MACHINE LEARNING-2(ML)

Coded Project Report

Easy Visa Project

Submitted to

greatlearning
Learning for Life

by

Subhadeep Seal

In Partial Fulfilment of PGP-DSBA

TEXAS McCombs
The University of Texas at Austin
McCombs School of Business

Contents

List of Tables

Easy Visa Project

1.1 Problem Definition

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 ForeignLabor Certification (OFLC).

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

*1.2 **Objective***

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

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

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

1.3 Data Dictionary

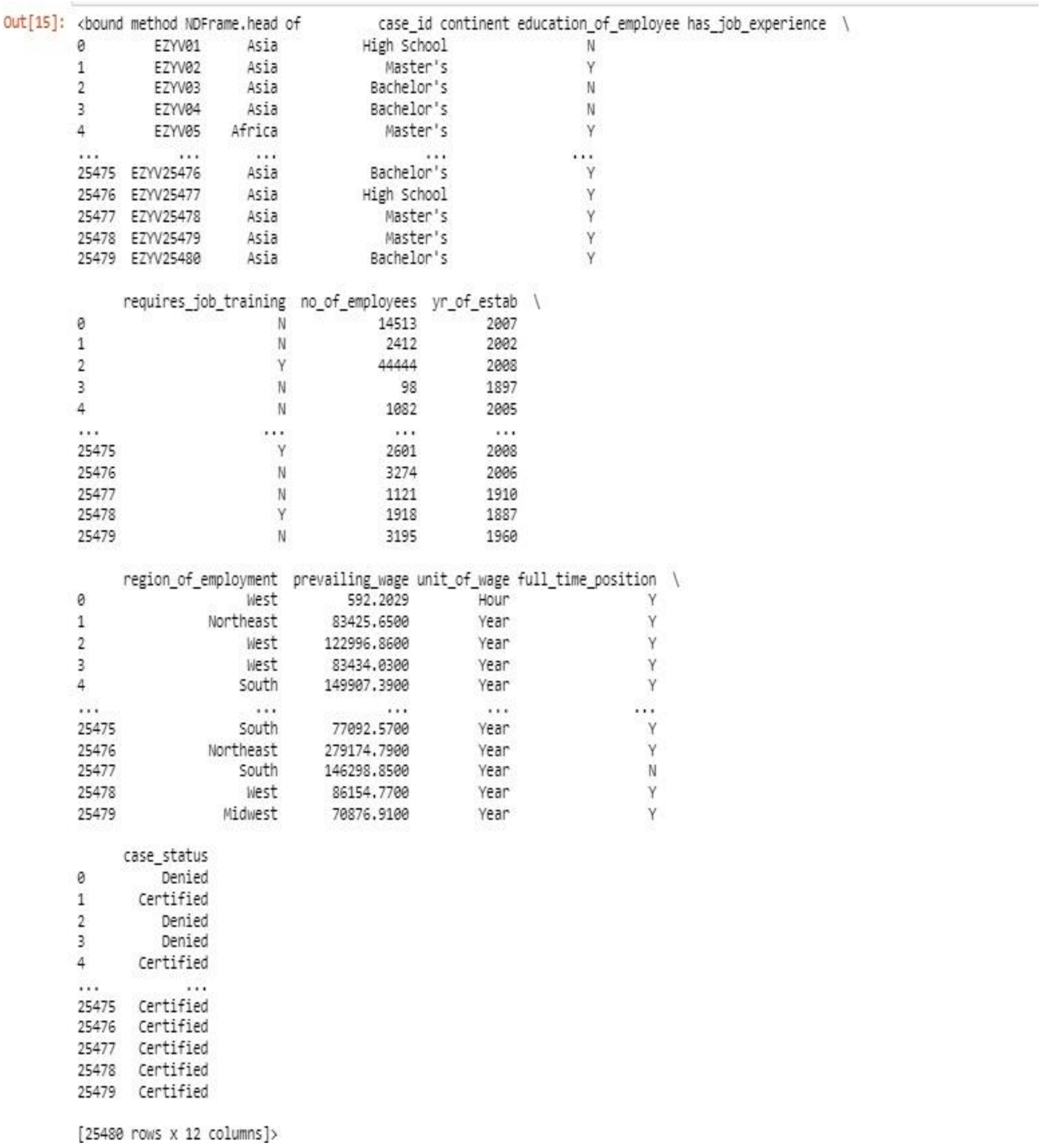
The data contains the different attributes of employee and the employer. The detailed data dictionary is given below.

- 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 ofthe 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

1.4 Data Overview

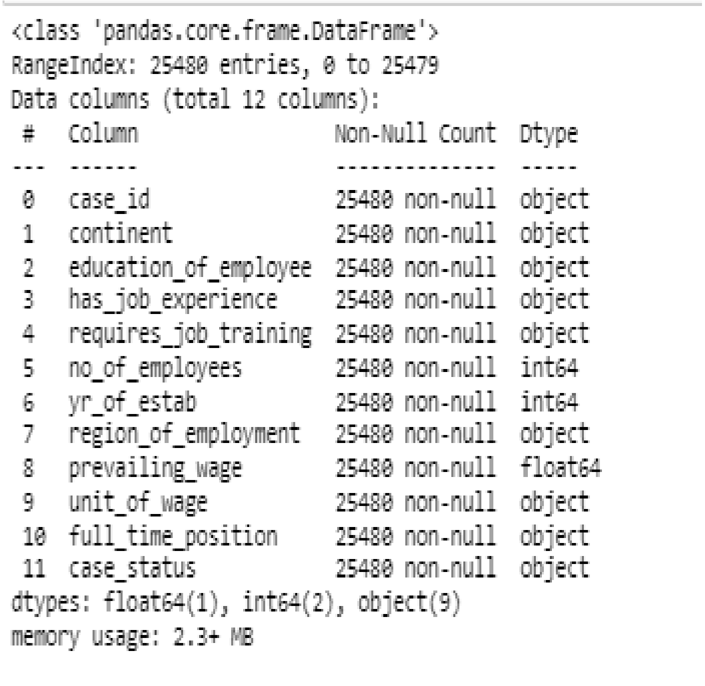
Structure of the data:

The data has 12 columns as mentioned in the data dictionary.



```
Out[15]: <bound method NDFrame.head of         case_id continent education_of_employee has_job_experience  \
0         EZYV01      Asia           High School                  N
1         EZYV02      Asia              Master's                  Y
2         EZYV03      Asia            Bachelor's                  N
3         EZYV04      Asia            Bachelor's                  N
4         EZYV05    Africa              Master's                  Y
...          ...       ...                   ...                ...
25475  EZYV25476      Asia            Bachelor's                  Y
25476  EZYV25477      Asia           High School                  Y
25477  EZYV25478      Asia              Master's                  Y
25478  EZYV25479      Asia              Master's                  Y
25479  EZYV25480      Asia            Bachelor's                  Y

      requires_job_training  no_of_employees  yr_of_estab  \
0                         N            14513         2007
1                         N             2412         2002
2                         Y            44444         2008
3                         N               98         1897
4                         N             1082         2005
...                     ...              ...          ...
25475                     Y             2601         2008
25476                     N             3274         2006
25477                     N             1121         1910
25478                     Y             1918         1887
25479                     N             3195         1960

      region_of_employment  prevailing_wage unit_of_wage full_time_position  \
0                     West         592.2029         Hour                  Y
1                Northeast       83425.6500         Year                  Y
2                     West      122996.8600         Year                  Y
3                     West       83434.0300         Year                  Y
4                    South      149907.3900         Year                  Y
...                    ...              ...          ...                ...
25475                South       77092.5700         Year                  Y
25476            Northeast      279174.7900         Year                  Y
25477                South      146298.8500         Year                  N
25478                 West       86154.7700         Year                  Y
25479              Midwest       70876.9100         Year                  Y

      case_status
0          Denied
1       Certified
2          Denied
3          Denied
4       Certified
...           ...
25475   Certified
25476   Certified
25477   Certified
25478   Certified
25479   Certified

[25480 rows x 12 columns]>
```

Table 1: Top 5 rows of the Data set



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

Table 2: Basic information of Data Set

Following are basic observations of data set after analyzing the data set:

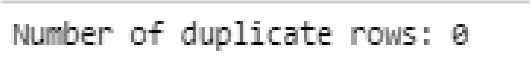
1.The data frame has 25480 rows and 12 columns.

2.There are no null values in data set

3. There are no missing values in data set

Duplicate Values

Inspecting the duplicate values in customer key column



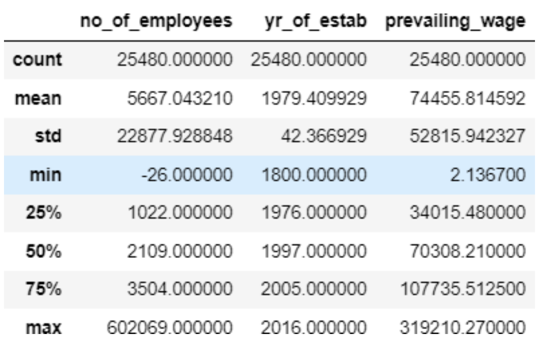
```
Number of duplicate rows: 0
```

Table 3: Duplicate data summary of the Dataset

Observation:

1. There are no duplicates in the dataset.

Statistical summary of the dataset



	no_of_employees	yr_of_estab	prevailing_wage
count	25480.000000	25480.000000	25480.000000
mean	5667.043210	1979.409929	74455.814592
std	22877.928848	42.366929	52815.942327
min	-26.000000	1800.000000	2.136700
25%	1022.000000	1976.000000	34015.480000
50%	2109.000000	1997.000000	70308.210000
75%	3504.000000	2005.000000	107735.512500
max	602069.000000	2016.000000	319210.270000

Table 4: Statistical summary of Categorical fields only

Observations:

- The average number of employees in the employer's organization are 5667 while the median number of employees in the employer's organization are 2109. This implies the attribute has a right skewed distribution with several positive outliers. The minimum number is negativewhich does not appear to be a valid data point
- There are companies in the dataset with years of establishment from 1800 to 2016
- The average prevailing wage for occupation in United States is USD 74,455 while the median (~50th percentile of wages) is USD 70,308. This indicates, slight right skewness in the data set. The minimum value of USD 2.1367 does not appear to be a valid data point. The attribute has tobe studied in union with unit_of_wage to gather further insight
- The case ID attribute can be dropped as it is a unique ID variable and is not expected to add any value to the status of a visa being accepted
- There are 6 continents in the database, with majority of applicants from Asia
- There are 4 different levels of eduction with Bachelor's being the highest education degree for majority of applicants
- Majority of applicants do not require further job training to perform the intended occupation in the US
- There are 5 different regions in the US requiring immigrants due to Human Resource shortages, the maximum being in the NorthEast US region
- There are 4 different units of wages with yearly being the most common. The prevailing wage and unit of wage may need to be studied in unionto gather further insight
- Majority of the occupation with employee shortages are full time positions
- Case status is the attribute of interest (which needs to be predicted by our ML model). As per dataset, 66.7% of all applicants have a certified visastatus and only 33.2% have a denied visa status

1.5 Exploratory Data Analysis

We have plotted histogram and boxplot all the data variables.
1. Number of employees in the employer's company:

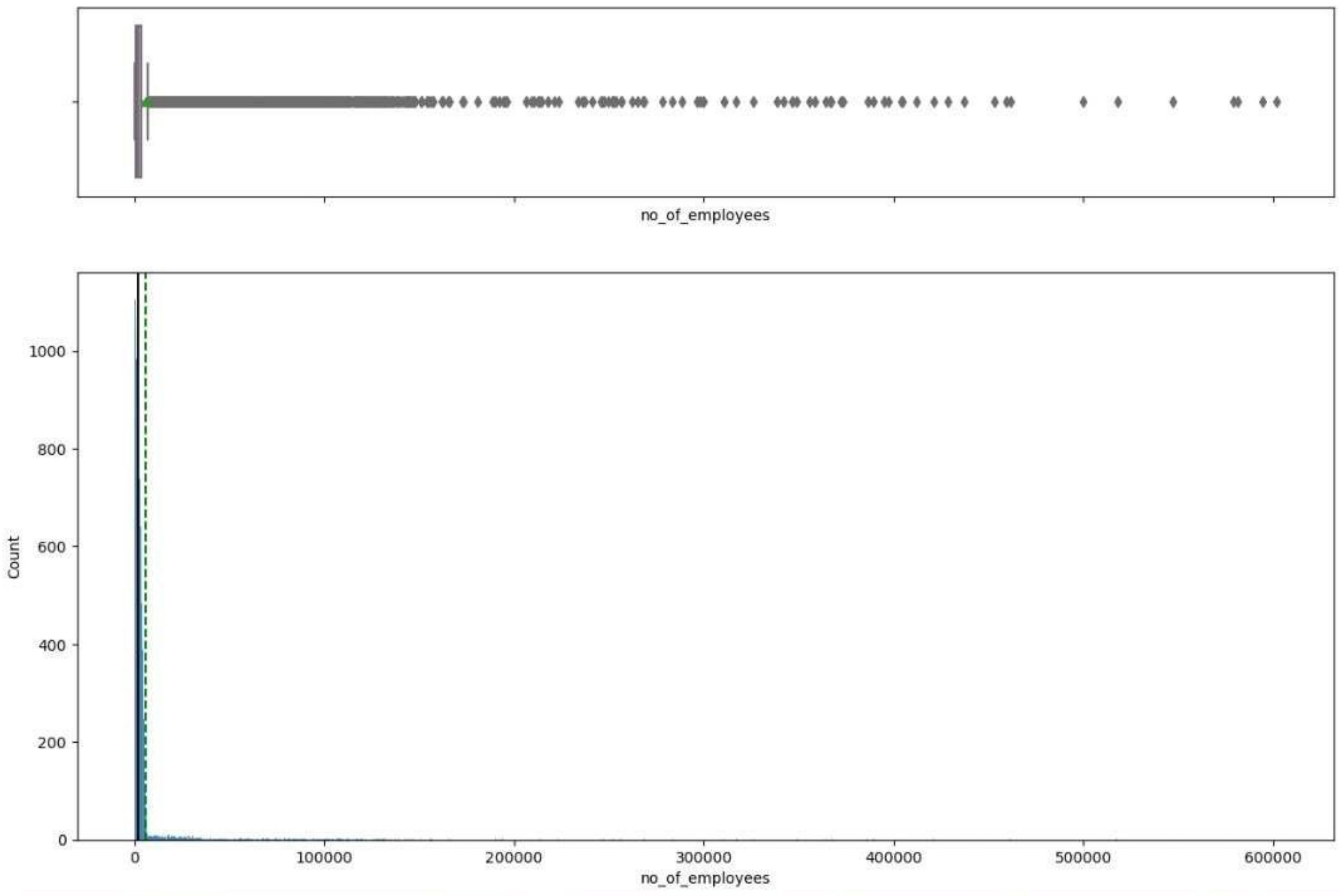


Table 6: Histogram and boxplot of Number of employees in the employer's company

2. Prevailing wage: Average wage paid to similarly employed workers in a specific occupation in the area of intended employment. The purpose of theprevailing 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.

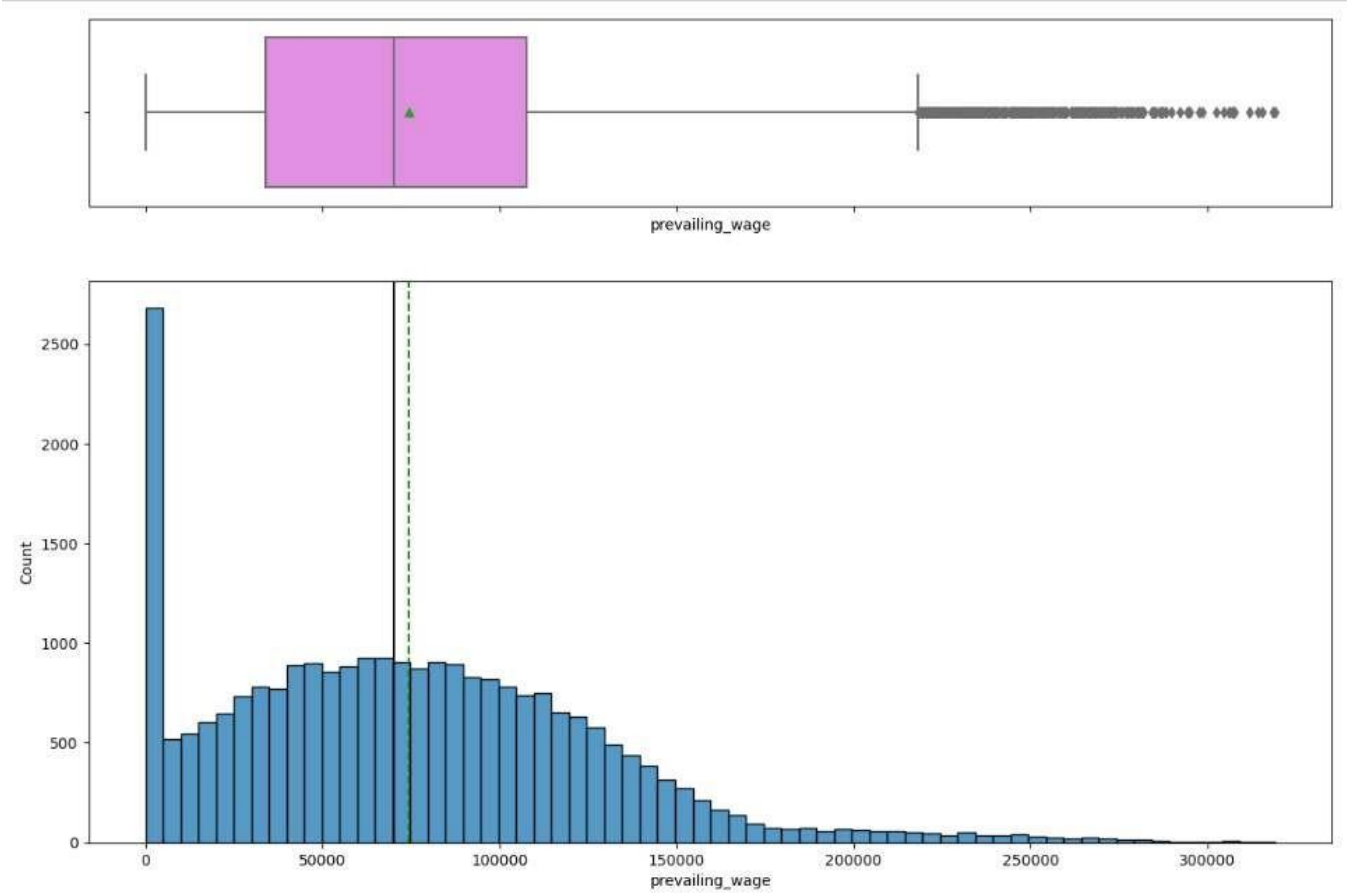
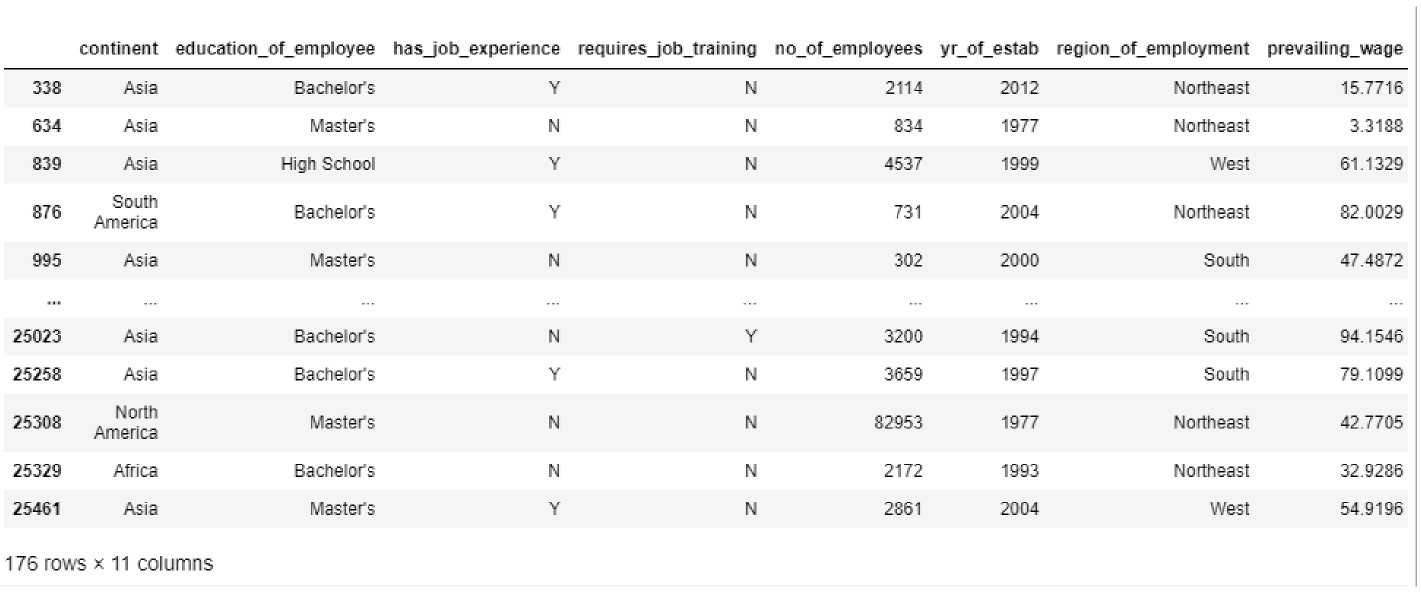


Table 7: Histogram and boxplot of prevailing wage

We are checking the observations which have less than 100 prevailing wage



	continent	education_of_employee	has_job_experience	requires_job_training	no_of_employees	yr_of_estab	region_of_employment	prevailing_wage
338	Asia	Bachelor's	Y	N	2114	2012	Northeast	15.7716
634	Asia	Master's	N	N	834	1977	Northeast	3.3188
839	Asia	High School	Y	N	4537	1999	West	61.1329
876	South America	Bachelor's	Y	N	731	2004	Northeast	82.0029
995	Asia	Master's	N	N	302	2000	South	47.4872
...	...	...	...	...	...	...	...	...
25023	Asia	Bachelor's	N	Y	3200	1994	South	94.1546
25258	Asia	Bachelor's	Y	N	3659	1997	South	79.1099
25308	North America	Master's	N	N	82953	1977	Northeast	42.7705
25329	Africa	Bachelor's	N	N	2172	1993	Northeast	32.9286
25461	Asia	Master's	Y	N	2861	2004	West	54.9196

176 rows × 11 columns

Table 8: observations which have less than 100 prevailing wage

Observations:

1.There are 176 observations which have less than prevailing wage

We will create bar plot of different variables

1. continent–Information of continent the employee:

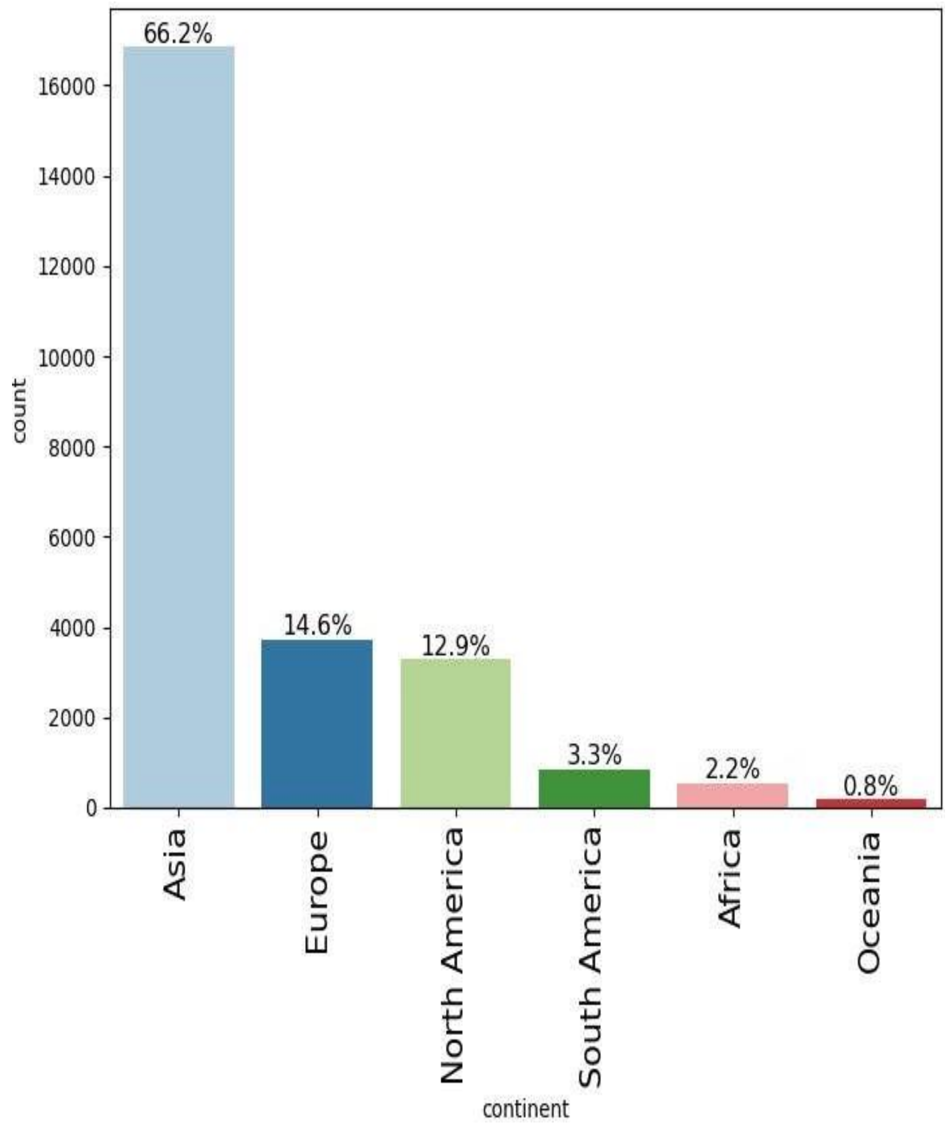


Table 9: Bar plot Information of continent the employee

2. Education of employee: Information of education of the employee

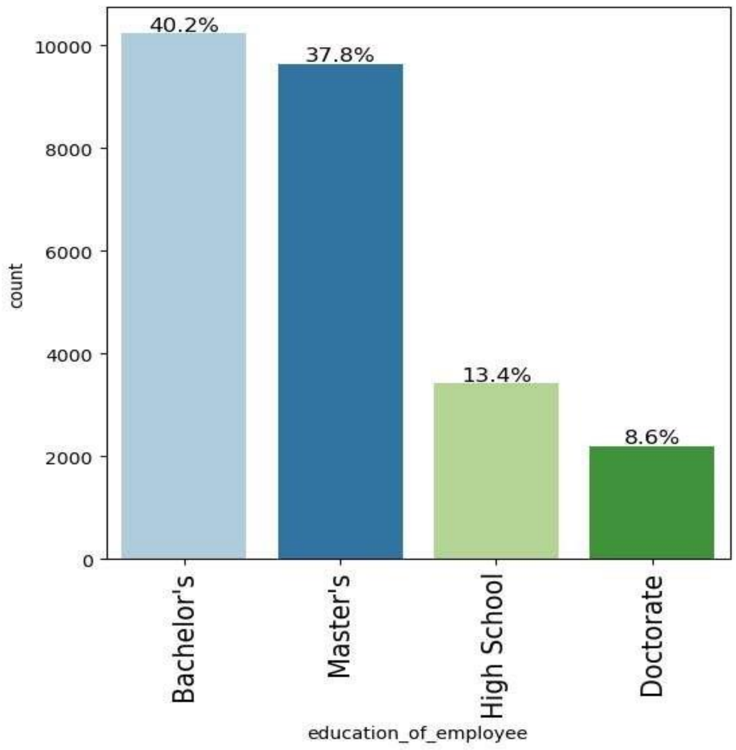


Table 10: Education of employee

3.has_job_experience: Does the employee has any job experience? Y= Yes; N = No

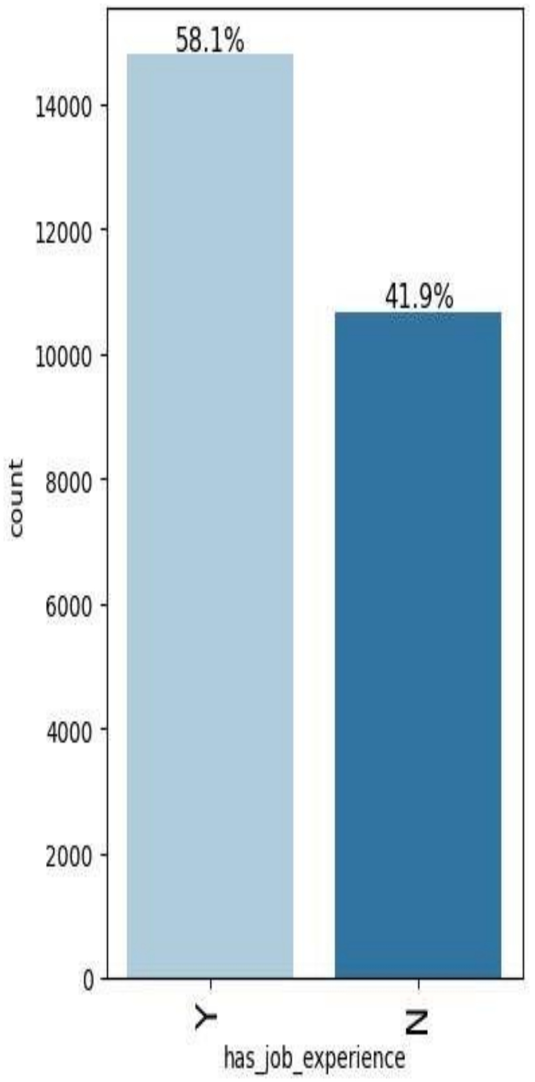


Table 11: Barplot of has_job_experience

4. requires_job_training: Does the employee require any job training?

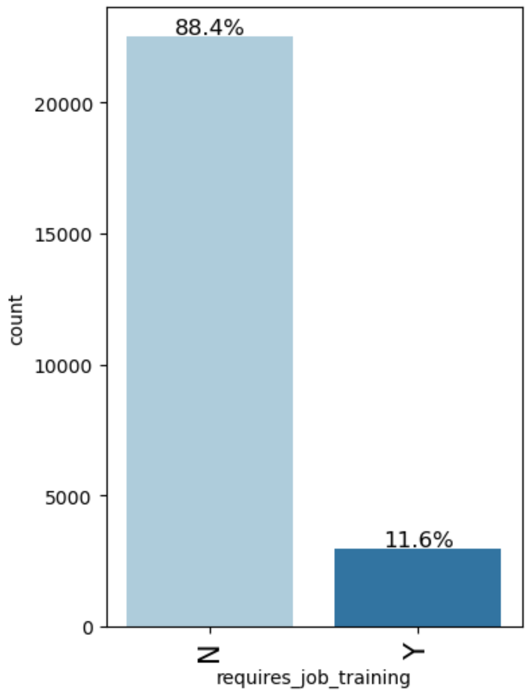


Table 12: Bar plot of requires_job_training

5. region_of_employment: Information of foreign worker's intended region of employment in the US

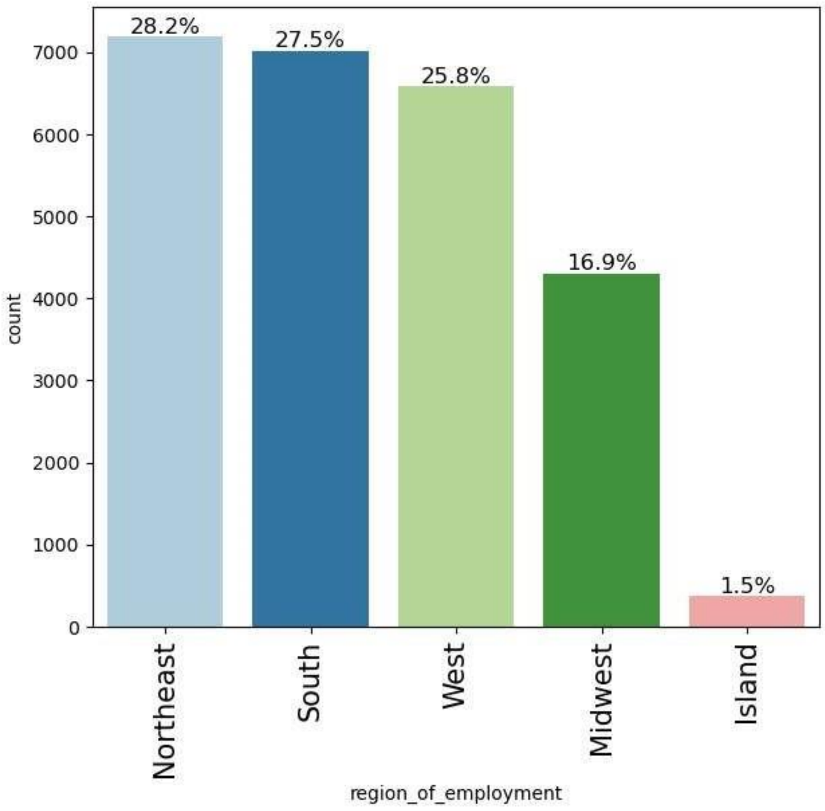


Table 13: Bar plot of region_of_employment

6. Unit_of_wage: Unit of prevailing wage. Values include Hourly, Weekly, Monthly, and Yearly.

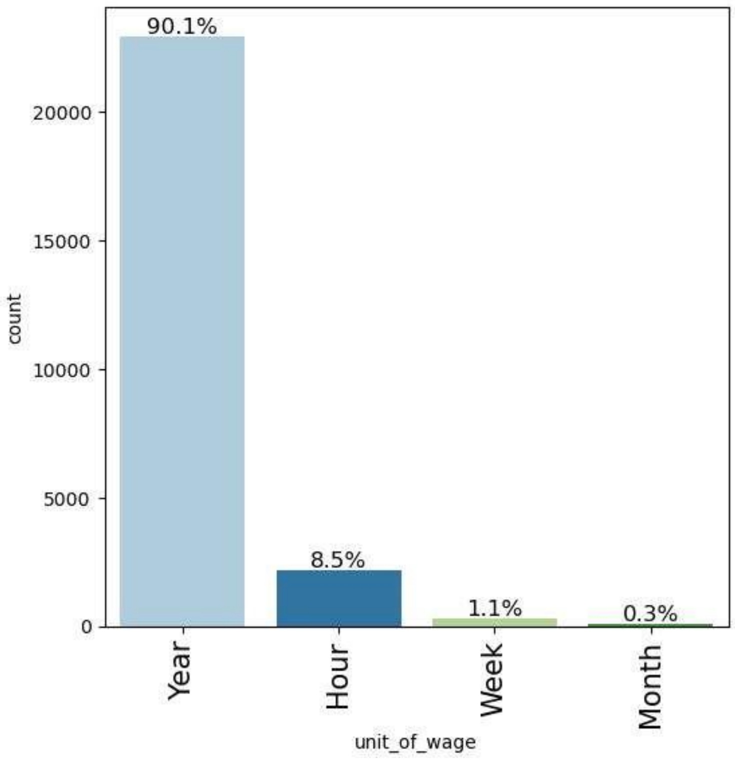


Table 14: Bar plot of Unit_of_wage

7.case_status: Flag indicating if the Visa was certified or denied

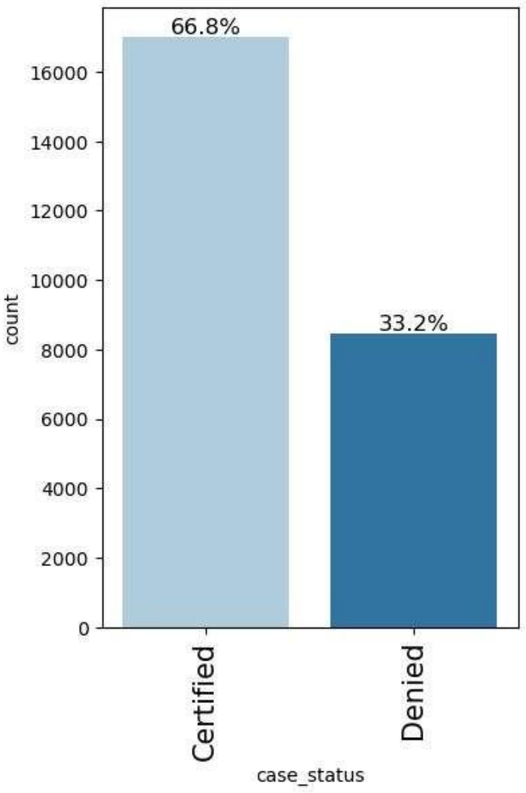


Table 15: Bar plot of case_status

Observations from Exploratory Data Analysis

- More than twice the number of cases were certified than denied irrespective of the number of employees in the employer's organization & the year of establishment of the employer's organization. These attributes are hence, not thought to have animpact on case statuses
    - Both these attributes are heavily skewed, the no_of_employees is skewed right but yr_of_estab is skewed left
    - From the EDA, we infer 58% of all cases were for smaller organizations (<2500 employees) and 61% of all cases werefor employer's established after 1990
- Only 35% of the cases were certified when the unit_of_wage is Hour−ly but 70% were certified when the unit_of_wage is notHour−ly (i.e., Week−ly, Month−ly or Year−ly). This indicates unit_of_wage is an important attribute that can influence case statuses
    - From the EDA, we infer only 8.5% of all cases were for unit_of_wage Hour−ly and the remaining 91.5% of all cases werefor unit_of_wage not Hourly (i.e., Weekly, Monthly or Yearly)
- Majority of cases are from applicants in Asia (66%), then Europe (15%), N. America (13%) & S. America (3%); however, casesgetting certified is highest for Europe (80% of such cases), then Africa (72% of such cases), then Asia (65% of such cases), & least for S. America & N. America (around 60% of such cases). More cases are certified than denied irrespective of the continent. Being from Europe is thought to be an important attribute to have an impact on case statuses

- Majority of applicants have a bachelor's (40%) or a master's degree (37.87%). A small number have only high school certification (13.4%) or are very highly educated/ doctorate (8.6%). However, cases getting certified is highest for doctorate degree (>86%), followed by master degree (>76%), then bachelor's (~62%). The cases getting certified is very low for those applicants with only a high school certification (<35%). The trend observed is intuitive and one can expect attributes having adoctorate degrees & having only a high school certification to significantly contribute to a case being certified and denied respectively

- From the EDA, we infer that 58% of all applicants have prior job experience and 42% do not. The cases getting certified is highfor applicants with prior job experience (75% of such cases) and low for applicants without prior job experience (~56% of suchcases). This is again an important attribute with an applicant having prior job experience significantly contributing to a case being certified

- Majority do not require the employee to receive any additional job training. This attribute was not found to have an impact onthe case statuses

- Majority of the applications are to Northeast (28.3%), then South (27.5%), then West (25.8%), Midwest (16.9%) and least to Island (1.5%) regions of the US. However, the cases certified follows the trend Midwest (75% of such cases), then South (70%of such cases), then Northeast, West, & Island (60% of such cases). Region of employment being Midwest hence is an important attribute contributing positively to a case being certified

- Majority of the jobs are full time rather than part time. This attribute was not found to have an impact on the case statuses

1.6 Bivariate Analysis

Correlation matrix of variables and Creating a pair plot with a kernel density estimate on the diagonal

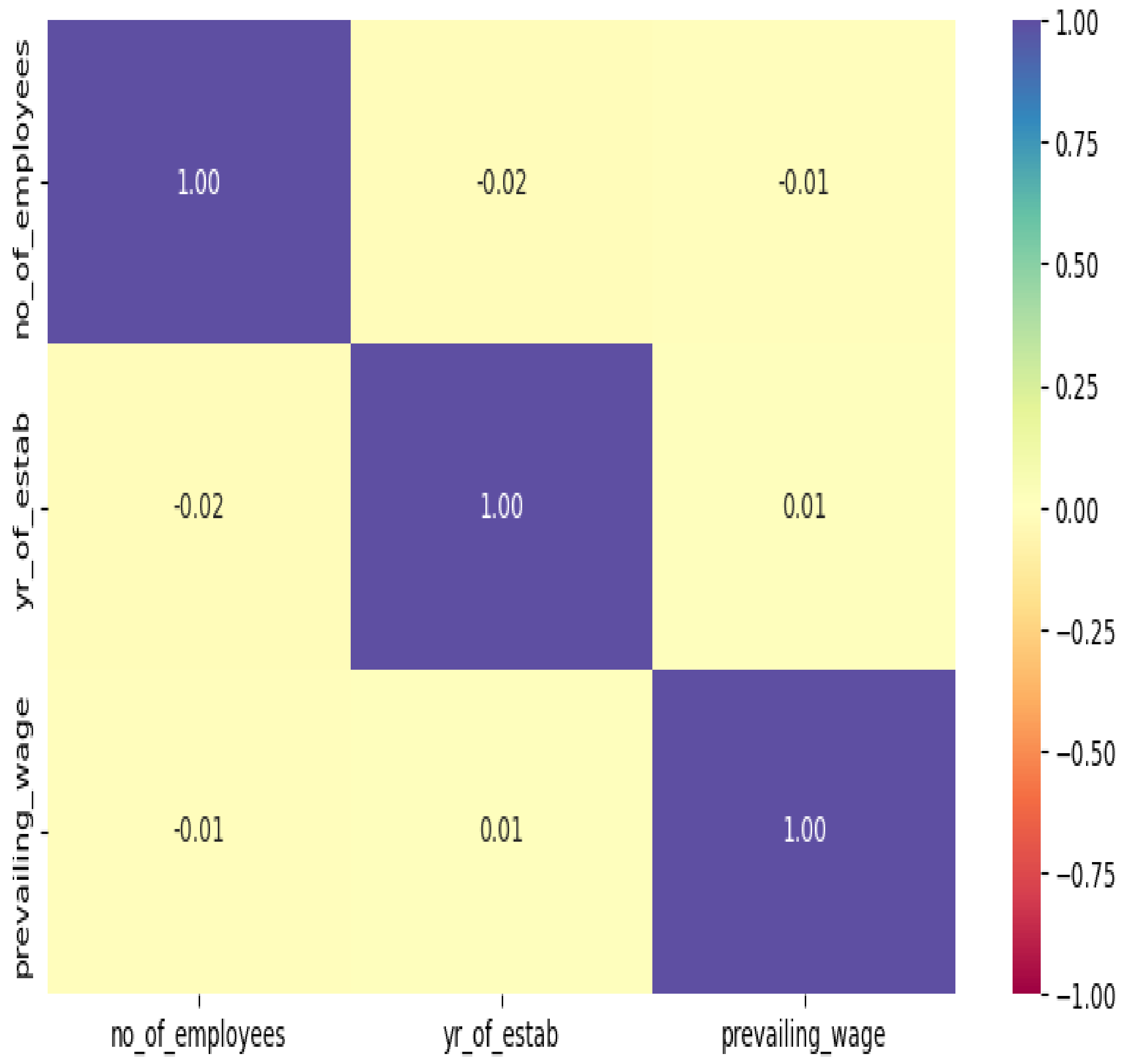
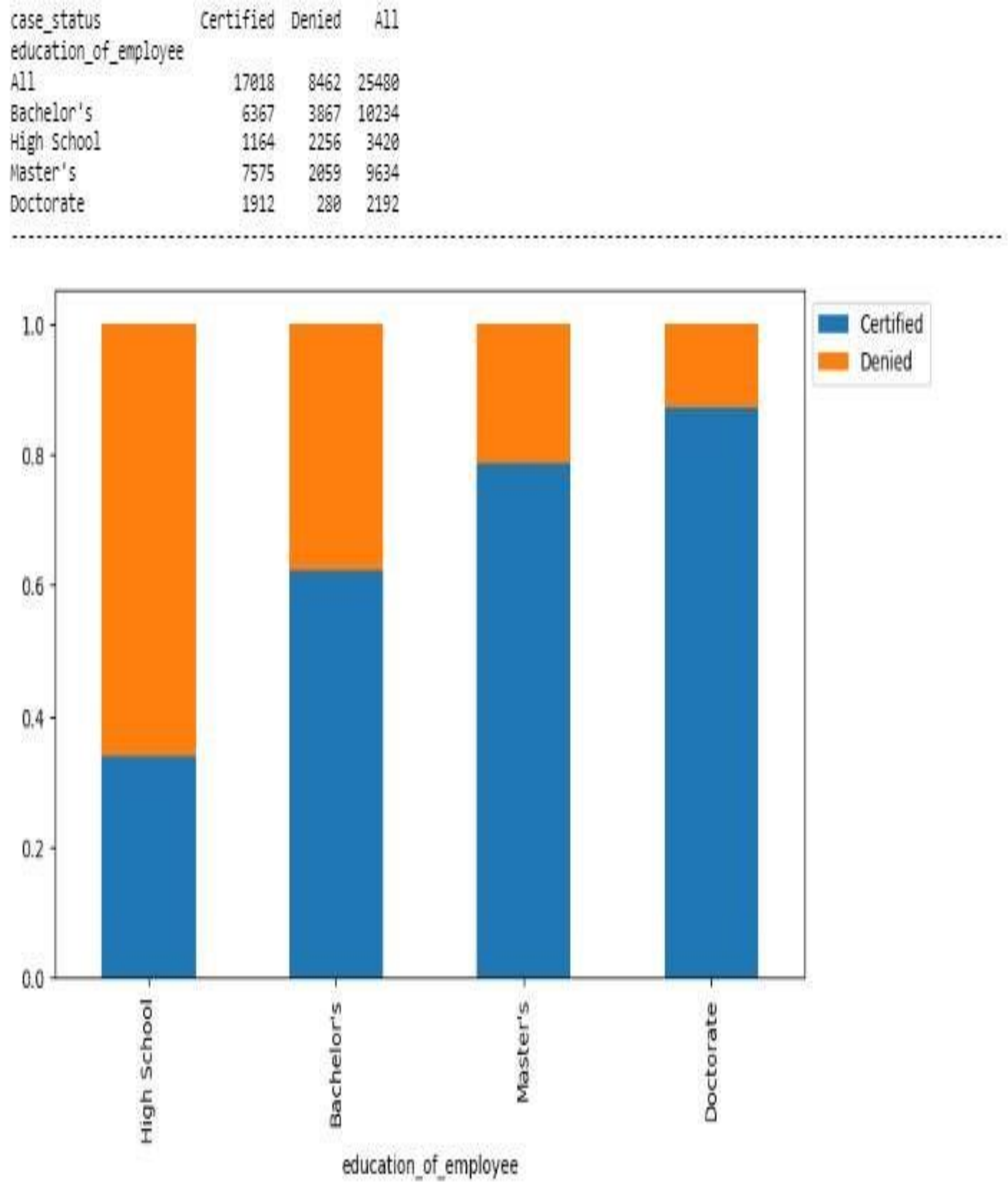


Table 16: Correlation matrix of variables

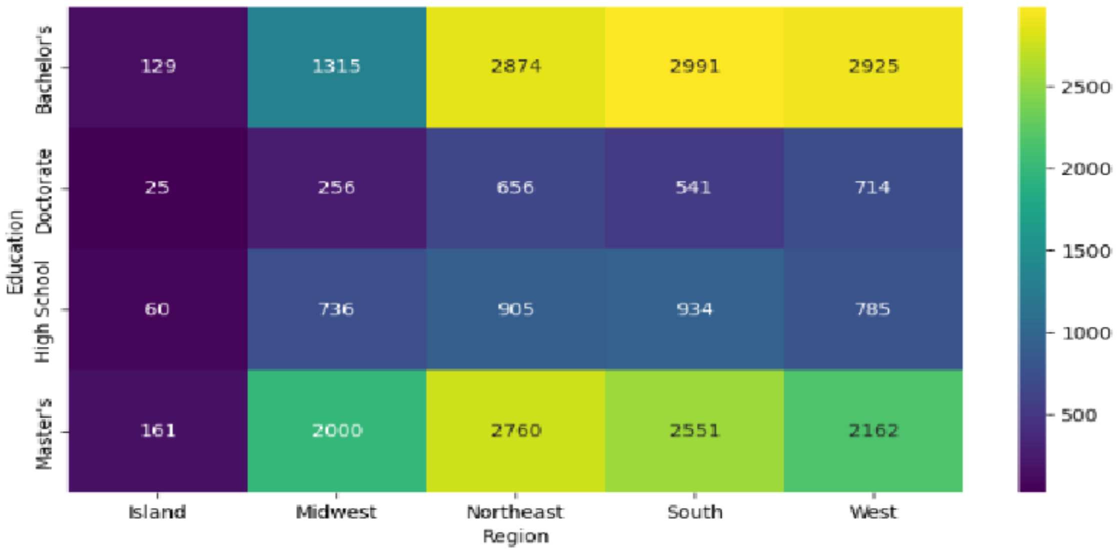
Those with higher education may want to travel abroad for a well−paid job. Let's find out if education has any impact on visa certification



```
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
------------------------------------------------------------------------------------------------------------------------
```

Table 17: education has any impact on visa certification

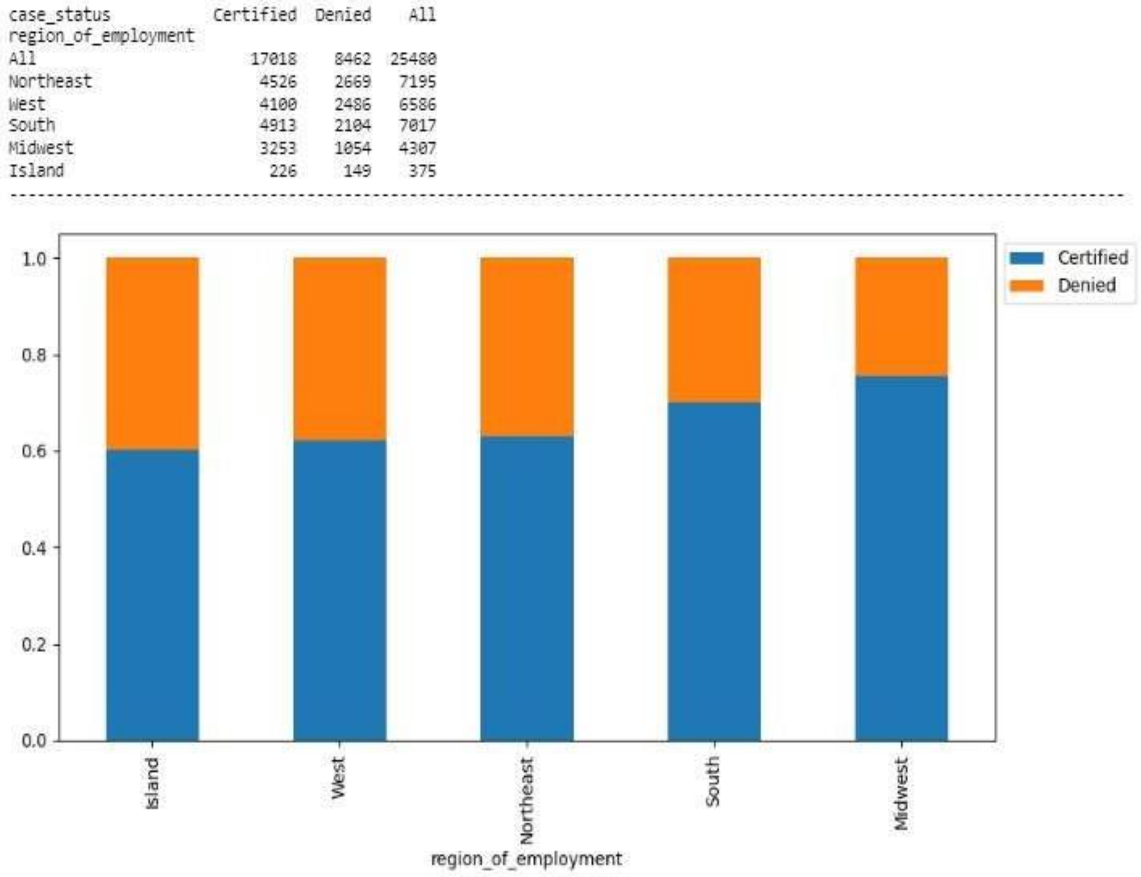
Different regions have different requirements of talent having diverse educational backgrounds. Let's analyze it further



Education	Island	Midwest	Northeast	South	West
Bachelor's	129	1315	2874	2991	2925
Doctorate	25	256	656	541	714
High School	60	736	905	934	785
Master's	161	2000	2760	2551	2162

Table 18: Different regions have different requirements of talent having diverse educational backgrounds

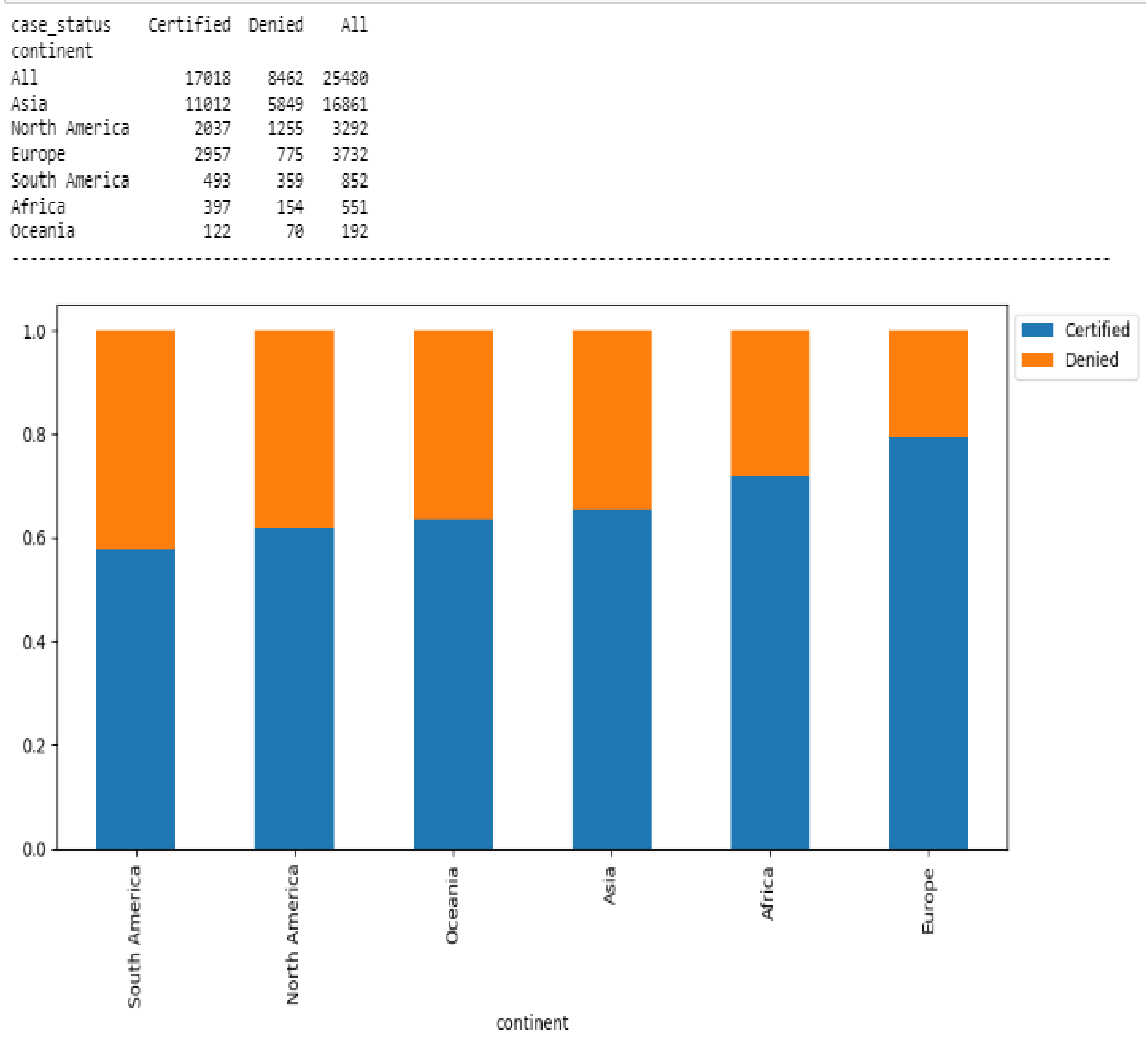
Let's have a look at the percentage of visa certifications across each region_



```
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
------------------------------------------------------------------------------------------------------------------------
```

Table 19: percentage of visa certifications across each region

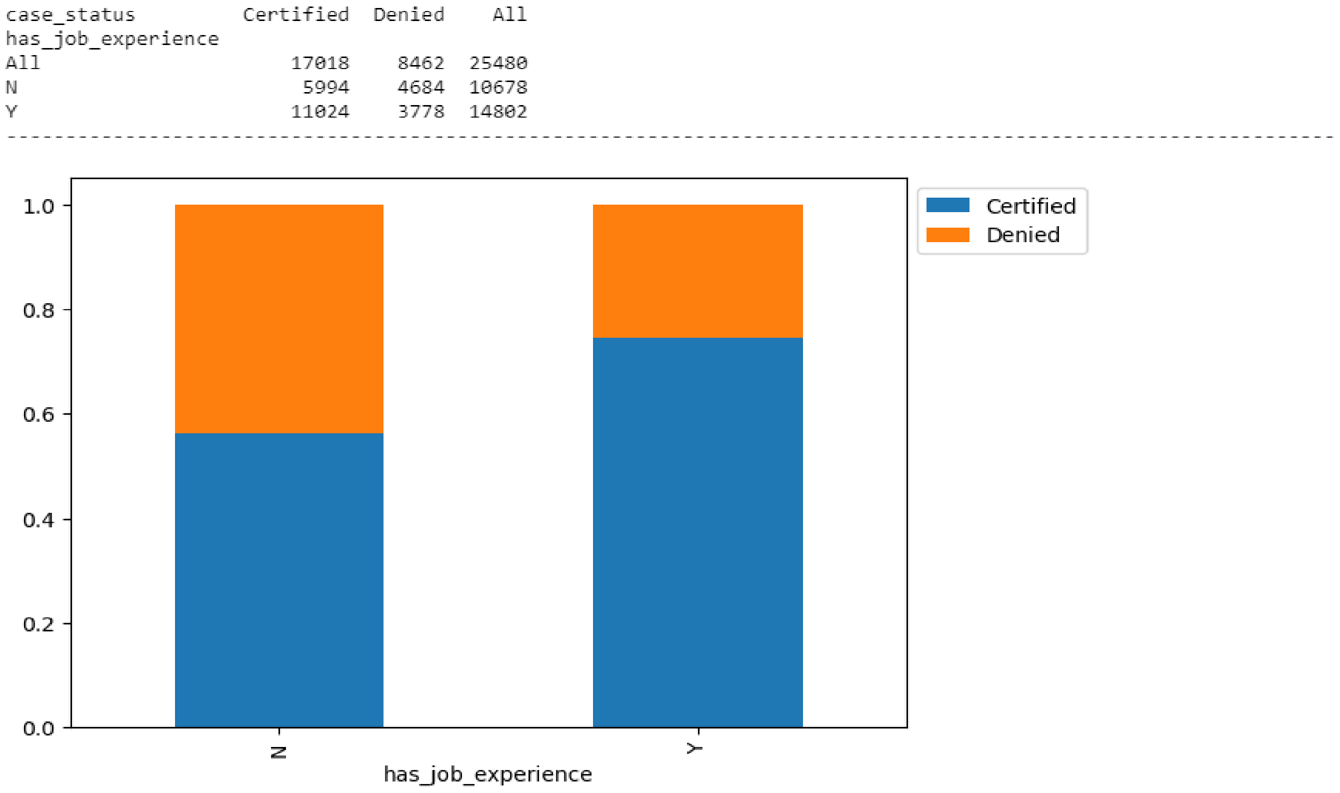
Lets' similarly check for the continents and find out how the visa status vary across different continents.



```
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
------------------------------------------------------------------------------------------------------------------------
```

Table 20: the visa status varies across different continents

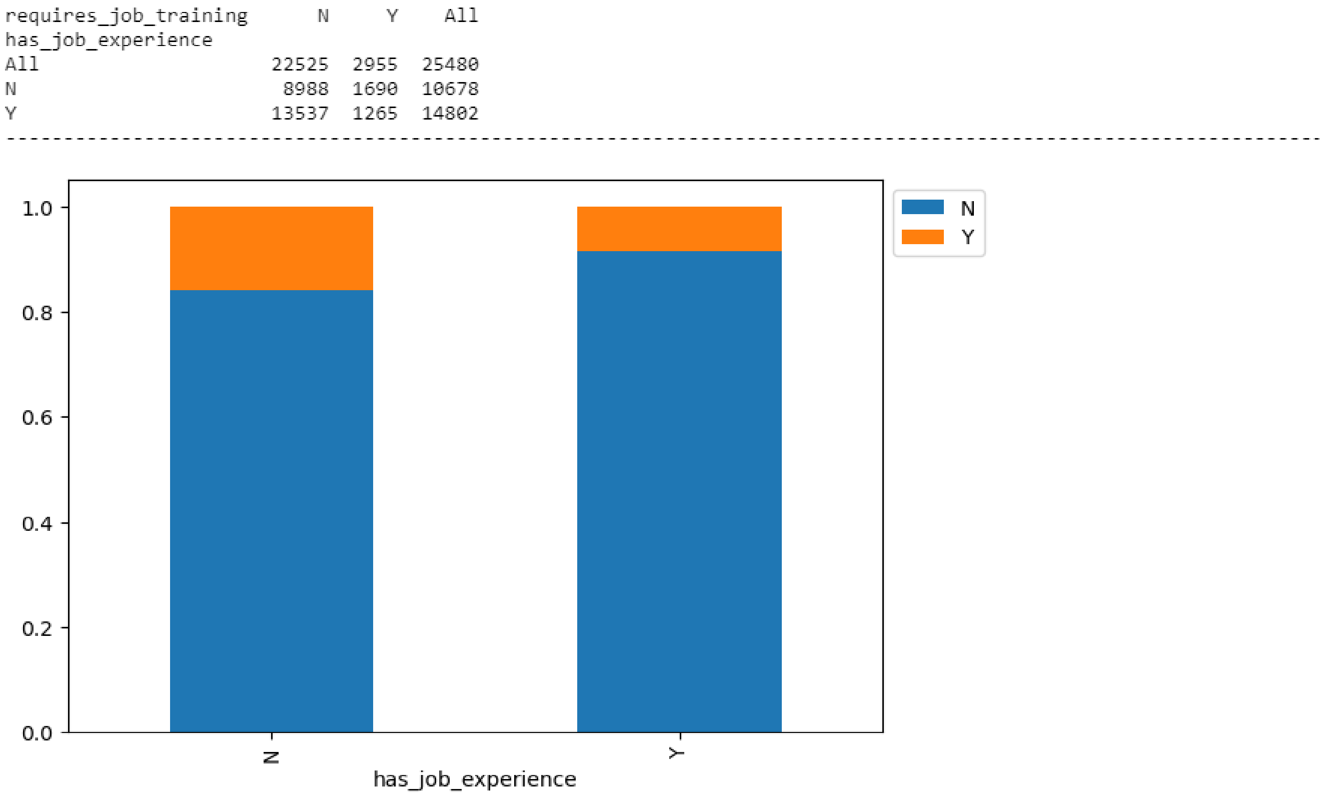
Experienced professionals might look abroad for opportunities to improve their lifestyles and career development. Let's see if having work experience hasany influence over visa certification



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

Table 21: having work experience has any influence over visa certification

Do the employees who have prior work experience require any job training?



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

Table 22: employees who have prior work experience require any job training

The US government has established a prevailing wage to protect local talent and foreign workers. Let's analyze the data and see if the visa status changeswith the prevailing wage

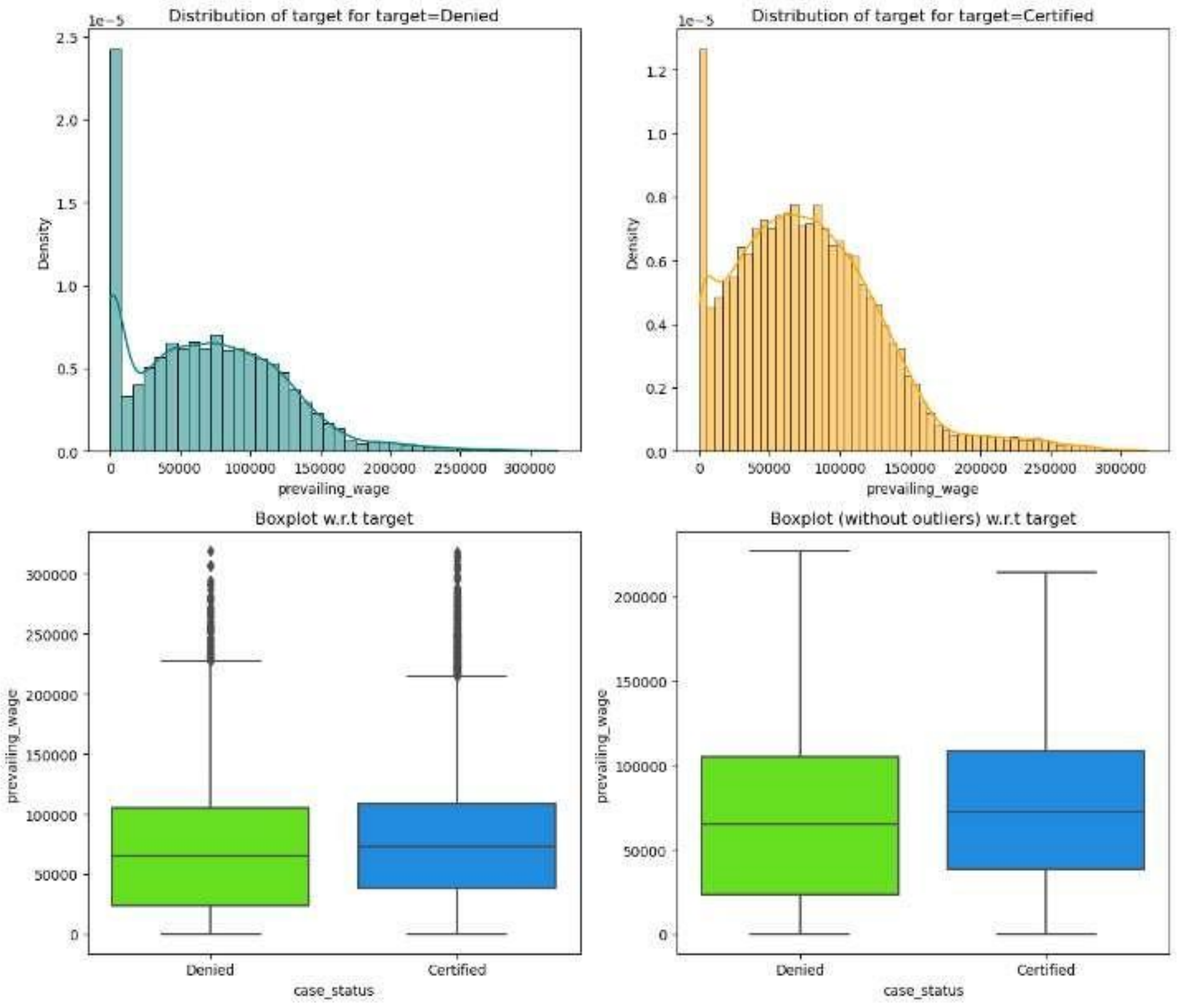


Table 23: see if the visa status changes with the prevailing wage

Checking if the prevailing wage is similar across all the regions of the US

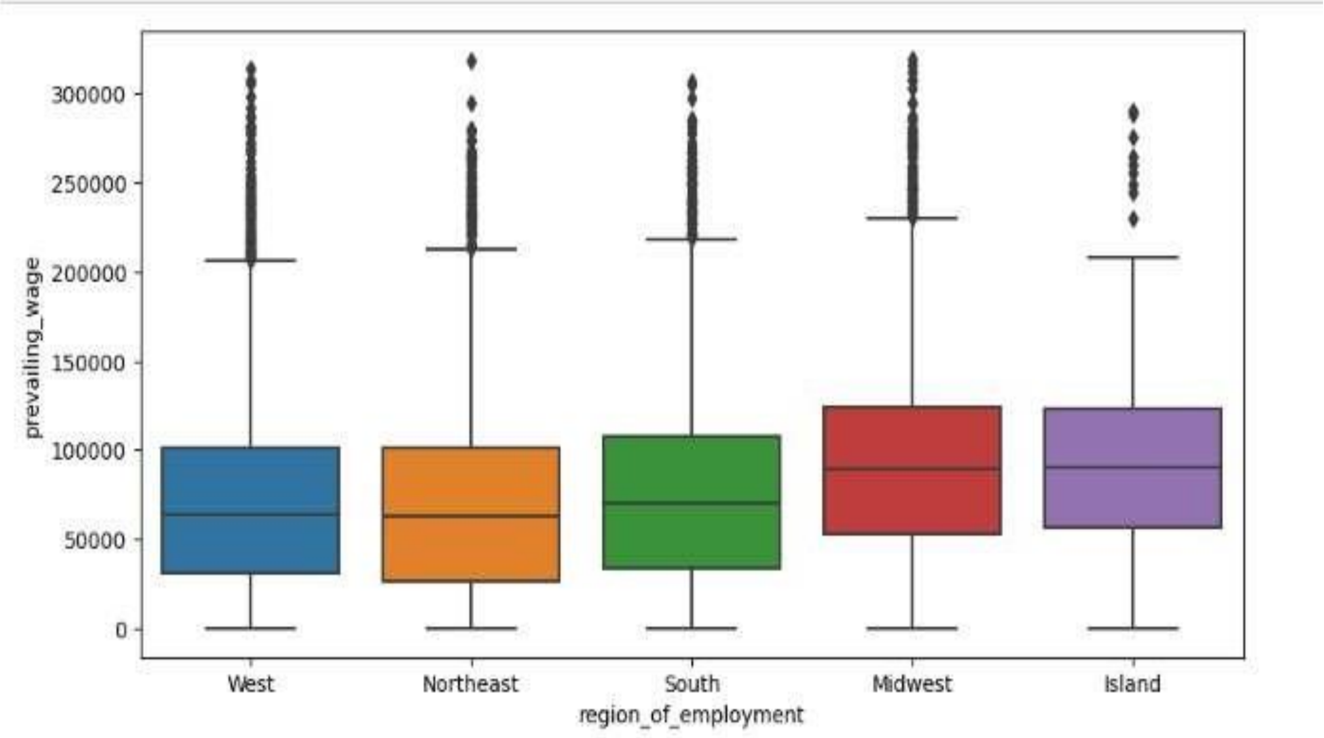


Table 24: Checking if the prevailing wage is similar across all the regions of the US

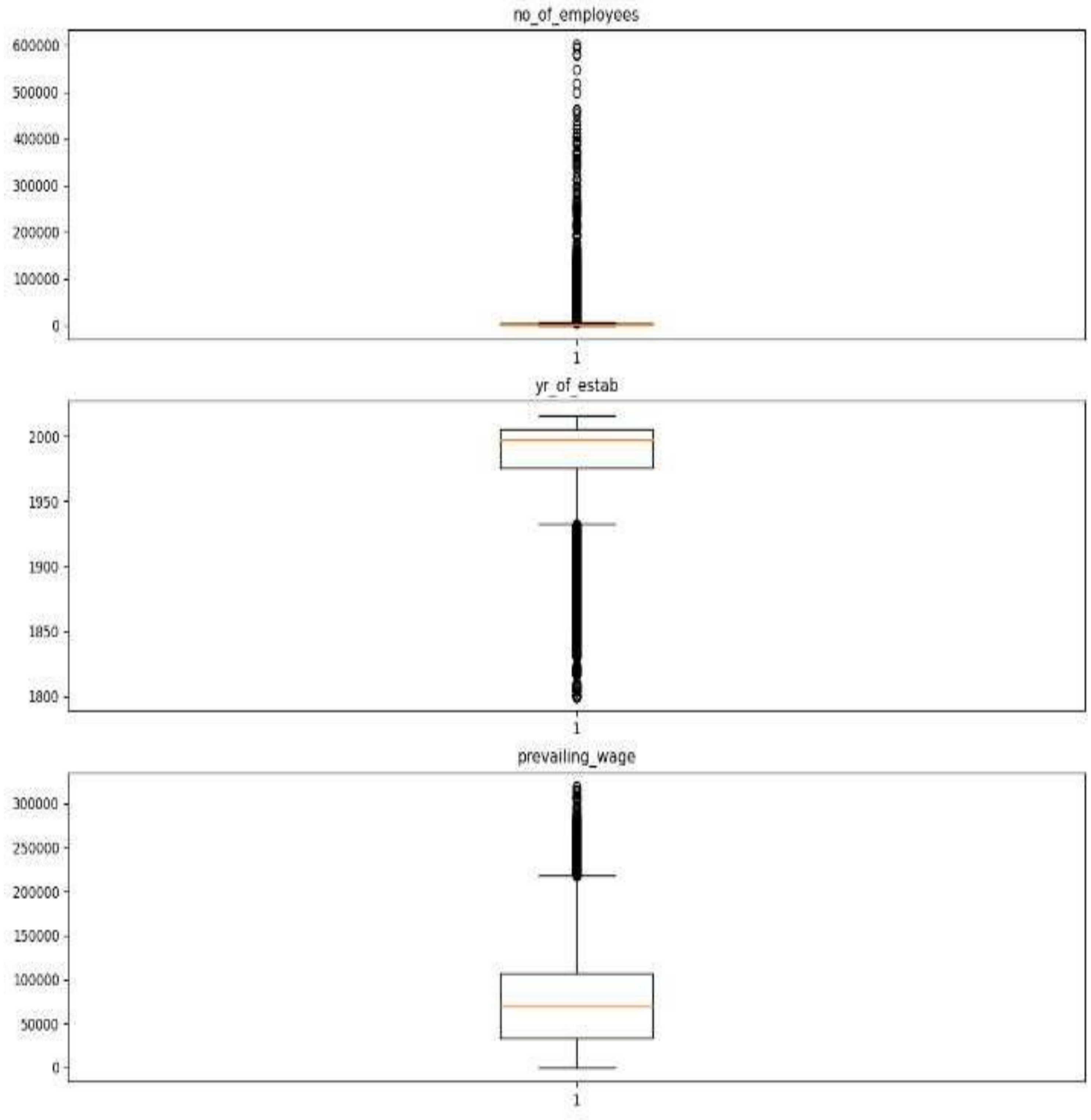
The prevailing wage has different units (Hourly, Weekly, etc.). Let's find out if it has any impact on visa applications getting certified.

1.7 Data Pre−processing

Outlier Check

Let's check for outliers in the data.

Table 25: Outliers in the data



Data Preparation for modelling

- We want to predict which visa will be certified.
- Before we proceed to build a model, we'll have to encode categorical features.
- We'll split the data into train and test to be able to evaluate the model that we build on the train data.

Model evaluation criterion

Model can make wrong predictions as:

1. Model predicts that the visa application will get certified but in reality, the visa application should get denied.
2. Model predicts that the visa application will not get certified but in reality, the visa application should get certified.

Which case is more important?

- Both the cases are important as:
- If a visa is certified when it had to be denied a wrong employee will get the job position while US citizens will miss the opportunity to work onthat position.

- If a visa is denied when it had to be certified the U.S. will lose a suitable human resource that can contribute to the economy.

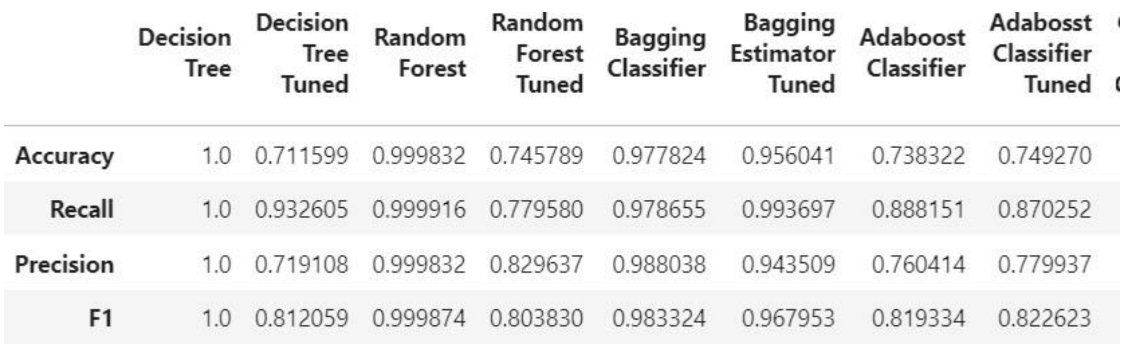
How to reduce the losses?

- F1 Score can be used the metric for evaluation of the model, greater the F1 score higher are the chances of minimizing False Negatives and FalsePositives.
- We will use balanced class weights so that model focuses equally on both classes.

1.8 Model Performance Comparison and Final Model Selection

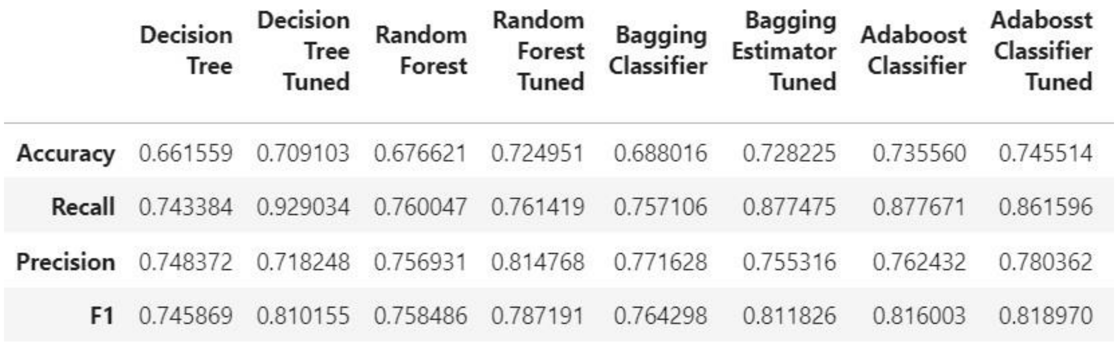
- Hyperparameter tuning has decreased the over fit and increased F1 score, however, this model is not performing as optimally as thehyperparameter tuned decision tree
- Bagging classifier is also overfitting the training data
- Bagging – Hyperparameter Tuning is still found to over fit the training data, as the training metrics are high but the testing metrics are not
- Unlike the decision tree, random forest, or the bagging classifier; the AdaBoost classifier is not found to over fit the training data. It is giving ageneralized performance on the training & testing data with a F1 score 0.819 & 0.816
- The hyperparameter tuned model is giving similar performance to the default AdaBoost model
- There is not much difference in the model performance after hyperparameter tuning
- Decision tree, Random forest (default & tuned) & Bagging classifier (default & tuned) were found to over fit the training dataset

- Decision tree tuned and Adaboost (default & tuned) were found to give generalized performance on the training & testing data sets.



	Decision Tree	Decision Tree Tuned	Random Forest	Random Forest Tuned	Bagging Classifier	Bagging Estimator Tuned	Adaboost Classifier	Adabosst Classifier Tuned
Accuracy	1.0	0.711599	0.999832	0.745789	0.977824	0.956041	0.738322	0.749270
Recall	1.0	0.932605	0.999916	0.779580	0.978655	0.993697	0.888151	0.870252
Precision	1.0	0.719108	0.999832	0.829637	0.988038	0.943509	0.760414	0.779937
F1	1.0	0.812059	0.999874	0.803830	0.983324	0.967953	0.819334	0.822623

Table 26: Training performance comparison



	Decision Tree	Decision Tree Tuned	Random Forest	Random Forest Tuned	Bagging Classifier	Bagging Estimator Tuned	Adaboost Classifier	Adabosst Classifier Tuned
Accuracy	0.661559	0.709103	0.676621	0.724951	0.688016	0.728225	0.735560	0.745514
Recall	0.743384	0.929034	0.760047	0.761419	0.757106	0.877475	0.877671	0.861596
Precision	0.748372	0.718248	0.756931	0.814768	0.771628	0.755316	0.762432	0.780362
F1	0.745869	0.810155	0.758486	0.787191	0.764298	0.811826	0.816003	0.818970

Table 27: Testing performance comparison

  - mind, and perhaps a revaluation of cases getting denied can be carried out in case there is a prevailing human resource shortage in the US. Themodel is still helpful, as only a small subset of data will need further re−evaluation significantly decreases time spent in the process

1.9 Actionable Insights and Recommendations

- Based on the EDA the following features were identified as important for visas getting certified than denied

  - (1) Education of employee; an employee with only a high school certification has over 65% chance of visa getting denied in comparison to anemployee with a doctorate degree with over a 85% chance of visa getting certified
  - (2) Unit of wage; an employee with an hourly pay likewise has over 65% chance of visa getting denied in comparison to an employee with anon−hourly pay (weekly, monthly or yearly) with over 70% chance of visa getting certified
  - (3) The continent the employee is from (e.g., if Europe, over 80% chance of visa getting certified), if the employee has prior job experience (over 75% chance of visa getting approved if an employee has prior work experience but 50% chance of visa getting denied if an employee hasno work experience) are other important attributes
  - (4) Likewise, the region of the US the employment opportunity is in is also an important deciding factor with over 70% cases getting certified ifthe region is Midwest or South
- Interestingly, attributes like if the job opportunity is full time/ part time; if an employee requires further job training; the annual prevailing wage of the occupation in the US; year of establishment of the employer or the number of employees in the organization are not important attributes & do not havemuch bearing on a case getting certified vs denied

- The hyperparameter tuned ML model is able to give generalized prediction on training & testing datasets (not prone to overfitting) and is able to explainover 80% of information (accuracy of 75% on test dataset & F1 score of 82% on test dataset).

  - The precision & recall are likewise both high (77% & 88% respectively)
  - The confusion matrix is able to identify a higher % of cases getting certified, but only a smaller % of cases getting denied correctly. This limitation has to be borne in mind, and perhaps a revaluation of cases getting denied can be carried out in case there is a prevailing human resource shortage in the US. The model is still helpful, as only a small subset of data will need further re−evaluation significantly decreases timespent in the process