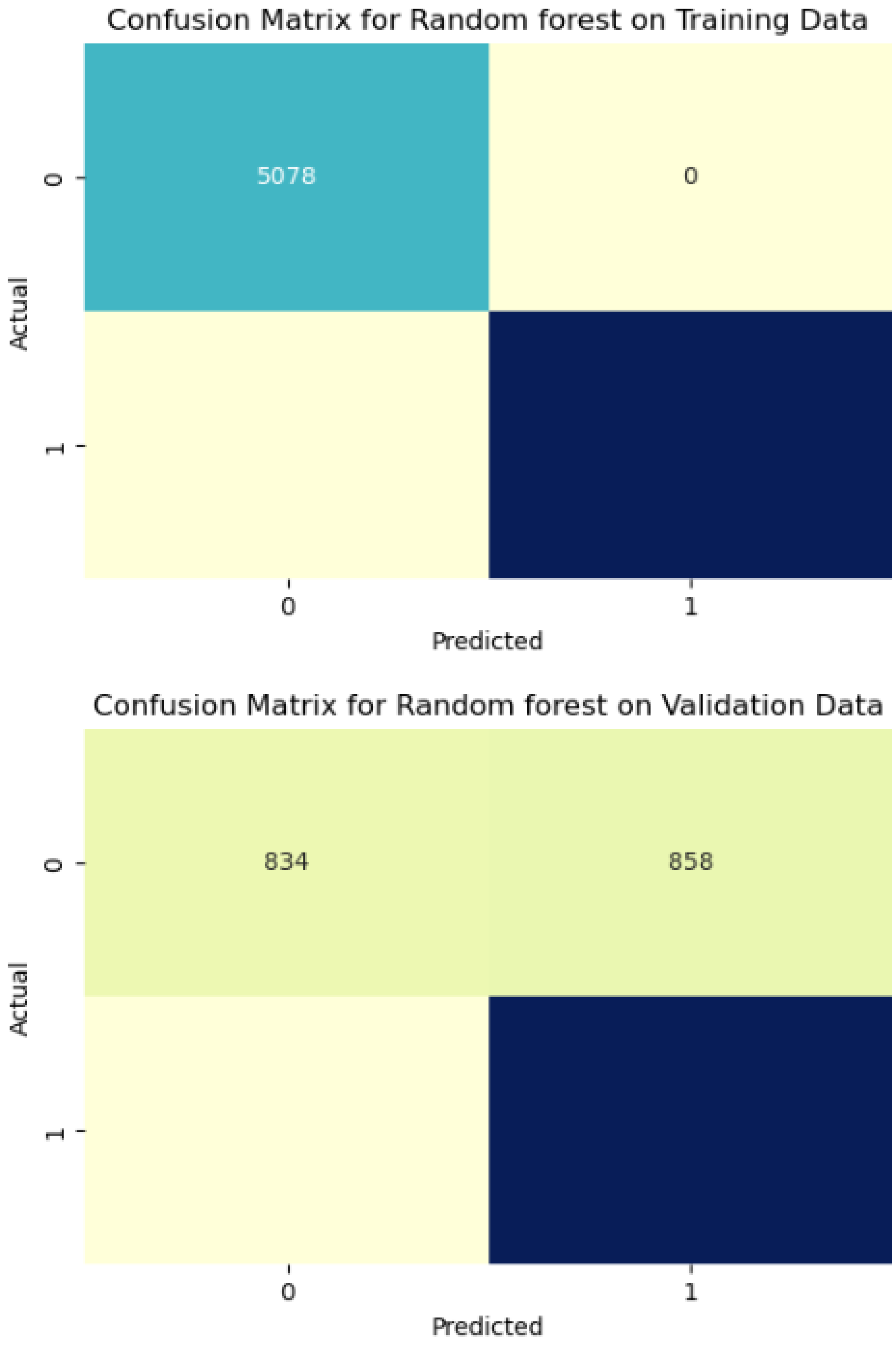
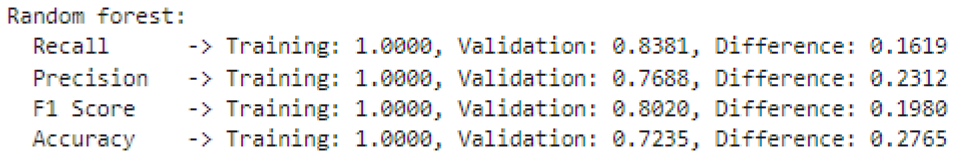


Figure 30 Training and Validation Performance of Random Forest Classifier

Random Forest:

- Achieves perfect scores on the training set (all metrics = 1.0000), indicating overfitting, as the model memorizes the training data.
- The validation performance declines, with recall (0.8381) and F1 score (0.8020) still relatively high, but not as strong as the training scores. This is a clear indication of overfitting, though it still performs reasonably well on unseen data.

8.1.2.2 Training and Validation Performance Difference



```
Random forest:
  Recall      -> Training: 1.0000, Validation: 0.8381, Difference: 0.1619
  Precision   -> Training: 1.0000, Validation: 0.7688, Difference: 0.2312
  F1 Score    -> Training: 1.0000, Validation: 0.8020, Difference: 0.1980
  Accuracy    -> Training: 1.0000, Validation: 0.7235, Difference: 0.2765
```

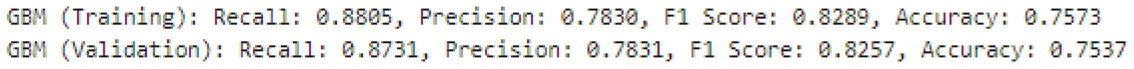
Table 5 Training and Validation Performance Difference of Random Forest Classifier

OBSERVATION:

- Like Bagging, Random Forest also shows substantial overfitting, with **recall** dropping by 0.1619 and **accuracy** by 0.2765.
- A particularly large drop in precision (0.2312) and **F1 score** (0.1980) further suggests that the model captures training data well but does not generalize as effectively to unseen cases.

8.1.3 GRADIENT BOOSTING CLASSIFIER

8.1.3.1 Training and Validation Performance



```
GBM (Training): Recall: 0.8805, Precision: 0.7830, F1 Score: 0.8289, Accuracy: 0.7573
GBM (Validation): Recall: 0.8731, Precision: 0.7831, F1 Score: 0.8257, Accuracy: 0.7537
```

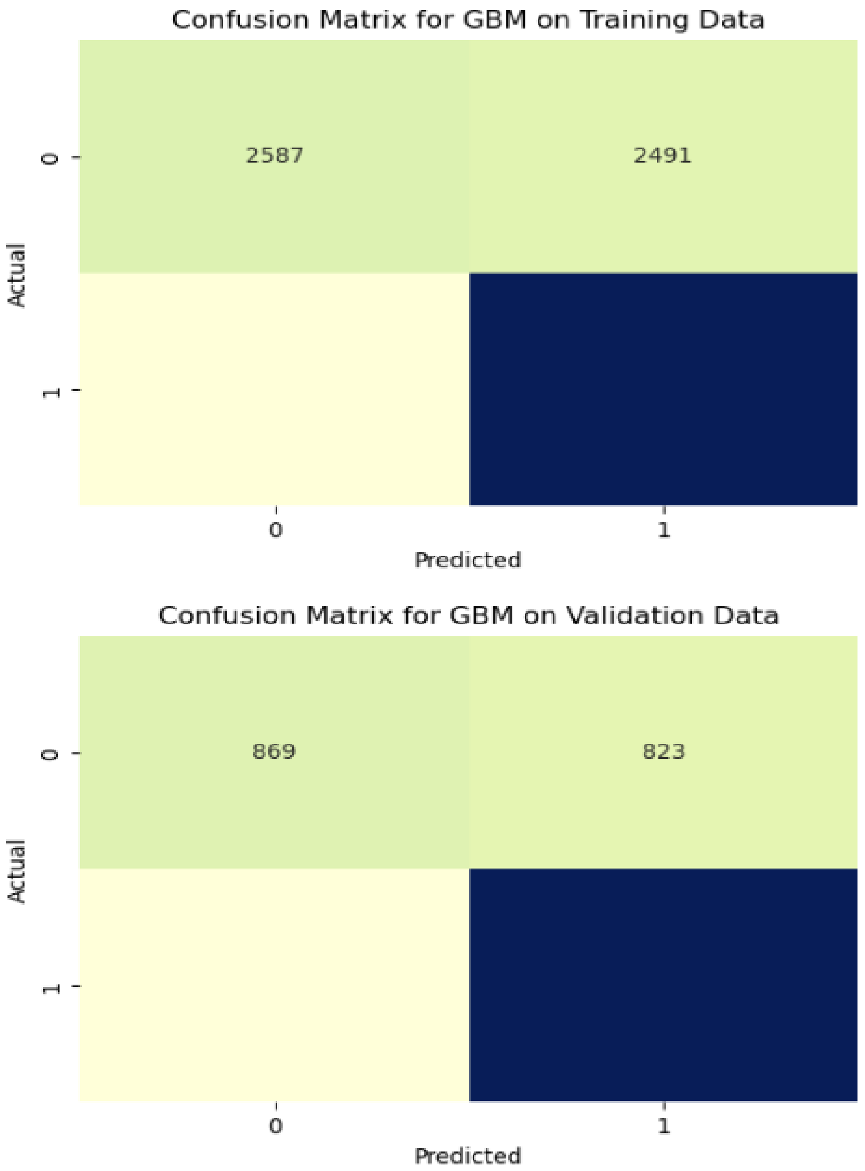
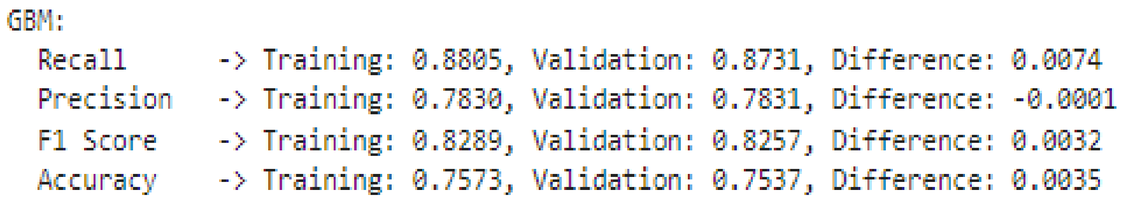


Figure 31 Training and Validation Performance of GBM Classifier

OBSERVATION:

- Exhibits balanced performance with minimal difference between training (Recall: 0.8805) and validation recall (0.8731). The F1 scores are also close, showing that this model generalizes better than Bagging and Random Forest.
- Given the lower training performance compared to other models, it might benefit from additional tuning but demonstrates a strong balance, indicating robustness.

8.1.3.2 Training and Validation Performance Difference



```
GBM:
  Recall      -> Training: 0.8805, Validation: 0.8731, Difference: 0.0074
  Precision   -> Training: 0.7830, Validation: 0.7831, Difference: -0.0001
  F1 Score    -> Training: 0.8289, Validation: 0.8257, Difference: 0.0032
  Accuracy    -> Training: 0.7573, Validation: 0.7537, Difference: 0.0035
```

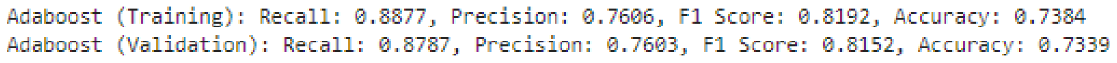
Table 6 Training and Validation Performance Difference of GBM Classifier

OBSERVATION:

- GBM demonstrates minimal performance difference across all metrics, indicating strong generalization.
- The **recall** difference is only 0.0074, and **F1 score** has a difference of just 0.0032, suggesting it is well-tuned and may require little adjustment.
- This consistency implies GBM can perform reliably on new data, making it a strong candidate for practical use.

8.1.4 ADA BOOST CLASSIFIER

8.1.4.1 Training and Validation Performance



```
Adaboost (Training): Recall: 0.8877, Precision: 0.7606, F1 Score: 0.8192, Accuracy: 0.7384
Adaboost (Validation): Recall: 0.8787, Precision: 0.7603, F1 Score: 0.8152, Accuracy: 0.7339
```

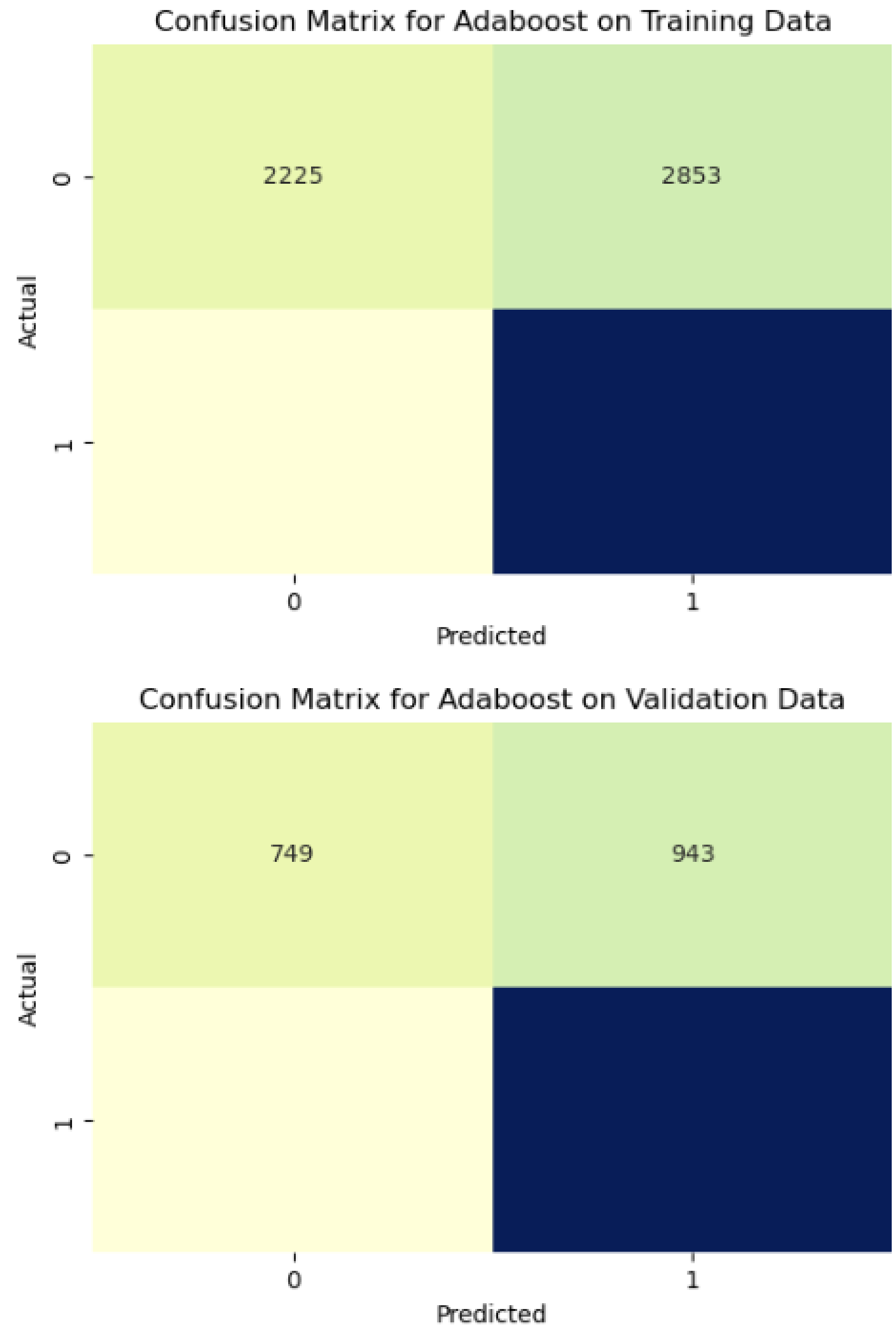
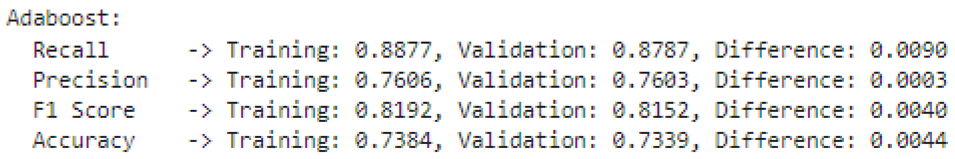


Figure 32 Training and Validation Performance of Adaboost Classifier

OBSERVATION:

- Similar to GBM, AdaBoost shows balanced performance between training (Recall: 0.8877, F1: 0.8192) and validation (Recall: 0.8787, F1: 0.8152) scores, suggesting good generalization.
- The close alignment of metrics across both sets indicates AdaBoost captures the data patterns well without severe overfitting.

8.1.4.2 Training and Validation Performance Difference



```
Adaboost:
  Recall      -> Training: 0.8877, Validation: 0.8787, Difference: 0.0090
  Precision   -> Training: 0.7606, Validation: 0.7603, Difference: 0.0003
  F1 Score    -> Training: 0.8192, Validation: 0.8152, Difference: 0.0040
  Accuracy    -> Training: 0.7384, Validation: 0.7339, Difference: 0.0044
```

Table 7 Training and Validation Performance Difference of Adaboost Classifier

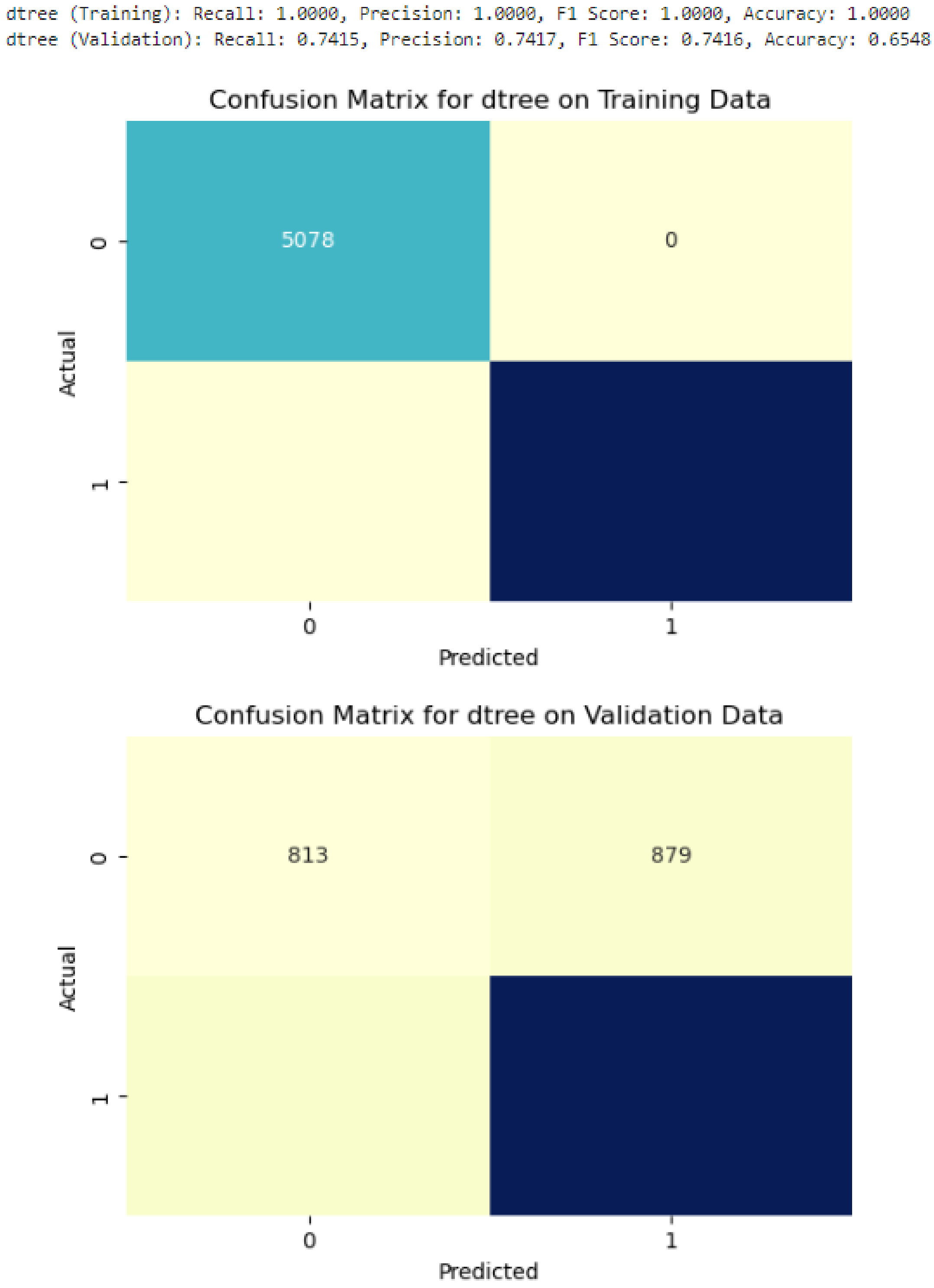
OBSERVATION:

- AdaBoost also shows excellent generalization, with very low metric differences across training and validation.
- The **recall** difference is only 0.0090, and **accuracy** has a small difference of 0.0044, similar to GBM.

This model's consistency suggests it is effective at handling variance in the dataset and adapting well to unseen data

8.1.5 DECISION TREE CLASSIFIER

8.1.5.1 Training and Validation Performance



```
dtree (Training): Recall: 1.0000, Precision: 1.0000, F1 Score: 1.0000, Accuracy: 1.0000
dtree (Validation): Recall: 0.7415, Precision: 0.7417, F1 Score: 0.7416, Accuracy: 0.6548
```

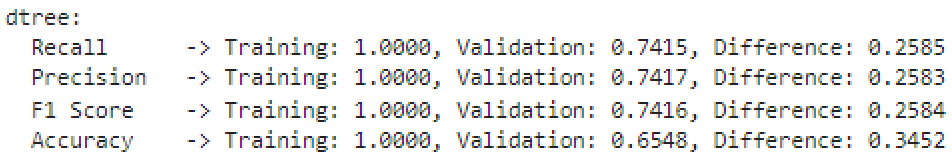
Figure 33 Training and Validation Performance of Decision Tree Classifier

OBSERVATION:

- Achieves perfect scores on the training set, indicating severe overfitting, as it likely memorizes the data.

- Performance on the validation set (Recall: 0.7415, F1: 0.7416) is much lower, which is indicative of its weak generalization. This model would benefit from pruning or additional tuning to improve validation performance.

8.1.5.2 Training and Validation Performance Difference



```
dtree:
  Recall      -> Training: 1.0000, Validation: 0.7415, Difference: 0.2585
  Precision   -> Training: 1.0000, Validation: 0.7417, Difference: 0.2583
  F1 Score    -> Training: 1.0000, Validation: 0.7416, Difference: 0.2584
  Accuracy    -> Training: 1.0000, Validation: 0.6548, Difference: 0.3452
```

Table 8 Training and Validation Performance Difference of Decision Tree Classifier

OBSERVATION:

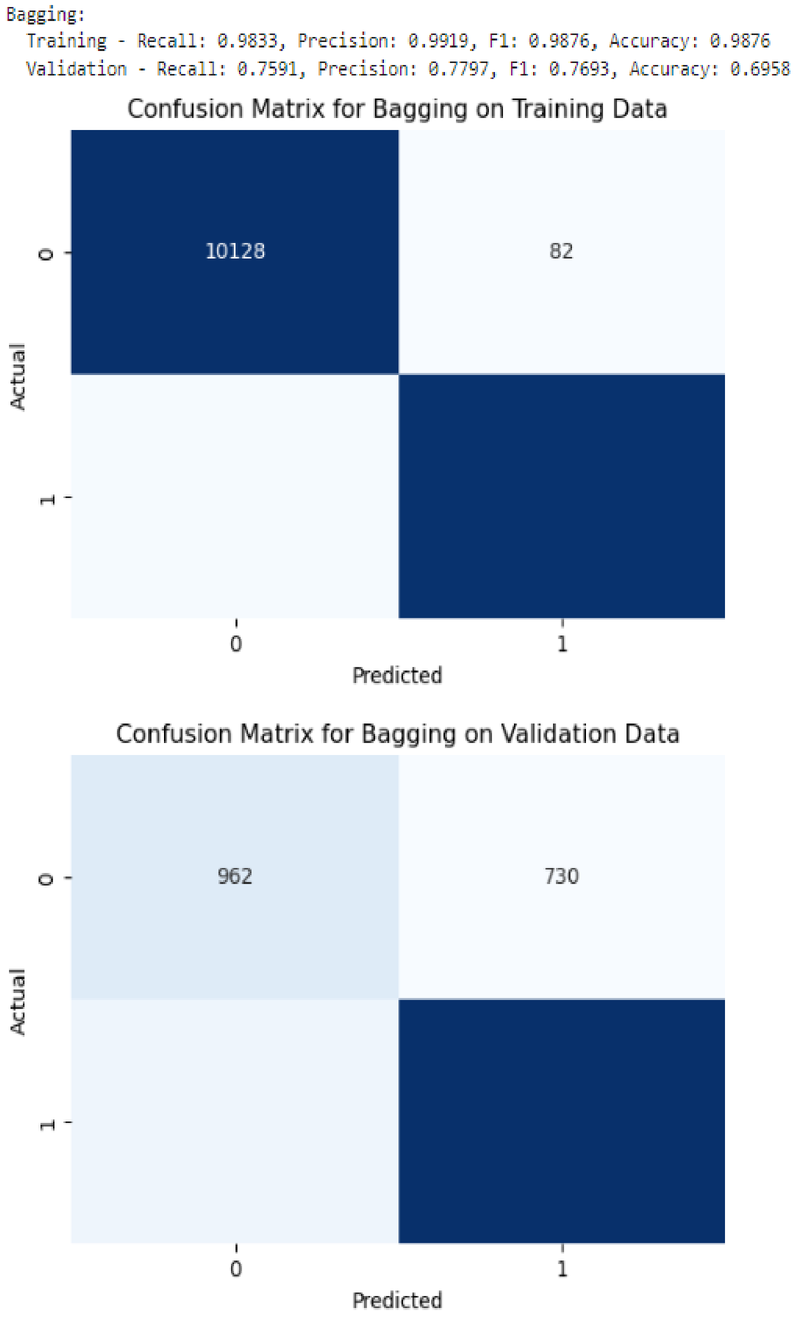
- Decision Tree exhibits the highest level of overfitting among all models, with perfect training metrics and considerable drops in validation performance.
- The **accuracy** difference (0.3452) and **recall** difference (0.2585) indicate a heavy reliance on memorization rather than generalization.
- This model would need significant tuning (such as pruning) to improve generalization and reduce its sensitivity to training data.

GBM and AdaBoost are the best-performing models in terms of generalization, with minimal differences between training and validation metrics, making them ideal candidates for deployment.

8.2. MODEL BUILDING - OVERSAMPLED DATA

8.2.1 BAGGING CLASSIFIER

8.2.1.1 Training and Validation Performance



```
Bagging:
  Training - Recall: 0.9833, Precision: 0.9919, F1: 0.9876, Accuracy: 0.9876
  Validation - Recall: 0.7591, Precision: 0.7797, F1: 0.7693, Accuracy: 0.6958
```

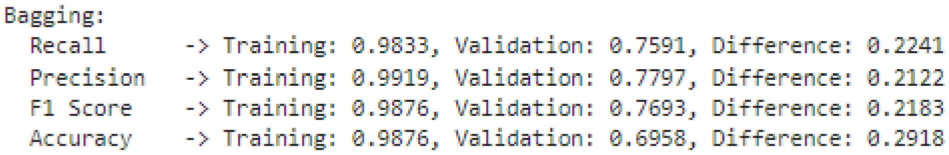
Figure 34 Training and Validation Performance of Bagging Classifier

Observations

- **Training** metrics are high, especially in **recall** (0.9833) and **F1 score** (0.9876), but there is a significant drop in validation performance, with **recall** decreasing to 0.7591 and **F1 score** to 0.7693.

- The **accuracy** difference of 0.2918 between training and validation indicates some overfitting, suggesting Bagging performs better on training data but loses generalizability on new data.

8.2.1.2 Training and Validation Performance Difference



```
Bagging:
  Recall      -> Training: 0.9833, Validation: 0.7591, Difference: 0.2241
  Precision   -> Training: 0.9919, Validation: 0.7797, Difference: 0.2122
  F1 Score    -> Training: 0.9876, Validation: 0.7693, Difference: 0.2183
  Accuracy    -> Training: 0.9876, Validation: 0.6958, Difference: 0.2918
```

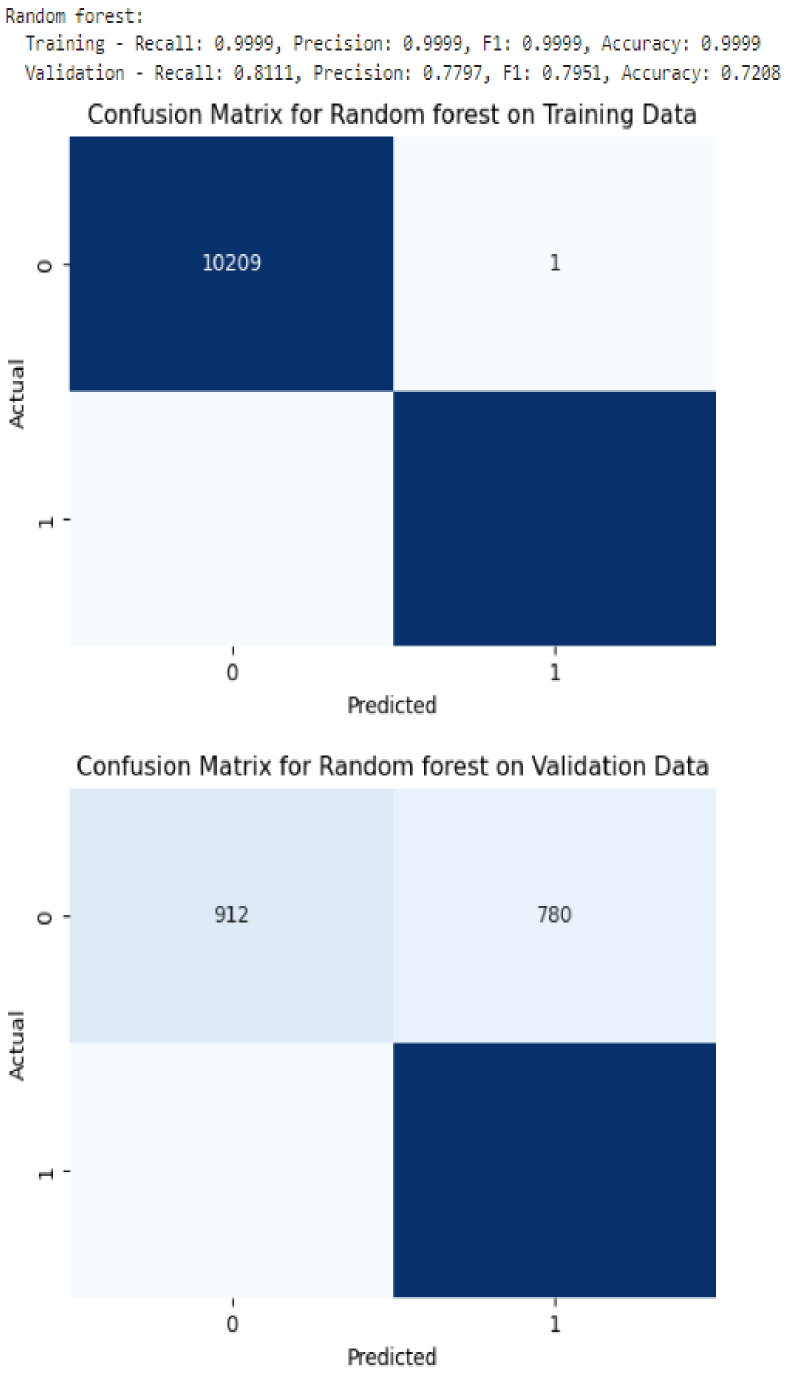
Table 9 Training and Validation Performance Difference of Bagging Classifier

Observation

- Shows a considerable difference in metrics, especially in **recall** (0.2241) and **accuracy** (0.2918).
- This high discrepancy suggests **overfitting** to the training data, with the model struggling to generalize on unseen data.

8.2.2 RANDOM FOREST CLASSIFIER

8.2.2.1 Training and Validation Performance



```
Random forest:
  Training - Recall: 0.9999, Precision: 0.9999, F1: 0.9999, Accuracy: 0.9999
  Validation - Recall: 0.8111, Precision: 0.7797, F1: 0.7951, Accuracy: 0.7208
```

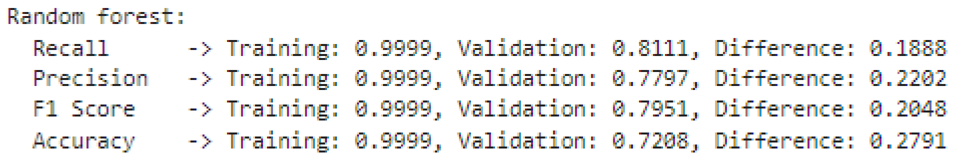
Confusion Matrix for Random forest on Training Data
10209
1
0
1
Actual
0
1
Predicted
Confusion Matrix for Random forest on Validation Data
912
780
0
1
Actual
0
1
Predicted

Figure 35 Training and Validation Performance of Random Forest Classifier

OBSERVATION:

- **Training** performance is nearly perfect (Recall, Precision, F1, Accuracy ~1.0), pointing to potential overfitting.
- Validation metrics are notably lower, with **recall** at 0.8111 and **F1 score** at 0.7951, and an **accuracy** of 0.7208. The model may struggle to generalize effectively without further tuning or regularization to prevent overfitting.

8.2.2.2 Training and Validation Performance Difference



```
Random forest:
  Recall      -> Training: 0.9999, Validation: 0.8111, Difference: 0.1888
  Precision   -> Training: 0.9999, Validation: 0.7797, Difference: 0.2202
  F1 Score    -> Training: 0.9999, Validation: 0.7951, Difference: 0.2048
  Accuracy    -> Training: 0.9999, Validation: 0.7208, Difference: 0.2791
```

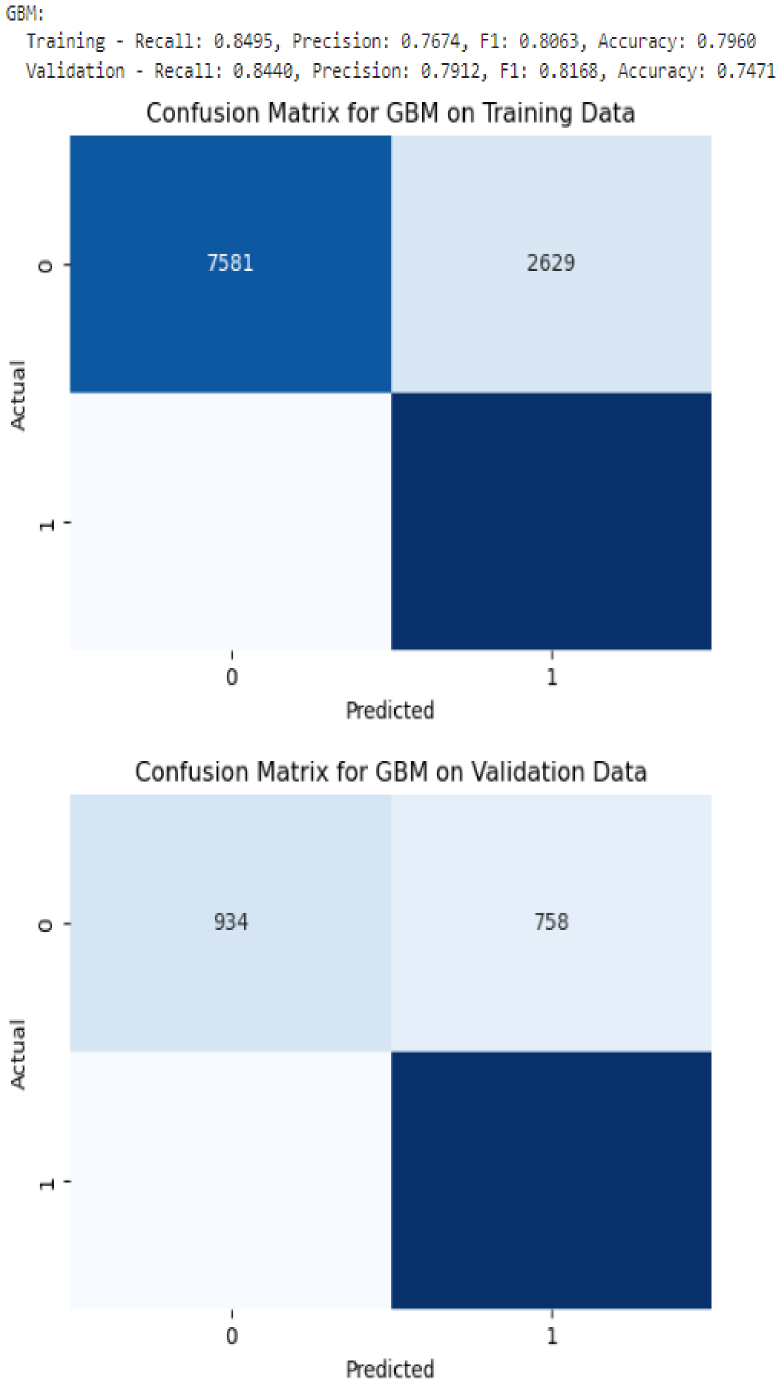
Table 10 Training and Validation Performance Difference of Random Forest Classifier

Observation

- Despite near-perfect training performance, validation metrics decline notably, with a **recall** difference of 0.1888 and **accuracy** difference of 0.2791.
- Indicates overfitting, though slightly less severe than Bagging. The model would benefit from **hyperparameter tuning or regularization** to improve generalizability.

8.2.3 GRADIENT BOOSTING CLASSIFIER

8.2.3.1 Training and Validation Performance



```
GBM:
  Training - Recall: 0.8495, Precision: 0.7674, F1: 0.8063, Accuracy: 0.7960
  Validation - Recall: 0.8440, Precision: 0.7912, F1: 0.8168, Accuracy: 0.7471
```

Confusion Matrix for GBM on Training Data
7581
2629
Actual
0
1
Predicted
Confusion Matrix for GBM on Validation Data
934
758
Actual
0
1
Predicted

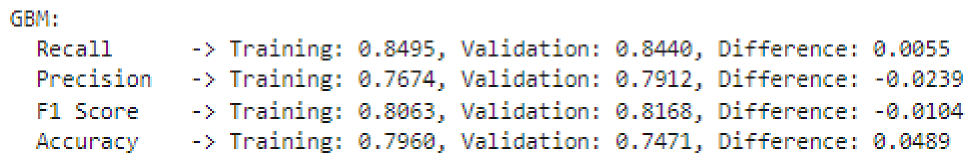
Figure 36 Training and Validation Performance of GBM Classifier

OBSERVATION:

- GBM shows balanced performance between training and validation datasets, with minor drops across all metrics.
- **Recall** difference between training (0.8495) and validation (0.8440) is only 0.0055, and **F1 score** is similarly stable (training: 0.8063, validation: 0.8168).

- This stability suggests strong generalization capabilities, making GBM a good choice for reliable performance on new data.

8.2.3.2 Training and Validation Performance Difference



```
GBM:
  Recall      -> Training: 0.8495, Validation: 0.8440, Difference: 0.0055
  Precision   -> Training: 0.7674, Validation: 0.7912, Difference: -0.0239
  F1 Score    -> Training: 0.8063, Validation: 0.8168, Difference: -0.0104
  Accuracy    -> Training: 0.7960, Validation: 0.7471, Difference: 0.0489
```

Table 11 Training and Validation Performance Difference of GBM Classifier

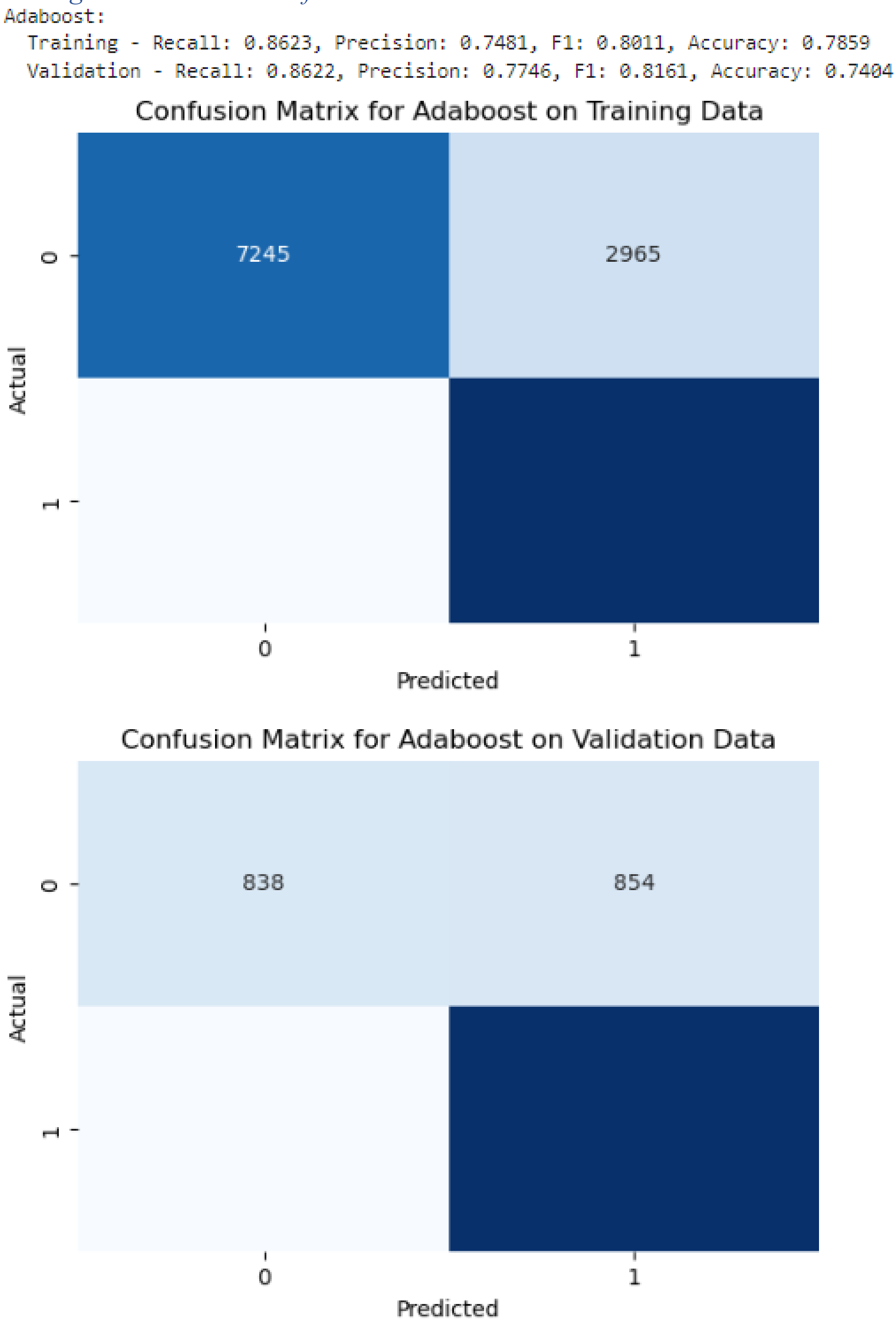
Observation

- Very small performance differences between training and validation, particularly in **recall** (0.0055) and **F1 score** (-0.0104).

This balance highlights **strong generalization**, with GBM adapting well to new data without significant overfitting.

8.2.4 ADA BOOST CLASSIFIER

8.2.4.1 Training and Validation Performance



```
Adaboost:
  Training - Recall: 0.8623, Precision: 0.7481, F1: 0.8011, Accuracy: 0.7859
  Validation - Recall: 0.8622, Precision: 0.7746, F1: 0.8161, Accuracy: 0.7404
```

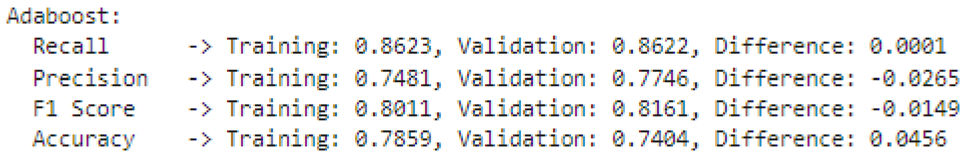
Figure 37 Training and Validation Performance of Adaboost Classifier

OBSERVATION:

- Similar to GBM, AdaBoost demonstrates consistent performance across both datasets, with minimal difference between training and validation.

- **Recall** and **F1 score** in training and validation are nearly identical (training recall: 0.8623 vs. validation recall: 0.8622).
- These results suggest AdaBoost has robust generalization, making it another strong candidate for deployment due to reliable performance on unseen data.

8.2.4.2 Training and Validation Performance Difference



```
Adaboost:
  Recall      -> Training: 0.8623, Validation: 0.8622, Difference: 0.0001
  Precision   -> Training: 0.7481, Validation: 0.7746, Difference: -0.0265
  F1 Score    -> Training: 0.8011, Validation: 0.8161, Difference: -0.0149
  Accuracy    -> Training: 0.7859, Validation: 0.7404, Difference: 0.0456
```

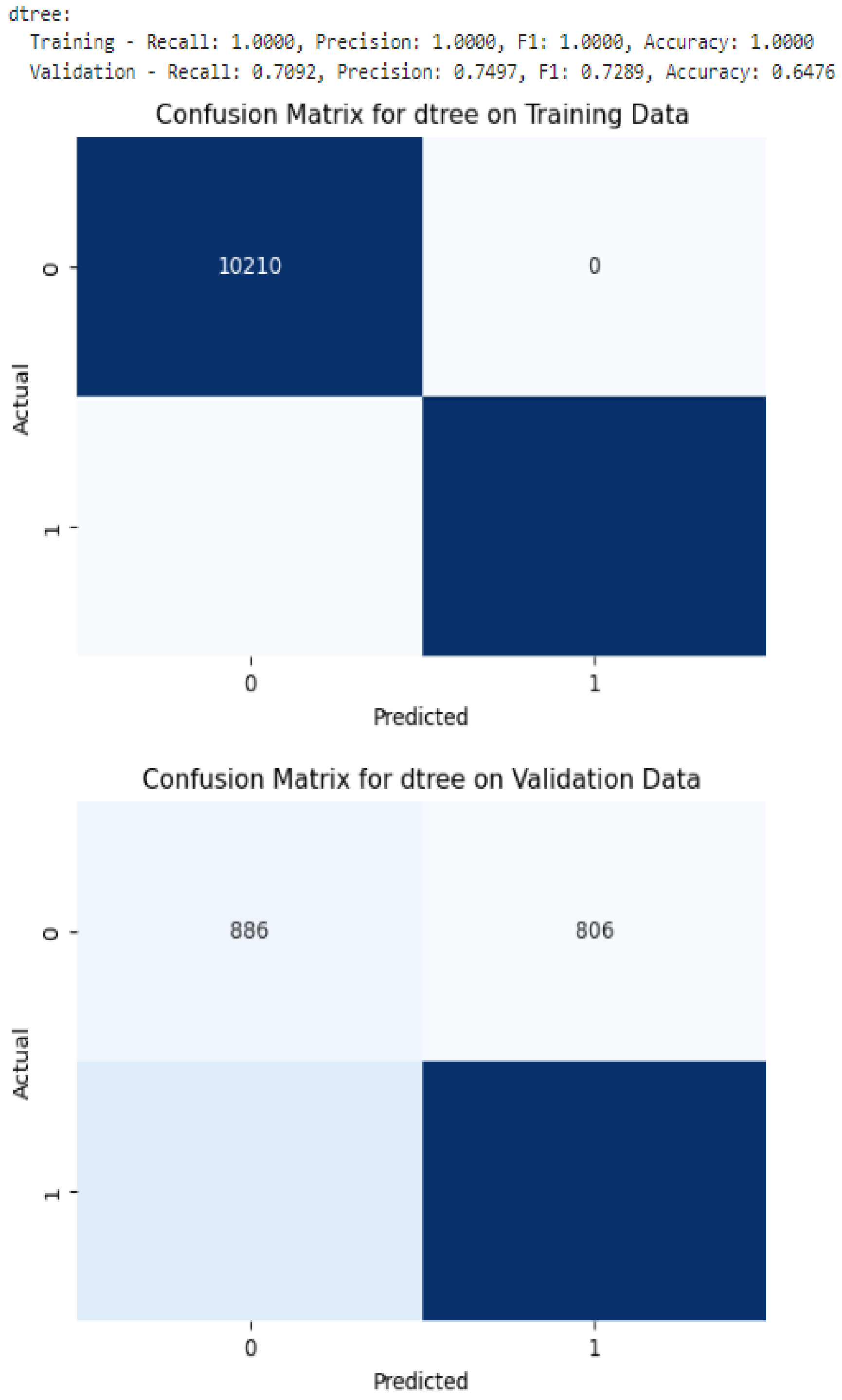
Table 12 Training and Validation Performance Difference of Adaboost Classifier

Observation

- Consistently low difference across metrics, with **recall** (0.0001) and **F1 score** (-0.0149) showing minimal deviation.
- Similar to GBM, AdaBoost demonstrates **excellent generalization** capabilities, suggesting it's well-suited for deployment on unseen data.

8.2.5 DECISION TREE CLASSIFIER

8.2.5.1 Training and Validation Performance



```
dtree:
  Training - Recall: 1.0000, Precision: 1.0000, F1: 1.0000, Accuracy: 1.0000
  Validation - Recall: 0.7092, Precision: 0.7497, F1: 0.7289, Accuracy: 0.6476
```

Confusion Matrix for dtree on Training Data
10210
0
Actual
0
1
0
1
Predicted
Confusion Matrix for dtree on Validation Data
886
806
Actual
0
1
0
1
Predicted

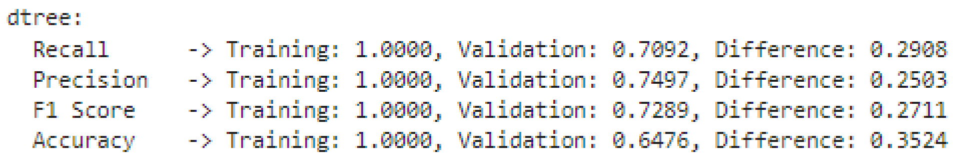
Figure 38 Training and Validation Performance of Decision Tree Classifier

OBSERVATION:

- Decision Tree performs perfectly on the training set (Recall, Precision, F1, Accuracy = 1.0), a classic indicator of overfitting.

- Validation metrics show a considerable drop, with **recall** at 0.7092, **F1 score** at 0.7289, and **accuracy** at 0.6476.
- This significant difference highlights that the model memorizes the training data rather than generalizing well, indicating a need for tuning or regularization.

8.2.5.2 Training and Validation Performance Difference



```
dtree:
  Recall      -> Training: 1.0000, Validation: 0.7092, Difference: 0.2908
  Precision   -> Training: 1.0000, Validation: 0.7497, Difference: 0.2503
  F1 Score    -> Training: 1.0000, Validation: 0.7289, Difference: 0.2711
  Accuracy    -> Training: 1.0000, Validation: 0.6476, Difference: 0.3524
```

Table 13 Training and Validation Performance Difference of Decision Tree Classifier

Observation

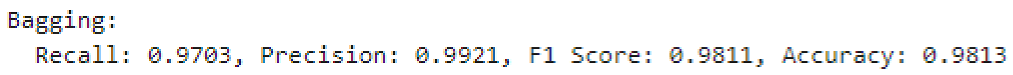
- Exhibits extreme overfitting with substantial differences in **recall** (0.2908) and **accuracy** (0.3524) between training and validation.
- Indicates that Decision Tree captures too much specific detail from training data, performing poorly on new samples. **Pruning or limiting tree depth** could help address overfitting.

8.3. MODEL BUILDING - UNDERSAMPLED DATA

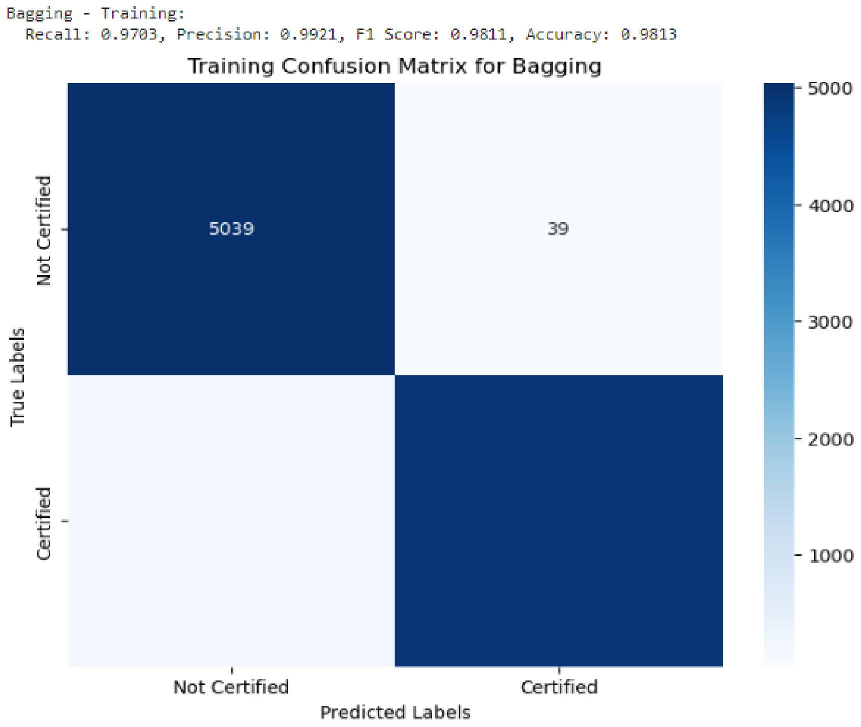
8.3.1 BAGGING CLASSIFIER

8.3.1.1 Training and Validation Performance

Training:



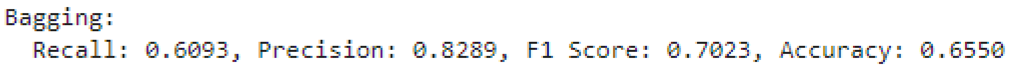
```
Bagging:
  Recall: 0.9703, Precision: 0.9921, F1 Score: 0.9811, Accuracy: 0.9813
```



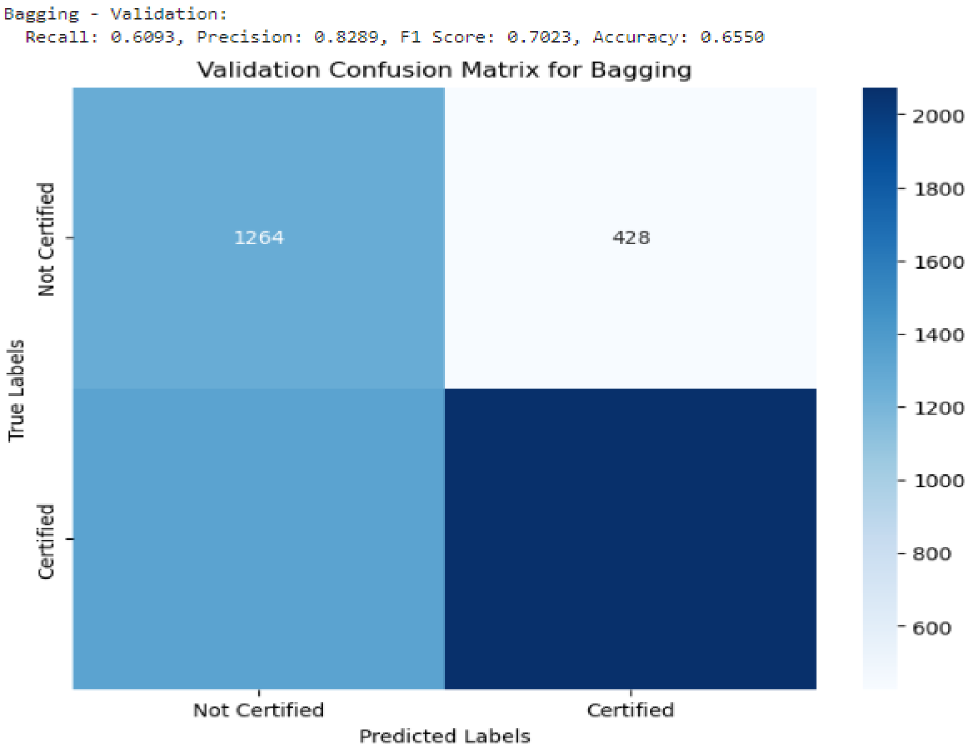
```
Bagging - Training:
  Recall: 0.9703, Precision: 0.9921, F1 Score: 0.9811, Accuracy: 0.9813
```

Figure 39 Training Performance of Bagging Classifier

Validation:



```
Bagging:
  Recall: 0.6093, Precision: 0.8289, F1 Score: 0.7023, Accuracy: 0.6550
```



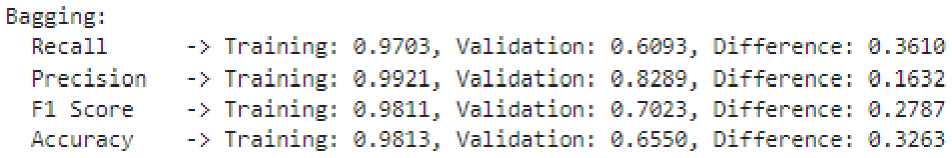
```
Bagging - Validation:
  Recall: 0.6093, Precision: 0.8289, F1 Score: 0.7023, Accuracy: 0.6550
```

Figure 40 Validation Performance of Bagging Classifier

Observations:

- **Training** metrics are high, especially recall (0.9703) and precision (0.9921), indicating the model fits the training data well.
- **Validation** metrics, however, show a significant drop, particularly in **recall** (0.6093) and **accuracy** (0.6550).
- This decline suggests **overfitting**, as the model generalizes poorly to new data.

8.3.1.2 Training and Validation Performance Difference



```
Bagging:
  Recall      -> Training: 0.9703, Validation: 0.6093, Difference: 0.3610
  Precision   -> Training: 0.9921, Validation: 0.8289, Difference: 0.1632
  F1 Score    -> Training: 0.9811, Validation: 0.7023, Difference: 0.2787
  Accuracy    -> Training: 0.9813, Validation: 0.6550, Difference: 0.3263
```

Table 14 Training and Validation Performance Difference of Bagging Classifier

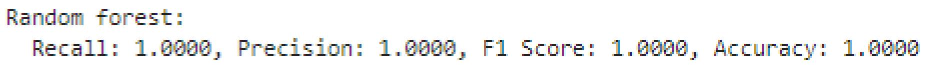
Observations:

- Large difference across metrics, particularly **recall** (0.3610) and **accuracy** (0.3263), indicates that while it performs well on training data, it generalizes poorly on validation data.
- The high discrepancy suggests **significant overfitting**, with a particularly noticeable drop in recall and F1 score on validation.

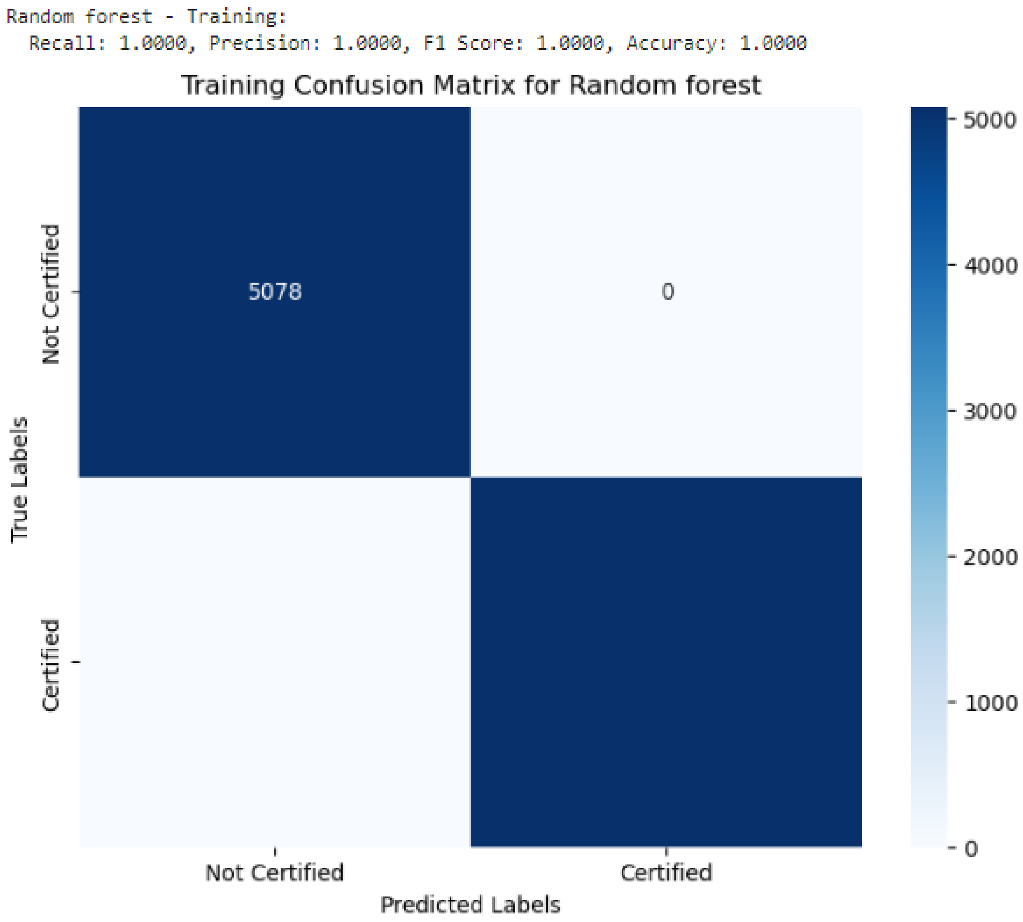
8.3.2 RANDOM FOREST CLASSIFIER

8.3.2.1 Training and Validation Performance

Training:



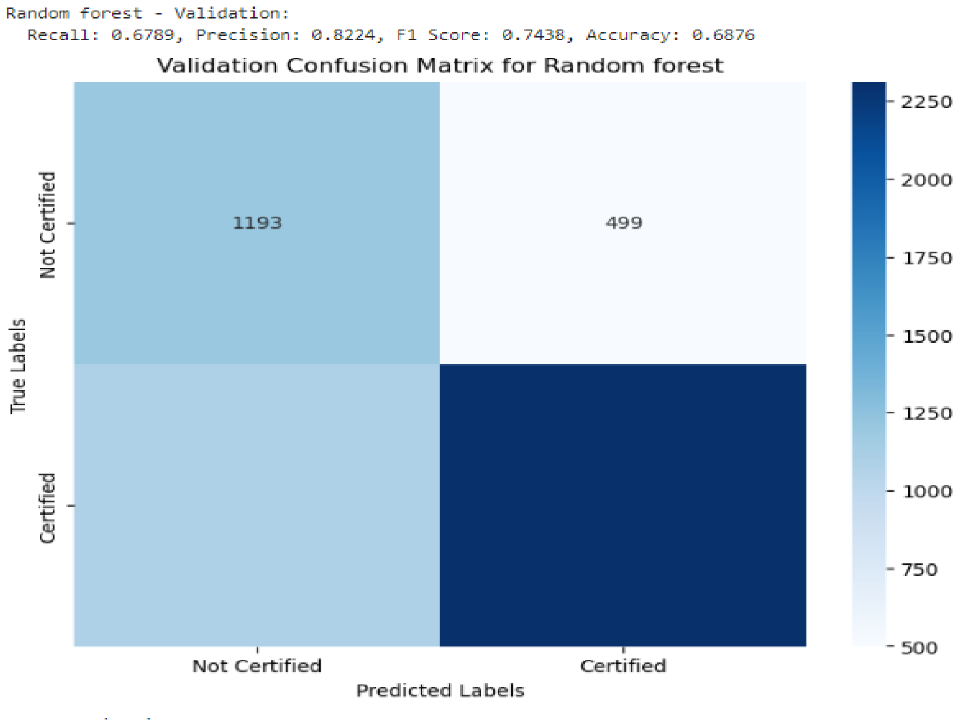
```
Random forest:
  Recall: 1.0000, Precision: 1.0000, F1 Score: 1.0000, Accuracy: 1.0000
```



```
Random forest - Training:
  Recall: 1.0000, Precision: 1.0000, F1 Score: 1.0000, Accuracy: 1.0000
```

Figure 41 Training Performance of Random Forest Classifier

Validation:



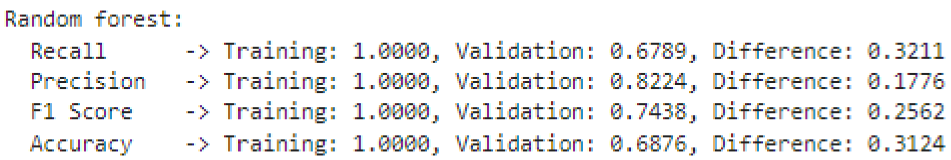
```
Random forest - Validation:
  Recall: 0.6789, Precision: 0.8224, F1 Score: 0.7438, Accuracy: 0.6876
```

Figure 42 Validation Performance of Random Forest Classifier

Observations:

- **Perfect training performance** across all metrics (Recall, Precision, F1 Score, Accuracy all at 1.0000), which is a strong indicator of overfitting.
- Validation scores drop considerably, with recall at 0.6789 and F1 at 0.7438.
- Though validation metrics are slightly higher than Bagging's, the large discrepancy points to overfitting.

8.3.2.2 Training and Validation Performance Difference



```
Random forest:
  Recall      -> Training: 1.0000, Validation: 0.6789, Difference: 0.3211
  Precision   -> Training: 1.0000, Validation: 0.8224, Difference: 0.1776
  F1 Score    -> Training: 1.0000, Validation: 0.7438, Difference: 0.2562
  Accuracy    -> Training: 1.0000, Validation: 0.6876, Difference: 0.3124
```

Table 15 Training and Validation Performance Difference of Random Forest Classifier

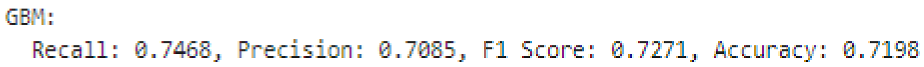
Observations:

- Similar to Bagging, there is a considerable **performance drop** on validation, especially in recall (0.3211) and accuracy (0.3124).
- Though precision's difference is smaller (0.1776), the high F1 score difference (0.2562) also indicates **overfitting** to the training set.

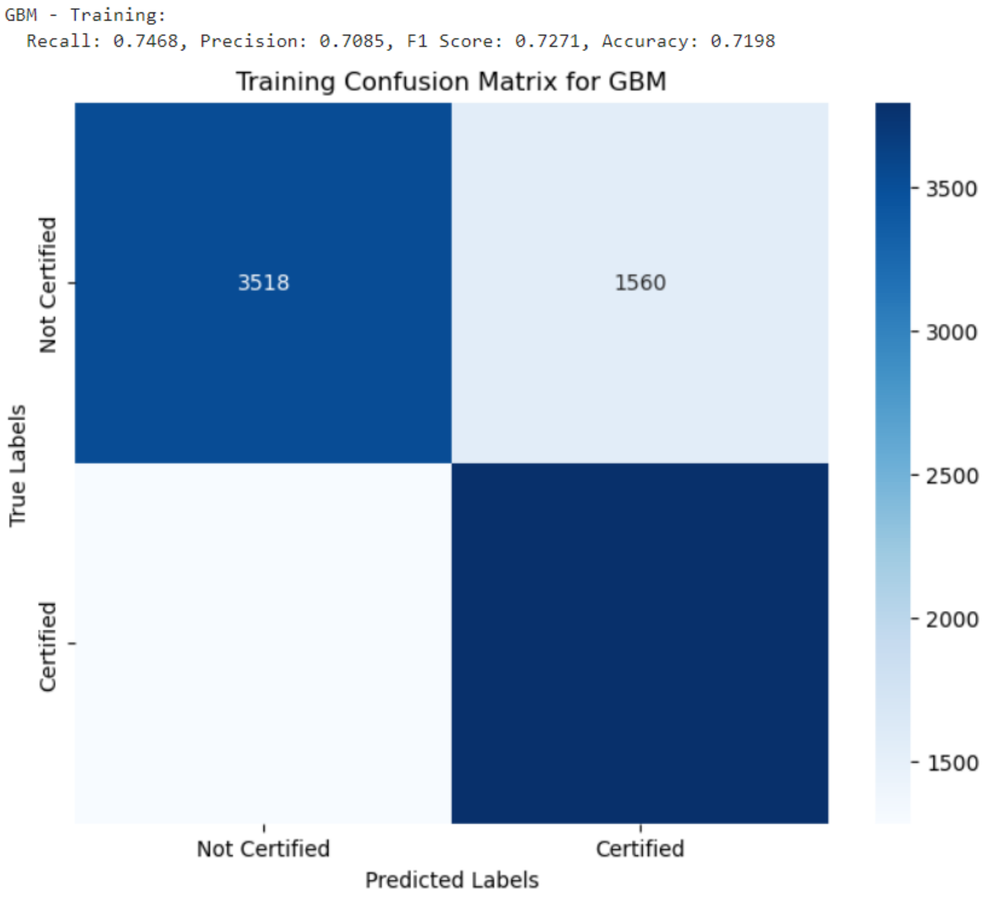
8.3.3 GRADIENT BOOSTING CLASSIFIER

8.3.3.1 Training and Validation Performance

Training:



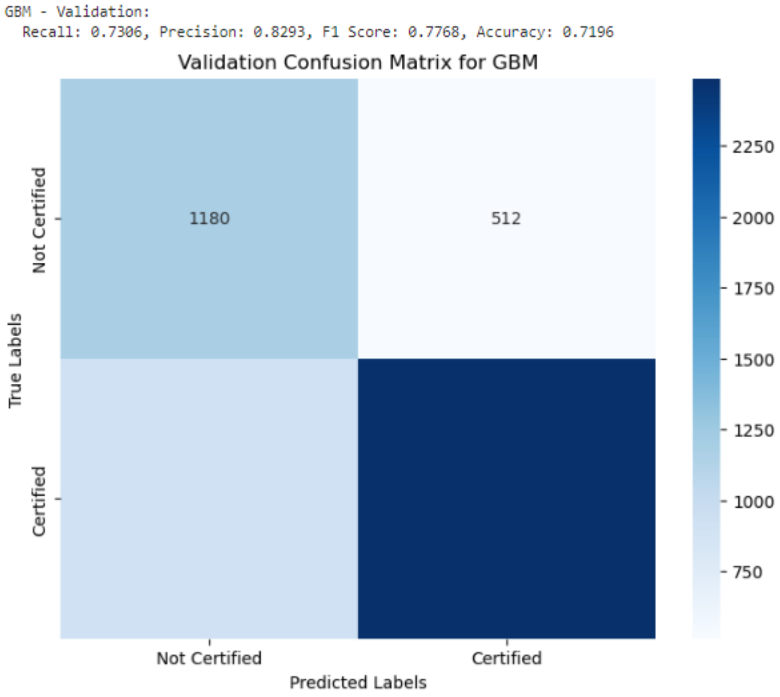
```
GBM:
  Recall: 0.7468, Precision: 0.7085, F1 Score: 0.7271, Accuracy: 0.7198
```



```
GBM - Training:
  Recall: 0.7468, Precision: 0.7085, F1 Score: 0.7271, Accuracy: 0.7198
```

Figure 43 Training Performance of GBM Classifier

Validation:



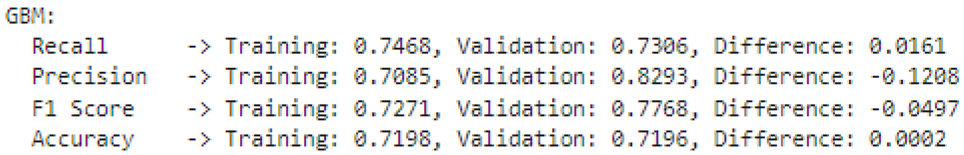
```
GBM - Validation:
  Recall: 0.7306, Precision: 0.8293, F1 Score: 0.7768, Accuracy: 0.7196
```

Figure 44 Validation Performance of GBM Classifier

Observations:

- Lower **training performance** compared to Bagging and Random Forest (Recall: 0.7468, Precision: 0.7085, F1: 0.7271, Accuracy: 0.7198), indicating it is less prone to memorizing the training data.
- **Validation metrics** are relatively stable and close to training performance, with **F1 score** at 0.7768 and **accuracy** at 0.7196.
- GBM demonstrates **strong generalization**, balancing training and validation performance well.

8.3.3.2 Training and Validation Performance Difference



```
GBM:
  Recall      -> Training: 0.7468, Validation: 0.7306, Difference: 0.0161
  Precision   -> Training: 0.7085, Validation: 0.8293, Difference: -0.1208
  F1 Score    -> Training: 0.7271, Validation: 0.7768, Difference: -0.0497
  Accuracy    -> Training: 0.7198, Validation: 0.7196, Difference: 0.0002
```

Table 16 Training and Validation Performance Difference of GBM Classifier

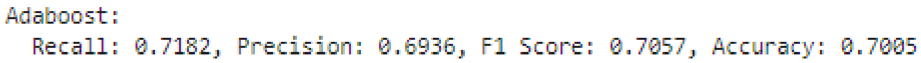
Observations:

- Performance differences between training and validation are minimal, with only a slight drop in recall (0.0161) and accuracy (0.0002), indicating **strong generalization**.
- Negative differences in precision (-0.1208) and F1 score (-0.0497) suggest slightly better performance on validation than training, which may be due to variability in data but reflects good **generalization ability**.

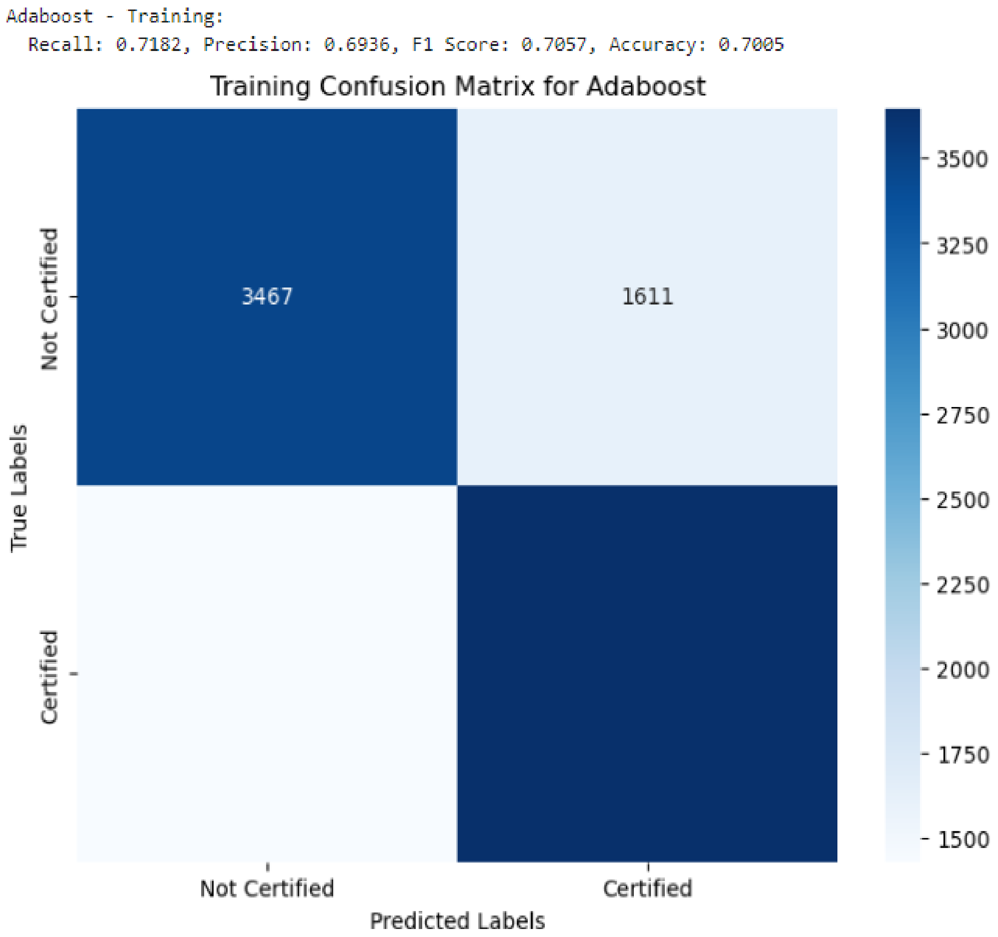
8.3.4 ADA BOOST CLASSIFIER

8.3.4.1 Training and Validation Performance

Training:



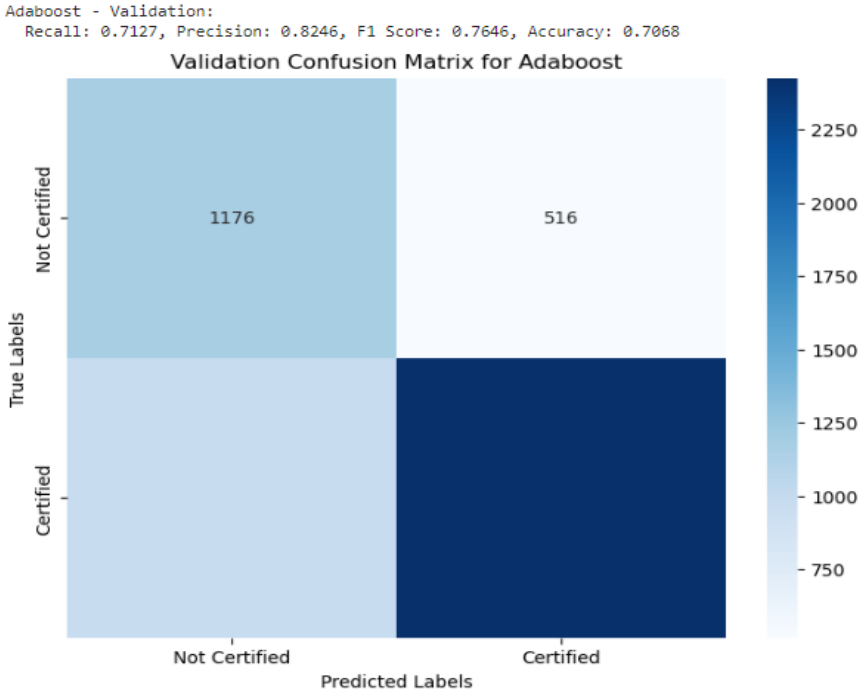
```
Adaboost:
  Recall: 0.7182, Precision: 0.6936, F1 Score: 0.7057, Accuracy: 0.7005
```



```
Adaboost - Training:
  Recall: 0.7182, Precision: 0.6936, F1 Score: 0.7057, Accuracy: 0.7005
```

Figure 45 Training Performance of Adaboost Classifier

Validation:



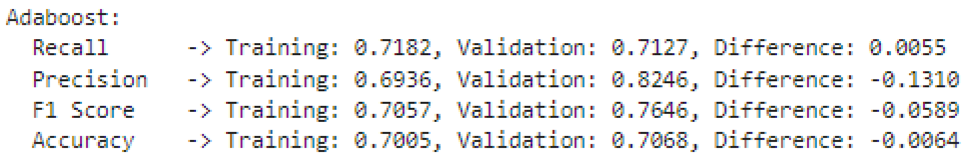
```
Adaboost - Validation:
  Recall: 0.7127, Precision: 0.8246, F1 Score: 0.7646, Accuracy: 0.7068
```

Figure 46 Validation Performance of Adaboost Classifier

Observations:

- Similar to GBM, **moderate training performance** with F1 score of 0.7057 and accuracy of 0.7005, indicating less focus on memorizing training data.
- **Validation performance** is close to the training results, with a recall of 0.7127 and an F1 score of 0.7646, suggesting **good generalization**.
- AdaBoost appears to handle the balance between fitting and generalizing effectively.

8.3.4.2 Training and Validation Performance Difference



```
Adaboost:
  Recall      -> Training: 0.7182, Validation: 0.7127, Difference: 0.0055
  Precision   -> Training: 0.6936, Validation: 0.8246, Difference: -0.1310
  F1 Score    -> Training: 0.7057, Validation: 0.7646, Difference: -0.0589
  Accuracy    -> Training: 0.7005, Validation: 0.7068, Difference: -0.0064
```

Table 17 Training and Validation Performance Difference of Adaboost Classifier

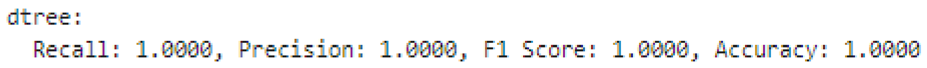
Observations:

- Similar to GBM, AdaBoost shows small differences in recall (0.0055) and accuracy (-0.0064), demonstrating stable performance across datasets.
- Negative precision (-0.1310) and F1 score (-0.0589) differences, where validation metrics slightly outperform training, suggest **good generalization** with a slight underfitting tendency rather than overfitting.

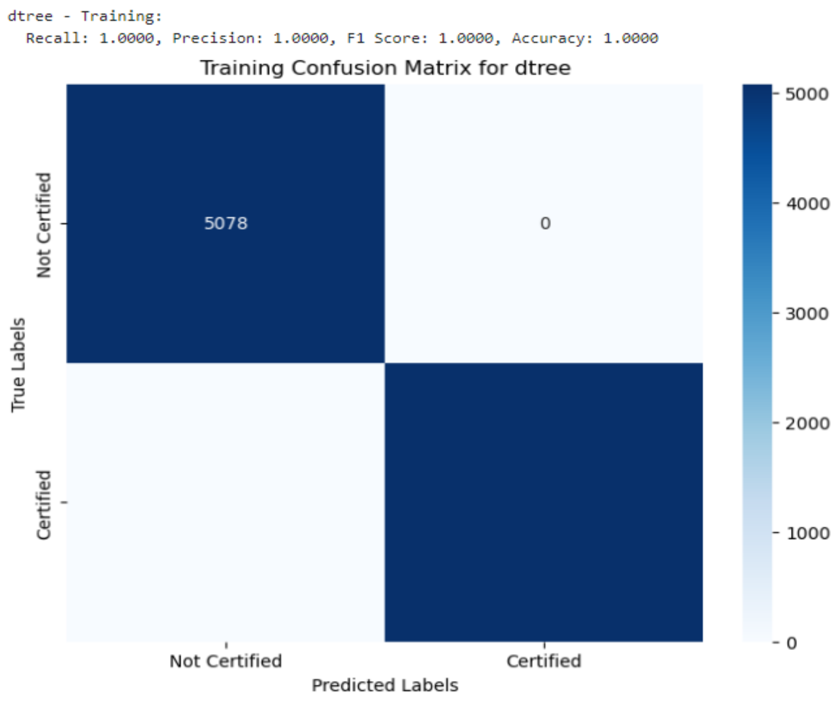
8.3.5 DECISION TREE CLASSIFIER

8.3.5.1 Training and Validation Performance

Training:



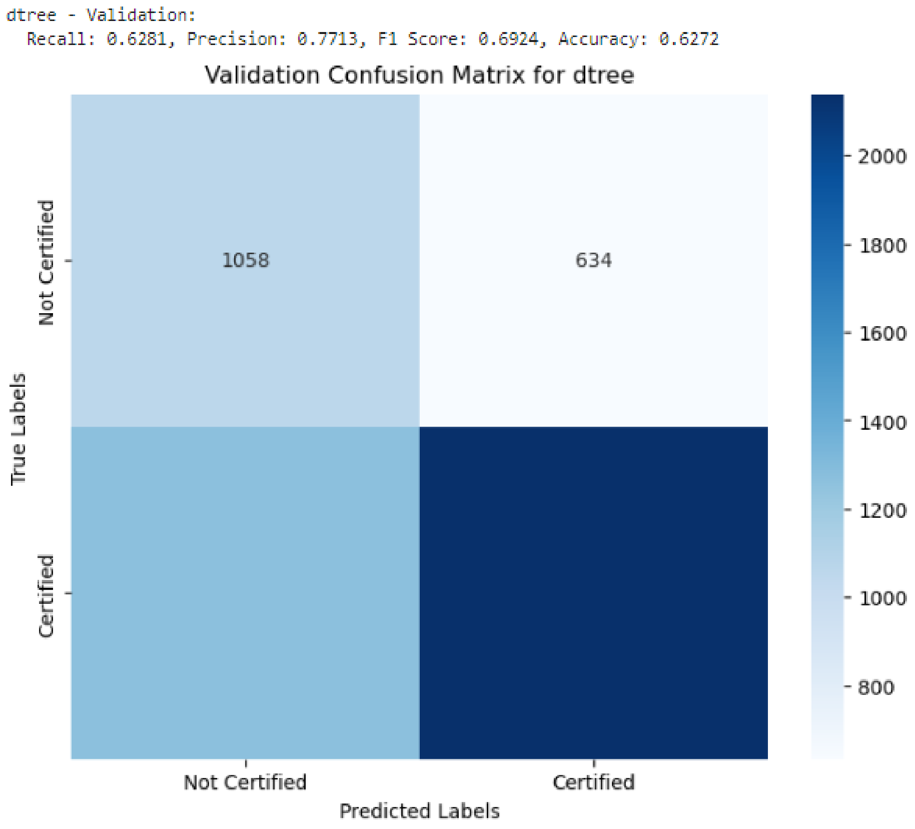
```
dtree:
  Recall: 1.0000, Precision: 1.0000, F1 Score: 1.0000, Accuracy: 1.0000
```



```
dtree - Training:
  Recall: 1.0000, Precision: 1.0000, F1 Score: 1.0000, Accuracy: 1.0000
```

Figure 47 Training Performance of Decision Tree Classifier

Validation:



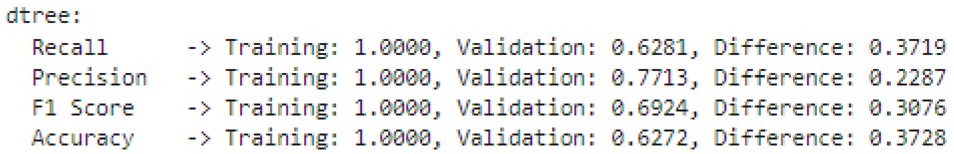
```
dtree - Validation:
  Recall: 0.6281, Precision: 0.7713, F1 Score: 0.6924, Accuracy: 0.6272
```

Figure 48 Validation Performance of Decision Tree Classifier

Observations:

- **Perfect training performance** across all metrics (1.0000), which, like Random Forest, indicates a clear case of **overfitting**.
- In validation, scores decrease significantly, with **recall** at 0.6281 and **accuracy** at 0.6272, confirming that this model is prone to poor generalization on new data.

8.3.5.2 Training and Validation Performance Difference



```
dtree:
  Recall      -> Training: 1.0000, Validation: 0.6281, Difference: 0.3719
  Precision   -> Training: 1.0000, Validation: 0.7713, Difference: 0.2287
  F1 Score    -> Training: 1.0000, Validation: 0.6924, Difference: 0.3076
  Accuracy    -> Training: 1.0000, Validation: 0.6272, Difference: 0.3728
```

Table 18 Training and Validation Performance Difference of Decision Tree Classifier

Observations:

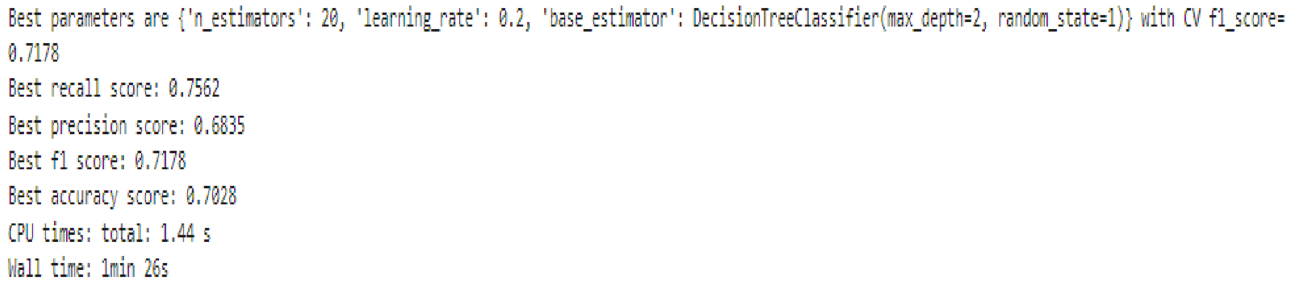
- The highest performance differences across all metrics, especially **recall** (0.3719) and **accuracy** (0.3728), indicate severe overfitting.
- This overfitting is reflected in the large discrepancy between training and validation F1 scores (0.3076), making it unsuitable for generalization.

Summary:

- Adaboost has the best performance followed by GBM model as per the validation performance
- After building 15 models, it was observed that both the GBM and Adaboost models, trained on an undersampled dataset, as well as the GBM model trained on an oversampled dataset, exhibited strong performance on both the training and validation datasets.
- Sometimes models might overfit after undersampling and oversampling, so it's better to tune the models to get a generalized performance
- We will tune these 3 models using the same data (undersampled or oversampled) as we trained them on before.

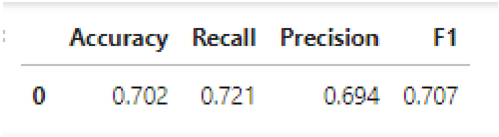
9. HYPERPARAMETER TUNING

9.1. TUNING ADABOOSTCLASSIFIER MODEL WITH UNDERSAMPLED DATA



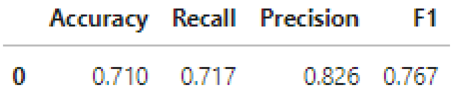
```
Best parameters are {'n_estimators': 20, 'learning_rate': 0.2, 'base_estimator': DecisionTreeClassifier(max_depth=2, random_state=1)} with CV f1_score=
0.7178
Best recall score: 0.7562
Best precision score: 0.6835
Best f1 score: 0.7178
Best accuracy score: 0.7028
CPU times: total: 1.44 s
Wall time: 1min 26s
```

Model performance on training set:



	Accuracy	Recall	Precision	F1
0	0.702	0.721	0.694	0.707

Model performance on validation set:



	Accuracy	Recall	Precision	F1
0	0.710	0.717	0.826	0.767

Table 19 Training and Validation Performance of AdaBoost Performance with Undersampled Data

Observation for AdaBoost Performance with Undersampled Data:

1. **Accuracy**:
   - The **validation accuracy** of 0.710 is moderate but shows a slight improvement over the initial performance (0.702), suggesting some benefit from hyperparameter tuning but with potential for further optimization.
2. **Recall**:
   - **Validation recall** is 0.717, which is slightly lower than the training recall of 0.721. While recall is reasonably close, it may indicate that the model is not capturing all positive cases as effectively as desired.
   - Boosting recall can be critical for minimizing false negatives. Adjustments in hyperparameters like the **learning rate** or **number of estimators** might improve recall.
3. **Precision**:
   - A high **precision** of 0.826 on the validation set compared to 0.694 initially suggests that the tuned model is better at correctly identifying positive cases and avoiding false positives.

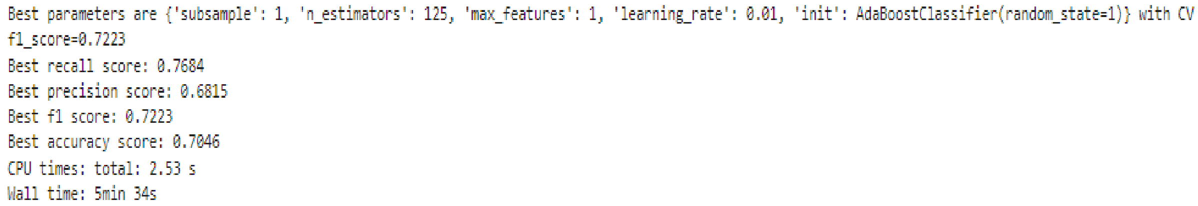
- To further refine the balance between precision and recall, you might adjust **base_estimator parameters** in AdaBoost, potentially switching to a more complex estimator if feasible.

4. **F1 Score**:

   - The **F1 score** of 0.767 is an improvement over the previous 0.707, indicating a better balance between precision and recall but with room for further fine-tuning to achieve optimal results.

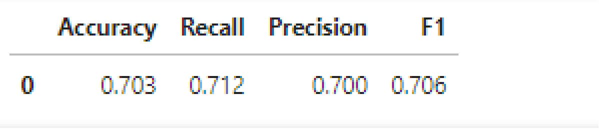
The stable F1 score improvement highlights that tuning has improved the balance between recall and precision, leading to more consistent and dependable performance across both metrics.

9.2. TUNING GRADIENT BOOSTING MODEL WITH UNDERSAMPLED DATA



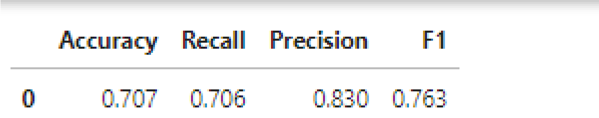
```
Best parameters are {'subsample': 1, 'n_estimators': 125, 'max_features': 1, 'learning_rate': 0.01, 'init': AdaBoostClassifier(random_state=1)} with CV
f1_score=0.7223
Best recall score: 0.7684
Best precision score: 0.6815
Best f1 score: 0.7223
Best accuracy score: 0.7046
CPU times: total: 2.53 s
Wall time: 5min 34s
```

Model performance on training set:



	Accuracy	Recall	Precision	F1
0	0.703	0.712	0.700	0.706

Model performance on validation set:



	Accuracy	Recall	Precision	F1
0	0.707	0.706	0.830	0.763

Table 20 Training and Validation Performance of GBM Performance with Undersampled Data

Observation for Gradient Boosting Model Performance with Undersampled Data

Accuracy:

- The validation accuracy of 0.707 is a modest improvement, indicating stable performance that aligns closely with the training accuracy. This shows that the model generalizes reasonably well but may still benefit from additional tuning to improve accuracy further.

Recall:

- The recall on the validation set is 0.706, only slightly lower than in training. This shows that the model is capturing a good portion of positive cases but could improve in minimizing false negatives.

Precision:

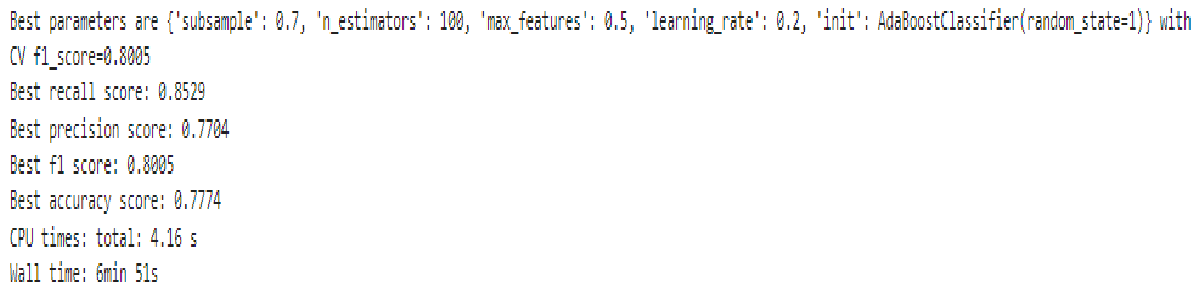
- The validation precision of 0.830, significantly higher than the training precision of 0.700, shows the model's increased accuracy in correctly identifying true positives and reducing false positives. This shift suggests that the tuned model is effectively refining its ability to distinguish positive cases.

F1 Score:

- The F1 score of 0.763 is an improvement from previous levels, indicating a more balanced performance between recall and precision.

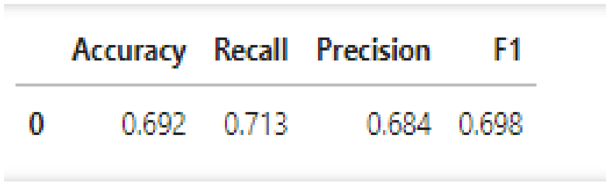
The Gradient Boosting model with undersampled data shows improved precision and a solid recall, indicating a balanced approach to classification.

9.3. TUNING GRADIENT BOOSTING MODEL WITH OVERSAMPLED DATA



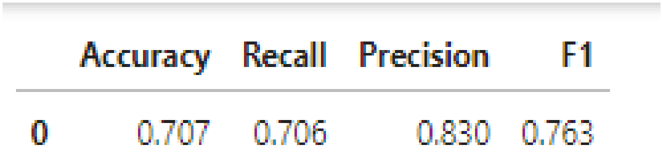
```
Best parameters are {'subsample': 0.7, 'n_estimators': 100, 'max_features': 0.5, 'learning_rate': 0.2, 'init': AdaBoostClassifier(random_state=1)} with
CV f1_score=0.8005
Best recall score: 0.8529
Best precision score: 0.7704
Best f1 score: 0.8005
Best accuracy score: 0.7774
CPU times: total: 4.16 s
Wall time: 6min 51s
```

Model performance on training set:



	Accuracy	Recall	Precision	F1
0	0.692	0.713	0.684	0.698

Model performance on validation set:



	Accuracy	Recall	Precision	F1
0	0.707	0.706	0.830	0.763

Table 21 Training and Validation Performance of GBM Performance with Over Sampled Data

Observations for Gradient Boosting Model with oversampled Data

Accuracy:

- The validation accuracy of 0.707 shows a notable increase from the training accuracy of 0.692, indicating that the model generalizes reasonably well to the validation data. This gain suggests that tuning has enhanced the model's predictive capability, although further optimization might still be possible to improve accuracy.

Recall:

- Validation recall is 0.706, close to the training recall of 0.713, suggesting a well-tuned model that captures a consistent proportion of positive cases in both sets.

Precision:

- Precision on the validation set (0.830) is significantly higher than the training precision (0.684). This improvement implies that the model, post-tuning, effectively reduces false positives, enhancing the accuracy of positive classifications.

F1 Score:

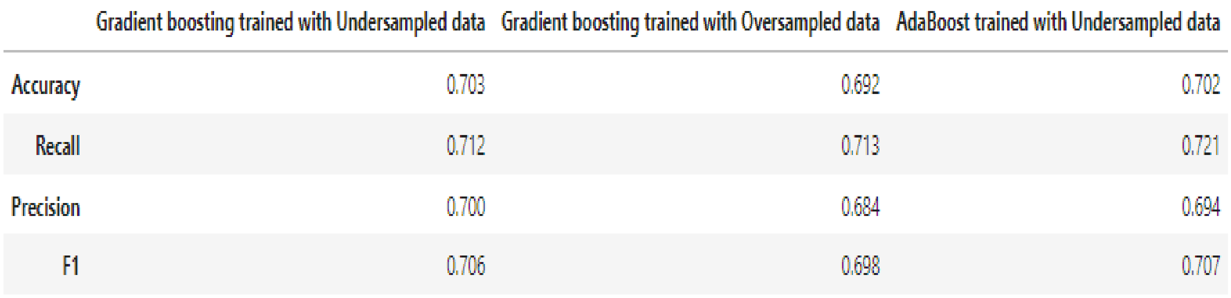
- The F1 score on the validation set has risen to 0.763 from 0.698 in training, reflecting an enhanced balance between precision and recall post-tuning.

Summary

The Gradient Boosting model with oversampled data shows improved model precision and F1 score, enhancing its ability to balance recall and precision for both training and validation data.

10. MODEL COMPARISON AND FINAL MODEL SELECTION

10.1. TRAINING PERFORMANCE COMPARISON



	Gradient boosting trained with Undersampled data	Gradient boosting trained with Oversampled data	AdaBoost trained with Undersampled data
Accuracy	0.703	0.692	0.702
Recall	0.712	0.713	0.721
Precision	0.700	0.684	0.694
F1	0.706	0.698	0.707

Table 22 Training Performance Comparison

Observations

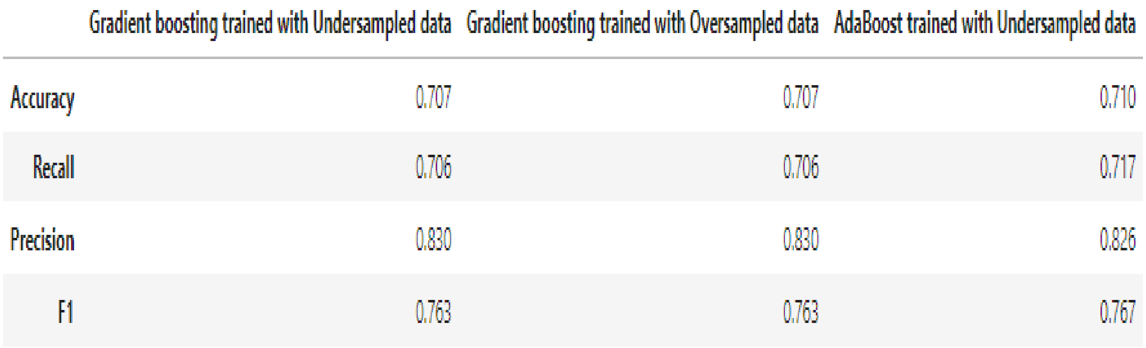
1. **Accuracy**:
   - Gradient Boosting with undersampling shows the highest accuracy (0.703) among the three approaches, slightly outperforming both oversampled Gradient Boosting (0.692) and undersampled AdaBoost (0.702).
   - This suggests that undersampling may have helped the Gradient Boosting model balance predictions effectively, especially in terms of avoiding bias toward majority classes.
2. **Recall**:
   - AdaBoost with undersampling achieved the highest recall at 0.721, indicating that it is capturing a larger proportion of positive cases compared to both versions of Gradient Boosting.
   - Both Gradient Boosting models (undersampled and oversampled) have slightly lower recall values of around 0.712 and 0.713, indicating relatively close but slightly reduced sensitivity to positive cases.
3. **Precision**:
   - The undersampled Gradient Boosting model shows the highest precision (0.700), followed closely by AdaBoost (0.694). The oversampled Gradient Boosting model has the lowest precision (0.684).
   - This pattern suggests that Gradient Boosting with undersampling may be more conservative with its positive classifications, resulting in fewer false positives.
4. **F1 Score**:
   - The highest F1 Score is achieved by AdaBoost with undersampling (0.707), closely followed by Gradient Boosting with undersampling (0.706). The oversampled Gradient Boosting model has the lowest F1 Score (0.698).
   - This balance of precision and recall in AdaBoost may make it a strong candidate for scenarios requiring a reliable balance between positive case sensitivity and precision.

Summary

- **Undersampled AdaBoost** appears to perform well in terms of recall and F1 score, indicating that it is well-suited for cases requiring high recall without sacrificing too much precision.
- **Undersampled Gradient Boosting** achieves high accuracy and precision, making it suitable for situations where avoiding false positives is a priority.
- **Oversampled Gradient Boosting** performs slightly lower across metrics, which may indicate that this approach introduces some imbalance in positive case identification.

Overall, **AdaBoost with undersampling** is the best performer when balancing recall and precision is essential, while **Gradient Boosting with undersampling** may be more effective if the focus is on overall accuracy and precision.

10.2. VALIDATION PERFORMANCE COMPARISON



	Gradient boosting trained with Undersampled data	Gradient boosting trained with Oversampled data	AdaBoost trained with Undersampled data
Accuracy	0.707	0.707	0.710
Recall	0.706	0.706	0.717
Precision	0.830	0.830	0.826
F1	0.763	0.763	0.767

Table 23 Validation Performance Comparison

Observations

1. **Accuracy**:
   - All models achieved similar accuracy levels, with both versions of Gradient Boosting recording 0.707 and AdaBoost slightly ahead at 0.710.
   - The performance is consistent across models, indicating that they are all effectively identifying a significant proportion of the correct predictions.
2. **Recall**:
   - AdaBoost outperforms both Gradient Boosting models in recall, with a score of 0.717 compared to 0.706 for both undersampled and oversampled Gradient Boosting.
   - This suggests that AdaBoost is better at capturing positive cases, which is critical in scenarios where false negatives are more impactful.
3. **Precision**:
   - All models maintain high precision, with both Gradient Boosting models achieving a precision of 0.830 and AdaBoost closely following at 0.826.
   - This indicates that all models are relatively good at minimizing false positives, ensuring that most predicted positive cases are indeed correct.
4. **F1 Score**:
   - AdaBoost has the highest F1 score at 0.767, slightly ahead of both versions of Gradient Boosting, which have an F1 score of 0.763.

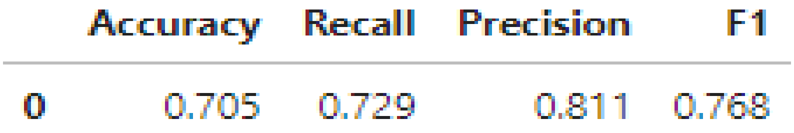
- The F1 score reflects a good balance between precision and recall, and AdaBoost's edge here indicates it might provide a more balanced performance when both false positives and false negatives are considered.

Summary

- **AdaBoost** shows a slight advantage over Gradient Boosting in recall and F1 score, making it the more favorable model in contexts where capturing positive cases is essential while maintaining a solid precision level.
- **Both Gradient Boosting models** have the same performance metrics, which indicates that changing the sampling method may not have significantly impacted their effectiveness at the validation stage.
- The consistent high precision across all models suggests that while they vary in their ability to capture positive cases (recall), they are reliable in their predictions when they do classify instances as positive.

Overall, **AdaBoost trained with undersampled data** stands out as the most balanced model in this comparison, combining good recall with high precision, making it well-suited for applications where both metrics are crucial.

10.3 PERFORMANCE ON TEST SET



	Accuracy	Recall	Precision	F1
0	0.705	0.729	0.811	0.768

Table 24 Performance on Test Set

Performance Observation for Test Data

1. **Accuracy**:
   - The model achieved an accuracy of **0.705**, indicating that it correctly classified approximately 70.5% of the test instances. This suggests a relatively strong overall performance, as the model is able to classify more than two-thirds of the cases correctly.
2. **Recall**:
   - The recall score is **0.729**, meaning the model successfully identified about 72.9% of the actual positive cases. This is a solid result, reflecting the model's capability to capture a significant proportion of the positive instances and minimize false negatives. A recall above 0.70 is generally considered effective, especially in scenarios where identifying positive cases is critical.
3. **Precision**:

  - With a precision of **0.811**, the model demonstrates a strong ability to correctly classify positive cases, as it indicates that approximately 81.1% of instances predicted as positive are indeed true positives. This high precision signifies that the model has a low rate of false positives, which is advantageous in applications where misclassification can lead to significant consequences.

4. **F1 Score**:
    - The F1 score of **0.768** represents a good balance between precision and recall. This score indicates that the model is effectively managing the trade-off between capturing positive cases and ensuring that those predictions are correct. An F1 score above 0.75 is typically seen as a strong performance, demonstrating that the model's predictions are reliable.

- The Adaboost model trained on undersampled data has given ~77% F1 score on the test set
- This performance is in line with what we achieved with this model on the train and validation sets
- So, this is a generalized model.

11. FEATURE IMPORTANCE

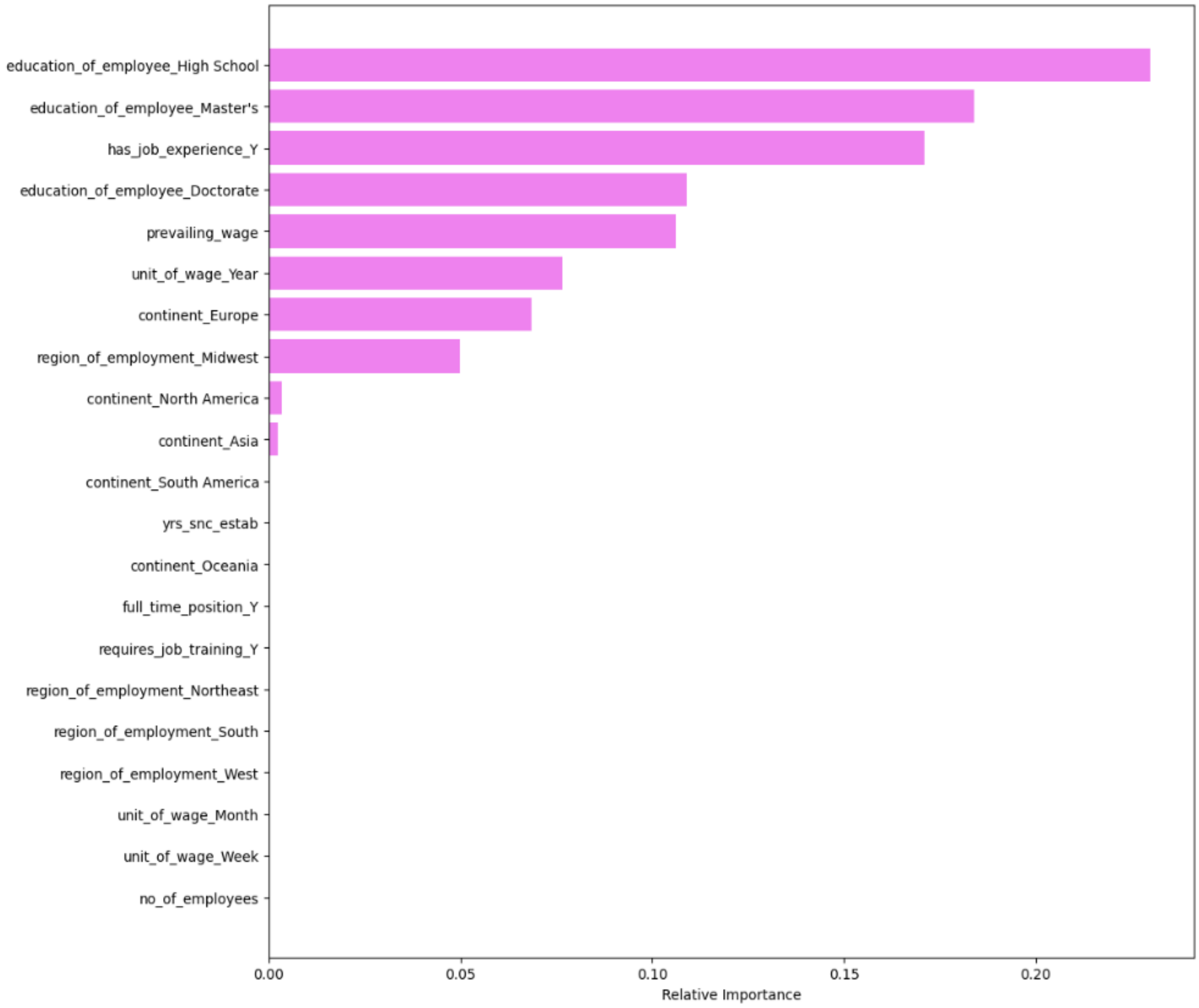


Figure 49 Feature Importance

- We can see that education of employee high school, education of employee masters, has job experience, prevailing wage and education of employee doctorate are the most important features for making predictions

12. ACTIONABLE INSIGHTS AND RECOMMENDATIONS

12.1. ACTIONABLE INSIGHTS

- Balanced F1 Scores: The F1 scores across models indicate a balance between precision and recall, essential for maintaining an effective application review process.
- Certifications vs. Denials: The difference in metrics between certified and denied visa applications indicates key factors influencing approvals. Understanding these factors can enhance predictive accuracy.

- High Precision in Certifications: The models show high precision when predicting certifications, indicating the potential to minimize unnecessary denials by focusing on strong candidate features.
- Recall Analysis: A lower recall in denied cases suggests the models may be missing some applications that should be denied, highlighting a need for refinement in identifying negative cases.
- Feature Importance: Certain features might significantly influence whether a visa is certified or denied. Analyzing these can provide insights into common traits among successful and unsuccessful applicants.

12.2. RECOMMENDATIONS

- Refine Feature Selection: Conduct a thorough analysis of the features impacting certification and denial rates. Focus on enhancing features that have shown to improve recall for denied cases.
- Threshold Adjustments for Denials: Consider adjusting classification thresholds for denied applications to reduce false negatives, ensuring that the model effectively identifies cases that should be denied.
- Targeted Training for Visa Officers: Provide training sessions for visa officers to help them understand model outputs and how to leverage insights for improving the decision-making process.
- Implement Risk Scoring: Develop a risk scoring system that uses model predictions to assess potential visa applicants based on identified features, enabling pre-screening and focused reviews.
- Regularly Update Models: Continuously update the models with new data to ensure they adapt to changing applicant patterns and external factors, such as policy changes or socio-economic conditions.
- Monitor Trends in Certification and Denial: Regularly analyze trends in certified versus denied applications to identify shifts in applicant behavior or potential areas of bias in decision-making.
- Collaborate with Legal Experts: Work closely with immigration law experts to understand the legal implications of denied cases and refine model predictions accordingly, ensuring compliance and fairness.
- Establish Feedback Mechanisms: Create a feedback loop with visa officers to gather insights on model performance, particularly on denied cases, which can inform future model training and adjustments.

By acting on these insights and recommendations, the business can improve its visa application processing efficiency and accuracy, ultimately leading to a more streamlined approval process.