EV

EASY VISA

—Immigration Consultants—

MACHINE LEARNING -2

Coded Project Report

EasyVisa – Tourist Visa & Education Consultant

Submitted by – Yash Juneja

Submitted to – Great Learning

Great Learning

POWER AHEAD

Table of Contents

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

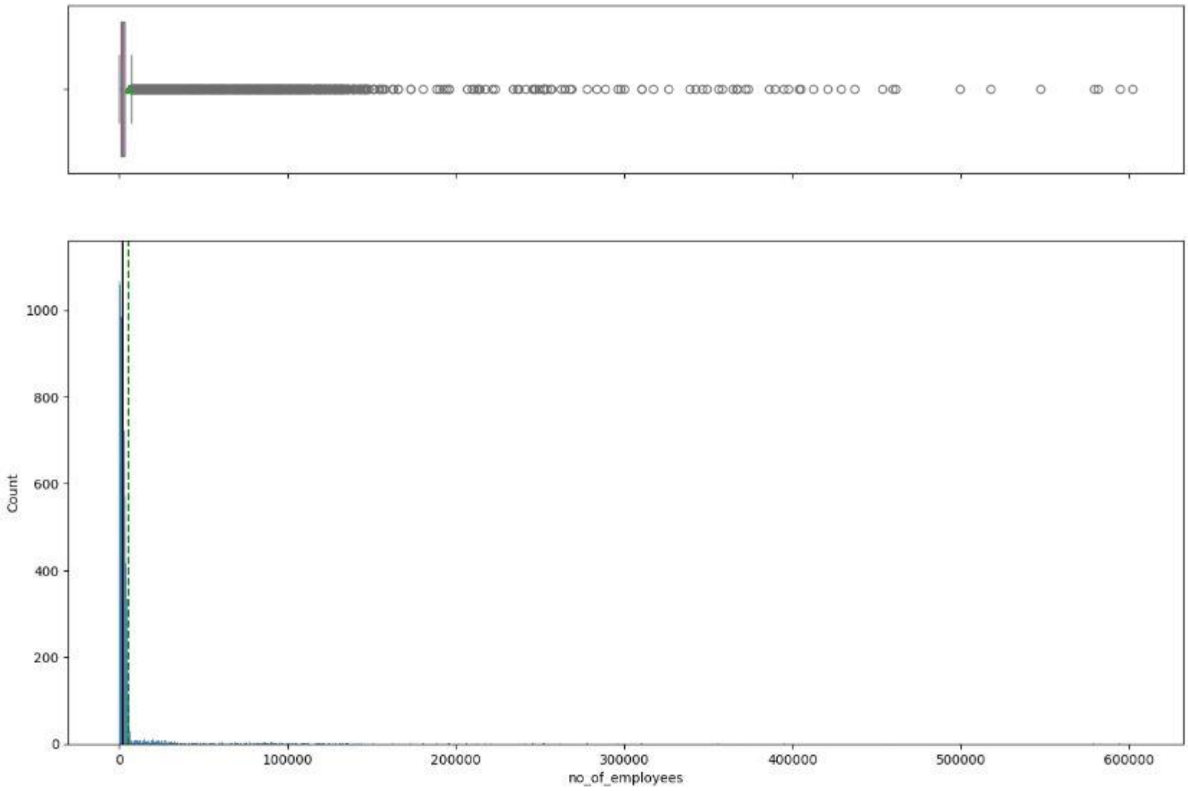
The data contains the different attributes of the employee and the employer. The detailed data dictionary is given below.

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

Exploratory Data Analysis

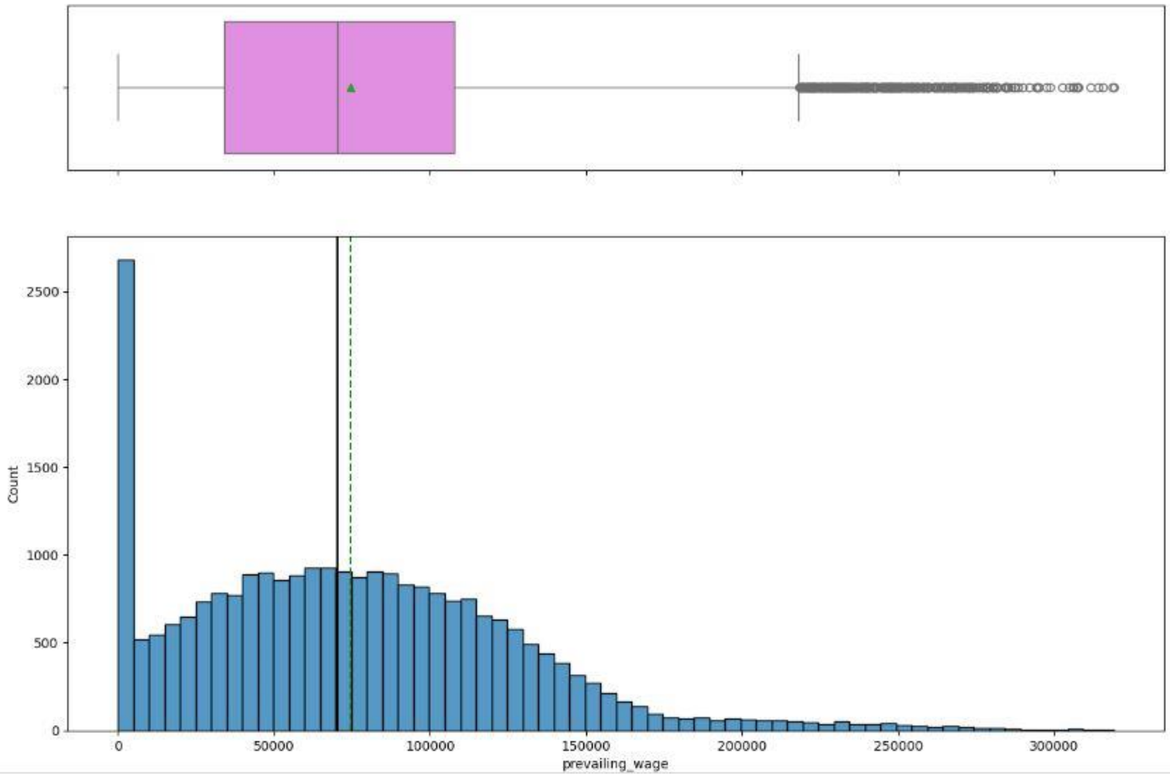
Univariate analysis

Observations on number of employees



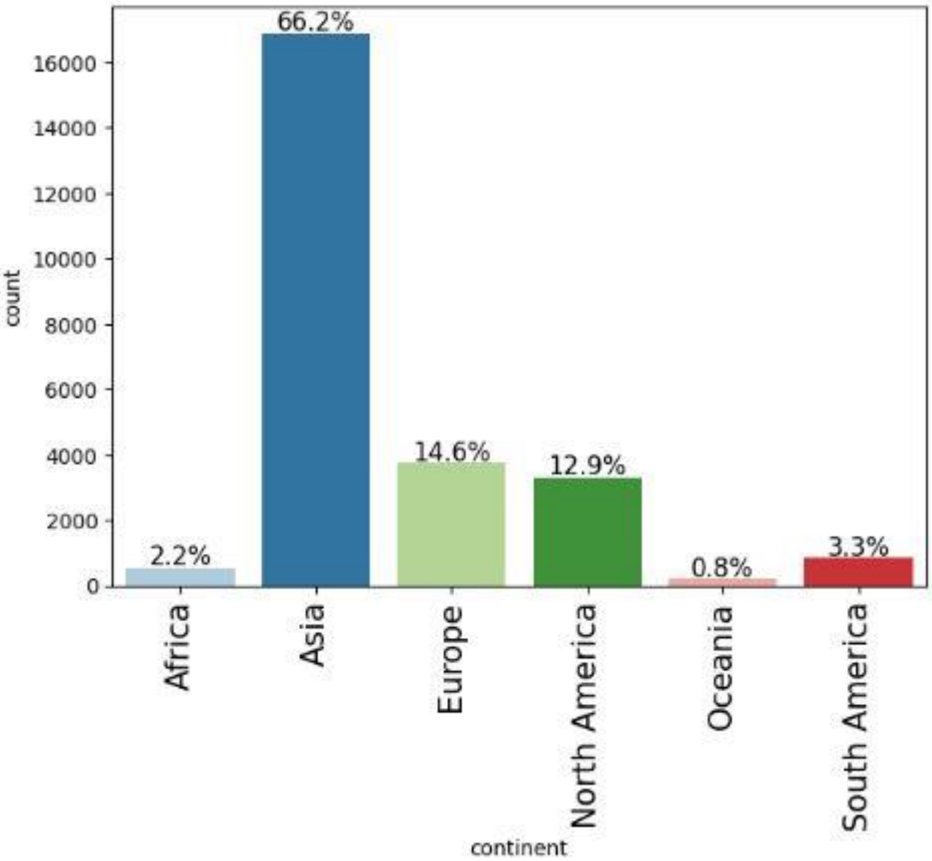
- The distribution of age is right-skewed
- The boxplot shows that there are outliers at the right end
- We will not treat these outliers as they represent the real market trend

Observations on prevailing wage



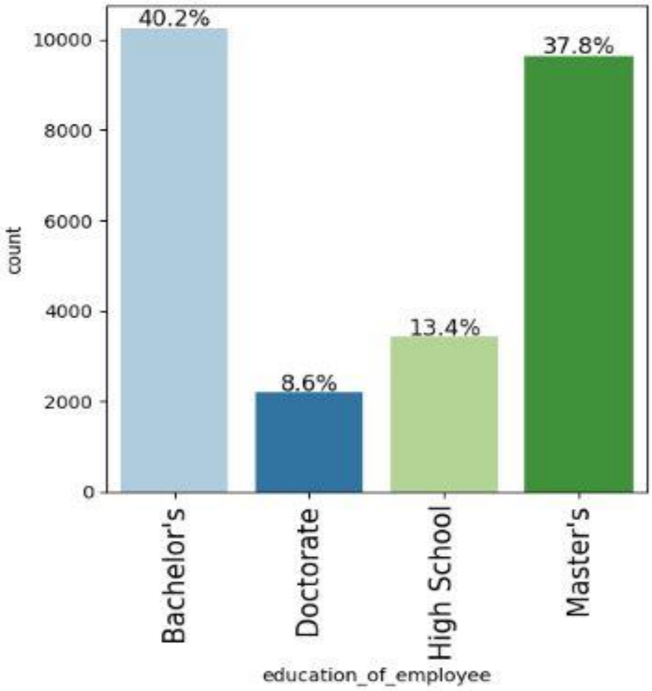
- The distribution of the prevailing wage is skewed to the right
- There is a huge difference between wages among applicants
- There are applicants whose wage is more than 150,000
- There are applicants whose wage is around 0.
- It could be that some wages are entered as hourly base while others as yearly base

Observations on continent



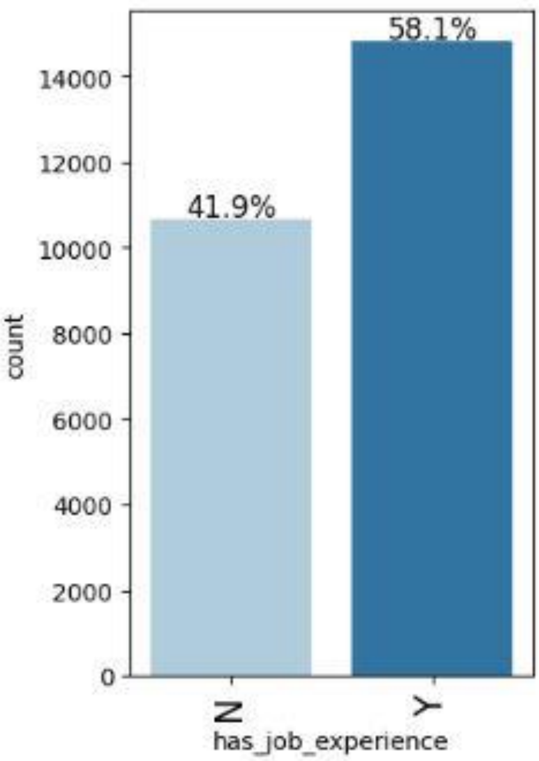
66.2% are coming from Asia, 15% are coming from Europe, and 13% are coming from North America

Observations on education of employee



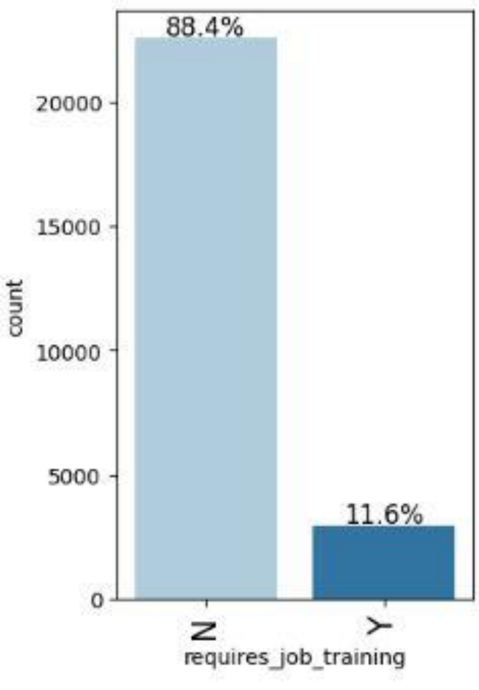
- 40.2% of the applicants have a bachelor's degree, followed by 37.8% having a master's degree.
- 8.6% of the applicants have a doctorate degree.

Observations on job experience



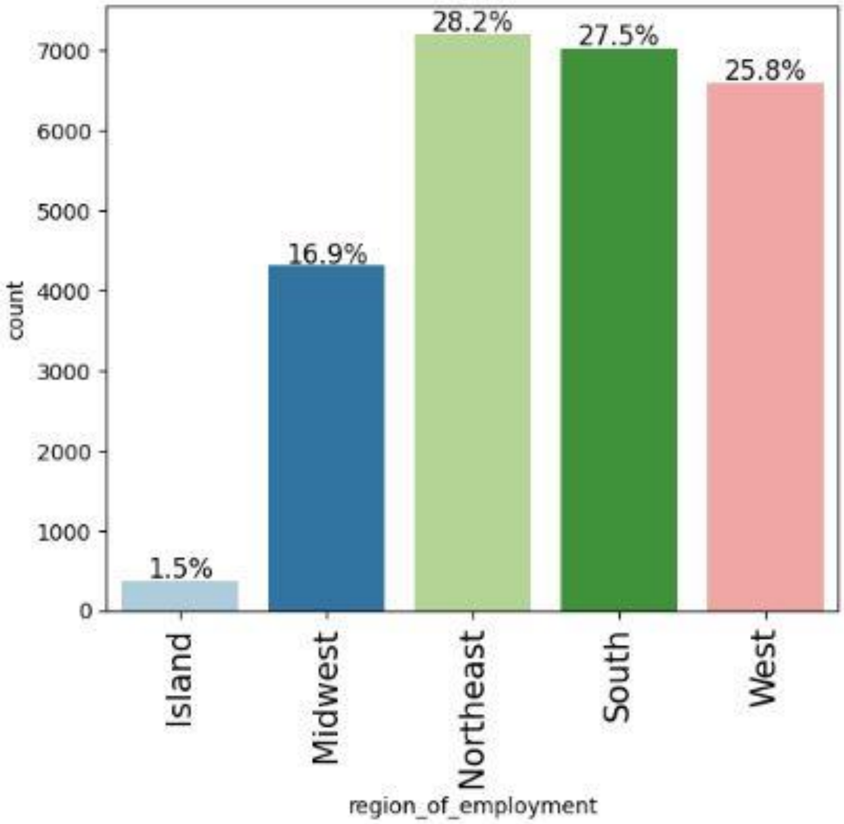
58.1% of the applicants have job experience

Observations on job training



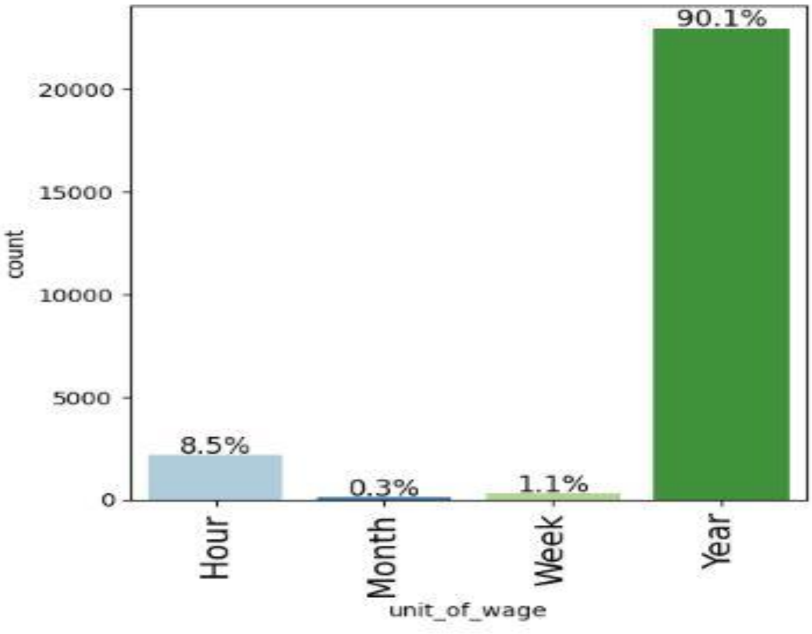
88.4% of the applicants do not require any job training

Observations on region of employment



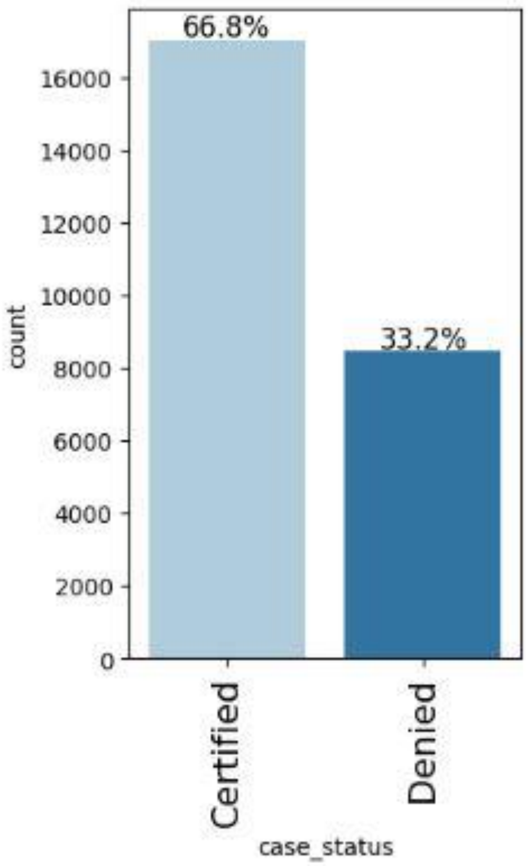
- Northeast, South, and West are approximate equally distributed
- The Island regions have only 1.5% of the applicants

Observations on unit of wage



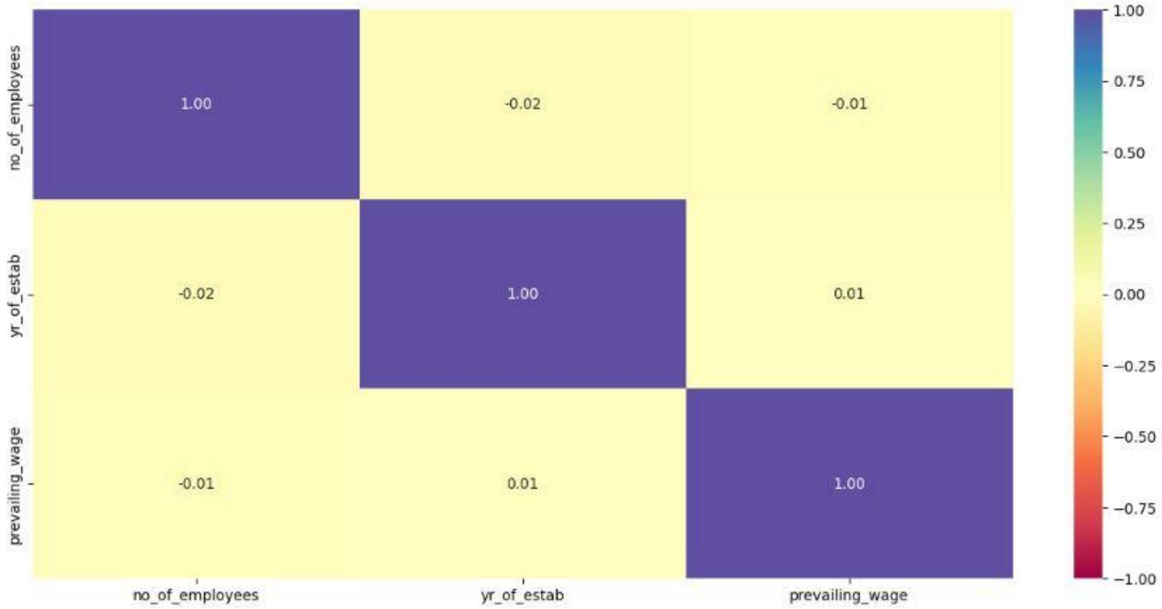
90.1% of the applicants have a yearly unit of the wage

Observations on case status



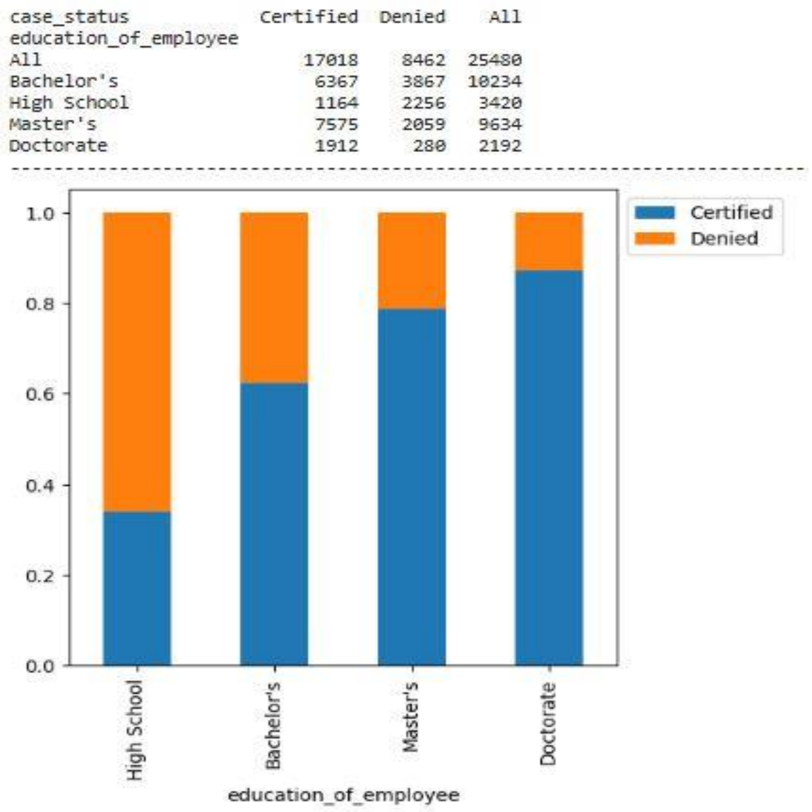
66.8% of the visas were certified

Bivariate Analysis



we cannot see any correlation between the numerical variables

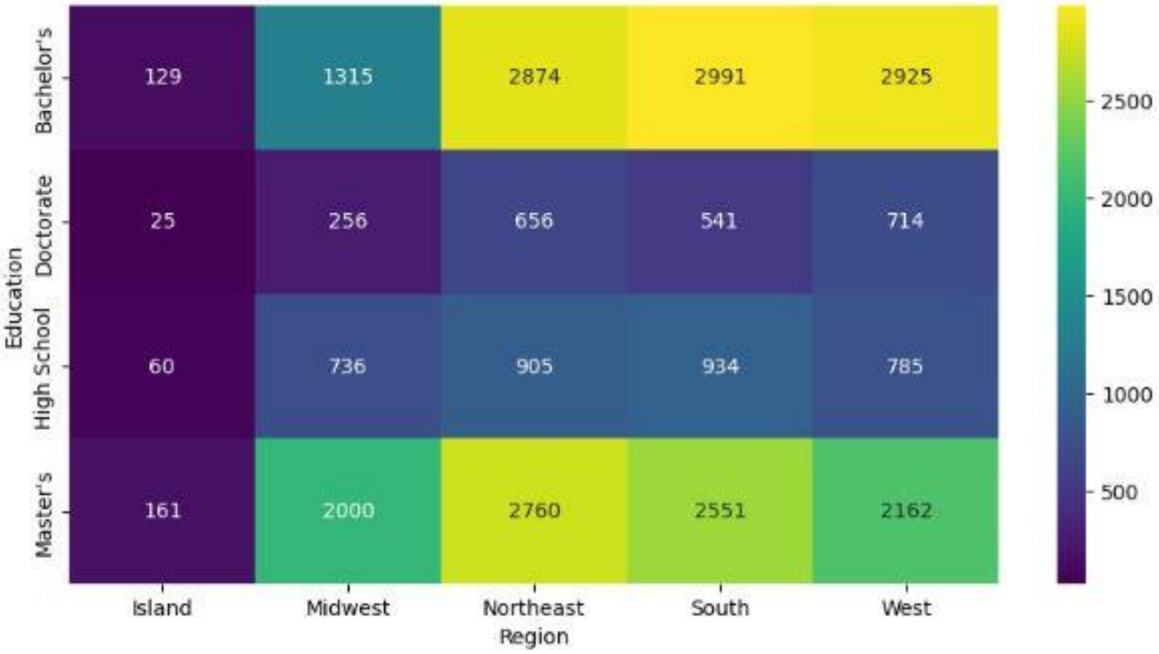
Education vs case status

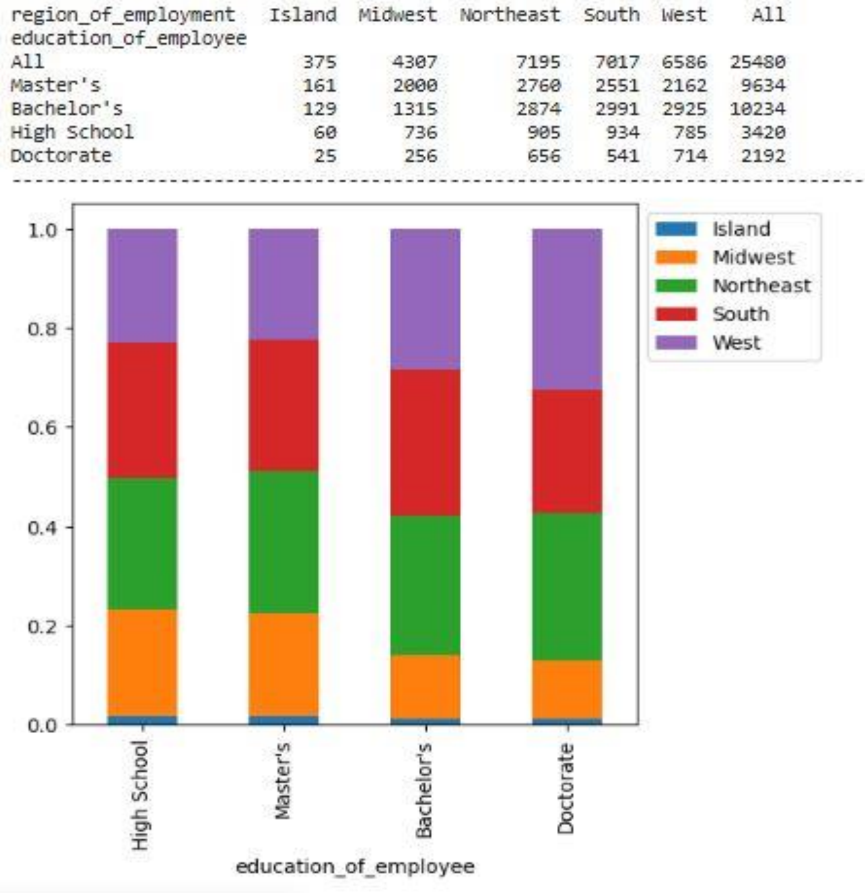


```
case_status            Certified  Denied    All
education_of_employee
All                        17018    8462  25480
Bachelor's                  6367    3867  10234
High School                 1164    2256   3420
Master's                    7575    2059   9634
Doctorate                   1912  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

The higher the education, the more chance to get certified

Education vs Region

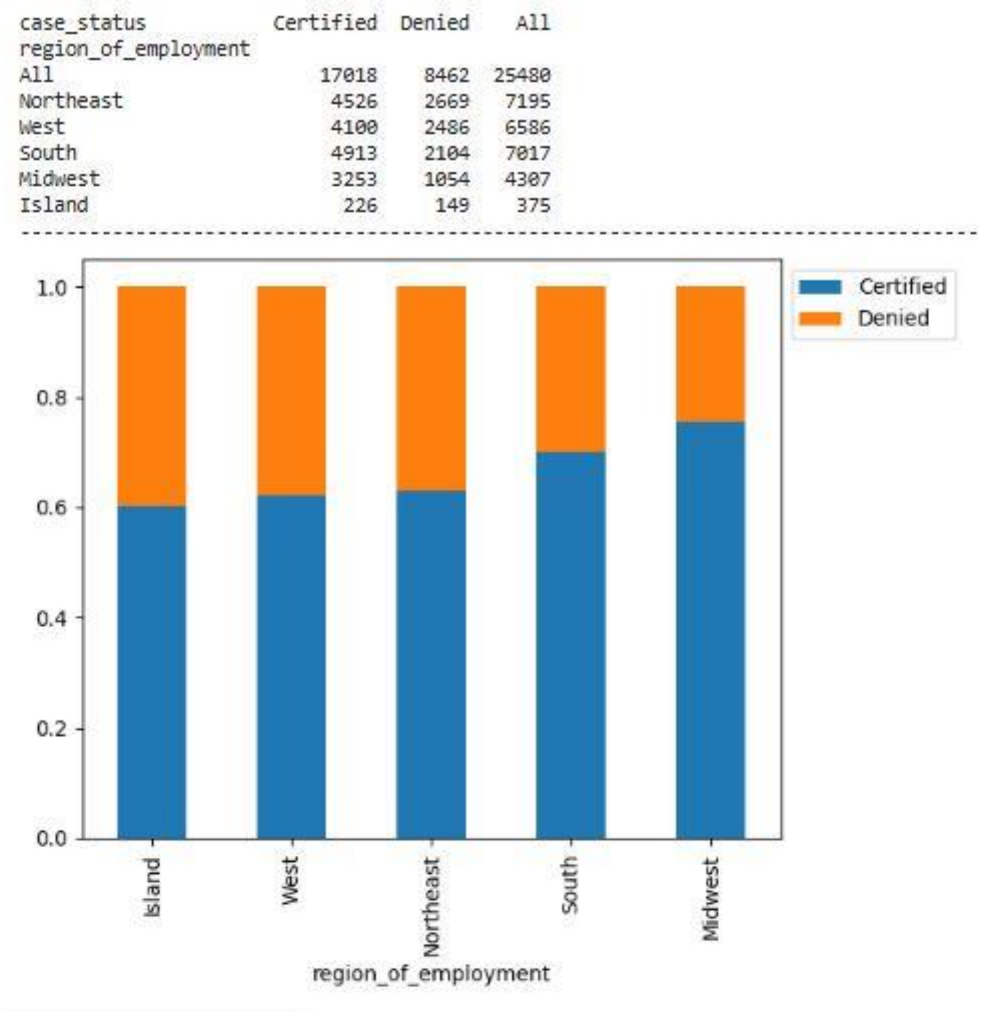




```
region_of_employment  Island  Midwest  Northeast  South  West    All
education_of_employee
All                      375     4307       7195   7017  6586  25480
Master's                 161     2000       2760   2551  2162   9634
Bachelor's               129     1315       2874   2991  2925  10234
High School               60      736        905    934   785   3420
Doctorate                 25      256        656    541   714   2192
---------------------------------------------------------------------------
```

- The requirement for the applicants who have passed high school is most in the South region, followed by Northeast region.
- The requirement for Bachelor's is mostly in South region, followed by West region.
- The requirement for Master's is most in Northeast region, followed by South region.
- The requirement for Doctorate's is mostly in West region, followed by Northeast region

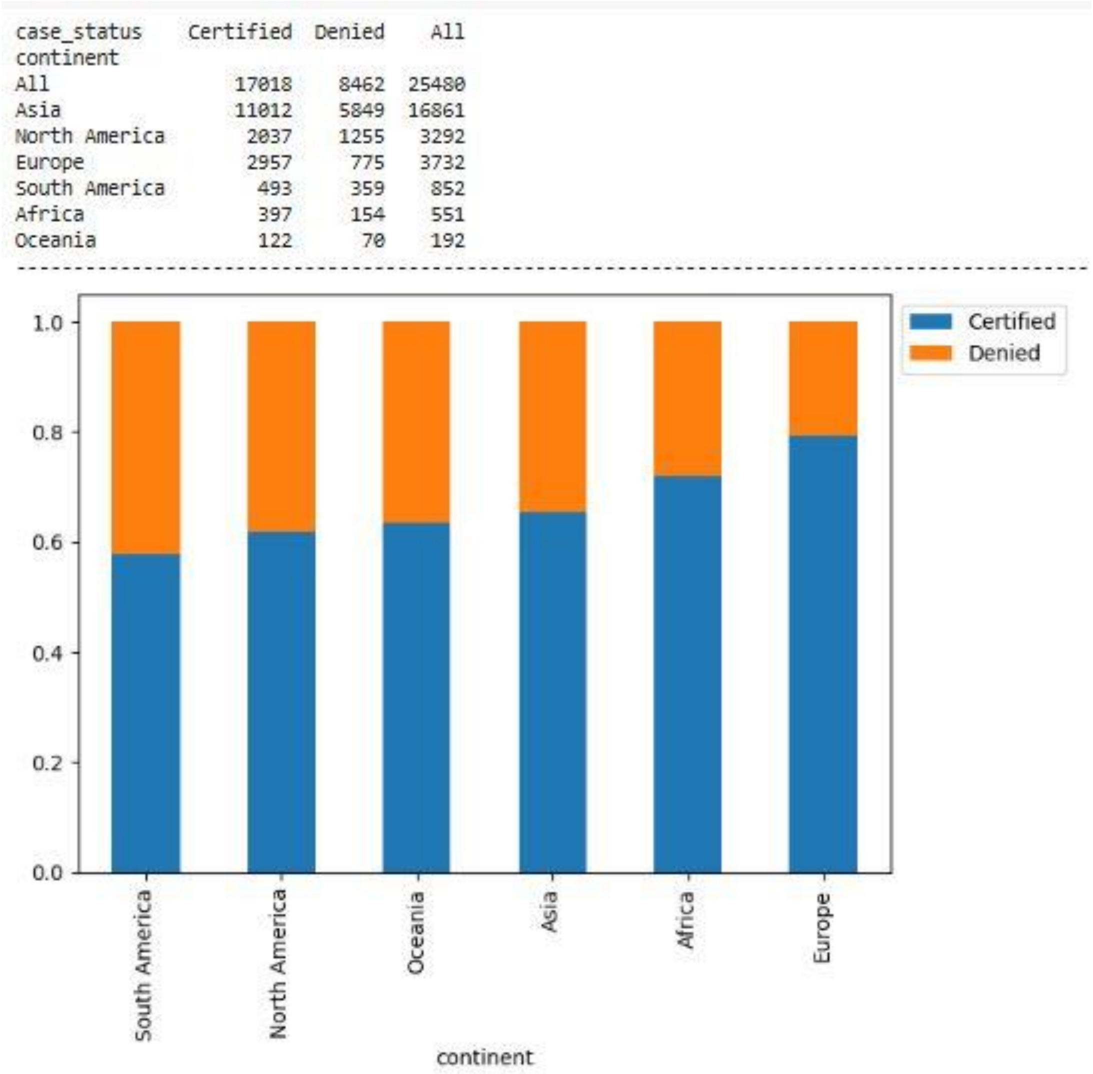
<u>Region vs Case status</u>



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

- Midwest has the highest positive case chance
- Island, in the opposite side, has the lowest positive case chance

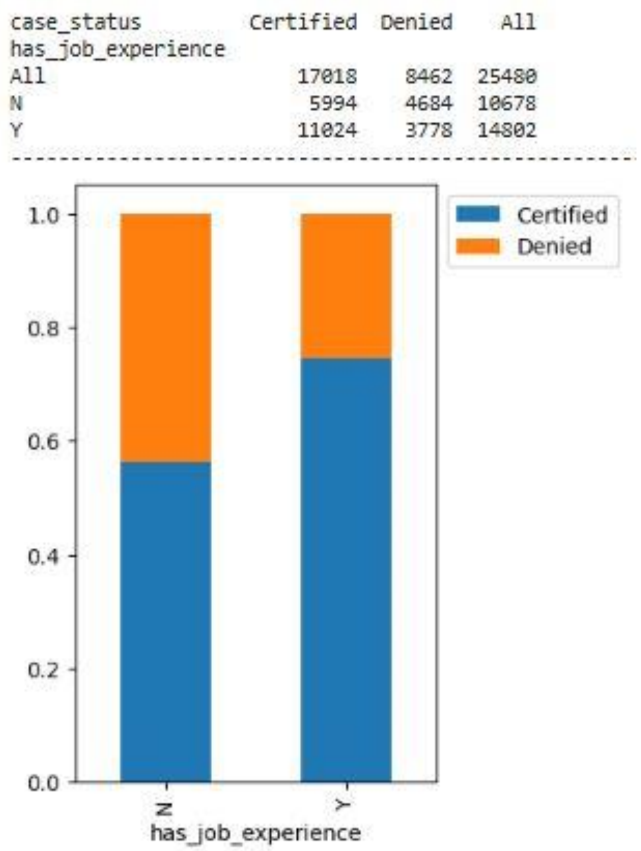
Continent vs Case status



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

Europe has the highest chance of getting certified, while South America has the lowest chance

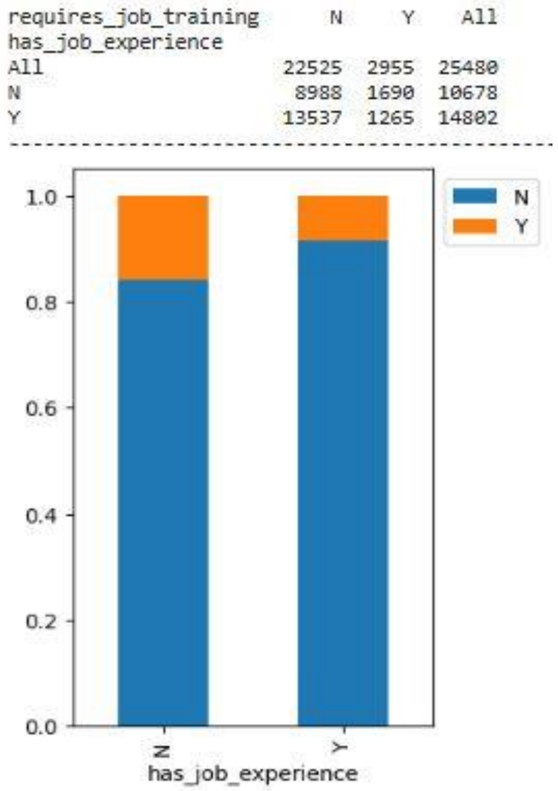
Job experience vs Case status



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

Applicants with job experience have more chances of getting certified

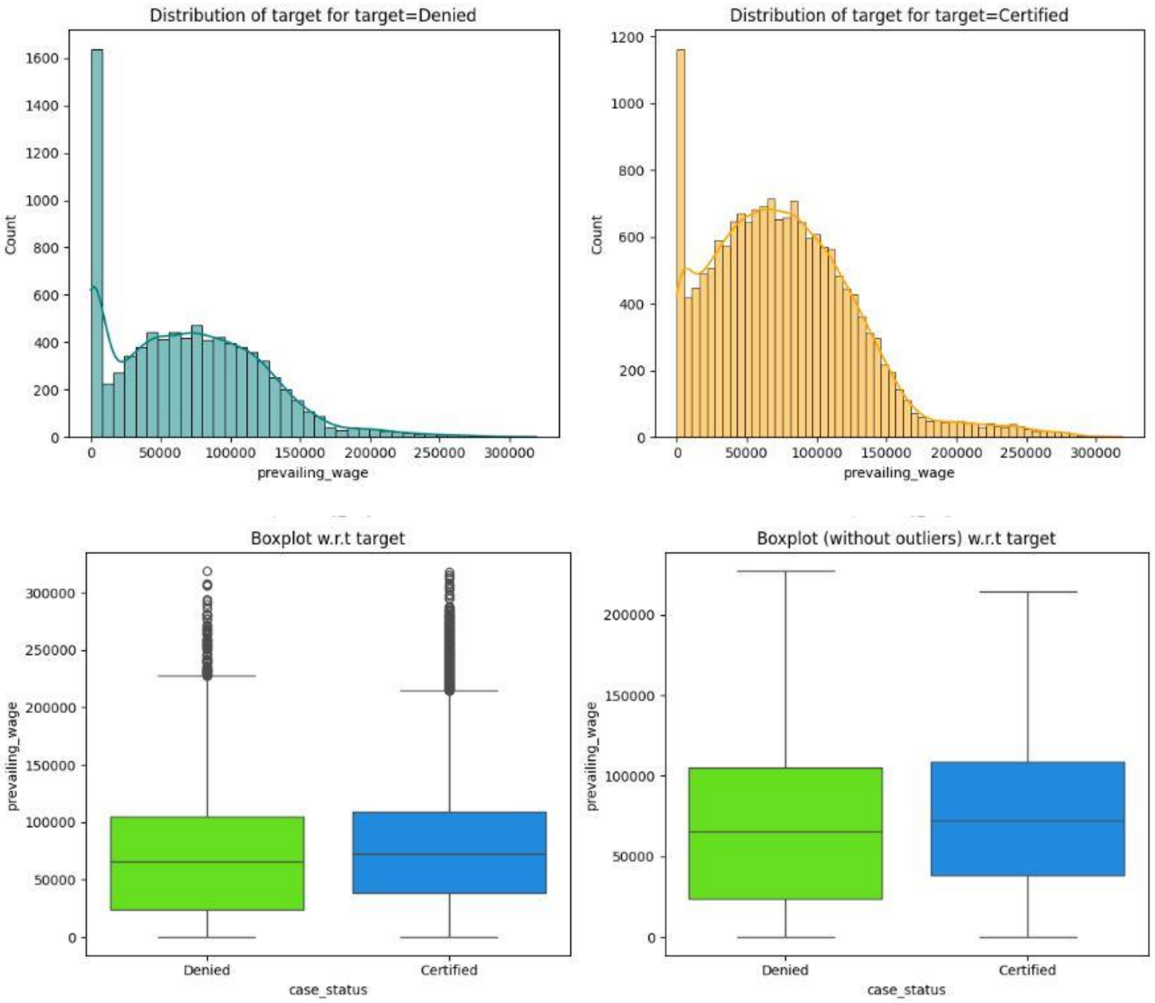
Job experience vs Training required



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

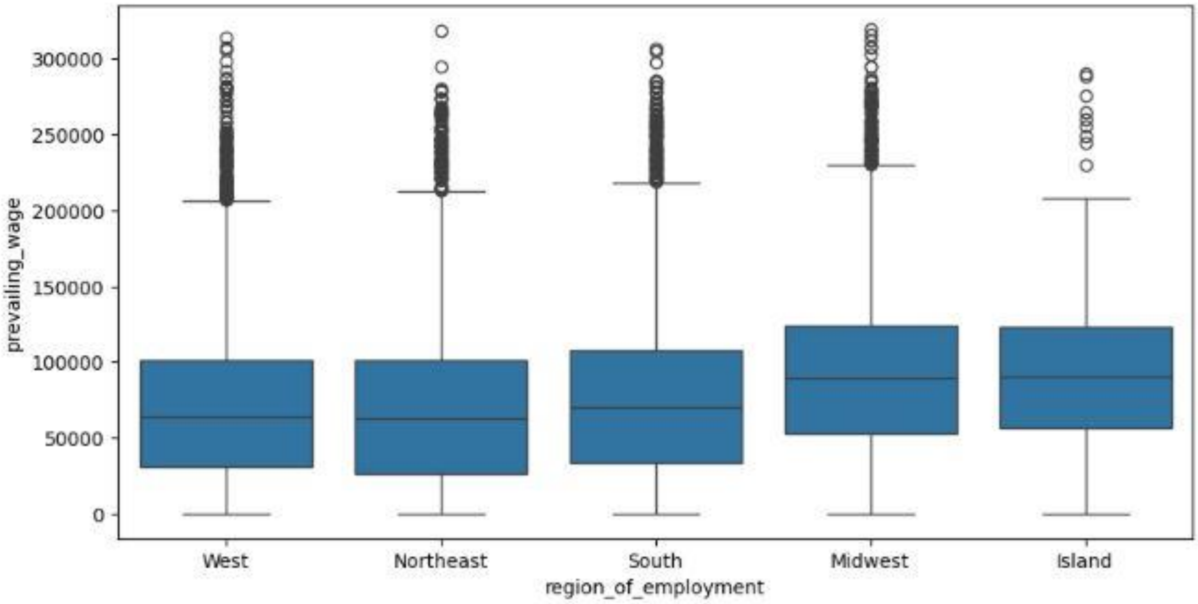
If the applicant has a job experience, they are less likely to require training

Wage vs Case status



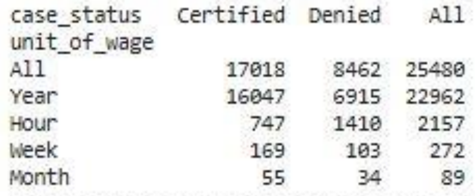
The median prevailing wage for the certified applications is slightly higher as compared to denied applications.

Region vs Wage

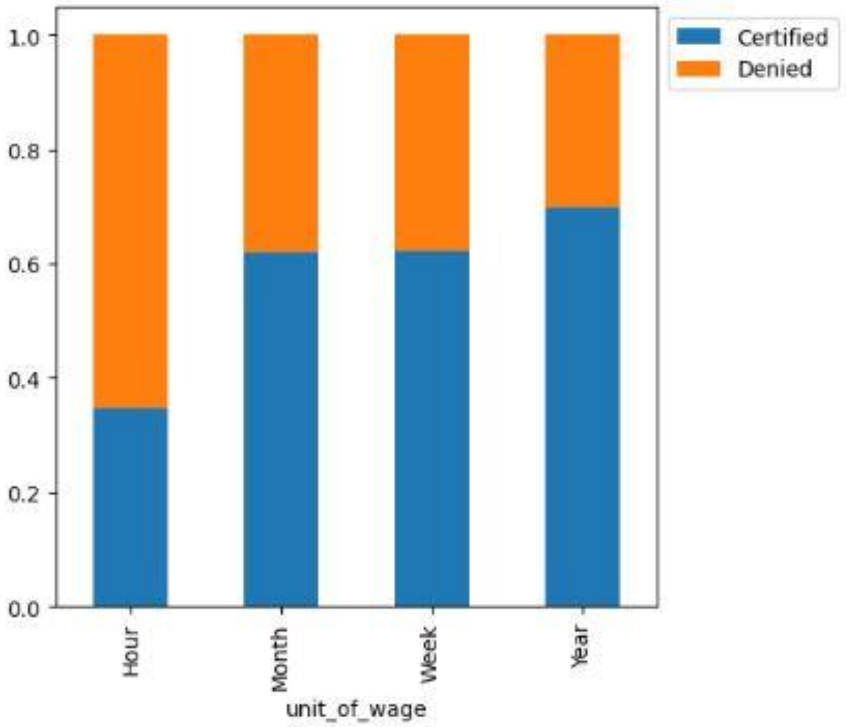


Prevailing wages is higher in Midwest and Island

Unit of wage vs case status



```
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
```



Hourly waged applicants are more likely to get denied while yearly are more likley to get certified

Observations from Exploratory Data Analysis

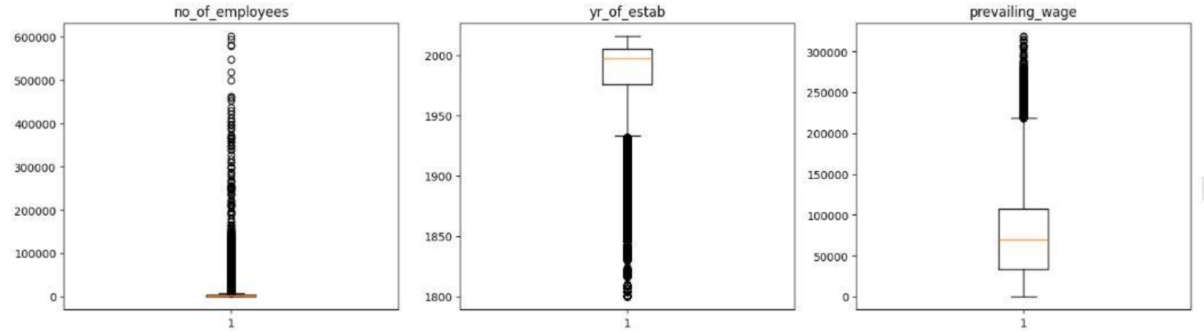
- More than twice the number of cases were certified than denied irrespective of the number of employees in the employer's organization & the year of establishment of the employer's organization. These attributes are hence, not thought to have an impact on case statuses
- Both these attributes are heavily 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 were for employer's established after 1990 • Only 35% of the cases were certified when the unit_of_wage is Hour–but 70% were certified when the unit_of_wage is not Hour– (i.e., Weekly, Monthly or Yearly). 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 Hourly and the remaining 91.5% of all cases were for 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, cases getting 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 a doctorate 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 high for applicants with prior job experience (75% of such cases) and low for applicants without prior job experience (~56% of such cases). 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 on the 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 status

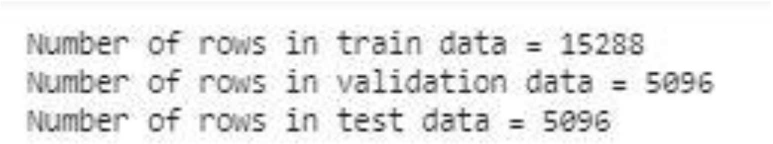
Data Preprocessing

Outlier Check

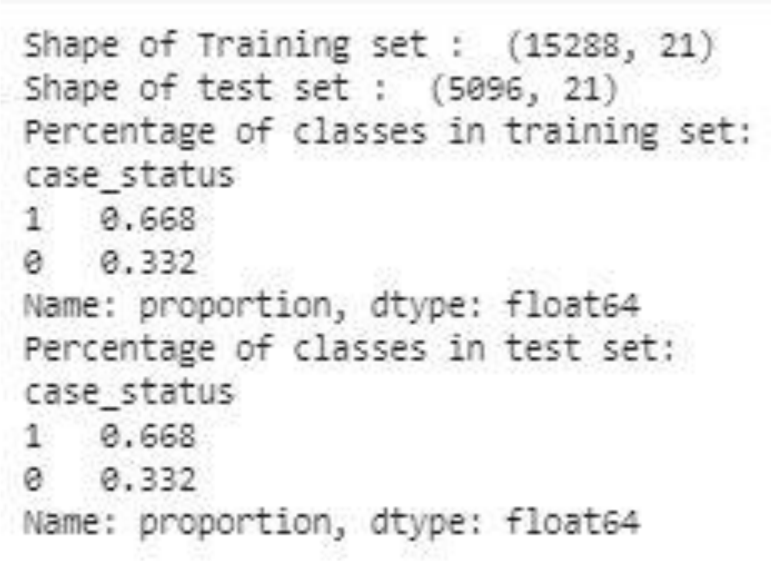


Although there are outliers, we will keep them as they have a valuable input

Data Preparation for modelling

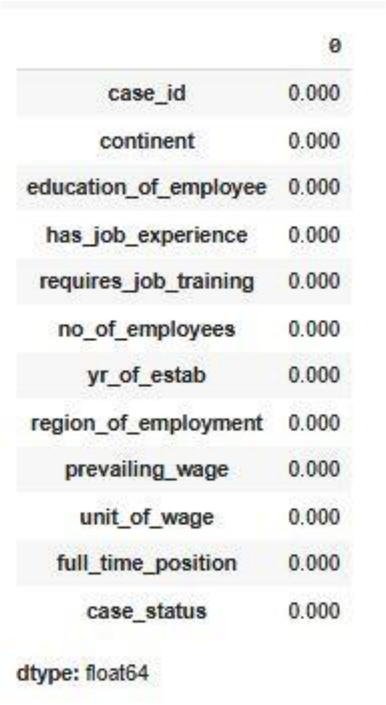


```
Number of rows in train data = 15288
Number of rows in validation data = 5096
Number of rows in test data = 5096
```



```
Shape of Training set :  (15288, 21)
Shape of test set :  (5096, 21)
Percentage of classes in training set:
case_status
1   0.668
0   0.332
Name: proportion, dtype: float64
Percentage of classes in test set:
case_status
1   0.668
0   0.332
Name: proportion, dtype: float64
```

Checking for Missing Values

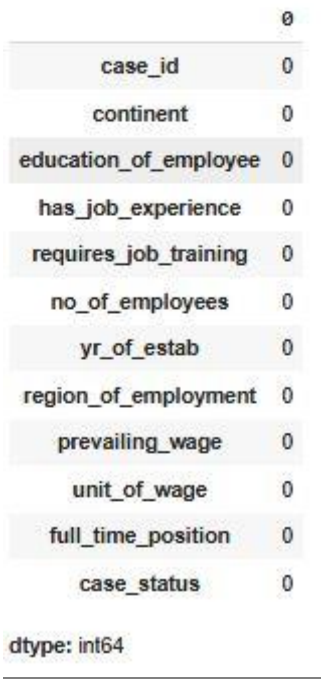


	0
case_id	0.000
continent	0.000
education_of_employee	0.000
has_job_experience	0.000
requires_job_training	0.000
no_of_employees	0.000
yr_of_estab	0.000
region_of_employment	0.000
prevailing_wage	0.000
unit_of_wage	0.000
full_time_position	0.000
case_status	0.000

dtype: float64

No Missing Values

Checking for Duplicate Values



	0
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

No Duplicate Values

Model Building

Model can make wrong predictions as:

- Model predicts that the visa application will get certified but in reality, the visa application should get denied.
- 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 on that 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.

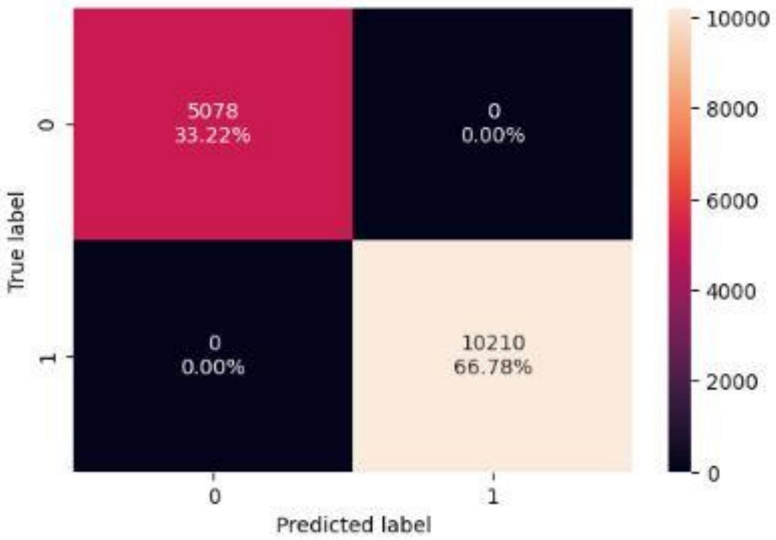
How to reduce the losses?

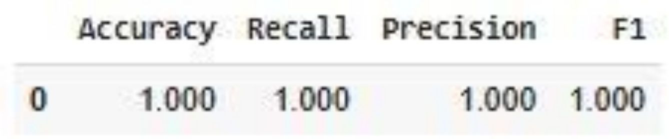
- F1 Score can be used as a the metric for evaluation of the model, greater the F1 score higher are the chances of minimizing False Negatives and False Positives.
- We will use balanced class weights so that model focuses equally on both classes

1.) Decision Tree – Model Building & Hyperparameter Tuning

- **Model Building**

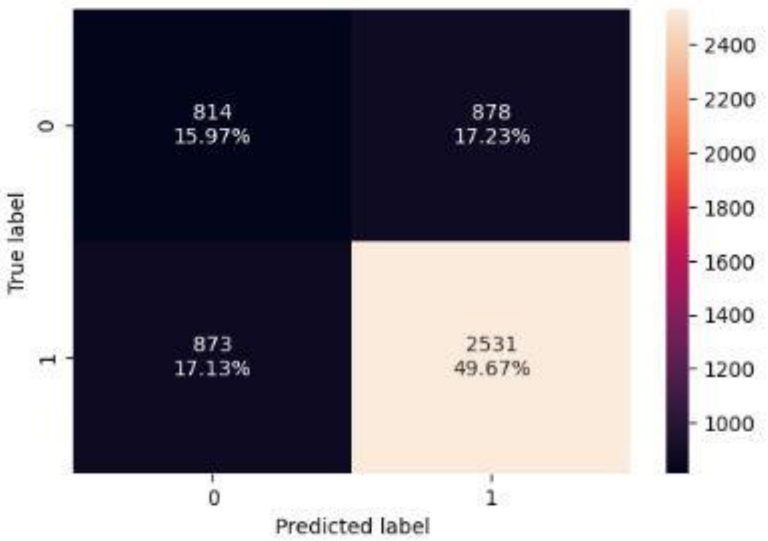
Checking model performance on test set

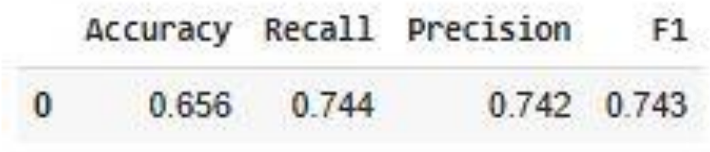




	Accuracy	Recall	Precision	F1
0	1.000	1.000	1.000	1.000

Checking model performance on training set



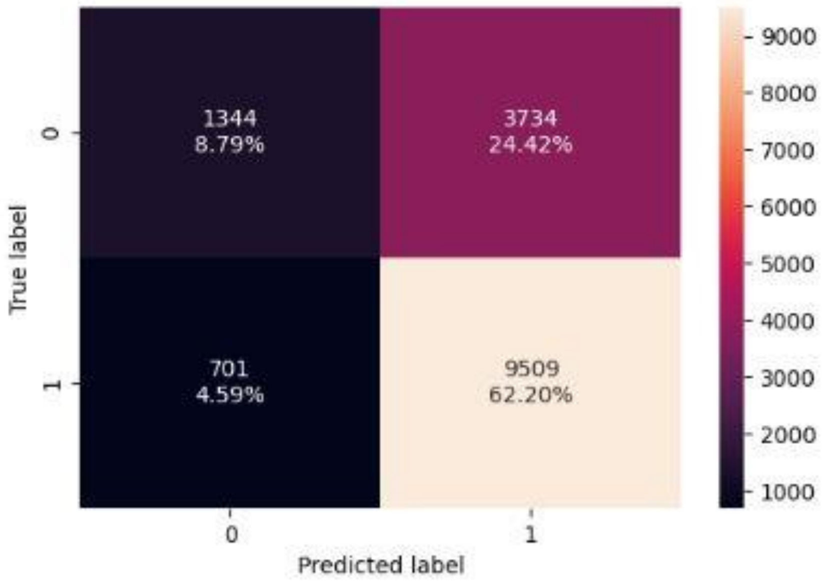


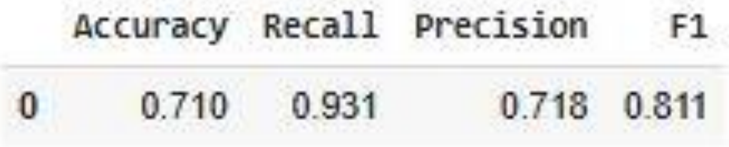
	Accuracy	Recall	Precision	F1
0	0.656	0.744	0.742	0.743

The decision tree is overfitting

- **Hyperparameter Tuning – Decision Tree**

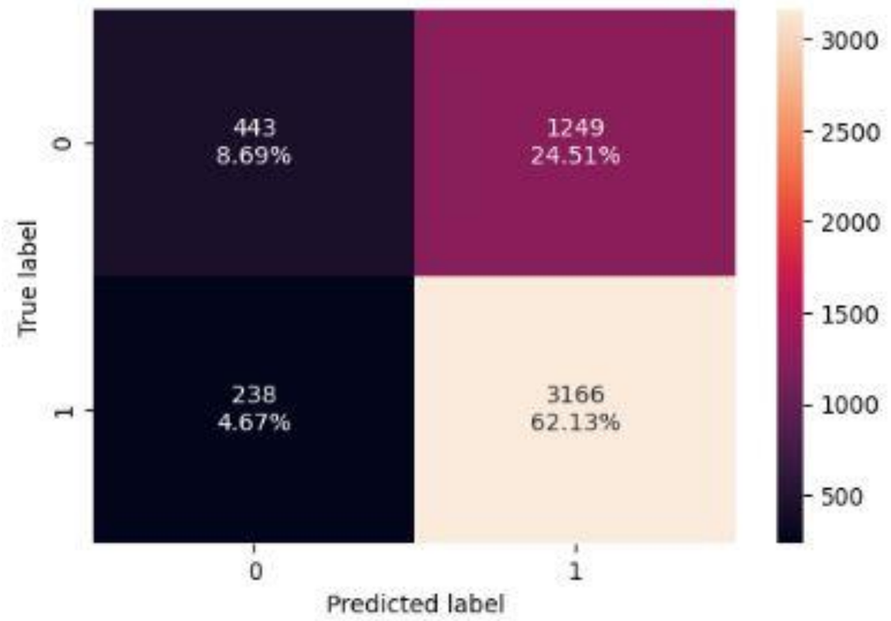
Checking model performance on training set

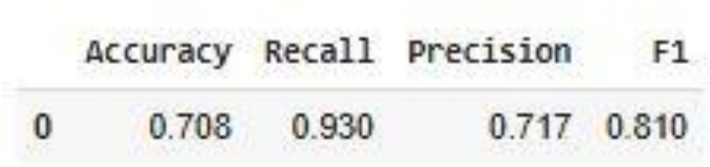




	Accuracy	Recall	Precision	F1
0	0.710	0.931	0.718	0.811

Checking model performance on test set





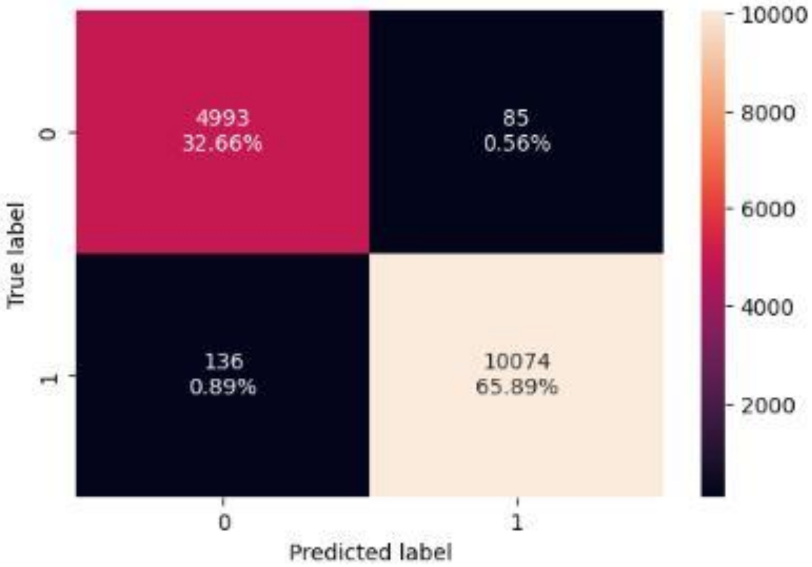
	Accuracy	Recall	Precision	F1
0	0.708	0.930	0.717	0.810

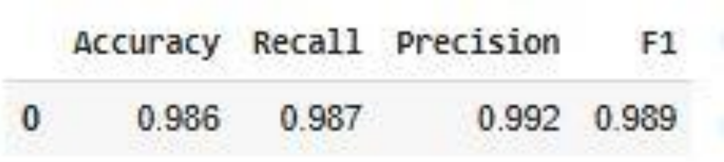
This model is a good fit and not suffering from overfitting and it can be generalized

2.) <u>Bagging - Model Building and Hyperparameter Tuning</u>

- **Model Building**

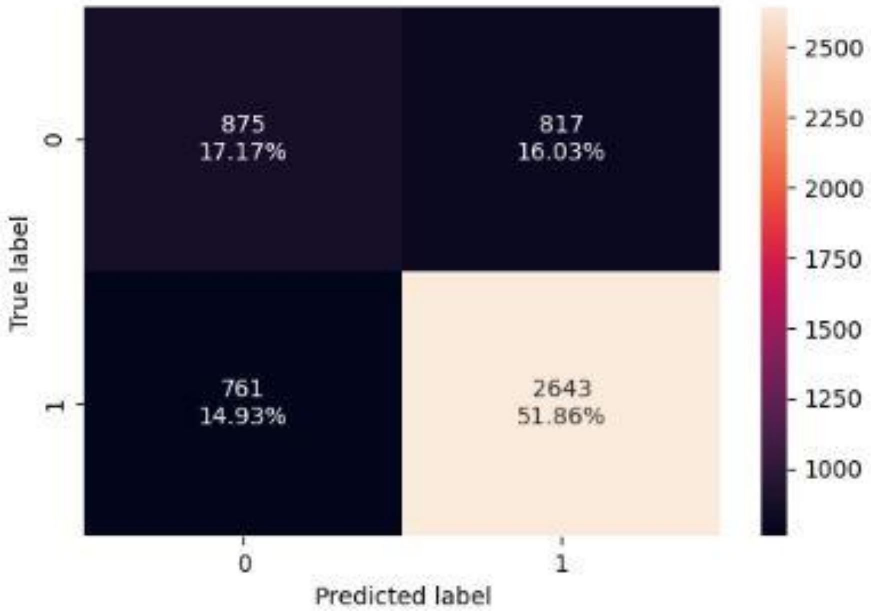
Checking model performance on training set

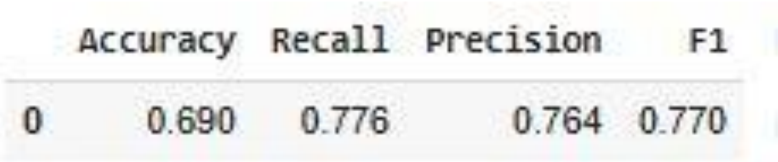




	Accuracy	Recall	Precision	F1
0	0.986	0.987	0.992	0.989

Checking model performance on test set



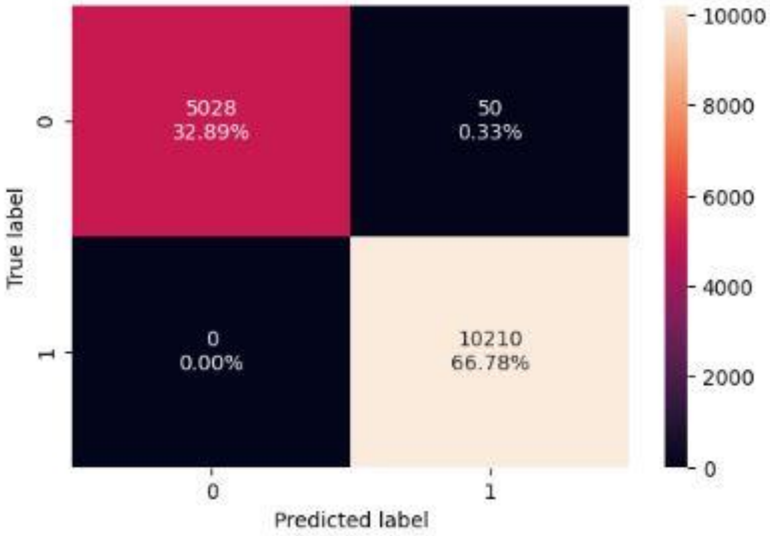


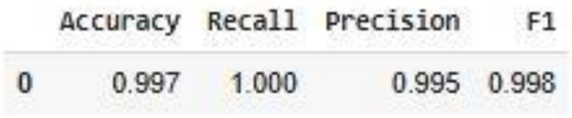
	Accuracy	Recall	Precision	F1
0	0.690	0.776	0.764	0.770

The bagging classifier is also overfitting

- **Hyperparameter Tuning - Bagging Classifier**

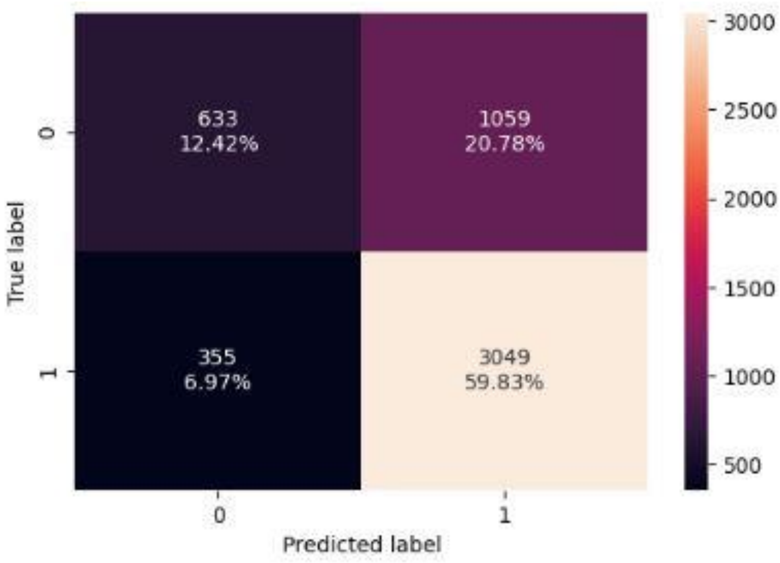
Checking model performance on training set

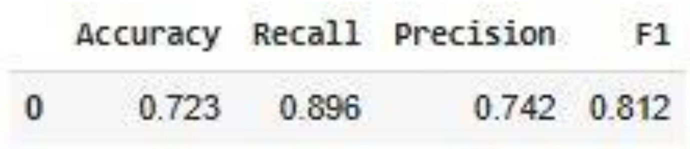




	Accuracy	Recall	Precision	F1
0	0.997	1.000	0.995	0.998

Checking model performance on test set





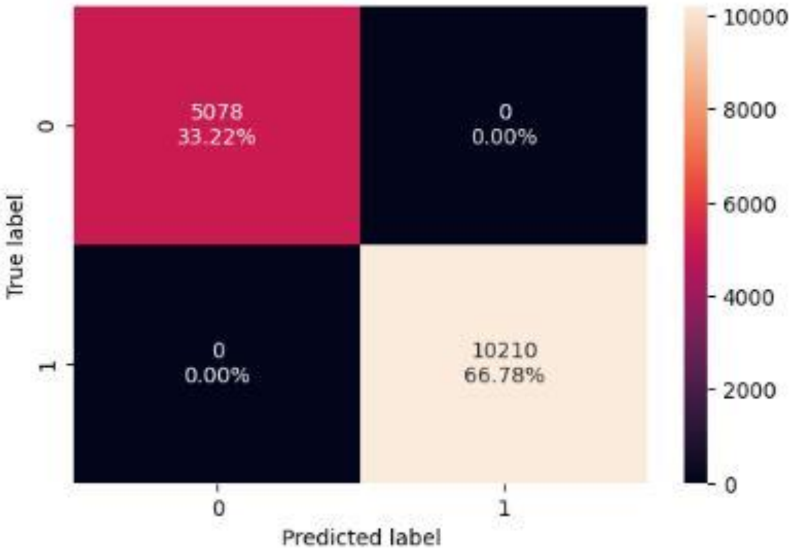
	Accuracy	Recall	Precision	F1
0	0.723	0.896	0.742	0.812

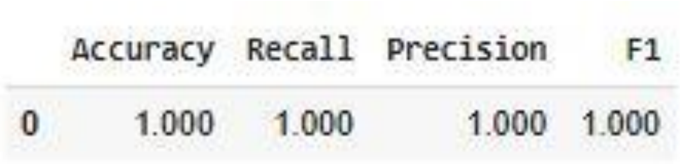
The bagging classifier is still overfitting even after tuning

3.) <u>Random Forest - Model Building and Hyperparameter Tuning</u>

- **Model Building**

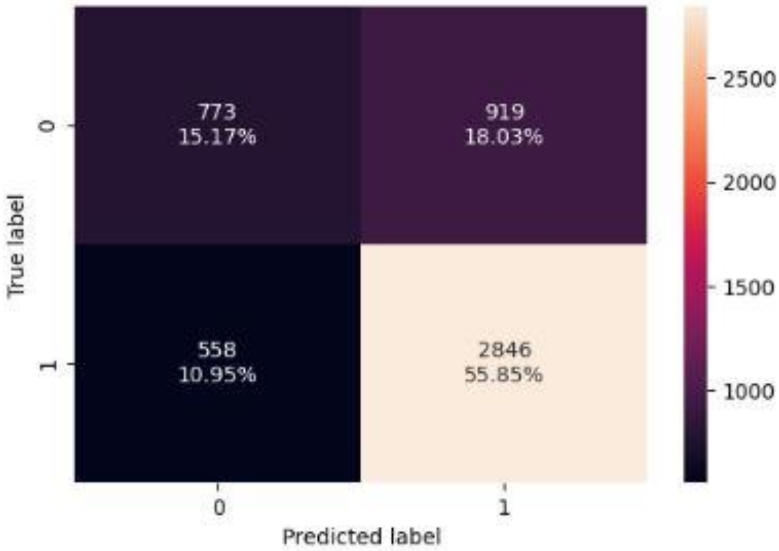
Checking model performance on training set

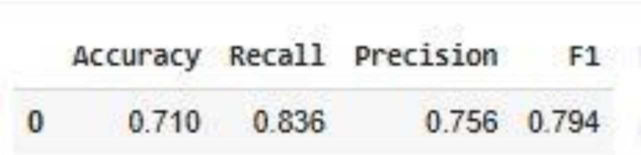




	Accuracy	Recall	Precision	F1
0	1.000	1.000	1.000	1.000

Checking model performance on test set



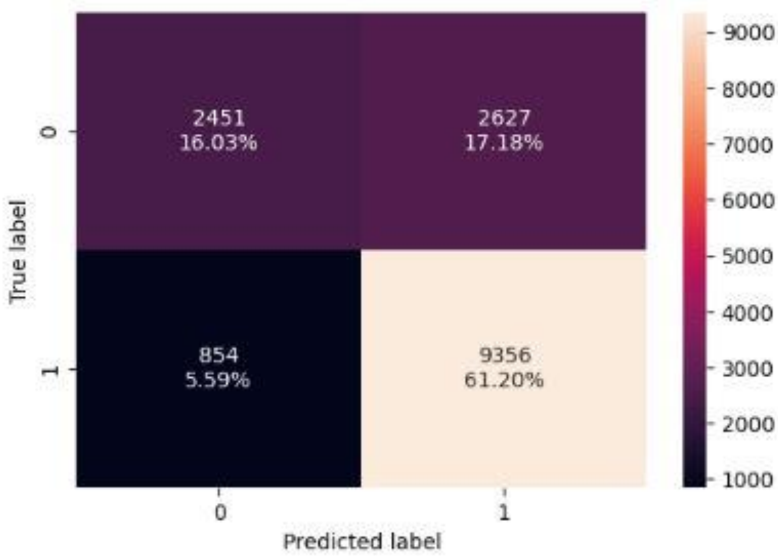


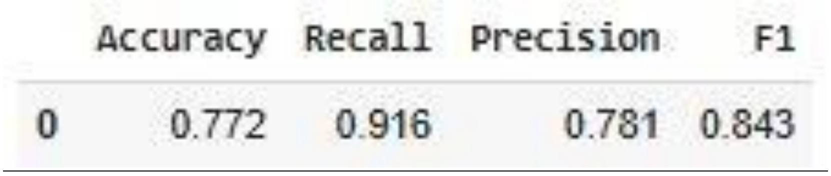
	Accuracy	Recall	Precision	F1
0	0.710	0.836	0.756	0.794

Random forest is overfitting

Hyperparameter Tuning - Random Forest

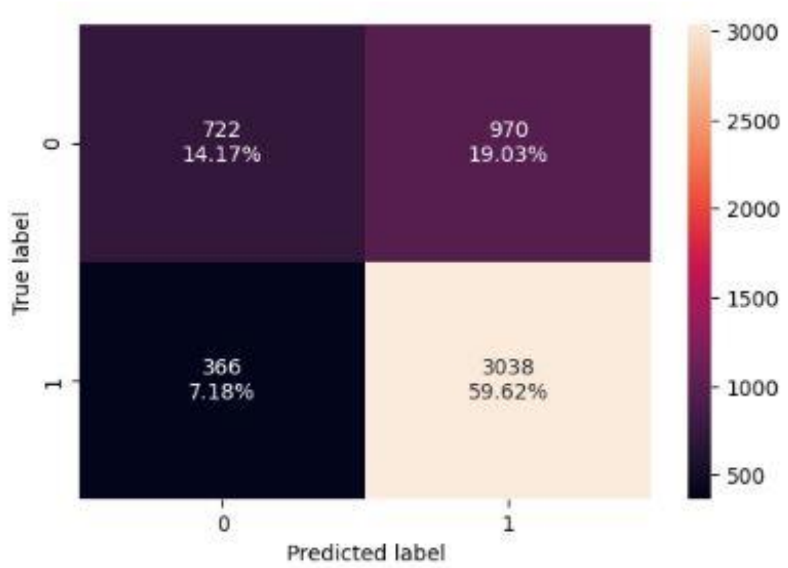
Checking model performance on training set

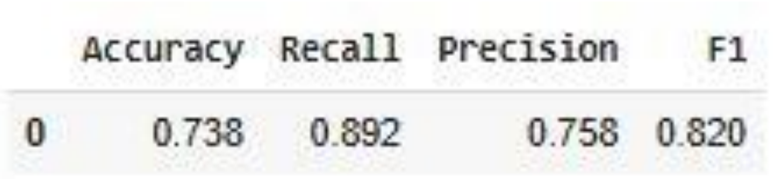




	Accuracy	Recall	Precision	F1
0	0.772	0.916	0.781	0.843

Checking model performance on test set





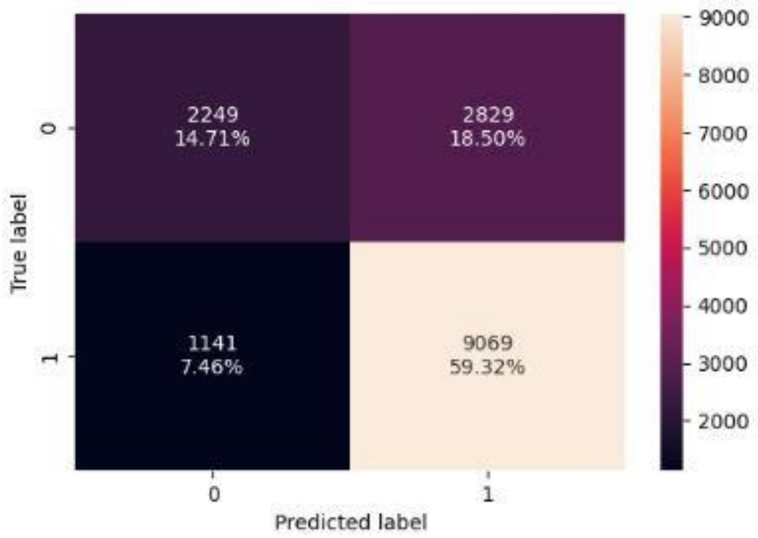
	Accuracy	Recall	Precision	F1
0	0.738	0.892	0.758	0.820

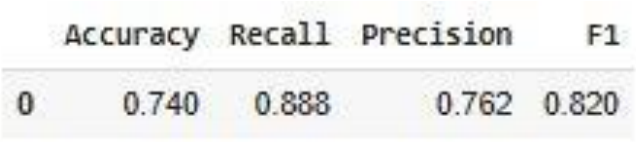
After hyperparameter tuning the model performance has generalized

4.) **Boosting - Model Building and Hyperparameter Tuning**

- **AdaBoost Classifier-Model Building**

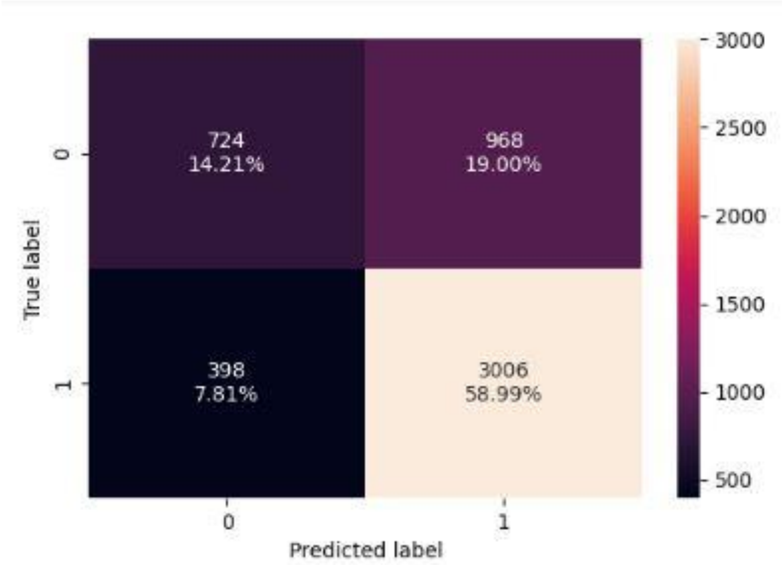
Checking model performance on training set

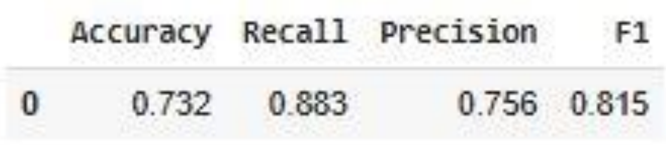




	Accuracy	Recall	Precision	F1
0	0.740	0.888	0.762	0.820

Checking model performance on test set



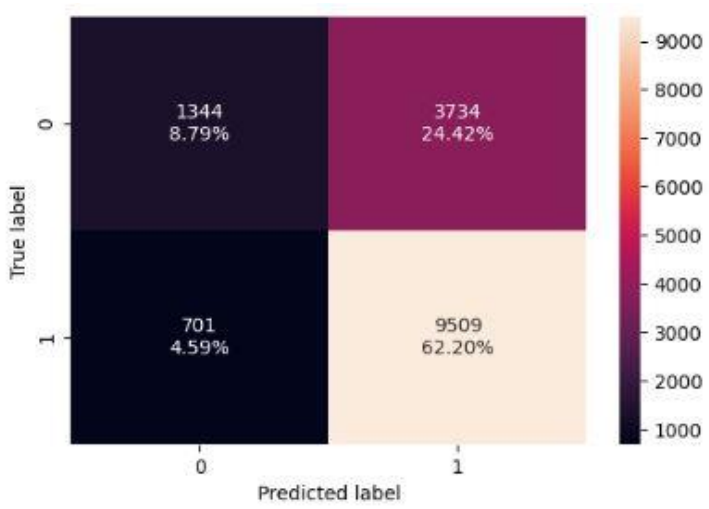


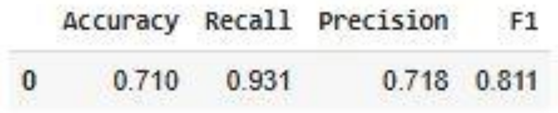
	Accuracy	Recall	Precision	F1
0	0.732	0.883	0.756	0.815

The model is giving a generalized performance

- **Hyperparameter Tuning - AdaBoost Classifier**

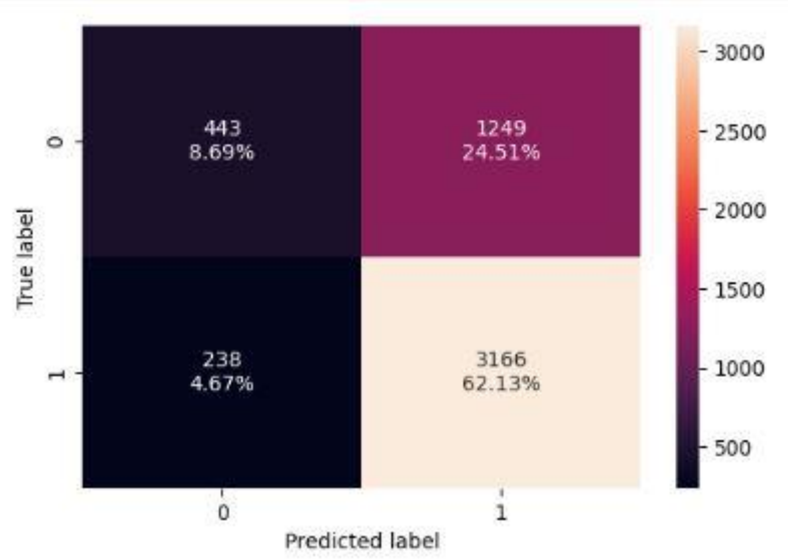
Checking model performance on test set

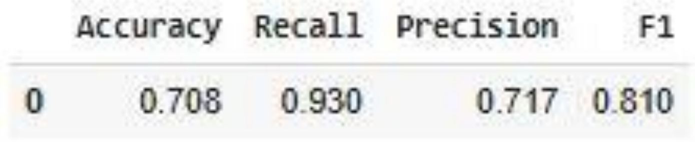




	Accuracy	Recall	Precision	F1
0	0.710	0.931	0.718	0.811

Checking model performance on training set





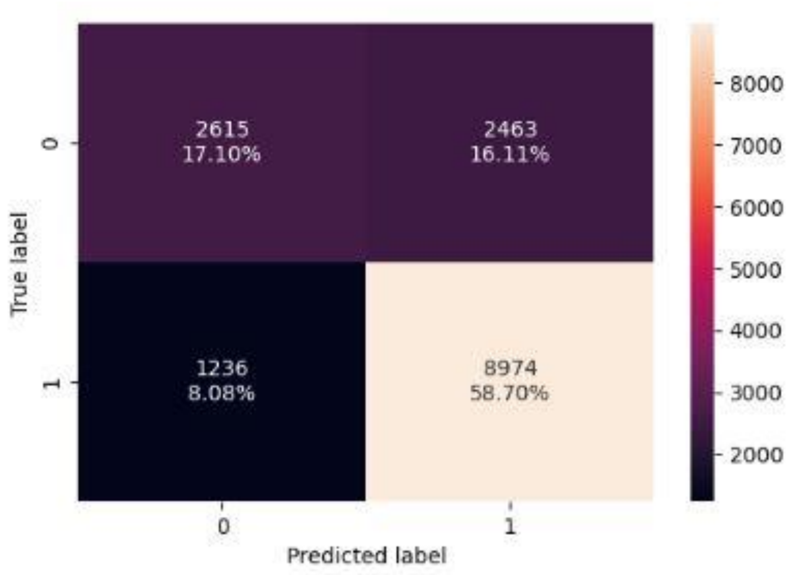
	Accuracy	Recall	Precision	F1
0	0.708	0.930	0.717	0.810

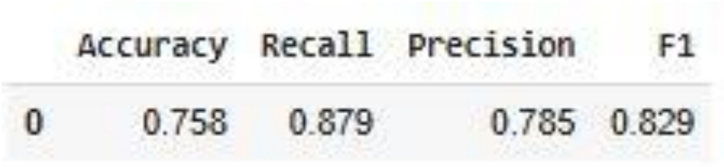
- After tuning the F1 score has minor reduced.
- The recall of the model has improved but the precision has decreased.

5.) <u>Gradient Boosting Classifier - Model Building and Hyperparameter Tuning</u>

- **Model Building – Gradient Boosting Classifier**

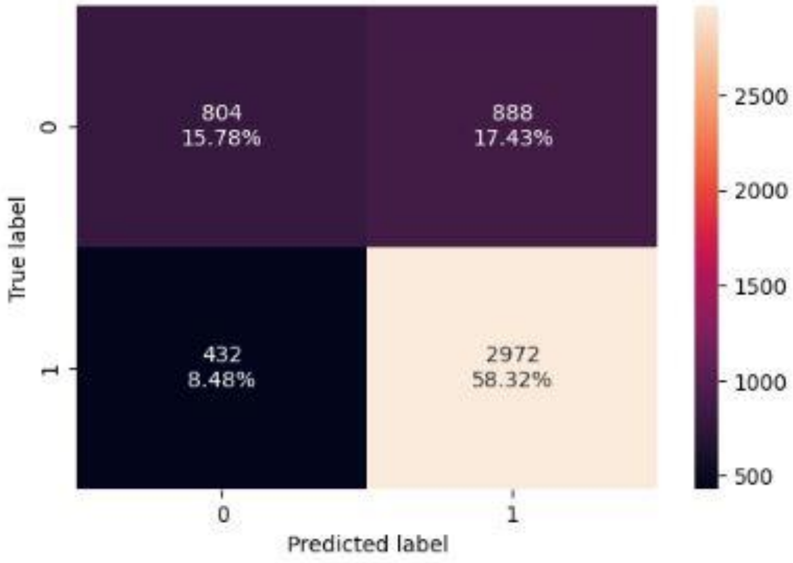
Checking model performance on training set

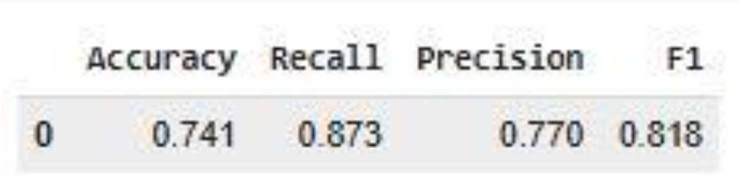




	Accuracy	Recall	Precision	F1
0	0.758	0.879	0.785	0.829

Checking model performance on training set



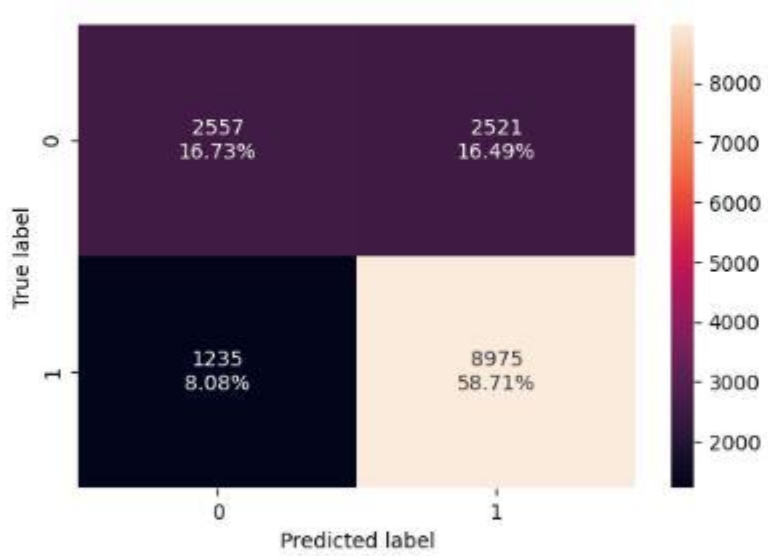


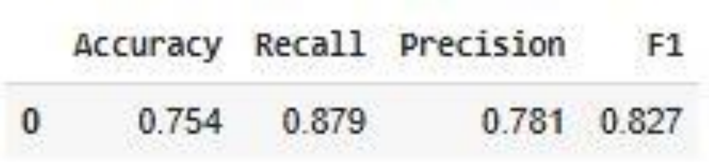
	Accuracy	Recall	Precision	F1
0	0.741	0.873	0.770	0.818

- The model is giving a good and generalized performance
- We are getting the F1 score of 0.83 and 0.82 on the training and test set

- **Hyperparameter Tuning - Gradient Boosting Classifier**

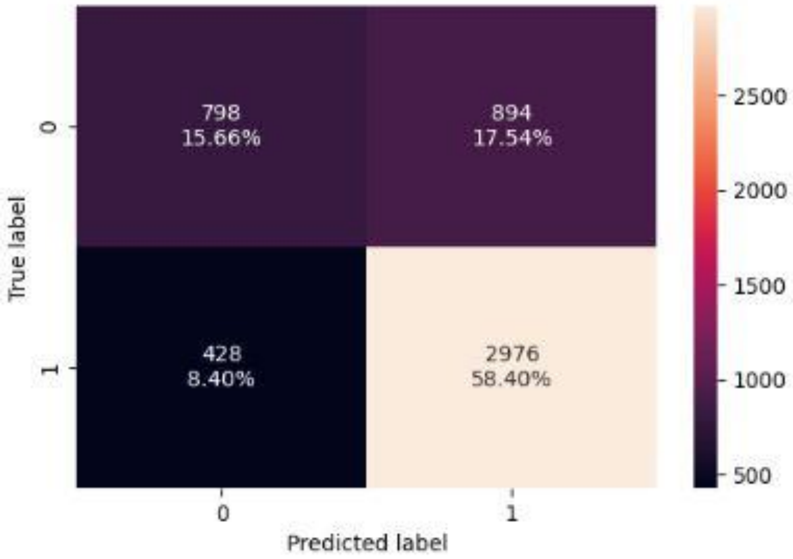
Checking model performance on test set

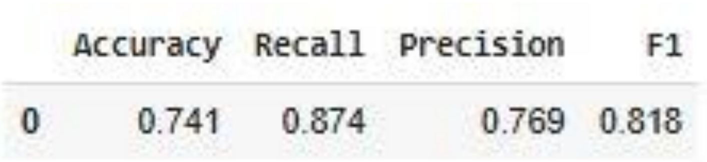




	Accuracy	Recall	Precision	F1
0	0.754	0.879	0.781	0.827

Checking model performance on training set





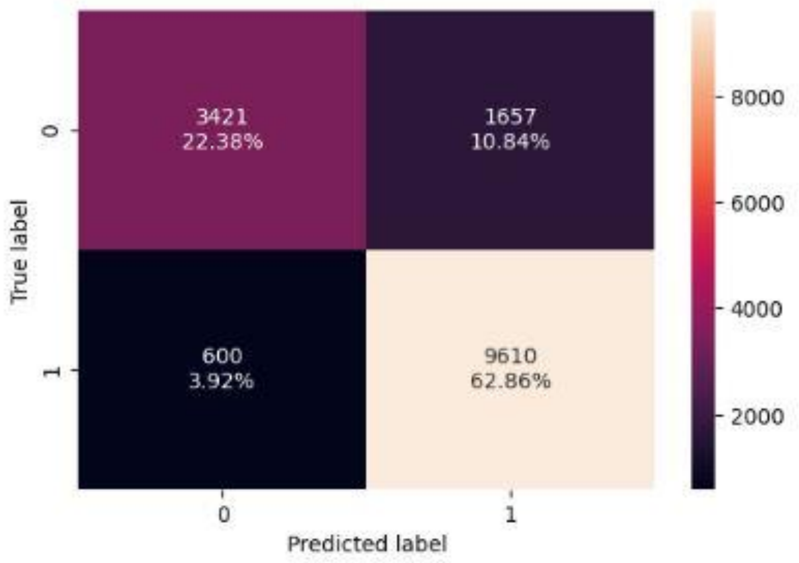
	Accuracy	Recall	Precision	F1
0	0.741	0.874	0.769	0.818

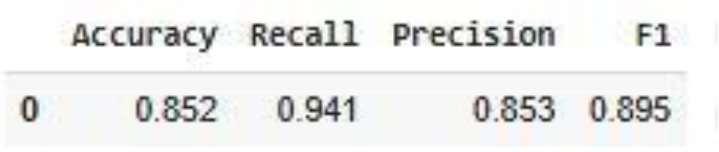
F1 is almost Equal after hyperparameter tunning

6.) <u>XGBoost Classifier - Model Building and Hyperparameter Tuning</u>

- **Model Building - XGBoost Classifier**

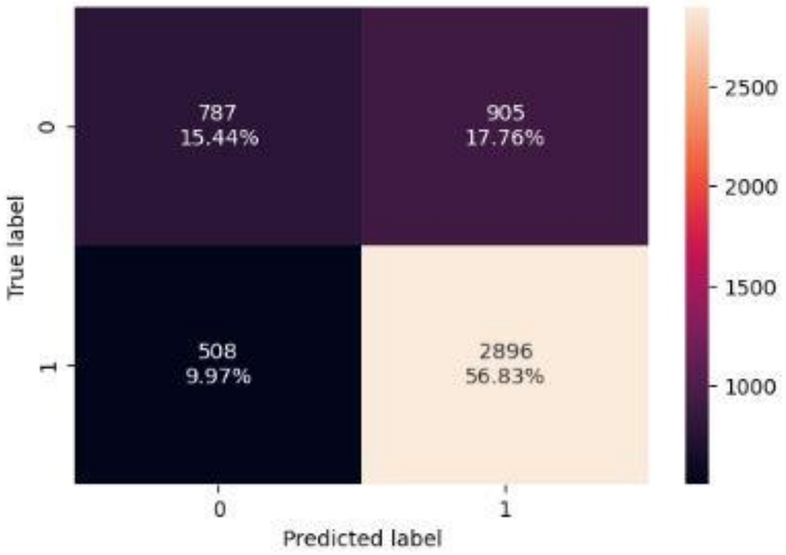
Checking the Performance on Training test

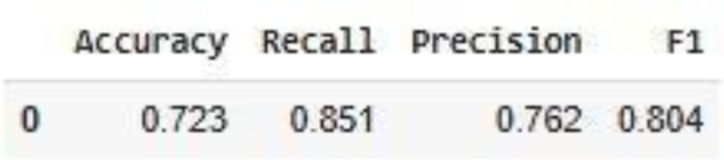




	Accuracy	Recall	Precision	F1
0	0.852	0.941	0.853	0.895

Checking the Performance on Test test





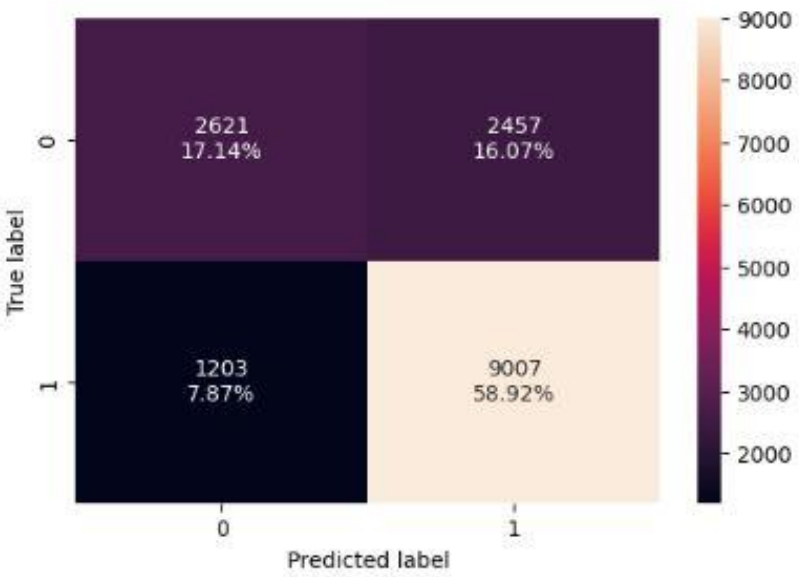
	Accuracy	Recall	Precision	F1
0	0.723	0.851	0.762	0.804

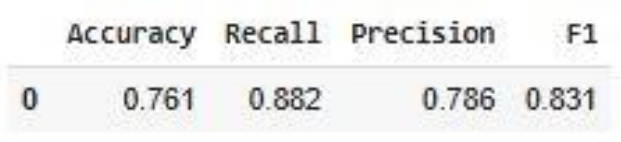
The XGB seems to be also overfitting

7.) Stacking Classifier - Model Building and Hyperparameter Tuning

- **Model Building - Stacking Classifier**

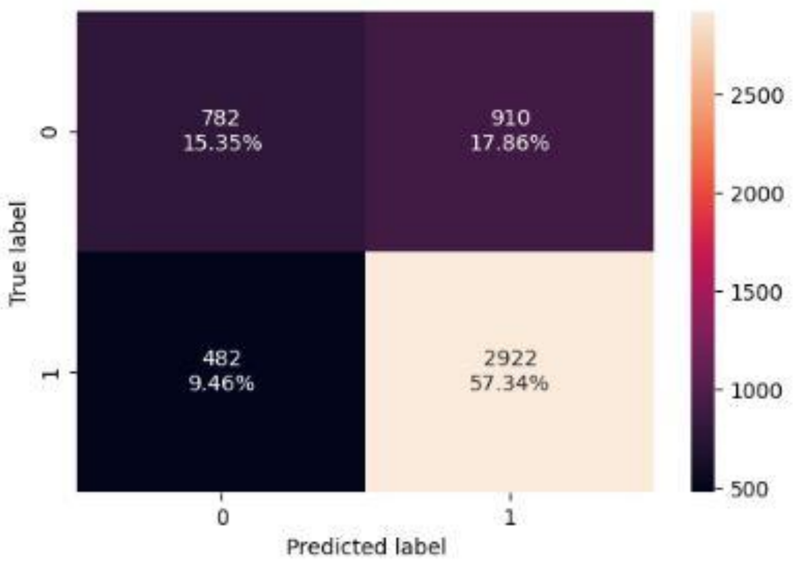
Checking model performance on training set

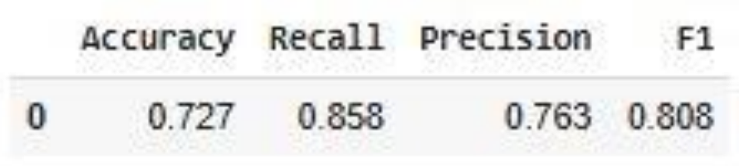




	Accuracy	Recall	Precision	F1
0	0.761	0.882	0.786	0.831

Checking model performance on test set



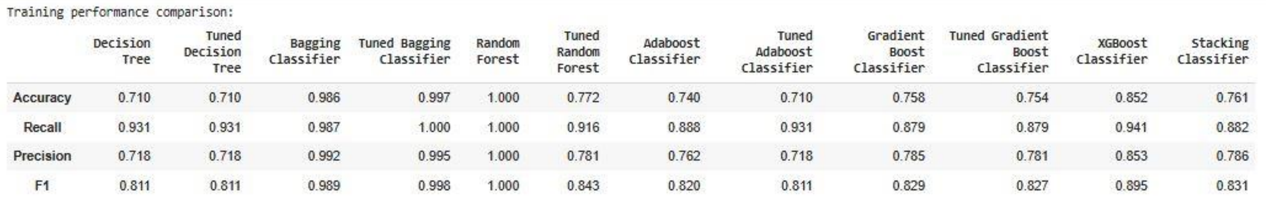


	Accuracy	Recall	Precision	F1
0	0.727	0.858	0.763	0.808

- Stacking model has also given a good and generalized performance.
- We have received F1 scores of 0.83 and 0.81 on the training and test set, respectively.

Model Performance Comparison and Final Model Selection

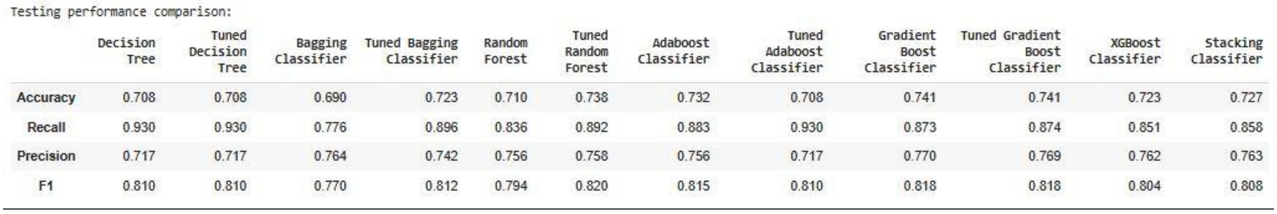
Test Performance



Training performance comparison:

	Decision Tree	Tuned Decision Tree	Bagging Classifier	Tuned Bagging Classifier	Random Forest	Tuned Random Forest	Adaboost Classifier	Tuned Adaboost Classifier	Gradient Boost Classifier	Tuned Gradient Boost Classifier	XGBoost Classifier	Stacking Classifier
Accuracy	0.710	0.710	0.986	0.997	1.000	0.772	0.740	0.710	0.758	0.754	0.852	0.761
Recall	0.931	0.931	0.987	1.000	1.000	0.916	0.888	0.931	0.879	0.879	0.941	0.882
Precision	0.718	0.718	0.992	0.995	1.000	0.781	0.762	0.718	0.785	0.781	0.853	0.786
F1	0.811	0.811	0.989	0.998	1.000	0.843	0.820	0.811	0.829	0.827	0.895	0.831

Test Performance

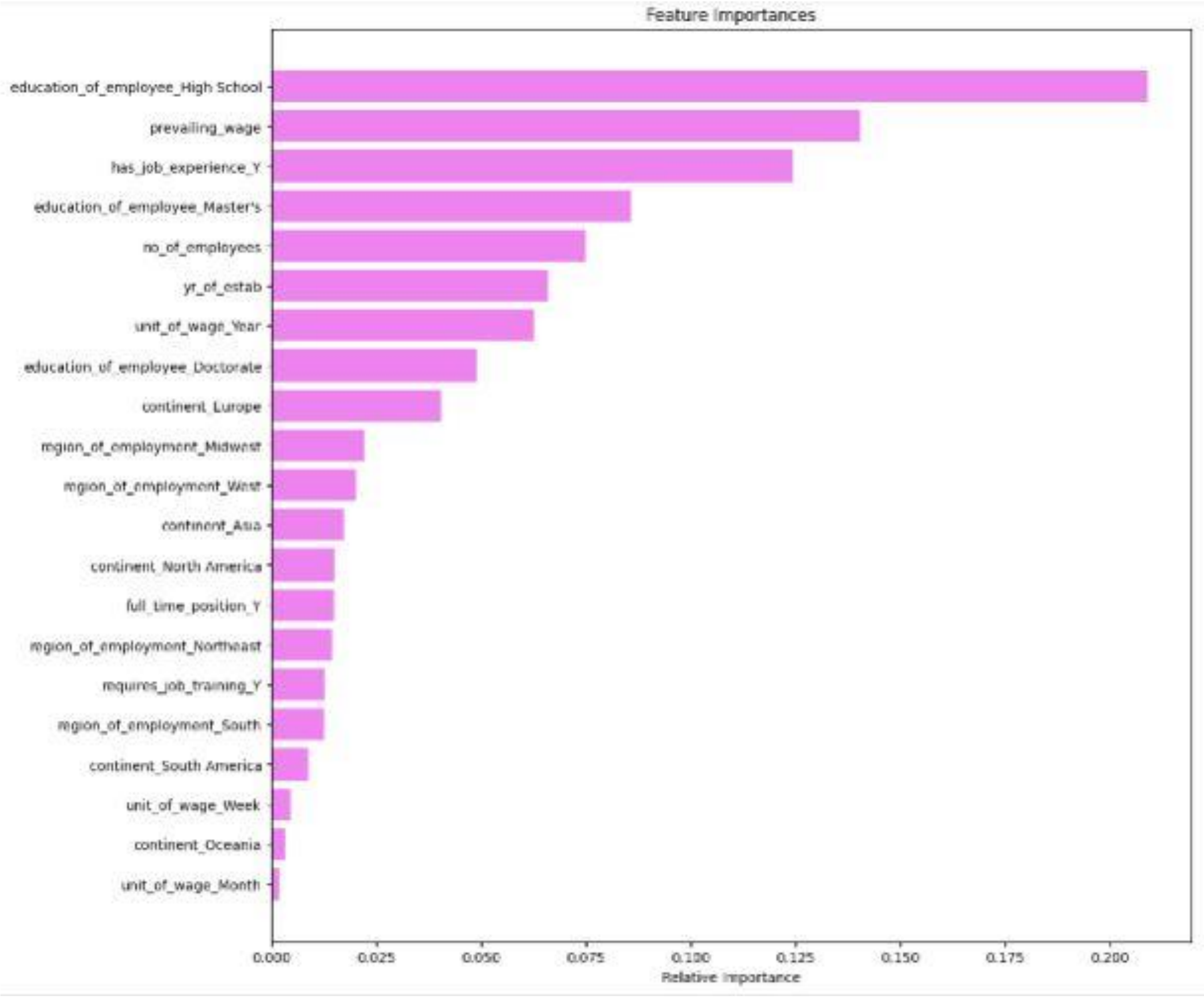


Testing performance comparison:

	Decision Tree	Tuned Decision Tree	Bagging Classifier	Tuned Bagging Classifier	Random Forest	Tuned Random Forest	Adaboost Classifier	Tuned Adaboost Classifier	Gradient Boost Classifier	Tuned Gradient Boost Classifier	XGBoost Classifier	Stacking Classifier
Accuracy	0.708	0.708	0.690	0.723	0.710	0.738	0.732	0.708	0.741	0.741	0.723	0.727
Recall	0.930	0.930	0.776	0.896	0.836	0.892	0.883	0.930	0.873	0.874	0.851	0.858
Precision	0.717	0.717	0.764	0.742	0.756	0.758	0.756	0.717	0.770	0.769	0.762	0.763
F1	0.810	0.810	0.770	0.812	0.794	0.820	0.815	0.810	0.818	0.818	0.804	0.808

• Hyperparameter tuning has decreased the over fit and Stabled F1 score, however, this model is not performing as optimally as the hyperparameter 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 a generalized performance on the training & testing data with a F1 score 0.82 & 0.815
• 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
• Random forest (default & tuned) & Bagging classifier (default & tuned) were found to over fit the training dataset

- <u>Tuned Random Forest model is performing very well</u>
- <u>We will use tuned random forest as the final model</u>



Actionable Insights and Recommendations¶

Observation

The profile of the applicants for whom the visa status can be approved:

the best fit profile is:

- Education level - (Higher education the better) At least has a Bachelor's degree - Master's and doctorate are preferred
- Job Experience - Should have some job experience.
- Prevailing wage - The median prevailing wage of the employees for whom the visa got certified is around 72k.

Secondary information to look at:

- Unit of Wage - Applicants having a yearly unit of wage.
- Continent - Applicants from Europe, Africa, and Asia have higher chances of visa certification.
- Region of employment - Our analysis suggests that the applications to work in the Mid-West region have more chances of visa approval. The approvals can also be made based on requirement of talent, from our analysis we see that:
    - The requirement for the applicants who have passed high school is most in the South region, followed by Northeast region.
    - The requirement for Bachelor's is mostly in South region, followed by West region.
    - The requirement for Master's is most in Northeast region, followed by South region.
    - The requirement for Doctorate's is mostly in West region, followed by Northeast region.

The profile of the applicants for whom the visa status can be denied:

Primary information to look at:

- Education level - Doesn't have any degree and has completed high school.
- Job Experience - Doesn't have any job experience.
- Prevailing wage - The median prevailing wage of the employees for whom the visa got certified is around 65k.

Secondary information to look at:

- Unit of Wage - Applicants having an hourly unit of wage.
- Continent - Ideally the nationality and ethnicity of an applicant shouldn't matter to work in a country but previously it has been observed that applicants from South America, North America, and Oceania have higher chances of visa applications getting denied.
- Additional information of employers and employees can be collected to gain better insights. Information such as:
    - Employers: Information about the wage they are offering to the applicant, Sector in which company operates in, etc
    - Employee's: Specialization in their educational degree, Number of years of experience, etc

Insights

For the Office of Foreign Labor Certification (OFLC), the three most critical components for pre-screening an applicant are:

- Education level
    - An applicant applying for a job requiring a high school diploma will more than likely be denied. Conversely, applications for jobs requiring a Master's degree or doctorate are very likely to be approved.
- Prior job experience
    - An applicant applying for a job without any previous job experience is more likley to be denied than an applicant for a job with experience.
- Prevailing wage
    - The higher the prevailing wage of the job an applicant is applying for, the more likely the application will be approved. This is especially true for applications for jobs with an hourly unit of wage.

Recommendations

- To prioritize limited resources towards screening a batch of applications for those most likely to be approved, the OFLC can:
    - Sort applications by level of education and review the higher levels of education first.
    - Sort applications by previous job experience and review those with experience first.
    - Divide applications for jobs into those with an hourly wage and those with an annual wage, sort each group by the prevailing wage, then review applications for salaried jobs first from highest to lowest wage.
- As stated previously, the Gradient Boosting classifier performs the best of all the models created. However, as shown above, the tuned Decision-Tree model performs barely worse by F1 score and is a far simpler model. This model may be preferable if post-hoc explanations of OFLC decision-making is expected to be required.
    - Furthermore, OFLC should examine more thoroughly why whether an application will be certified or denied can be very well predicted through just three nodes as shown above.

- For those in less skilled, entry-level, and/or hourly jobs, the system would appear to be biased against these applications being certified.