MACHINE LEARNING - II
BUSINESS REPORT

Submitted by

Dr. JEMIMAH J P P

BUSINESS REPORT

Contents

List of figures

List of Tables

1. EXPLORATORY DATA ANALYSIS (EDA)

1.1 Context

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

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

OFLC processes job certification applications for employers seeking to bring foreign workers into the United States and grants certifications in those cases where employers can demonstrate that there are not sufficient US workers available to perform the work at wages that meet or exceed the wage paid for the occupation in the area of intended employment.

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 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:

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

1.3 Data description and information

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 immigration programs are administered by the Office of Foreign Labor Certification (OFLC). OFLC has hired the firm EasyVisa for data-driven solutions in order to help in shortlisting the candidates having a higher chance of VISA approval and to 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.

From a data scientist view, the provided dataset can be used to analyse various factors that influence the recommendation of visa approvals and predict suitable profiles to help with a classification model. The information about the different variables mentioned in the data set is elaborated in Table 1.

Information

Predictor Variables	**Description**
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
Target Variable	**Description**
case_status	Flag indicating if the Visa was certified or denied

Table 1: Variables and its description

1.4 Data overview

The necessary packages need to be imported, the working directory is set and the data file is loaded to understand and describe the overview of the provided dataset.

Displaying the first few rows and last few columns of the dataset

The dataset consists of 25480 rows and 12 columns. The 25480 rows represents the case_status of visa applications who apply for US immigrant visas. The 12 columns that give the details on various driving factors are continent, education_of_employee, has_job_experience, requires_job_training, no_of_employees, yr_of_estab, region_of_employment, prevailing_wage, unit_of_wage, full_time_position. These 10 columns drive the target variable, the case_status. The "case_id" column shows the unique identification number given to each visa application and this column has no role to play in the exploratory data analysis and in the model prediction, so is not considered as a driving factor.

Tables 1and 2 show the details of the list of first and last five rows available in the dataset of the EasyVisa data driven solutions provider respectively.

	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	full_time_position	case_status
0	EZYV01	Asia	High School	N	N	14513	2007	West	592.203	Hour	Y	Denied
1	EZYV02	Asia	Master's	Y	N	2412	2002	Northeast	83425.650	Year	Y	Certified
2	EZYV03	Asia	Bachelor's	N	Y	44444	2008	West	122996.860	Year	Y	Denied
3	EZYV04	Asia	Bachelor's	N	N	98	1897	West	83434.030	Year	Y	Denied
4	EZYV05	Africa	Master's	Y	N	1082	2005	South	149907.390	Year	Y	Certified

Table 2: Top five rows of the dataset

	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	full_time_position	case_status
25475	EZYV25476	Asia	Bachelor's	Y	Y	2601	2008	South	77092.570	Year	Y	Certified
25476	EZYV25477	Asia	High School	Y	N	3274	2006	Northeast	279174.790	Year	Y	Certified
25477	EZYV25478	Asia	Master's	Y	N	1121	1910	South	146298.850	Year	N	Certified
25478	EZYV25479	Asia	Master's	Y	Y	1918	1887	West	86154.770	Year	Y	Certified
25479	EZYV25480	Asia	Bachelor's	Y	N	3195	1960	Midwest	70876.910	Year	Y	Certified

Table 3: Bottom five rows of the dataset

<u>Checking the data types of the columns for the dataset</u>

The dataset consists of 3 numerical columns and 9 object type columns. The no_of_employees, yr_of_estab, prevailing_wage are the numerical columns of the dataset.

The continent, education_of_employee, has_job_experience, requires_job_training,, region_of_employment, unit_of_wage, full_time_position, case_id and case_status are the object type columns in the dataset. The case_status columns describe the details if the visa applications have been "Certified" or "Denied" and hence can be encoded as "1" and "0" respectively. From the information obtained it is observed that there is no missing values in the dataset.

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

Table 4: Information about the columns of the dataset

Observations

- The dataset consists of 3 numerical columns and 9 object type columns.
- The `no_of_employees` , `yr_of_estab` and `prevailing_wage` are the numerical columns of the dataset.
- The `case_id`,`continent`, `education_of_employee`, `has_job_experience`, `requires_job_training`, `region_of_employment`, `unit_of_wage`, `full_time_position` and `case_status ` are the object type columns in the dataset.
- The columns in the dataset are in the preferred format as per the respective values
- From the information obtained it is observed that there is no missing values in the dataset.

Checking for missing values

The table 5 shows that the provided dataset does not contain any missing values.

```
case_id                  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
dtype: int64
```

Table 5: Checking for missing values

Checking for duplicate values

- It is also observed that there are no duplicate entries in the dataset.

Checking the number of distinct values in the dataset

```
case_id                  25480
continent                    6
education_of_employee        4
has_job_experience           2
requires_job_training        2
no_of_employees           7105
yr_of_estab                199
region_of_employment         5
prevailing_wage          25454
unit_of_wage                 4
full_time_position           2
case_status                  2
dtype: int64
```

Table 6: Unique values in the dataset

- Among the variables in the dataset, no_of_employees, yr_of_estab and prevailing_wage have the highest counts of unique values.
- It is noted that the column `case_id` has `25480` unique values and this column does not play any significant role in analysis, hence can be removed

Statistical summary of the numerical columns of the dataset

	count	mean	std	min	25%	50%	75%	max
no_of_employees	25480.000	5667.043	22877.929	-26.000	1022.000	2109.000	3504.000	602069.000
yr_of_estab	25480.000	1979.410	42.367	1800.000	1976.000	1997.000	2005.000	2016.000
prevailing_wage	25480.000	74455.815	52815.942	2.137	34015.480	70308.210	107735.513	319210.270

Table 7: Description of the numerical columns of the dataset

The table 7 shows the statistical summary of the numerical columns present in the data set

Observations

- From the statistical summary of the numerical columns, it can be seen that some employers have as much as `602,069 employees` working in their firm.
- It is to be noted that the minimum no. of employees is `-26`and this clearly shows that it is an error, as the no. of employees cannot be of negative value.
- The average no. of employees (mean) is much larger than its median (50%) value, indicating a longer right tailed (positively skewed) distribution. It can also be seen that there is presence of outliers in this column
- The mean year of establishment is 1979, while the median is 1997 which indicates that the distribution of `yr_of_estab` is left skewed. It is noted that the minimum value is 1800 and the maximum value is 2016, showing that the dataset has details about firms established from 1800 to 2016, where 1800 being the oldest firm in the dataset.
- The prevailing wage ranges from a minimum value of USD 2.137 to a maximum of USD 319,210.270, which shows there is a huge gap in between and the reason can be analysed further.
- The mean `prevailing_wage` is USD 74,455.815 which is higher that it's median, which is USD 70,308.210, indicating a positively skewed distribution.

The distribution of these predictor variables can be best understood using a box plot and histograms. Their impact against the target variable is also visualized using the same and against the categorical variables are analysed using bar plots etc.

<u>Statistical summary of the categorical/ object columns of the dataset</u>

	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 8: Description of the categorical columns of the dataset

<u>Checking for anomalous values in categorical variables</u>

The unique values are determined for each categorical variable to check if any junk/garbage values present in the dataset. This check helps us to identify if any data entry issues are present. From the determined unique values it's concluded that there is no data entry issues present.

```
case_id
EZYV01       1
EZYV16995    1
EZYV16993    1
EZYV16992    1
EZYV16991    1
            ..
EZYV8492     1
EZYV8491     1
EZYV8490     1
EZYV8489     1
EZYV25480    1
Name: count, Length: 25480, dtype: int64
--------------------------------------------------
continent
Asia            16861
Europe           3732
North America    3292
South America     852
Africa            551
Oceania           192
Name: count, dtype: int64
--------------------------------------------------
education_of_employee
Bachelor's     10234
Master's        9634
High School     3420
Doctorate       2192
Name: count, dtype: int64
--------------------------------------------------
has_job_experience
Y    14802
N    10678
Name: count, dtype: int64
--------------------------------------------------
requires_job_training
N    22525
Y     2955
Name: count, dtype: int64
--------------------------------------------------
region_of_employment
Northeast    7195
South        7017
West         6586
Midwest      4307
Island        375
Name: count, dtype: int64
--------------------------------------------------
unit_of_wage
Year     22962
Hour      2157
Week       272
Month       89
Name: count, dtype: int64
--------------------------------------------------
full_time_position
Y    22773
N     2707
Name: count, dtype: int64
--------------------------------------------------
--------------------------------------------------
case_status
Certified    17018
Denied        8462
Name: count, dtype: int64
--------------------------------------------------
```

Table 9: Value counts of the categorical variables of the dataset

Observations

- **continent:** The dataset shows that there are applicants from `6 continents` throughout the globe, of which `16861` visa applications are from Asia.
- **education_of_employee:** The education level of the applicants were mentioned in `4 levels`, of which most of the applicants nearly `10234` have completed their Bachelor's degree
- **has_job_experience:** A high number of applicants say `14802` were seen to have a previous job experience and almost `10678` applications were from freshers.
- **requires_job_training:** A mojory of the applicants say `22525` for visa application did not require `job_training`
- **region_of_employment:** There are `5 different region` where the applicants were employed, of which fewer applications (`375`) were from applicants employed in an`Island`
- **unit_of_wage:** The `unit of wage` of the employees is categorized under `4 divisions`, `Year`, `Hour`,`Week`, and `Month`. This is the reason behind the huge variation in `Mean`,`Median`,`Minimum` and `Maximum` values in this column.
- **full_time_position:** It is also seen that most of the applicants were from applicants (`22773`) who were employed in full_time
- **case_status:** The `case_status` is the `target` of this analysis, and it can be seen that most of the applications were `Certified` with visa approval and the `Denied` count is less. Here the category `Certified` is encoded as “1” and the category `Denied` is encoded as “0” and proceeded with univariate and bivariate analysis.

It is also noted that there is no anomalous values in these categorical variables. It is also seen that the `"case_id"` column can be dropped before proceeding with model building.

1.5 Univariate analysis

The univariate analysis is carried out to explore all the variables and their distributions are observed. Generally, histograms, boxplots, countplots, etc. are used for univariate exploration. The categorical variables are explored using labelled_barplots and the numerical variables are explored using histograms and boxplots respectively.

Numerical variables

- no_of_employees
- yr_of_estab
- prevailing_wage

no_of_employees

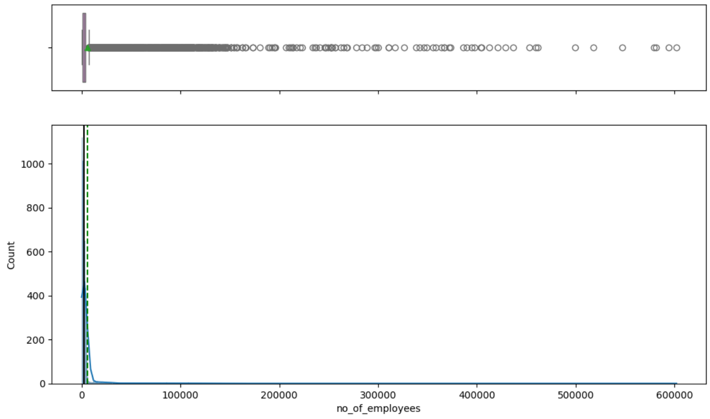


Figure 1: Histogram and Box plot for no_of_employees

Observation:

- Here, it is seen that the distribution is right skewed and has a large number of outliers.
- From the data description we see that the applications are from applicants who are employed in various firms, which were established from 1800 to 2016.
- So, there is a possibility of having more number of employees by the firms which were established long before than the companies that were established recently.

yr_of_estab

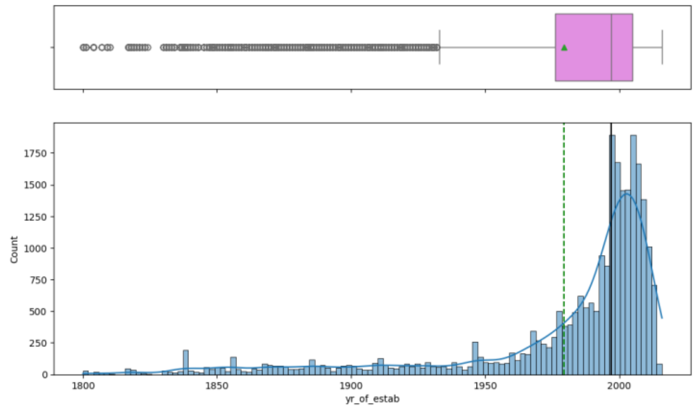


Figure 2: Histogram and Box plot for yr_of_estab

Observation:

- Here, it is seen that the distribution is left skewed and has a large number of outliers.
- The mean year of establishment is 1979, while the median is 1997 which indicates that the distribution of yr_of_estab is left skewed. It is noted that the minimum value is 1800 and the maximum value is 2016, showing that the dataset has details

about firms established from 1800 to 2016, where 1800 being the oldest firm in the dataset.
- It is observed that there are high number of applications from applicants who are employed in companies established by 2000 and later.

prevailing_wage

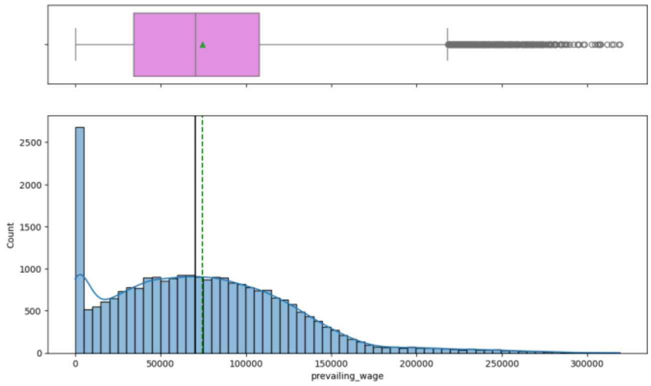


Figure 3: Histogram and Box plot for prevailing_wage

Observation:

- The distribution is right skewed with the mean slightly larger than the median
- To understand the variation on values in this distribution, it is essential to focus on `unit_of_wage`. It is also seen that it is under `4 categories` such as `Year`, `Hour`,`Week`, and `Month`. Hence data falling above larger right tail cannot be considered as outliers

Distribution of numeric variables in the dataset

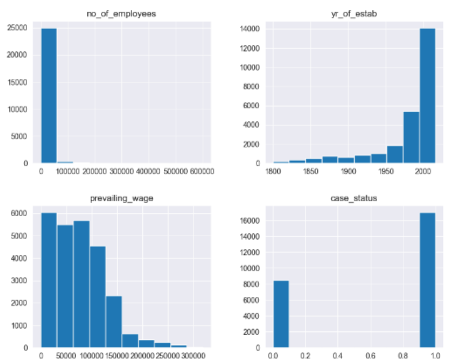


Figure 4: Visual representation of numerical data in tabular form

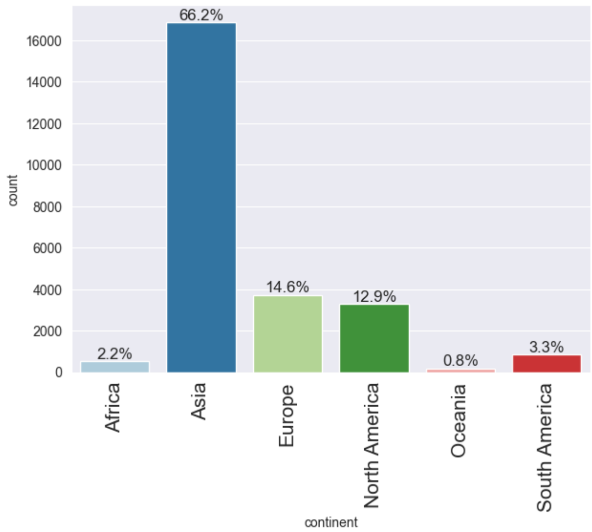
Observation:

- The employer's company profile of the visa applicants is shown such that majority of the companies have less than 1 lakh employees working in their firms.
- The firms in this dataset have been established from 1800 to above 2000.
- There is a high dense of applications with their prevailing wage less than 1 lakh USD
- We see that the certified visas outnumbers the denials

Categorical variables

- continent
- education_of_employee
- has_job_experience
- requires_job_training
- region_of_employment
- unit_of_wage
- full_time_position
- case_status

continent



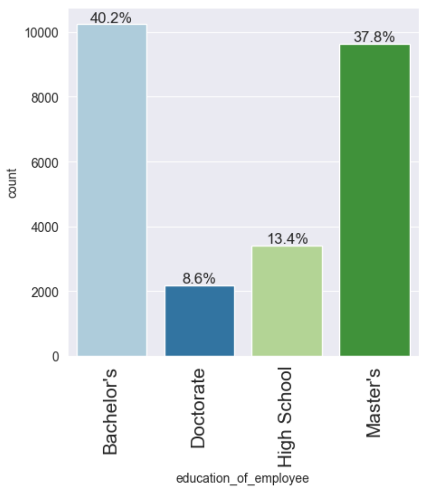

```
continent
Asia             16861
Europe            3732
North America     3292
South America      852
Africa             551
Oceania            192
Name: count, dtype: int64
```

Figure 5: Labelled bar plot and value counts for continent

Observation:

- Most of the employees belong to Asian continent
- Very less no. of applicants (≈0.8%) are from Oceania

education_of_employee



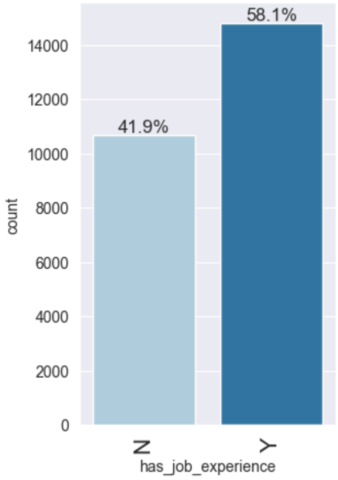

```
education_of_employee
Bachelor's     10234
Master's        9634
High School     3420
Doctorate       2192
Name: count, dtype: int64
```

Figure 6: Labelled bar plot and value counts for education_of_employee

Observation:

- Most of the employees have completed bachelors degree, the percentage of employees who have a masters degree is slightly lower than the bachelor's
- Very few employees have doctorate degrees

has_job_experience



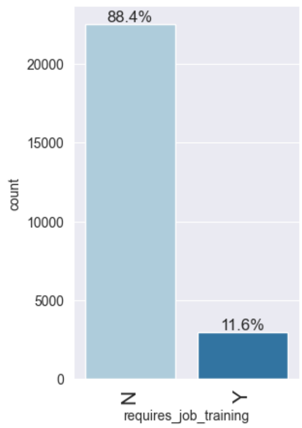

```
has_job_experience
Y     14802
N     10678
Name: count, dtype: int64
```

Figure 7: Labelled bar plot and value counts for has_job_experience

Observation:

- Nearly 58.1% of visa applications have a previous job experience while 41.9% were without an experience.

requires_job_training



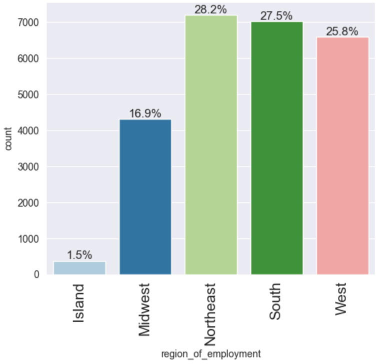

```
requires_job_training
N     22525
Y      2955
Name: count, dtype: int64
```

Figure 8: Labelled bar plot and value counts for requires_job_training

Observation:

- Almost 88.4% of the visa applicants do not require a job_training while a least percentage of about 11.6% require training for their profession.

region_of_employment



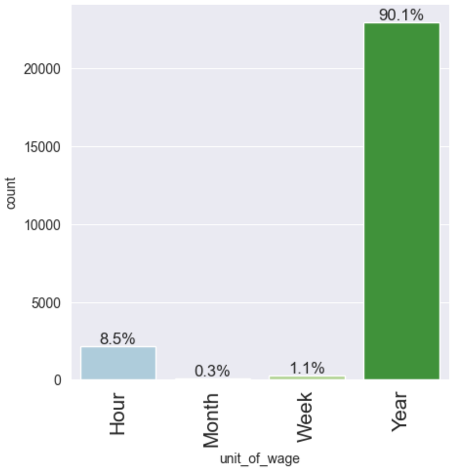

```
region_of_employment
Northeast    7195
South        7017
West         6586
Midwest      4307
Island        375
Name: count, dtype: int64
```

Figure 9: Labelled bar plot and value counts for region_of_employment

Observation:

- There are only 1.5% of visa applications from employees placed in the island region being the the lowest.
- The visa applications from employees working in the Northeast region tops with 28.2% followed by South and West regions

unit_of_wage



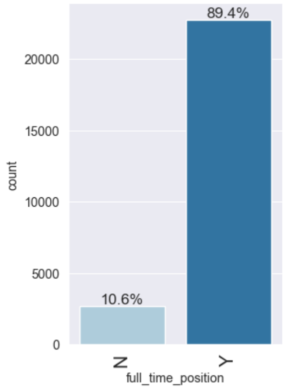

```
unit_of_wage
Year     22962
Hour      2157
Week       272
Month       89
Name: count, dtype: int64
```

Figure 10: Labelled bar plot and value counts for unit_of_wage

Observation:

- The most used unit of wage is the `Year` unit, and this explains the right skewed distribution of the prevailing wage distribution.
- Only 8.5% of the prevailing_wage is mentioned with `Hour` units, where the `Month` and `Week` units are minimally mentioned

full_time_position



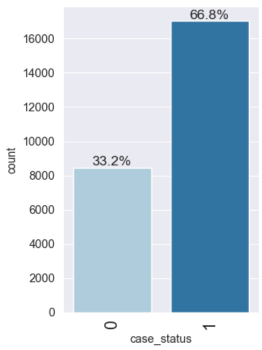

```
full_time_position
Y    22773
N     2707
Name: count, dtype: int64
```

Figure 11: Labelled bar plot and value counts for full_time_position

Observation:

- The employees who are working in `full_time_position` are the major ones to apply for visa, whereas part_time employees are the least (≈10.6%)

case_status




```
case_status
1    17018
0     8462
Name: count, dtype: int64
```

Figure 12: Labelled bar plot and value counts for case_status

Observation:

- The majority of the applicants (≈66.8%) were `Certified` with visa
- While less no. of applicants (≈33.2%) were `Denied`.

1.6 Bivariate analysis

Let's see the attributes that have a strong correlation with each other

Correlation between numerical variables

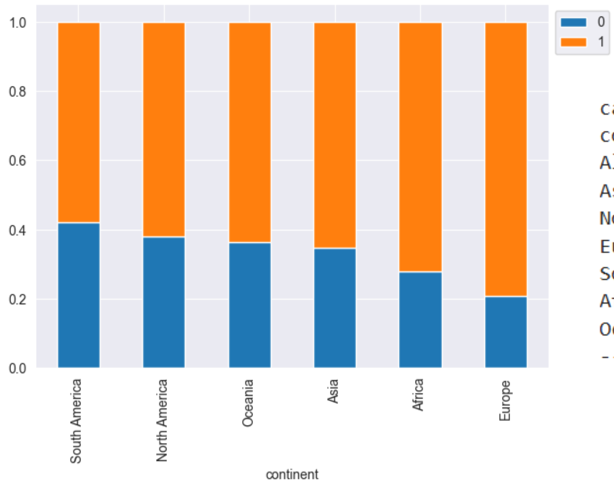
Figure 13: Heatmap between numerical variables

Observation:

- It is understood that there is no much correlation between the numerical columns of the dataset.

Relationship between categorical variables vs target variable-case_status

continent vs case_status



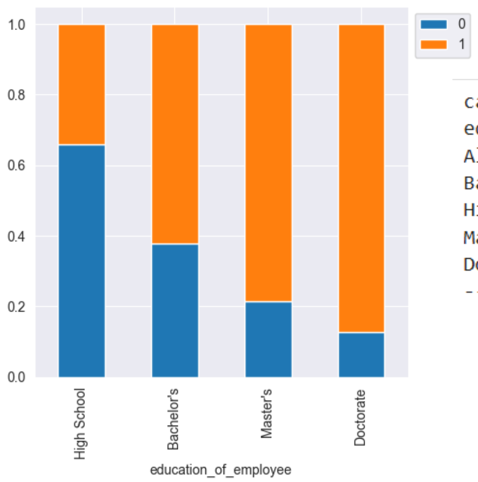
case_status continent	0	1	All
All	8462	17018	25480
Asia	5849	11012	16861
North America	1255	2037	3292
Europe	775	2957	3732
South America	359	493	852
Africa	154	397	551
Oceania	70	122	192

Figure 14: Stacked bar plot for continent vs case_status

Observation:

- Nearly 79.2% of applications from Europe have received the certificate and tops all the continents with only 20.7% of denials
- Followed by Africa having approximately 72.1% of total certified applications with 27.9% of denials
- Then comes Asia with approximately 65.3% accepted applications with 34.7 % of denials and has the majority of applicants from all over the globe.

education_of_employee vs case_status



case_status education_of_employee	0	1	All
All	8462	17018	25480
Bachelor's	3867	6367	10234
High School	2256	1164	3420
Master's	2059	7575	9634
Doctorate	280	1912	2192

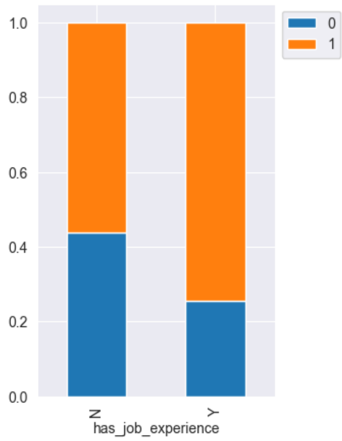
Figure 15: Stacked bar plot for education_of_employee vs case_status

Observation:

- Nearly 87.2% of doctorate holder applications were certified while ≈12.7% were denied.

- Followed by Master degree holders having approximately 78.6% of certified applications with only 21.4% of denials
- Then comes Bachelor degree holders with approximately 62.2% accepted applications with ≈37.7% of denials and has around 10234 of the total applications from all the educational background

<u>has_job_experience vs case_status</u>



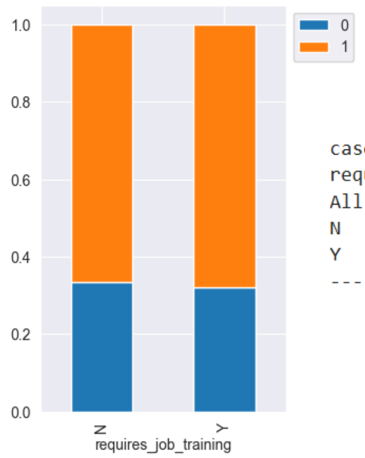
case_status has_job_experience	0	1	All
All	8462	17018	25480
N	4684	5994	10678
Y	3778	11024	14802

Figure 16: Stacked bar plot for has_job_experience vs case_status

Observation:

- Nearly 74.5% of applicants who has previous job experience were certified while ≈25.5% were denied, and has a majority of applications of about 14802.
- Approximately 56.13% of applicants who does not have previous working experience were certified while ≈43.9% were denied, but has less no. of applications around 10678 compared to the other category

<u>requires_job_training vs case_status</u>



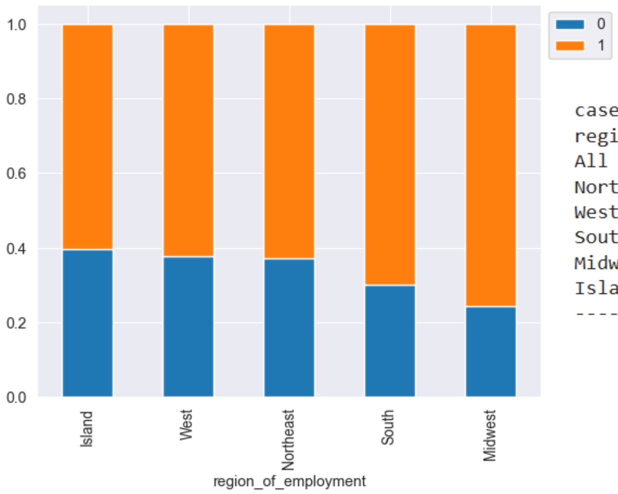
case_status requires_job_training	0	1	All
All	8462	17018	25480
N	7513	15012	22525
Y	949	2006	2955

Figure 17: Stacked bar plot for requires_job_training vs case_status

Observation:

- Nearly 66.6% of applicants who do not require a job training were certified while ≈33.3% were denied, and has a majority of applications of about 22525.
- Approximately 67.8% of applicants who require a job training were certified while ≈32.1% were denied, but has very less no. of applications around 2955

region_of_employment vs case_status



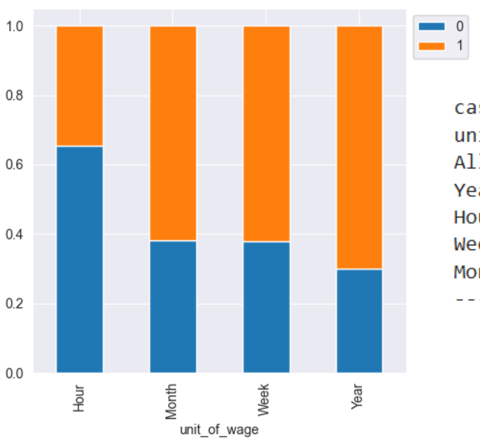
case_status	0	1	All
region_of_employment			
All	8462	17018	25480
Northeast	2669	4526	7195
West	2486	4100	6586
South	2104	4913	7017
Midwest	1054	3253	4307
Island	149	226	375

Figure 18: Stacked bar plot for region_of_employment vs case_status

Observation:

- The applications from employees who were appointed at midwest region have higher rate of visas certified, which is about 75.5% and lesser denial rates of about 24.5%
- Followed by applicants employed at south, with 70% of certifications and 30% of denials.
- Island seems to have the highest denials with 39.7% with lowest acceptance of 60.3%
- Employments at Northeast and West almost have similar acceptance rates such as 62.9% and 62.2% respectively.

unit_of_wage vs case_status



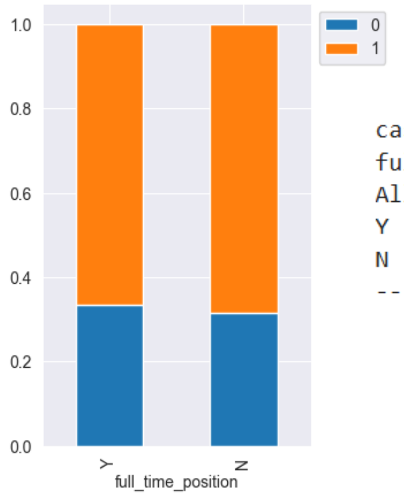
case_status	0	1	All
unit_of_wage			
All	8462	17018	25480
Year	6915	16047	22962
Hour	1410	747	2157
Week	103	169	272
Month	34	55	89

Figure 19: Stacked bar plot for unit_of_wage vs case_status

Observation:

- It is noted that applicants who receive their wages in yearly basis have higher visa certifications rate (69.8%) and 30.1% of denials
- Where as applicants whose wages are determined on monthly and weekly basis were said to have almost similar acceptance rates of about 61.8% and 62.1% respectively.
- And the applicants whose wages are of hourly basis face the highest denial rates of about 65.3%

full_time_position vs case_status



case_status full_time_position	0	1	All
All	8462	17018	25480
Y	7610	15163	22773
N	852	1855	2707

Figure 20: Stacked bar plot for full_time_position vs case_status

Observation:

- It seems that the employees positioned in full time face higher denial rates compared to part time employees.
- The acceptance rate for full time employees is 66.6% and for the other category is 68.5%, while still majority of the visa applications come from employers who hire employees for full time which is about 22773 (≈89.3%) of total applications

Relationship between numerical variables vs target variable-case_status
Distribution plot on prevailing_wage Vs case_status

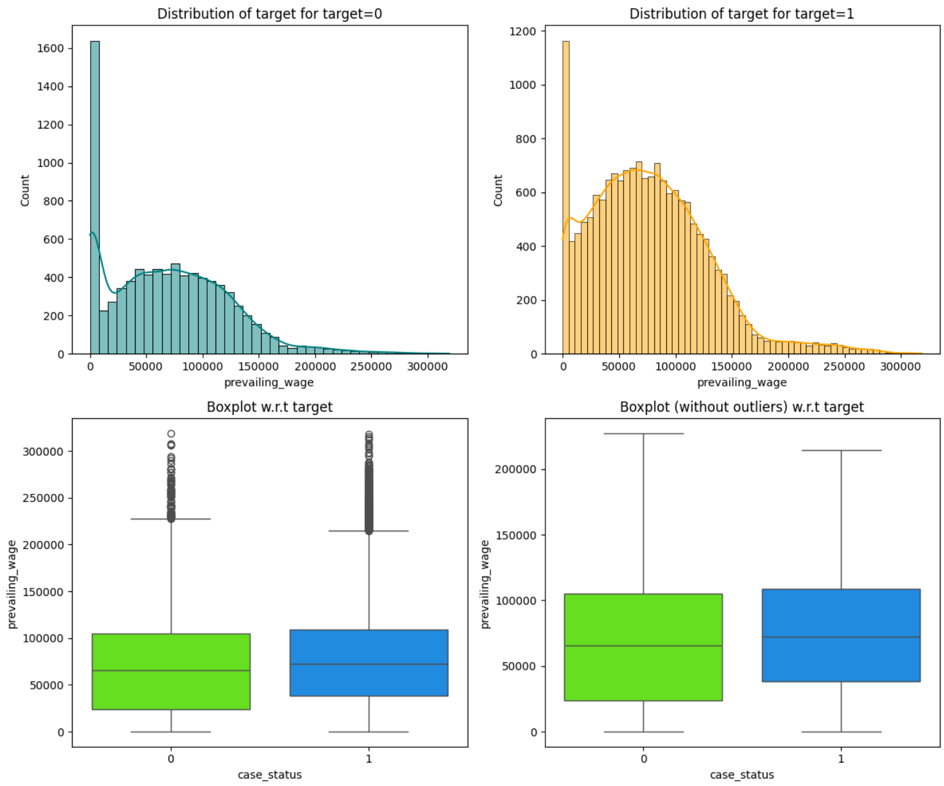


Figure 21: Distribution plot on prevailing_wage Vs case_status

Observation:

- We can observe similar densities in both the case_status and prevailing wage of the employees.
- We can also see,a significantly higher density in the denial for prevailing wage between 0 and 100
- The boxplot of prevailing wage with respect to case_status shows that the median prevailing wage of the employees who are certified is slightly higher.
- Observing the boxplot without outliers, it is seen that the maximum prevailing wage of employees which were denied is higher than those that were certified.

Distribution plot on no_of_employees Vs case_status

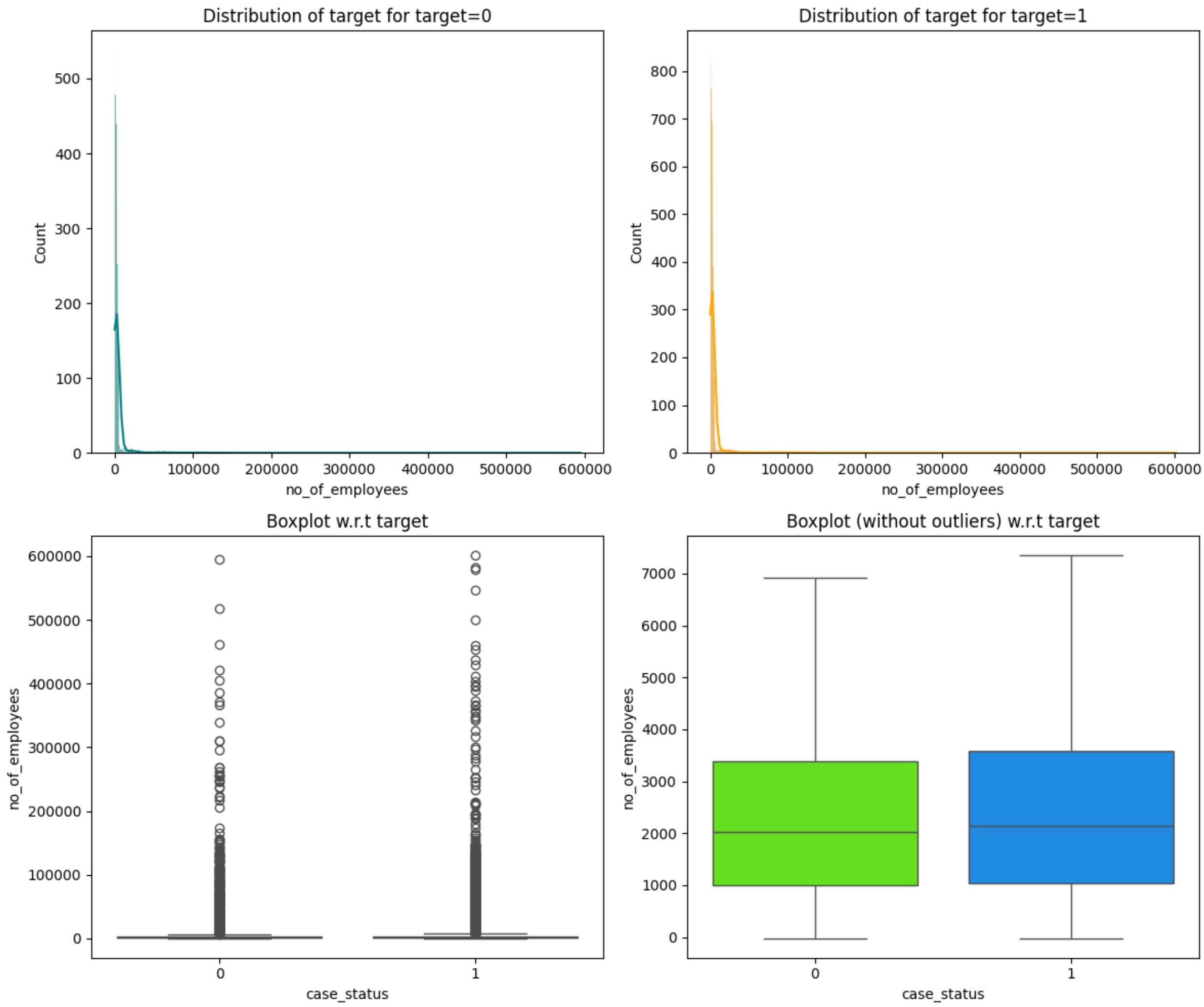


Figure 22: Distribution plot on no_of_employees Vs case_status

Observation:

- From the density plot and box plots, the applicant's employer company whose visa were certified have slightly higher number of employees than the applicant's employer company that were denied.
- It is evident that company with higher no. of employees working has the feasibility to get the acceptance of their employee's visa applications

Distribution plot on yr_of_estab Vs case_status

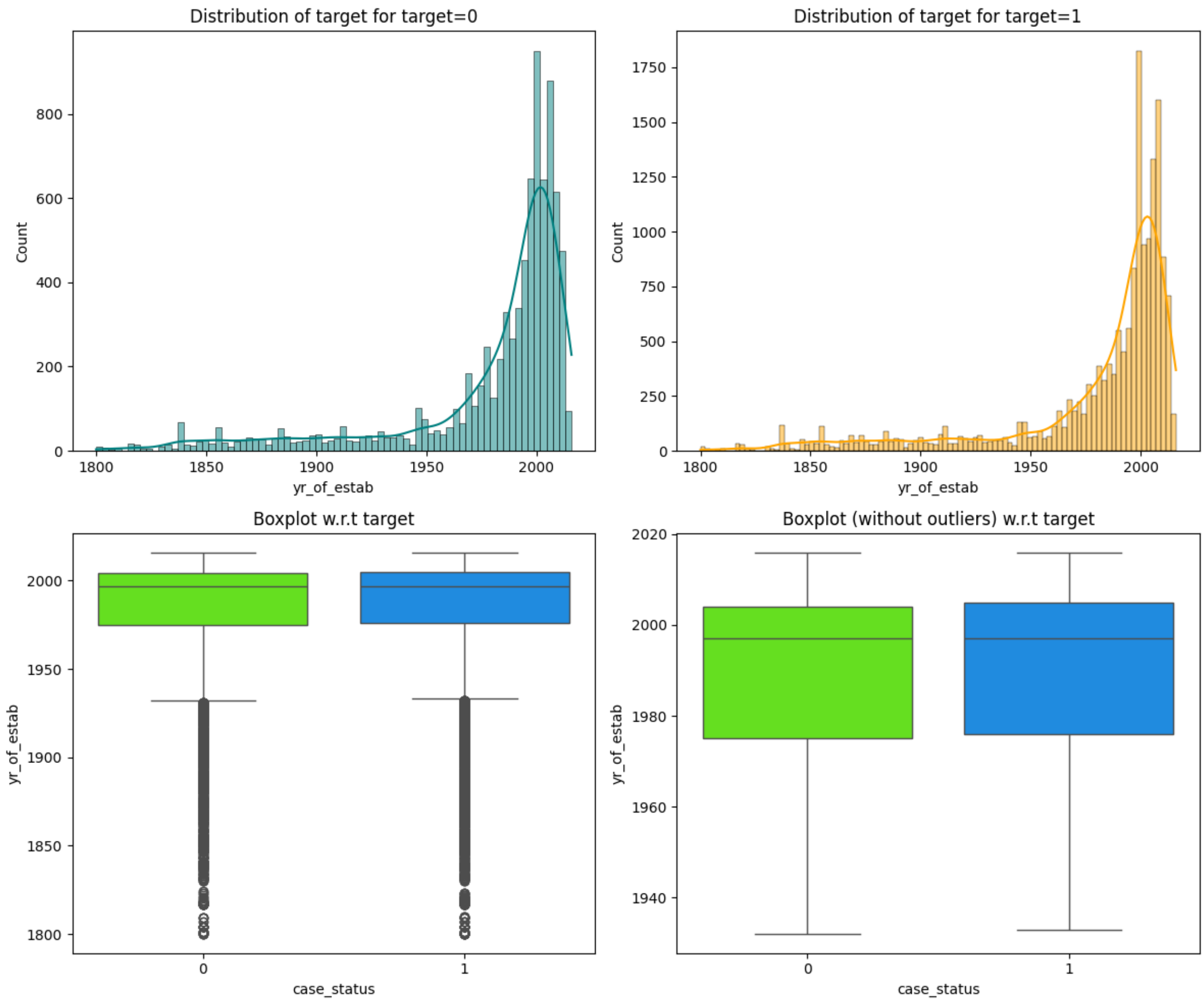


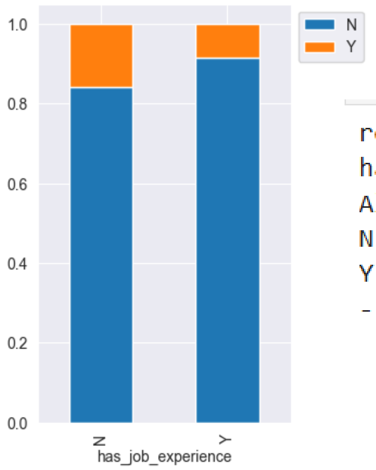
Figure 23: Distribution plot on yr_of_estab Vs case_status

Observation:

- From the density plots, we can see that there is no significant difference between the densities of the certified and denied applications with respect to their year of establishment
- From the box plot, with and without outliers, the same observations is made.
- There is a slight difference between the year of establishment of the companies that had certified applications and those that had denials

Let's now try to find out some relationship between the other columns

has_job_experience Vs requires_job_training



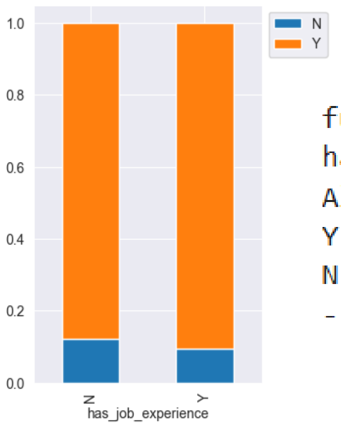
requires_job_training	N	Y	All
has_job_experience			
All	22525	2955	25480
N	8988	1690	10678
Y	13537	1265	14802

Figure 24: Stacked bar plot for has_job_experience Vs requires_job_training

Observation:

- Most employees who have job experience did not require job training, that is about 91.4% of total employees who has job experience did not require training, while 8.5% of employees who had previous experience required training

has_job_experience Vs full_time_position



full_time_position	N	Y	All
has_job_experience			
All	2707	22773	25480
Y	1410	13392	14802
N	1297	9381	10678

Figure 25: Stacked bar plot for has_job_experience Vs full_time_position

Observation:

- Most of the employees who have a previous job experience were offered full time positions, about 58.8% of full time positions were offered to employees with prior job experience, while 41.2% of full time workers did not have the previous job experience
- Even for non full time positions, employees with prior experience were opted for about 52.1%.
- It is seen that, people with prior experience were preferred most by the employers in either of the positions.

requires_job_training Vs prevailing_wage with hue=case_status

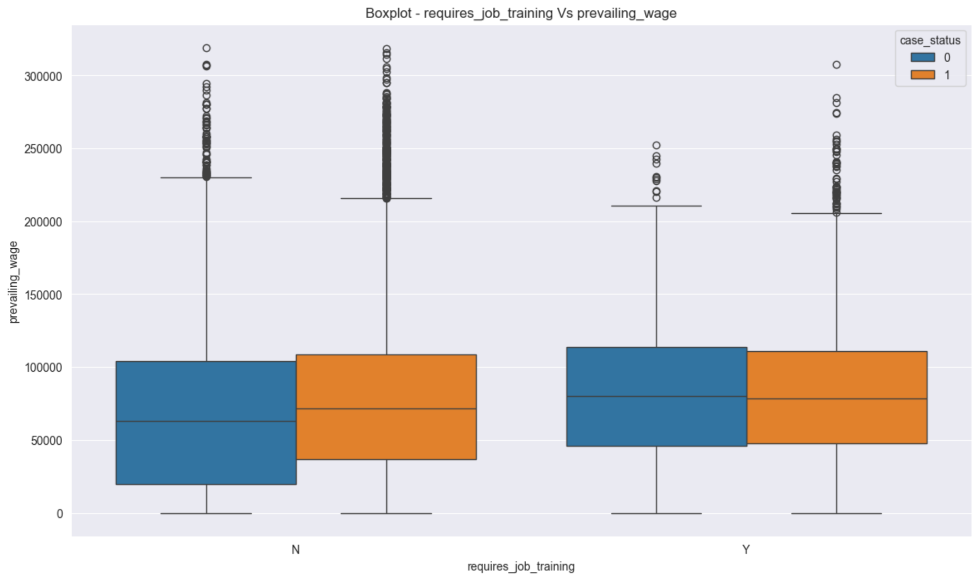


Figure 26: Box plot for requires_job_training Vs prevailing_wage with hue=case_status

Observation:

- The 75th percentile of prevailing wage for a certified visa applicant who does not require job training is almost equal to the certified visa applicant does require job training.
- The prevailing wages for "no_job_training" category is slightly more but their visa certifications in both cases looks feasible.
- Hence prevailing wage with respect to job training requirement does not seem to influence visa certifications

has_job_experience Vs prevailing_wage with hue=case_status

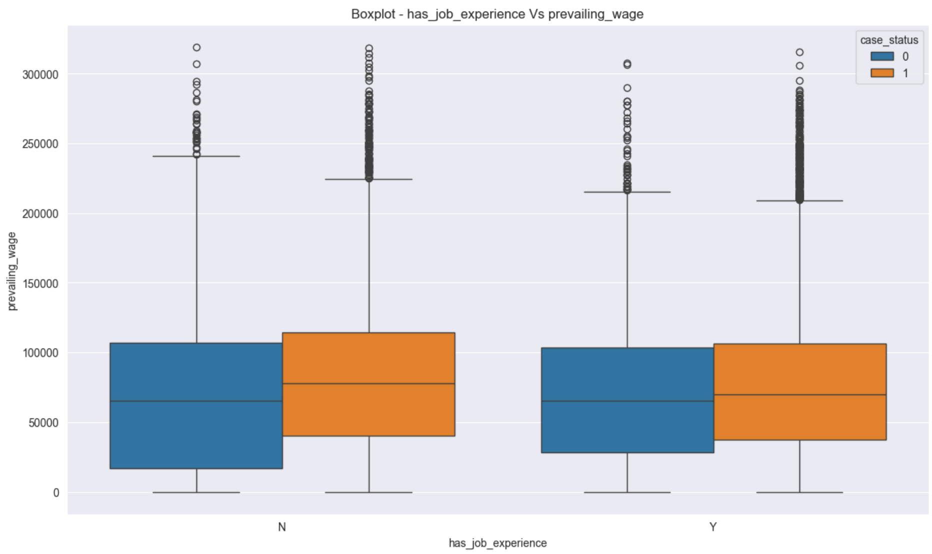


Figure 27: Box plot for has_job_experience Vs prevailing_wage with hue=case_status

Observation:

- The 75th percentile of prevailing wage for a certified visa applicant who does not have a previous job experience is lower to the certified visa applicant who does not have a previous experience.
- The 75th percentile of prevailing wage for visa certified employees is higher compared to the prevailing age of visa denied employees.
- The maximum prevailing wages for "no_experience" category is slightly more but their visa certifications in both cases looks feasible.
- Hence prevailing wage with respect to previous_job_experience does not seem to influence visa certifications

continent Vs prevailing_wage with hue=case_status

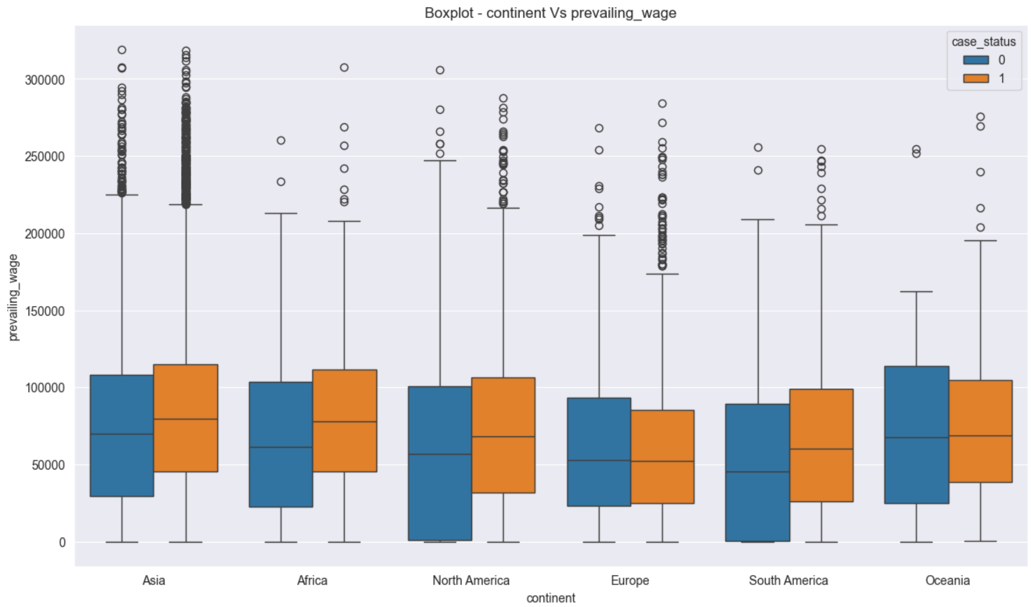


Figure 28: Box plot for continent Vs prevailing_wage with hue=case_status

Observation:

- The 75th percentile of prevailing_wage for Asian visa certified applicants looks seemingly high compared to all other continents while the prevailing wages for Europeans seems to be the lowest
- For most of the continents with respect to the prevailing wage, the acceptance of visas outnumbers its denials, except for Europe and Oceania
- For Oceania and Europe, the mean prevailing wage for visa accepted and the denied applications remains the same., while for other continents the mean prevailing wage of the accepted applications is high compared to the respective denials of that continent.
- Though the prevailing wages vary with respect to continents, it does not widely influence the visa certifications.

region_of_employment Vs prevailing_wage with hue=case_status

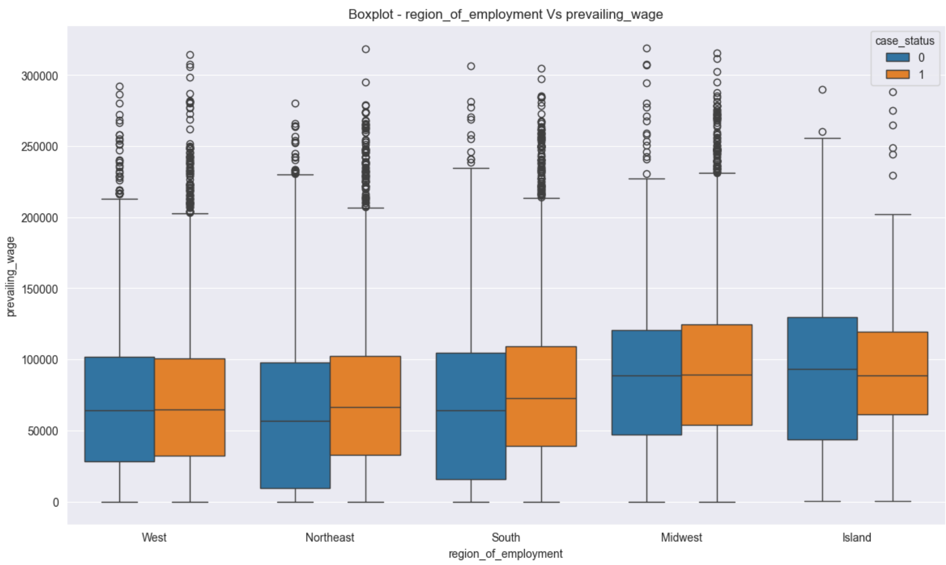


Figure 29: Box plot for region_of_employment Vs prevailing_wage with hue=case_status

Observation:

- The mean prevailing_wage for employees employed at `Midwest` and `Island` regions looks high compared the rest other regions.
- While the mean prevailing wage for employees at the west is the lowest.
- Looking into the outliers of the prevailing wage from all the regions of employment, it is noted that it is higher for visa certified employees.
- Though the prevailing wages vary with respect to regions, it does not influence the visa certifications.

education_of_employee Vs no_of_employees with hue=case_status

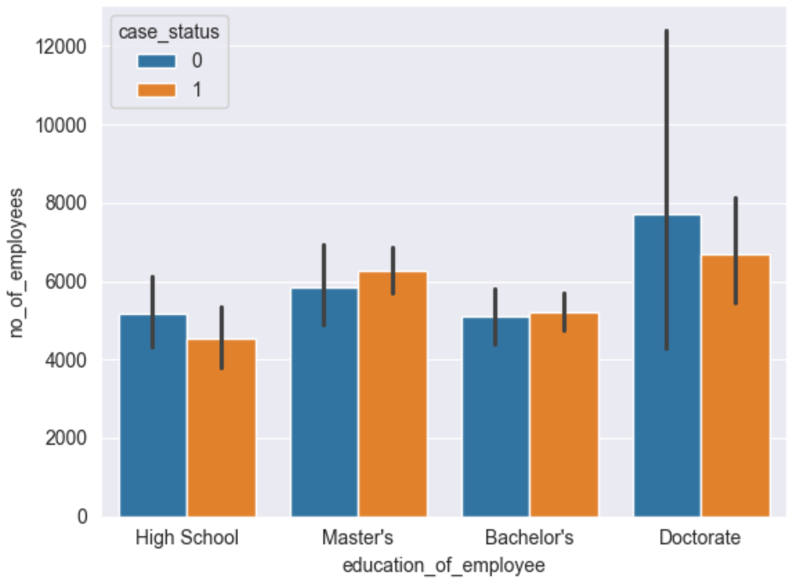


Figure 30: Bar plot for education_of_employee Vs no_of_employees with hue=case_status

Observation:

- It is seen that the doctorates are the highest among the accepted visa applicants, followed by Masters, and the maximum no_of_employees in the employers company are also doctorates
- Similarly the maximum no. of applicants who were denied were also doctorates
- The no_of_employees with Masters, has a quite high acceptance than any other education level and next to Doctorates, master degree for employees were much preferred by the employers.

<u>Let's analyze the education_of_employee of the visa applicants from different continents</u>

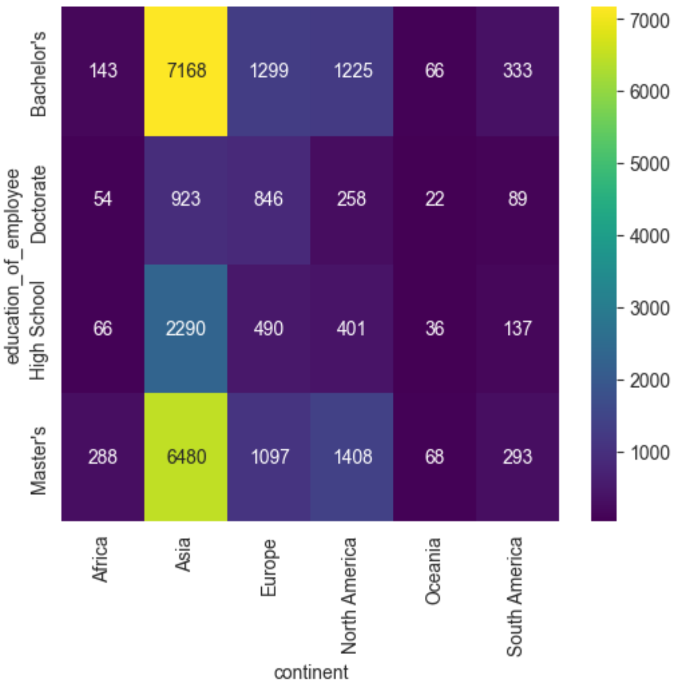


Figure 31: Heatmap on education_of_employee Vs continent

Observation:

- Nearly 6480 visa applications with Masters degree holders are from Asian applicants followed by 1408 applications from North Americans
- Similarly 7168 visa applications with Bachelors degree holders are from Asian applicants followed by 1299 applications from Europeans
- While there are about 293 and 288 Master's from South America and Africa respectively.
- Coming to Doctorates, Europe (846) is just less to Asia (923)

Let's analyze the education_of_employee of the visa applicants placed in different region_of_employment in US

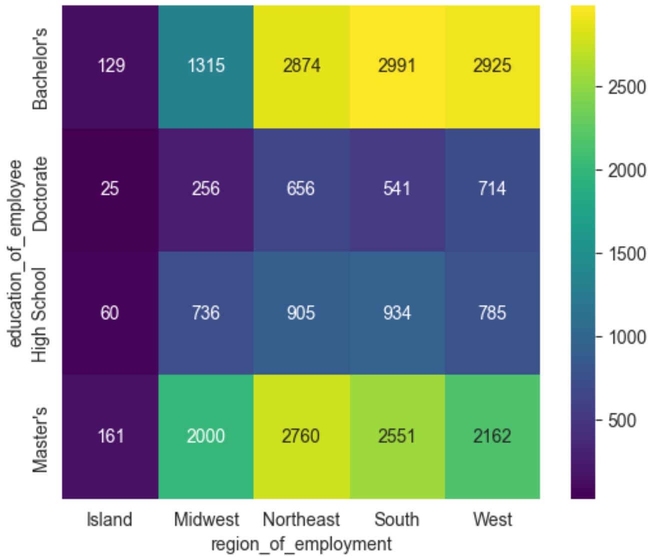


Figure 32: Heatmap on education_of_employee Vs region_of_employment

Observation:

- Master degree holders were placed widely in different in different regions in US, while majority (2760) were placed in Northeast
- Bachelors were also equally preferred with their placing higher than Masters in Northeast (2874), South (2991) and West (2925) while Midwest (2000) and Island (161) preferred Master's

region_of_employment Vs no_of_employees with hue=case_status

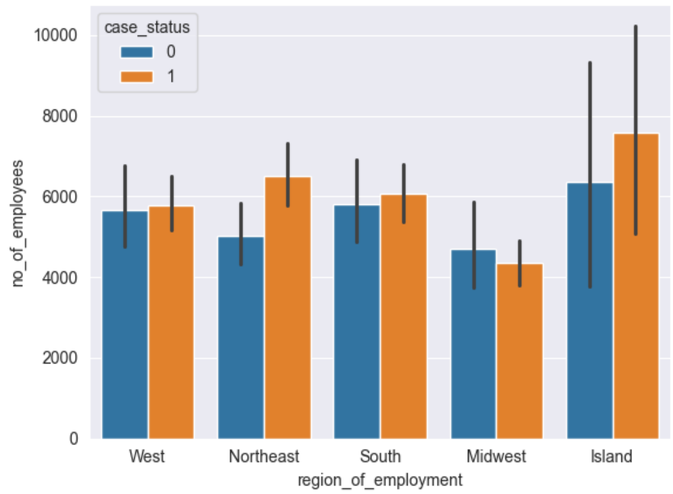


Figure 33: Bar plot for region_of_employment Vs no_of_employees with hue=case_status

Observation:

- The no_of employees with accepted visa applications were high compared to denials in almost all regions like `West`, `Northeast`, `South` and `Island`, except for `Midwest`, where the denials are higher than acceptance.

- Most of employees approximately higher than 10k were placed in `Islands`

continent Vs no_of_employees with hue=case_status

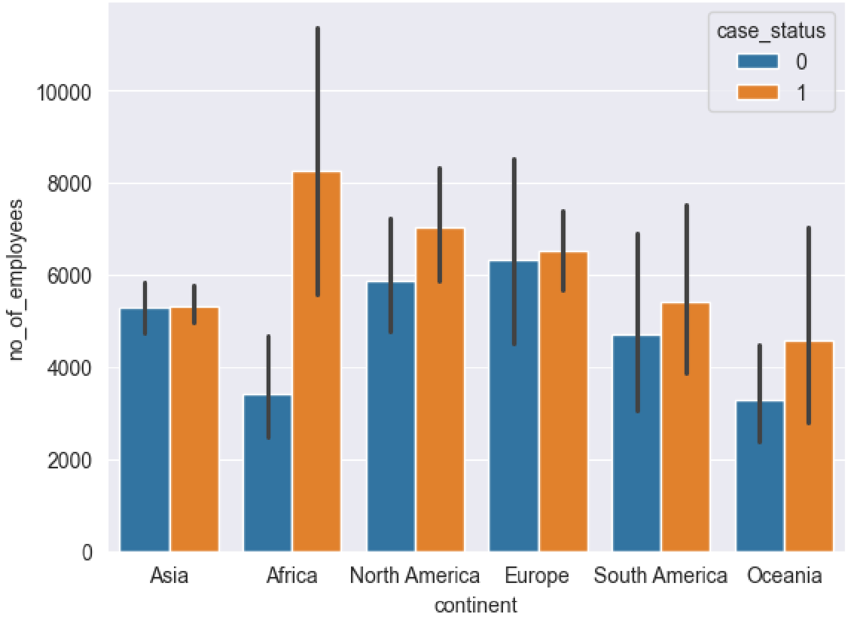


Figure 34: Bar plot for continent Vs no_of_employees with hue=case_status

Observation:

- Most of the employees in company's profile, whose visas were certified were Africans, followed by North America, Europe, Asia, South America and Oceania.
- It is seen that the certified visas outnumber the denials for employees from all over the globe

1.7 Insights based on EDA

- It is observed that the visa applications are from applicants who are employed in various firms, which were established from `1800` to `2016`.
- It is observed that there are high number of applications from applicants who are employed in companies established by 2000 and later, and it is also observed from the density plots, that there is no significant difference between the densities of the certified and denied applications with respect to their year of establishment
- The prevailing_wage falls under 4 categories such as `Year`, `Hour`, `Week`, and `Month`. Hence all the observations are found to be meaningful and the prevailing wage with respect to case_status shows that the median prevailing wage of the employees who are certified is slightly higher.
- It is observed that there is significantly higher density in the denial for prevailing wage between 0 and 100 and there is also a high dense of applications with their prevailing wage less than 1 lakh USD.
- The employer's company profile of the visa applicants is shown such that majority of the companies have less than 1 lakh employees working in their firms, but from the density plot and box plots, it is evident that company with higher no. of

employees working has the feasibility to get the acceptance of their employee's visa applications.

- Most of the visa applicants belong to Asian continent, and nearly 79.2% of applications from Europe have received the certificate and tops all the continents with only 20.7% of denials.
- While most of the employees in company's profile, whose visas were certified were Africans, followed by North America, Europe, Asia, South America and Oceania.
- Most of the visa applicants have completed bachelors degree, the percentage of employees who have a masters degree is slightly lower than the bachelor's.
- It is also observed that the no_of_employees with Masters, has a quite high acceptance than any other education level and next to Doctorates, master degree for employees were much preferred by the employers.
- Nearly 87.2% of doctorate holder applications were certified while ≈12.7% were denied, while 10234 applications were from Bachelor degree holders
- Nearly 58.1% of visa applications have a previous job experience while 41.9% were without an experience.
- Nearly 74.5% of applicants who has previous job experience were certified while ≈25.5% were denied, and has a majority of applications of about 14802. Most employees who have job experience did not require job training.
- Almost 88.4% of the visa applicants do not require a job_training while a least percentage of about 11.6% require training for their profession.
- Nearly 66.6% of applicants who do not require a job training were certified while ≈33.3% were denied, and has a majority of applications of about 22525.
- The visa applications from employees working in the Northeast region tops with 28.2% followed by South and West regions, while most of the employees approximately higher than 10k were placed in `Islands`
- The applications from employees who were appointed at midwest region have higher rate of visas certified, which is about 75.5% and lesser denial rates of about 24.5%
- The most used unit of wage is the `Year` unit, and this explains the right skewed distribution of the prevailing wage distribution. It is noted that applicants who receive their wages in yearly basis have higher visa certifications rate (69.8%) and 30.1% of denials
- The employees who are working in `full_time_position` are the major ones to apply for visa, whereas part_time employees are the least ≈10.6%)
- The acceptance rate for full time employees is 66.6% and for the other category is 68.5%, while still majority of the visa applications come from employers who hire employees for full time which is about 22773 (≈89.3%) of total applications. Even for non full time positions, employees with prior experience were opted for about 52.1%.
- From the heatmap, it is understood that there is no much correlation between the numerical columns of the dataset.
- On the overall, it is observed that the certified visas outnumber the denials, that is a majority of the applicants (≈66.8%) were Certified with visa, while less no. of applicants (≈33.2%) were Denied.

2. DATA PRE-PROCESSING

2.1 Duplicate Value check

In order to build an efficient model it is essential to know that if the data set does not contain any duplicate values from the pre-existing rows. The command to check duplicate entries is duplicated().sum(). This returns the total number of duplicated entries in the data set. The provided dataset from EasyVisa for data-driven solutions does not contain any duplicated entries.

2.2 Missing value treatment

Another pre-processing step is to check if the provided data set has missed any values in any of the columns by using the isnull().sum() command. This command counts the missing values in each columns and returns the sum of missing values in each of the column respectively. From Table 5, it can be understood that there is no values missing in this data set.

2.3 Outlier detection and treatment

We see there are so many outliers in each of the numerical column in the data set. So it's indeed essential to carefully examine the data set before treating the outliers. Upon observing and examining the dataset, the following conclusion is made with respect to outliers.

- From the above histogram and box plots, we see there are so many outliers in no_of_employees, yr_of_estab and prevailing_wage.
- These extreme values can be considered for model building, as treating them does not produce an efficient prediction on the model based on the following reasons:

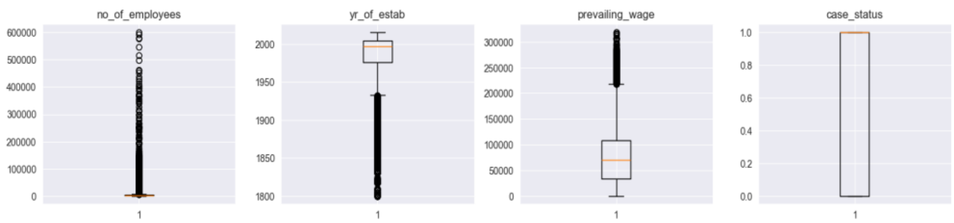
  1. The no of employees can vary depending on the `yr_of_estab` of the companies and the type of business the company is involved in.

  So it is not a weird thing to find companies with lakhs of employees, especially when the company has been established long before. So the outliers showing extreme count of employees can't be treated.

  2. For the year of establishment, it is not unusual to see companies being established in 1800 and still running over 200 years.

  Some companies have been for generations, while others have just been started. Thus the outliers in `yr_of_estab` contain valuable information about the employer company.

  3. The prevailing wage is recorded without considering the unit of wages, thus, if outliers were treated, the adequate information cannot be captured efficiently. Prevailing wage also can vary based on many factors such as the `region_of_employment`, `education_of_employee`, `continent`, level of experience and so on. Hence, the outliers in this column is not treated.




```
no_of_employees     6.107
yr_of_estab        12.794
prevailing_wage     1.676
case_status         0.000
dtype: float64
```

Figure 35: Outlier checks and percentage of outliers for numerical variables of the dataset

2.4 Feature engineering

Dropping of the column `case_id`

The “case_id” column is unique and has no significance in model prediction. Hence the column is dropped.

	continent	education_of_employee	has_job_experience	requires_job_training	no_of_employees	yr_of_estab	region_of_employment	prevailing_wage	unit_of_wage	full_time_position	case_status
0	Asia	High School	N	N	14513	2007	West	592.203	Hour	Y	Denied
1	Asia	Master's	Y	N	2412	2002	Northeast	83425.650	Year	Y	Certified
2	Asia	Bachelor's	N	Y	44444	2008	West	122996.860	Year	Y	Denied
3	Asia	Bachelor's	N	N	98	1897	West	83434.030	Year	Y	Denied
4	Africa	Master's	Y	N	1082	2005	South	149907.390	Year	Y	Certified

Table 10: First five rows showing dropped columns on the feature engineered dataset

Encoding `Denied` and `Certified` "case_status" to `0` and `1` respectively, for analysis

	continent	education_of_employee	has_job_experience	requires_job_training	no_of_employees	yr_of_estab	region_of_employment	prevailing_wage	unit_of_wage	full_time_position	case_status
0	Asia	High School	N	N	14513	2007	West	592.203	Hour	Y	0
1	Asia	Master's	Y	N	2412	2002	Northeast	83425.650	Year	Y	1
2	Asia	Bachelor's	N	Y	44444	2008	West	122996.860	Year	Y	0
3	Asia	Bachelor's	N	N	98	1897	West	83434.030	Year	Y	0
4	Africa	Master's	Y	N	1082	2005	South	149907.390	Year	Y	1

Table 11: First five rows showing encoded `case_status` column on the feature engineered dataset

Correcting inconsistencies in the dataset

From the data description of numerical columns in the dataset, the minimum value of number of employees is found to be ‘-26’, which is clearly seen as an error. This has to be corrected, and in order to do that, the rows carrying such negative value is observed and then replaced using abs() function

	continent	education_of_employee	has_job_experience	requires_job_training	no_of_employees	yr_of_estab	region_of_employment	prevailing_wage	unit_of_wage	full_time_position	case_status
245	Europe	Master's	N	N	-25	1980	Northeast	39452.990	Year	Y	1
378	Asia	Bachelor's	N	Y	-11	2011	Northeast	32506.140	Year	Y	0
832	South America	Master's	Y	N	-17	2002	South	129701.940	Year	Y	1
2918	Asia	Master's	Y	N	-26	2005	Midwest	112799.460	Year	Y	1
6439	Asia	Bachelor's	N	N	-14	2013	South	103.970	Hour	Y	0

Table 12: First five rows showing no_of_employees column with negative values

Observation:

- We see that there are 33 rows which has negative values for the data about `no_of_employees`
- This need to be corrected as positive value using abs(data["no_of_employees"])

2.5 Data preparation for modelling

- We want to predict which of the independent factors is more likely to help in determining the recommendation of “Certified/Denied” case_status of the visa applications.
- Before we proceed to build a model, we'll have to assign the dependent and independent factors to ‘X’ and ‘y’ variables respectively. From Table 1, the target

variable is the dependent variable and the predictor variables are the independent variables

- Then the dataset is split in to train, validation and test sets before carrying out any further process like missing value imputation, outlier treatment and any other feature engineering steps that involves all the rows of the data set.

The data split done before any of the process on data ensures no data leakage among the train-test and validation sets

From the exploratory data analysis and outlier checks we find that the EasyVisa data set does not contain any missing values and outlier values to be treated. Hence we proceed building of the models after splitting of data, followed with one-hot encoding.

Split the data

The dataset is split in to train, validation and test sets in the following percentages. The entire dataset is split into temp and test in the ratio 80:20. Then the temp is split into train and validation in the ratio 75:25 as shown below.

```
Percentage of Train set: 60.0 %
Percentage of Validation set: 20.0 %
Percentage of Test set: 20.0 %
```

Train dataset

No. of rows in train data=15288
No. of columns in train data=10

```
continent
Asia             10085
Europe            2285
North America     1944
South America      528
Africa             333
Oceania            113
Name: count, dtype: int64
******************************
education_of_employee
Bachelor's     6141
Master's       5792
High School    2045
Doctorate      1310
Name: count, dtype: int64
******************************
has_job_experience
Y    8845
N    6443
Name: count, dtype: int64
******************************
requires_job_training
N    13477
Y     1811
Name: count, dtype: int64
******************************
```

```
region_of_employment
Northeast    4312
South        4248
West         3920
Midwest      2576
Island        232
Name: count, dtype: int64
******************************
unit_of_wage
Year     13786
Hour      1286
Week       156
Month       60
Name: count, dtype: int64
******************************
full_time_position
Y    13678
N     1610
Name: count, dtype: int64
******************************
```

Table 13: Value counts of the categorical variables in the train dataset

Validation Dataset

No. of rows in train data=5096

No. of columns in train data=10

```
continent
Asia             3395
Europe            713
North America     655
South America     173
Africa            121
Oceania            39
Name: count, dtype: int64
******************************
education_of_employee
Bachelor's     2033
Master's       1886
High School     694
Doctorate       483
Name: count, dtype: int64
******************************
has_job_experience
Y    2963
N    2133
Name: count, dtype: int64
******************************
requires_job_training
N    4501
Y     595
Name: count, dtype: int64
******************************
```

```
region_of_employment
Northeast    1430
South        1389
West         1352
Midwest       855
Island         70
Name: count, dtype: int64
******************************
unit_of_wage
Year     4576
Hour      452
Week       57
Month      11
Name: count, dtype: int64
******************************
full_time_position
Y    4552
N     544
Name: count, dtype: int64
******************************
```

Table 14: Value counts of the categorical variables in the validation dataset

Test Dataset

No. of rows in train data=5096

No. of columns in train data=10

```
continent
Asia             3381
Europe            734
North America     693
South America     151
Africa             97
Oceania            40
Name: count, dtype: int64
******************************
education_of_employee
Bachelor's     2060
Master's       1956
High School     681
Doctorate       399
Name: count, dtype: int64
******************************
has_job_experience
Y    2994
N    2102
Name: count, dtype: int64
******************************
requires_job_training
N    4547
Y     549
Name: count, dtype: int64
******************************
```

```
region_of_employment
Northeast    1453
South        1380
West         1314
Midwest       876
Island         73
Name: count, dtype: int64
******************************
unit_of_wage
Year     4600
Hour      419
Week       59
Month      18
Name: count, dtype: int64
******************************
full_time_position
Y    4543
N     553
Name: count, dtype: int64
******************************
```

Table 15: Value counts of the categorical variables in the test dataset

Create dummy variables

Values under categorical columns cannot be read into an equation. So one-hot encoding technique is applied to these categorical columns and it is established using a `get-dummies()` function in the pandas dataframe.

Dummy created train dataset

No. of rows in train data=15288

No. of columns in train data=21

	no_of_employees	yr_of_estab	prevailing_wage	continent_Asia	continent_Europe	continent_North America	continent_Oceania	continent_South America
5008	1020	2008	70919.850	True	False	False	False	False
12951	1624	2003	59082.940	False	True	False	False	False
3214	438	1991	22235.800	True	False	False	False	False
18876	211	1911	18937.370	False	True	False	False	False
21939	2696	2007	65906.820	True	False	False	False	False

education_of_employee_Doctorate	education_of_employee_High School	education_of_employee_Master's	has_job_experience_Y	requires_job_training_Y
False	False	False	True	False
False	False	True	True	True
False	False	False	False	False
False	False	False	False	False
False	False	False	False	False

region_of_employment_Midwest	region_of_employment_Northeast	region_of_employment_South	region_of_employment_West	unit_of_wage_Month	unit_of_wage_Week
False	False	True	False	False	False
False	True	False	False	False	False
True	False	False	False	False	False
False	False	False	True	False	False
False	False	True	False	False	False

unit_of_wage_Year	full_time_position_Y
True	True
True	True
True	True
True	True
True	True

Table 16: First five rows of dummy created train data set

```
continent_Asia
True      10085
False      5203
Name: count, dtype: int64
******************************
continent_Europe
False     13003
True       2285
Name: count, dtype: int64
******************************
continent_North America
False     13344
True       1944
Name: count, dtype: int64
******************************
continent_Oceania
False     15175
True        113
Name: count, dtype: int64
******************************
continent_South America
False     14760
True        528
Name: count, dtype: int64
******************************
education_of_employee_Doctorate
False     13978
True       1310
Name: count, dtype: int64
******************************
```

```
education_of_employee_High School
False     13243
True       2045
Name: count, dtype: int64
******************************
education_of_employee_Master's
False      9496
True       5792
Name: count, dtype: int64
******************************
has_job_experience_Y
True       8845
False      6443
Name: count, dtype: int64
******************************
requires_job_training_Y
False     13477
True       1811
Name: count, dtype: int64
******************************
region_of_employment_Midwest
False     12712
True       2576
Name: count, dtype: int64
******************************
region_of_employment_Northeast
False     10976
True       4312
Name: count, dtype: int64
******************************
```

```
region_of_employment_South
False     11040
True       4248
Name: count, dtype: int64
******************************
region_of_employment_West
False     11368
True       3920
Name: count, dtype: int64
******************************
unit_of_wage_Month
False     15228
True         60
Name: count, dtype: int64
******************************
unit_of_wage_Week
False     15132
True        156
Name: count, dtype: int64
******************************
unit_of_wage_Year
True      13786
False      1502
Name: count, dtype: int64
******************************
full_time_position_Y
True      13678
False      1610
Name: count, dtype: int64
******************************
```

Table 17: Value counts of the Boolean variables in the dummy created train dataset

Dummy created validation dataset

No. of rows in validation data=5096

No. of columns in validation data=21

	no_of_employees	yr_of_estab	prevailing_wage	continent_Asia	continent_Europe	continent_North America	continent_Oceania	continent_South America
6360	1282	2008	117135.280	False	False	True	False	False
16248	2586	1984	7242.390	True	False	False	False	False
5828	877	2012	36973.670	False	True	False	False	False
22590	3822	1992	112220.650	False	True	False	False	False
20335	2995	1969	64695.100	True	False	False	False	False

education_of_employee_Doctorate	education_of_employee_High School	education_of_employee_Master's	has_job_experience_Y	requires_job_training_Y
False	False	True	False	False
False	True	False	False	False
False	False	True	False	False
False	True	False	True	False
False	False	True	True	False

region_of_employment_Midwest	region_of_employment_Northeast	region_of_employment_South	region_of_employment_West	unit_of_wage_Month	unit_of_wage_Week
False	False	False	True	False	False
False	False	False	True	False	False
False	True	False	False	False	False
False	True	False	False	False	False
True	False	False	False	False	False

unit_of_wage_Year	full_time_position_Y
True	False
True	True
True	False
True	True
True	False

Table 18: First five rows of dummy created validation data set

```
continent_Asia
True     3395
False    1701
Name: count, dtype: int64
******************************
continent_Europe
False    4383
True      713
Name: count, dtype: int64
******************************
continent_North America
False    4441
True      655
Name: count, dtype: int64
******************************
continent_Oceania
False    5057
True       39
Name: count, dtype: int64
******************************
continent_South America
False    4923
True      173
Name: count, dtype: int64
******************************
education_of_employee_Doctorate
False    4613
True      483
Name: count, dtype: int64
******************************
```

```
education_of_employee_High School
False    4402
True      694
Name: count, dtype: int64
******************************
education_of_employee_Master's
False    3210
True     1886
Name: count, dtype: int64
******************************
has_job_experience_Y
True     2963
False    2133
Name: count, dtype: int64
******************************
requires_job_training_Y
False    4501
True      595
Name: count, dtype: int64
******************************
region_of_employment_Midwest
False    4241
True      855
Name: count, dtype: int64
******************************
region_of_employment_Northeast
False    3666
True     1430
Name: count, dtype: int64
******************************
```

```
region_of_employment_South
False    3707
True     1389
Name: count, dtype: int64
******************************
region_of_employment_West
False    3744
True     1352
Name: count, dtype: int64
******************************
unit_of_wage_Month
False    5085
True       11
Name: count, dtype: int64
******************************
unit_of_wage_Week
False    5039
True       57
Name: count, dtype: int64
******************************
unit_of_wage_Year
True     4576
False     520
Name: count, dtype: int64
******************************
full_time_position_Y
True     4552
False     544
Name: count, dtype: int64
******************************
```

Table 19: Value counts of the Boolean variables in the dummy created validation dataset

Dummy created test dataset

No. of rows in test data=5096

No. of columns in test data=21

	no_of_employees	yr_of_estab	prevailing_wage	continent_Asia	continent_Europe	continent_North America	continent_Oceania	continent_South America
6726	287	2005	72125.460	True	False	False	False	False
9404	708	2005	110222.490	True	False	False	False	False
12977	1524	1928	72723.490	True	False	False	False	False
16089	3928	1973	516.505	False	False	True	False	False
15284	3081	2000	107725.690	True	False	False	False	False

education_of_employee_Doctorate	education_of_employee_High School	education_of_employee_Master's	has_job_experience_Y	requires_job_training_Y
False	False	False	False	True
False	False	False	True	False
False	False	False	False	False
False	False	True	False	False
False	True	False	True	False

region_of_employment_Midwest	region_of_employment_Northeast	region_of_employment_South	region_of_employment_West	unit_of_wage_Month	unit_of_wage_Week
False	False	True	False	False	False
False	False	False	True	False	False
False	False	False	True	False	False
False	True	False	False	False	False
False	True	False	False	False	False

unit_of_wage_Year	full_time_position_Y
True	True
True	True
True	True
False	True
True	True

Table 20: First five rows of dummy created test data set

```
continent_Asia
True      3381
False     1715
Name: count, dtype: int64
******************************
continent_Europe
False     4362
True       734
Name: count, dtype: int64
******************************
continent_North America
False     4403
True       693
Name: count, dtype: int64
******************************
continent_Oceania
False     5056
True        40
Name: count, dtype: int64
******************************
continent_South America
False     4945
True       151
Name: count, dtype: int64
******************************
education_of_employee_Doctorate
False     4697
True       399
Name: count, dtype: int64
******************************
education_of_employee_High School
False     4415
True       681
Name: count, dtype: int64
******************************
education_of_employee_Master's
False     3140
True      1956
Name: count, dtype: int64
******************************
has_job_experience_Y
True      2994
False     2102
Name: count, dtype: int64
******************************
requires_job_training_Y
False     4547
True       549
Name: count, dtype: int64
******************************
region_of_employment_Midwest
False     4220
True       876
Name: count, dtype: int64
******************************
region_of_employment_Northeast
False     3643
True      1453
Name: count, dtype: int64
******************************
region_of_employment_South
False     3716
True      1380
Name: count, dtype: int64
******************************
region_of_employment_West
False     3782
True      1314
Name: count, dtype: int64
******************************
unit_of_wage_Month
False     5078
True        18
Name: count, dtype: int64
******************************
unit_of_wage_Week
False     5037
True        59
Name: count, dtype: int64
******************************
unit_of_wage_Year
True      4600
False      496
Name: count, dtype: int64
******************************
full_time_position_Y
True      4543
False      553
Name: count, dtype: int64
******************************
```

Table 21: Value counts of the Boolean variables in the dummy created test dataset

3. MODEL BUILDING-ORIGINAL DATA

3.1 Model evaluation criterion

Model can make wrong predictions as:

1. Predicting the recommendation of the visa application as “Certified”, but in reality it has to be denied – False Positive - Loss of opportunity for US citizens
2. Predicting the recommendation of the visa application as “Denied” but in reality it has to be certified - False Negative - Loss of valuable resource

Which case is more important?

Both are important:

- If the visa application is recommended as “Certified”, but it has to be “Denied”, then the US embassy would end up giving the opportunity to a wrong person who would not contribute to the growth of the company and in turn to the country’s economy. The wrong person would also grab the job opportunity of an US citizen, for whom that position would have been of great benefit.

- If the visa application is recommended as “Denied”, but it has to be “Certified”, then the US embassy would end up missing a valuable human resource who would contribute to the development of the organization and in turn for the economy of the country.

How to reduce these costs i.e maximize True Positives?

- We need to reduce both False Negatives and False Positives

- **F1_score** should be maximized, as greater the f1_score, higher the chances of reducing both False Negatives and False Positives and identifying both the classes correctly

- F1_score is computed as

$$f1_{score} = \frac{2 * Precision * Recall}{Precision + Recall}$$

3.2 Build the Model –Original data

Before building of the model, certain functions were defined to analyse the different output metrics such as Accuracy, Precision, Recall and F1_score, and also to visualize the confusion matrix. Initially the models are built on the original data. The following models are built on the original data set.

1. Bagging Classifier
2. Random Forest Classifier
3. Gradient Boosting Classifier

4. Adaptive Boosting Classifier
5. Extreme Gradient Boosting Classifier
6. Decision Tree Classifier

Cross-validation Performance

As per the model evaluation criteria, it is important to maximize the true positives, hence F1_score is calculated for all the models by dividing the training data into k folds and the cross validation performance is done to analyse the best performing model on the training data set.

```
Cross-Validation Performance:

Bagging: 77.49819034445665
Random forest: 80.43381835409008
GBM: 82.23176915133634
Adaboost: 82.09082175550402
Xgboost: 80.88109618665464
dtree: 74.23560177028313
```

Table 22: Cross-validation performance evaluation metric (F1_score) using all the models on training set

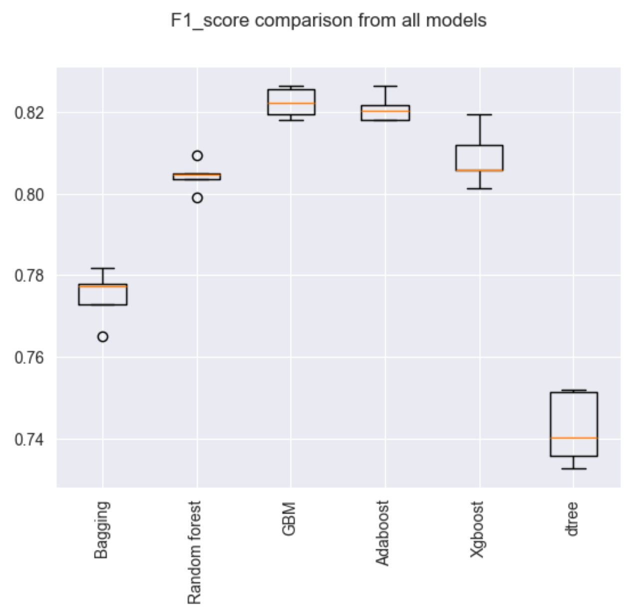


Figure 36: Box plot of Cross-validation evaluation metric (F1_score) using all the models on training set

Observation

- We can see that the Gradient Boosting mechanism is giving the highest cross-validated f1_score followed by Adaboost and then XGBoost.
- The boxplot shows that the performance of GradientBoost and Adaboost is consistent and their performance on the validation set is also good with a very low difference of 0.0025 and 0.0024 respectively (Table 24).

Model Building using original data

1. The BaggingClassifier, RandomForestClassifier, GradientBoostingClassifier, AdaBoostClassifier, XGBClassifier and the DecisionTreeClassifier models are defined with random_state=1, class_weight=balanced, and eval_metric= “logloss” for the respective models.
2. Then they are built with .fit() command accordingly
3. The evaluation metrics such as Accuracy, Recall, Precision and F1_score are calculated and tabulated as below.

```
Performance of the models on train and validation sets:

Bagging:
                Accuracy  Recall  Precision    F1
0   Train Perf     0.986   0.987      0.992 0.989
1  Valset Perf     0.699   0.771      0.776 0.774


Random forest:
                Accuracy  Recall  Precision    F1
0   Train Perf     1.000   1.000      1.000 1.000
1  Valset Perf     0.727   0.842      0.771 0.805


GBM:
                Accuracy  Recall  Precision    F1
0   Train Perf     0.758   0.879      0.785 0.829
1  Valset Perf     0.755   0.873      0.785 0.827


Adaboost:
                Accuracy  Recall  Precision    F1
0   Train Perf     0.740   0.888      0.762 0.820
1  Valset Perf     0.738   0.881      0.764 0.818

Xgboost:
                Accuracy  Recall  Precision    F1
0   Train Perf     0.855   0.939      0.857 0.896
1  Valset Perf     0.729   0.852      0.768 0.808


dtree:
                Accuracy  Recall  Precision    F1
0   Train Perf     1.000   1.000      1.000 1.000
1  Valset Perf     0.664   0.748      0.749 0.749
```

Table 23: Performance evaluation metrics using all the models on training and validation set

```
Training and Validation Performance Difference:

Bagging: Training Score: 0.9892, Validation Score: 0.7737, Difference: 0.2155
Random forest: Training Score: 1.0000, Validation Score: 0.8050, Difference: 0.1950
GBM: Training Score: 0.8291, Validation Score: 0.8266, Difference: 0.0025
Adaboost: Training Score: 0.8204, Validation Score: 0.8180, Difference: 0.0024
Xgboost: Training Score: 0.8963, Validation Score: 0.8079, Difference: 0.0884
dtree: Training Score: 1.0000, Validation Score: 0.7486, Difference: 0.2514
```

Table 24: Difference of F1_score between training and validation sets on all the models

4. Then the confusion matrix is created for the validation sets to analyse the models performance as shown in Figure 37.

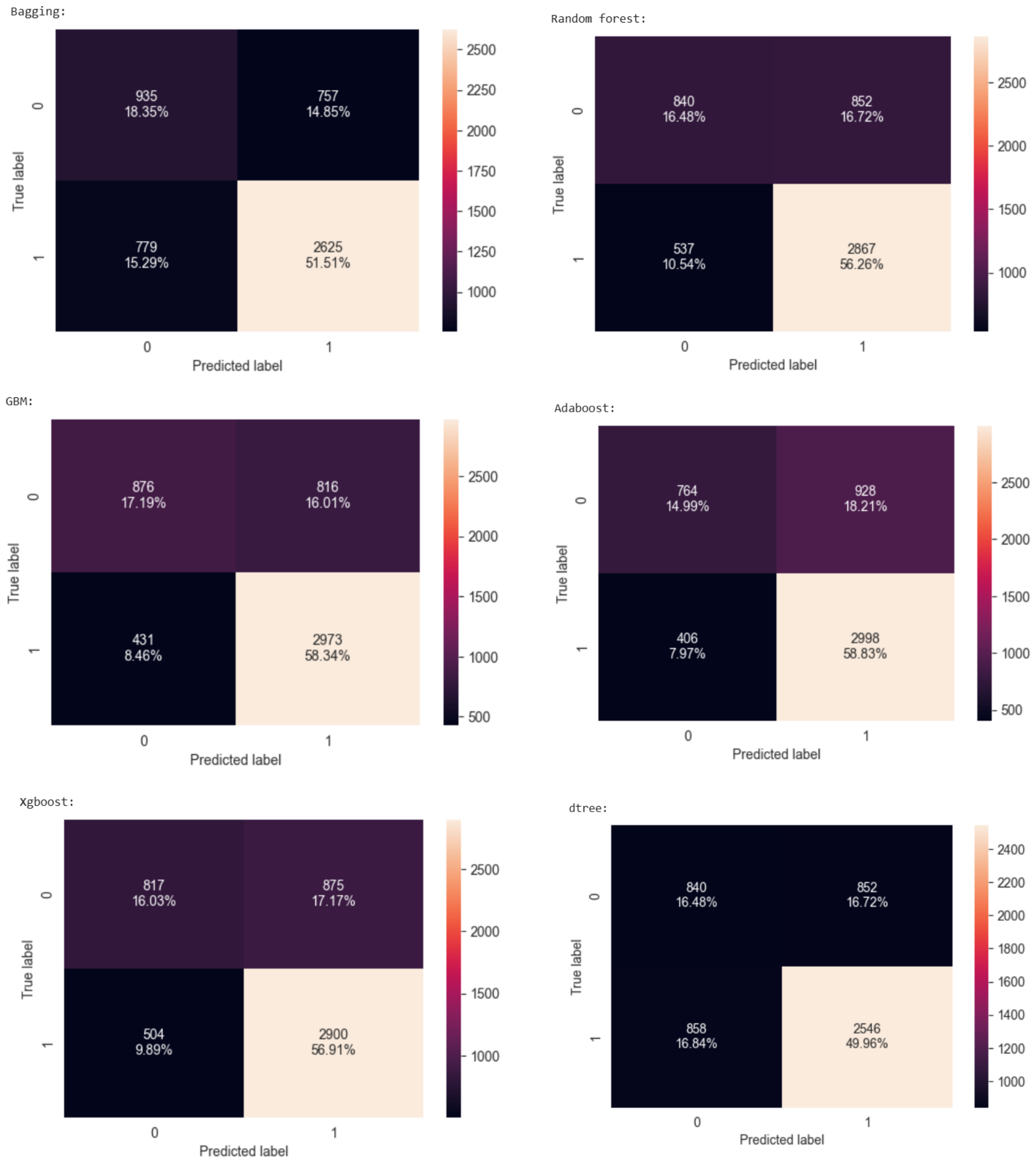


Figure 37: Confusion matrix for all the models using the validation set

3.3 Comments on the model performance using original data

Bagging: The bagging classifier seems to be overfitting as it performs good on training set but not that efficient on the validation set, as all the performance evaluation metrics shows ≈0.2 difference between the train and validation sets. It predicts the "True Positives" with about 51.51% accuracy.

Random Forest: The Random Forest classifier is definitely overfitting as it has the value 1.0 for all the evaluation metrics but does not perform on the validation set. This model also shows a difference F1_score of 0.195 and predicts the "True Positives" with about 56.26% accuracy.

Gradient Boost: The GB classifier performs extremely good on the validation test and gives the maximum F1_score of 0.827 and a F1_score difference of only 0.0025 between the training and validation sets. It also predicts the "True Positives" with a good accuracy of 58.34%

AdaBoost: The AdaptiveBoosting classfier also performs equally and better to GB classifier both on the train and validation sets. Though GB's F1_score is maximum, AdaBoost(F1_score = 0.818) gives the minimum F1_score difference between the train and validation sets with a value of 0.0024 and tops all the models in its prediction of "True Positives" having 58.83% accuracy.

XGBoost: The XGBoost classifier performs next to GB classifier and AdaBoost classifier with a good F1_score of about 0.808 and the difference F1_score of ≈0.08 between the training and validation sets. It does "True Positive" prediction with 56.91% accuracy.

Decision Tree: The decision tree model performs very poor than all the model with the lowest of all F1_scores (0.749) and the maximum difference F1_score (0.2514). It also shows a poor "True Positive" prediction of about 49.96% accuracy.

By analysing the various evaluation metrics and confusion matrix on the validation set the following models are ranked in order below.

1. **GradientBoost**
2. **AdaBoost**
3. **XGBoost**

4. MODEL BUILDING-OVERSAMPLED DATA

4.1 Oversampling the original data

To build the model with oversampled dataset, the original split portion of training set is oversampled using the SMOTE (Synthetic Minority Over-sampling Technique) which oversamples the dataset with the help of k-nearest neighbour algorithm as a part of its process. The oversampled dataset has increased no. of rows both in the X_train samples and in y_train samples as shown below.

```
Before Oversampling, counts of label 'Yes': 10210
Before Oversampling, counts of label 'No': 5078

After Oversampling, counts of label 'Yes': 10210
After Oversampling, counts of label 'No': 10210

After Oversampling, the shape of train_X: (20420, 21)
After Oversampling, the shape of train_y: (20420,)
```

Table 25: Shape and size of the oversampled training dataset

4.2 Build the model-Oversampled data

After oversampling the data, the models are built on the oversampled data. The following models are built on the oversampled training data set.

1. Bagging Classifier
2. Random Forest Classifier
3. Gradient Boosting Classifier
4. Adaptive Boosting Classifier
5. Extreme Gradient Boosting Classifier
6. Decision Tree Classifier

Cross-validation Performance

As per the model evaluation criteria, it is important to maximize the true positives, hence F1_score is calculated for all the models by dividing the oversampled training data into k folds and the cross validation performance is done to analyse the best performing model on the oversampled training data set.

```
Cross-Validation Performance on oversampled training set:

Bagging: 75.55503868710014
Random forest: 79.3178171785329
GBM: 80.28773059477831
Adaboost: 79.79150439247013
Xgboost: 79.93652879332656
dtree: 72.66966032310629
```

Table 26: Cross-validation performance evaluation metric (F1_score) using all the models on oversampled training set

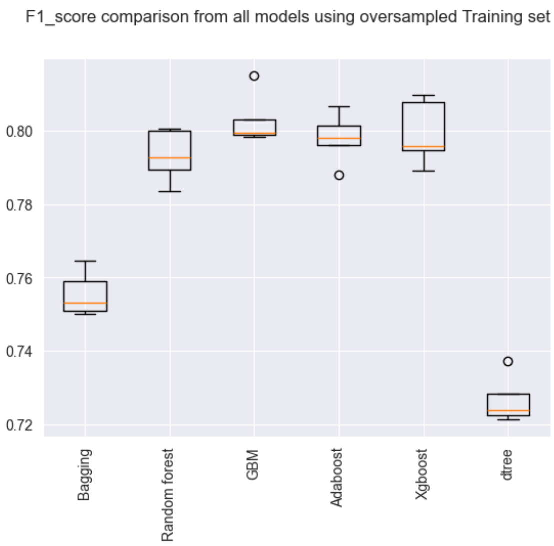


Figure 38: Box plot of Cross-validation evaluation metric (F1_score) using all the models on oversampled training set

Observation

- We can see that the Gradient Boosting mechanism is giving the highest cross-validated f1_score of 80.28 followed by XGBoost (79.93)

- The boxplot shows that the performance of GradientBoost, XGBoost and Adaboost is consistent and their performance on the validation set is also good with a very low difference of -0.0101, 0.0579 and -0.0190 respectively (Table 28).
- From the F1_scores on the validation test AdaBoost gives the highest F1_score and the lowest difference of F1_score between the train and validation, thus tops all the models in performance while training using an oversampled training set

Model Building using oversampled data

1. The BaggingClassifier, RandomForestClassifier, GradientBoostingClassifier, AdaBoostClassifier, XGBClassifier and the DecisionTreeClassifier models are defined with random_state=1, class_weight=balanced, and eval_metric= "logloss" for the respective models.
2. Then they are built with .fit() command accordingly
3. The evaluation metrics such as Accuracy, Recall, Precision and F1_score are calculated and tabulated as below.

```
Performance of the models on oversampled train and validation sets:

Bagging:
              Accuracy  Recall  Precision    F1
0   Train Perf     0.988   0.983      0.993 0.988
1  Valset Perf     0.693   0.753      0.780 0.767


Random forest:
              Accuracy  Recall  Precision    F1
0   Train Perf     1.000   1.000      1.000 1.000
1  Valset Perf     0.723   0.813      0.781 0.797


GBM:
              Accuracy  Recall  Precision    F1
0   Train Perf     0.796   0.854      0.766 0.807
1  Valset Perf     0.747   0.846      0.790 0.817


Adaboost:
              Accuracy  Recall  Precision    F1
0   Train Perf     0.780   0.884      0.731 0.801
1  Valset Perf     0.740   0.882      0.765 0.820


Xgboost:
              Accuracy  Recall  Precision    F1
0   Train Perf     0.864   0.916      0.830 0.871
1  Valset Perf     0.739   0.850      0.779 0.813

dtree:
              Accuracy  Recall  Precision    F1
0   Train Perf     1.000   1.000      1.000 1.000
1  Valset Perf     0.650   0.715      0.749 0.732
```

Table 27: Performance evaluation metrics using all the models on oversampled training and validation set

```
Oversampled Training and Validation Performance Difference in F1_score:

Bagging: Training Score: 0.9875, Validation Score: 0.7665, Difference: 0.2210
Random forest: Training Score: 1.0000, Validation Score: 0.7965, Difference: 0.2034
GBM: Training Score: 0.8072, Validation Score: 0.8173, Difference: -0.0101
Adaboost: Training Score: 0.8005, Validation Score: 0.8195, Difference: -0.0190
Xgboost: Training Score: 0.8709, Validation Score: 0.8129, Difference: 0.0579
dtree: Training Score: 1.0000, Validation Score: 0.7320, Difference: 0.2680
```

Table 28: Difference of F1_score between oversampled training and validation sets on all the models

4. Then the confusion matrix is created for the validation sets to analyse the models performance using oversampled training set is shown in Figure 39.

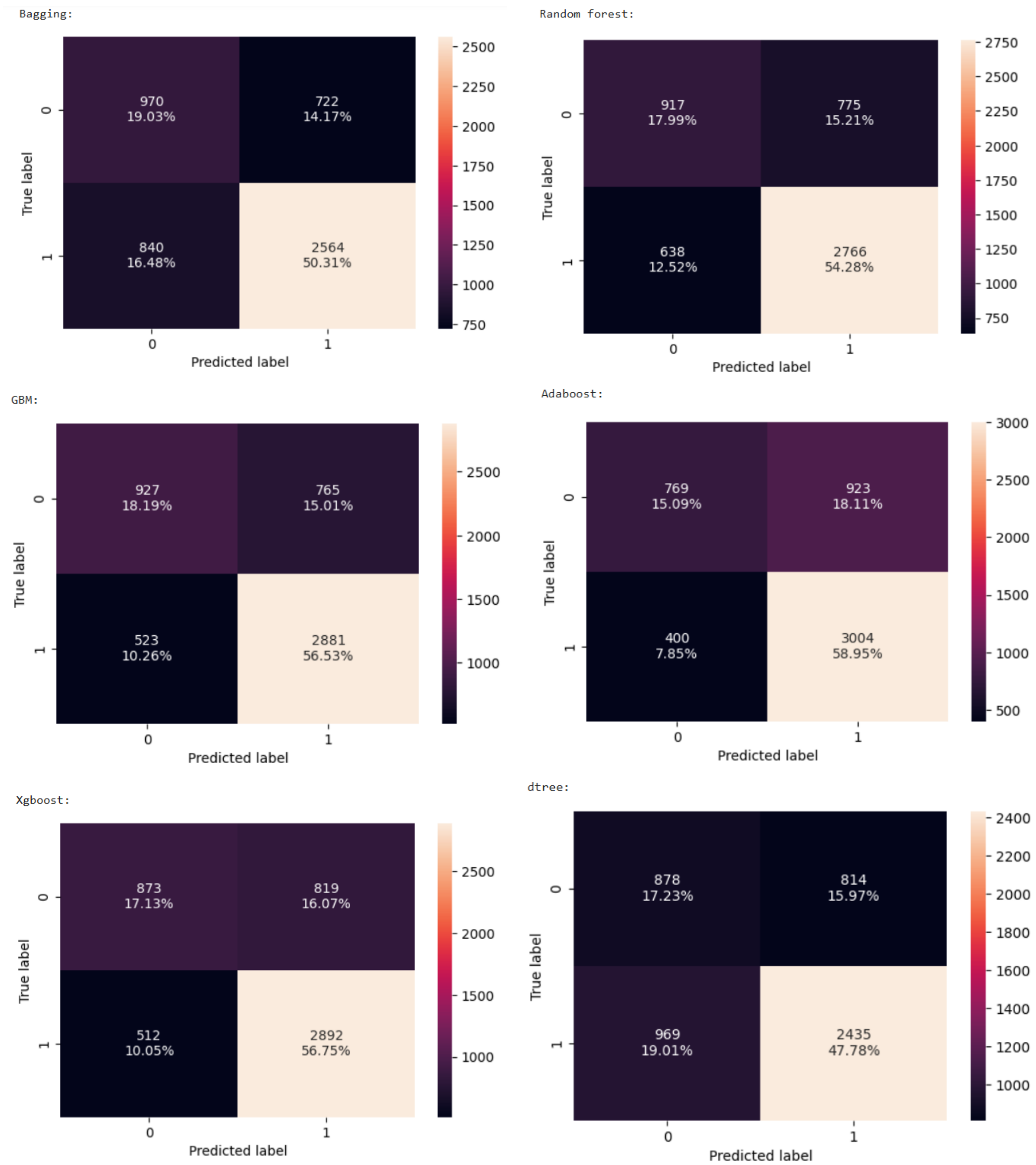


Figure 39: Confusion matrix for all the models on the validation set trained using oversampled training set

4.3 Comments on the model performance using oversampled data

Bagging: The bagging classifier seems to be overfitting as it performs well on training set but not that efficient on the validation set, as all the performance evaluation metrics shows ≈0.2 difference between the train and validation sets. It predicts the “True Positives” with about 50.31% accuracy. Bagging is lagging in performance when trained using oversampled data compared to the original data

Random Forest: The Random Forest classifier is definitely overfitting as it has the value 1.0 for all the evaluation metrics but does not perform on the validation set. This model also shows a difference F1_score of 0.203 and predicts the “True Positives” with about

54.28% accuracy. Random Forest is also lagging in performance when trained using oversampled data compared to the original data

Gradient Boost: The GB classifier performs well on the validation test and gives the F1_score of 0.817 on the validation set and with an F1_score difference of -0.0101 between the oversampled training and validation sets. This shows that the F1_score of the model, on the validation set is better compared to that on the oversampled training set. It also predicts the "True Positives" with a good accuracy of 56.53%. Gradient Boost also lags in predicting the "True Positives", when trained using oversampled data compared to the original data

AdaBoost: The AdaptiveBoosting classfier performs better to GB classifier on the validation sets. The F1_score is maximum for AdaBoost(F1_score = 0.820). It also gives the minimum F1_score difference between the oversampled train and validation sets with a value of -0.0190 and tops all the models in its prediction of "True Positives" having 58.95% accuracy. AdaBoost seems to show an improvement in "True Positive" predictions when trained with oversampled training set compared to the original data

XGBoost: The XGBoost classifier performs next to AdaBoost classifier and GB classifier with a good F1_score of about 0.813 and the difference F1_score of ≈0.057 between the oversampled training and validation sets. It does "True Positive" prediction with 56.75% accuracy and shows a decrease in prediction accuracy compared to original data training. But as per the validation performance the F1_score is better than the original data.

Decision Tree: The decision tree model performs very poor than all the models with the lowest of all F1_scores (0.732) and the maximum difference F1_score (0.2680). It also shows a poor "True Positive" prediction of about 47.78% accuracy and a decrease in accuracy compared to original data training

By analysing the various evaluation metrics and confusion matrix on the validation set when trained the following models using oversampled training set, their rankings in order is concluded as below.

1. **AdaBoost**
2. **GradientBoost**
3. **XGBoost**

5. MODEL BUILDING-UNDERSAMPLED DATA

5.1 Undersampling the original data

To build the model with undersampled dataset, the original split portion of training set is undersampled using Random Under Sampler.
The undersampled dataset has decreased no. of rows both in the X_train samples and in y_train samples as shown below.

```
Before Under Sampling, counts of label 'Yes': 10210
Before Under Sampling, counts of label 'No': 5078

After Under Sampling, counts of label 'Yes': 5078
After Under Sampling, counts of label 'No': 5078

After Under Sampling, the shape of train_X: (10156, 21)
After Under Sampling, the shape of train_y: (10156,)
```

Table 29: Shape and size of the undersampled training dataset

5.2 Build the model-Undersampled data

After undersampling the data, the models are built on the undersampled data. The following models are built on the undersampled training data set.

1. Bagging Classifier
2. Random Forest Classifier
3. Gradient Boosting Classifier
4. Adaptive Boosting Classifier
5. Extreme Gradient Boosting Classifier
6. Decision Tree Classifier

Cross-validation Performance

As per the model evaluation criteria, it is important to maximize the true positives, hence F1_score is calculated for all the models by dividing the undersampled training data into k folds and the cross validation performance is done to analyse the best performing model on the undersampled training data set.

```
Cross-Validation Performance on undersampled training set:

Bagging: 63.58910991015145
Random forest: 67.70437575849454
GBM: 71.23880772066684
Adaboost: 70.71916646661187
Xgboost: 67.87320411742058
dtree: 62.06808909417422
```

Table 30: Cross-validation performance evaluation metric (F1_score) using all the models on undersampled training set

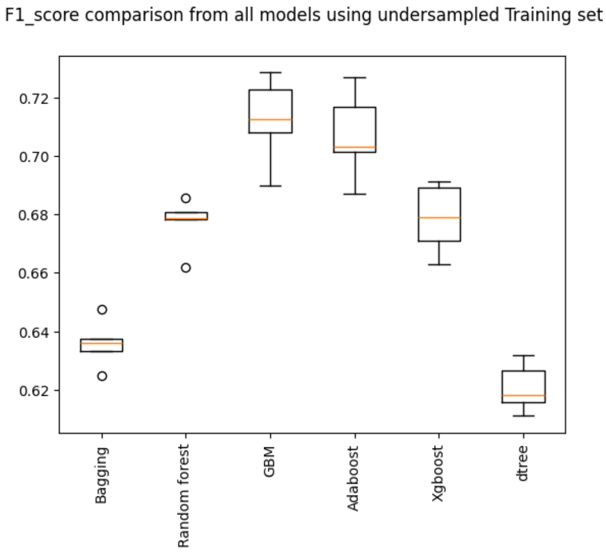


Figure 40: Box plot of Cross-validation evaluation metric (F1_score) using all the models on undersampled training set

Observation

- We can see that the Gradient Boosting mechanism is giving the highest cross-validated f1_score of 71.23 followed by AdaBoost (70.719)
- The boxplot shows that the performance of GradientBoost is consistent and their performance on the validation set is also good with a low difference of -0.04. The performance of GBM is followed by AdaBoost showing a lower difference in F1_score (-.06).
- From the F1_scores on the validation test, GBM gives the highest F1_score (0.777) and the AdaBoost gives the lowest difference of F1_score between the train and validation, thus both the models goes hand in hand exhibiting good performance while training using an undersampled training set.

Model Building using undersampled data

1. The BaggingClassifier, RandomForestClassifier, GradientBoostingClassifier, AdaBoostClassifier, XGBClassifier and the DecisionTreeClassifier models are defined with random_state=1, class_weight=balanced, and eval_metric= “logloss” for the respective models.
2. Then they are built with .fit() command accordingly
3. The evaluation metrics such as Accuracy, Recall, Precision and F1_score are calculated and tabulated as below.

```
Performance of the models on undersampled train and validation sets:

Bagging:
              Accuracy  Recall  Precision    F1
0   Train Perf     0.981   0.969      0.992 0.980
1  Valset Perf     0.656   0.618      0.823 0.706


Random forest:
              Accuracy  Recall  Precision    F1
0   Train Perf     1.000   1.000      1.000 1.000
1  Valset Perf     0.686   0.675      0.824 0.742


GBM:
              Accuracy  Recall  Precision    F1
0   Train Perf     0.721   0.748      0.709 0.728
1  Valset Perf     0.720   0.729      0.831 0.777


Adaboost:
              Accuracy  Recall  Precision    F1
0   Train Perf     0.695   0.716      0.688 0.702
1  Valset Perf     0.708   0.716      0.824 0.766


Xgboost:
              Accuracy  Recall  Precision    F1
0   Train Perf     0.871   0.880      0.865 0.872
1  Valset Perf     0.688   0.687      0.817 0.746

dtree:
              Accuracy  Recall  Precision    F1
0   Train Perf     1.000   1.000      1.000 1.000
1  Valset Perf     0.631   0.631      0.774 0.696
```

Table 31: Performance evaluation metrics using all the models on undersampled training and validation set

```
Undersampled Training and Validation Performance Difference in F1_score:

Bagging: Training Score: 0.9804, Validation Score: 0.7057, Difference: 0.2747
Random forest: Training Score: 1.0000, Validation Score: 0.7417, Difference: 0.2583
GBM: Training Score: 0.7281, Validation Score: 0.7766, Difference: -0.0485
Adaboost: Training Score: 0.7015, Validation Score: 0.7660, Difference: -0.0645
Xgboost: Training Score: 0.8720, Validation Score: 0.7459, Difference: 0.1261
dtree: Training Score: 1.0000, Validation Score: 0.6956, Difference: 0.3044
```

Table 32: Difference of F1_score between undersampled training and validation sets on all the models

4. Then the confusion matrix is created for the validation sets to analyse the models performance using undersampled training set is shown in Figure 41.

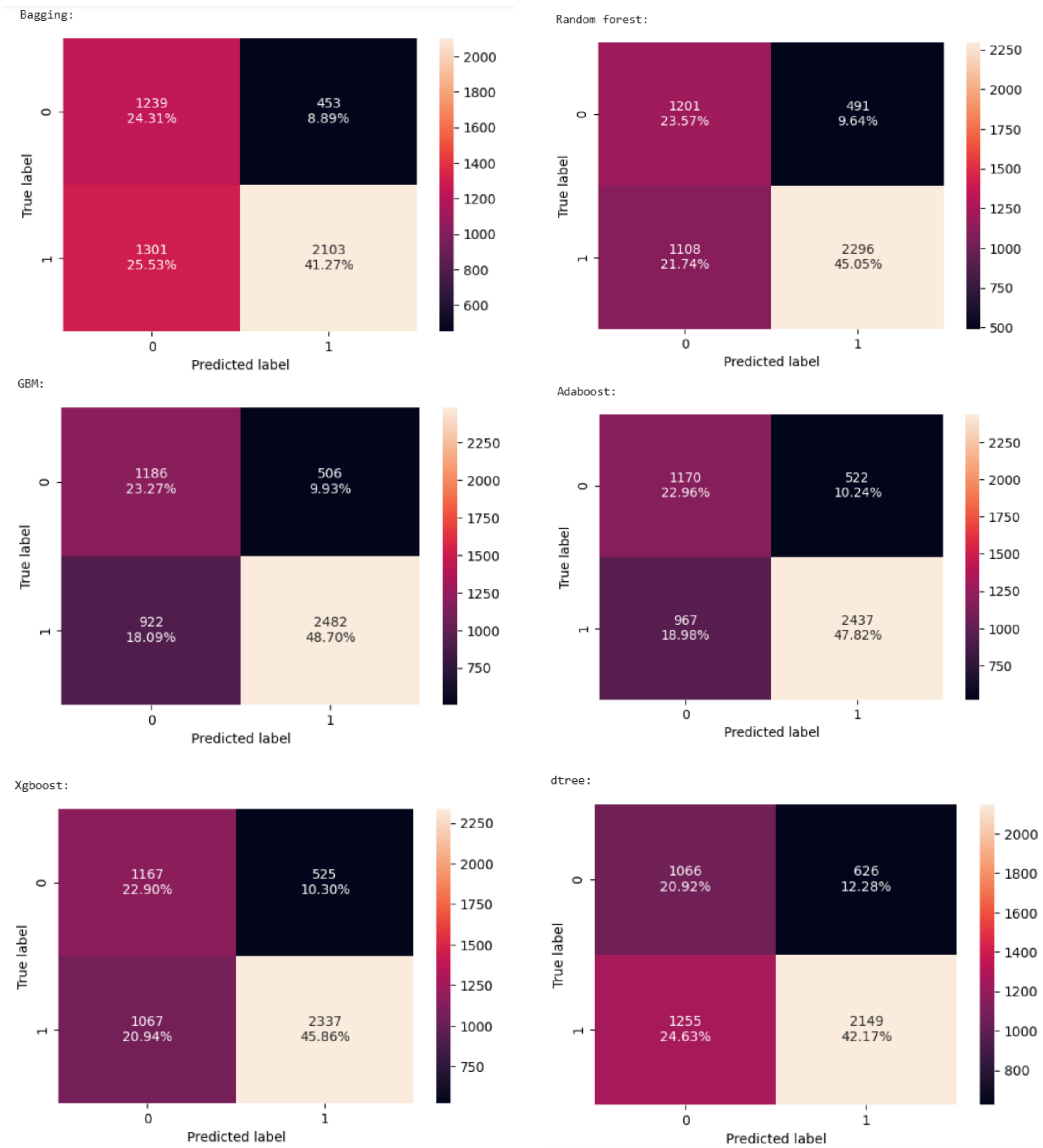


Figure 41: Confusion matrix for all the models on the validation set trained using undersampled training set

5.3 Comments on the model performance using undersampled data

Bagging: The bagging classifier seems to be overfitting as it performs good on undersampled training set but not that efficient on the validation set, as all the performance evaluation metrics shows ≈0.27 difference between the train and validation sets. It predicts the "True Positives" with about 41.27% accuracy. Bagging is lagging in performance when trained using undersampled data compared to the original and oversampled datasets.

Random Forest: The Random Forest classifier is definitely overfitting as it has the value 1.0 for all the evaluation metrics but does not perform on the validation set. This model also shows a difference F1_score of ≈0.26 and predicts the "True Positives" with about 45.05% accuracy. Random Forest is also lagging in performance when trained using undersampled data compared to the original and oversampled datasets.

Gradient Boost: The GB classifier performs well on the validation test and gives the F1_score of 0.777 on the validation set and with an F1_score difference of -0.0485 between the undersampled training and validation sets. This shows that the F1_score of the model, on the validation set is better compared to that on the undersampled training set. It also predicts the "True Positives" with a good accuracy of 48.70%. Gradient Boost also lags in predicting the "True Positives", when trained using undersampled data compared to the original and oversampled datasets

AdaBoost: The AdaptiveBoosting classfier performs well on validation sets similar to GB classifier. It also gives the minimum F1_score difference between the undersampled train and validation sets with a value of -0.0645 and makes prediction of "True Positives" with 47.82% accuracy. AdaBoost seems to show an improvement in "True Positive" predictions when trained with oversampled training set but poorer with undersampled training set, and still poor than with original data.

XGBoost: The XGBoost classifier performs next to AdaBoost classifier and GB classifier with a good F1_score of about 0.746 and the difference F1_score of ≈0.12 between the undersampled training and validation sets. It does "True Positive" prediction with 45.86% accuracy and shows a decrease in prediction accuracy compared to original and oversampled data training.

Decision Tree: The decision tree model performs very poor than all the models with the lowest of all F1_scores (0.696) and the maximum difference F1_score (0.3044). The training set evaluation metrics shows that the model is clearly overfitting. It also shows a poor "True Positive" prediction of about 42.17% accuracy and slightly better to bagging, but a decrease in accuracy compared to original and oversampled data training

By analysing the various evaluation metrics and confusion matrix on the validation set when trained the following models using undersampled training set, their rankings in order is concluded as below.

1. **GradientBoost**
2. **AdaBoost**
3. **XGBoost**

6. MODEL PERFORMANCE IMPROVEMENT USING HYPERPARAMETER TUNING

6.1 Reasoning

Considering the Tables 24, 28 and 32 along with figures 36, 38 and 40 the following reasonable conclusions are made based on which the hyper parameter tuning is carried out to improve the performance of the model.

- The Box plot shows that the performance of AdaBoost and GBM is consitent followed by XGBoost and their performance on the validation set is also good
- After building 18 models, and analysing its performance on the validation set it was observed that the GBM trained on an original data exhibits strong performance on both training and validation sets
- Next to GBM, both AdaBoost and XGBoost models, trained on an oversampled dataset exhibits a consistent performance.
- Sometimes models might overfit after oversampling, so it's better to tune the models to get a generalized performance
- We will tune these 5 models using the same data as we trained them on before.
    1. GBM trained on original data
    2. AdaBoost trained on an oversampled dataset
    3. AdaBoost trained on original dataset
    4. GBM trained on oversampled dataset
    5. XGBoost trained on an oversampled dataset

6.2 Tuned GBM trained on original data

The GBM model with original data is tuned with the best parameter as shown in table below.

- subsample: 0.7
- n_estimators: 50
- max_features: 0.7
- learning_rate: 0.05
- init: AdaBoostClassifier (random_state=1)

The tuned GBM trained on original data gives a CV score = 0.825

```
Best parameters are {'subsample': 0.7, 'n_estimators': 50, 'max_features': 0.7, 'learning_rate': 0.05, 'init': AdaBoostClassifier(random_state=1)} with C
V score=0.8253298966061878:
CPU times: total: 6.72 s
Wall time: 1min 17s
```

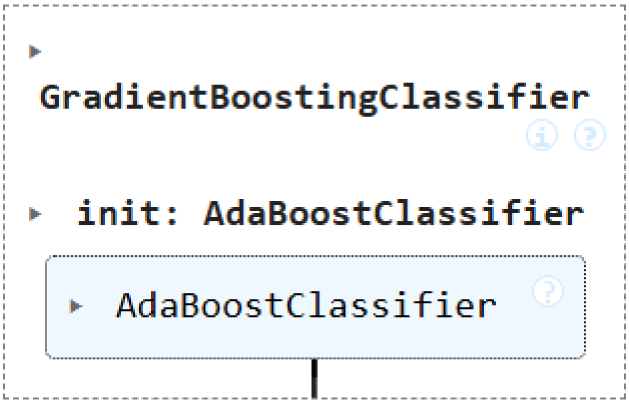


Figure 42: Tuned GBM trained on original data

	Accuracy	Recall	Precision	F1
0	0.750	0.899	0.767	0.828

Table 33: Training performance of the tuned GBM with original data

	Accuracy	Recall	Precision	F1
0	0.749	0.897	0.767	0.827

Table 34: Validation performance of the tuned GBM with original data

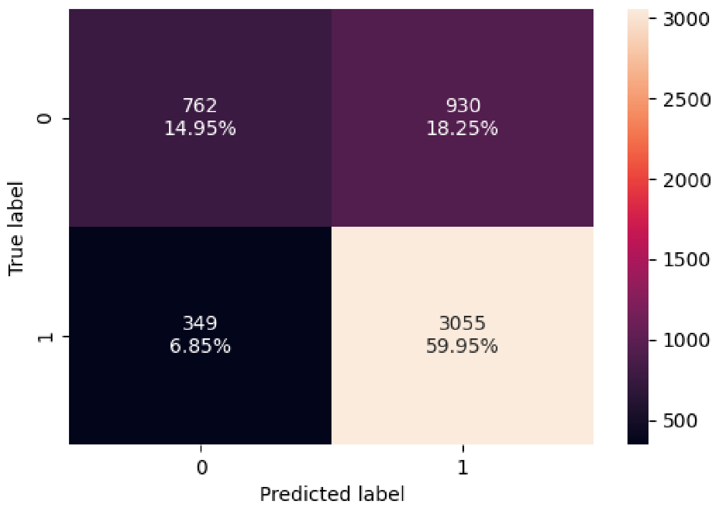


Figure 43: Confusion matrix of the Tuned GBM trained on original data

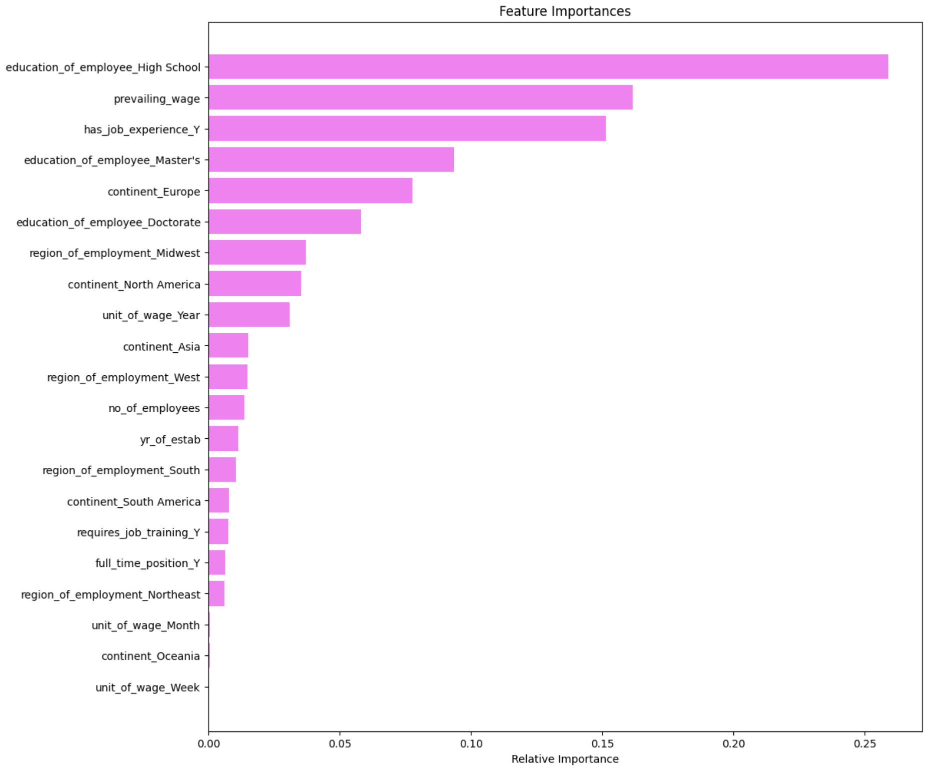


Figure 44: Feature importances of the Tuned GBM trained on original data

Observations

- The performance of the GBM model gives approximately the same F1 score before and after hyperparameter tuning.
- In terms of precision and accuracy, the default model has a slightly better score.Recall is however better in this tuned model.
- The top 5 important features here are eduction of employee (high school and masters), has job experience (y), prevailing wage and continent Europe.
- The tuned model predicts the "True Positives" with approximately 60% accuracy, which is better than the default model.

6.3 AdaBoost trained on an oversampled dataset

The AdaBoost model with oversampled data is tuned with the best parameter as shown in table below.

- n_estimators: 30
- learning_rate: 1
- estimator: DecisionTreeClassifier (max_depth=3, random_state=1)

The tuned AdaBoost trained on an oversampled data gives a CV score = 0.7897

```
Best parameters are {'n_estimators': 30, 'learning_rate': 1, 'estimator': DecisionTreeClassifier(max_depth=3, random_state=1)} with CV score=0.7897563703
598921:
CPU times: total: 1.58 s
Wall time: 8.47 s
```

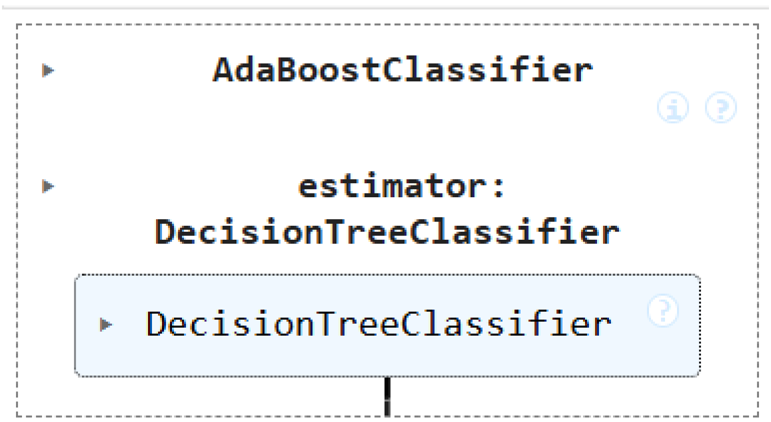


Figure 45: Tuned AdaBoost trained on oversampled data

	Accuracy	Recall	Precision	F1
0	0.786	0.854	0.752	0.800

Table 35: Training performance of the tuned AdaBoost trained on oversampled data

	Accuracy	Recall	Precision	F1
0	0.742	0.848	0.783	0.814

Table 36: Validation performance of the tuned AdaBoost trained on oversampled data

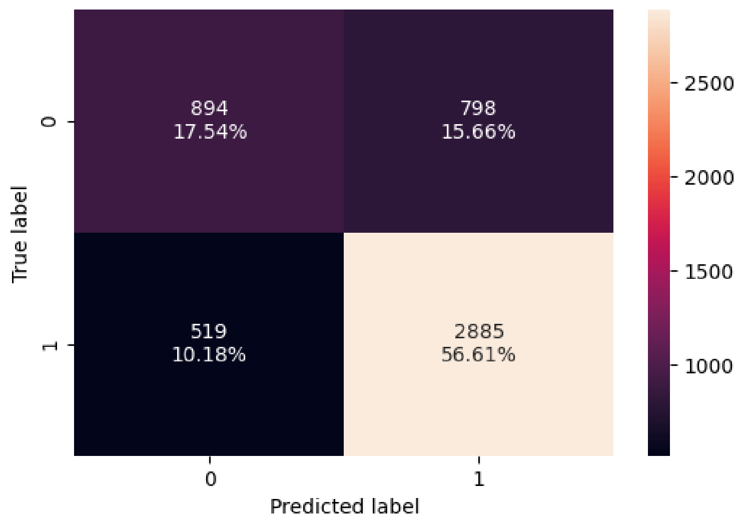


Figure 46: Confusion matrix of the tuned AdaBoost trained on oversampled data

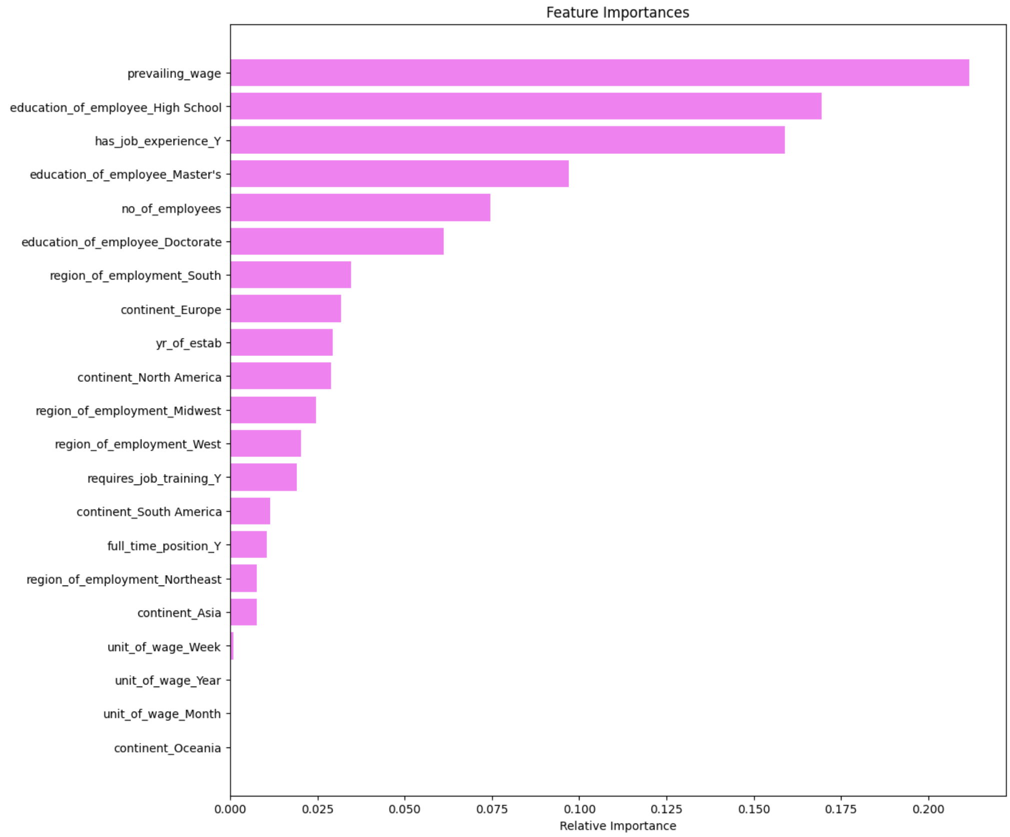


Figure 47: Feature importances of the tuned AdaBoost trained on oversampled data

Observations

- The performance of the AdaBoost model gives a slightly lower F1 score after hyperparameter tuning the model which was trained with oversampled data.
- In terms of precision and accuracy, the tuned model with oversampled data has a slightly better score. However recall has a better score in the default model trained with oversampled data.
- The top 5 important features here are eduction of employee (high school and masters), has job experience (y), prevailing wage and no. of employees.
- The "True Positive" prediction rate was better on the default model with oversampled data than the tuned model.

6.4 AdaBoost trained on Original dataset

The AdaBoost model with original data is tuned with the best parameter as shown in table below.

- n_estimators: 20
- learning_rate: 0.2
- estimator: DecisionTreeClassifier (max_depth=3, random_state=1)

The tuned AdaBoost trained on original data gives a CV score = 0.822

```
Best parameters are {'n_estimators': 20, 'learning_rate': 0.2, 'estimator': DecisionTreeClassifier(max_depth=3, random_state=1)} with CV score=0.82233229
12021356:
CPU times: total: 2.33 s
Wall time: 14.5 s
```

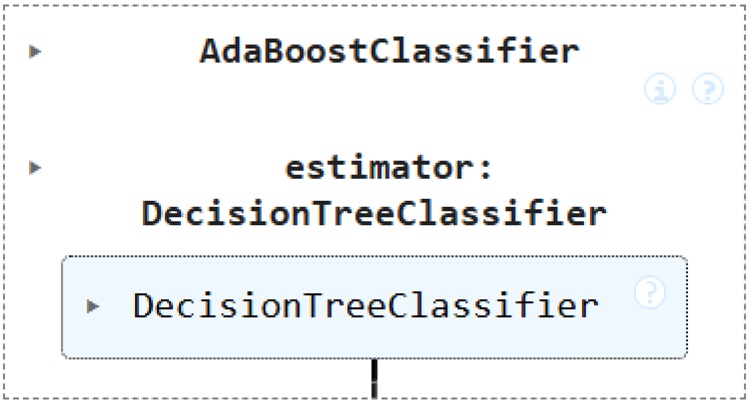


Figure 48: Tuned AdaBoost trained on original data

	Accuracy	Recall	Precision	F1
0	0.747	0.882	0.772	0.823

Table 37: Training performance of the tuned AdaBoost trained on original data

	Accuracy	Recall	Precision	F1
0	0.742	0.871	0.772	0.818

Table 38: Validation performance of the tuned AdaBoost trained on original data

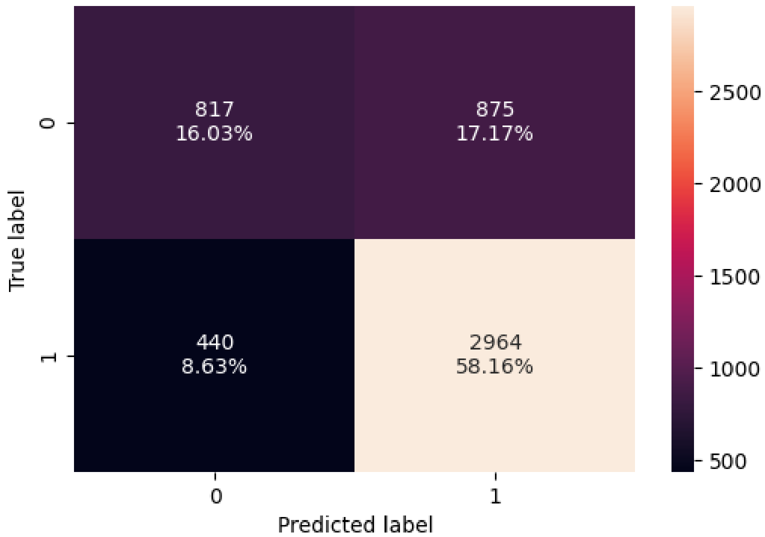


Figure 49: Confusion matrix of the tuned AdaBoost trained on original data

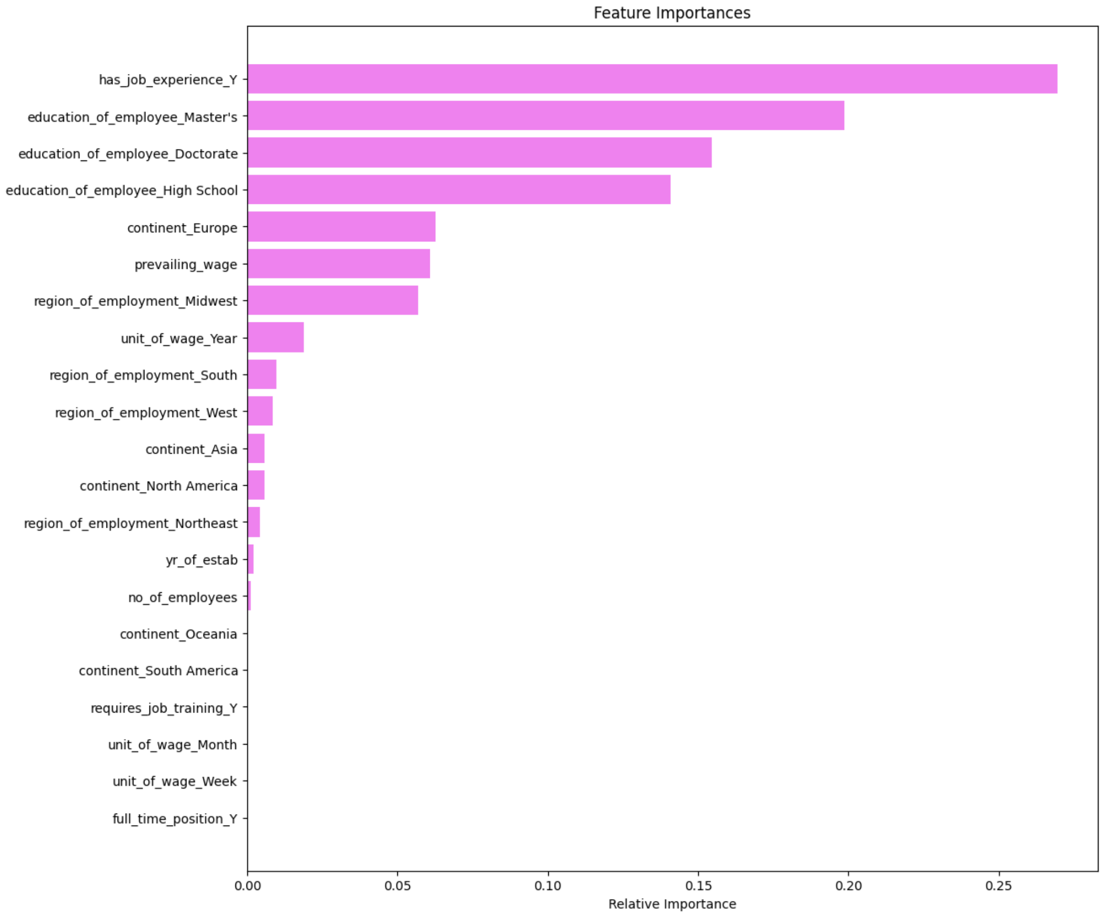


Figure 50: Feature importances of the tuned AdaBoost trained on original data

Observations

- The performance of the AdaBoost model gives approximately the same F1 score before and after hyperparameter tuning.
- In terms of precision and accuracy, the tuned model has a slightly better score. Recall is however better in the default model.
- The top 5 important features here are eduction of employee (high school, Doctorate and masters), has job experience (y) and continent_Europe.
- The tuned model predicts the "True Positives" with approximately 58.16% accuracy, which is slightly lower than the default model which predicts with 58.83% accuracy.

6.5 GBM trained on oversampled dataset

The GBM model with oversampled data is tuned with the best parameter as shown in table below.

- subsample: 0.9
- n_estimators: 150
- max_features: 0.7
- learning_rate: 0.2
- init: AdaBoostClassifier (random_state=1)

The tuned GBM trained on oversampled data gives a CV score = 0.802

```
Best parameters are {'subsample': 0.9, 'n_estimators': 150, 'max_features': 0.7, 'learning_rate': 0.2, 'init': AdaBoostClassifier(random_state=1)} with C
V score=0.8022275753211767:
CPU times: total: 4.12 s
Wall time: 1min 3s
```

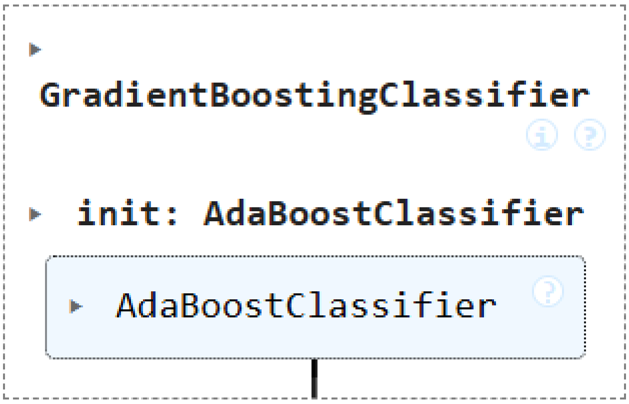


Figure 51: Tuned GBM trained on oversampled data

	Accuracy	Recall	Precision	F1
0	0.808	0.866	0.776	0.818

Table 39: Training performance of Tuned GBM trained on oversampled data

	Accuracy	Recall	Precision	F1
0	0.746	0.851	0.786	0.818

Table 40: Validation performance of Tuned GBM trained on oversampled data

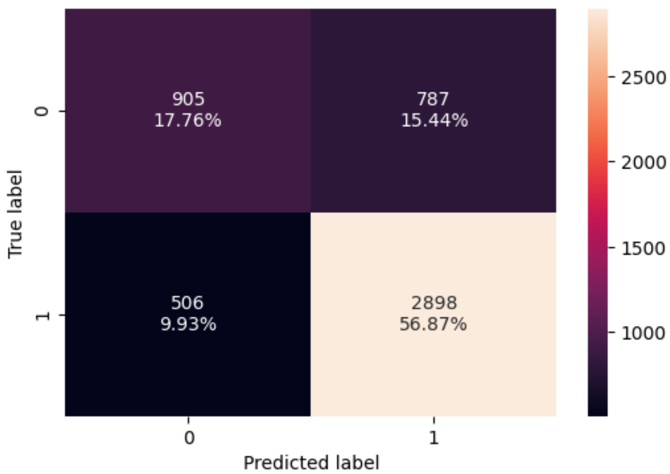


Figure 52: Confusion matrix of the Tuned GBM trained on oversampled data

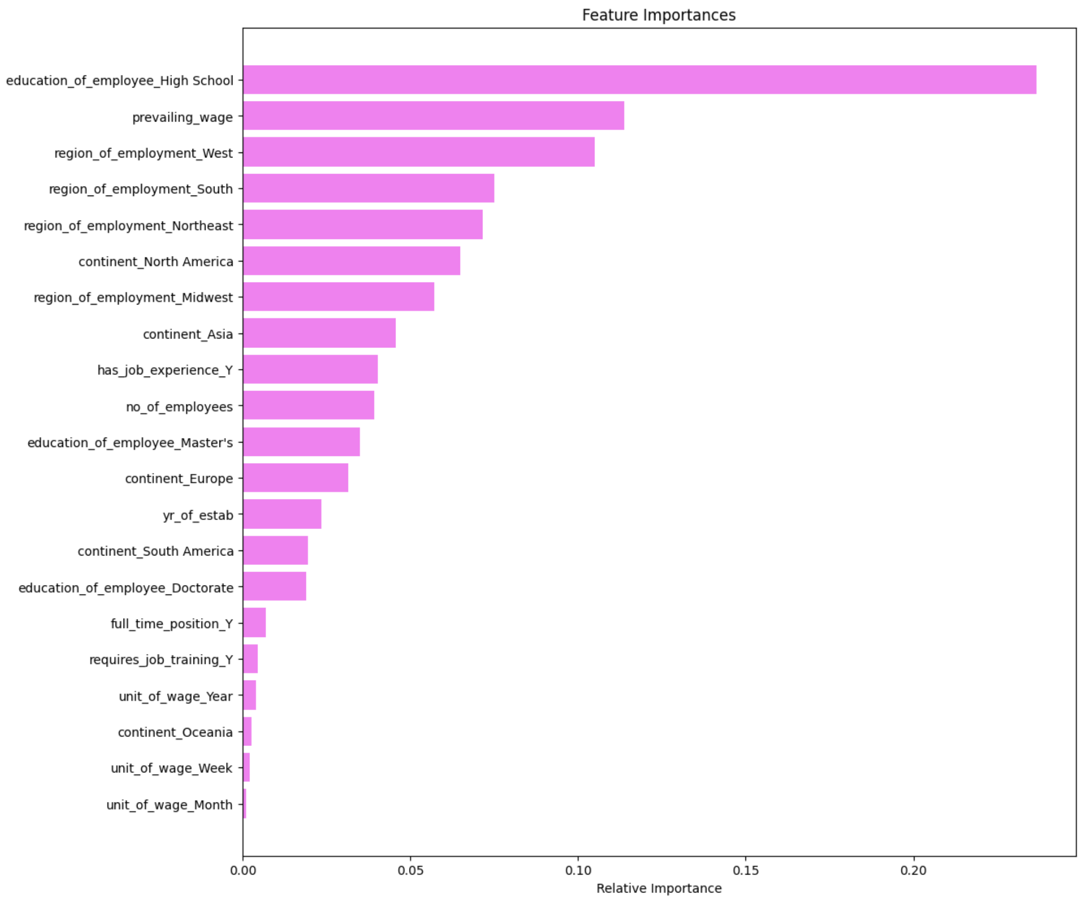


Figure 53: Feature importances of the Tuned GBM trained on oversampled data

Observations

- The performance of the GMB model with oversampled data gives approximately the same F1 score and accuracy before and after hyperparameter tuning.
- In terms of precision, the tuned model has a slightly lower score. Recall is however better in the tuned model.
- The top 5 important features here are eduction of employee (high school)), region of employment (West, South, Northeast) and prevailing wage.
- The tuned model predicts the "True Positives" with approximately 56.87% accuracy, which is slightly better than the default model which predicts with 56.53% accuracy.

6.6 XGBoost trained on an oversampled dataset

The XGBoost model with oversampled data is tuned with the best parameter as shown in table below.

- subsample: 0.9
- scale_pos_weight: 2
- n_estimators: 100
- learning_rate: 0.1
- gamma: 1

The tuned XGBoost trained on oversampled data gives a CV score = 0.814

```
Best parameters are {'subsample': 0.9, 'scale_pos_weight': 2, 'n_estimators': 100, 'learning_rate': 0.1, 'gamma': 1} with CV score=0.8141464166173762:
CPU times: total: 2.41 s
Wall time: 7.61 s
```

```
XGBClassifier

XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=None, device=None, early_stopping_rounds=None,
              enable_categorical=False, eval_metric='logloss',
              feature_types=None, feature_weights=None, gamma=1,
              grow_policy=None, importance_type=None,
              interaction_constraints=None, learning_rate=0.1, max_bin=None,
              max_cat_threshold=None, max_cat_to_onehot=None,
              max_delta_step=None, max_depth=None, max_leaves=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              multi_strategy=None, n_estimators=100, n_jobs=None,
              num_parallel_tree=None, ...)
```

Figure 54: Tuned XGBoost trained on oversampled data

	Accuracy	Recall	Precision	F1
0	0.801	0.961	0.728	0.828

Table 41: Training performance of Tuned XGBoost trained on oversampled data

	Accuracy	Recall	Precision	F1
0	0.729	0.933	0.734	0.821

Table 42: Validation performance of Tuned XGBoost trained on oversampled data

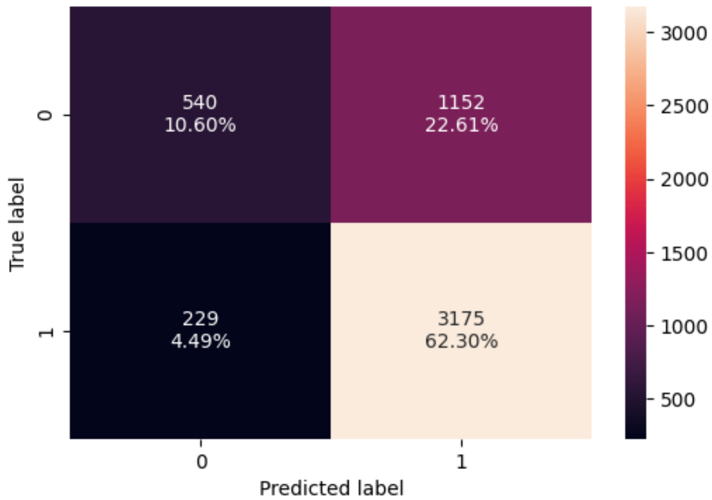


Figure 55: Confusion matrix of the Tuned XGBoost trained on oversampled data

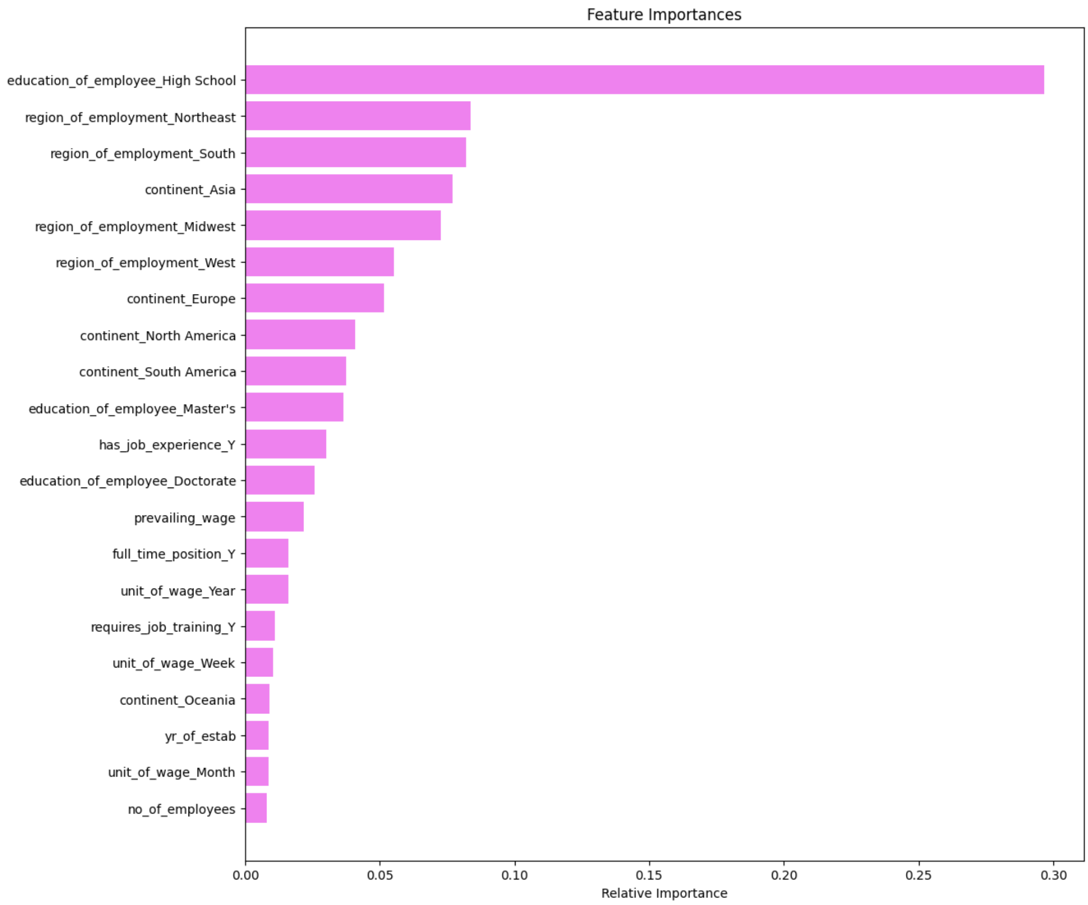


Figure 56: Feature importances of the Tuned XGBoost trained on oversampled data

Observations

- The performance of the XGB model with oversampled data gives higher F1 score and Recall after hyperparameter tuning.
- In terms of precision and accuracy, the tuned model has a slightly lower score compared to the untuned model with oversampled data
- The top 5 important features here are eduction of employee (high school)), region of employment (South, Northeast and Midwest), continent Asia.
- The tuned model predicts the "True Positives" with approximately 62.30% accuracy thus giving the best of all untuned and tuned models.

7. MODEL PERFORMANCE COMPARISON AND FINAL MODEL SELECTION

7.1 Model Performance comparison

Out of all the 18 models built, 5 models were chosen to be tuned with the best parameters obtained through RandomizedSearchCV. They are as follows,

1. GBM trained on original data
2. GBM trained on oversampled dataset
3. AdaBoost trained on original dataset
4. AdaBoost trained on an oversampled dataset
5. XGBoost trained on an oversampled dataset

Their performance on the training data set is tabulated as follows,

Training performance comparison:

	GBM with original data	GBM with Oversampled data	AdaBoost with original data	AdaBoost with Oversampled data	XGB with Oversampled data
Accuracy	0.750	0.808	0.747	0.786	0.801
Recall	0.899	0.866	0.882	0.854	0.961
Precision	0.767	0.776	0.772	0.752	0.728
F1	0.828	0.818	0.823	0.800	0.828

Table 43: Training performance comparison of all the 5 tuned models

Their performance on the validation data set is tabulated as follows,

Validation performance comparison:

	GBM with original data	GBM with Oversampled data	AdaBoost with original data	AdaBoost with Oversampled data	XGB with Oversampled data
Accuracy	0.749	0.746	0.742	0.742	0.729
Recall	0.897	0.851	0.848	0.848	0.933
Precision	0.767	0.786	0.783	0.783	0.734
F1	0.827	0.818	0.814	0.814	0.821

Table 44: Validation performance comparison of all the 5 tuned models

Observation

- It is observed that the Tuned GBM with original data gives the best F1_score of all the tuned models and it shows a consistent performance on the validation set as well. Hence it is selected as the best final model to be applied on the test set

7.2 Final model selection

- It is observed that the Gradient Boosting model trained on the original data is efficient in predicting the true positives with the maximum F1_score of 0.827
- The XGBoost model trained on the oversampled data also shows a very good recall with a recall score of 0.933, thus minimizing the false negatives efficiently. This model also shows a consistent performance.
- But the tuned Gradient Boosting model maximises the "True Positives" and minimizes the "False Positives" efficiently with an Accuracy score of 0.749 and Precision score of 0.769 respectively. Hence the Tuned Gradient Boosting model trained with original data is finalized to be the best model, that can be applied on the test set.

7.3 Performance of the best model on the test set

	Accuracy	Recall	Precision	F1
0	0.737	0.892	0.758	0.819

Table 45: Test performance of the final best model

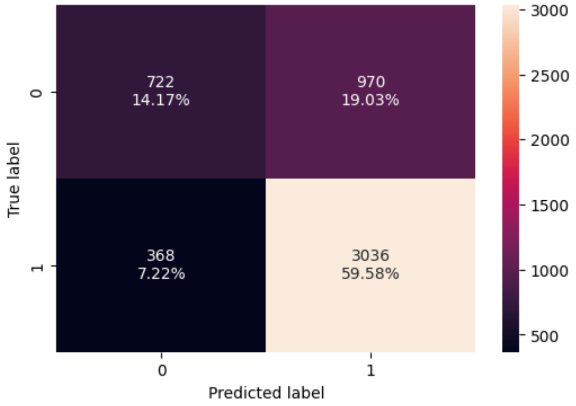


Figure 57: Confusion matrix of the final best model

The tuned model predicts the "True Positives" with approximately 60% accuracy on the test data

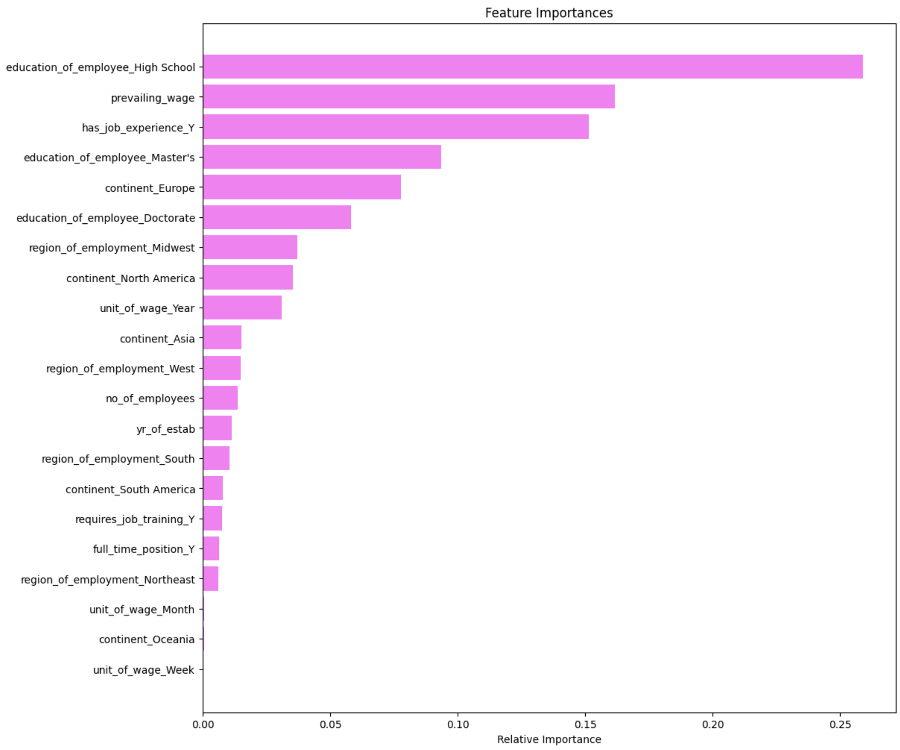


Figure 58: Feature importances of the final best model

It is observed that the education of the employee (High school), prevailing wage and previous experience (Y) are the most important features that are required to make predictions

8. ACTIONABLE INSIGHTS AND RECOMMENDATIONS

ANALYSIS INSIGHTS

- Considering the EDA and the GB Classifier model, the top 10 most important features to decide whether a visa gets certified or denied are Education of employee (High school, masters and doctorate), Has Job experience (Y), prevailing wage, Continent (Europe, North America, Asia), Unit of wage (Year), Region of employment (Midwest).
- The most important feature with a magnitude of 0.27 is the education of employee. From the EDA, it is observed that if an employee possesses a high school as the highest qualification, then that employee is more likely to be denied than accepted. Employees with doctorate degrees and with Masters have the highest chances of getting certified, so they can be encouraged to apply for visas.
- If an employee has job experience, they are more likely to be certified than those who do not have prior job experience. And those without experience are likely to be denied.
- If an employee is from Europe/ North America or Asia, he/she has higher chance of getting visa certifications than any other continent.
- Employees with a higher prevailing wage are likely to be certified, and it is noted that unit of wage is years is also an important feature. So the prevailing wage goes in line with unit of wage in years.
- If the region of employment is Midwest, employees have higher chances of being certified than denied. Employees who apply to other regions chances of being denied.
- With respect to the number of employees, the employer company with higher the number of employees, have good chance of getting approved.
- Surprisingly, other factors like requirement for job training, year of establishment and full time or part time employment, do not appear to affect the denial or certification significantly

ACTIONABLE RECOMMENDATIONS

- In order to reduce the tiresome task of screening quite a lot of visa applications, the OFLC can prioritize employees from Europe with higher degrees, who have some job experience and with the intention of getting employed in Midwest at a company with large number of employees, which uses years as the pay unit. The applications that fall under this category can be shortlisted first and then the rest can be reviewed.
- The employing companies can also be encouraged to set a standard for job application requirements based on the important features, so that they attract candidates that have better chances of getting their visas certified.
- The OFLC can implement a point based system for visa applicants, considering these important features as the decision criteria. Based on these details a score card can be generated automatically as the visa application is submitted and OFLC can begin screening based on the highest points in the score card.
- The OFLC can also organize special programs designed based on some of these important features for the respective domains for which they need immigrant workers and can make it a mandatory for US visa applicants. This also helps employers to choose talented employees who have completed or enrolled themselves in these special programmes and find suitable candidates who can contribute for both their business and country's economy.