ML-2 Coded Project Report

Prepared By: Parthasarathi Behura

CONTENTS: EasyVisa Page

LIST OF FIGURES:

LIST OF TABLES:

ML-2 Project

Business 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.

Objective

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

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

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

Data Description

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

- case_id: ID of each visa application
- continent: Information of continent the employee
- education_of_employee: Information of education of the employee
- has_job_experience: Does the employee has any job experience? Y= Yes; N = No
- requires_job_training: Does the employee require any job training? Y = Yes; N = No
- no_of_employees: Number of employees in the employer's company
- yr_of_estab: Year in which the employer's company was established
- region_of_employment: Information of foreign worker's intended region of employment in the US.
- prevailing_wage: Average wage paid to similarly employed workers in a specific occupation in the area of intended employment. The purpose of the prevailing wage is to ensure that the foreign worker is not underpaid compared to other workers offering the same or similar service in the same area of employment.
- unit_of_wage: Unit of prevailing wage. Values include Hourly, Weekly, Monthly, and Yearly.
- full_time_position: Is the position of work full-time? Y = Full-Time Position; N = Part-Time Position
- case_status: Flag indicating if the Visa was certified or denied

Imported the libraries for the Data are

- import pandas as pd
- import numpy as np
- import matplotlib.pyplot as plt
- import seaborn as sns
- from sklearn.metrics import

    f1_score,

    accuracy_score,

    recall_score,

    precision_score,

    confusion_matrix,

    roc_auc_score,

    ConfusionMatrixDisplay,

- from sklearn.model_selection import train_test_split, StratifiedKFold, cross_val_score
- from sklearn.preprocessing import StandardScaler, MinMaxScaler, OneHotEncoder
- from sklearn.impute import SimpleImputer
- from sklearn import metrics
- from imblearn.over_sampling import SMOTE
- from imblearn.under_sampling import RandomUnderSampler
- from sklearn.model_selection import RandomizedSearchCV
- pd.set_option("display.max_columns", None)

- pd.set_option("display.float_format", lambda x: "%.3f" % x)
- from sklearn.tree import DecisionTreeClassifier
- from sklearn.ensemble import (
  AdaBoostClassifier,
  GradientBoostingClassifier,
  RandomForestClassifier,
  BaggingClassifier,
- from xgboost import XGBClassifier
- from sklearn.linear_model import LogisticRegression
- pd.set_option("display.float_format", lambda x: "%.3f" % x)
- import warnings
- warnings.filterwarnings("ignore")

DATA PROCESSING:

1. There are some information about the dataset, decision makers should have a look.

- The dataset is having 25480 rows and 12 columns.
- There is a look on the 5 sample rows to check the data type.

	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 1: Top five rows of the dataset

2. While having a look on the data set information, it is found that there are 14 numerical and 5 categorial variables. The below table contains the same information.

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

Table 2: Basic information of the data type

3. Checking the duplicate values.

Ans:- There are no duplicate values present in the data.

4. Checking the missing values.

Ans:- There are no missing values present in the data.

```
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 3: Missing values in the data

5. Checking statistical summery 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 4: Statistical summary of the data

Observations:

- The mean and median values of no_of_employees are 5667 and 2109, respectively, implying a right-skewed distribution.
- The maximum value of no_of_employees is above 600000, which is quite high but possible.
- The minimum value of no_of_employees is -26, i.e., negative, which is unreasonable. The negative values should be treated as missing values.

- The oldest and newest employers have been established since (yr_of_estab) 1800 and 2016, respectively.
- The distribution of prevailing_wage is difficult to interpret at this point, because its unit varies across the rows. However, the minimum value is above zero, which is reasonable.ble.

6. Checking the statistics of the categorial columns.

	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 5: Statistical summary of the categorial

7. Inspecting the categorial columns with value counts.

```
case_id
EZYV01      0.000
EZYV16995   0.000
EZYV16993   0.000
EZYV16992   0.000
EZYV16991   0.000
            ...
EZYV8492    0.000
EZYV8491    0.000
EZYV8490    0.000
EZYV8489    0.000
EZYV25480   0.000
Name: proportion, Length: 25480, dtype: float64
```

```
--------------------------------------------------
continent
Asia            0.662
Europe          0.146
North America   0.129
South America   0.033
Africa          0.022
Oceania         0.008
Name: proportion, dtype: float64
--------------------------------------------------
education_of_employee
Bachelor's    0.402
Master's      0.378
High School   0.134
Doctorate     0.086
Name: proportion, dtype: float64
--------------------------------------------------
has_job_experience
Y   0.581
N   0.419
Name: proportion, dtype: float64
--------------------------------------------------
requires_job_training
N   0.884
Y   0.116
Name: proportion, dtype: float64
--------------------------------------------------
no_of_employees
183     0.001
854     0.001
724     0.001
766     0.001
1476    0.001
        ...
```

5876 0.000

5536 0.000

47866 0.000

4700 0.000

40224 0.000

Name: proportion, Length: 7105, dtype: float64

---

yr_of_estab

1998 0.045

2005 0.041

2001 0.040

2007 0.039

1999 0.034

...

1842 0.000

1846 0.000

1822 0.000

1810 0.000

1824 0.000

Name: proportion, Length: 199, dtype: float64

---

region_of_employment

Northeast 0.282

South 0.275

West 0.258

Midwest 0.169

Island 0.015

Name: proportion, dtype: float64

---

prevailing_wage

82560.280 0.000

122.650 0.000

60948.150 0.000

64357.580 0.000

108.120 0.000

...

25713.980 0.000

101656.640 0.000

65665.550 0.000

50.881 0.000

70876.910 0.000

Name: proportion, Length: 25454, dtype: float64

---

unit_of_wage

Year 0.901

Hour 0.085

Week 0.011

Month 0.003

Name: proportion, dtype: float64

---

full_time_position

Y 0.894

N 0.106

Name: proportion, dtype: float64

---

case_status

Certified 0.668

Denied 0.332

Name: proportion, dtype: float64

Table 6: Categorial columns with value counts.

Observations: -

- The majority of employees are from Asia.
- The majority of employees have a Bachelor's degree.
- Most of the employees have job experience.
- The vast majority of the jobs do not require training.
- The regions Northeast, South, and West need most of the employees.

- The available units for wage are Year, Hour, Week, and Month. The majority of the wage values in the data are per year.
- The vast majority of the applications are for full-time positions.
- Near 2/3 of the visa applications are certified.

8. Exploratory Data Analysis

Univariate Analysis:

Checking Continent of Origin:-

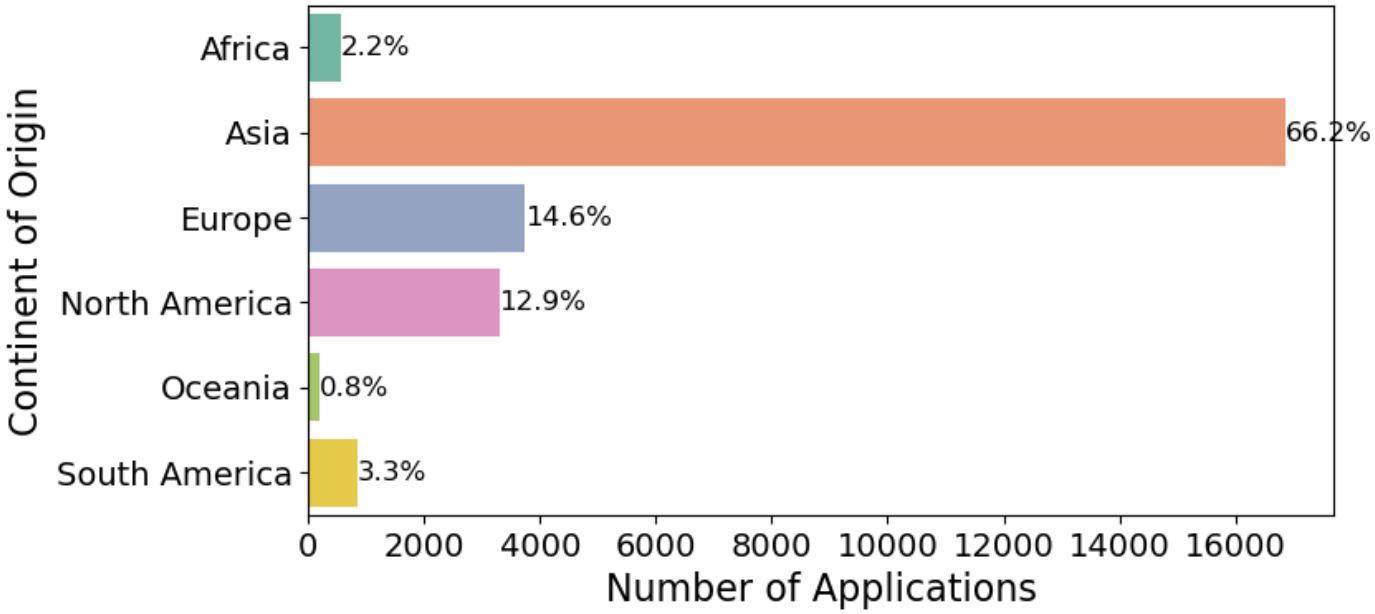


Figure-1 : Bar plot of categorial variable continent of Origin

Observations:-

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

Checking Education Level:-

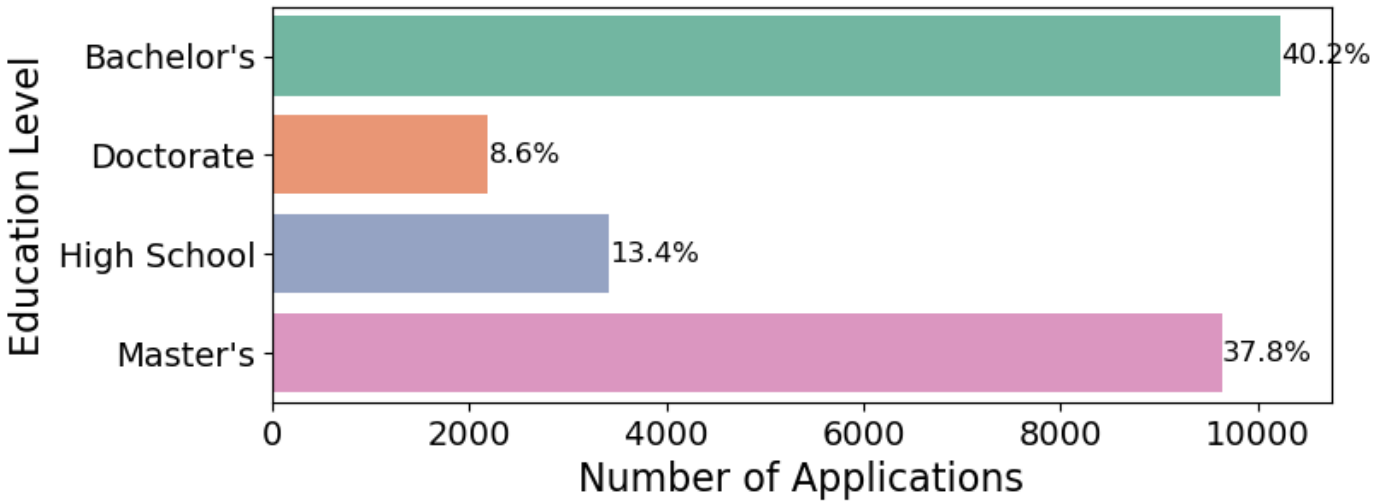


Figure-2 : Bar plot of categorial variable Education level

Observations: -

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

Checking Job Experience:-

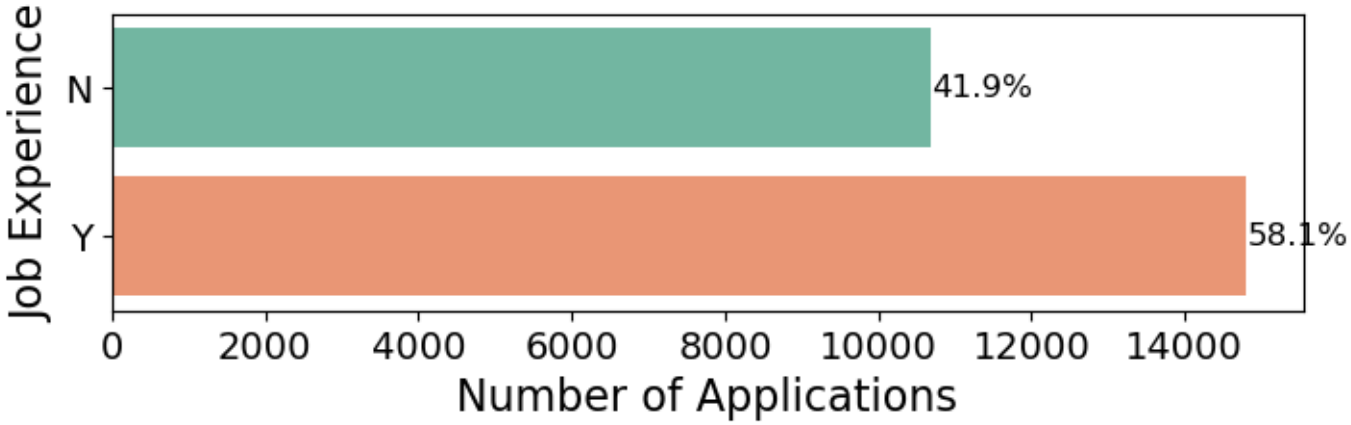


Figure-3 : Bar plot of categorial variable Job Experience

Observations:-

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

Checking Job Training Requirement:-

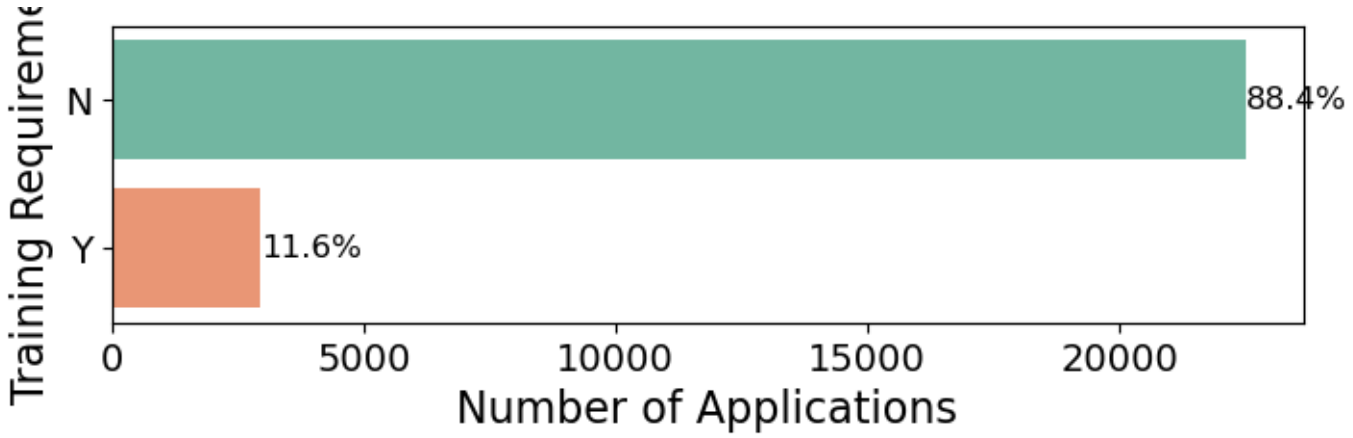


Figure-4 : Bar plot of categorial variable Job Training Requirement

Observations:-

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

Checking Employer Region:-

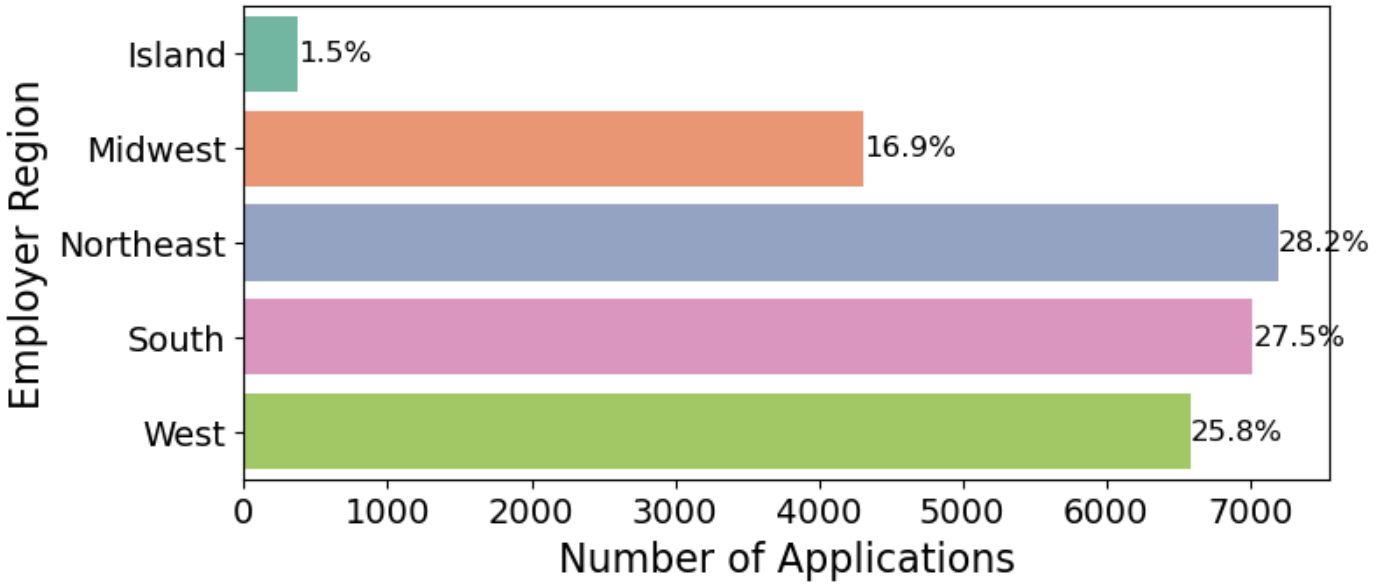


Figure-5 : Bar plot of categorial variable Employer Region

Observations:-

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

Checking Job Type:-

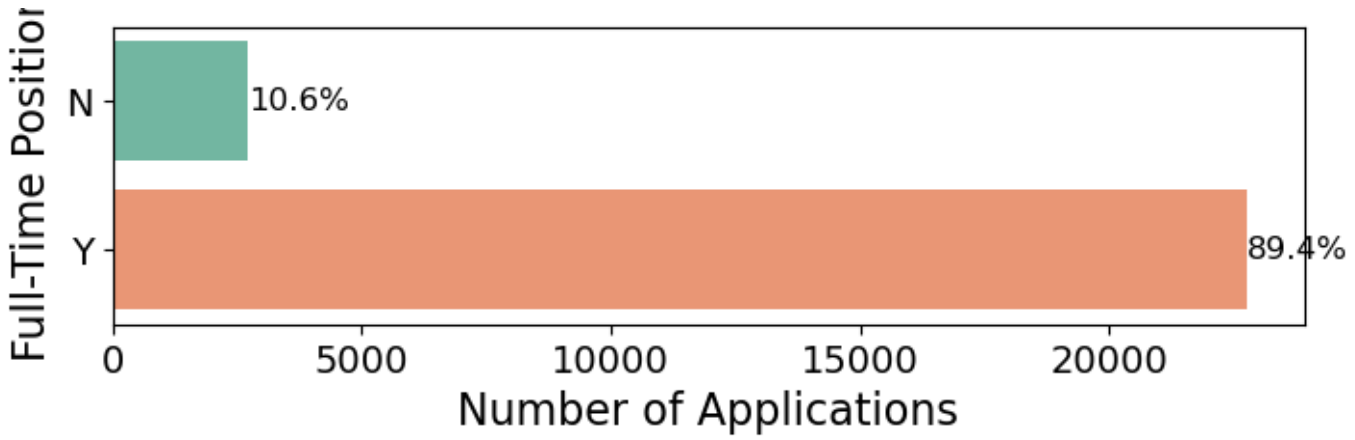


Figure-6 : Bar plot of categorial variable Job Type

Observations:-

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

Checking Wage Unit:-

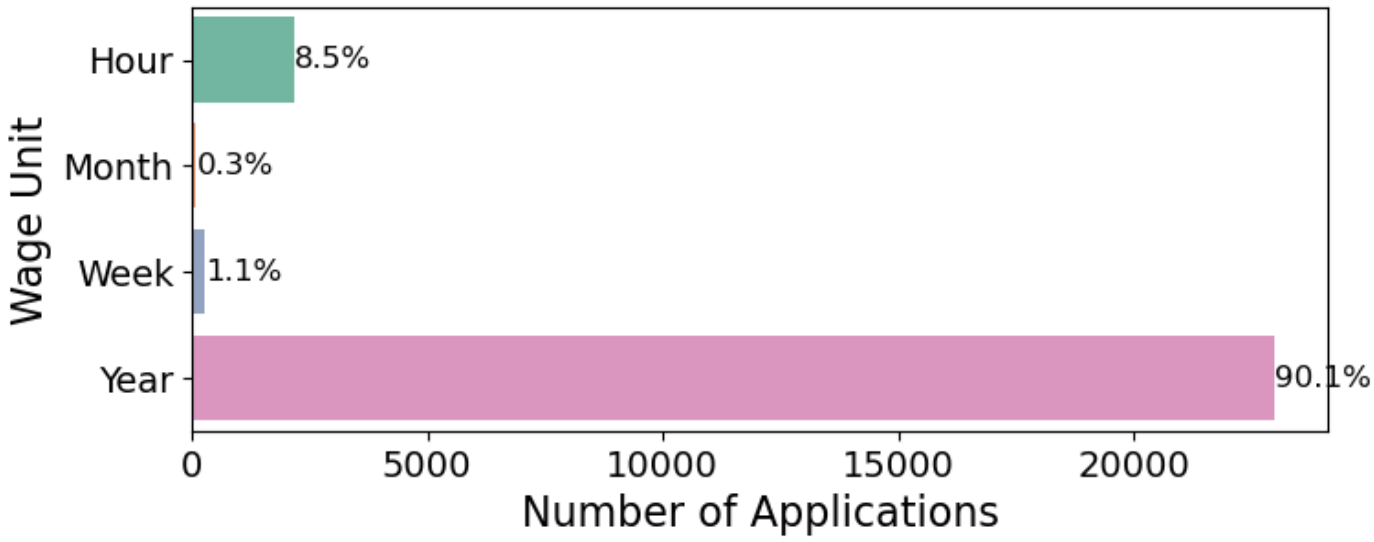


Figure-7 : Bar plot of categorial variable Wage Unit

Observations:-

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

Checking Case Status:-

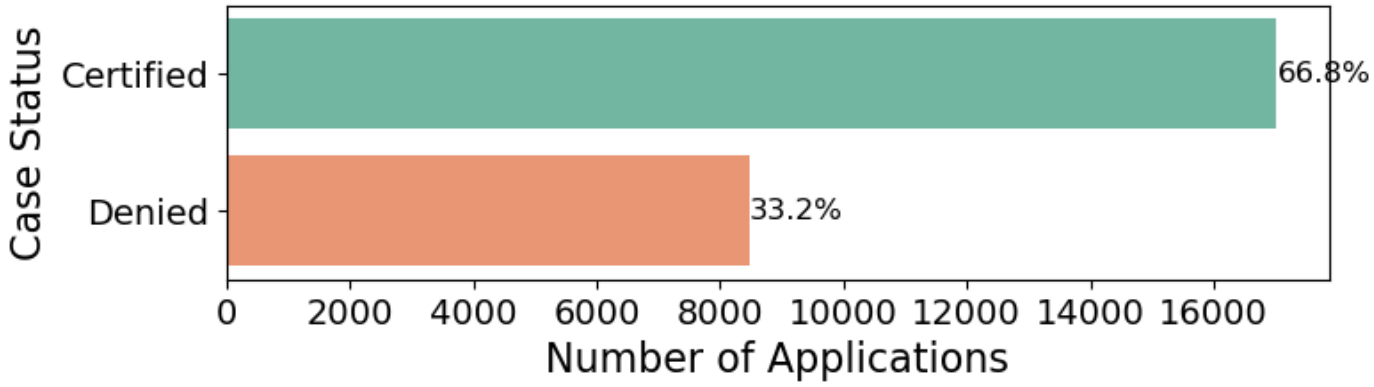


Figure-8 : Bar plot of categorial variable case status

Observations:-

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

Checking No of Employees:-

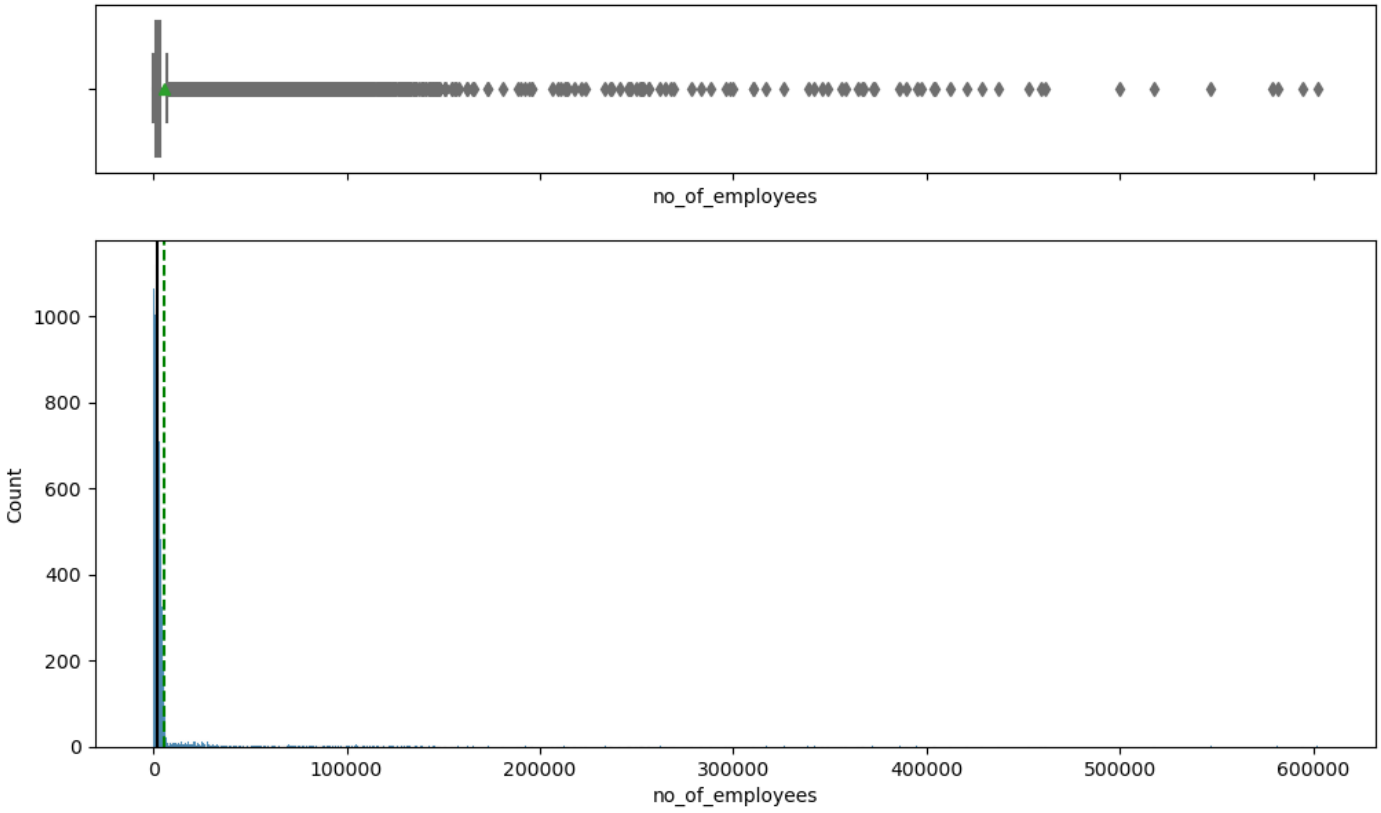


Figure-9 : Hist-Box plot of numerical variable no of employees

Observations:-

- There is a large variation in the number of employees of the employers.

- The distribution is highly right-skewed
- Not all the detected outliers per 1.5-IQR rule shall be treated as outliers, because, in 2016, there existed employers in the United States that actually had hundreds of thousands of employees. Here, per the shown distribution, a cut-off value of 450000 is considered for the number of employees.

Bivariate Analysis:-

Case Status vs. Continent of Origin:-

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

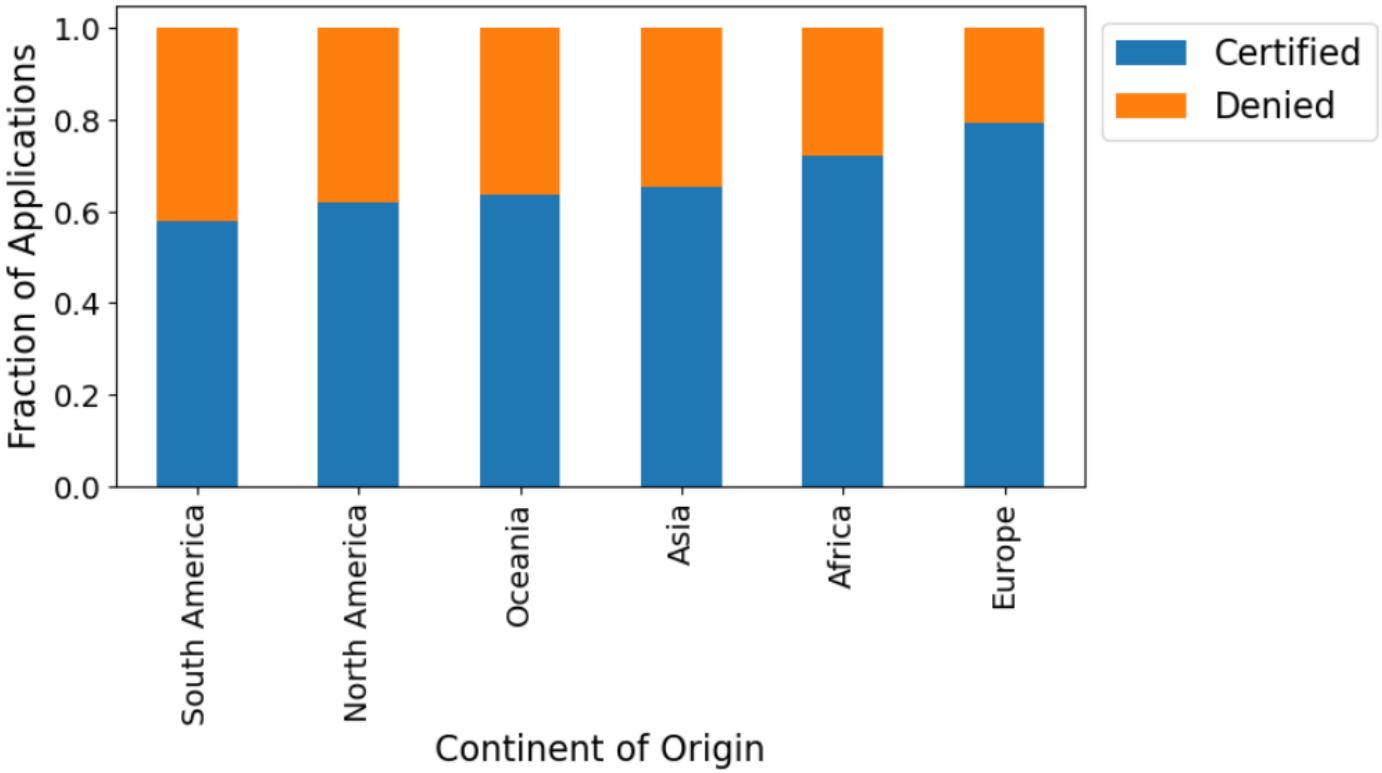


Figure-10 : Bar plot of Continent vs Case status

Observations:-

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

Case Status vs. Education Level:-

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

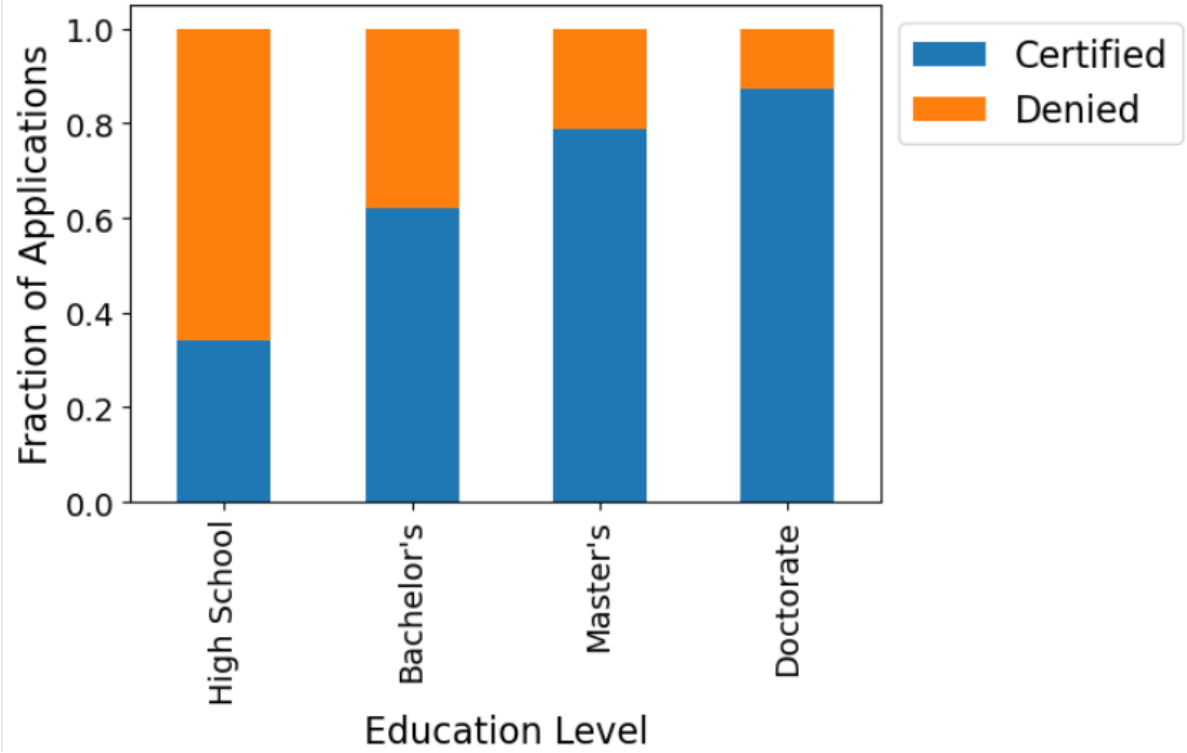


Figure-11 : Bar plot of Education level vs Case status

Observations:-

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

Case Status vs. Job Experience:-

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

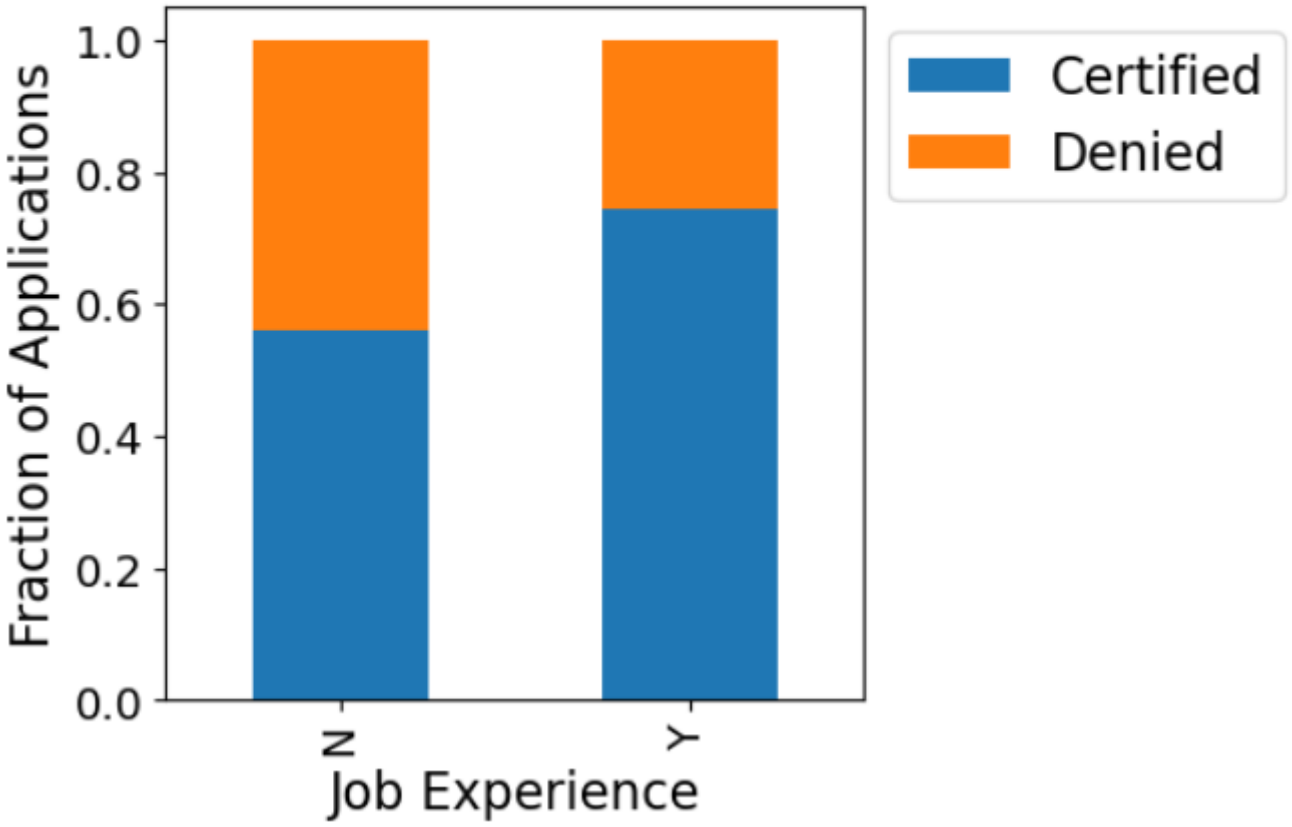


Figure-12 : Bar plot of Job Experience vs Case status

Observation:-

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

Case Status vs. Job Training Requirement:-

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

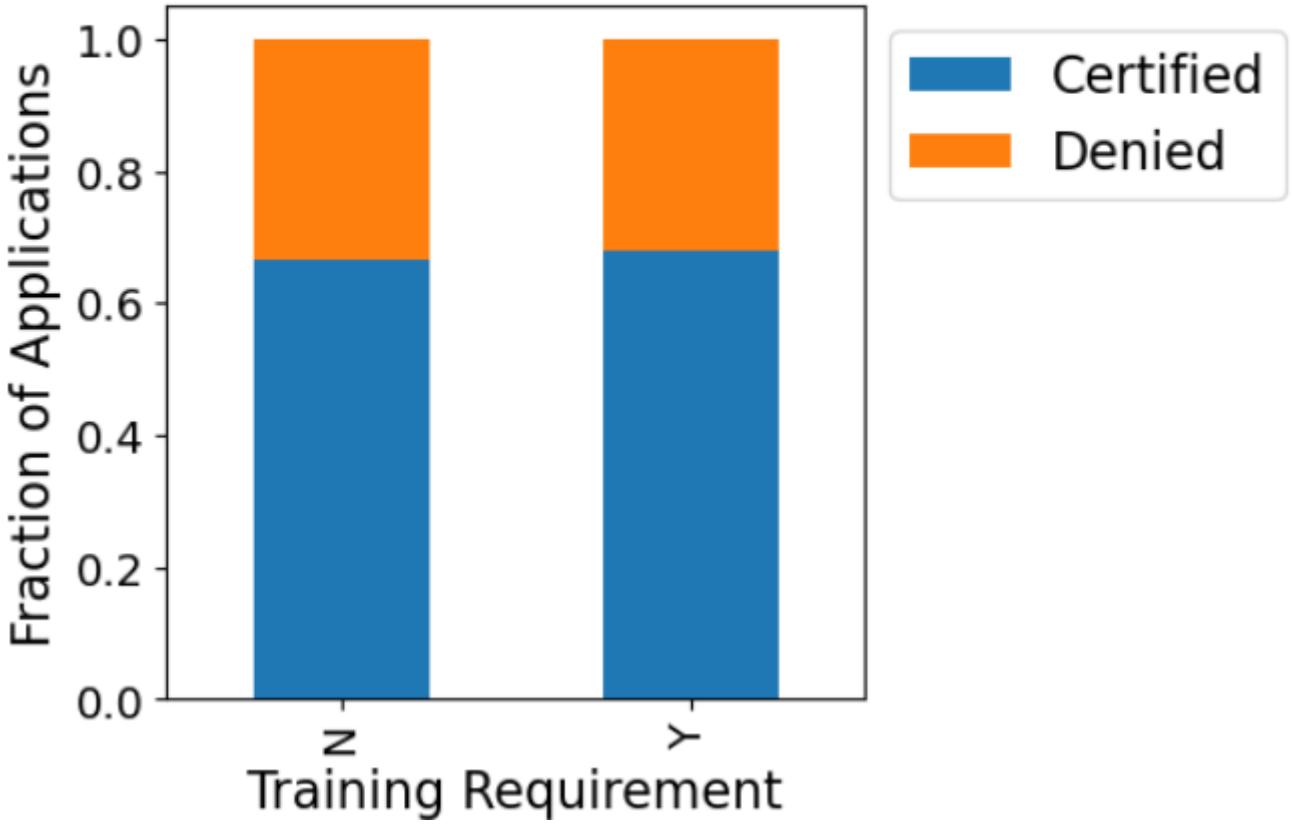


Figure-13 : Bar plot of Job Experience vs Case status

Observations:-

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

Case Status vs. Employer Region:-

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

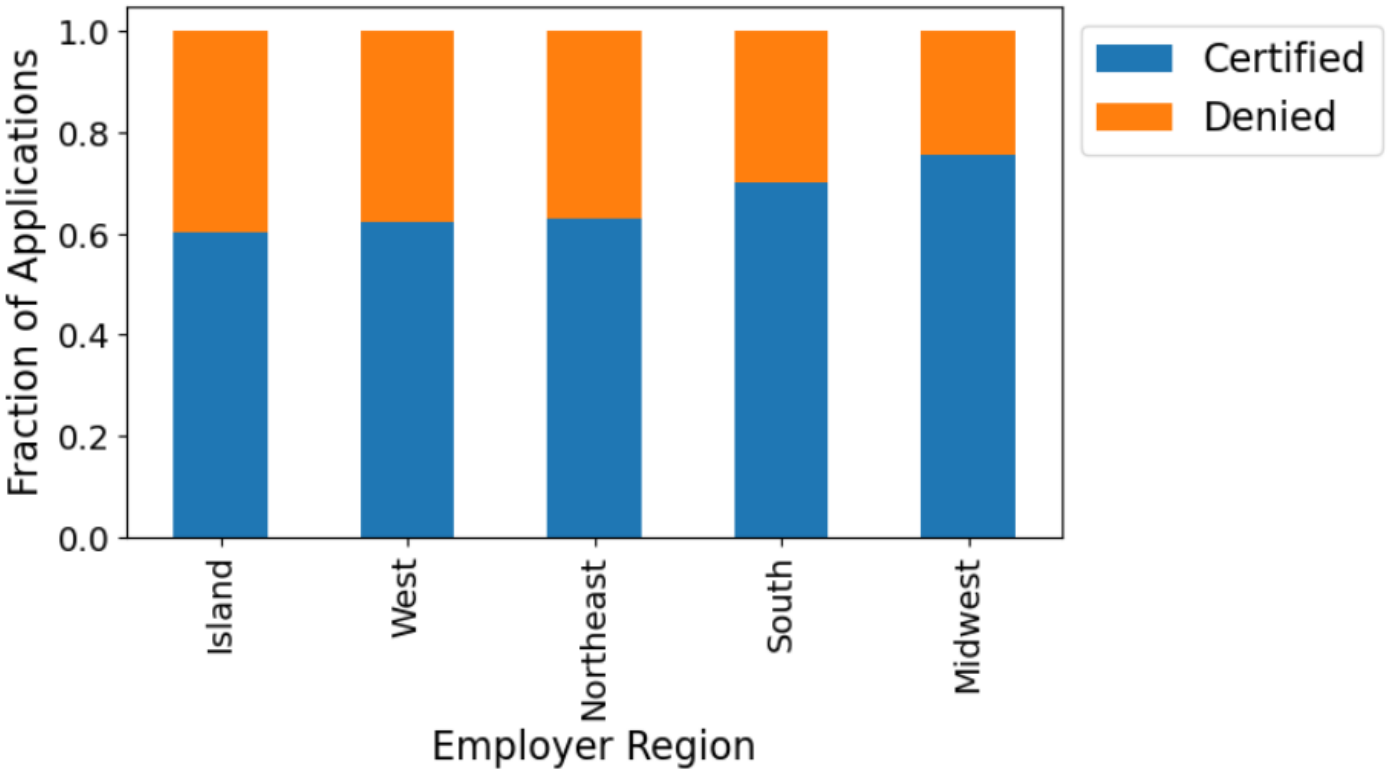


Figure-14 : Bar plot of Job Experience vs Case status

Observations:-

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

Case Status vs. Position Type:-

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

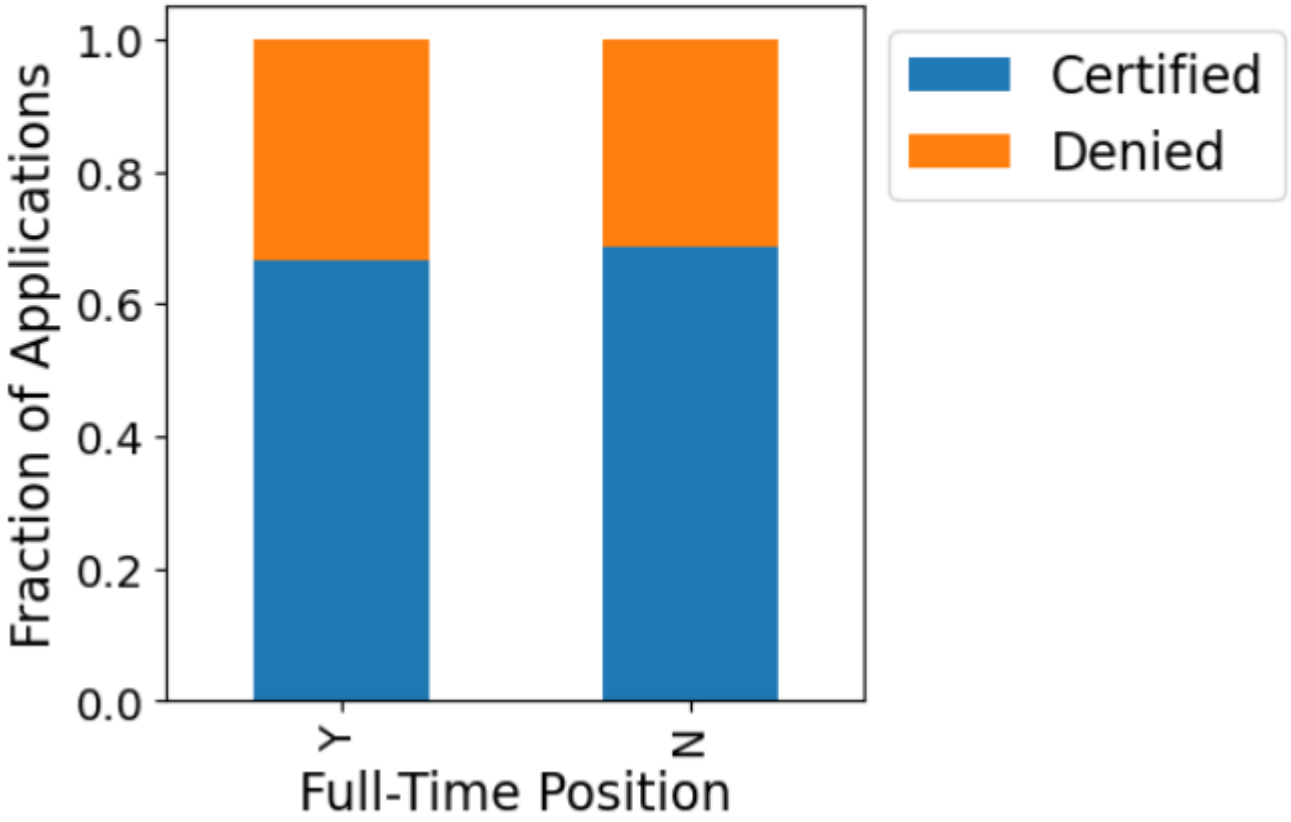


Figure-15 : Bar plot of Position Type vs Case status

Observations:-

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

Case Status vs. Wage Unit:-

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

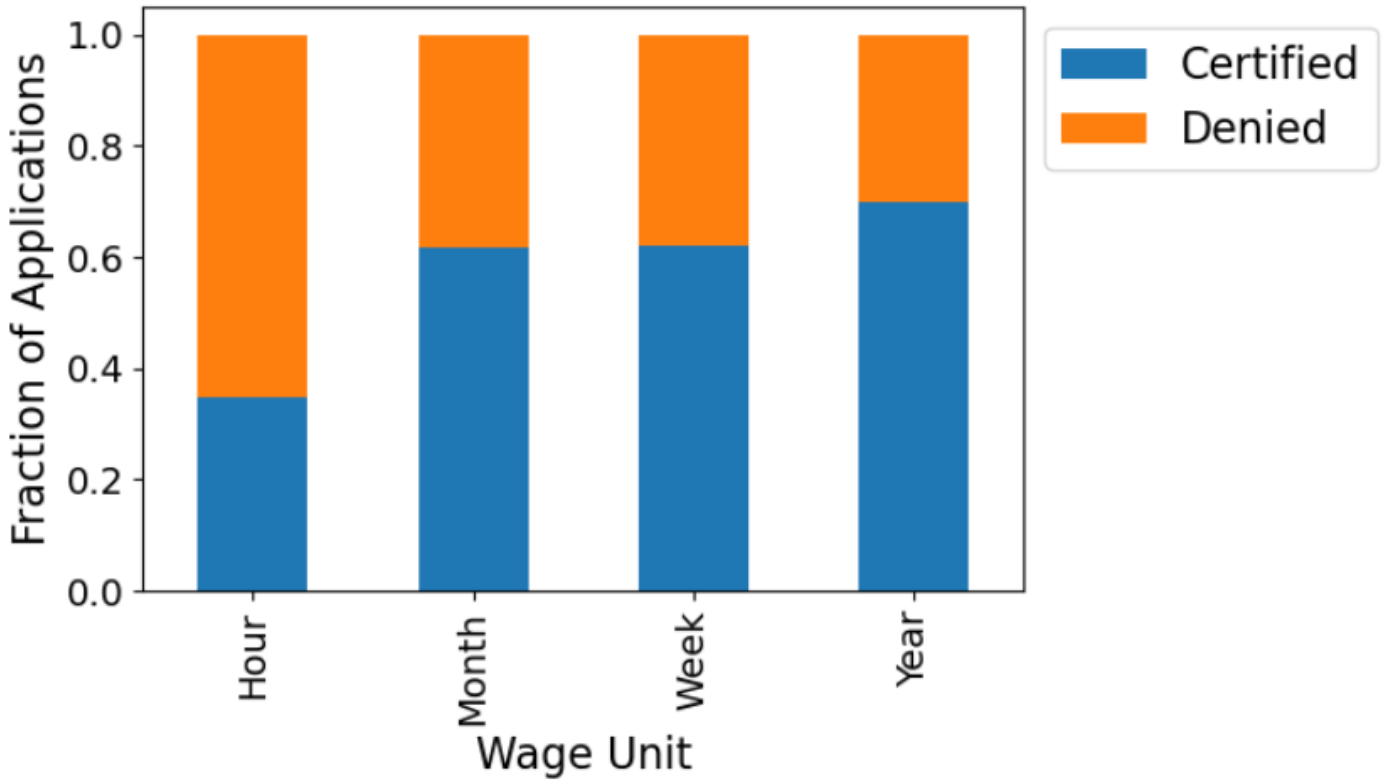


Figure-16 : Bar plot of Wage Unit vs Case status

Observations:-

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

Case Status vs. No of Employees:-

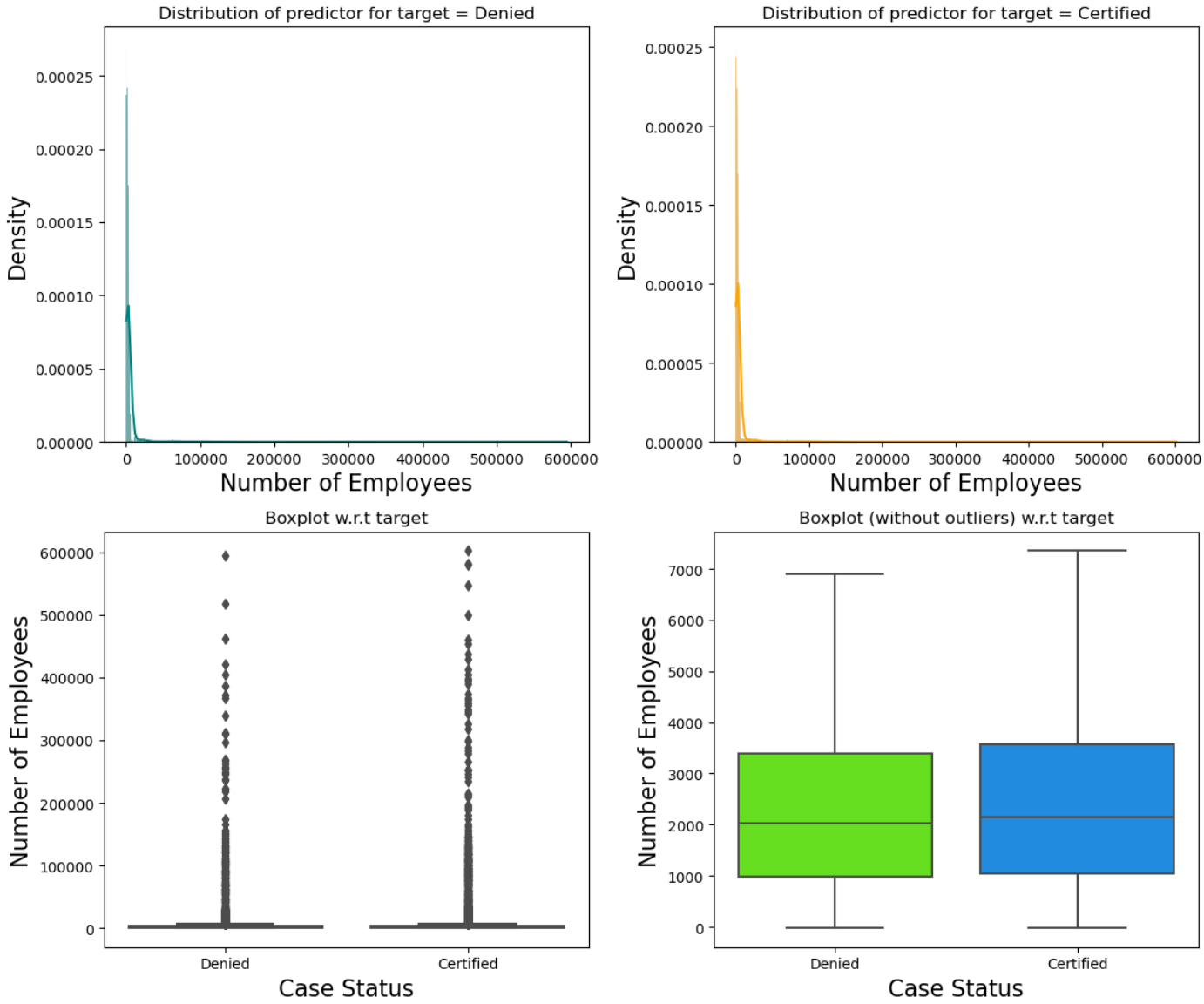


Figure-17 : Distribution plot of No of Employees vs Case status

Observations:-

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

Training Requirement vs. Job Experience:-

Figure-18 : Heat Map of Training Requirement vs. Job Experience

Observations:-

- Reasonably, a higher percentage of the applicants who have no job experience require job training than the applicants who have job experience (16% vs. ~9%).

Training Requirement vs. Continent:-

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

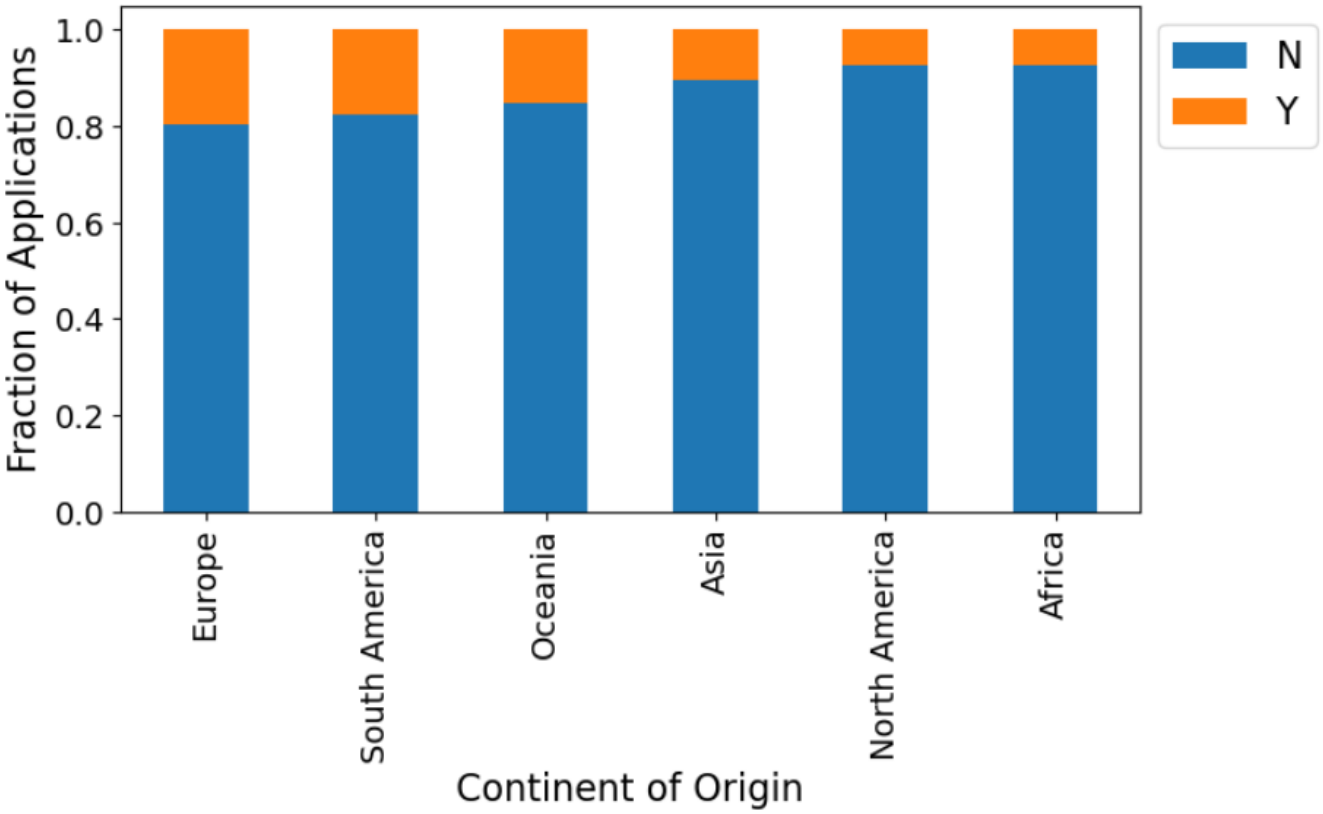


Figure-19 : Bar plot of Training Requirement vs. Continent

Observations:-

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

Data Preprocessing:-

Treatment of Missing Values:-

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

```
There are 33 rows with non-positive number of employees.
The new minimum number of employees is 12.
```

Feature Engineering:-

The feature yr_of_estab is converted to yrs_snc_estab, containing the years since establishment. Also, to make the prevailing wages (in the column prevailing_wage) interpretable across the rows, they are all transformed into an equivalent hourly wage and are saved in a new column, hourly_wage. The columns yr_of_estab and prevailing_wage are dropped subsequently.

	count	mean	std	min	25%	50%	75%	max
no_of_employees	25480.000	5669.798	22877.372	12.000	1028.000	2109.000	3504.000	602069.000
yrs_snc_estab	25480.000	36.590	42.367	0.000	11.000	19.000	40.000	216.000
hourly_wage	25480.000	94.903	278.177	0.048	22.648	39.827	60.012	7004.399

Table 7: Statistics of new Dataset

Observations:-

- The mean and median values of yrs_snc_estab are ~37 and 19 years, respectively. The oldest employer was established 216 years before the data collection.
- The minimum and maximum values of hourly_wage are 0.05 and ~7004 (probably in dollars), respectively, so the variation of this variable is very large. The mean hourly wage is ~95.

Detection and Treatment of Outliers:-

Initially, the 1.5-IQR rule is used to detect potential outliers. However, it is noted that all the values detected as outlier by this method are not always outliers.

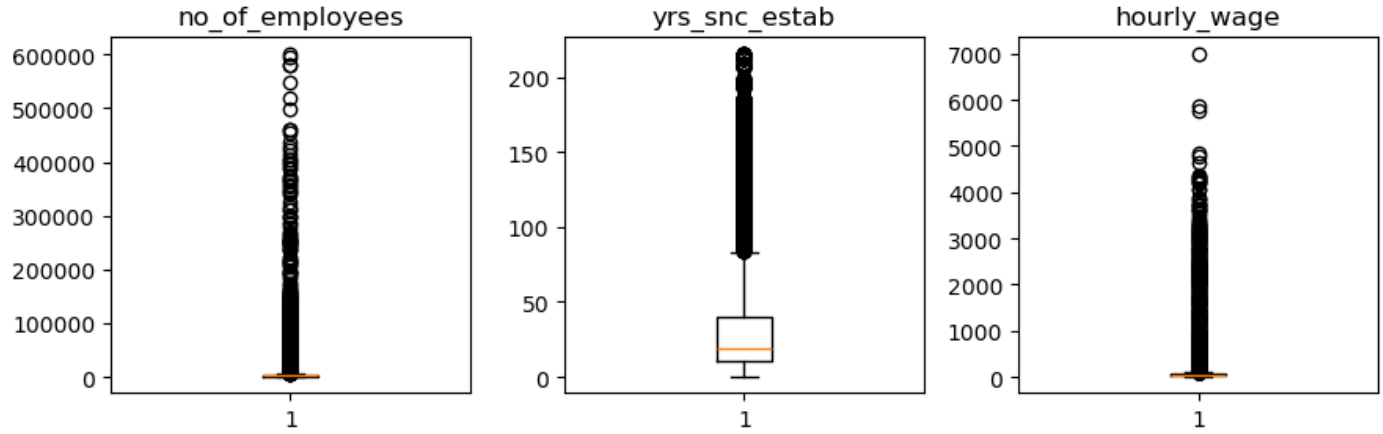


Figure-20 : Box plot for Outlier detection

Observations:-

Given the discussions provided in the initial EDA section, not all the outliers detected based on the 1.5-IQR rule are actual outliers. Here, merely to remove very large infrequent values, the following maximum cut-off values are considered for the above three variables:

- no_of_employees: 450000
- yrs_snc_estab: 200
- hourly_wage: 4000

Treatment of Outliers:-

The detected upper outliers are replaced with the maximum values of the respective columns in the absence of the outliers.

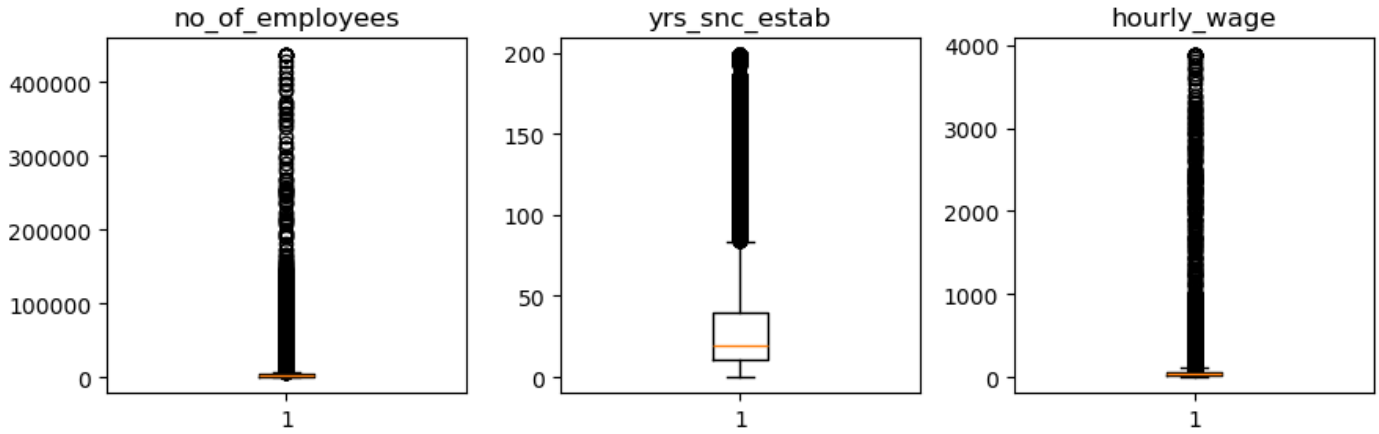


Figure-21 : Box plot for Outlier treatment

Secondary EDA:-

The focus of the secondary EDA is on the new variables created in the section Data Preprocessing.

Univariate Analysis:-

Years Since Establishment:-

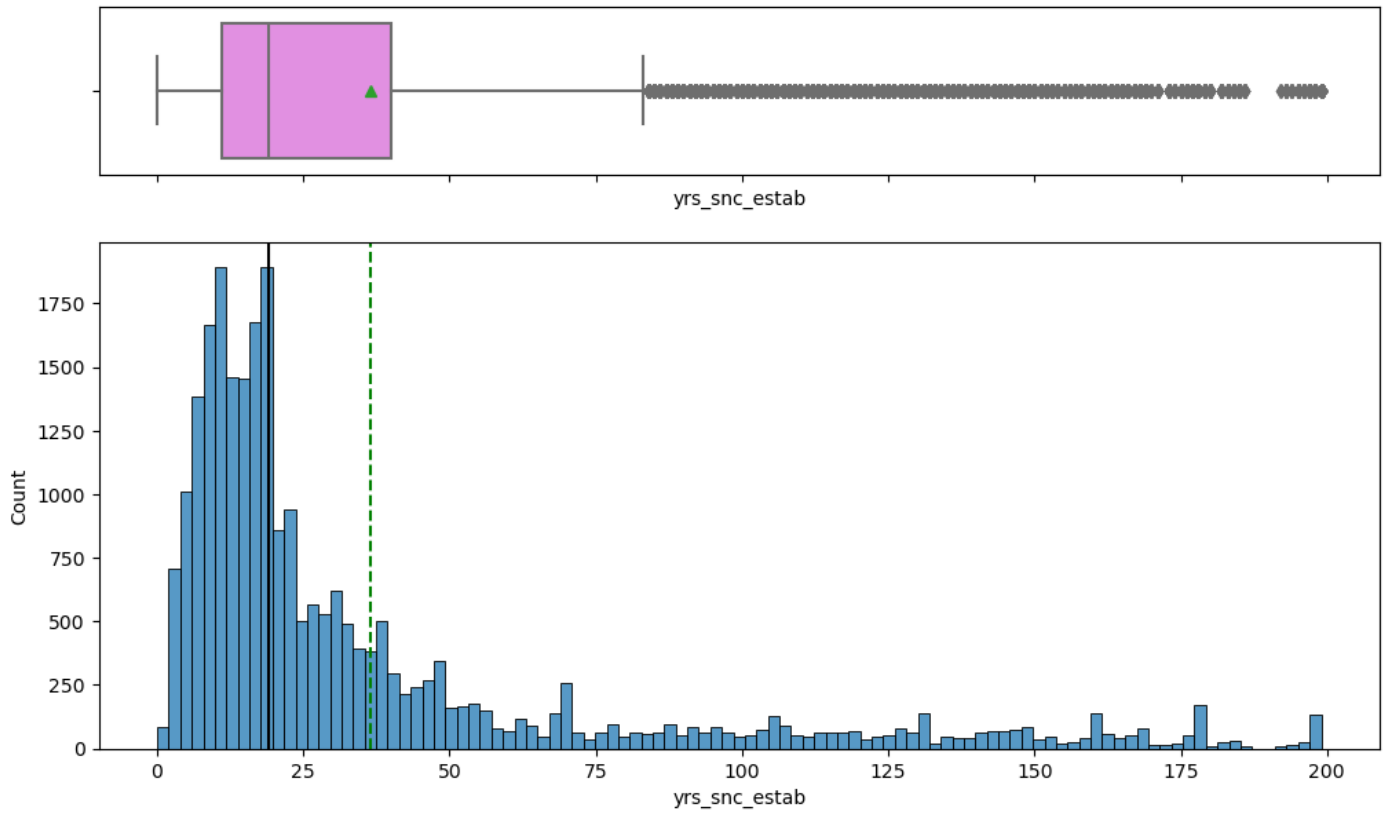


Figure-22 : Hist-Box plot for Year Since establishment

Observations:-

The distribution is quite right-skewed and the majority of the employers are less than 40 years old. As mentioned in the previous section on the treatment of outliers, the detected outliers per 1.5-IQR rule are not actually outliers.

Hourly Wage:-

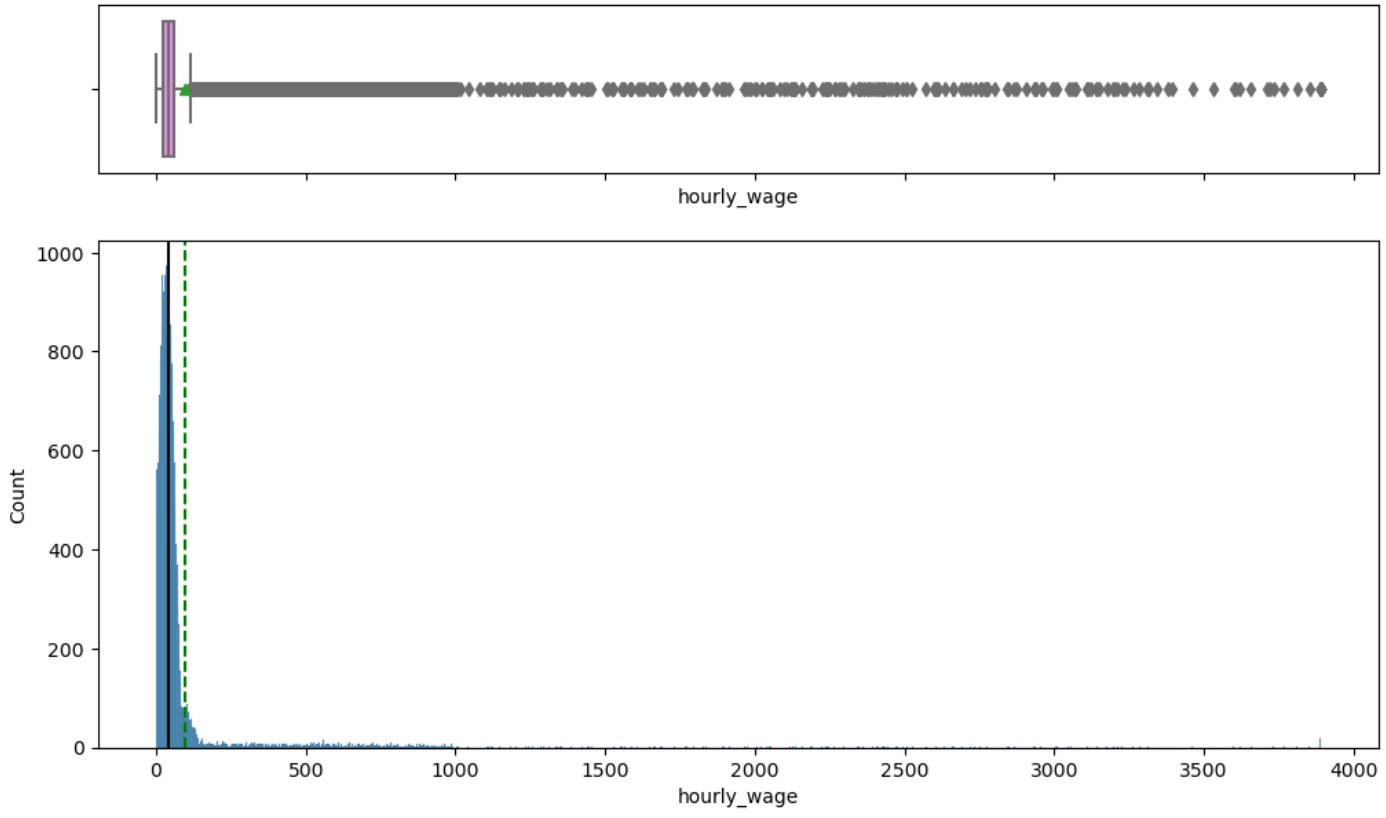


Figure-23 : Hist-Box plot for Hourly Wage

Observations:-

- The distribution of the computed equivalent hourly wage is highly right-skewed and the majority of the applications are for the positions with less than 100 (dollars) of equivalent hourly wage.
- Since there are certain positions in certain industries that are paid millions of dollars per year, the detected outliers are not actual outliers.

Bivariate Analysis:-

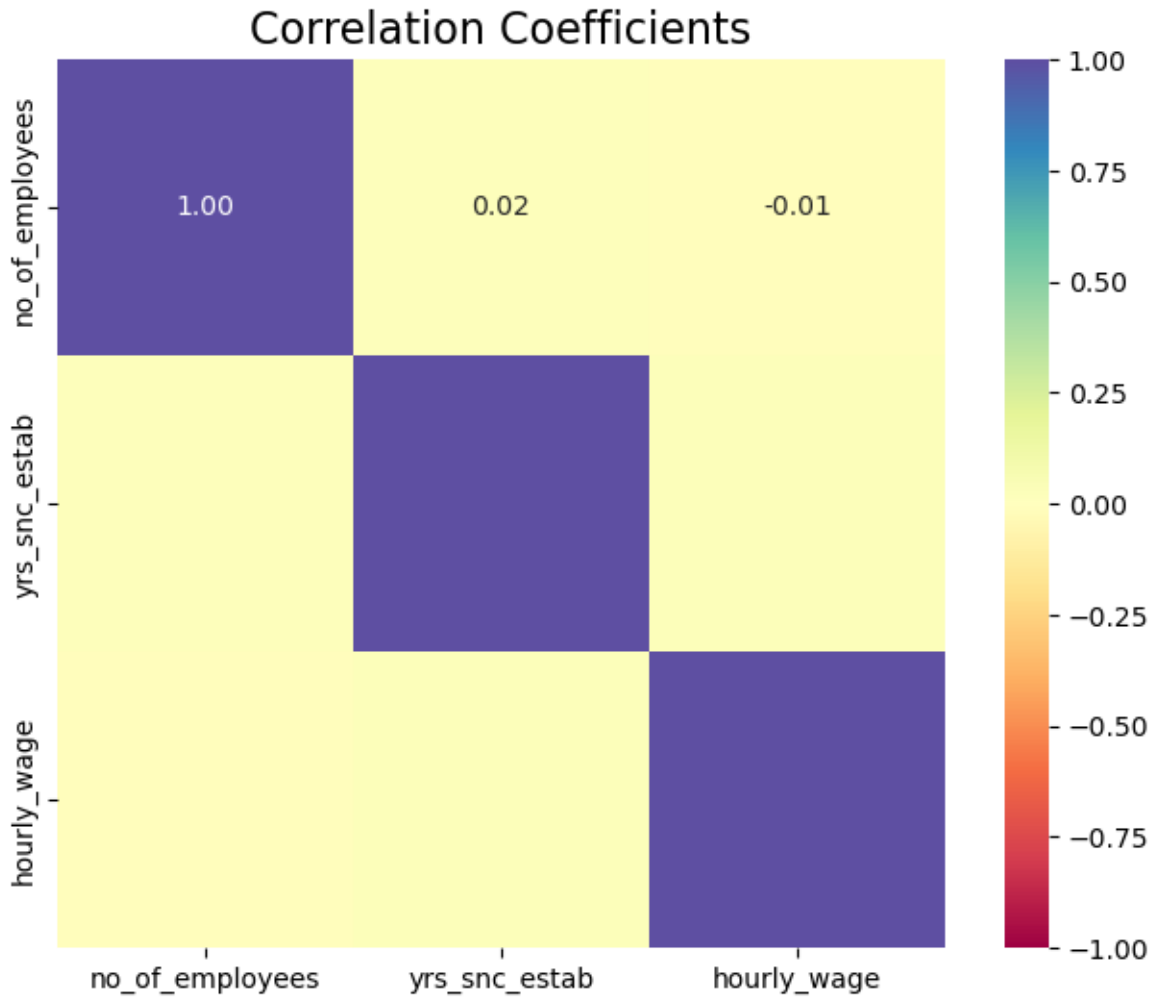


Figure-24 : Heat map for numeric variable's correlation

Observations:-

- Negligible linear correlation is observed between the numeric variables.

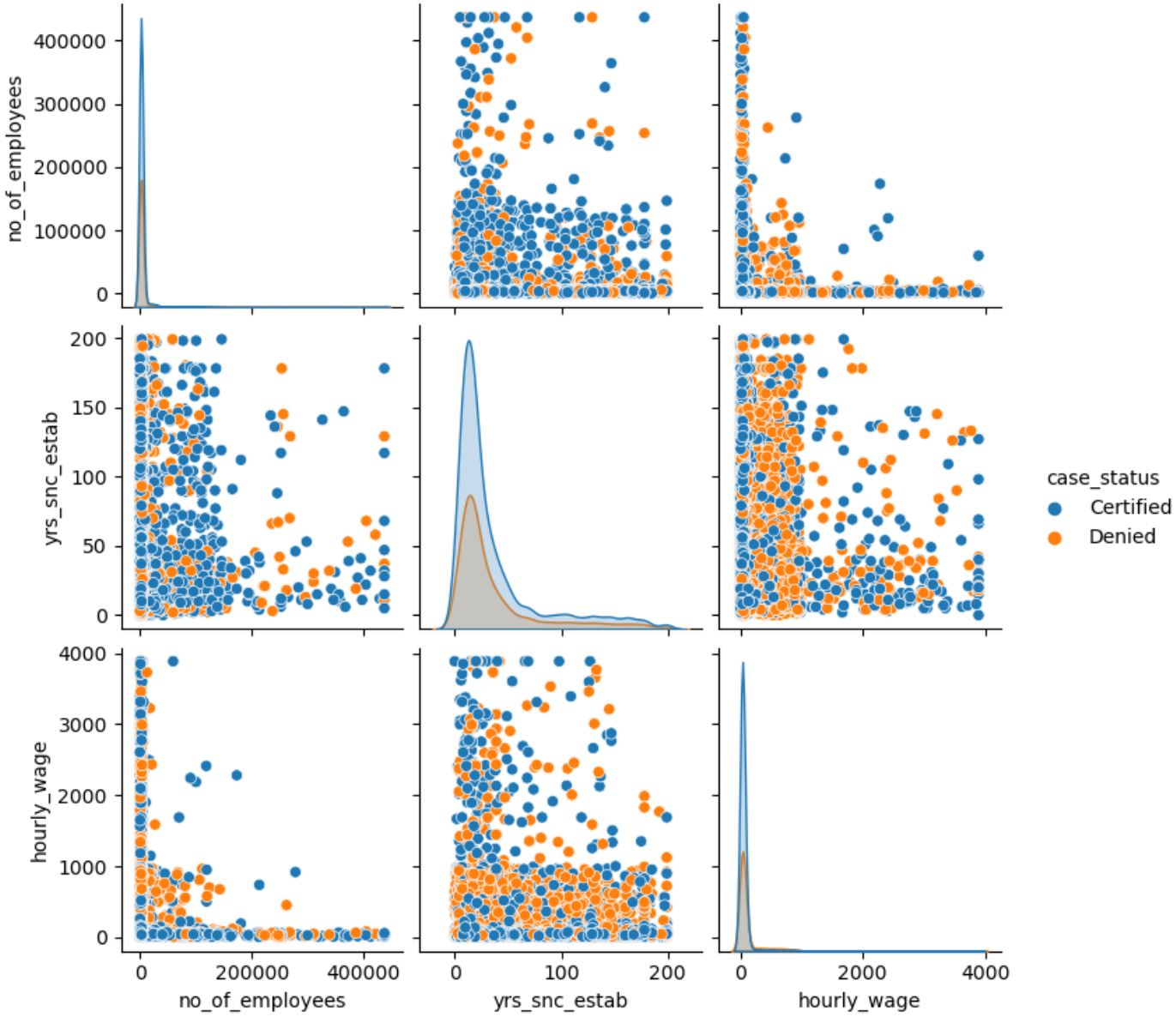


Figure-25 : Pair-plot for numeric variables vs case status

Observations:-

- No linear correlation is observed between the numeric variables.
- It is hard to identify the effects of the above variables on the visa certification likelihood.

Case Status vs Hourly Wage:-

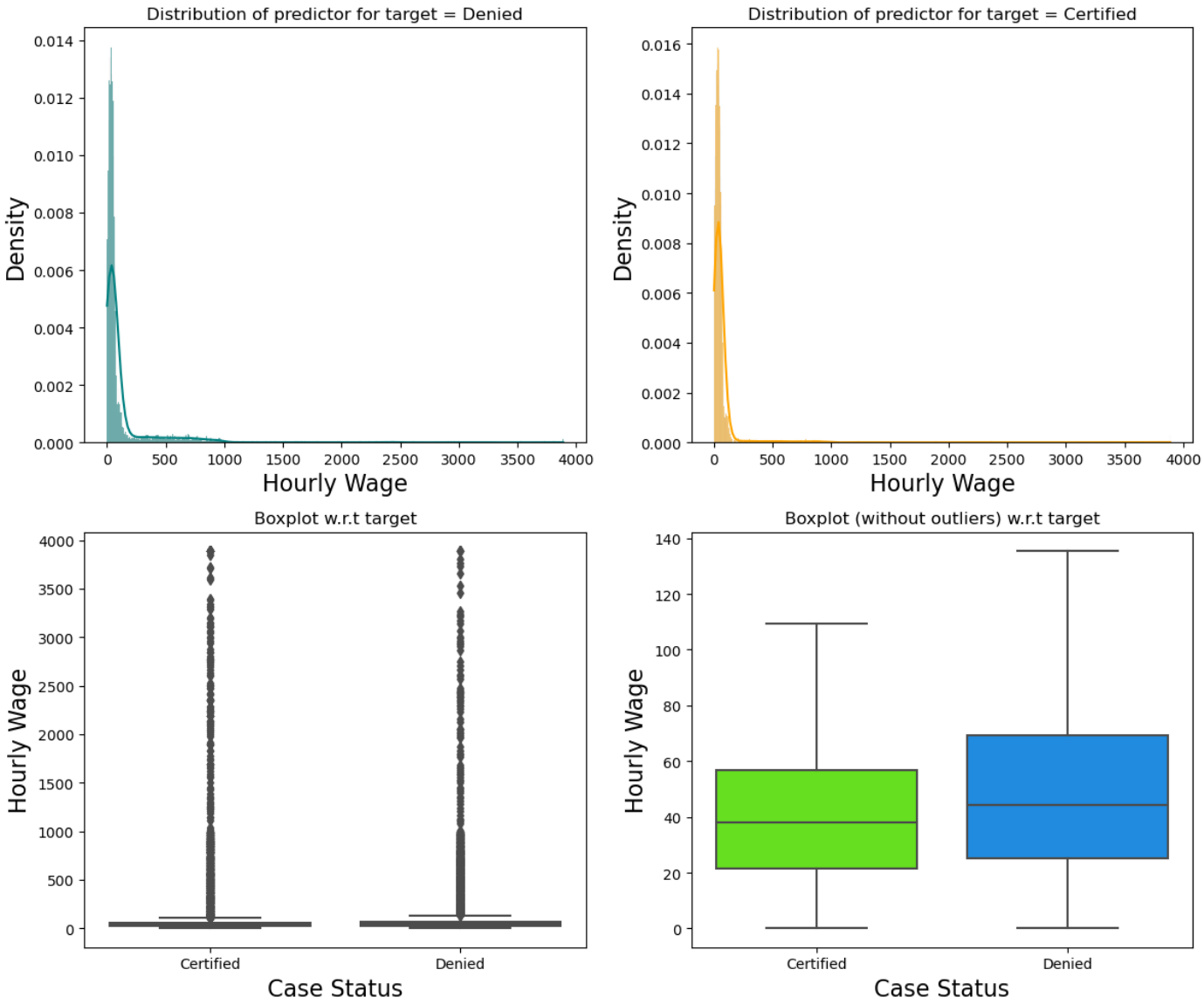


Figure-26 : Dist-plot for Hourly Wage vs case status

Observations:-

- It appears that a decrease in the equivalent hourly wage would lead to an increase in the likelihood of visa certification. This could be justified by the fact that the jobs that are paid higher could be more easily filled by American workers, making the employment of outsiders unjustifiable.

Hourly Wage vs. Education Level:-

```
(array([0, 1, 2, 3]),
 [Text(0, 0, 'High School'),
  Text(0, 1, "Master's"),
  Text(0, 2, "Bachelor's"),
  Text(0, 3, 'Doctorate')])
```

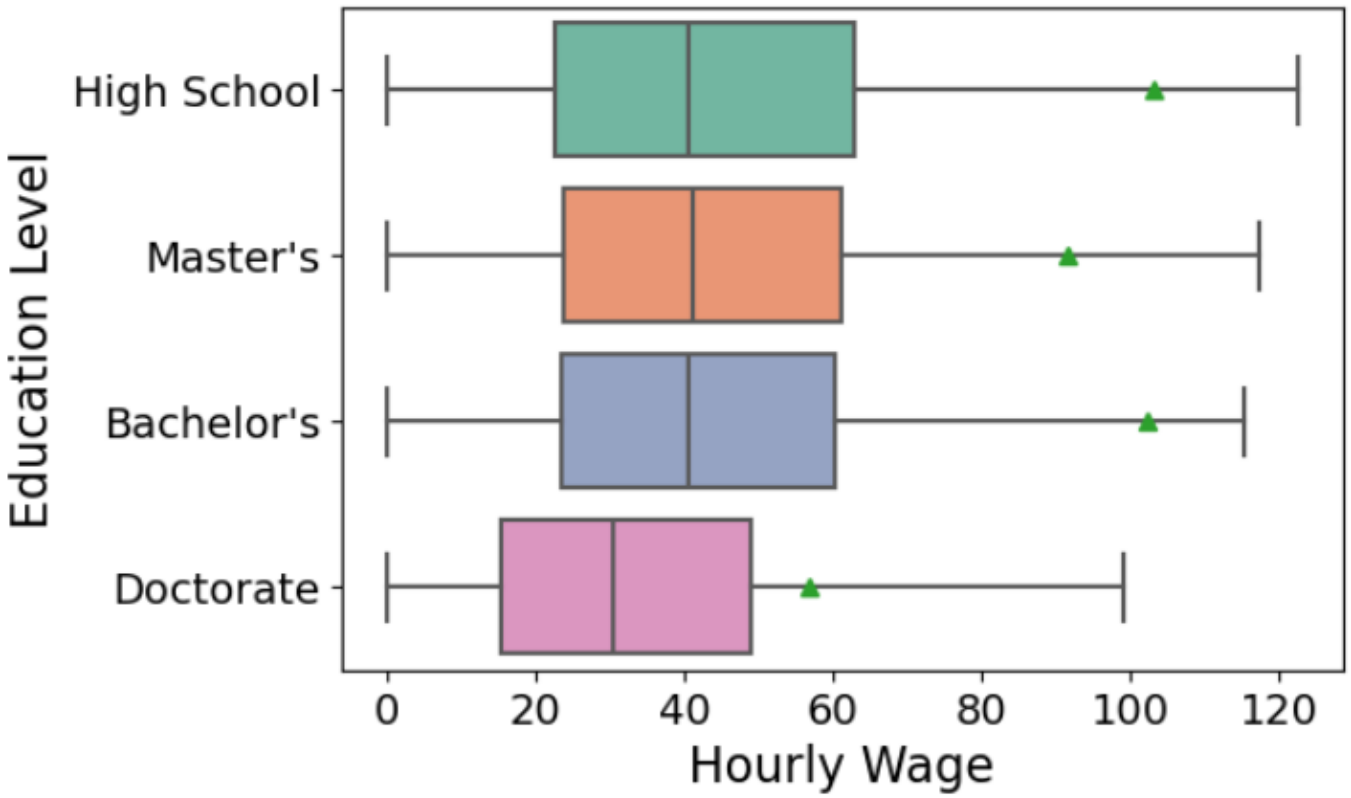


Figure-27 : Box plot for Hourly Wage vs Education Level

Observations:-

On average, the employees of less education (e.g., high school and bachelor's degree) seem to be paid more in terms of equivalent hourly wage than the employees of higher education, particularly, those of a doctorate degree.

Hourly Wage vs. Job Experience:-

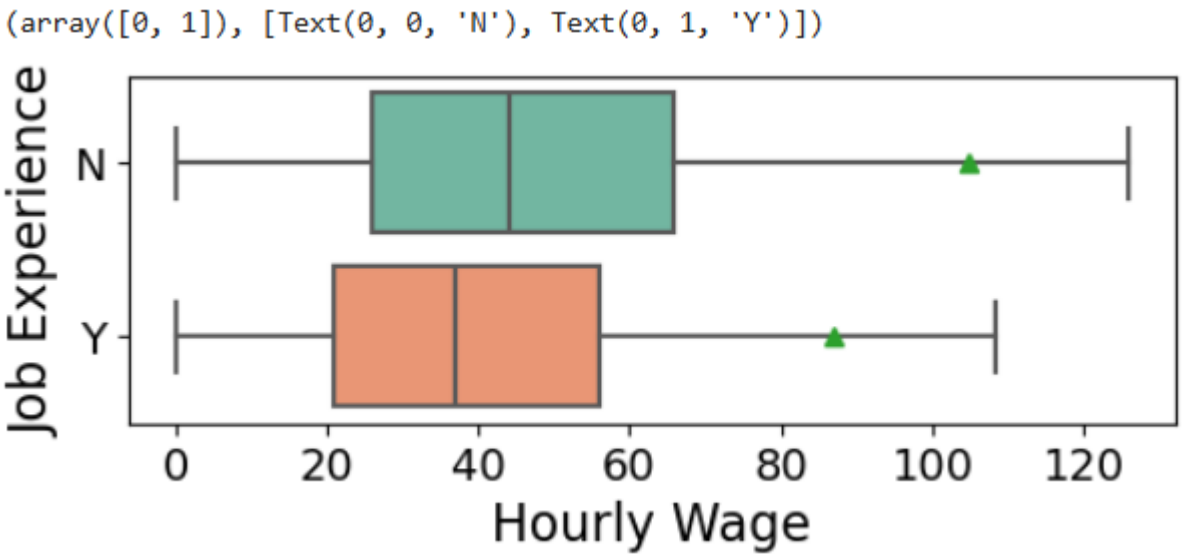


Figure-28 : Box plot for Hourly Wage vs Job Experience

Observations:-

On average, those employees that have job experience seem to receive lower equivalent hourly wage than those who have no job experience.

Hourly Wage vs. Job Training:-

Figure-29 : Box plot for Hourly Wage vs Job Training

Observations:-

On average, the equivalent hourly wage of the applicants who do not require training is higher than those who require training.

Case Status vs. Years Since Establishment:-

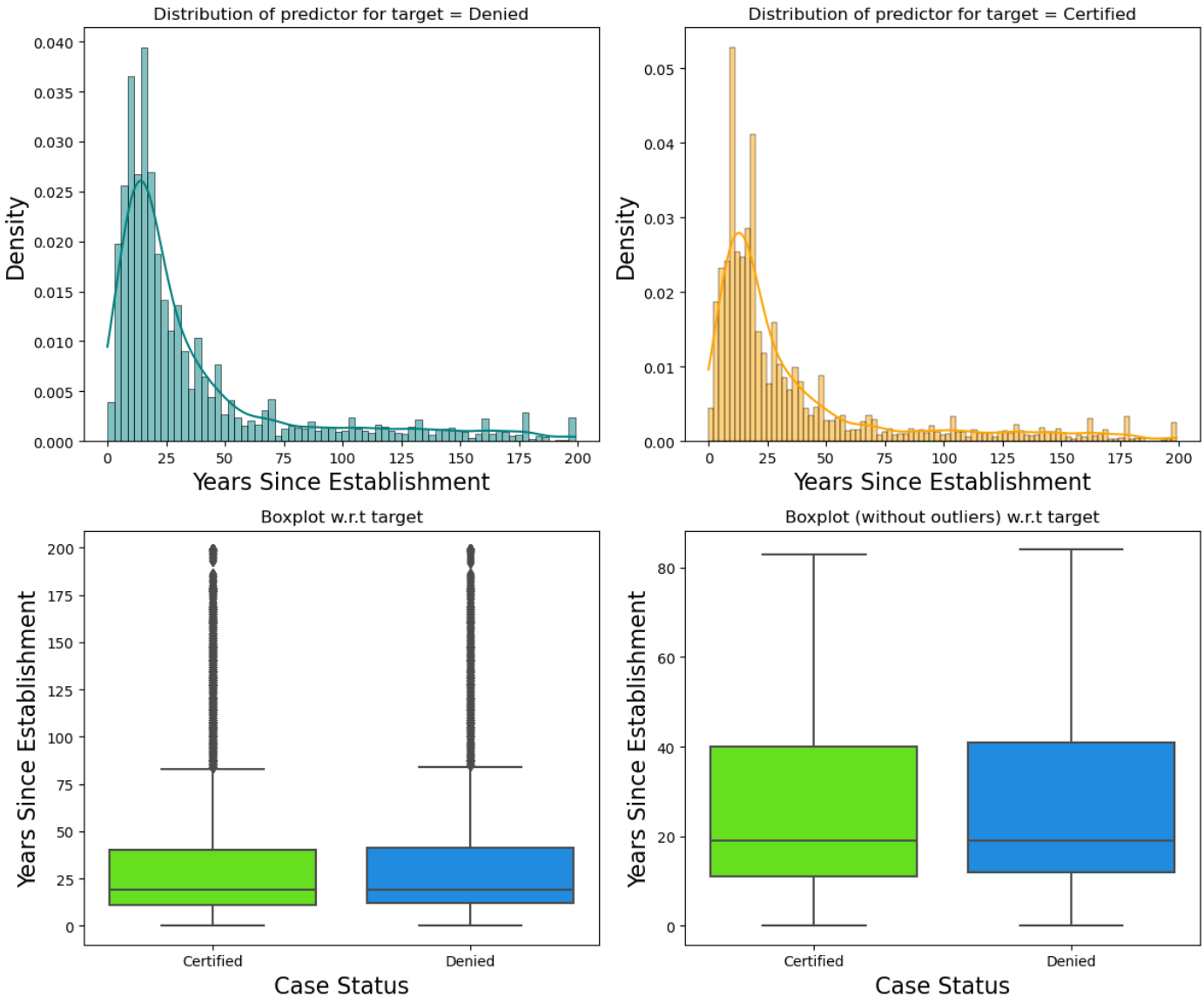


Figure-30 : Dist plot for Case status vs Yrs. Of establishment

Observations:-

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

Number of Employees vs. Years Since Establishment:-

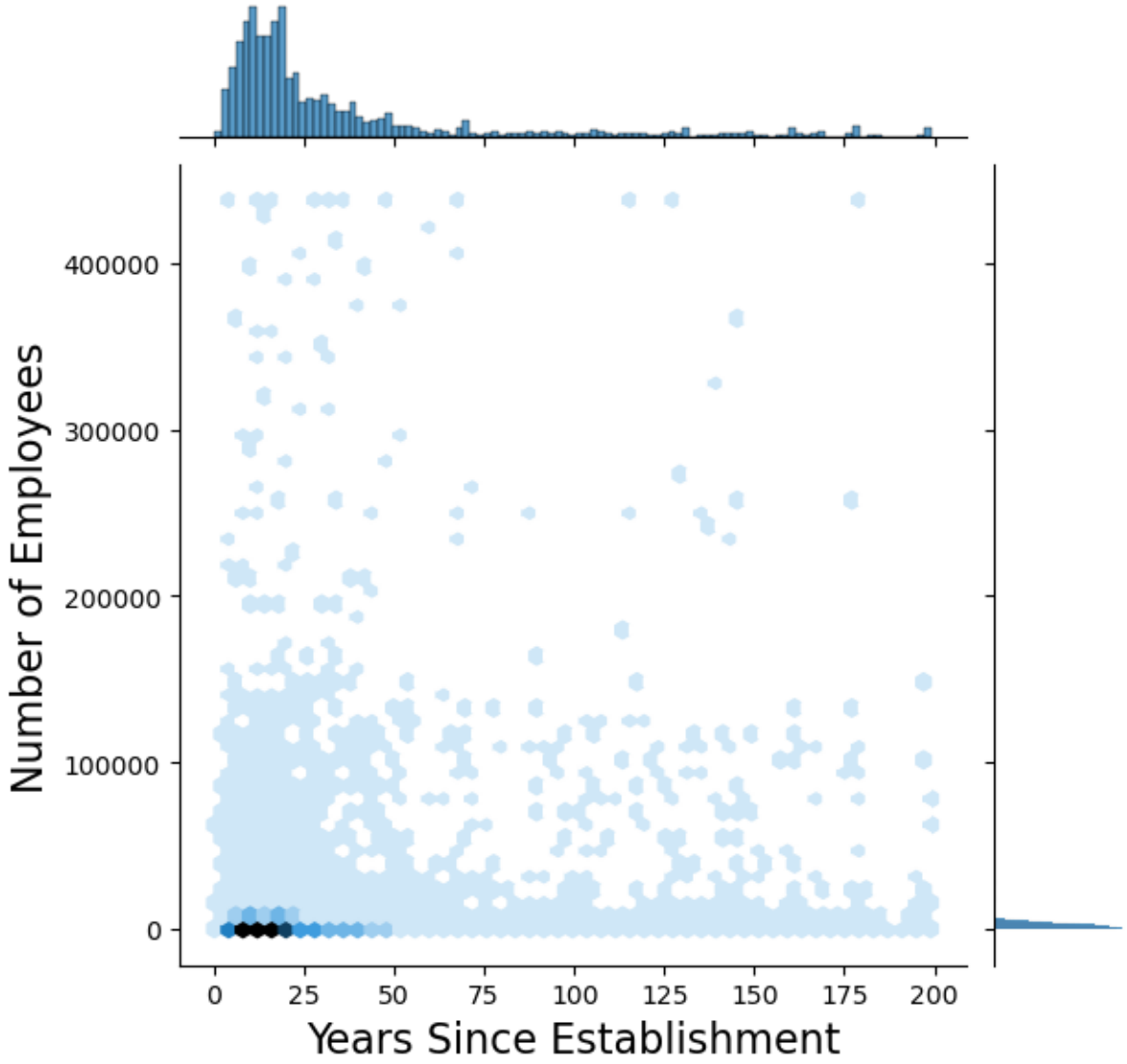


Figure-31 : Joint plot for No of employees vs Yrs. since establishment

Observations:-

Older employers seem to tend to have slightly smaller number of employees compared to the younger employers.

Data Preparation for Modeling:-

Encoding categorial Data:-

	continent	education_of_employee	has_job_experience	requires_job_training	no_of_employees	region_of_employment	unit_of_wage	full_time_position	case_status	yrs_snc_estab	hourly_wage
17639	Asia	2	1	0	567	Midwest	Year	1	1	24	12.905
23951	Oceania	2	0	0	619	Midwest	Year	1	1	78	31.933
8625	Asia	3	0	0	2635	South	Hour	1	1	11	887.292
20206	Asia	2	1	1	3184	Northeast	Year	1	1	30	23.767
7471	Europe	2	1	0	4681	West	Year	1	0	88	23.974
3433	Asia	2	1	0	222	South	Hour	1	1	27	813.726
24440	Europe	1	0	1	3278	South	Year	1	0	22	98.533
12104	Asia	3	1	0	1359	West	Year	0	1	19	97.229
15656	Asia	2	0	0	2081	West	Year	1	0	13	53.708
23110	North America	2	1	0	854	Northeast	Hour	1	0	18	444.826

Table 8: Encoding categorial data

Separation of Dependent and Independent Variables:-

```
Independent Variables
 ================================================================================
       continent  education_of_employee  has_job_experience  \
17639      Asia                       2                   1
23951   Oceania                       2                   0
8625       Asia                       3                   0
20206      Asia                       2                   1
7471     Europe                       2                   1

       requires_job_training  no_of_employees region_of_employment  \
17639                      0              567              Midwest
23951                      0              619              Midwest
8625                       0             2635                South
20206                      1             3184            Northeast
7471                       0             4681                 West

      unit_of_wage  full_time_position  yrs_snc_estab  hourly_wage
17639         Year                   1             24       12.905
23951         Year                   1             78       31.933
8625          Hour                   1             11      887.292
20206         Year                   1             30       23.767
7471          Year                   1             88       23.974


Dependent Variables
 ================================================================================
 17639    1
23951    1
8625     1
20206    1
7471     0
Name: case_status, dtype: category
Categories (2, int64): [1, 0]
```

Table 9: Separation of dependent & Independent

Creating Dummy Variables:-

	education_of_employee	has_job_experience	requires_job_training	no_of_employees	full_time_position	yrs_snc_estab	hourly_wage	continent_Asia
17639	2	1	0	567	1	24	12.905	True
23951	2	0	0	619	1	78	31.933	False
8625	3	0	0	2635	1	11	887.292	True
20206	2	1	1	3184	1	30	23.767	True
7471	2	1	0	4681	1	88	23.974	False

continent_Europe	continent_North America	continent_Oceania	continent_South America	region_of_employment_Midwest	region_of_employment_Northeast
False	False	False	False	True	False
False	False	True	False	True	False
False	False	False	False	False	False
False	False	False	False	False	True
True	False	False	False	False	False

region_of_employment_South	region_of_employment_West	unit_of_wage_Month	unit_of_wage_Week	unit_of_wage_Year
False	False	False	False	True
False	False	False	False	True
True	False	False	False	False
False	False	False	False	True
False	True	False	False	True

Table 10: Creating Dummy variables

Splitting Data into Training and Test Set:-

```
(15288, 19) (5096, 19) (5096, 19)
```

Model Building:-

Initial Model Building:-

- Model Building - Original Data

```
Training Performance:

Bagging: 0.9866797257590597
Random forest: 1.0
GBM: 0.8781586679725759
Adaboost: 0.8905974534769834
dtree: 1.0

Validation Performance:

Bagging: 0.7884841363102233
Random forest: 0.8393066980023501
GBM: 0.8783783783783784
Adaboost: 0.8821974148061105
dtree: 0.7441245593419507
```

Table 11: Training Performance

```
Training and Validation Performance Difference:

Bagging: Training Score: 0.9867, Validation Score: 0.7885, Difference: 0.1982
Random forest: Training Score: 1.0000, Validation Score: 0.8393, Difference: 0.1607
GBM: Training Score: 0.8782, Validation Score: 0.8784, Difference: -0.0002
Adaboost: Training Score: 0.8906, Validation Score: 0.8822, Difference: 0.0084
dtree: Training Score: 1.0000, Validation Score: 0.7441, Difference: 0.2559
```

Table 12: Training & Validation Performance

Observation:-

- AdaBoost has the best performance followed by GBM model as per the validation performance

- Model Building - Oversampled Data:-

```
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, 19)
After Oversampling, the shape of train_y: (20420,)
```

Table 13: Training Performance

```
Training Performance:

Bagging: 0.9815866797257591
Random forest: 1.0
GBM: 0.8431929480901077
Adaboost: 0.8283055827619981
dtree: 1.0

Validation Performance:

Bagging: 0.7502937720329025
Random forest: 0.8119858989424207
GBM: 0.8304935370152762
Adaboost: 0.81786133960047
dtree: 0.7394242068155111
```

Table 14: Training & Validation Performance

```
Training and Validation Performance Difference:

Bagging: Training Score: 0.9816, Validation Score: 0.7503, Difference: 0.2313
Random forest: Training Score: 1.0000, Validation Score: 0.8120, Difference: 0.1880
GBM: Training Score: 0.8432, Validation Score: 0.8305, Difference: 0.0127
Adaboost: Training Score: 0.8283, Validation Score: 0.8179, Difference: 0.0104
dtree: Training Score: 1.0000, Validation Score: 0.7394, Difference: 0.2606
```

Table 15: Training & Validation Performance difference

Observations:-

- GBM has the best performance on validation followed by Adaboost.

➢ Model Building - Undersampled Data:-

```
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, 19)
After Under Sampling, the shape of train_y: (10156,)
```

Table 16: Training Performance

```
Training Performance:

Bagging: 0.9706577392674282
Random forest: 0.9998030720756204
GBM: 0.7534462386766444
Adaboost: 0.7046081134304845
dtree: 1.0

Validation Performance:

Bagging: 0.6075205640423031
Random forest: 0.6700940070505288
GBM: 0.732373678025852
Adaboost: 0.7006462984723855
dtree: 0.6204465334900118
```

Table 17: Training & Validation Performance

```
Training and Validation Performance Difference:

Bagging: Training Score: 0.9707, Validation Score: 0.6075, Difference: 0.3631
Random forest: Training Score: 0.9998, Validation Score: 0.6701, Difference: 0.3297
GBM: Training Score: 0.7534, Validation Score: 0.7324, Difference: 0.0211
Adaboost: Training Score: 0.7046, Validation Score: 0.7006, Difference: 0.0040
dtree: Training Score: 1.0000, Validation Score: 0.6204, Difference: 0.3796
```

Table 18: Training & Validation Performance difference

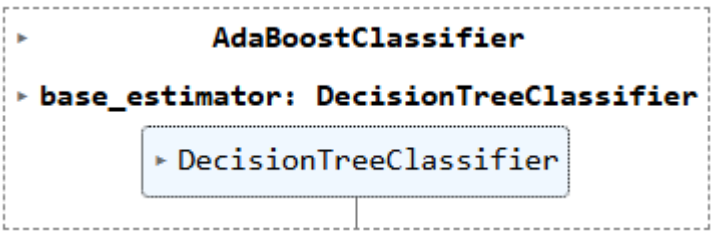
Observations:-

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

Hyperparameter Tuning:-

Tuning AdaBoostClassifier model with Undersampled data:-

```
Best parameters are {'n_estimators': 30, 'learning_rate': 0.05, 'base_estimator': DecisionTreeClassifier(max_depth=2, random_state=1)} with CV score=0.7
908595477289477:
CPU times: total: 2.94 s
Wall time: 37.6 s
```



Checking model's performance on training set:-

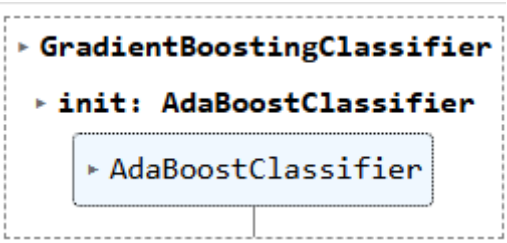
	Accuracy	Recall	Precision	F1
0	0.703	0.718	0.697	0.707

Checking model's performance on validation set:-

	Accuracy	Recall	Precision	F1
0	0.709	0.713	0.828	0.766

Tuning Gradient Boosting model with Undersampled Data:-

```
Best parameters are {'subsample': 1, 'n_estimators': 150, 'max_features': 0.5, 'learning_rate': 0.01, 'init': AdaBoostClassifier(random_state=1)} with C
V score=0.7469487219270005:
CPU times: total: 7.28 s
Wall time: 2min 42s
```



Checking model's performance on training set:-

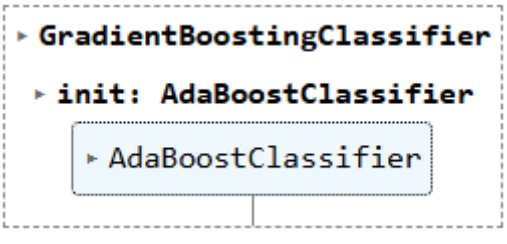
	Accuracy	Recall	Precision	F1
0	0.705	0.723	0.697	0.710

Checking model's performance on validation set:-

	Accuracy	Recall	Precision	F1
0	0.711	0.718	0.827	0.768

Tuning Gradient Boosting model with Oversampled data:-

```
Best parameters are {'subsample': 1, 'n_estimators': 75, 'max_features': 0.7, 'learning_rate': 0.1, 'init': AdaBoostClassifier(random_state=1)} with CV
score=0.8345739471106759:
CPU times: total: 9.97 s
Wall time: 3min 59s
```



Checking model's performance on training set:-

	Accuracy	Recall	Precision	F1
0	0.742	0.723	0.751	0.737

Checking model's performance on validation set:-

	Accuracy	Recall	Precision	F1
0	0.711	0.718	0.827	0.768

Model Comparison and Final Model Selection:-

Training performance comparison:

	Gradient boosting trained with Undersampled data	Gradient boosting trained with Oversampled data	AdaBoost trained with Undersampled data
Accuracy	0.705	0.742	0.703
Recall	0.723	0.723	0.718
Precision	0.697	0.751	0.697
F1	0.710	0.737	0.707

Validation performance comparison

Validation performance comparison:

	Gradient boosting trained with Undersampled data	Gradient boosting trained with Oversampled data	AdaBoost trained with Undersampled data
Accuracy	0.711	0.711	0.709
Recall	0.718	0.718	0.713
Precision	0.827	0.827	0.828
F1	0.768	0.768	0.766

Observations:-

- AdaBoost model trained with undersampled data has generalised performance, so let's consider it as the best model.

Let's check the performance on test set:-

	Accuracy	Recall	Precision	F1
0	0.706	0.726	0.813	0.767

Observations:-

- The Adaboost model trained on undersampled data has given ~73% recall on the test set
- This performance is in line with what we achieved with this model on the train and validation sets
- So, this is a generalized model

Feature Importance:-

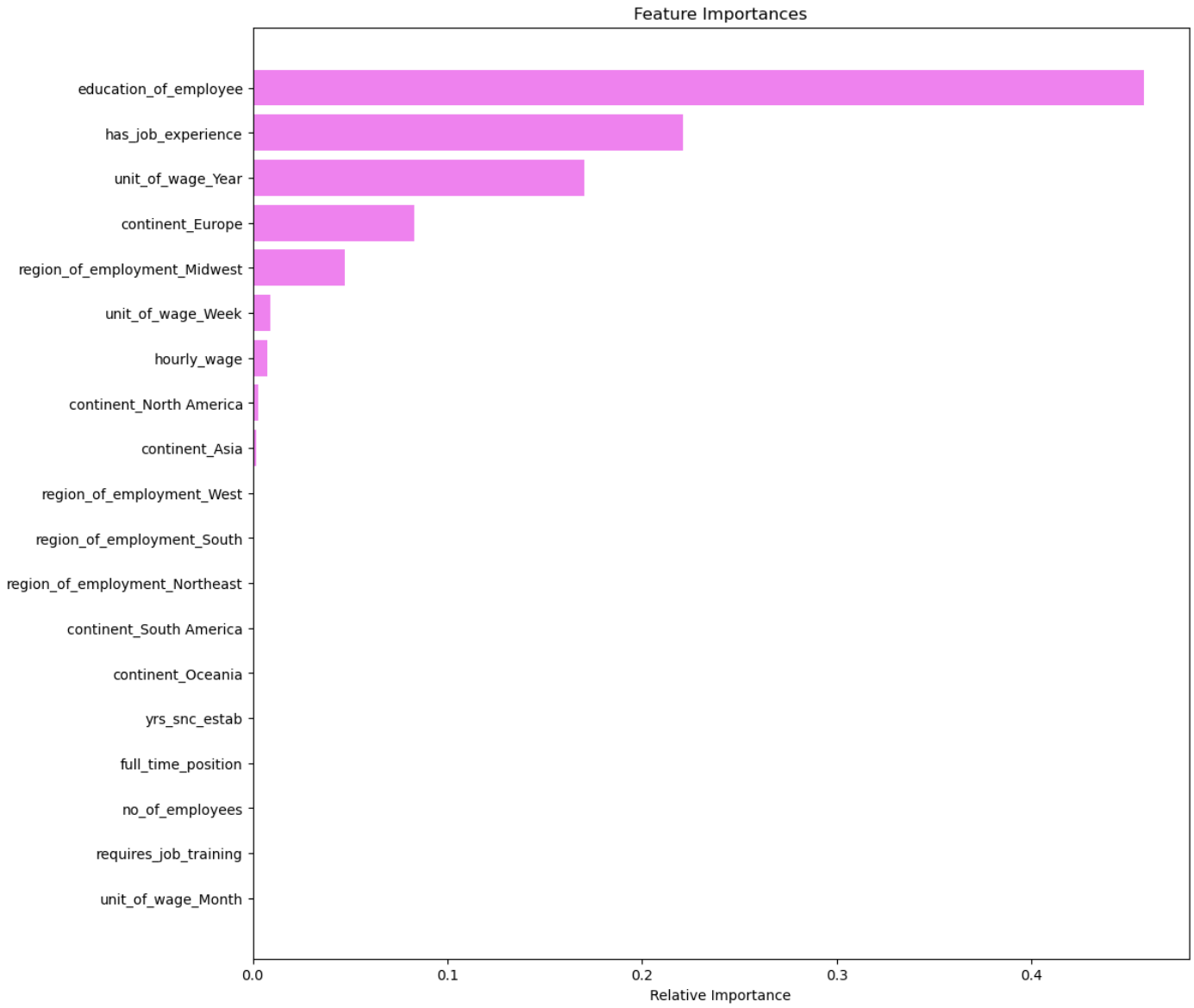


Figure-32 : Bar graph for Tuned Adaboost Important features

Observations:-

- We can see that education_of_employee, has_job_experience, unit_of_wage_Year are the most important features for making predictions

Profile Recommendations Based on Model Outcomes

Q. Facilitate the process of visa approvals.

Ans:

- Certified Applicants:
    - Applicants with higher salaries and education levels (e.g., graduate degrees).
    - Applicants working in technical fields (like software engineering) where there is higher demand.
    - Applicants with substantial work experience in multinational companies.
- Denied Applicants:
    - Applicants in job categories with lower salaries or where there are high rejection rates.
    - Applicants from industries facing more scrutiny, such as outsourcing.
    - Profiles with discrepancies or incomplete documentation.

Insights from the Analysis Conducted

Q. 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.

Ans:

- Key Drivers for Approval:
    - Salary Level: Higher salaries are positively correlated with visa approval, indicating that applicants with competitive wages have better chances of approval.
    - Job Role: Specialized technical or managerial roles often have higher approval rates due to skill shortages in the job market.
    - Work Experience: More experienced applicants, especially those working in multinational or well-established companies, tend to have more success.
    - Education: Advanced degrees (Master's, PhDs) are a major positive factor, as they signal higher skill levels.
    - Employer's Profile: Larger and more reputed companies have higher approval rates compared to smaller or less established firms.

Actionable Business Recommendations and Insights:-

According to the EDA:

- The majority (66%) of work via applications are from Asia.
- A large portion (78%) of the applicants have a bachelor's or a master's degree and only less than 9% have a doctrate degree.
- Most (58%) of the applicants have job experience.
- The vast majority of offerred jobs (88%) do not require training.

- The majority (>81%) of the offered jobs are for Northeast, South, and West regions of the US.
- The majority (89%) of the offered positions are full-time.
- Merely about 10% of the positions have a wage unit other than Year.
- About 2/3 of the work visa applications are certified.
- The European and South American applicants have the highest and the lowest chances of visa certification, respectively.
- The higher the applicant's education level is, the more their chances of visa certification are.
- Having job experience increases the chances of visa certification.
- Job training requirement has a negligible effect on visa certification likelihood.
- The visa applications for the employment in the Midwest region are more likely to be certified than the applications for the employment in other regions.
- Being a full- or part-time position does not observably affect the visa certification likelihood.
- The offered positions with the wage units of Year and Hour have the highest and the lowest chances of visa certification, respectively.
- The employer's number of employees has an insignificant impact on the chances of visa certification for its potential foreign employees.
- The majority of employers applying for work visas are less than 40 years old.
- The majority of the applications are for the jobs with an equivalent hourly wage of less than 100 (probably in dollars).
- The positions with certified visa applications are on average of lower equivalent hourly wages than the positions with denied visa applications.
- The age of an employer has negligible effect on the likelihood of visa certification.

- **For Employers**: Focus on filing visa applications for roles that require specialized skills, offer competitive salaries, and are in high demand. Ensure documentation is complete and accurate to minimize denial risks.
- **For Applicants**:
    - Enhance qualifications through advanced degrees or certifications relevant to their field.
    - Apply to roles that offer higher wages, as salary is a significant factor in approval.
    - Gain experience in multinational companies, as this improves chances of approval.
- **For Immigration Consultants**: Use data-driven insights to pre-screen applicants and flag profiles that may face higher rejection risks. Recommend strategies to strengthen their profiles, such as improving documentation or pursuing advanced education.

These insights should guide targeted strategies for increasing visa approval success rates based on key drivers from the analysis.

END