GREAT LEARNING

PROJECT ON MACHINE LEARNING -2

Project designer: *DURGESH KUMAR JHA*

Session: 2024-25

From the business perspective we have to find the driving factors which influences the trends which is affecting the approval rate for the visa for the different categories of the employees the data so we can provide the recommendation to the Office of Foreign Labor Certification (OFLS).

PROBLEM STATEMENT

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.

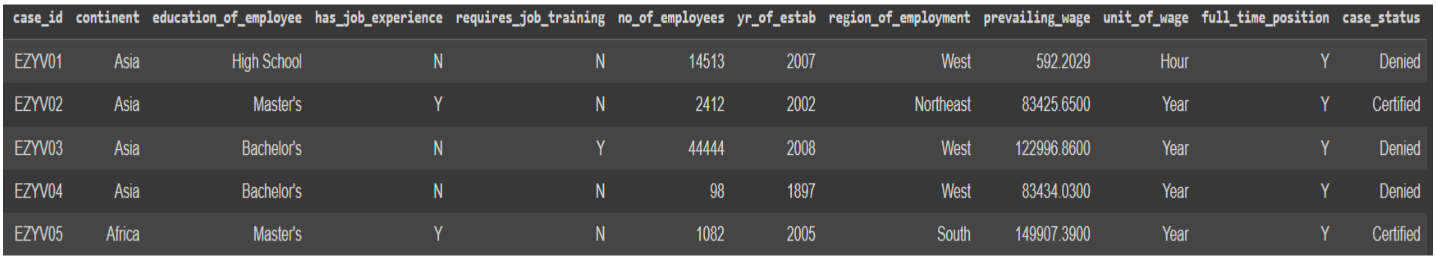
2. Recommend a suitable profile for the applicants for whom the visa should be certified or denied based on the drivers that significantly influence the case status.

DATA DESCRIPTION

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

List of Image

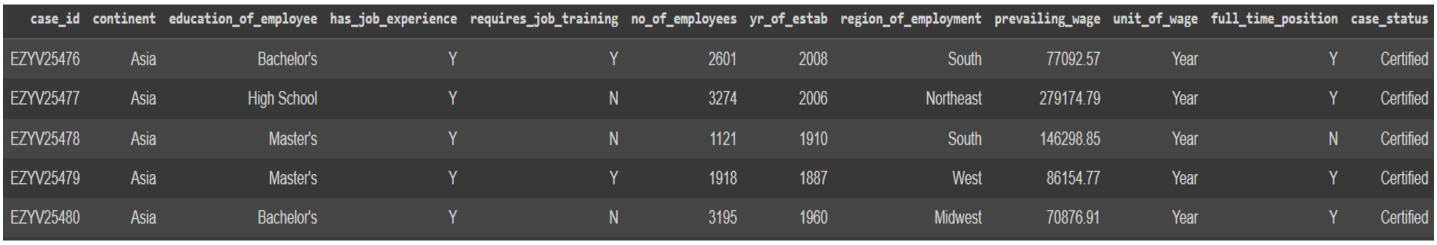
Data Overview



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

Image 1

Above Image shows top 5 rows of the data set.



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
EZYV25476	Asia	Bachelor's	Y	Y	2601	2008	South	77092.57	Year	Y	Certified
EZYV25477	Asia	High School	Y	N	3274	2006	Northeast	279174.79	Year	Y	Certified
EZYV25478	Asia	Master's	Y	N	1121	1910	South	146298.85	Year	N	Certified
EZYV25479	Asia	Master's	Y	Y	1918	1887	West	86154.77	Year	Y	Certified
EZYV25480	Asia	Bachelor's	Y	N	3195	1960	Midwest	70876.91	Year	Y	Certified

Image 2

Above image shows bottom 5 rows of the data set.

```
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
```

Image 3

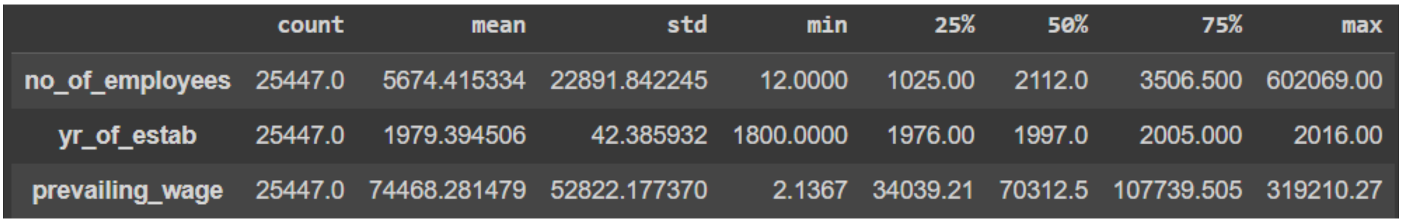
```
(25480, 12)
```

Above image shows that we have total 12 columns and 25480 rows.

Above image also shows that we have no missing values available in our data set.

We have total 9 columns with object data type, 2 as integers, and 1 with float type.

Our target variable is of object type.



	count	mean	std	min	25%	50%	75%	max
no_of_employees	25447.0	5674.415334	22891.842245	12.0000	1025.00	2112.0	3506.500	602069.00
yr_of_estab	25447.0	1979.394506	42.385932	1800.0000	1976.00	1997.0	2005.000	2016.00
prevailing_wage	25447.0	74468.281479	52822.177370	2.1367	34039.21	70312.5	107739.505	319210.27

Image 4

Above image shows the statistical summary of the data set.

Above image show that the average prevailing wage for the employees is around 70K.

```
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
```

Image 5

Above image shows we not have any duplicate values in our data set.

Exploratory Data Analysis Univariate analysis

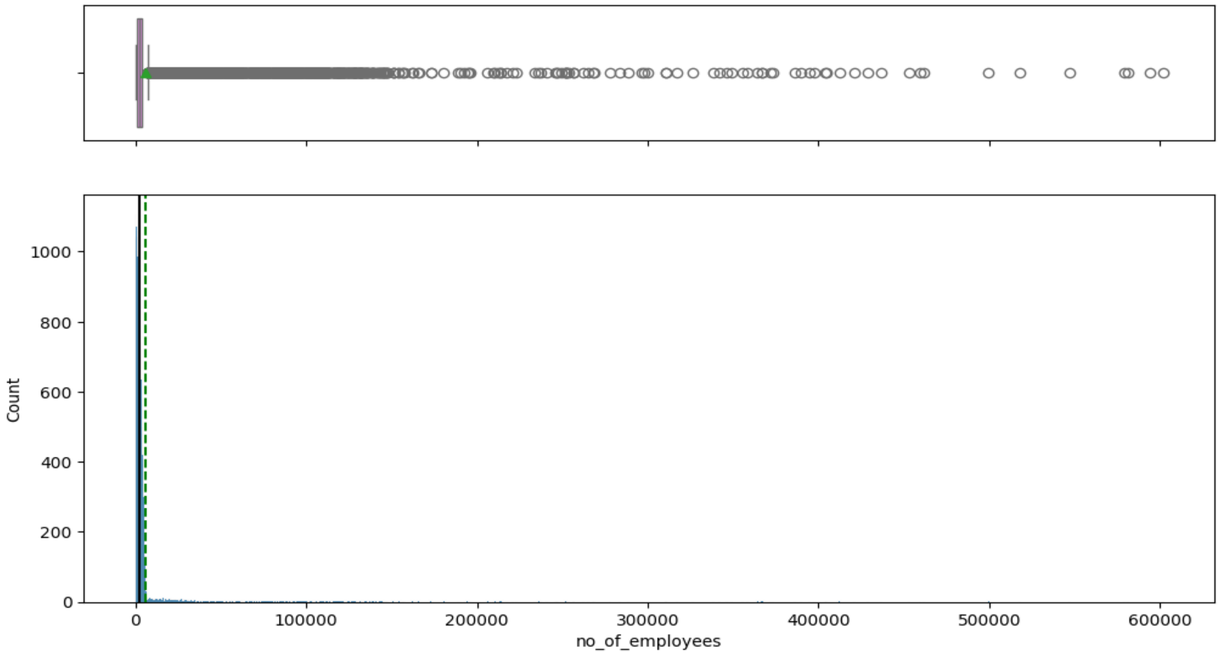


Image 6

Above image shows that for the number of employees our data is heavily right skewed.

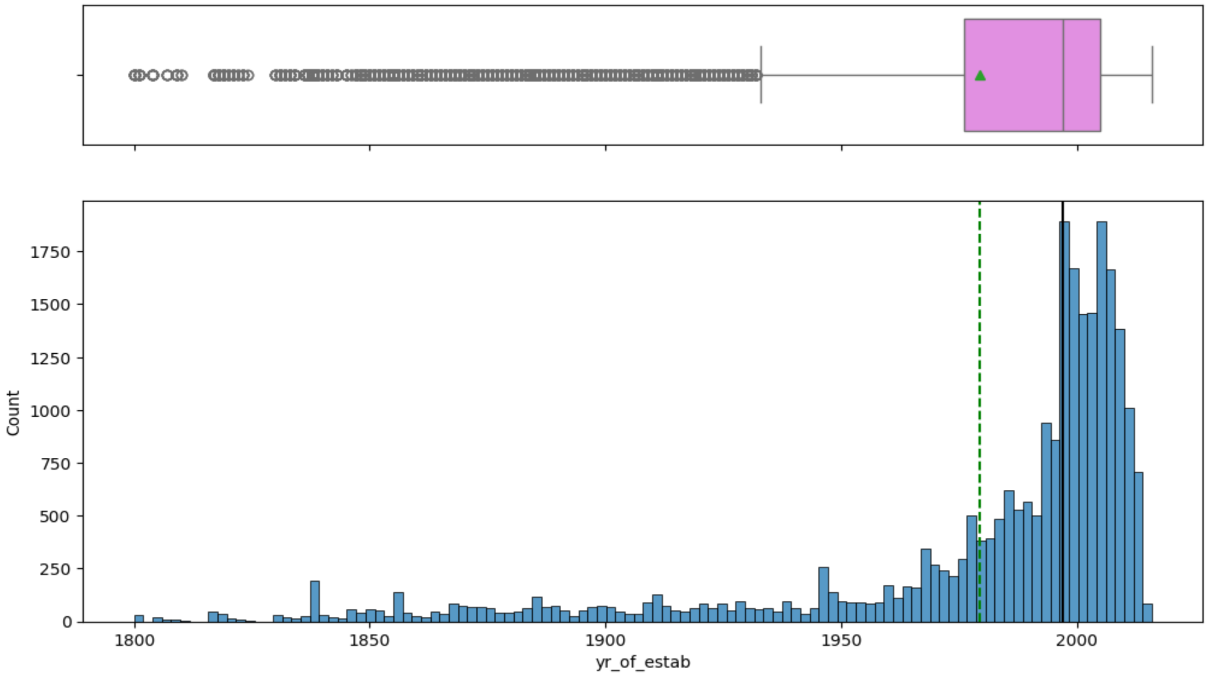


Image 7

Above image show that for the establishment year for the industries are left skewed.

Above image shows mostly the companies hires work force from foreign countries are newly established.

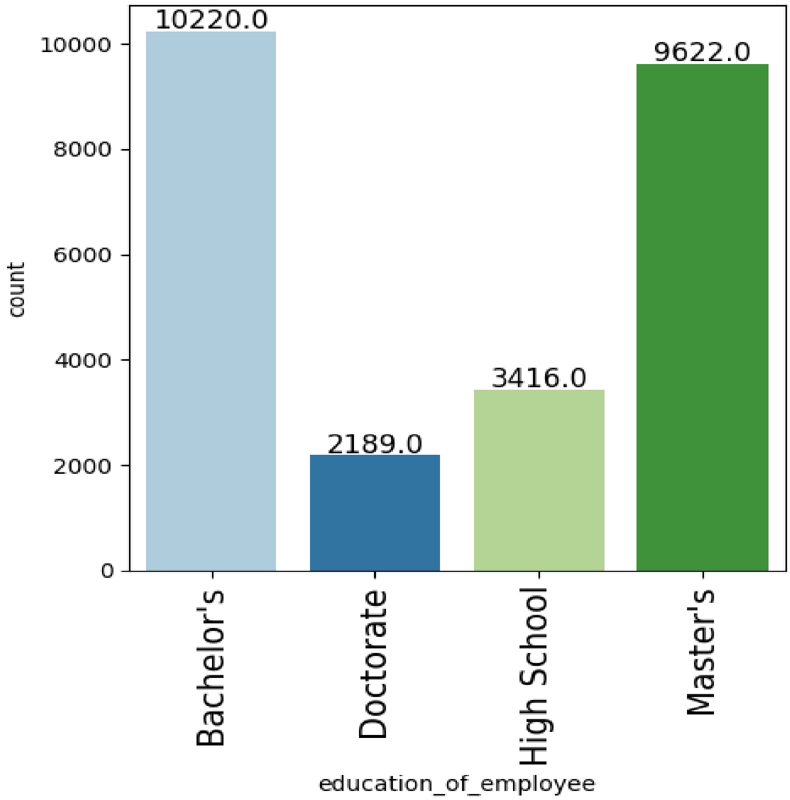


Image 8

Above image shows mostly the people apply for the visa are who have completed their Bachelor's, followed by Master's then High School lastly Doctorate.

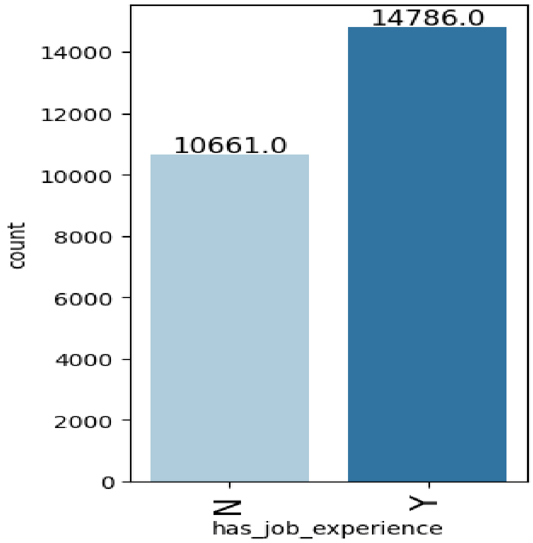


Image 9

Above image shows mostly the persons apply for the visa and jobs are freshers with no work experience.

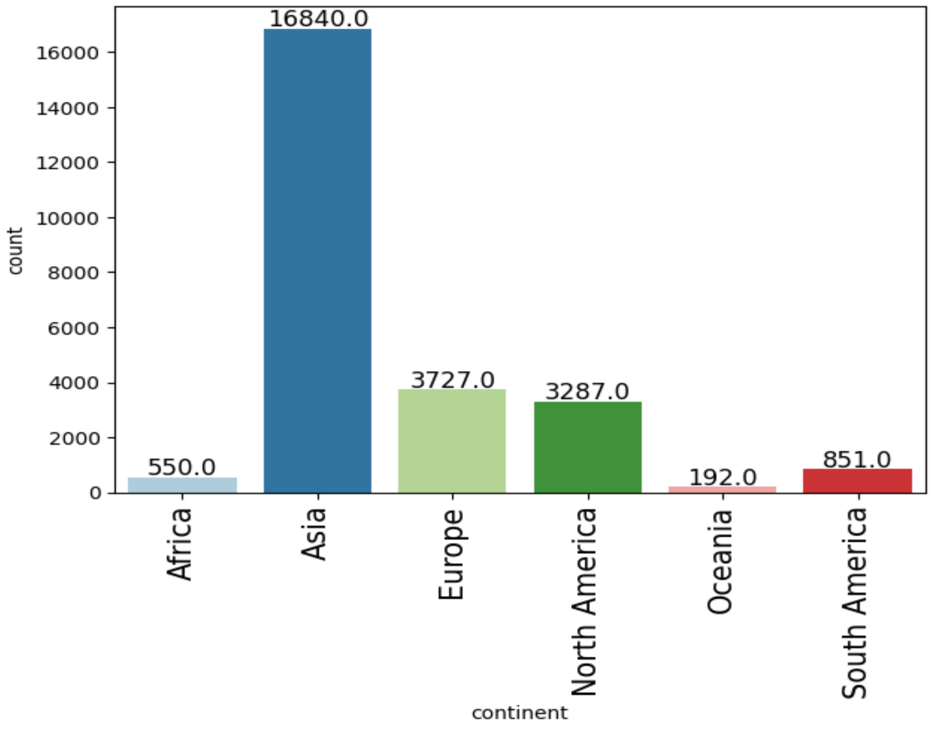


Image 10

Above image shows mostly the Employees are from Asia they are around 66% of total applied for visa followed by Europe and North America.

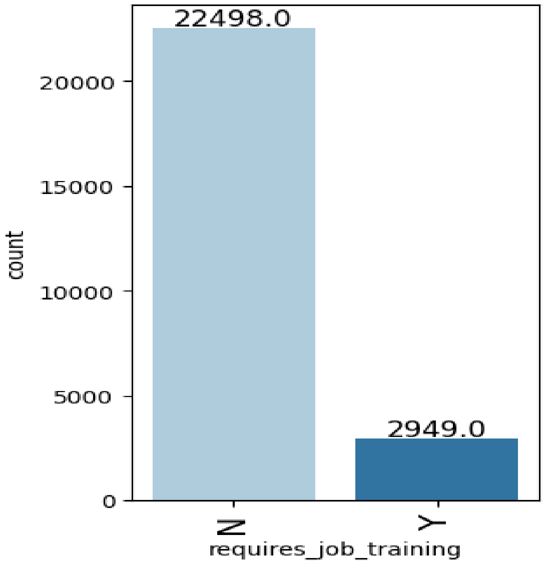


Image 11

Above image shows mostly in the jobs they not need any kind of specific training.

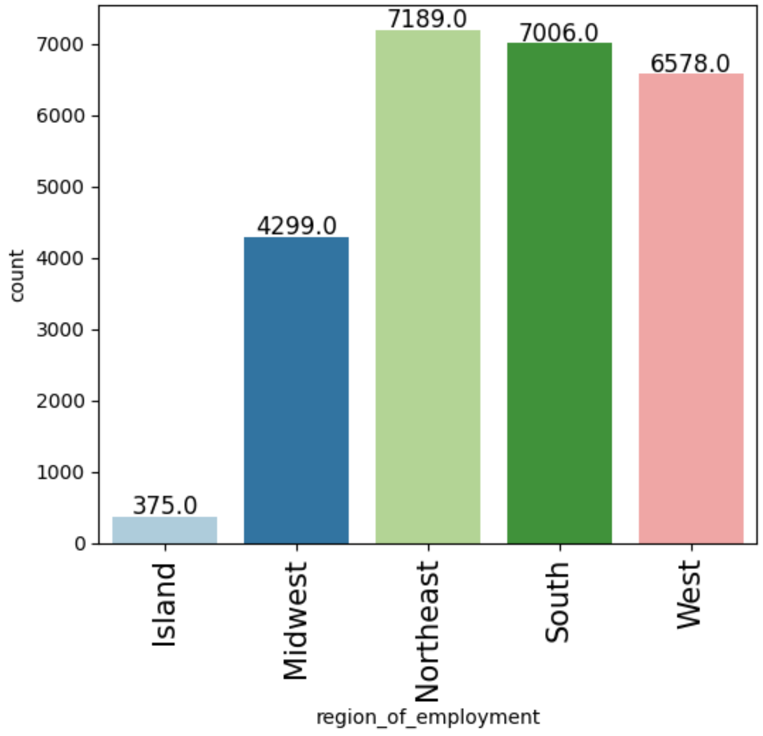


Image 12

Above image shows mostly the employment is given in the Northeastern side, then followed by South, West side of the country which has mostly needs for Human Resource.

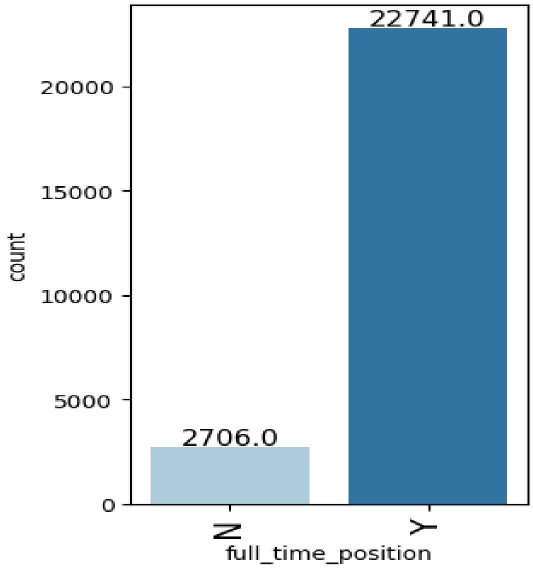


Image 13

Above image shows mostly the industries are providing full time position which is around 88%.

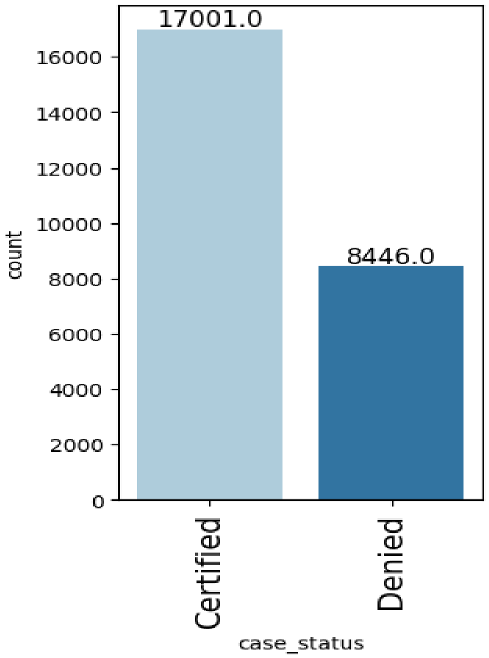


Image 14

Above image shows only around 66% cases are approved and roughly 34% are denied.

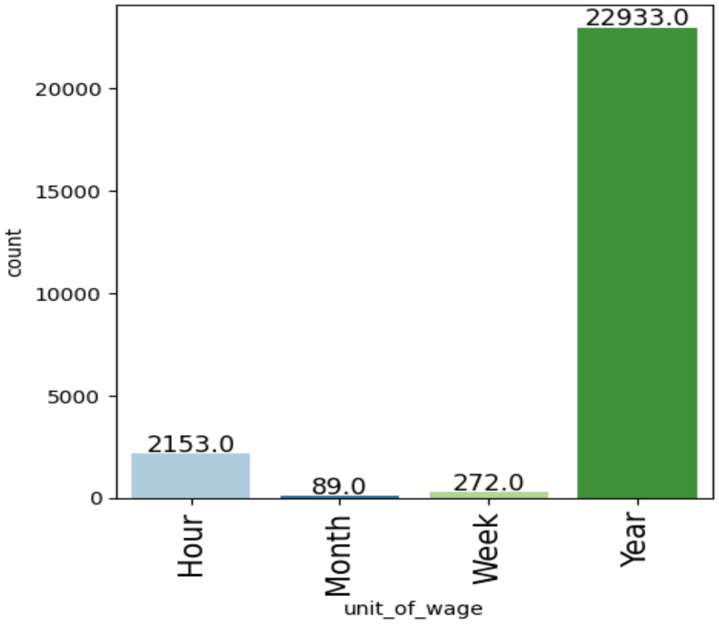


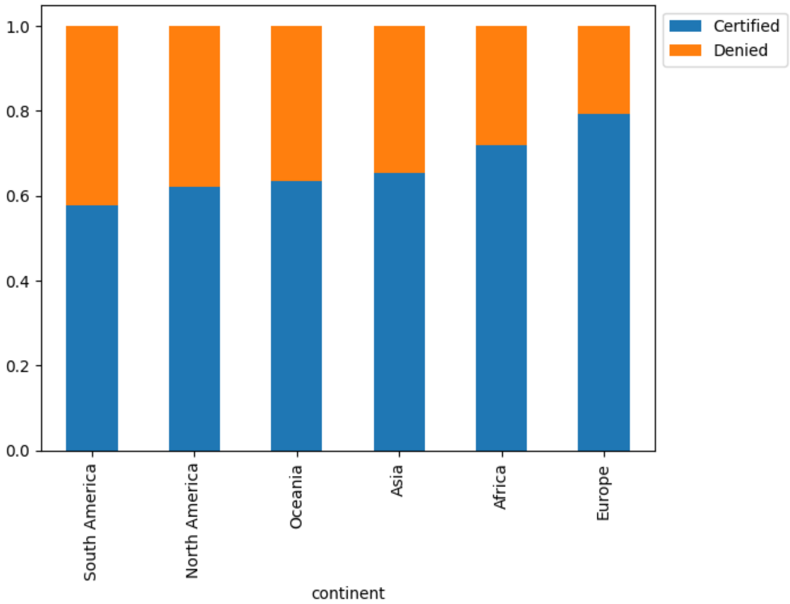
Image 15

Above image shows mostly they offers the compensation to their employees in Yearly basis around 97%.

Indicates they mostly companies prefers the employees who is willing to be in contract around a year.

Bivariate Distributions

```
case_status    Certified  Denied    All
continent
All                17001    8446  25447
Asia               11001    5839  16840
North America       2037    1250   3287
Europe              2953     774   3727
South America        492     359    851
```



```
Africa                    396      154     550
Oceania                   122       70     192
```

Image 16

Above image shows mostly the request for the visa are being approved with the highest rate from Europe > Africa > Asia > Oceania > North America > South America.

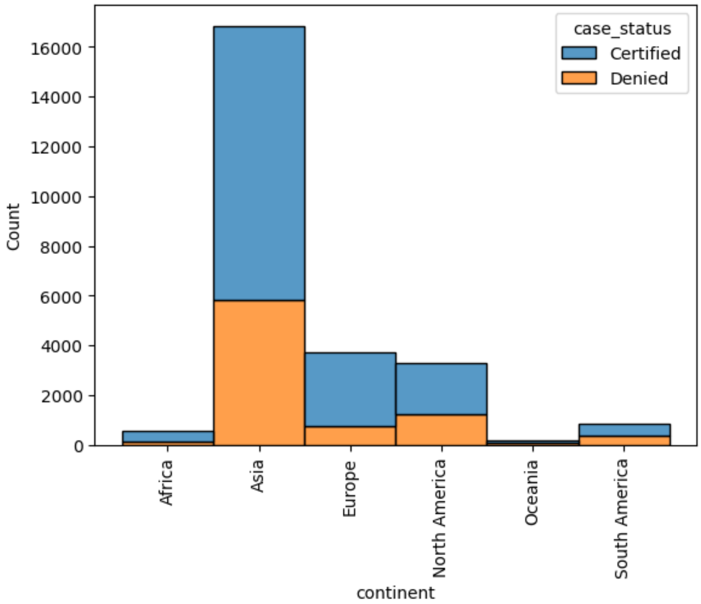


Image 17

Above both images show mostly the approval rate for visa is for Europe but then with number of counts Asian have the number of visas approved.

```
case_status             Certified  Denied    All
education_of_employee
All                         17001    8446  25447
Bachelor's                   6362    3858  10220
High School                  1164    2252   3416
Master's                     7565    2057   9622
Doctorate                    1910     279   2189
------------------------------------------------------------------------------------------------------------------
```

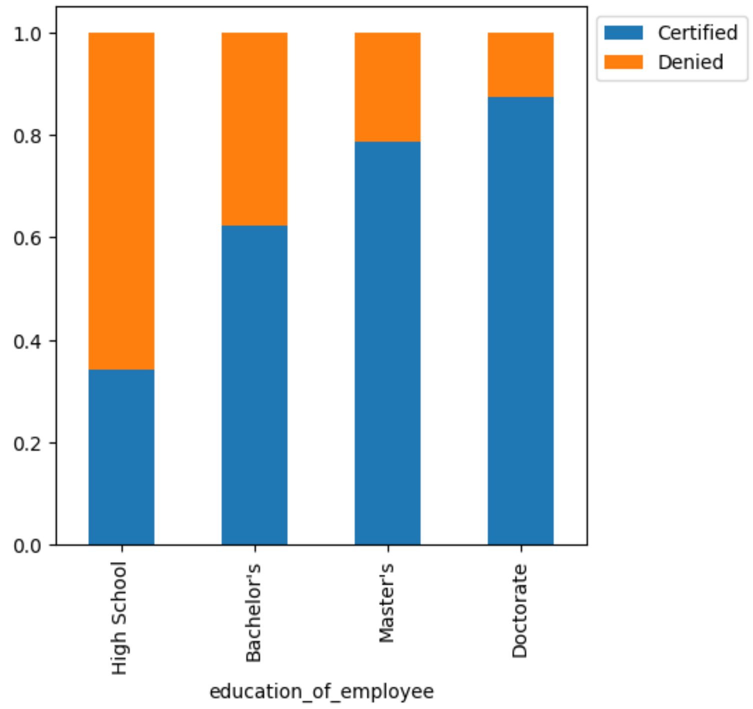


Image 18

Above image shows the mostly the approval rate for the visas are for employee with the higher education mostly Doctorate > Master’s > Bachelor’s > High School.

```
case_status         Certified  Denied    All
full_time_position
All                     17001    8446  25447
Y                       15146    7595  22741
N                        1855     851   2706
------------------------------------------------------------------------------------------
-----------------------------------
```

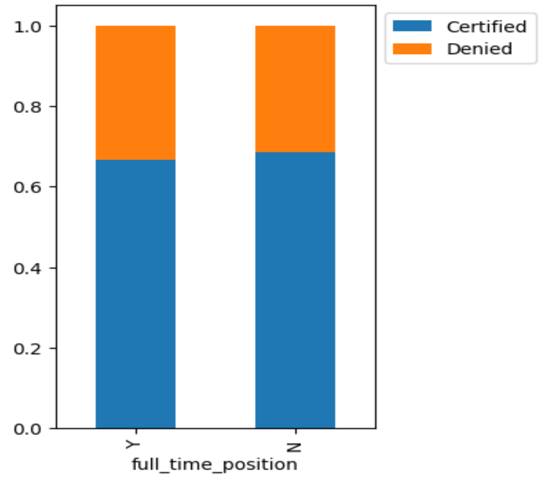


Image 19

Above image shows visa approval rate has not issue with full time position and part time position.

```
case_status          Certified  Denied    All
region_of_employment
All                      17001    8446  25447
```

```
Northeast                 4524    2665   7189
West                      4097    2481   6578
South                     4908    2098   7006
Midwest                   3246    1053   4299
Island                     226     149    375
---------------------------------------------------------------------------------------
----------------------------------
```

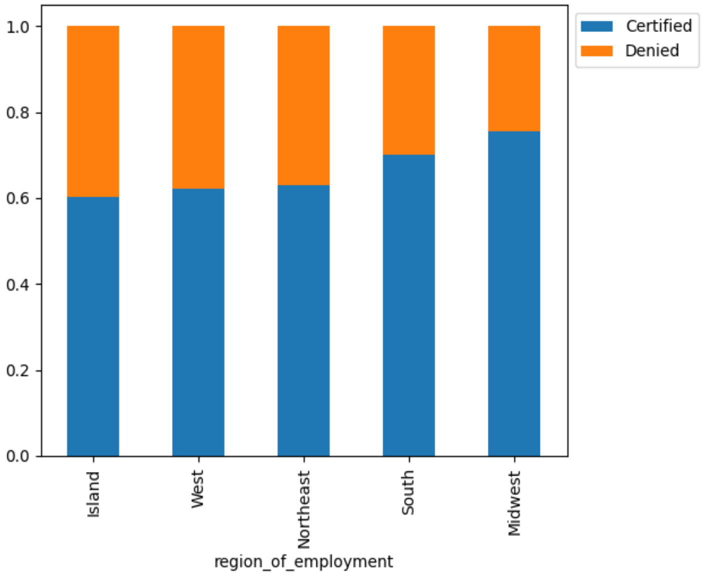


Image 20

Above image shows mostly for the Midwest region Visa are approved followed by South and Northeast, as they are in mostly needs of Human Resourse.

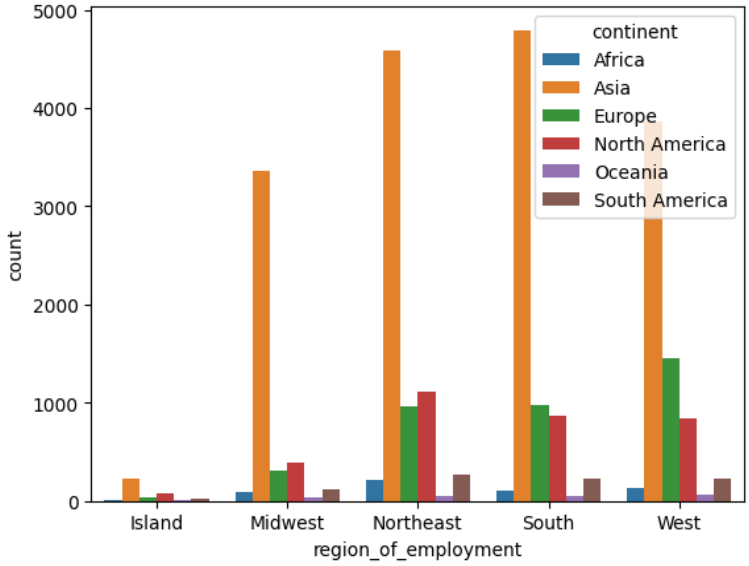


Image 21

Above image shows the distribution of the employees from the different continent region wise.

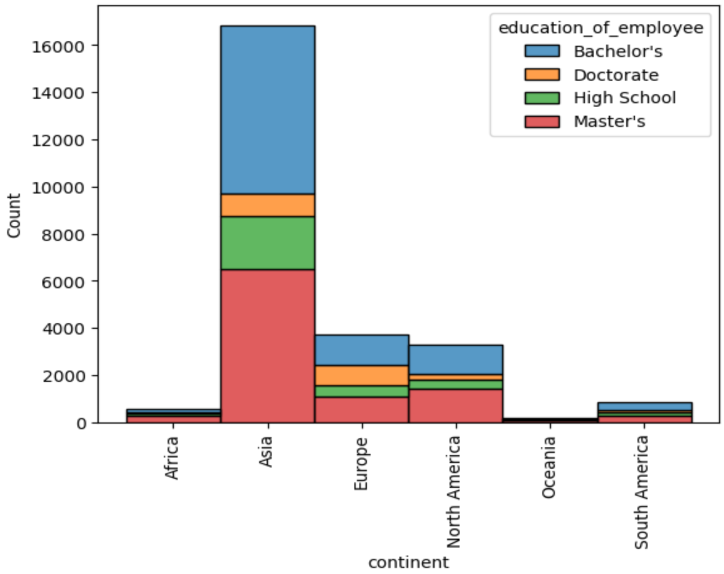


Image 22

Above image shows the employees with their education and their respective continents.

```
case_status   Certified  Denied    All
unit_of_wage
All               17001    8446  25447
Year              16030    6903  22933
Hour                747    1406   2153
Week                169     103    272
Month                55      34     89
------------------------------------------------------------------------------------------
----------------------------------
```

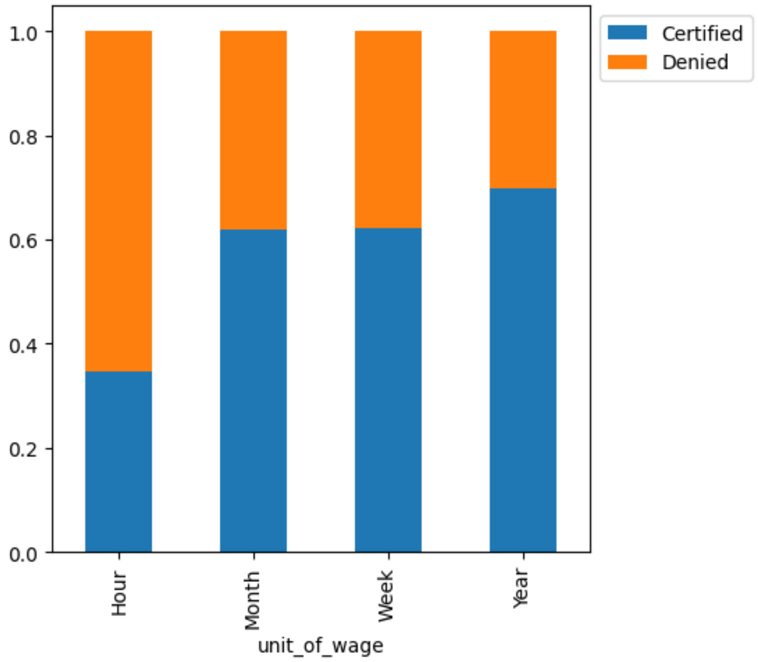


Image 23

Above image shows the approval rate for the visa is higher for the Yearly wages then the weekly followed by Monthly and Hourly.

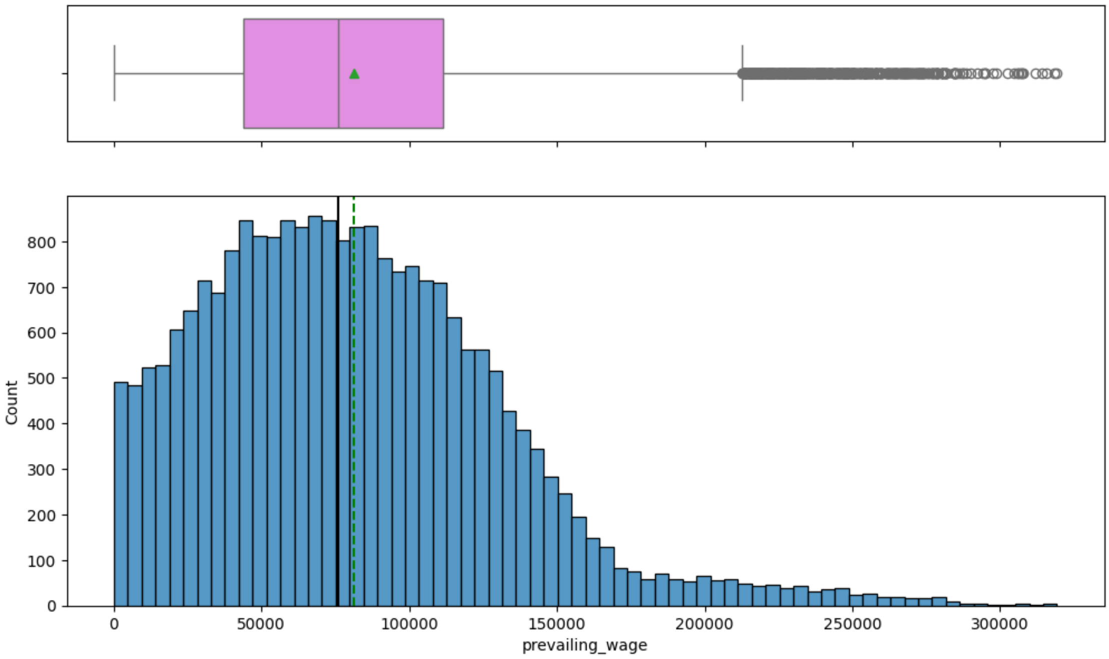


Image 24

Above image shows the average wage for the employees is around 70K on yearly basis.

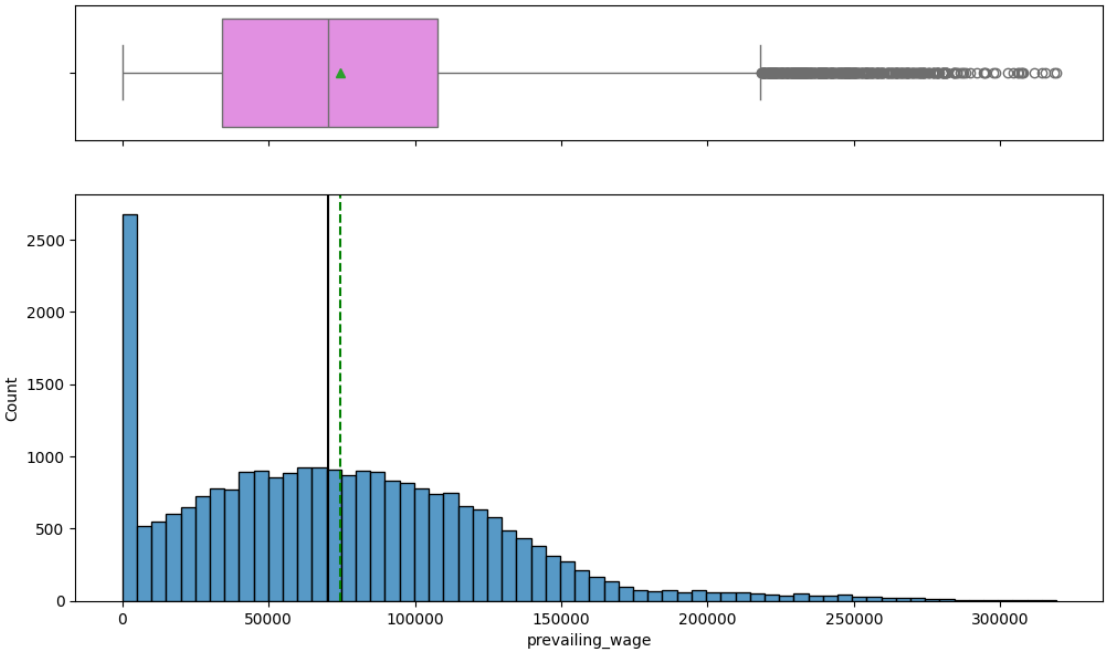


Image 25

Above image shows there are some outliers in upper and lower end which needs further reserch.

Data Preprocessing

```
((17812, 21), (7635, 21))
```

We have total 17812 rows and 21 columns in the training set and total 7635 rows and 21 columns in our test data.

```
case_status
1    0.668094
0    0.331906


case_status
1    0.668089
0    0.331911


case_status
1    0.668107
0    0.331893
```

We would want F1-Score to be maximized, the greater the F1-Score higher the chances of predicting both the classes correctly

Decision Tree Classifier

```
Training performance:
    Accuracy  Recall  Precision   F1
0       1.0     1.0        1.0  1.0
Testing performance:
    Accuracy    Recall  Precision        F1
0  0.657367  0.735934   0.747362  0.741604
```

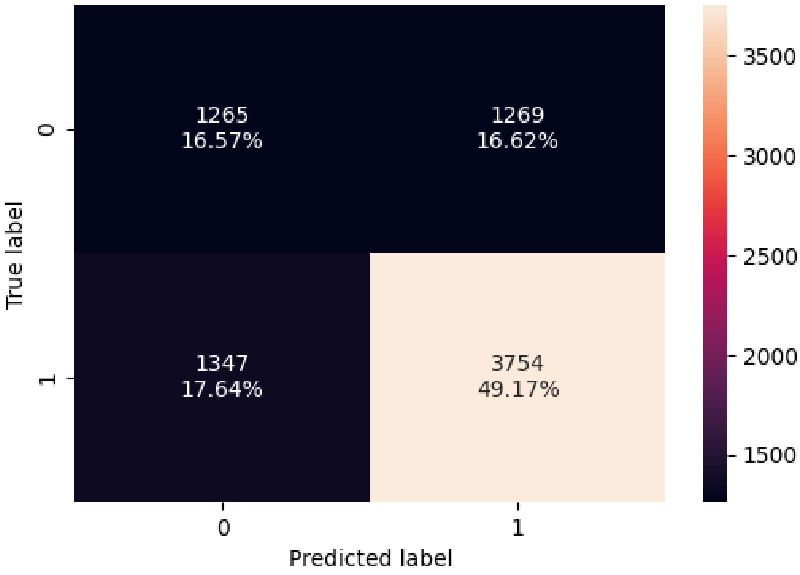


Image 26

The above image show that the model we have prepared is overfitting as our F1 score in training data is perfect but in our testing data it is around 0.74

So, we need to improve our model performance by hyperparameter tuning

Decision Tree - Hyperparameter tuning

```
Training performance:
    Accuracy    Recall  Precision        F1
0  0.711599  0.932605   0.719108  0.812059
Testing performance:
    Accuracy    Recall  Precision        F1
0  0.709103  0.929034   0.718248  0.810155
```

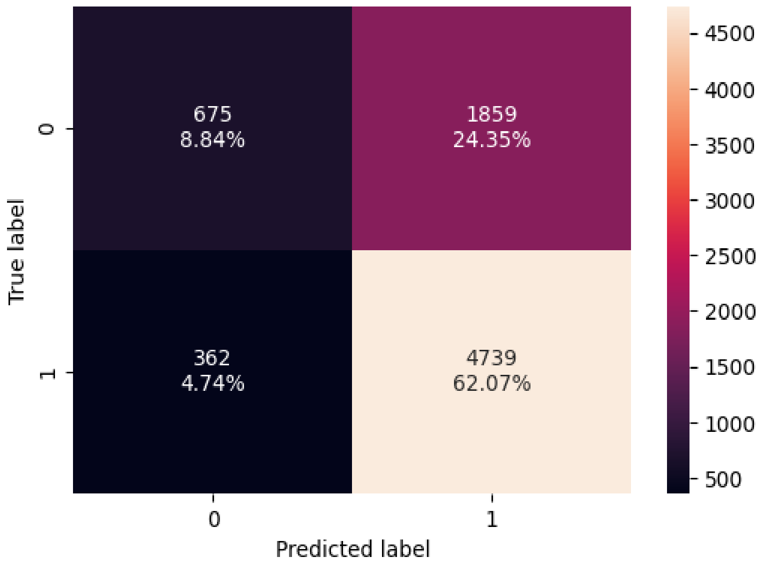


Image 27

As now we have use the hyperparameter tuning our decision tree is not overfitting the data set. As our F1 score has improved for both training and test set

Random Forest Classifier

```
Training performance:
    Accuracy  Recall  Precision   F1
0        1.0     1.0        1.0  1.0
Testing performance:
    Accuracy    Recall  Precision        F1
0  0.721022  0.832974   0.768771  0.799586
```

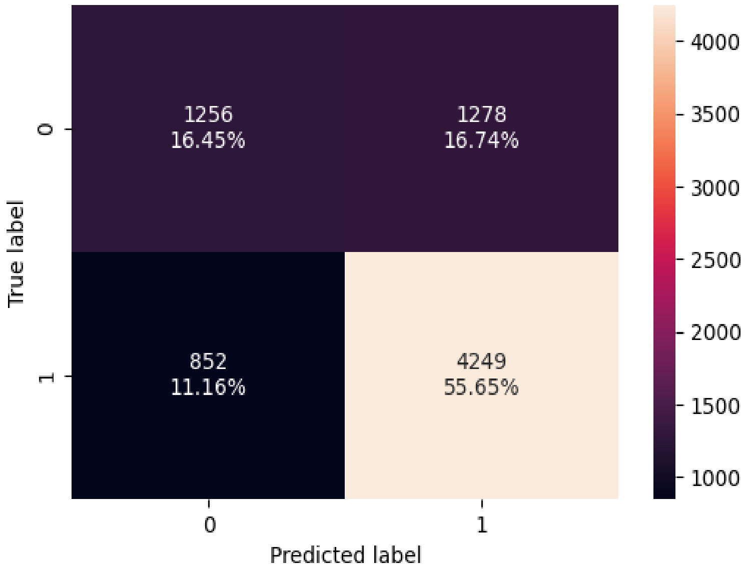


Image 28

The above image shows the model we have built with help of our random forest classifier is overfitting as our F1 score in training set is perfect but, in our testing, set it is around 0.79.

Random Forest - Hyperparameter tuning

```
Training performance:
    Accuracy    Recall  Precision        F1
0  0.895857  0.899076   0.942394  0.920225
```

```
Testing performance:
    Accuracy    Recall  Precision        F1
0  0.724427  0.790433   0.795737  0.793076
```

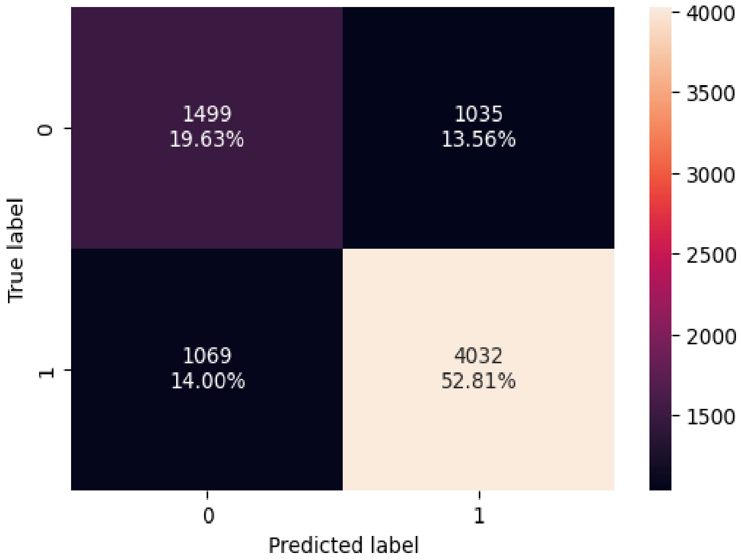


Image 29

As we have tuned the model it has improved slightly the performance as in our training set it is around 0.92 and in test set it is 0.79.

Our performance has not improved which is needed with tuning the model.

Bagging Classifier

```
Training performance:
    Accuracy    Recall  Precision        F1
0  0.984673  0.985882    0.99113  0.988499
Testing performance:
    Accuracy    Recall  Precision       F1
0  0.701244  0.779259   0.774854  0.77705
```

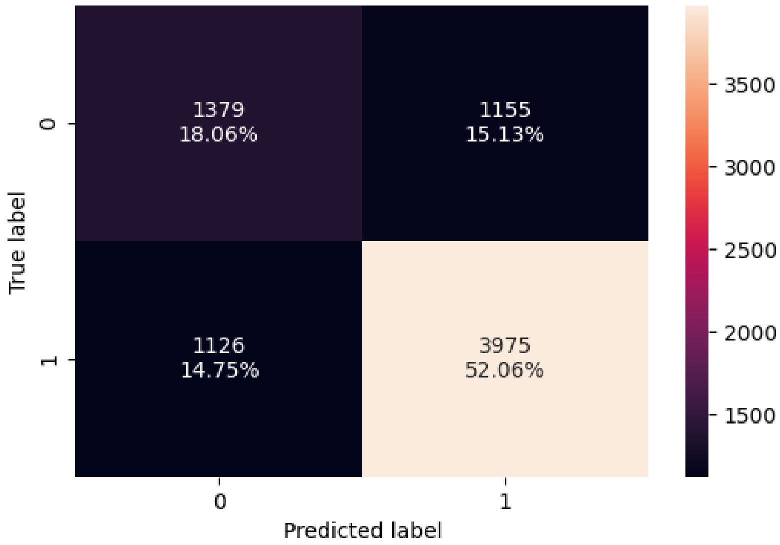


Image 30

The above image shows the model is overfitting as it is working fine in training set with F1 score with 0.98 but in the testing set it is only around 0.77.

Bagging- Hyperparameter Tuning

```
Training performance:
    Accuracy    Recall  Precision        F1
0  0.989894  0.999412   0.985662  0.992489
Testing performance:
    Accuracy    Recall  Precision        F1
0    0.7222  0.887865   0.745146  0.810269
```

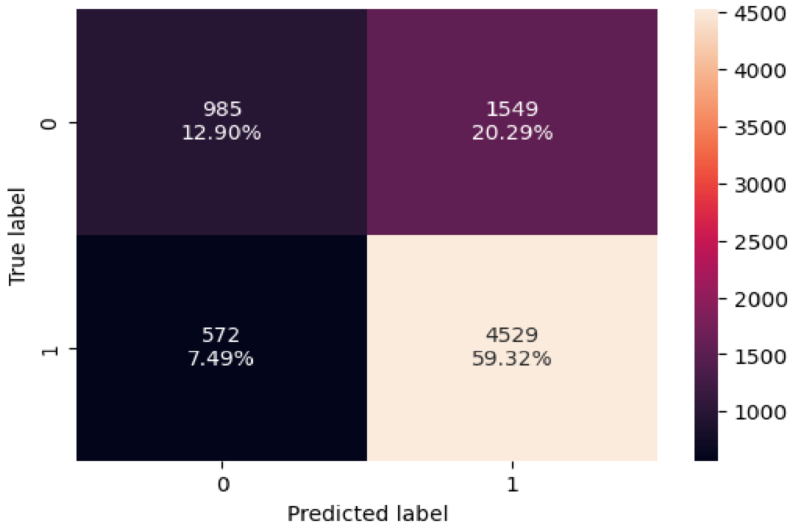


Image 31

As we have tuned the model it has improved our model performance as in our training set our F1 score is around 0.99 and in test it is 0.81.

AdaBoost Classifier

```
Training performance:
    Accuracy   Recall  Precision        F1
0  0.740568  0.89084   0.761402  0.821051
Testing performance:
    Accuracy    Recall  Precision        F1
0  0.733857  0.877475   0.760836  0.815004
```

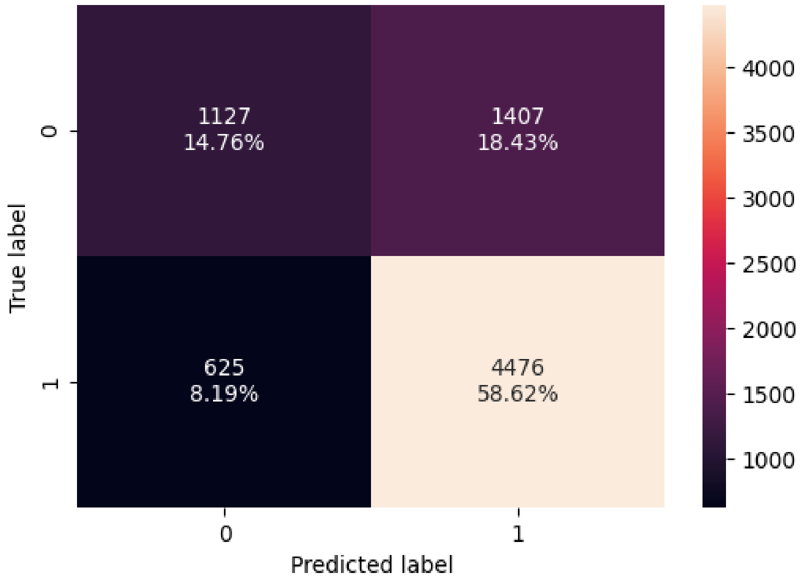


Image 32

Above image shows our performance in training set and test for AdaBoost is compatible with is around 0.82 in training set and 0.81 in our testing set.

AdaBoost - Hyperparamter tuning

```
Training performance:
    Accuracy    Recall  Precision        F1
0  0.750337  0.877143   0.777621  0.824389
Testing performance:
    Accuracy    Recall  Precision        F1
0  0.743026  0.862772    0.77715  0.817726
```

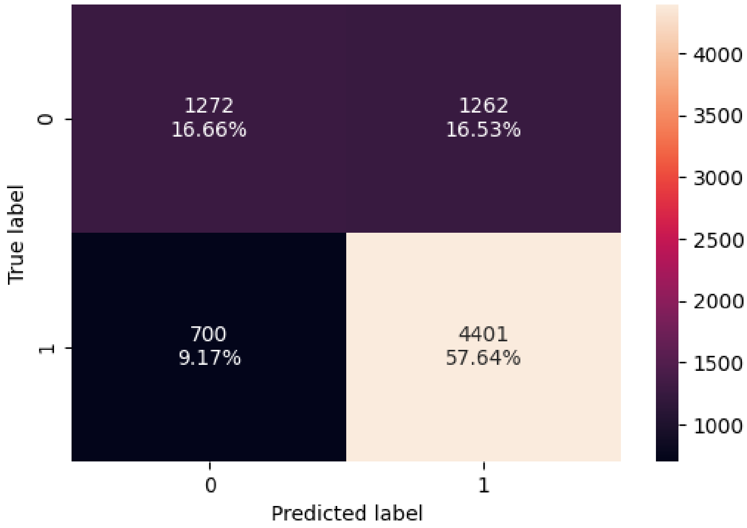


Image 33

After tuning the model also, we have the same result as before tuning the model for AdaBoost model

Gradient Boosting Classifier

```
Training performance:
    Accuracy    Recall  Precision        F1
0  0.757242  0.880504   0.783109  0.828956
Testing performance:
    Accuracy    Recall  Precision       F1
0   0.74761  0.867869   0.779401  0.82126
```

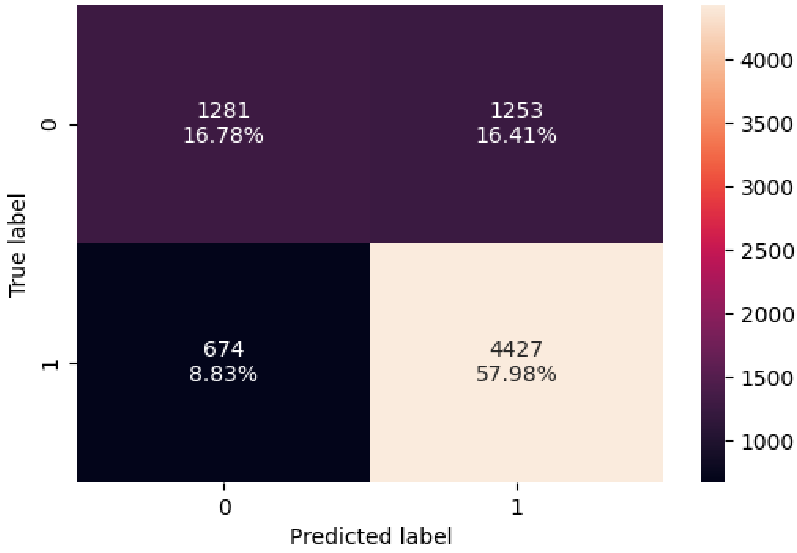


Image 34

Above image shows Gradient Boosting model in which F1 is very high score in training set is 0.82 and in test set it is almost same 0.82

Gradient Boosting - Hyperparameter Tuning

```
Training performance:
    Accuracy    Recall  Precision        F1
0  0.757804  0.879496   0.784205  0.829121
Testing performance:
    Accuracy    Recall  Precision        F1
0  0.748265  0.866497   0.780781  0.821409
```

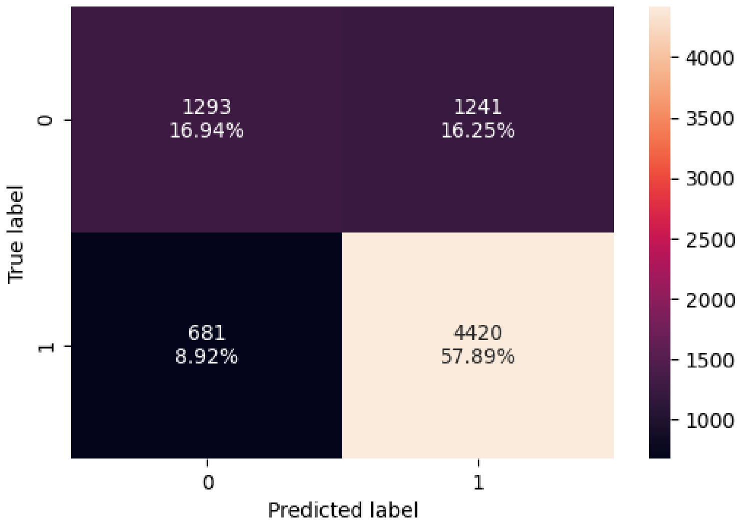


Image 35

As we have tuned the Gradient Boosting model but there is not much change in the model performance

XGBoost Classifier

```
Training performance:
    Accuracy    Recall  Precision        F1
0  0.840108  0.932017   0.844707  0.886217
Testing performance:
    Accuracy    Recall  Precision        F1
0  0.730845  0.850618   0.770419  0.808534
```

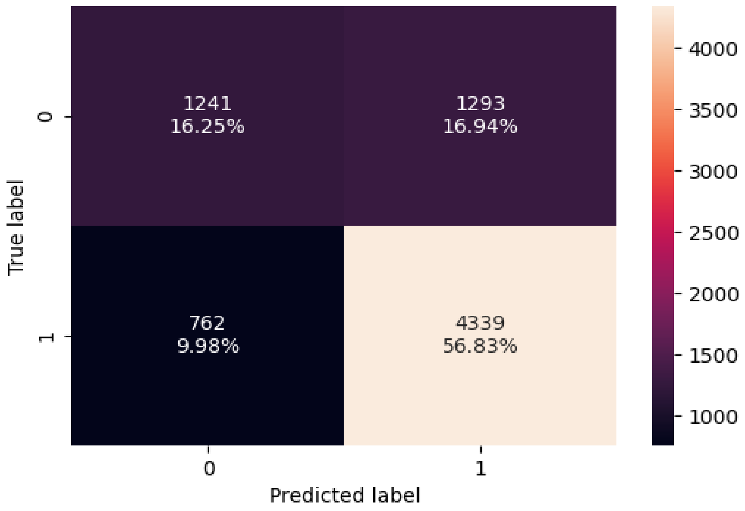


Image 36

Above image shows with XGBoost model our performance is slightly overfitting as our F1 score in training set is 0.89 and in test set it is around 0.81

```
Training performance:
    Accuracy     Recall  Precision        F1
0  0.747586  0.919664   0.755592  0.829594
Testing performance:
    Accuracy     Recall  Precision      F1
0  0.739358  0.908841   0.752475  0.8233
```

XGBoost - Hyperparameter tuning

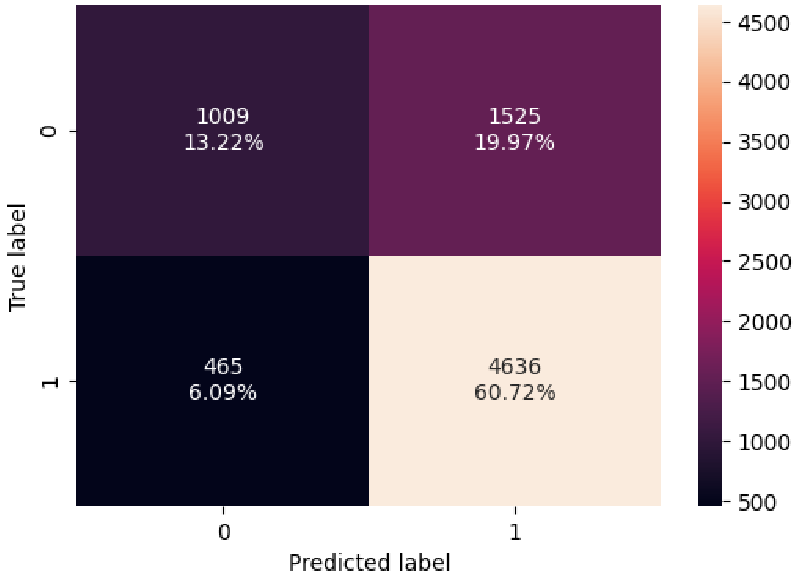


Image 37

As we have used hyperparameter tuning our performance with XGBoost has improved it is around 0.83 in training set and 0.82 in test set

Stacking Classfier

```
Training performance:
    Accuracy     Recall  Precision        F1
0  0.754491  0.894874   0.773292  0.829652
Testing performance:
    Accuracy     Recall  Precision        F1
0   0.74368  0.882964   0.768076  0.821523
```

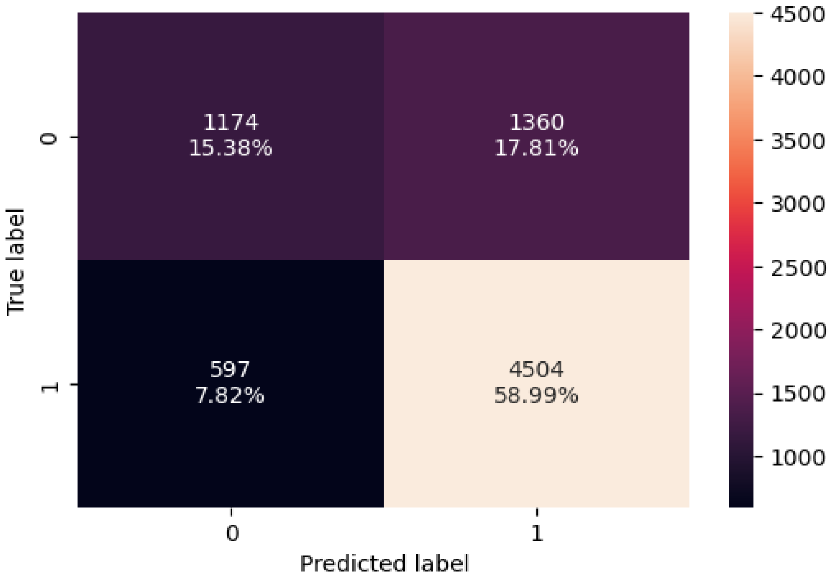
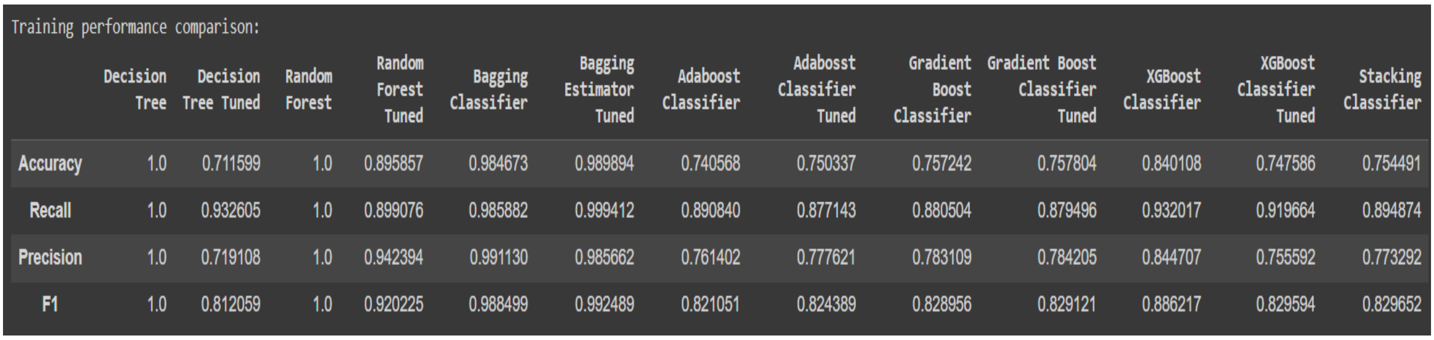


Image 38

Above image shows similar results after we tuned the model with hyperparameter as with Stacking classifier our F1 score in training set with is 0.83 test set it is 0.82

Comparing all models

`Training performance comparison:`

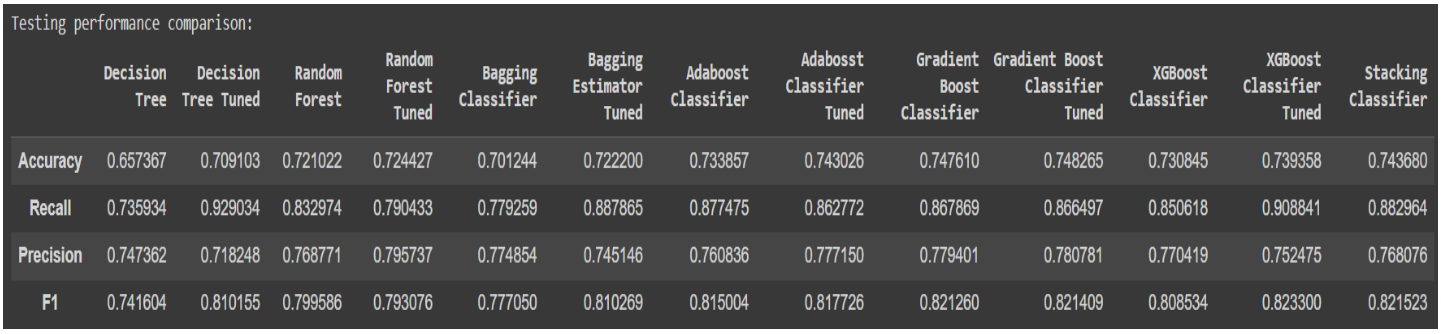


Training performance comparison:

	Decision Tree	Decision Tree Tuned	Random Forest	Random Forest Tuned	Bagging Classifier	Bagging Estimator Tuned	Adaboost Classifier	Adabosst Classifier Tuned	Gradient Boost Classifier	Gradient Boost Classifier Tuned	XGBoost Classifier	XGBoost Classifier Tuned	Stacking Classifier
Accuracy	1.0	0.711599	1.0	0.895857	0.984673	0.989894	0.740568	0.750337	0.757242	0.757804	0.840108	0.747586	0.754491
Recall	1.0	0.932605	1.0	0.899076	0.985882	0.999412	0.890840	0.877143	0.880504	0.879496	0.932017	0.919664	0.894874
Precision	1.0	0.719108	1.0	0.942394	0.991130	0.985662	0.761402	0.777621	0.783109	0.784205	0.844707	0.755592	0.773292
F1	1.0	0.812059	1.0	0.920225	0.988499	0.992489	0.821051	0.824389	0.828956	0.829121	0.886217	0.829594	0.829652

Image 39

`Testing performance comparison:`



Testing performance comparison:

	Decision Tree	Decision Tree Tuned	Random Forest	Random Forest Tuned	Bagging Classifier	Bagging Estimator Tuned	Adaboost Classifier	Adabosst Classifier Tuned	Gradient Boost Classifier	Gradient Boost Classifier Tuned	XGBoost Classifier	XGBoost Classifier Tuned	Stacking Classifier
Accuracy	0.657367	0.709103	0.721022	0.724427	0.701244	0.722200	0.733857	0.743026	0.747610	0.748265	0.730845	0.739358	0.743680
Recall	0.735934	0.929034	0.832974	0.790433	0.779259	0.887865	0.877475	0.862772	0.867869	0.866497	0.850618	0.908841	0.882964
Precision	0.747362	0.718248	0.768771	0.795737	0.774854	0.745146	0.760836	0.777150	0.779401	0.780781	0.770419	0.752475	0.768076
F1	0.741604	0.810155	0.799586	0.793076	0.777050	0.810269	0.815004	0.817726	0.821260	0.821409	0.808534	0.823300	0.821523

Image 40

Above image shows Decision tree, Random Forest (default & tuned), Bagging classifier (default & tuned) & XGBoost were found to overfit the training dataset

Decision tree tuned, Adaboost (default & tuned), Gradient boost (default & tuned) and XGBoost (tuned) were found to give generalized performance on training and testing set.

XGBoost (tuned) has the highest F1 score then all other model performance after tuning although their performance is almost same

Feature importance of XGBoost Hyperparameter Tuned Model

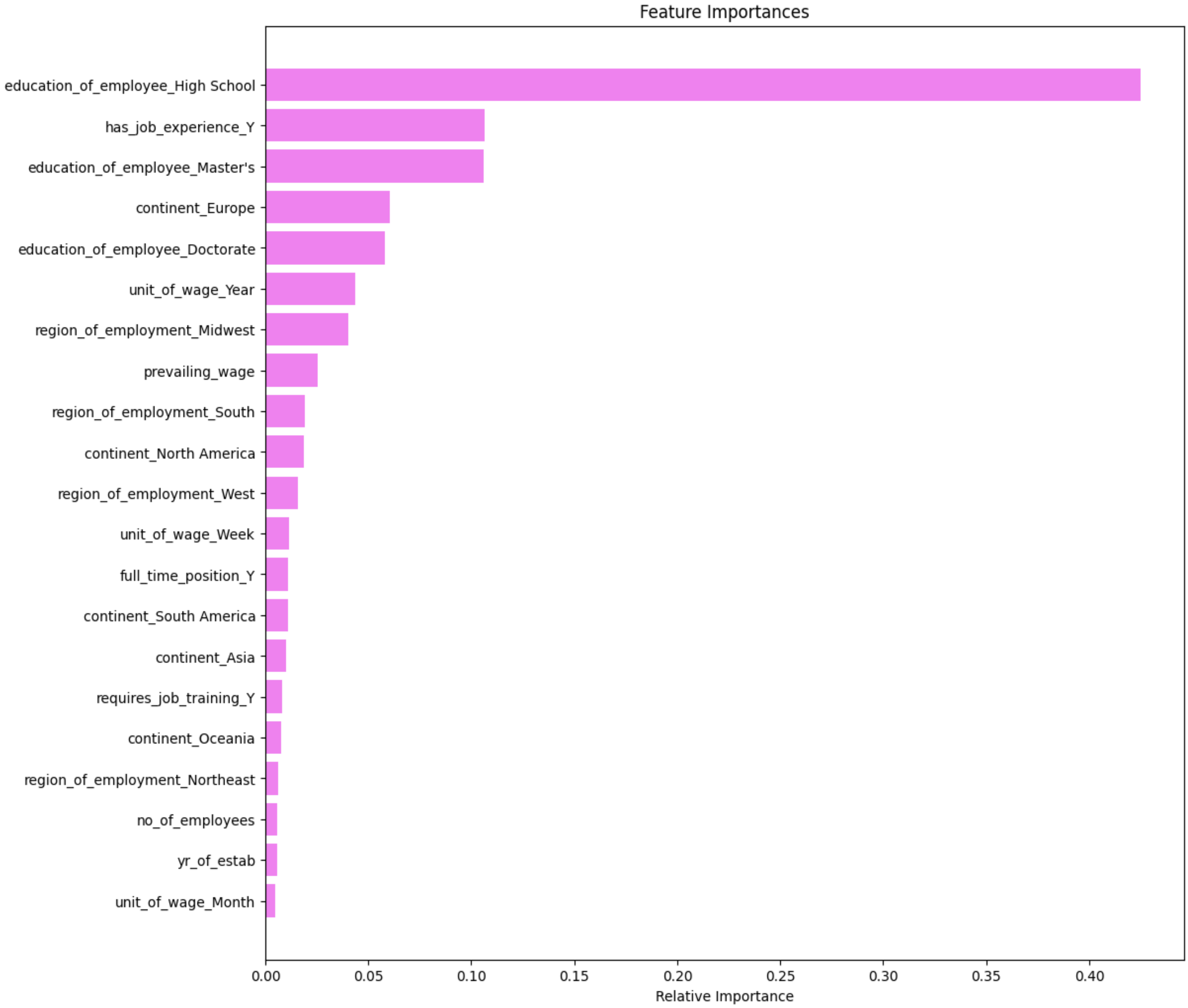


Image 41

After reviewing all the model and the results from the EDA shows the most important parameter is the Education of the employees as higher the educational degree more the approval rate for the visa

The other important attributes are if the employee have any prior working experience, then they having Master’s degree, then their continent of the employe, unit of wages and region of employment in US

Actionable Insights and Recommendations

More then 66% of the cases were certified and around 34% of the cases were not certified irrespective of employer's organization, or the year of establishment of the employer's organization. So, these have not much impact on the case status.

As found 35% of the cases were certified when the unit of wages were hourly, but 70% were certified when the unit of wages were yearly. So, this has impact on the case status.

Majority of applicants have Bachelor's degree or a master's degree. A very small number of having the High school degree, very less have very high degree Doctorate. But the approval rate for the visa is very high for the Doctorate which is around 86%, then having master's 76% then bachelor's with 62%, very less with high school has get the visa.

The trend for person having higher degree of qualification have high chances of case being certified.

Majority of the application are from the Asia 66%, followed by Europe 16%, N. America 13% & S. America 3%. However, the approval rate for the visa is for the Europe around 80%, then for Africa around 72%, then Asia around 66%. So, more cases are being certified or denied based on continent.

Being from Europe is an important attribute to get the case status certified.

Majority of the applications are to Northeast 28%, then South 27%, then West 25.8, Midwest 16% and least to Island 1.5% regions of the US. However, the cases certified follows the trend Midwest 75%, then South 70%.

Region of deployment is a important attribute for the cases being certified.

Based on the XGBoost model Tuned we have found important features for the visa get certified: -

1. Education of the Employee: - An employee with having only high school certification has over 66% chances of visa getting denied in comparison with a doctorate having more than 85% of chance of approval of visa application
2. Unit of wages: - An employee with an hourly pay likewise has over 65% of chance with the visa getting denied but with the monthly of yearly pay has more then 70% of visa approval rate
3. The continent of employee is from: - As we have found from the model and EDA that the approval rate for the person from the Europe is higher than from the other continent.

We have found from the data employers are preferring the person who has applied for the yearly pay then the monthly pay.