
MACHINE LEARNING - 2

ANALYSIS REPORT

Done By,
N. S. SRIRAM

TABLE OF CONTENTS

S.No	CONTENTS	PAGE NO
1	PROBLEM STATEMENT	1
2	DATA DESCRIPTION	3
	2.1. DUPLICATED VALUES	6
	2.2. MISSING VALUES	6
3	FEATURE ENGINEERING	7
4	EXPLORATORY DATA ANALYSIS	9
	4.1. UNIVARIATE ANALYSIS	9
	4.1.1. NUMERICAL DATA	9
	4.1.2. CATEGORICAL DATA	11
	4.2. BIVARIATE ANALYSIS	14
5	OUTLIERS DETECTION AND TREATMENT	21
6	DATA PREPARATION FOR MODELLING	22
7	MODEL BUILDING	23
	7.1. DECISION TREE CLASSIFIER	23
	7.2. RANDOM FOREST CLASSIFIER	25
	7.3. BAGGING CLASSIFIER	27
	7.4. ADAPTIVE BOOSTING CLASSIFIER	28
	7.5. XGBOOST CLASSIFIER	30
8	OVERSAMPLING AND UNDERSAMPLING OF DATA	32
	8.1. OVERSAMPLING	32
	8.2. UNDERSAMPLING	35
9	HYPERPARAMETER TUNING AND FINAL MODEL	39
10	OVERALL CONCLUSIONS AND RECOMMENDATIONS	42

LIST OF TABLES

S.No	TABLE NAME	PAGE NO.
2.1	FIRST 5 VALUES OF DATASET	3
2.2	DATA INFORMATION	3
2.3	DATA DESCRIPTION	4
2.4	CATEGORICAL DATA DESCRIPTION	4
2.5	PROPORTION OF CATEGORICAL DATA	6
2.6	MISSING VALUES IN THE DATASET	6
3.1	DATA DESCRIPTION OF NO_OF_EMPLOYEES	7
3.2	DATA INFO FOR THE CONVERTED COLUMNS.	8
3.3	DATA INFO AFTER CONVERTING THE DATATYPE OF THE COLUMNS.	8
6.1	PROPORTION OF TARGET VARIABLE ACROSS TRAINING, VALIDATIONS AND TEST SETS.	22
7.1	DECISION TREE PERFORMANCE ON THE TRAINING SET.	23
7.2	DECISION TREE PERFORMANCE ON THE VALIDATION SET.	23
7.3	DECISION TREE PERFORMANCE ON THE TEST SET.	23
7.4	RANDOM FOREST PERFORMANCE IN TRAINING DATA.	25
7.5	RANDOM FOREST PERFORMANCE IN VALIDATION DATA.	25
7.6	RANDOM FOREST PERFORMANCE IN TEST DATA.	25

S.No	TABLE NAME	PAGE NO.
7.7	BAGGING PERFORMANCE ON TRAINING DATA.	27
7.8	BAGGING PERFORMANCE ON VALIDATION DATA.	27
7.9	BAGGING PERFORMANCE ON TEST DATA.	27
7.10	ADAPTIVE BOOST PERFORMANCE ON TRAINING DATA	29
7.11	ADAPTIVE BOOST PERFORMANCE FOR VALIDATION DATA	29
7.12	ADAPTIVE BOOST PERFORMANCE FOR TEST DATA	29
7.13	XGBOOST PERFORMANCE ON TRAINING DATA.	30
7.14	XGBOOST PERFORMANCE ON VALIDATION DATA.	30
7.15	XGBOOST PERFORMANCE ON TEST DATA	31
8.1	OVERSAMPLED DECISION TREE PERFORMANCE ON TRAINING SET.	33
8.2	OVERSAMPLED DECISION TREE PERFORMANCE ON VALIDATION SET.	33
8.3	OVERSAMPLED DECISION TREE PERFORMANCE ON TEST SET.	33
8.4	OVERSAMPLED RANDOM FOREST PERFORMANCE ON TRAINING SET.	33
8.5	OVERSAMPLED RANDOM FOREST PERFORMANCE ON VALIDATION SET.	33
8.6	OVERSAMPLED RANDOM FOREST PERFORMANCE ON TEST SET.	33

S.No	TABLE NAME	PAGE NO.
8.7	OVERSAMPLED BAGGING PERFORMANCE ON TRAINING SET.	34
8.8	OVERSAMPLED BAGGING PERFORMANCE ON TEST SET.	34
8.9	OVERSAMPLED BAGGING PERFORMANCE ON TEST SET	34
8.10	OVERSAMPLED ADAPTIVE BOOST PERFORMANCE ON TRAINING SET.	34
8.11	OVERSAMPLED ADAPTIVE BOOST PERFORMANCE ON VALIDATION SET.	34
8.12	OVERSAMPLED ADAPTIVE BOOST PERFORMANCE ON TEST SET.	34
8.13	OVERSAMPLED XGBOOST PERFORMANCE ON TRAINING DATA.	35
8.14	OVERSAMPLED XGBOOST PERFORMANCE ON VALIDATION DATA.	35
8.15	OVERSAMPLED XGBOOST PERFORMANCE ON TEST DATA.	35
8.16	UNDERSAMPLED DECISION TREE ON TRAINING SET.	36
8.17	UNDERSAMPLED DECISION TREE ON VALIDATION SET.	36
8.18	UNDERSAMPLED DECISION TREE ON TEST SET.	36
8.19	UNDERSAMPLED RANDOM FOREST ON TRAINING SET.	36

S.No	TABLE NAME	PAGE NO.
8.20	UNDERSAMPLED RANDOM FOREST ON VALIDATION SET.	36
8.21	UNDERSAMPLED RANDOM FOREST ON TEST SET.	37
8.22	UNDERSAMPLED BAGGING ON TRAINING SET.	37
8.23	UNDERSAMPLED BAGGING ON VALIDATION SET.	37
8.24	UNDERSAMPLED BAGGING ON TEST SET.	37
8.25	UNDERSAMPLED ADAPTIVE BOOST ON TRAINING SET.	37
8.26	UNDERSAMPLED ADAPTIVE BOOST ON VALIDATION SET.	38
8.27	UNDERSAMPLED ADAPTIVE BOOST ON TEST SET.	38
8.28	UNDERSAMPLED XGBOOST ON TRAINING SET.	38
8.29	UNDERSAMPLED XGBOOST ON VALIDATION SET.	38
8.30	UNDERSAMPLED XGBOOST ON TEST SET.	38
9.1	TUNED MODEL PERFORMANCES ON TRAINING SET.	39
9.2	TUNED MODEL PERFORMANCES ON VALIDATION SET.	39
9.3	TUNED MODEL PERFORMANCES ON TEST SET.	39

LIST OF FIGURES

S.No	FIGURE NAME	PAGE NO.
4.1	DISTRIBUTION AND BOX PLOT FOR NO_OF_EMPLOYEES.	9
4.2	DISTRIBUTION AND BOXPLOT FOR YR_OF_ESTABLISHMENT.	10
4.3	DISTRIBUTION AND BOXPLOT FOR PREVAILING WAGE	10
4.4	COUNT PLOT FOR CONTINENT	11
4.5	COUNT PLOT FOR EDUCATION_OF_EMPLOYEE.	11
4.6	COUNT PLOT FOR HAS_JOB_EXPERIENCE.	12
4.7	COUNT PLOT FOR REQUIRES_JOB_TRAINING.	12
4.8	COUNT PLOT FOR REGION_OF_EMPLOYMENT.	13
4.9	COUNT PLOT FOR UNIT_OF_WAGE.	13
4.10	COUNT PLOT FOR FULL_TIME_POSITION	14
4.11	COUNT PLOT FOR FULL_TIME_POSITION	14
4.12	HEAT MAP	15
4.13	NO_OF_EMPLOYEES VS CASE_STATUS.	15
4.14	PREVAILING_WAGE VS CASE STATUS.	16
4.15	PREVAILING_WAGE VS EDUCATION_OF_EMPLOYEE.	16
4.16	PREVAILING_WAGE VS CONTINENT.	17
4.17	PREVAILING_WAGE VS FULL_TIME_POSITION	17
4.18	FULL_TIME_POSITION VS CASE_STATUS	18
4.19	REQUIRES_JOB_TRAINING VS CASE_STATUS	19
4.20	CONTINENT VS CASE_STATUS	19
4.21	EDUCATION_OF_EMPLOYEE VS CASE_STATUS	20

S.No	FIGURE NAME	PAGE NO.
5.1	BOX PLOT FOR OUTLIER DETECTION.	21
7.1	DECISION TREE CONFUSION MATRIX FOR TRAINING DATA.	24
7.2	DECISION TREE CONFUSION MATRIX FOR VALIDATION DATA.	24
7.3	DECISION TREE CONFUSION MATRIX FOR TEST DATA	25
7.4	RANDOM FOREST CONFUSION MATRIX FOR TRAINING DATA.	26
7.5	RANDOM FOREST CONFUSION MATRIX FOR VALIDATION DATA.	26
7.6	RANDOM FOREST CONFUSION MATRIX FOR TEST DATA.	26
7.7	BAGGING CONFUSION MATRIX FOR TRAINING DATA	28
7.8	BAGGING CONFUSION MATRIX FOR VALIDATION DATA	28
7.9	BAGGING CONFUSION MATRIX FOR TEST DATA	28
7.10	ADAPTIVE BOOSTING CONFUSION MATRIX FOR TRAINING DATA	30
7.11	ADAPTIVE BOOSTING CONFUSION MATRIX FOR VALIDATION DATA	30
7.12	ADAPTIVE BOOSTING CONFUSION MATRIX FOR TEST DATA	30
9.1	CONFUSION MATRIX FOR THE FINAL MODEL ON TRAINING SET.	40

S.No	FIGURE NAME	PAGE NO.
9.2	CONFUSION MATRIX FOR THE FINAL MODEL ON VALIDATION SET.	40
9.3	CONFUSION MATRIX FOR THE FINAL MODEL ON TEST SET	41
9.4	FEATURE IMPORTANCES OF THE FINAL MODEL	41

1. PROBLEM STATEMENT

In FY 2016, the OFLC processed 775,979 employer applications for 1,699,957 positions for temporary and permanent labour 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 a higher chance 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.

DATA DICTIONARY

The data contains the following

- **case_id:** ID of each visa application
- **continent:** Information of continent the employee
- **education_of_employee:** Information of education of the employee
- **has_job_experience:** Does the employee have any job experience?
Y= Yes; N = No
- **requires_job_training:** Does the employee require any job training? Y
= Yes; N = No

- **no_of_employees:** Number of employees in the employer's company
- **yr_of_estab:** Year in which the employer's company was established
- **region_of_employment:** Information of foreign worker's intended region of employment in the US.
- **prevailing_wage:** Average wage paid to similarly employed workers in a specific occupation in the area of intended employment. The purpose of the prevailing wage is to ensure that the foreign worker is not underpaid compared to other workers offering the same or similar service in the same area of employment.
- **unit_of_wage:** Unit of prevailing wage. Values include Hourly, Weekly, Monthly, and Yearly.
- **full_time_position:** Is the position of work full-time? Y = Full-Time Position; N = Part-Time Position
- **case_status:** Flag indicating if the Visa was certified or denied

2. DATA DESCRIPTION

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_wage
EZYV01	Asia	High School	N	N	14513	2007	West	592.2029	Hour
EZYV02	Asia	Master's	Y	N	2412	2002	Northeast	83425.6500	Year
EZYV03	Asia	Bachelor's	N	Y	44444	2008	West	122996.8600	Year
EZYV04	Asia	Bachelor's	N	N	98	1897	West	83434.0300	Year
EZYV05	Africa	Master's	Y	N	1082	2005	South	149907.3900	Year

Table 2.1: First 5 values of the dataset

Table 2.1 shows the first 5 values of the dataset. It shows that there are 12 columns.

Initially there are 25480 rows and 12 columns.

```
<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 2.2: Data Information

Table 2.2 shows data information it states that,

- There are 0 missing values in the dataset. every column has 25480 non null values which is the shape of the dataset.
- Out of all the 12 columns no_of_employees and yr_of_establishment is of the int data type and prevailing_wage is of the float datatype and the rest of the columns is of the object datatype.
- Casae_status column is the target column which needs to converted into a categorical datatype.

	count	mean	std	min	25%	50%	75%	max
no_of_employees	25480.0	5667.043210	22877.928848	-26.0000	1022.00	2109.00	3504.0000	602069.00
yr_of_estab	25480.0	1979.409929	42.366929	1800.0000	1976.00	1997.00	2005.0000	2016.00
prevailing_wage	25480.0	74455.814592	52815.942327	2.1367	34015.48	70308.21	107735.5125	319210.27

Table 2.3: Data Description.

Table 2.3 shows the numerical description of the dataset. From that description the following inferences are made.

- **no_of_employees:** the minimum value is shown to be -26 which is not possible as the employee count cant be negative. the average number of employees is found to be 5667.
- **yr_of_estab:** the minimum values is 1800 meaning there are companies established in 1800 and as recent as 2016
- **prevailing_wage:** the average wage of the employees in the given dataset is 74455. The minimum value is 2.1367 which may be an hourly wage. the max value is 319210.

The description of the categorical data is shown in table 2.4.

	count	unique	top	freq
case_id	25480	25480	EZYV25480	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 2.4: Categorical data description.

From the above table we can arrive the following inferences,

- **case_id:** It is a unique identifier of all the people in the dataset so it can be dropped.
- **continent:** The most people belong to asia continent with 16861 people applying for visa.
- **education_of_employee:** more than 10234 people applying for visa have the bachelors degree
- **has_job_experience:** Most of the people applying for visa have a work experience.
- **requires_job_training:** the frequency is high for no which means more than 22000 people applying for visa doesnt require training.
- **region_of_employment:** there are 5 unique values in which northeast has the high frequency meaning most people have their region of employment as northeast.
- **unit_of_wage:** There are 4 unique values in which the frequency is high for year meaning most people have an yearly wage.
- **full_time_position:** most of the people applying for visa have a fulltime position in their employment as shown that 22773 people have a full time position.
- **case_status:** This is the target variable to be predicted. The frequency is high for certified indicating most of the people (more than 17000) has been certified for visa.

After dropping case id the percentages of the values in the categorical data is shown in the below table 2.5. The following conclusions can be arrived.

- The target variable shows that certified has the high frequency which is 66.8% while 33.2% of the dataset is denied for visa

```

continent          0.661735
Asia              0.146468
Europe             0.129199
North America      0.033438
South America       0.021625
Africa              0.007535
Oceania             0.000000
Name: proportion, dtype: float64
-----
education_of_employee 0.401648
Bachelor's          0.378100
Master's             0.134223
High School           0.086028
Doctorate            0.000000
Name: proportion, dtype: float64
-----
has_job_experience 0.580926
Y                  0.419074
N                  0.115973
Name: proportion, dtype: float64
-----
requires_job_training 0.884027
N                  0.115973
Y                  0.000000
Name: proportion, dtype: float64
-----
region_of_employment 0.282378
Northeast            0.275392
West                 0.258477
Midwest               0.169035
Island                0.014717
Name: proportion, dtype: float64
-----
unit_of_wage         0.901177
Year                 0.084655
Hour                 0.010675
Week                  0.003493
Month                 0.000000
Name: proportion, dtype: float64
-----
full_time_position   0.89376
Y                   0.10624
N                   0.000000
Name: proportion, dtype: float64
-----
case_status           0.667896
Certified             0.332104
Denied                 0.000000
Name: proportion, dtype: float64

```

Table 2.5: Proportion of categorical data.

2.1. DUPLICATED VALUES

There are no duplicated values in the dataset.

2.2. MISSING VALUES

	0
continent	0
education_of_employee	0
has_job_experience	0
requires_job_training	0
no_of_employees	0
yr_of_estab	0
region_of_employment	0
prevailing_wage	0
unit_of_wage	0
full_time_position	0
case_status	0

Table 2.6: Missing values in the dataset.

Table 2.6 shows that There are 0 missing values in the dataset.

3. FEATURE ENGINEERING

- There are some negative values in no_of_employees columns which is not possible.
- So the values are converted into positive values.

There are 33 negative values in the no_of_employees columns.

These values are converted into the absolute values of the same values which removes the negative number and replaces them with the positive value of the same number.

Figure 3.1 shows that the values have been changes to the positive number which changes the minimum values but the mean and standard deviation remains unchanged.

no_of_employees	
count	25480.000000
mean	5667.089207
std	22877.917453
min	11.000000
25%	1022.000000
50%	2109.000000
75%	3504.000000
max	602069.000000

Table 3.1: Data Description of no_of_employees.

- The columns has_job_experience, requires_job_training and full_time_position are categorical in nature and converted into categorical variables and are replaced by the following,
 - Y = 1
 - N = 0

```

RangeIndex: 25480 entries, 0 to 25479
Data columns (total 3 columns):
 #   Column            Non-Null Count Dtype  
--- 
 0   has_job_experience    25480 non-null category
 1   requires_job_training 25480 non-null category
 2   full_time_position    25480 non-null category
dtypes: category(3)
memory usage: 75.1 KB

```

Table 3.2: Data info for the converted columns.

Table 3.2 shows that The values in these columns are converted into 0 and 1 and are converted into categorical columns.

- The column case_status is the target variable and it needs to be encoded as a categorical variable.
- So it is encoded as the following.
 - Certified - 1
 - Denied – 0

Table 3.3 shows that the required columns are converted into categorical columns.

```

RangeIndex: 25480 entries, 0 to 25479
Data columns (total 11 columns):
 #   Column            Non-Null Count Dtype  
--- 
 0   continent          25480 non-null object 
 1   education_of_employee 25480 non-null object 
 2   has_job_experience    25480 non-null category
 3   requires_job_training 25480 non-null category
 4   no_of_employees       25480 non-null int64  
 5   yr_of_estab          25480 non-null int64  
 6   region_of_employment  25480 non-null object 
 7   prevailing_wage       25480 non-null float64
 8   unit_of_wage          25480 non-null object 
 9   full_time_position    25480 non-null category
10   case_status           25480 non-null category
dtypes: category(4), float64(1), int64(2), object(4)
memory usage: 1.5+ MB

```

Table 3.3: Data info after converting the datatype of the columns.

4. EXPLORATORY DATA ANALYSIS

4.1. UNIVARIATE ANALYSIS

4.1.1. NUMERICAL DATA

No_of_employees

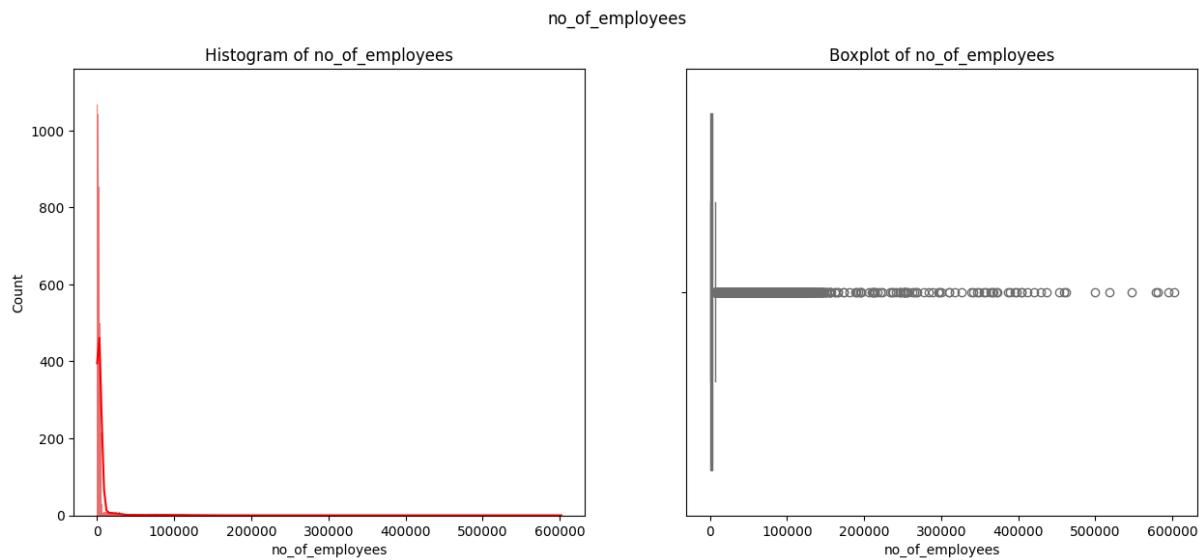


Fig.4.1. Distribution and box plot for no_of_employees.

From figure 4.1 it can infer the following

Shows a heavy right skewed distribution. The box plot shows that the data has a large number of outliers. The data also shows more than 75% of the data have the employee count less than 10.

Yr_of_estab

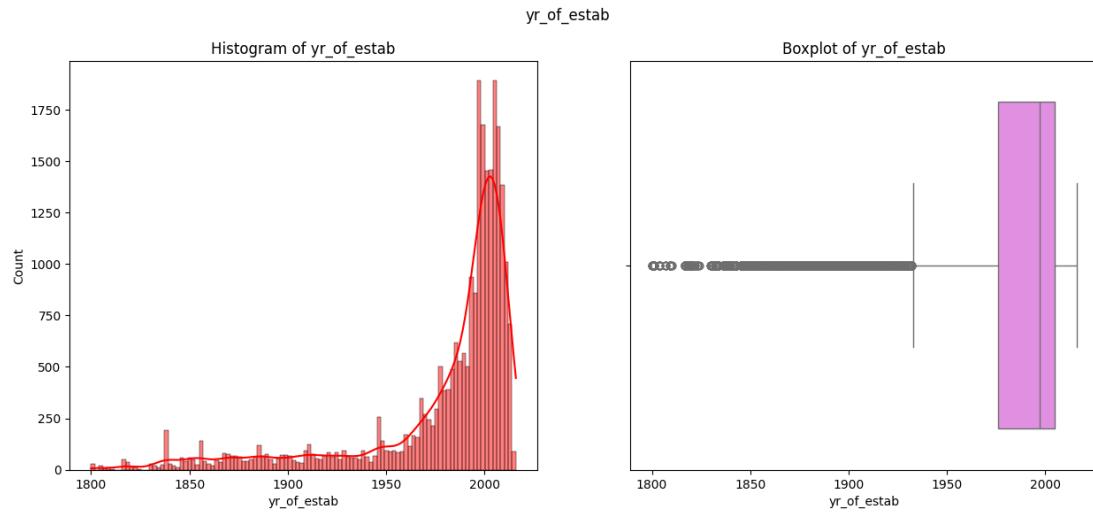


Fig.4.2. Distribution and boxplot for yr of establishment.

Figure 4.2 shows a heavy left skewed distribution. The box plot shows that more than 50% of the data have the year of establishment less than 2000. It also indicates a lot of outliers in the data.

Prevailing_wage

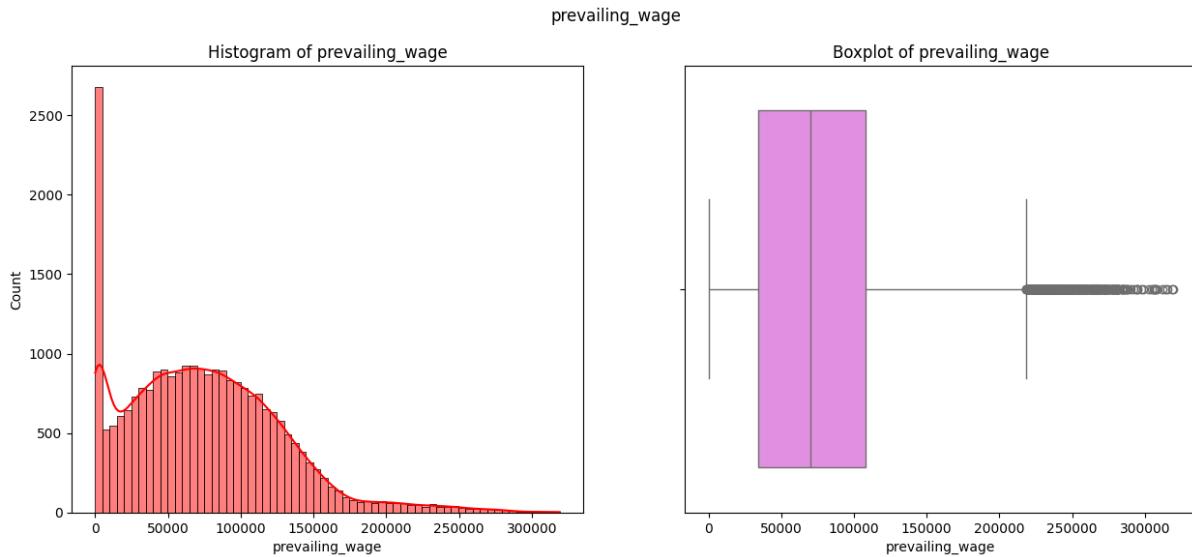


Fig.4.3. Distribution and boxplot for prevailing wage

Figure 4.3 shows a right skewed distribution. The boxplot indicates more than 25% of the data shows wages less than 30,000. More than 75% of the data

shows the prevailing wage of less than 125000. The wages are split between yearly, weekly, hourly and monthly so the presence of outliers is acceptable.

4.1.2. CATEGORICAL DATA

Continent

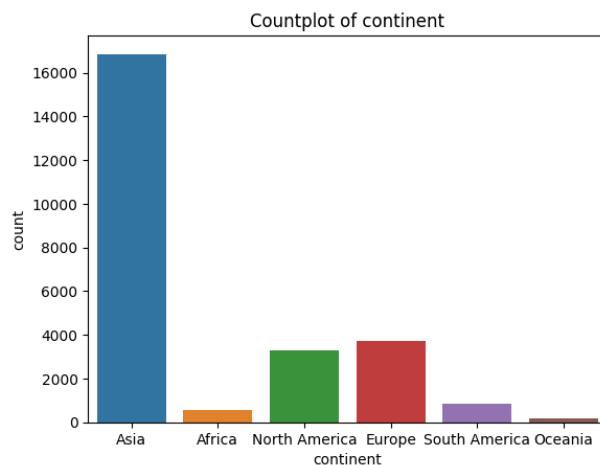


Fig.4.4. Count plot for Continent

Figure 4.4 Asia has the maximum number(more than 16000) of employees applying for visa, followed by Europe and North America.

Education_of_employee

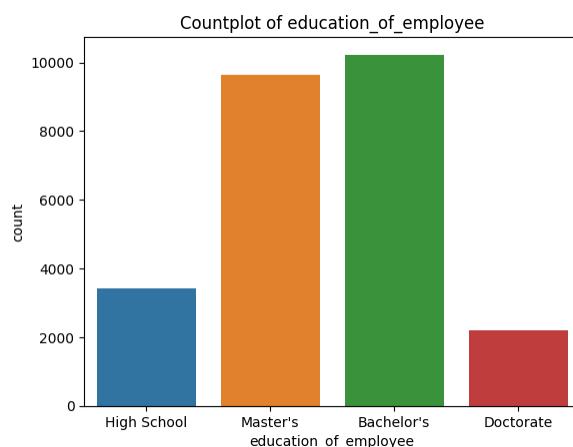


Fig.4.5. Count plot for education_of_employee.

Most people who applied for visa has completed their bachelors degree which has the count of more than 10000 followed by masters and high school. as shown in figure 4.5.

Has_job_experience

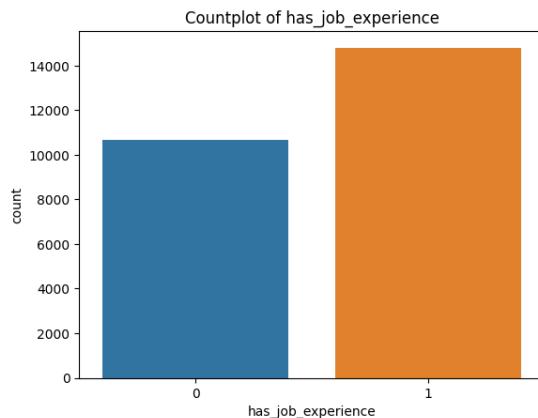


Fig.4.6. Count plot for has_job_experience.

Figure 4.6 shows that most of the employees applying for the visa currently has job experience.

Requires_job_training

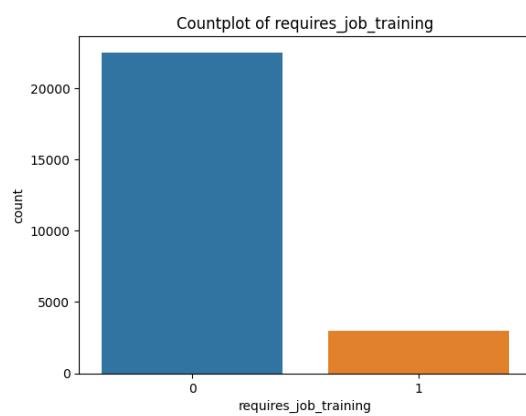


Fig.4.7. Count plot for requires_job_training.

More than 20000 of the employees who applied for visa doesnt require job training which is acceptable as they currently have job experience as seen in figure 4.7.

Region_of_employment

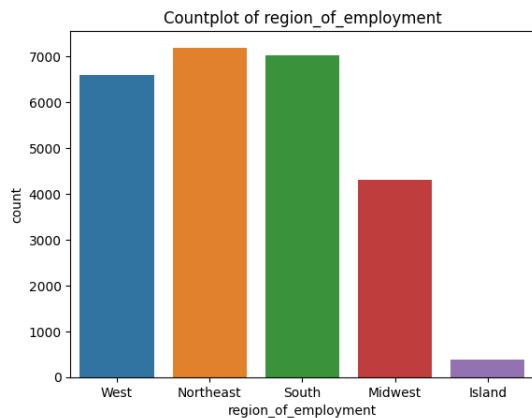


Fig.4.8. Count plot for region_of_employment.

7000 of the employees have their employment in the northeast region of the US followed by south and western regions as seen in figure 4.8.

Unit_of_wage

Most of the employees have given their yearly wage followed by hourly as shown in figure 4.9.

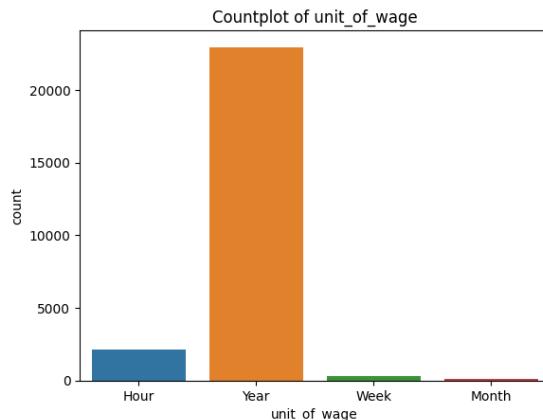


Fig.4.9. Count plot for unit_of_wage.

Full_time_position

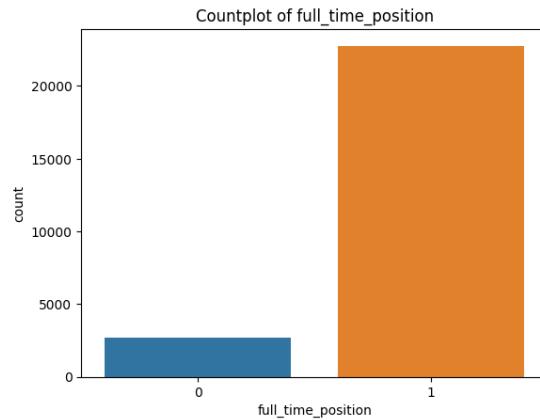


Fig.4.10. Count plot for full_time_position

More than 20000 employees applying for visa have been a full time employment as seen in figure 4.10.

Case_status

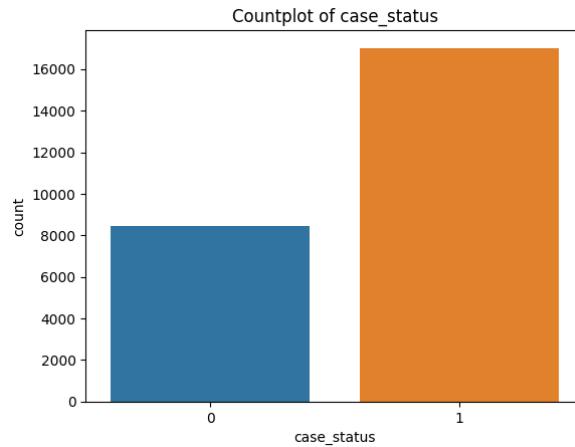


Fig.4.11. Count plot for case_status.

From figure 4.11 it shows that this is the target variable. It shows that more than 16000 of the employees have been certified for visa.

4.2. BIVARIATE ANALYSIS

Figure 4.12 shows the heatmap for the corelations of the numeric variables.

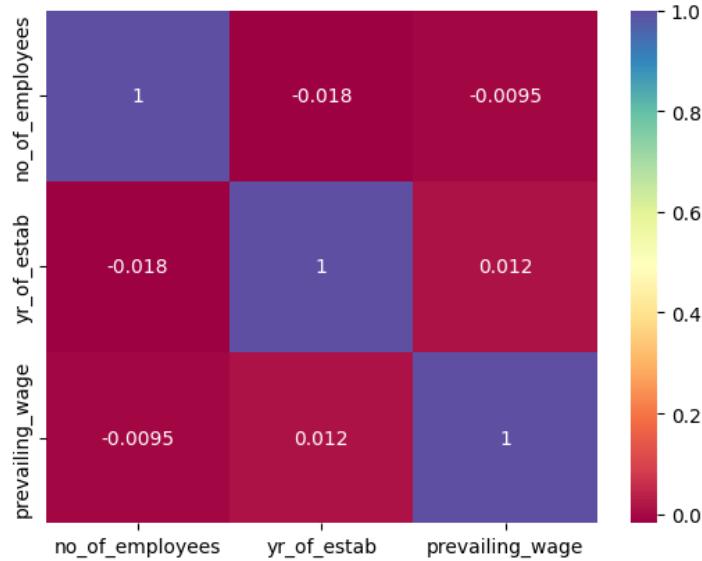


Fig.4.12. Heat map

From the above figure we can infer that the numerical columns doesn't have strong correlation towards each other.

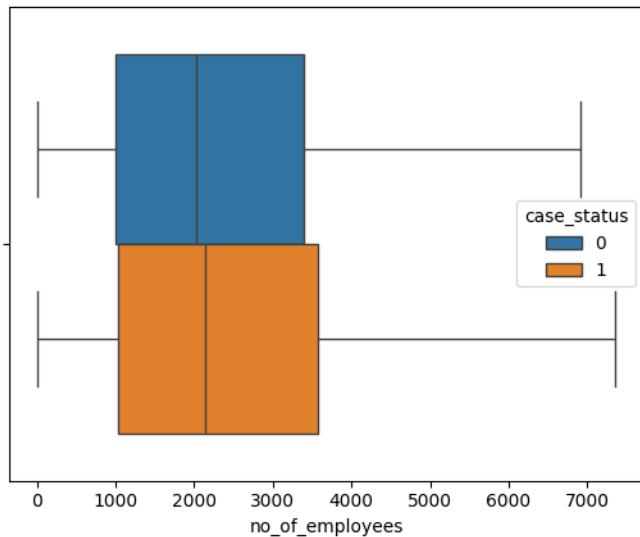


Fig.4.13. no_of_employees vs case_status.

Figure 4.13 shows that the box plot for both approved and denied cases are similar indicating that visa certificating doesn't rely on the no of employees in the company.

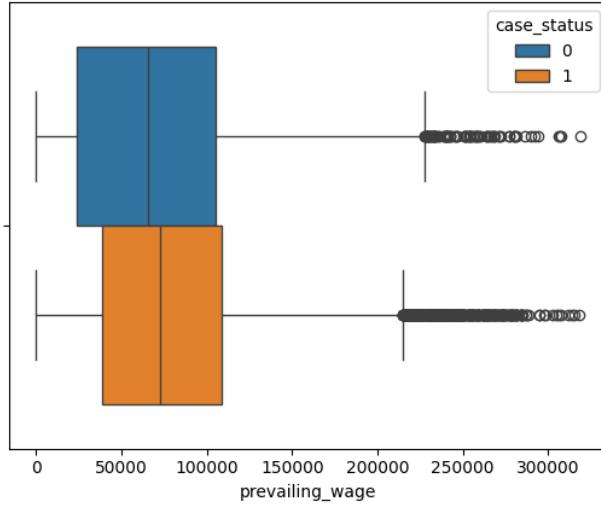


Fig.4.14. prevailing_wage vs case status.

The above plot 4.14 shows that,

- The wage spreads of the certified and denied cases are similar.
- The median lies between 75000 in both cases suggesting higher wage doesn't imply the visa will be certified.
- There are outliers in both cases.

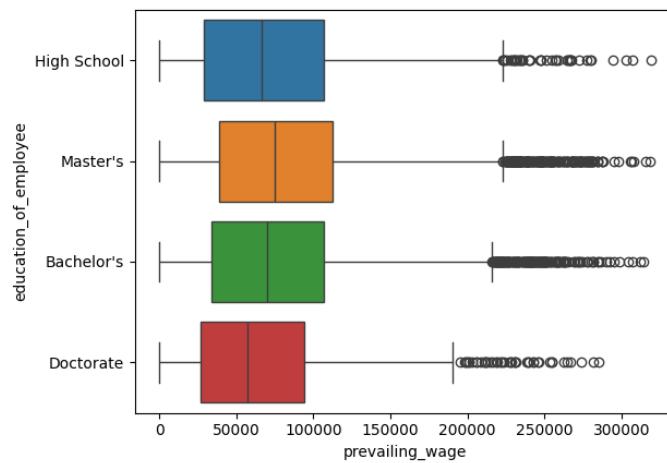


Fig.4.15. prevailing_wage vs education_of_employee.

The plot 4.15 shows the box plot indicates the employees with doctorate education gets a slightly lower wage than others. Higher salary is offered to the employees with a master's degree as the median value is slightly higher.

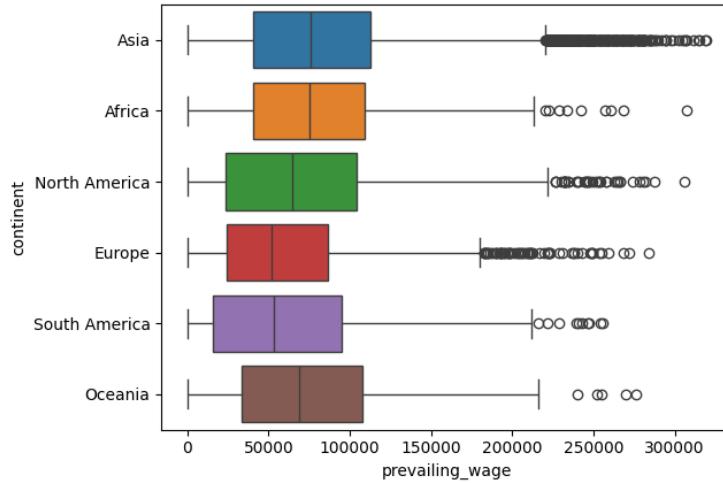


Fig.4.16. prevailing_wage vs continent.

The plot 4.16 shows employees from Asia and oceania gets a higher prevailing wage compared to others. But the outliers are higher for asia and europe indicating they get higher salary compared to others. Median values for all the continents are similar indicating there is no major difference in wages by the continents.

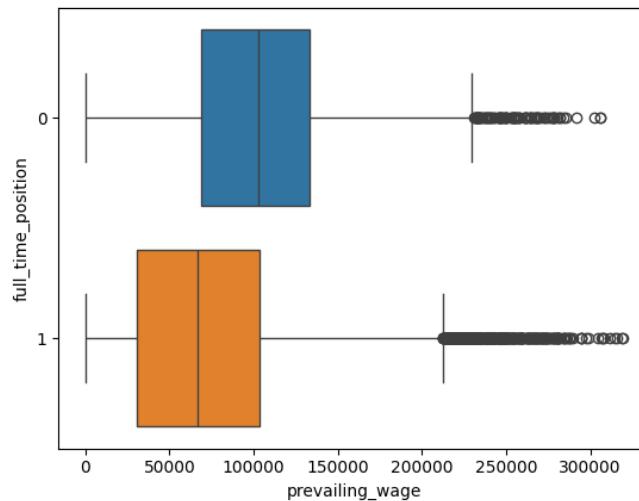


Fig.4.17. prevailing_wage vs full_time_position

Figure 4.17 shows the median values of the non-full time position is higher indicating more demand for their expertise leading to higher salary. Full time position employees has a higher value of outliers indicating that there are employees getting higher salary compared to non-full time employees.

25% of the non-full time employees gets a higher salary compared to a full time employees indicating their expertise and demand.

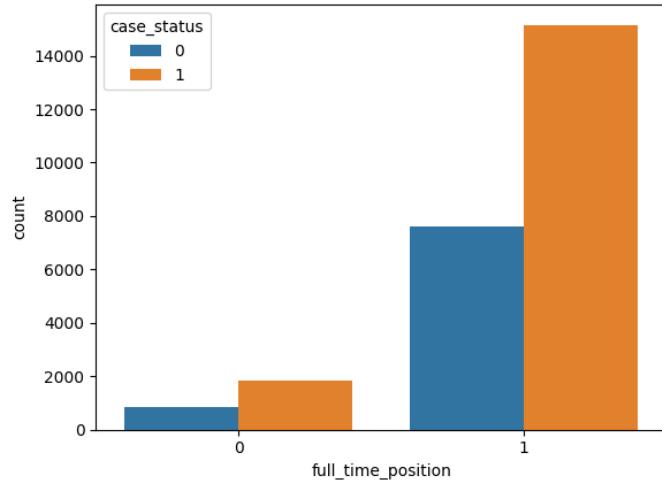


Fig.4.18. full_time_position vs case_status

The plot 4.18 shows,

- The bar plot shows that there are more visa applications for more full time role than non-full time role.
- More than 14000 of the full time role application resulted in approval indicating most employees having a full time job gets approved. This shows a fulltime role has a high likelihood for visa approval
- The number of denials for a full time role is also high. This can be due to various factors like training required etc.
- Non fulltime applications is low compared to the full time. The approved rate is slightly higher than denials but the difference is not too high compared to full time roles showing approval rate is higher for full time roles than non-full time roles.

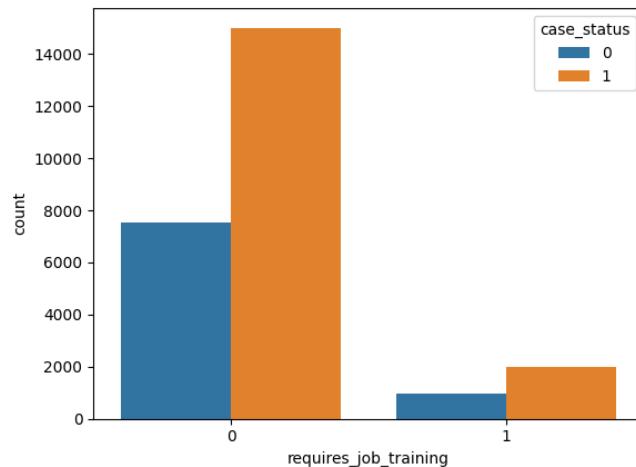


Fig.4.19. requires_job_training vs case_status

Figure 4.19 shows that the maximum visa applications are from employees with no training requirement. The difference between approved and denied is higher for 0 job training than needs job training indicating that the employees that requires no job training has a high likelihood for visa approval.

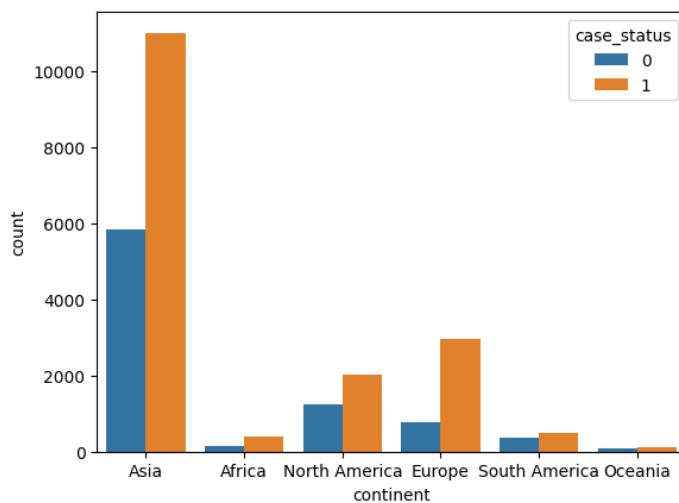


Fig.4.20. continent vs case_status

Figure 4.20 shows that Asia has the highest count for visa applications. The approval rate is higher for Asia and Europe indicating a likelihood for visa approval in that continent. For Oceania and South America the approval rate is very low indicating a low approval rate.

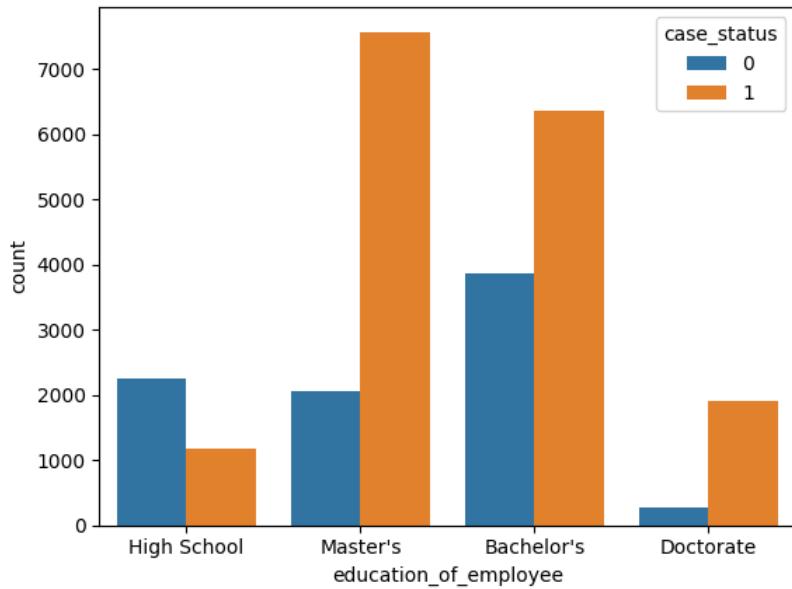


Fig.4.21. education_of_employee vs case_status

Figure 4.21 shows that masters education has a higher approval rate followed by bachelor's degree indicating a high likelihood for visa approval. Employees with high school education has a high likelihood for denial as the denial rate is high compared to others.

5. OUTLIERS DETECTION AND TREATMENT

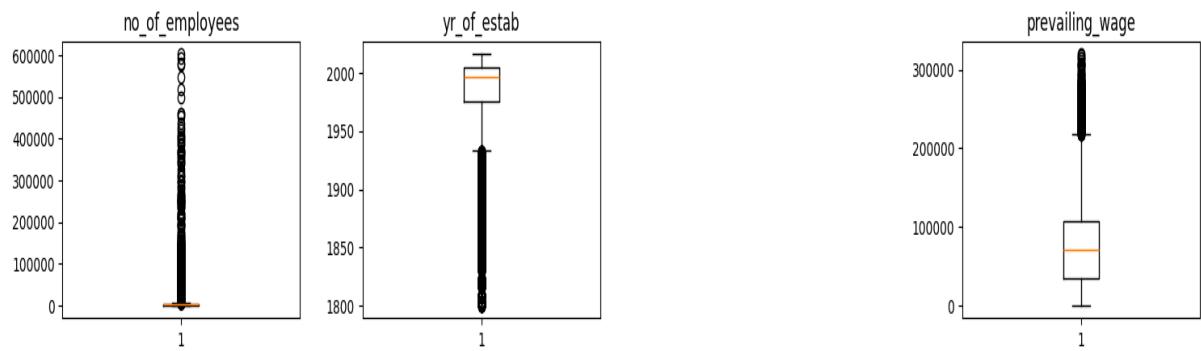


Fig.5.1. Box plot for outlier detection.

However, from the boxplot shown in figure 5.1 it can be inferred that,

- There are many outliers in the data
- However, the outliers are not treated as they are proper values.

6. DATA PREPARATION FOR MODELLING

The models can only made for numerical data and not of string data type. So, the categorical data is needed to be converted to numerical data for modelling.

The dataset is split into x and y variables x being the independent variable and y being the predictor variable which are further split into training, validation and test data to avoid data leakage.

After splitting the data the shape of the training data is 15288, 21. The shape of the validation data is 5096, 21 and the shape of the test data is 5096, 21.

```
case_status
1    0.667909
0    0.332091
Name: proportion, dtype: float64
case_status
1    0.667779
0    0.332221
Name: proportion, dtype: float64
case_status
1    0.667975
0    0.332025
Name: proportion, dtype: float64
```

Table 6.1: Proportion of target variable across training, validations and test sets.

Table 6.1 shows that the training and test set have an equal proportion of the target variable classification.

7. MODEL BUILDING

- For the above data 5 models is to be built namely,
 - Decision tree classifier
 - Random forest classifier
 - bagging classifier
 - Adaptive boosting classifier
 - XGboost classifier
- From the above 5 models the best model is selected by comparing the performance metrics on training, validation and test sets.

Metric to be used

The metrics to be used for comparing the dataset is f1 score because f1 score balances the models ability to find the true positives and true negatives respectively.

7.1. DECISION TREE CLASSIFIER

The decision tree classifier is built using the default parameters and trained on the training set. The model performance of the training, validation and test dataset are given in table 7.1, 7.2 and 7.3 respectively.

	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Table 7.1: Decision tree performance on the training set.

	Accuracy	Recall	Precision	F1
0	0.659537	0.734058	0.750601	0.742237

Table 7.2: Decision tree performance on the validation set.

	Accuracy	Recall	Precision	F1
0	0.663462	0.745006	0.749631	0.747311

Table 7.3: Decision tree performance on the test set.

From the above performance metrics, we can infer that the model is oversampled as the f1 score is varied between training, validation and test sets. F1 score is similar for validation and test sets. A f1 score of 0.74 is test dataset is acceptable for the model.

The confusion matrix for the training, validation and test dataset is shown in figure 7.1,7.2 and 7.3. The matrix shows the model predicted 49% of the true positives and 16% of the true negatives in test set.

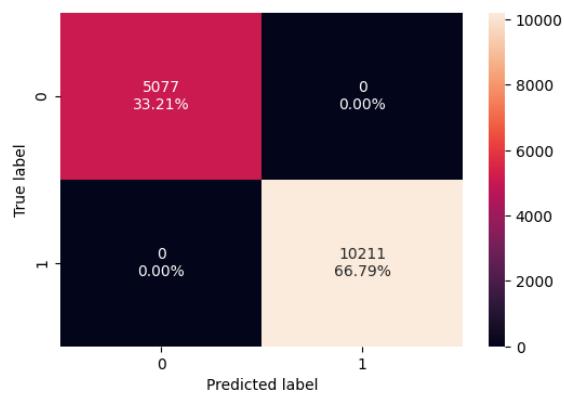


Fig.7.1. Decision tree confusion matrix for training data.

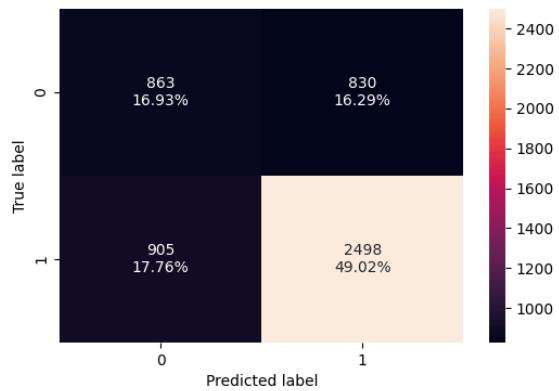


Fig.7.2. Decision tree confusion matrix for validation data.

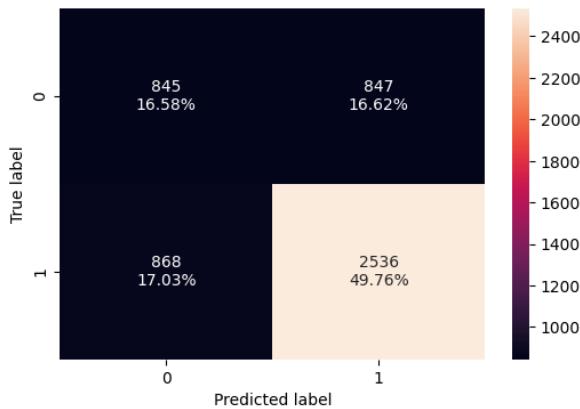


Fig.7.3. Decision tree confusion matrix for test data.

7.2. RANDOM FOREST CLASSIFIER

The random forest classifier is built using the default parameters and trained on the training set. The model performance of the training, validation and test dataset are given in table 7.4, 7.5 and 7.6 respectively.

	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Table 7.4: Random forest performance in training data.

	Accuracy	Recall	Precision	F1
0	0.730965	0.83779	0.776839	0.806164

Table 7.5: Random forest performance in validation data.

	Accuracy	Recall	Precision	F1
0	0.714482	0.827262	0.764594	0.794695

Table 7.6: Random forest performance in test data.

- From the above table it can be inferred that the performance metrics in test and validation sets is higher than decision tree classifier.
- The model is oversampled as the performance metrics across training, validation and test sets are different.

The confusion matrix for the same is given in the figure 7.4,7.5 and 7.6 which indicates that the model can identify 55% of the true positives and 16% of the true negatives in the test dataset.

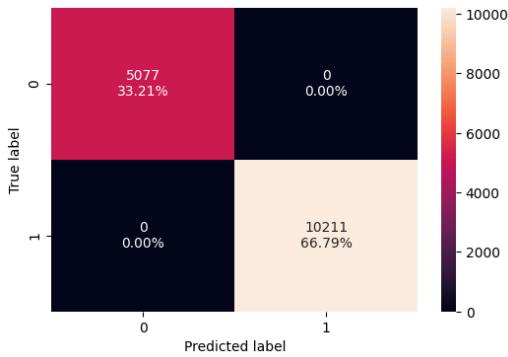


Fig.7.4. Random forest confusion matrix for training data.

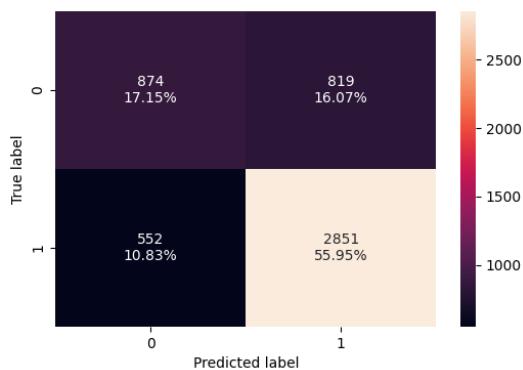


Fig.7.5. Random forest confusion matrix for validation data.

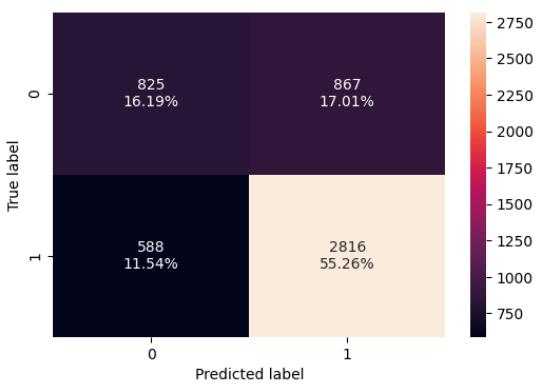


Fig.7.6. Random forest confusion matrix for test data.

7.3. BAGGING CLASSIFIER

The bagging classifier is built using the default parameters and trained on the training set. The model performance of the training, validation and test dataset are given in table 7.7, 7.8 and 7.9 respectively.

	Accuracy	Recall	Precision	F1
0	0.983582	0.986485	0.988906	0.987694

Table 7.7: Bagging performance on training data.

	Accuracy	Recall	Precision	F1
0	0.695251	0.77079	0.772379	0.771584

Table 7.8: Bagging performance on validation data.

	Accuracy	Recall	Precision	F1
0	0.695644	0.776146	0.770038	0.77308

Table 7.9: Bagging performance on test data.

- From the above table it can be inferred that the model is slightly oversampled as the training set performance is higher than validation and test set performance.
- The f1 score is higher than decision tree model but lesser than random forest model.

The confusion matrix for the same is given in the figure 7.7,7.8 and 7.9 which indicates that the model can identify 51% of the true positives and 17% of the true negatives in the test dataset.

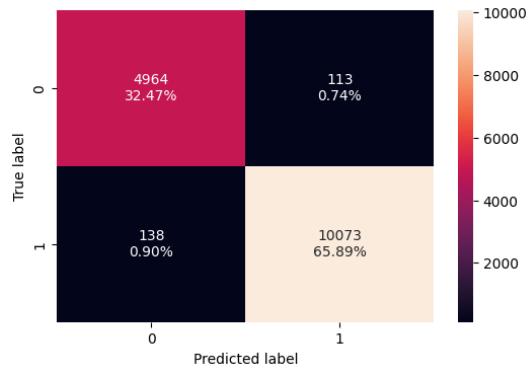


Fig.7.7. Bagging confusion matrix for training data.

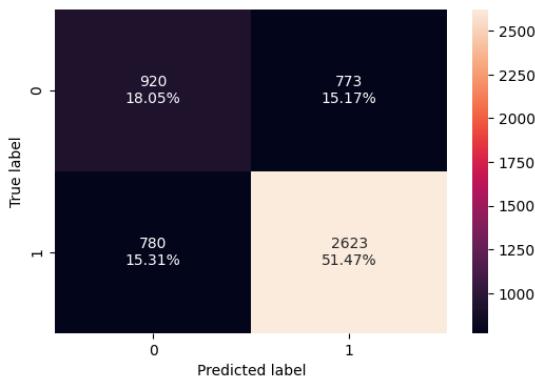


Fig.7.8. Bagging confusion matrix for validation data.

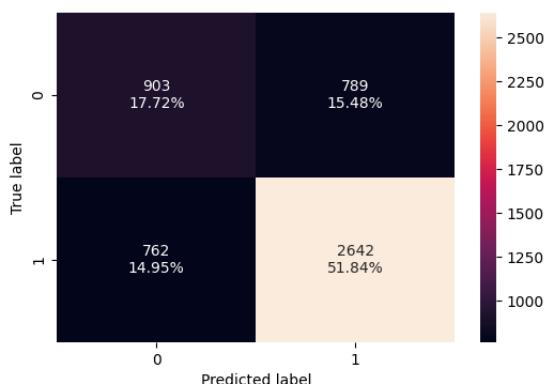


Fig.7.9. Bagging confusion matrix for test data.

7.4. ADAPTIVE BOOST CLASSIFIER

The adaptive boosting classifier is built using the default parameters and trained on the training set. The model performance of the training, validation and test dataset are given in table 7.10, 7.11 and 7.12 respectively.

	Accuracy	Recall	Precision	F1
0	0.744767	0.883851	0.768674	0.822249

Table 7.10: Adaptive boost performance on training data.

	Accuracy	Recall	Precision	F1
0	0.741954	0.880694	0.767281	0.820085

Table 7.11: Adaptive boost performance for validation data.

	Accuracy	Recall	Precision	F1
0	0.732143	0.87691	0.759349	0.813906

Table 7.12: Adaptive boost performance for test data.

- From the above table, it can be inferred that The model is not overfit or underfit as the performance in training, validation and test sets are similar.
- The f1 score on the validation and test sets are higher compared to previous models i.e., decision tree, random forest and adaptive boosting classifier.

The confusion matrix for the same is given in the figure 7.10, 7.11 and 7.12 which indicates that the model can identify 58% of the true positives and 14% of the true negatives in the test dataset.

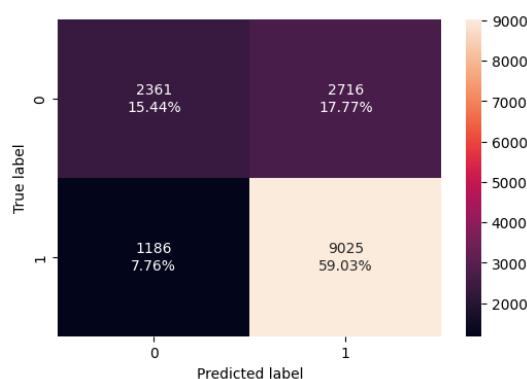


Fig.7.10. Adaptive boosting confusion matrix for training data.

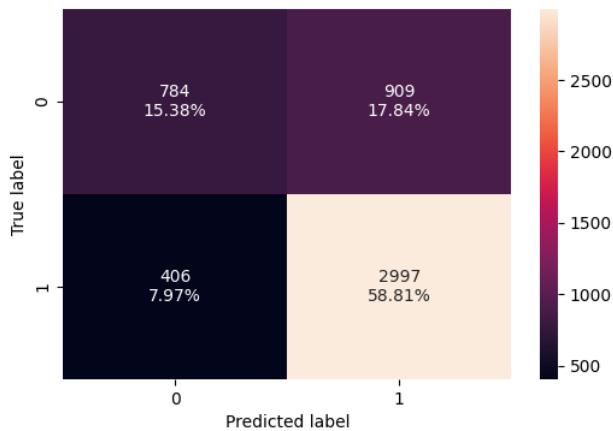


Fig.7.11. Adaptive boosting confusion matrix for validation data.

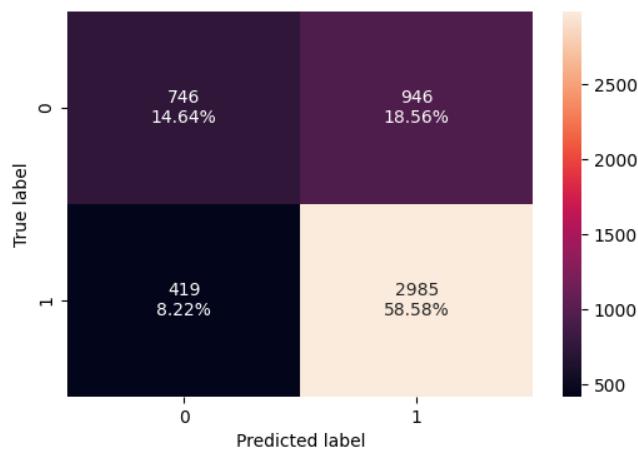


Fig.7.12. Adaptive boosting confusion matrix for test data.

7.5. XGBOOST CLASSIFIER

The xgboost classifier is built using the default parameters and trained on the training set. The model performance of the training, validation and test dataset are given in table 7.13, 7.14 and 7.15 respectively.

	Accuracy	Recall	Precision	F1
0	0.854657	0.94075	0.855921	0.896333

Table 7.13: xgboost performance on training data.

	Accuracy	Recall	Precision	F1
0	0.737637	0.856303	0.774588	0.813398

Table 7.14: xgboost performance on validation data.

	Accuracy	Recall	Precision	F1
0	0.726452	0.856639	0.762951	0.807086

Table 7.15: xgboost performance on test data.

- From the table above it can be inferred that The model is not over sampled or under sampled as the performance metrics are fairly similar.
- The f1 score is 0.80 showing the model can predict 80% of the visa applications certification or denial.

8. OVERSAMPLING AND UNDERSAMPLING OF THE DATA

- From the table 2.5 it shows that the target variable class is not balanced.
- So the model may be biased for the certified visa applications as the case status of certified has the highest frequency.
- To balance the data and making the model unbiased the training dataset is oversampled and under sampled. The model is selected based on the performance the f1 score.
- Oversampling is done using SMOTE and under sampling is done using random under sampler.

8.1. OVERSAMPLING

Oversampling is done only on training data and the models performance is checked in validation and test data.

- Before Oversampling, count of label '1': 10211.
- Before Oversampling, count of label '0': 5077.
- After Oversampling, count of label '1': 10211.
- After Oversampling, count of label '0': 8168.
- After Oversampling, the shape of train_X: (18379, 21).
- After Oversampling, the shape of train_y: (18379,).

Decision tree classifier

The model performance from training, validation and test set is shown in the table 8.1, 8.2 and 8.3.

	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Table 8.1: Oversampled decision tree performance on training set.

	Accuracy	Recall	Precision	F1
0	0.654239	0.730826	0.746175	0.73842

Table 8.2: Oversampled decision tree performance on validation set.

	Accuracy	Recall	Precision	F1
0	0.651688	0.72973	0.743935	0.736764

Table 8.3: Oversampled decision tree performance on test set.

From the above table, it can be inferred that the model is oversampled. The f1 score doesn't change much likely due to over sampling.

Random forest classifier

The model performance from training, validation and test set is shown in the table 8.4, 8.5 and 8.6.

	Accuracy	Recall	Precision	F1
0	0.999946	1.0	0.999902	0.999951

Table 8.4: Oversampled random forest performance on training set.

	Accuracy	Recall	Precision	F1
0	0.729396	0.83397	0.777108	0.804536

Table 8.5: Oversampled random forest performance on validation set.

	Accuracy	Recall	Precision	F1
0	0.717033	0.828731	0.766576	0.796443

Table 8.6: Oversampled random forest performance on test set.

From the above table, it can be inferred that the model is overfit. Similar to decision tree classifier the f1 score doesn't change compared to previous model likely due to oversampling.

Bagging Classifier

The model performance from training, validation and test set is shown in the table 8.7, 8.8 and 8.9.

	Accuracy	Recall	Precision	F1
0	0.987649	0.9858	0.99192	0.98885

Table 8.7: Oversampled bagging performance on training set.

	Accuracy	Recall	Precision	F1
0	0.701531	0.765501	0.782752	0.774031

Table 8.8: Oversampled bagging performance on test set.

	Accuracy	Recall	Precision	F1
0	0.693289	0.771445	0.769862	0.770653

Table 8.9: Oversampled bagging performance on test set

From the above table, it can be inferred that the model performance doesn't change after oversampling.

Adaptive boosting classifier

The model performance from training, validation and test set is shown in the table 8.10, 8.11 and 8.12.

	Accuracy	Recall	Precision	F1
0	0.774689	0.843796	0.771905	0.806251

Table 8.10: Oversampled adaptive boost performance on training set.

	Accuracy	Recall	Precision	F1
0	0.733124	0.842198	0.776904	0.808235

Table 8.11: Oversampled adaptive boost performance on validation set.

	Accuracy	Recall	Precision	F1
0	0.725471	0.84577	0.76712	0.804527

Table 8.12: Oversampled adaptive boost performance on test set.

From the above table it can be inferred that the f1 score is fairly similar in training validation and test sets.

XGboost classifier

The model performance from training, validation and test set is shown in the table 8.13, 8.14 and 8.15.

	Accuracy	Recall	Precision	F1
0	0.869525	0.927431	0.851083	0.887618

Table 8.13: Oversampled xgboost performance on training data.

	Accuracy	Recall	Precision	F1
0	0.737834	0.850132	0.777897	0.812412

Table 8.14: Oversampled xgboost performance on validation data.

	Accuracy	Recall	Precision	F1
0	0.718014	0.845182	0.759704	0.800167

Table 8.15: Oversampled xgboost performance on test data.

From the above table, it can be inferred that the model performance is similar in training and test sets. The performance of the model doesn't change after oversampling.

8.2. UNDERSAMPLING

Oversampling is done only on training data and the models performance is checked in validation and test data.

- Before Under Sampling, count of label '1': 10211.
- Before Under Sampling, count of label '0': 5077.
- After Under Sampling, count of label '1': 5077.
- After Under Sampling, count of label '0': 5077.
- After Under Sampling, the shape of train_X: (10154, 21).

- After Under Sampling, the shape of train_y: (10154,).

Decision tree classifier

The model performance from training, validation and test set is shown in the table 8.16, 8.17 and 8.18.

	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Table 8.16: Undersampled decision tree on training set.

	Accuracy	Recall	Precision	F1
0	0.62029	0.620041	0.766715	0.685621

Table 8.17: Undersampled decision tree on validation set.

	Accuracy	Recall	Precision	F1
0	0.624411	0.625441	0.769147	0.68989

Table 8.18: Undersampled decision tree on test set.

From the above table it is inferred that the model is overfit. The model performance is very low as the training dataset is undersampled which may be due to loss of information.

Random forest classifier

The model performance from training, validation and test set is shown in the table 8.19, 8.20 and 8.21.

	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Table 8.19: Undersampled random forest on training set.

	Accuracy	Recall	Precision	F1
0	0.680338	0.672348	0.81656	0.73747

Table 8.20: Undersampled random forest on validation set.

	Accuracy	Recall	Precision	F1
0	0.673666	0.674794	0.805117	0.734218

Table 8.21: Undersampled random forest on test set.

From the table above it can be inferred that the model is overfit. The performance is very low compared to previous oversampled and normal model. May be due to loss of information.

Bagging classifier

The model performance from training, validation and test set is shown in the table 8.22, 8.23 and 8.24.

	Accuracy	Recall	Precision	F1
0	0.979811	0.965531	0.993917	0.979518

Table 8.22: Undersampled bagging on training set.

	Accuracy	Recall	Precision	F1
0	0.642857	0.604173	0.81297	0.693189

Table 8.23: Undersampled bagging on validation set.

	Accuracy	Recall	Precision	F1
0	0.644623	0.607814	0.812967	0.695579

Table 8.24: Undersampled bagging on test set.

From the table above it is inferred that, the model performance is low compared to the previous model.

Adaptive boost classifier

The model performance from training, validation and test set is shown in the table 8.25, 8.26 and 8.27.

	Accuracy	Recall	Precision	F1
0	0.701989	0.715186	0.696795	0.705871

Table 8.25: Undersampled adaptive boost on training set.

	Accuracy	Recall	Precision	F1
0	0.695644	0.700558	0.817558	0.75455

Table 8.26: Undersampled adaptive boost on validation set.

	Accuracy	Recall	Precision	F1
0	0.691523	0.708872	0.805945	0.754298

Table 8.27: Undersampled adaptive boost on test set.

From the table above it can be inferred that the f1 score is low compared to the normal model.

XGboost classifier

The model performance from training, validation and test set is shown in the table 8.28, 8.29 and 8.30.

	Accuracy	Recall	Precision	F1
0	0.863404	0.862123	0.864336	0.863228

Table 8.28: Undersampled xgboost on training set.

	Accuracy	Recall	Precision	F1
0	0.684262	0.681164	0.815623	0.742354

Table 8.29: Undersampled xgboost on validation set.

	Accuracy	Recall	Precision	F1
0	0.681515	0.683901	0.809739	0.741519

Table 8.30: Undersampled xgboost on test set.

From the above table it can be inferred that the model performance is less than the oversampled XGboost model.

9. HYPERPARAMETER TUNING AND FINAL MODEL SELECTION

- Bagging classifier, adaptive boosting classifier and XGBoost classifier uses many weak learning models to reduce overfitting to the model.
- Since their f1 score in training and test sets is high in the range of 0.1 to 0.85 it is speculated that the mode performance could improve with hyperparameter tuning.
- The hyperparameter tuning is done with both grid search and randomized search and the best model is selected by comparing their f1 score.

The hyperparameter tuned models performance for the training, validation and test sets given in table 9.1, 9.2 and 9.3 respectively.

	Bagging classifier with grid search	Bagging classifier with random search	Adaptive boosting classifier with grid search	Adaptive boosting classifier with random search	XGBoost classifier with grid search	XGBoost classifier with random search
Accuracy	0.992111	0.998966	0.778008	0.785353	0.799826	0.804288
Recall	0.998433	0.999902	0.879640	0.849672	0.969053	0.967584
Precision	0.987505	0.998240	0.759064	0.782609	0.746342	0.751559
F1	0.992939	0.999070	0.814916	0.814763	0.843240	0.845999

Table 9.1: Tuned model performances on training set.

	Bagging classifier with grid search	Bagging classifier with random search	Adaptive boosting classifier with grid search	Adaptive boosting classifier with random search	XGBoost classifier with grid search	XGBoost classifier with random search
Accuracy	0.738422	0.733909	0.741170	0.745683	0.725667	0.728022
Recall	0.880694	0.856303	0.880694	0.850132	0.946812	0.940641
Precision	0.763761	0.770696	0.766496	0.786355	0.725839	0.729989
F1	0.818070	0.811247	0.819636	0.817001	0.821729	0.822034

Table 9.2: Tuned model performances on validation set.

	Bagging classifier with grid search	Bagging classifier with random search	Adaptive boosting classifier with grid search	Adaptive boosting classifier with random search	XGBoost classifier with grid search	XGBoost classifier with random search
Accuracy	0.720173	0.727826	0.730573	0.736068	0.724686	0.721546
Recall	0.868096	0.854289	0.874266	0.849001	0.941246	0.933608
Precision	0.751526	0.765465	0.758990	0.776673	0.727025	0.727065
F1	0.805616	0.807441	0.812560	0.811228	0.820382	0.817492

Table 9.3: Tuned model performances on test set.

- From the above performance metrics it is evident that XGboost with gridsearch hypertuning method has the highest f1 score and recall score.
- The performance of the xgboost model has improved significantly after model tuning compared to bagging and adaptive boosting classifier as the performance is similar compared to untuned models.
- So, **XGBoost with gridsearch** hyperparameter tuning is selected as the **final model**.

The confusion matrix for the above-mentioned final model is given in figure 9.1, 9.2 and 9.3 respectively for the training, validation and test sets. It shows that the model can predict up to 57 to 58% of the true positives and 15% of the true negatives respectively.

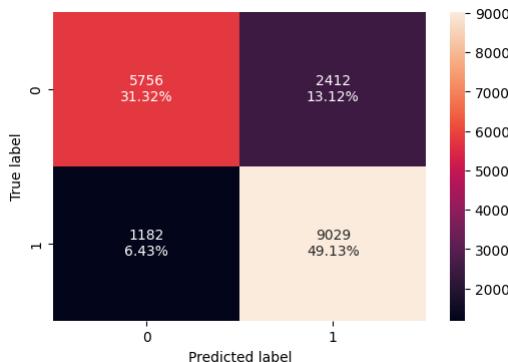


Fig.9.1. Confusion matrix for the final model on training set.

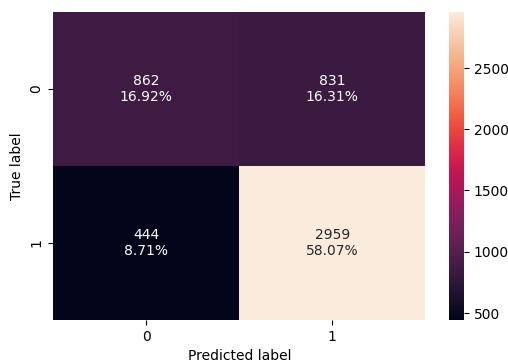


Fig.9.2. Confusion matrix for the final model on validation set.

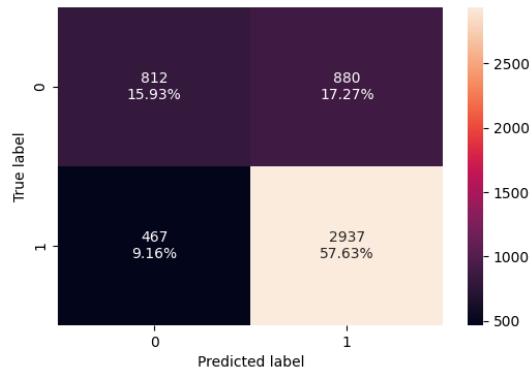


Fig.9.3. Confusion matrix for the final model on test set.

The feature importances of the final model is shown in the figure 9.4. It shows that,

- The employee education of high school has the more importance followed by masters education.
- This shows that education of the employee is the crucial deciding factor for visa certification and rejection followed by northeast employment and job experience.

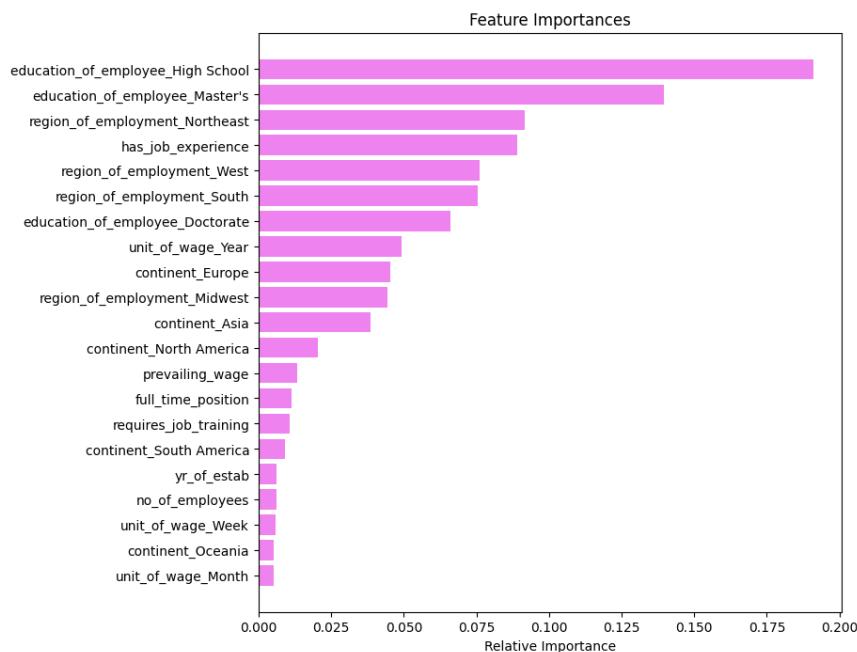


Fig.9.4. Feature importances of the final model.

10. OVERALL CONCLUSIONS AND RECOMMENDATIONS

- The final model has the f1 score of 0.82 which indicates that model can predict 82% of the visa certification and rejection.
- The feature importance shows the employees with at least high school education or master's education can get their visa approved.
- The exploratory data analysis showed that a fulltime employment has a high likelihood for visa approval. So its recommended to get a full-time job before applying for visa preferably in the northeast region.
- It is also recommended to get a experience on the job as job experience has a high likelihood for visa approval and so as the employee doesn't require job training.
- Employees from Asian continent have the highest frequency of job applications. The feature importances and exploratory data analysis also shows that employees from Asia and Europe continent have a high likelihood for visa approval.
- Employees having yearly wages tend to have a high likelihood for visa approval. Its recommended to have a yearly unit wage than hourly or weekly.