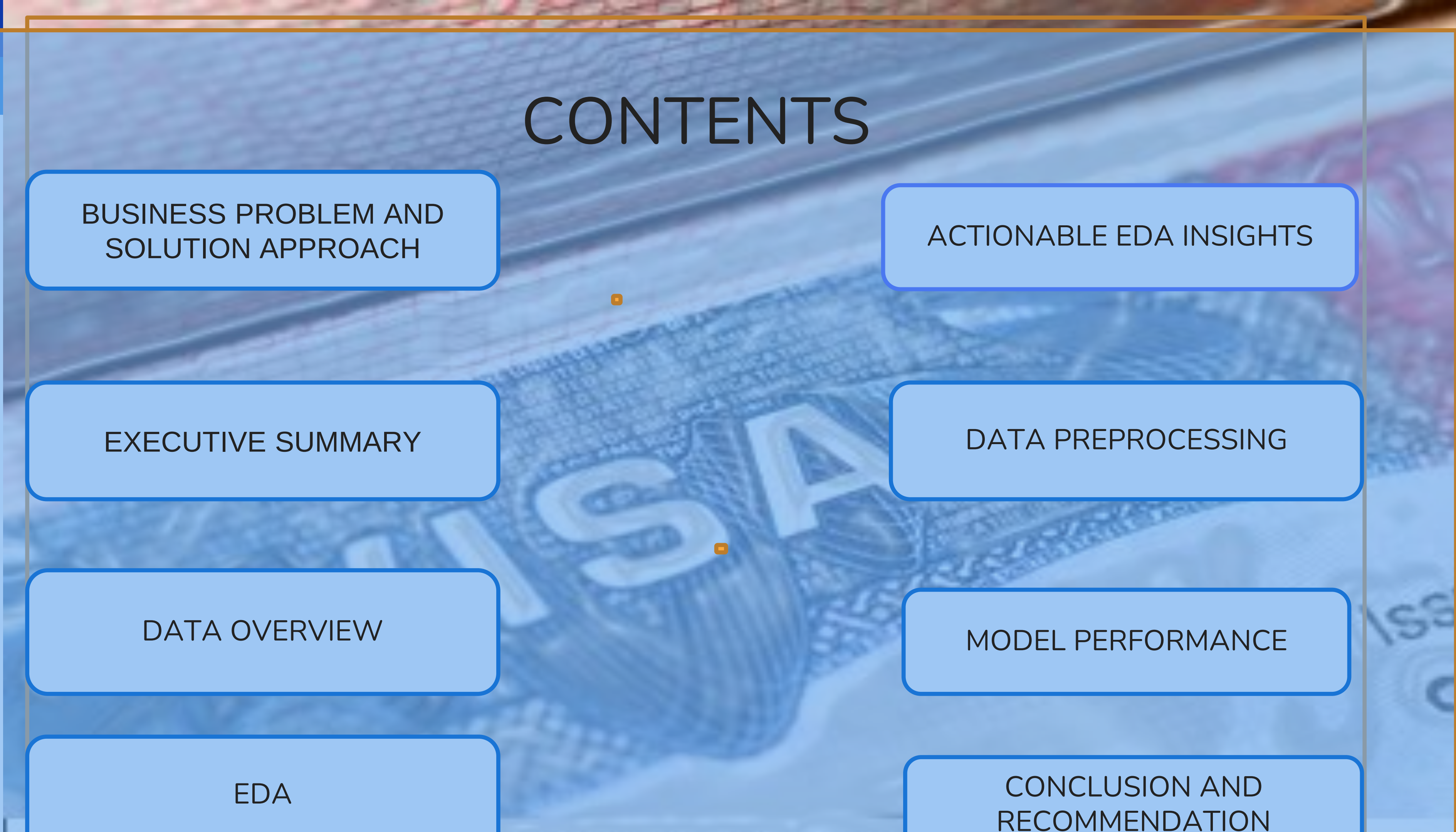
EASY VISA PROJECT ENSEMBLE TECHNIQUE

12/03/22

CONTENTS



- BUSINESS PROBLEM AND SOLUTION APPROACH
- EXECUTIVE SUMMARY
- DATA OVERVIEW
- EDA
- ACTIONABLE EDA INSIGHTS
- DATA PREPROCESSING
- MODEL PERFORMANCE
- CONCLUSION AND RECOMMENDATION

Business problem and Solution Approach

- ❑ The Immigration and Nationality Act (INA) of the US permits foreign workers to come to the United States to work on either a temporary or permanent basis.
- ❑ The immigration programs are administered by the Office of Foreign Labor Certification (OFLC).
- ❑ OFLC processes job certification applications for employers seeking to bring foreign workers into the United States.
- ❑ The process of reviewing every case is becoming a tedious task as the number of applicants is increasing every year.

Solution

As a data scientist at EasyVisa have to analyze the data provided and, 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 influence the case status..

Executive Summary

❑ **The Classification ML model** is able to give generalized prediction on training & testing datasets (not prone to overfitting) and is able to explain over 80% of information (accuracy of 75% on test dataset & F1 score of 82% on test dataset).

❑ The confusion matrix is able to identify a higher % of cases getting certified, but only a smaller % of cases getting denied correctly.

❑ Based on the EDA and the classification Model , the following features were identified as important for visas getting certified than denied

(1) Education of employee
(2) Unit of wage
(3) The continent the employee is from
(4) Region of Employment

DATA OVERVIEW

There **are 25480** observations and **12** features.
Number of employees, years of establishment & prevailing wages are of type integer or float. And 9 categorical values.

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

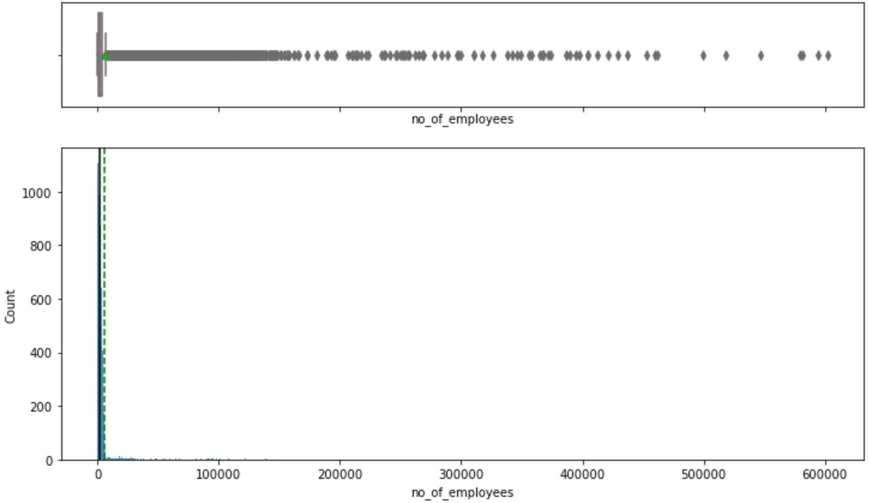
- case_id: ID of each visa application
- continent: Information of continent of the employee
- education_of_employee: Information of education of the employee
- has_job_experience: Does the employee has any job experience? Y= Yes; N = No
- requires_job_training: Does the employee require any job training?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

Statistical Summary

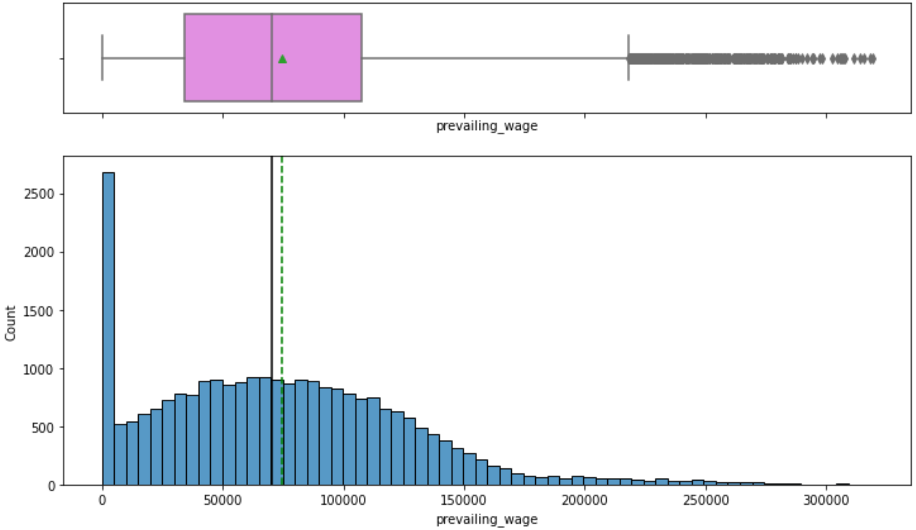
- ❑ The average number of employees in the employer's organization are 5667. The minimum number is negative which does not appear to be a valid data point.
- ❑ There are companies in the dataset with years of establishment from 1800 to 2016.
- ❑ The case ID attribute can be dropped as it is a unique 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 education 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.
- ❑ 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 visa status and only 33.2% have a denied visa status

EDA Results

The distribution of number of employees is highly skewed right.

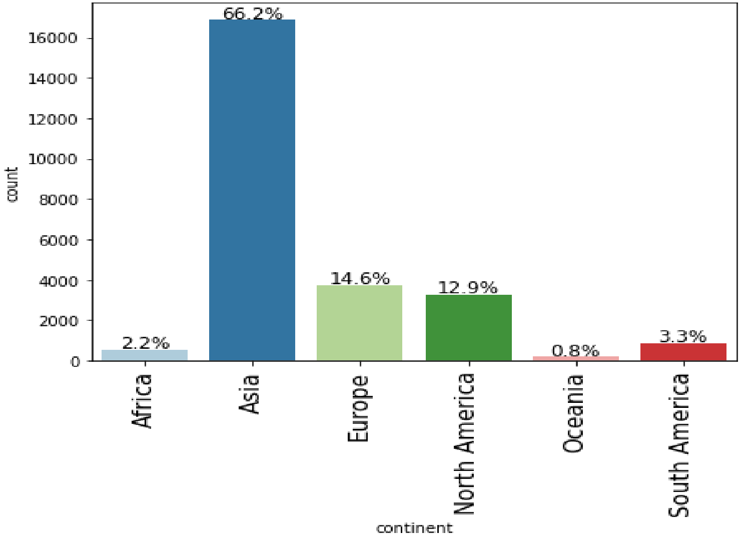


The average prevailing wages in the US is $74,455 while 50% of prevailing wages in the US is $70,308. This implies the date is slightly rightly skewed. The minimum prevailing wages is $2.1367 which is an invalid data point.

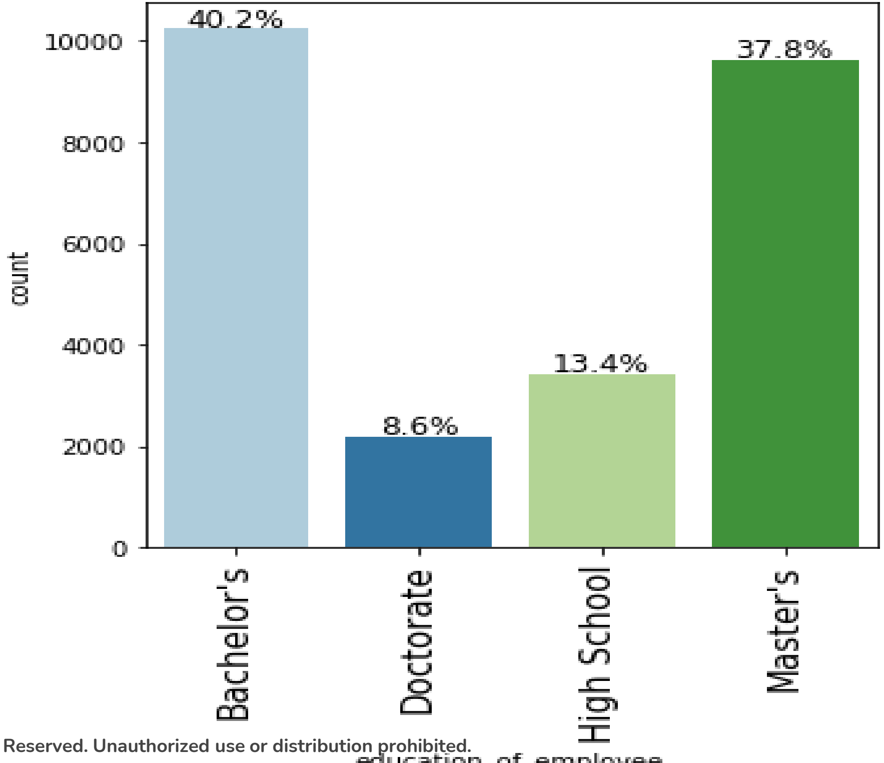


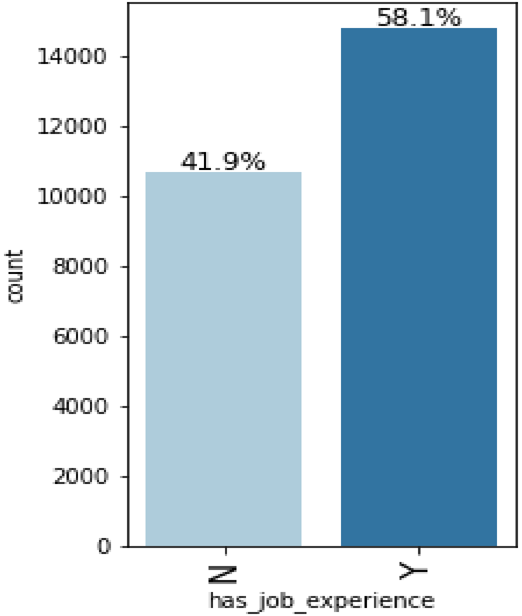
Labelled_BARPLOT

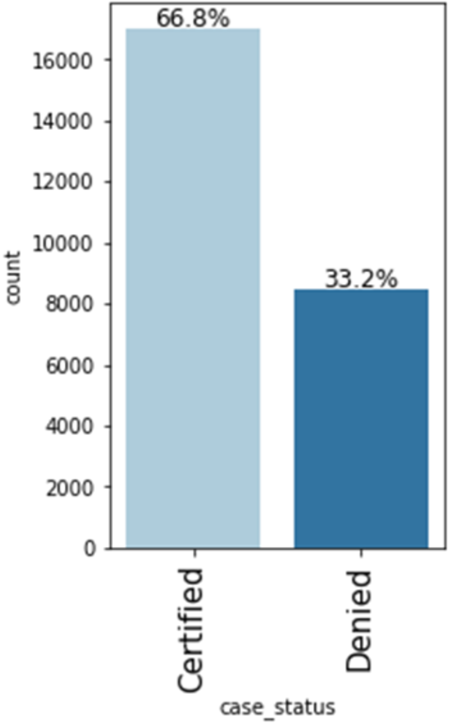
Majority of employees (>50%) are from Asia

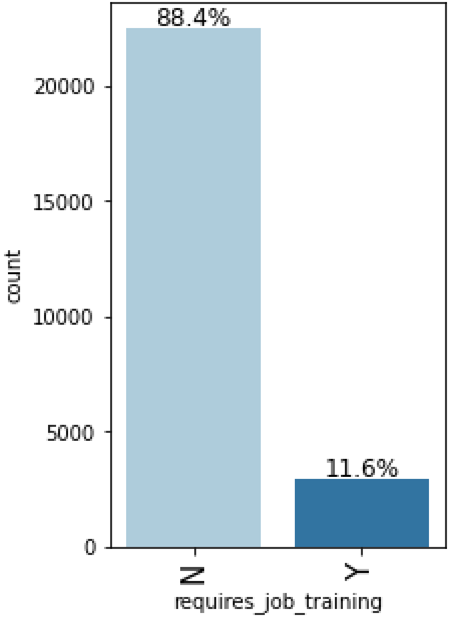


- Majority of employees have either a bachelor's (40%) or a master's (38%)





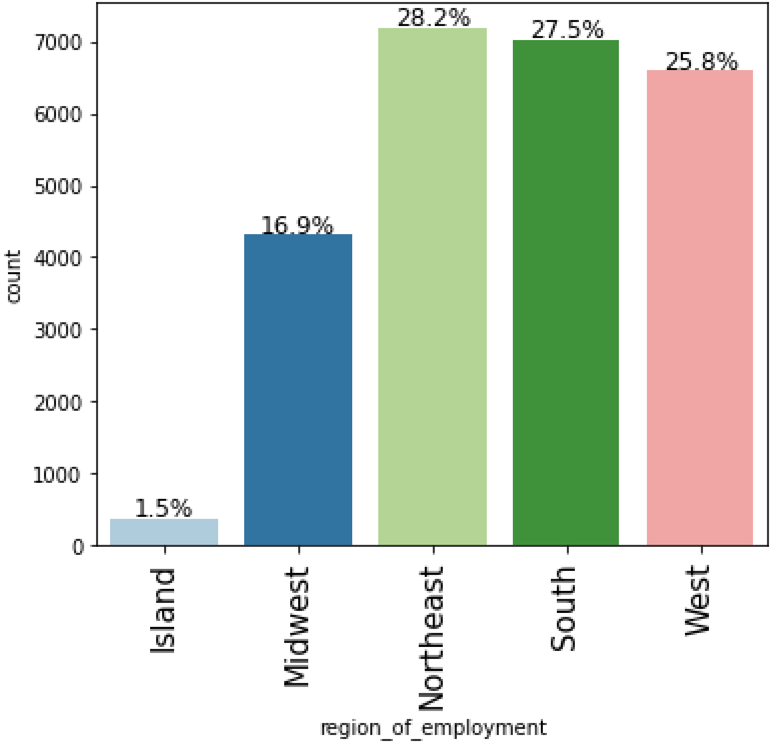


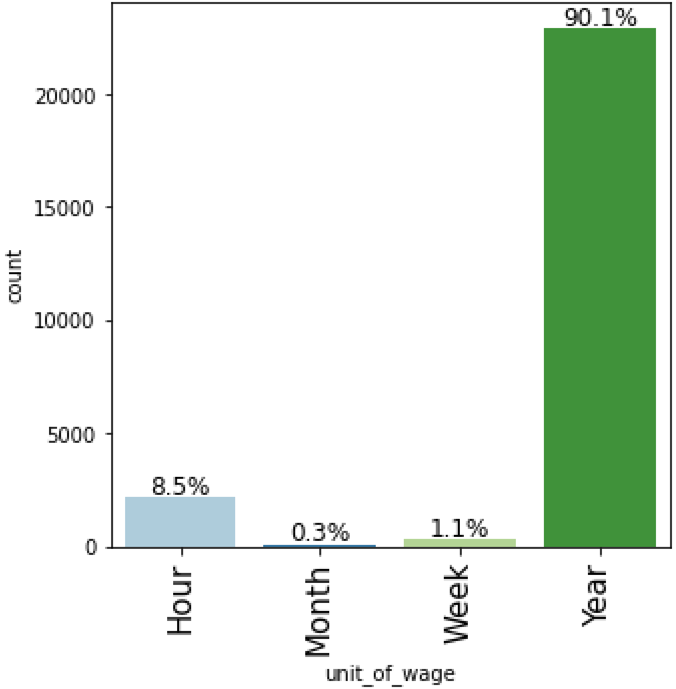


Majority of Applicants have job experience

Around 88.4% of Applicants do not require job training

The % of certified cases are 66.8 and denied cases are 33.2%





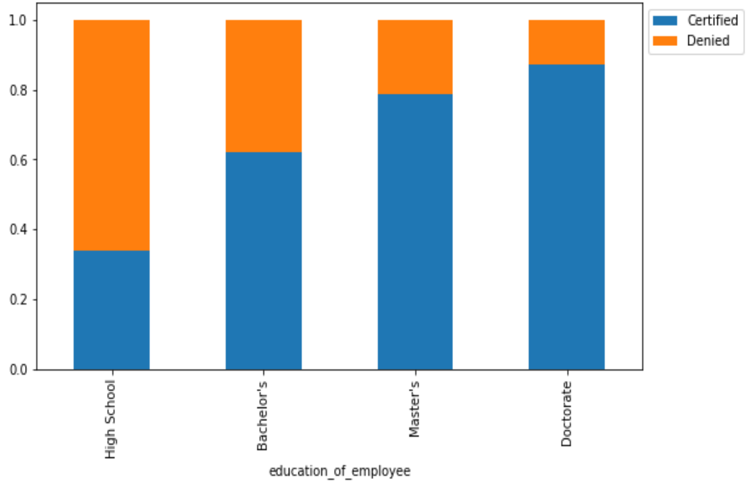
•Northeast, South and West equally have employment opportunities with Human Resource shortages with 25-28% employees applying for visa approval to these regions, followed by Midwest (18%) and Island (1.5%)

•From the EDA, we infer only 8.5% of all cases were for unit_of_wage Hour-ly and the remaining 90.5% of all cases were for unit_of_wage Yearly

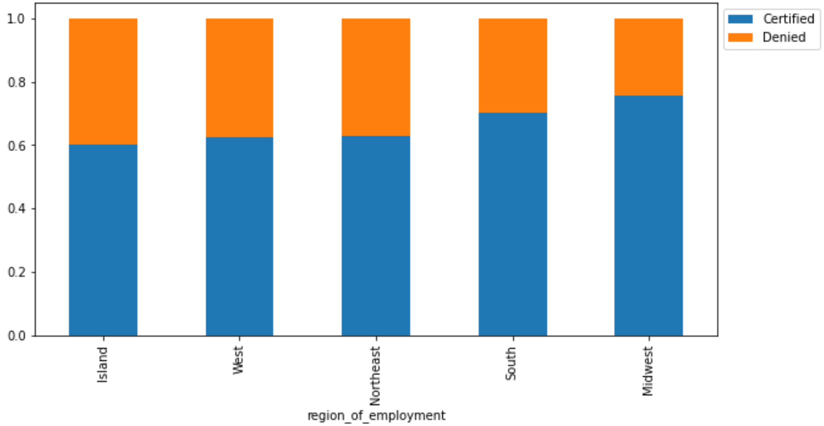
STACKED_BARPLOT

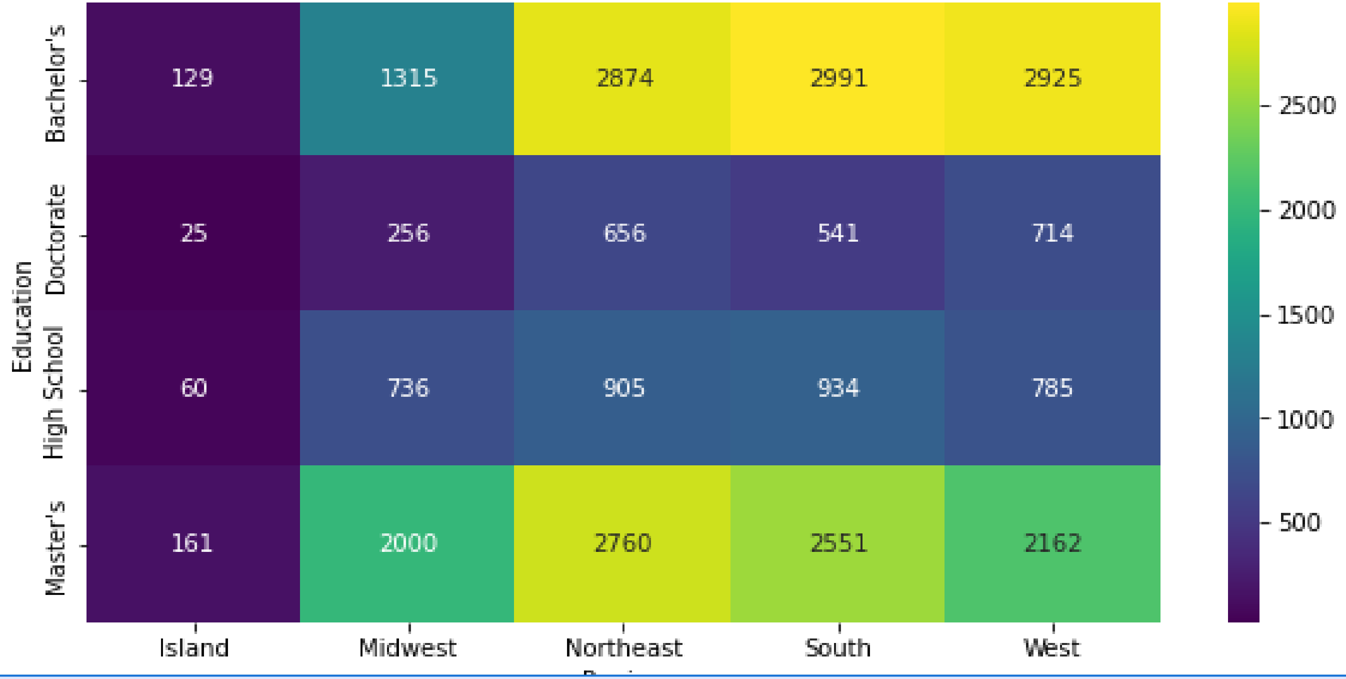
Cases getting certified have the following Qualifications:

- Doctoral (86%), Master (76%), Bachelor (62%),Other (35%)

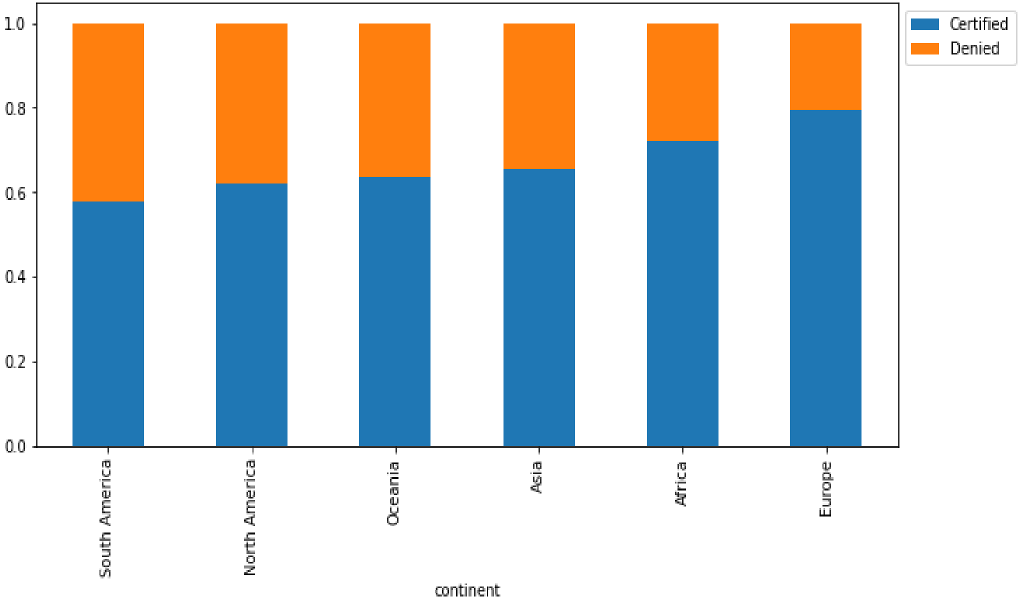


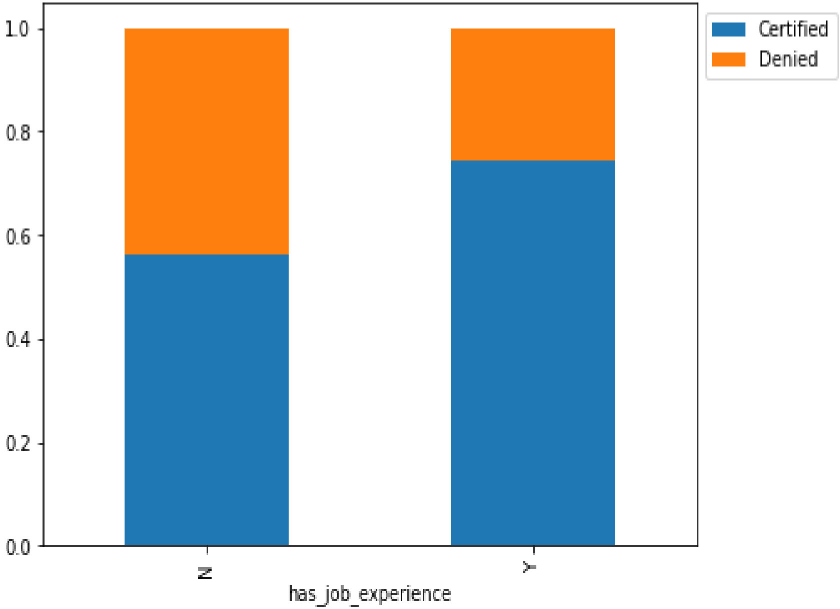
- The cases certified follows the trend Midwest (75% of such cases), then South (70% of such cases), then Northeast, West, & Island (60% of such cases).
- Region of employment being Midwest hence is an important attribute contributing positively to a cases being certified.



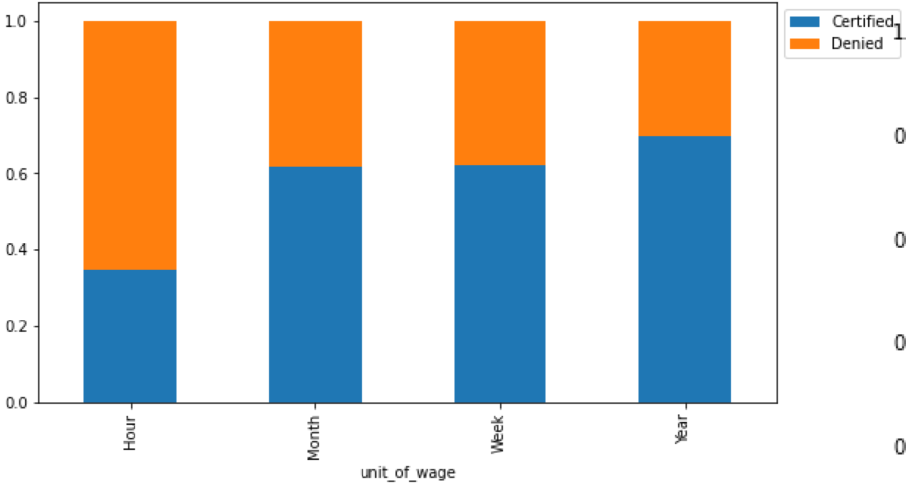


- The NorthEast Region has high number of applicants(2760) with Master’s degree and Bachelor’s(2874),followed by South with applicants(2551) having Master’s and 2991 having Bachelor’s
- 2162 applicants having master’s are from West and 2925 Bachelor’s
- The Island Region has least number of applicants with high level of Education.

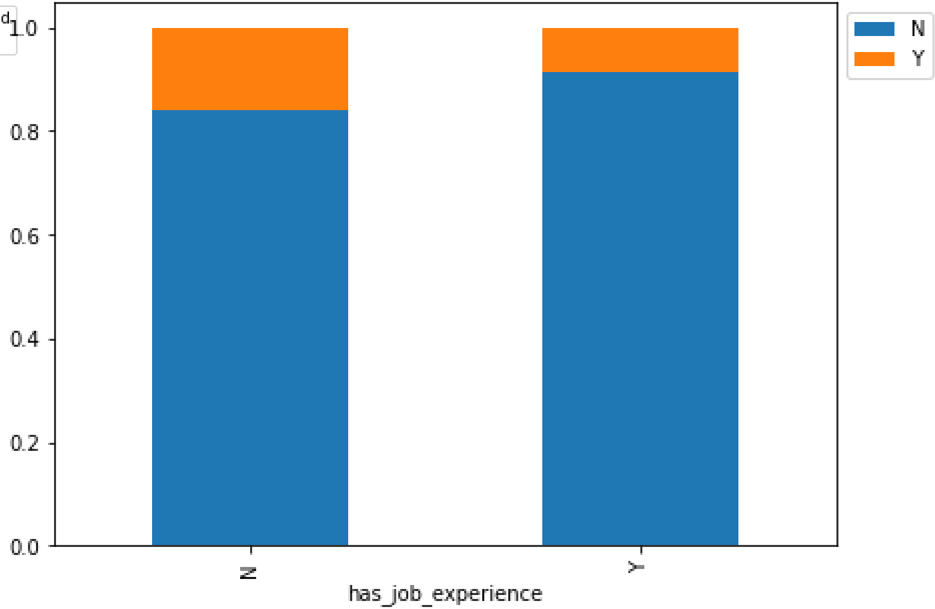




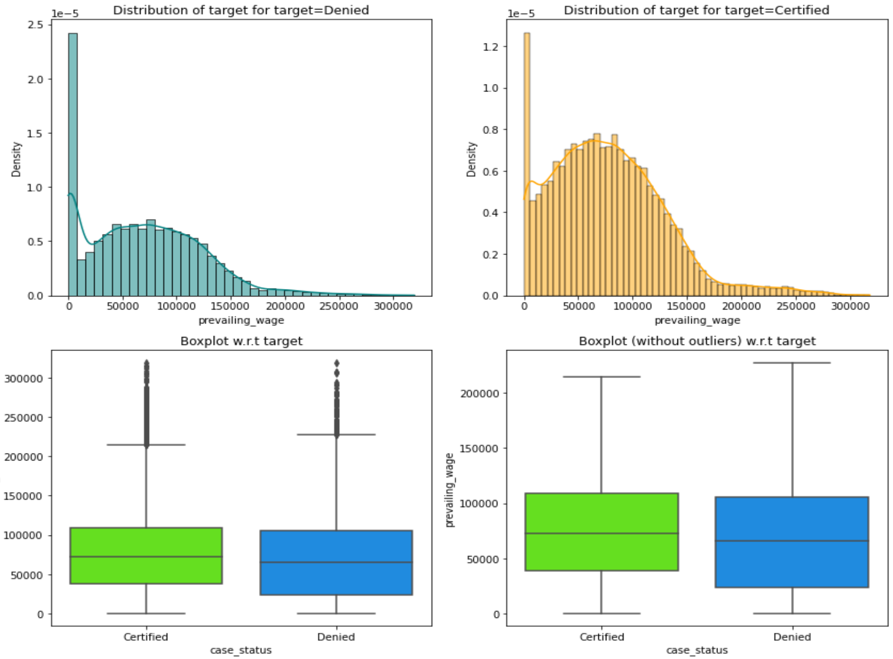
- ❑ Around(70%) the cases certified are of those applicants who has job experience..
- ❑ The cases getting certified is highest for Europe (80% of such cases), then Africa (72% of such cases), then Asia (65% of such cases),

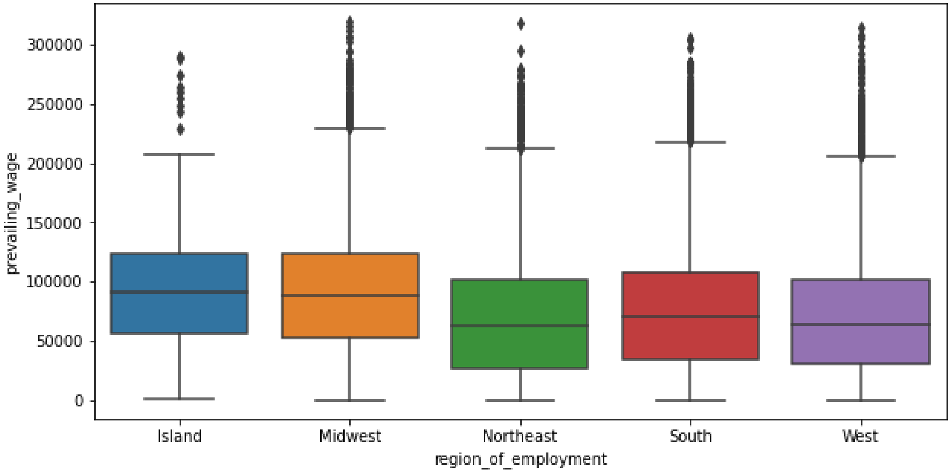


Only 35% of the cases were certified when the unit_of_wage is Hour-ly but 70% were certified when the unit_of_wage are not Hour-ly (i.e, Week-ly, Month-ly or Year-ly). This indicates unit_of_wage is an important attribute that can influence case statuses



We observe that applicants with prior job experience do not require job training





- There are outliers on the higher end (USD 200,000 or more annually) which reference highly skilled positions.
- % certifications in comparison to % denied drops slightly on the lower end of the prevailing_wage and increases slightly on the upper end of the prevailing_wage
- Most of cases certified with high prevailage_wage. The Midwest and Island region of employment have high prevailing_wage.

EDA INSIGHTS

- ❑ 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.
- ❑ 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 Year-ly. This indicates unit_of_wage is an important attribute that can influence case status.
- ❑ Majority of cases are from applicants in Asia (66%), Europe (15%) and then AFRICA(12.9%).however, cases getting certified is highest for Europe (80% of such cases), then Africa (72% of such cases), then Asia (65% of such cases)
- ❑ 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%).
- ❑ **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 status.

•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

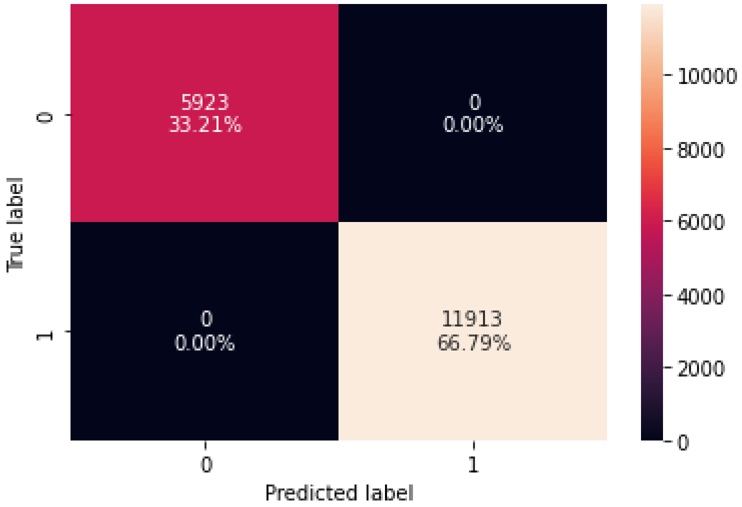
•Majority of the jobs are full time rather than part time. This attribute was not found to have an impact on the case statuses

Data preprocessing

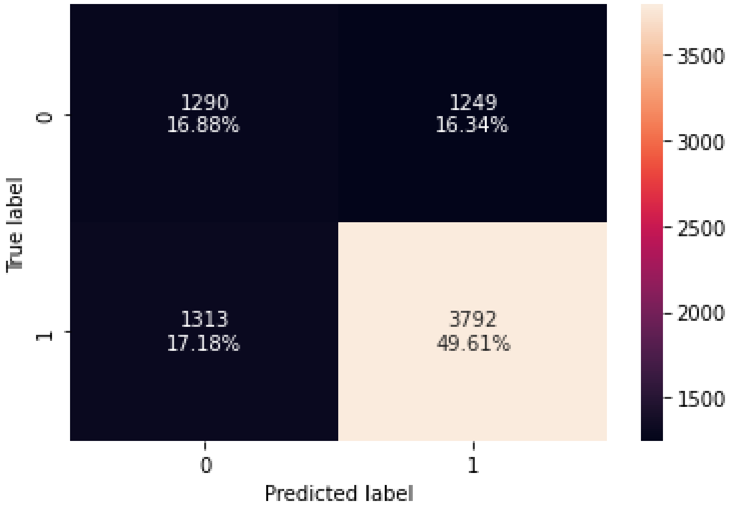
- ❑ There are no **missing** values nor **duplicate** values in the dataset.
- ❑ The are few outliers in the dataset. They are not treated as they are proper values.
- ❑ There are **32 negative values** in the number_of_employee columns. These values were converted to positive one using ABS() Function.
- ❑ The case ID attribute was dropped as it is a unique ID variable and is not expected to add any value to the status of a visa being accepted
- ❑ 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.
- ❑ **F1 Score** can be used 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.

DECISION TREE MODEL PERFORMANCE

Training model



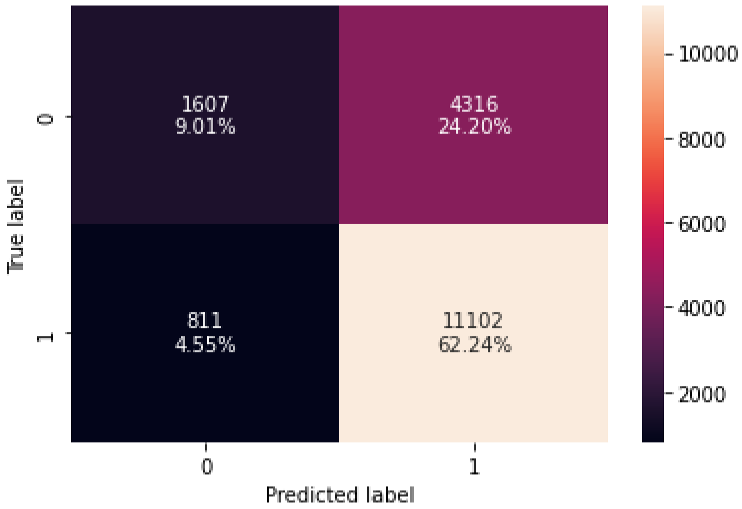
Testing Model



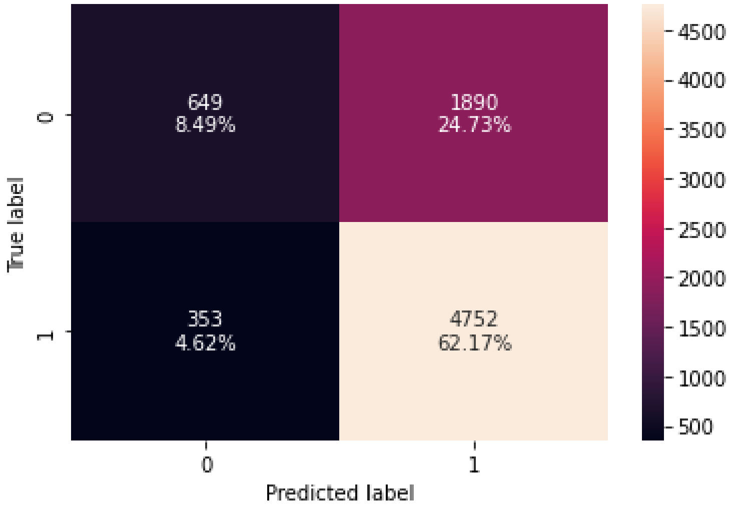
- The decision tree is overfitting the training data. Training metrics are high but the testing metrics are not. F1_score for the test set is only 0.75.
- We can improve model performance by hyperparameter tuning

Hyperparameter Tuning- Decision Tree

Training dataset

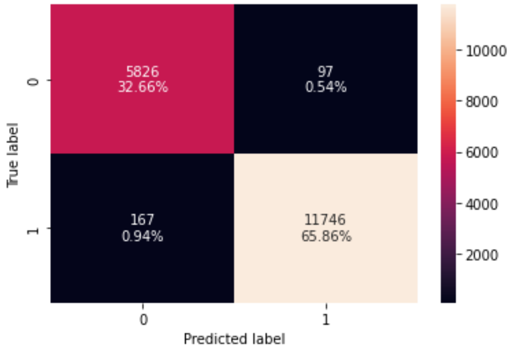


Test dataset

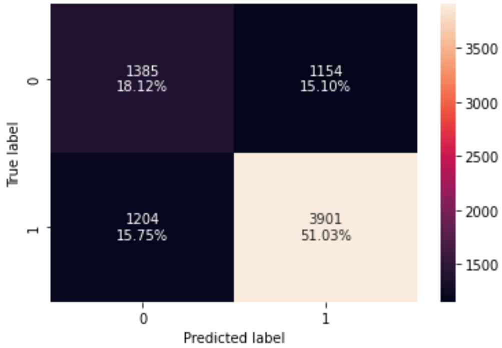


- The hyperparameter tuned decision tree is not overfiting the dataset, as well the F1 score has improved.
- F1 score for both the train and test datasets are 0.812 & 0.809 respectively

BAGGING CLASSIFIER

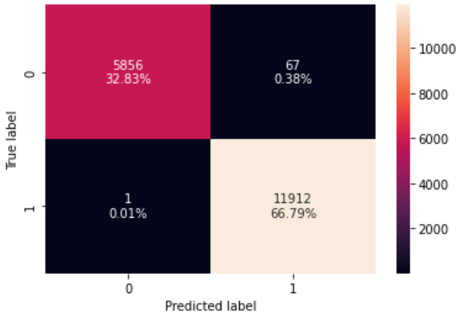


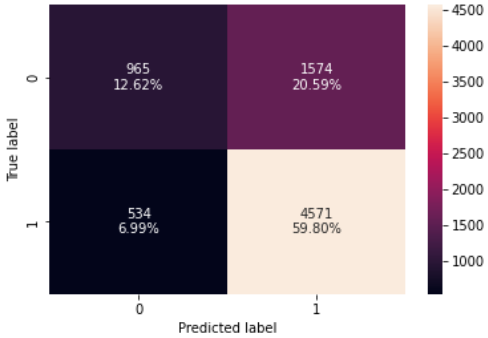
Confusion matrix Training for dataset



Confusion matrix for test dataset

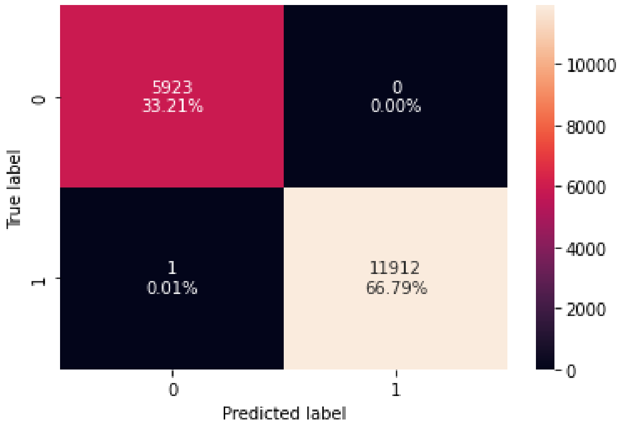
- Bagging classifier is also overfiting the training data.
- F1 score for both the train and test datasets are 0.982 & 0.769 respectively



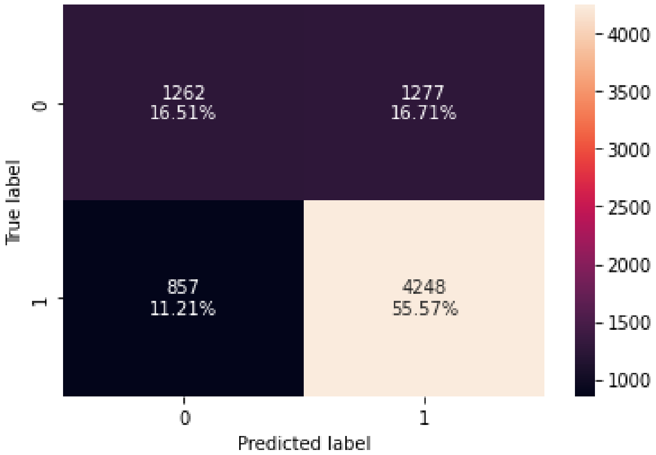


- The model is still found to overfit the training data, as the training metrics are high but the testing metrics are not

RANDOM FOREST CLASSIFIER CONFUSION MATRIX



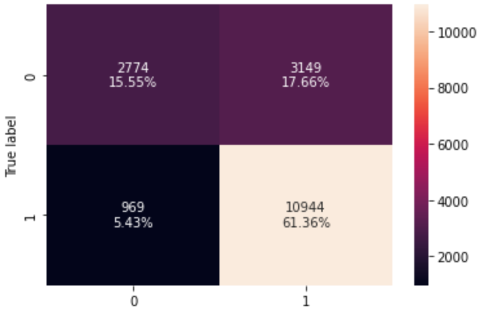
TRAINING DATASET

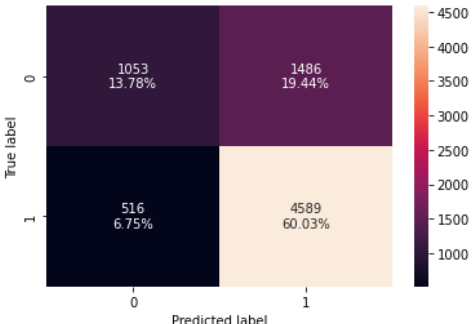


TEST DATASET

- Random forest model is also overfitting the training data
- Hyperparameter tuning has decreased the overfit and increased F1 score, however, this model is not performing as optimally as the hyperparameter tuned decision tree

HYPERPARAMETER TUNING RANDOM FOREST

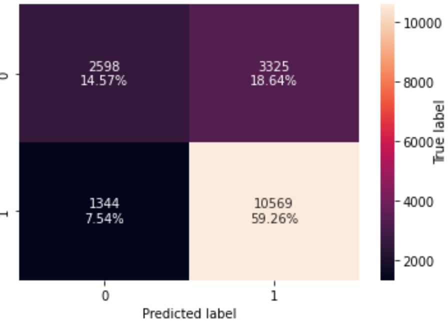


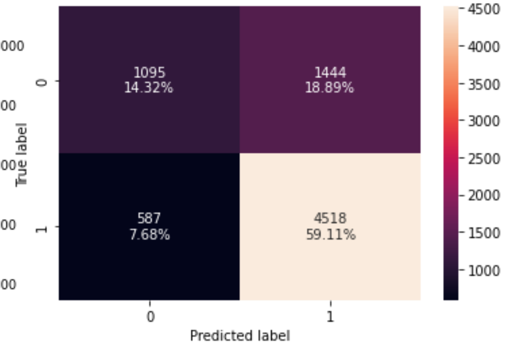


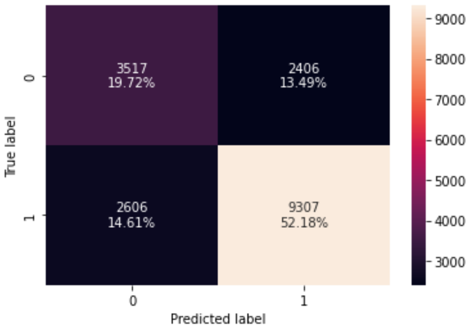
ADABOOST CLASSIFIER MODEL

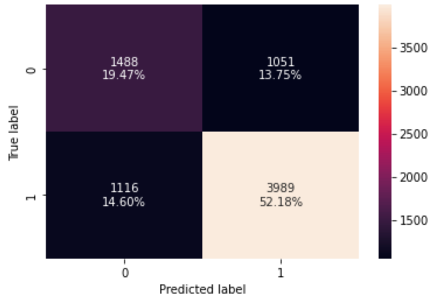
CONFUSION MATRIX

HYPERTUNING MODEL







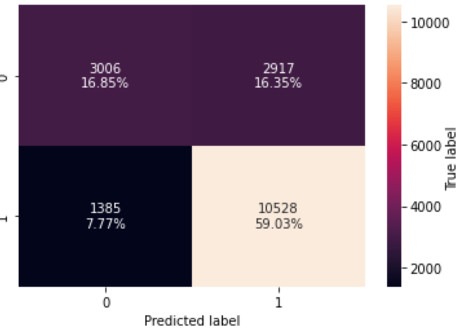


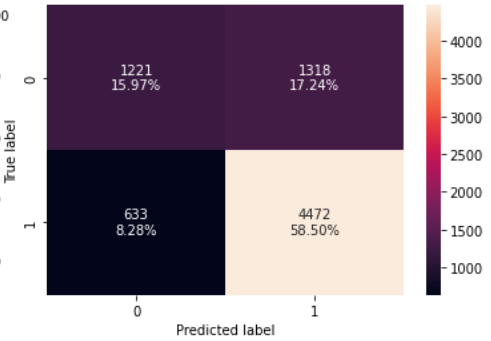
- Unlike the decision tree, random forest, or the bagging classifier; the AdaBoost classifier is not found to overfit the training data. It is giving a generalized performance on the training & testing data with a F1 score 0.819 & 0.816
- The hyperparameter tuned model is giving similar performance to the default AdaBoost model

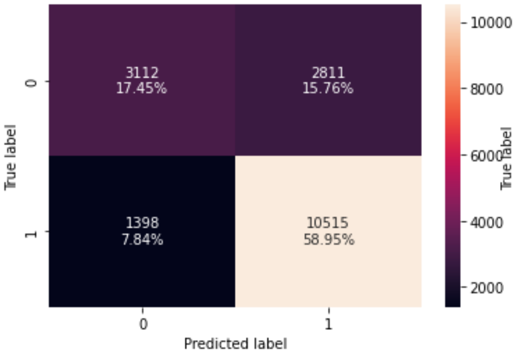
GRADIENT BOOSTING CLASSIFIER

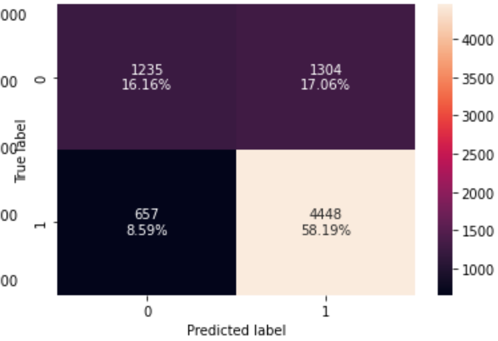
CONFUSION MATRIX FOR TRAIN AND TEST

HYPERTUNING MODEL









COMPARING MODEL PERFORMANCE OF ALL THE MODELS

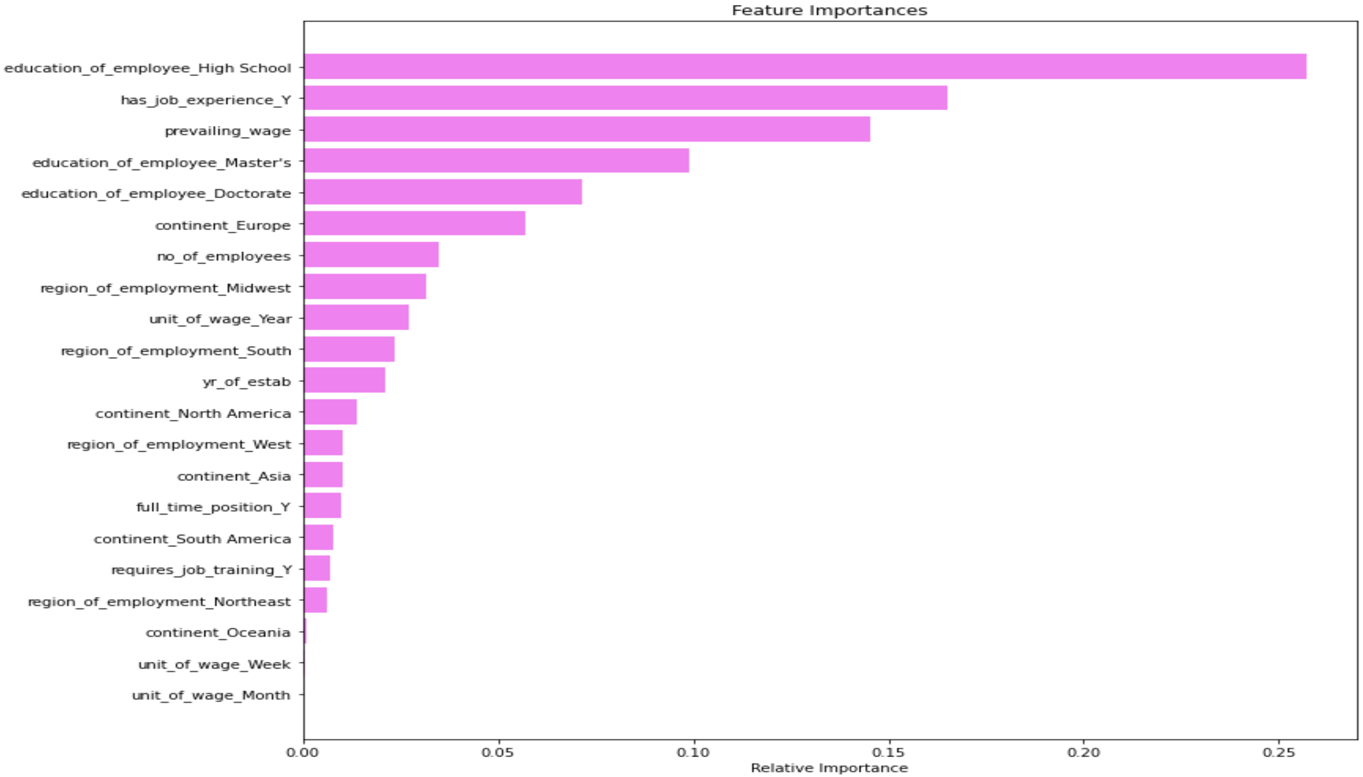
Training performance comparison:

METRICS	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	Stacking Classifier
Accuracy	1.0	0.712548	0.985198	0.996187	0.999944	0.769119	0.738226	0.718995	0.758802	0.764017	0.770576
Recall	1.0	0.931923	0.985982	0.999916	0.999916	0.918660	0.887182	0.781247	0.883740	0.882649	0.854361
Precision	1.0	0.720067	0.991810	0.994407	1.000000	0.776556	0.760688	0.794587	0.783042	0.789059	0.811966
F1	1.0	0.812411	0.988887	0.997154	0.999958	0.841652	0.819080	0.787861	0.830349	0.833234	0.832624

COMPARING MODEL PERFORMANCE ON TEST DATASET

METRICS	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	Stacking Classifier
Accuracy	0.664835	0.706567	0.720827	0.738095	0.691523	0.724228	0.734301	0.716510	0.744767	0.743459	0.739403
Recall	0.742801	0.930852	0.832125	0.898923	0.764153	0.895397	0.885015	0.781391	0.876004	0.871303	0.836239
Precision	0.752232	0.715447	0.768869	0.755391	0.771711	0.743857	0.757799	0.791468	0.772366	0.773296	0.786912
F1	0.747487	0.809058	0.799247	0.820930	0.767913	0.812622	0.816481	0.786397	0.820927	0.819379	0.810826

Important features of the final model



- ❑ We observe that education of an employee is the important feature in getting visa certified
- ❑ It also confirms important attributes such as employee having prior job experience, unit of wage and continent of the employee factor in determining is a VISA would be certified.

CONCLUSION AND RECOMMENDATION

- ❑ The GradientBoost hyperparameter tuned ML model is able to give generalized prediction on training & testing datasets (not prone to overfitting) and is able to explain over 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 can 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 reevaluation of cases getting denied can be carried out in case there is a 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 time spent in the process.

CONCLUSION AND RECOMMENDATION

❑ Based on the EDA and the GradientBoost(tuned) model, 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 an employee 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 a non-hourly pay (week-ly, month-ly or year-ly) 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 has no 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 if the 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 have much bearing on a case getting certified vs denied.

Great Learning

Happy Learning !