EASY VISA

BY: DR. CYNTHIA OFFOHA

AGENDA

- EXECUTIVE SUMMARY
- BUSINESS PROBLEM OVERVIEW & SOLUTION APPROACH
- DATA OVERVIEW
- EDA – UNIVARIATE ANALYSIS & KEY QUESTIONS
- EDA - BIVARIATE ANALYSIS & KEY QUESTIONS
- DATA PRE-PROCESSING
- MODEL BUILDING & IMPROVEMENT
- RECOMMENDATIONS & CONCLUSIONS

EXECUTIVE SUMMARY

- Communities' business within the United States are experiencing increased demand for human resources but the greatest challenge is acquiring the right talent
- United States institutions are aggressively looking for hard-working, talented and qualified candidates that reside locally and abroad
- It is shown that the U. S INA (Immigration and Nationality Act) allows foreign employees to enter the U.S to work on a temporary or permanent basis
    - Protects U.S employees against adverse impacts on wages or working conditions
    - Administered by Office of Foreign Labor Certification (OFLC)
        - Processes job certification applications for employers wanting to bring foreign workers into U.S
        - Grants certifications to those when enough U. S workers are not available

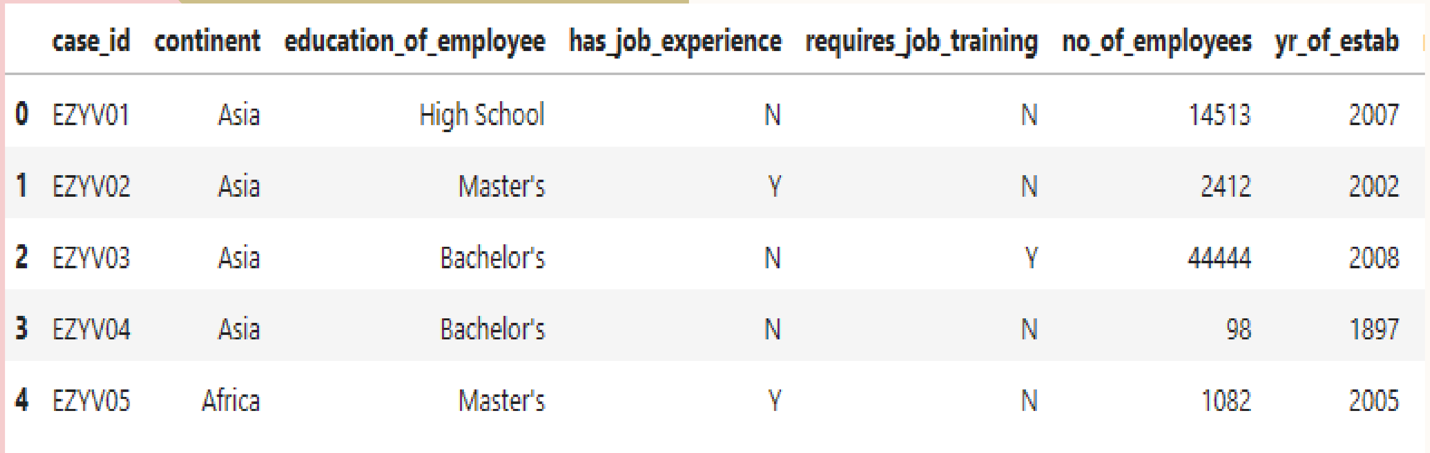
BUSINESS PROBLEM OVERVIEW & SOLUTION APPROACH

- Within FY 2016, OFLC analyzed 775,979 employer applications for 1,699,957 positions for temporary and permanent labor certifications
- ~ 9% increase in total #processed applications from previous year
  - The process of case review became very tedious as there is an increased in the number of applicants each year
    - This led to the intervention for Machine Learning based solution to assist in shortlisting candidates having greater probability of VISA approval
- OFLC gave contract to firm EasyVisa for data-driven solutions
  - EasyVisa data scientist must examine data provided with the assistance of clarification model with the objective to:
    - Enhance the process of visa approvals
    - Recommend suitable profile for applicants for which the visa should be certified or denied based upon drivers that influence the case status

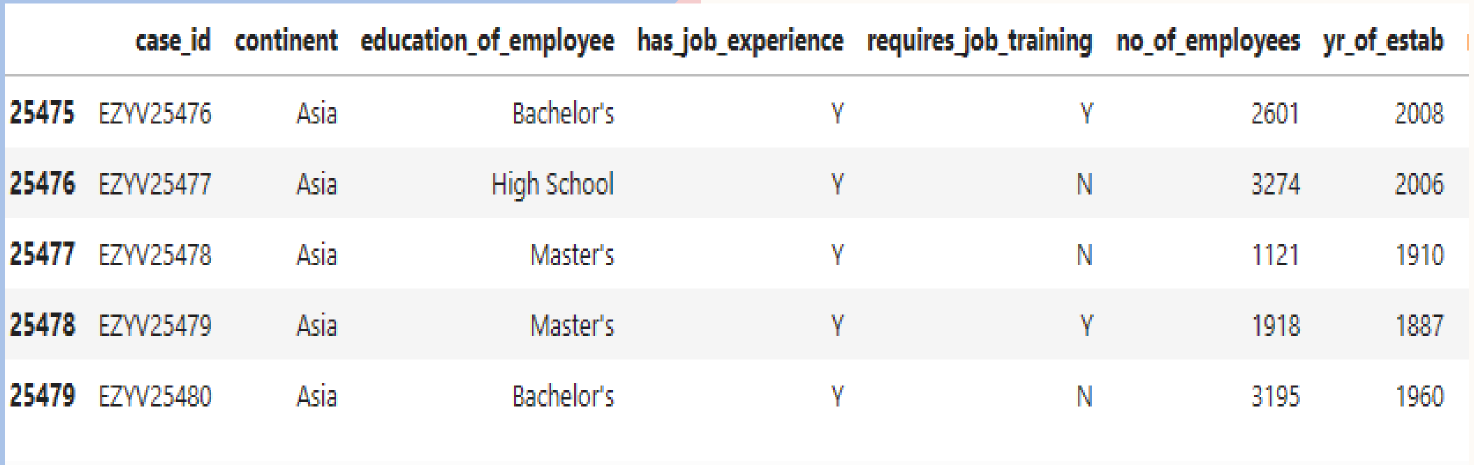
DATA OVERVIEW

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

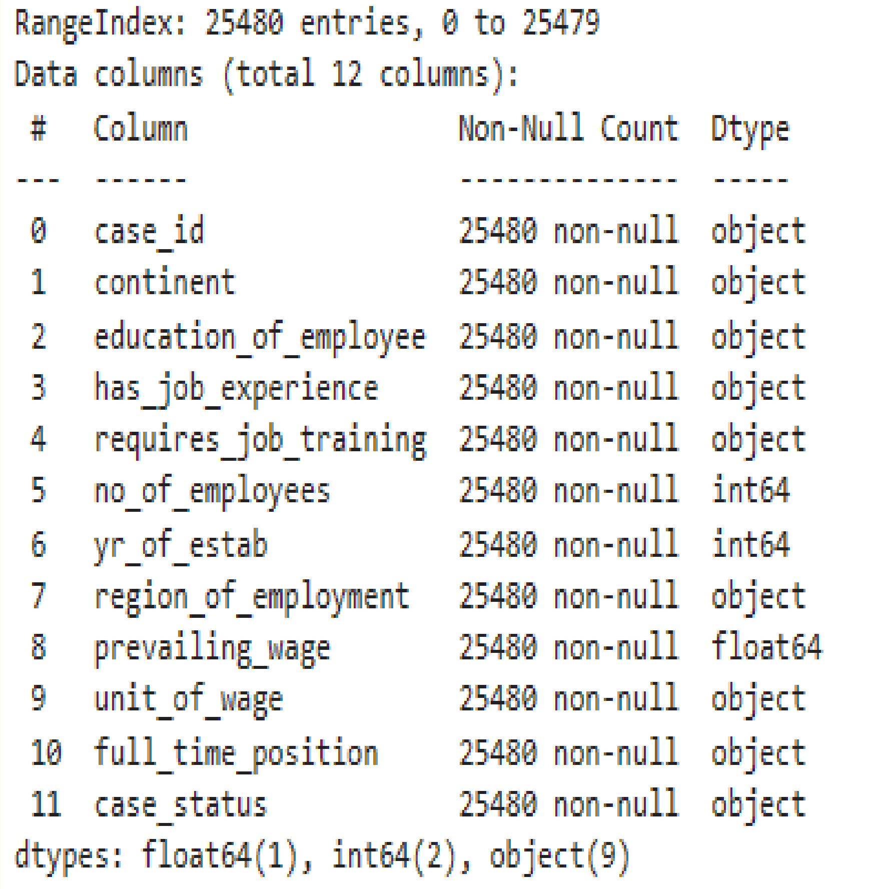
DATA STRUCTURE



	case_id	continent	education_of_employee	has_job_experience	requires_job_training	no_of_employees	yr_of_estab
0	EZYV01	Asia	High School	N	N	14513	2007
1	EZYV02	Asia	Master's	Y	N	2412	2002
2	EZYV03	Asia	Bachelor's	N	Y	44444	2008
3	EZYV04	Asia	Bachelor's	N	N	98	1897
4	EZYV05	Africa	Master's	Y	N	1082	2005



	case_id	continent	education_of_employee	has_job_experience	requires_job_training	no_of_employees	yr_of_estab
25475	EZYV25476	Asia	Bachelor's	Y	Y	2601	2008
25476	EZYV25477	Asia	High School	Y	N	3274	2006
25477	EZYV25478	Asia	Master's	Y	N	1121	1910
25478	EZYV25479	Asia	Master's	Y	Y	1918	1887
25479	EZYV25480	Asia	Bachelor's	Y	N	3195	1960

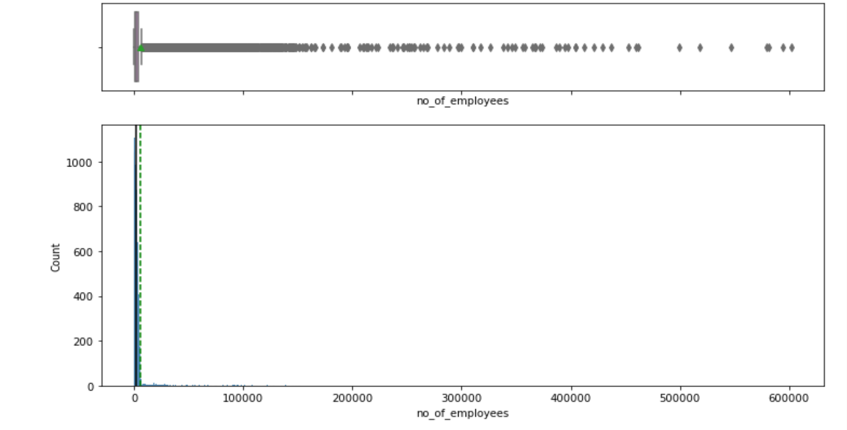


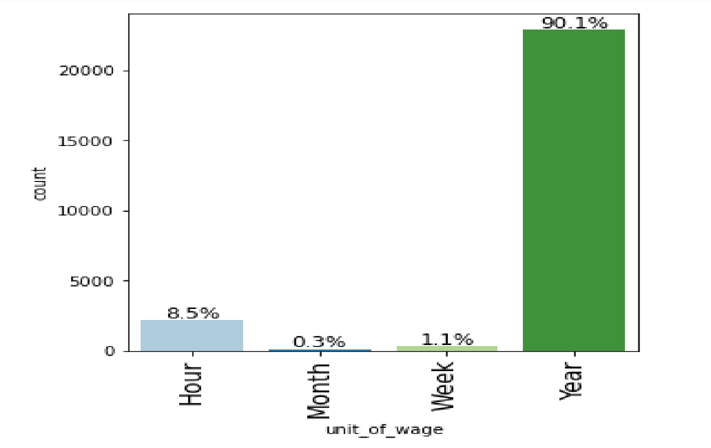
```
RangeIndex: 25480 entries, 0 to 25479
Data columns (total 12 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   case_id                25480 non-null  object
 1   continent              25480 non-null  object
 2   education_of_employee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

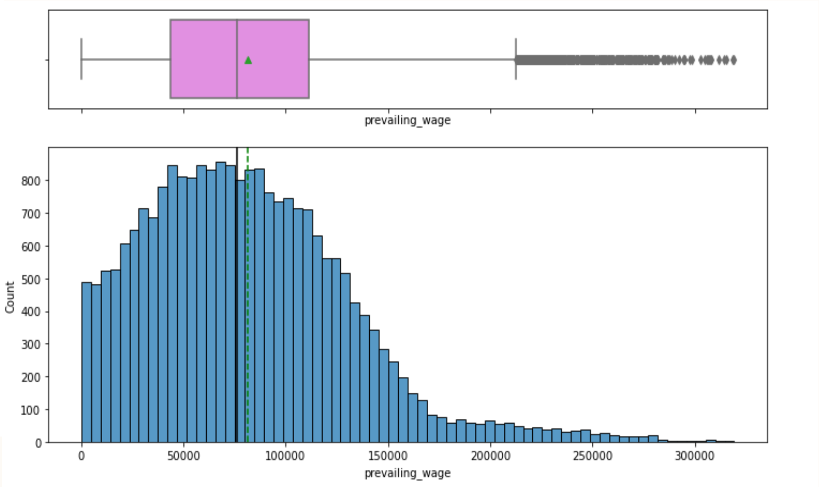
OBSERVATIONS

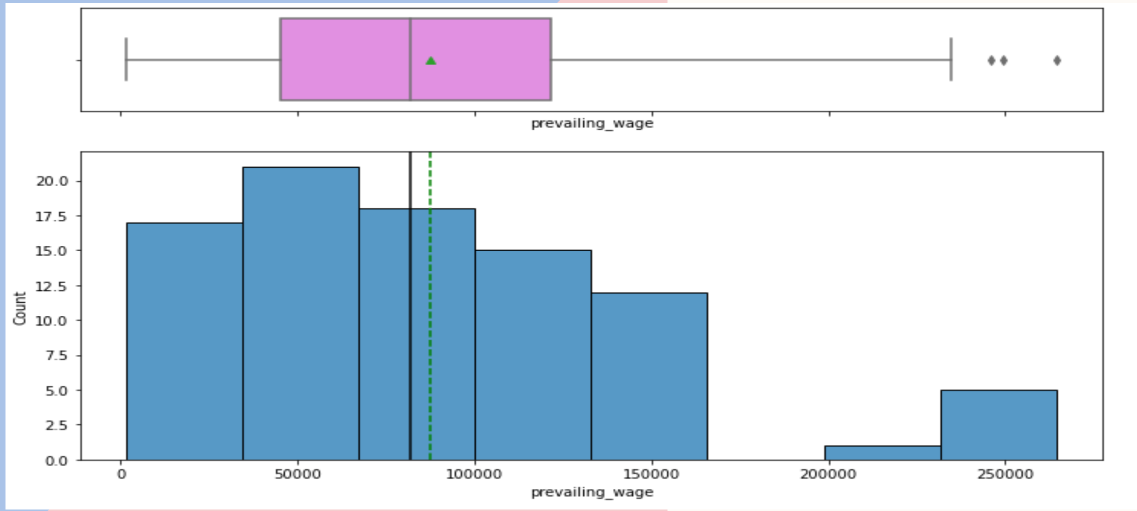
- Datasets consist of 25480 entries which correspond to the #of rows and has 12 columns
- No missing values were seen in the data set
- # of employees, years of establishment & prevailing wages compromise of integer or float. Whereas other variables are object type and need to be converted to the correct datatype
- It is shown to have 9 columns of dtype object, 1 columns of dtype float64, and 2 columns of dtype int64; and target variable for the models is *case_status*
- *case_id* is randomly assigned by INA

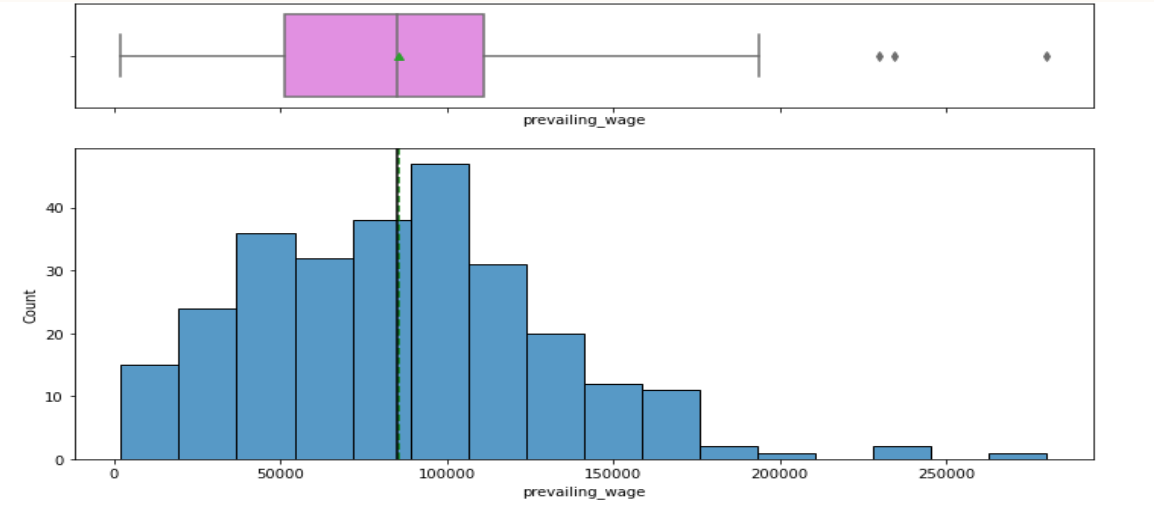
EDA: UNIVARIATE ANALYSIS



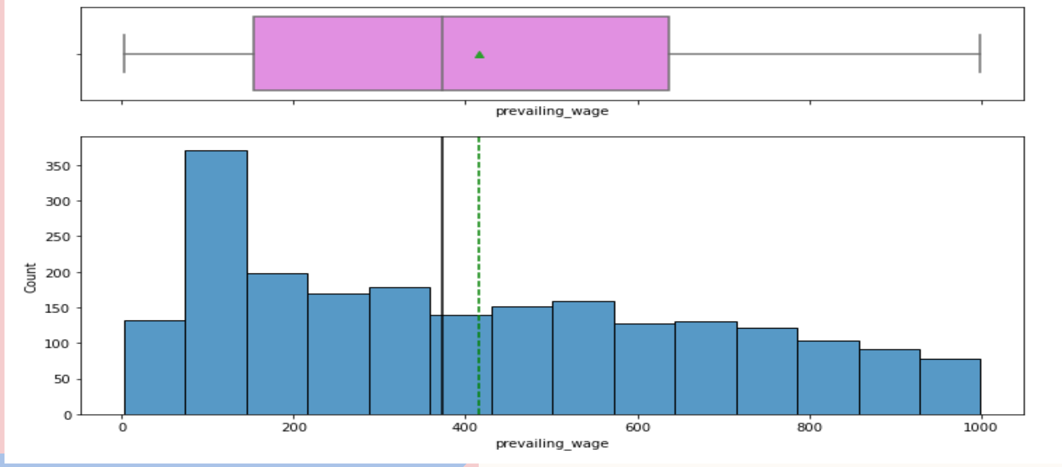


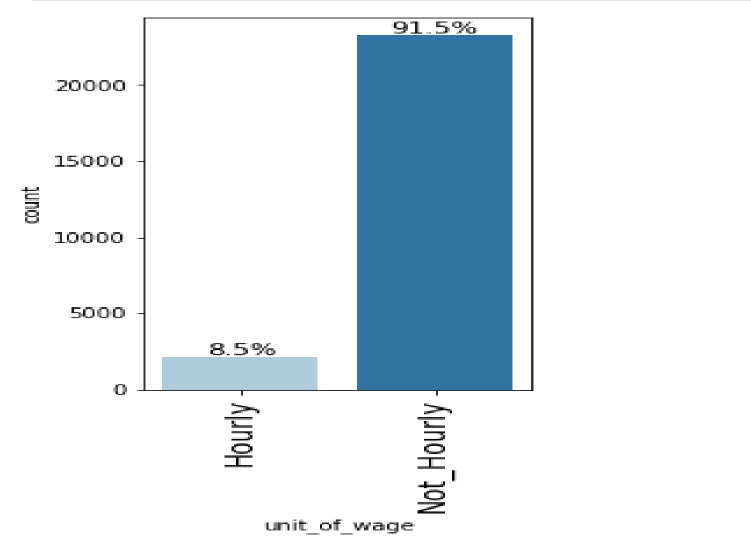


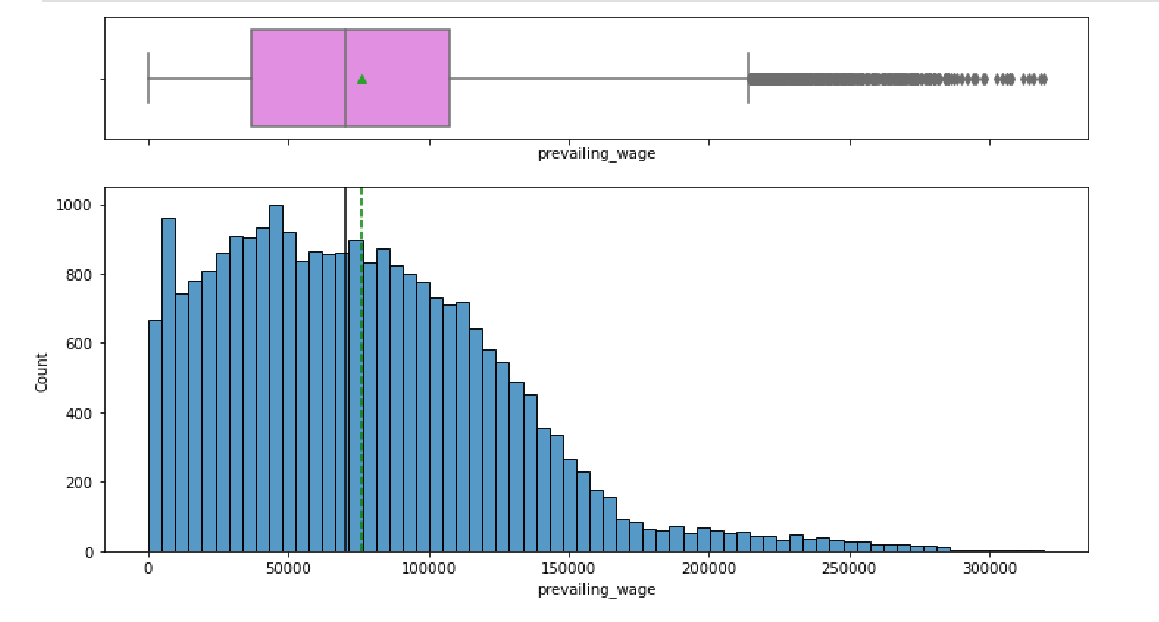




EDA: UNIVARIATE ANALYSIS



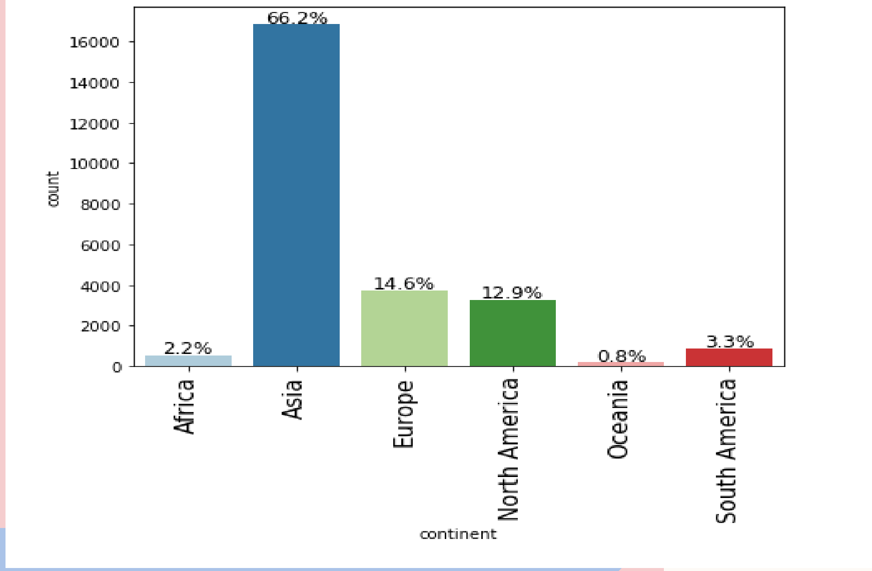


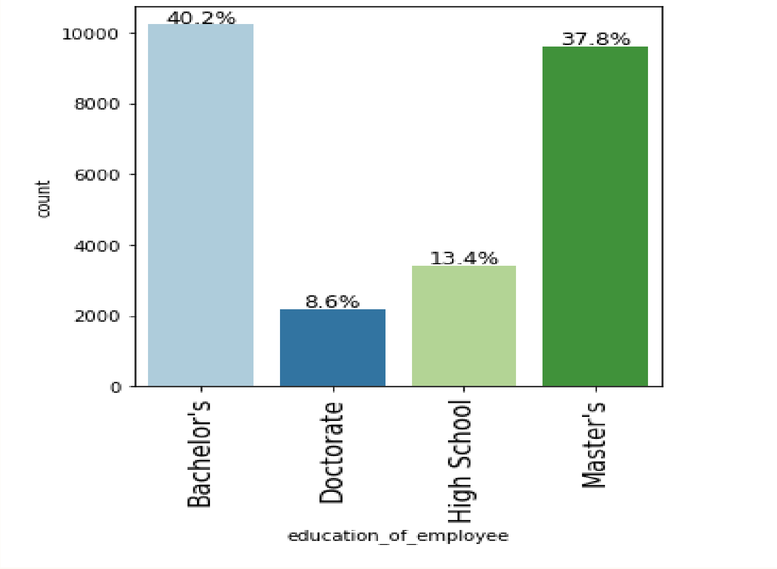


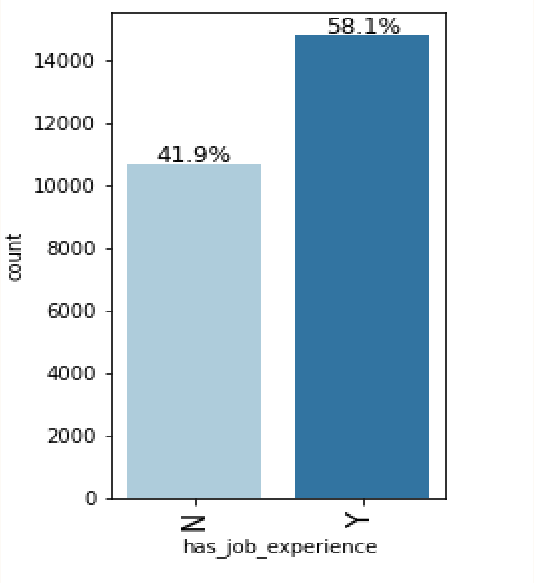
OBSERVATIONS

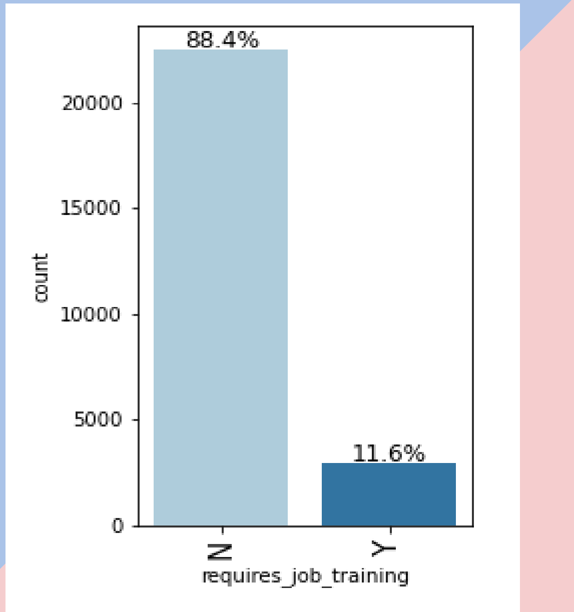
- The distribution for # of employees is right skewed whereas the on the year established is left skewed
- ~90% of entries seen within unit of wage is yearly and ~8.5% is hourly
- Avg. and median annual salary ~$70,000
- The unit wages is shown to be Not-Hourly when employees is paid fixed salary and Hourly when employee is paid on number of hours worked
- It is shown that there are several outliers within annual prevailing wages in which further analysis is needed

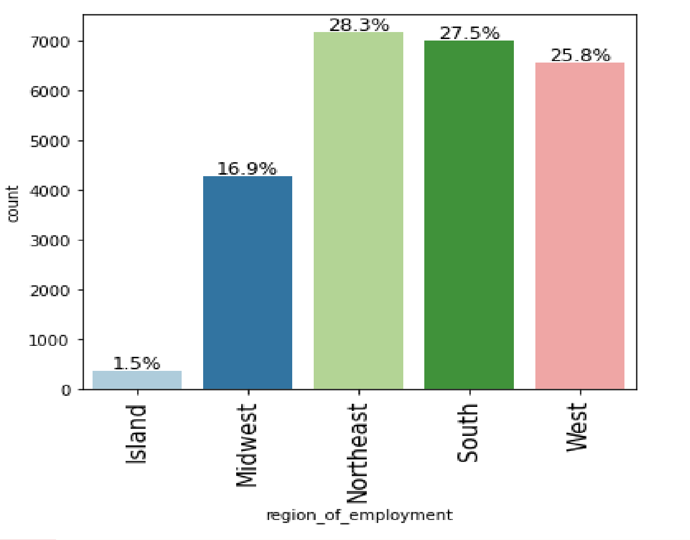
EDA: UNIVARIATE ANALYSIS

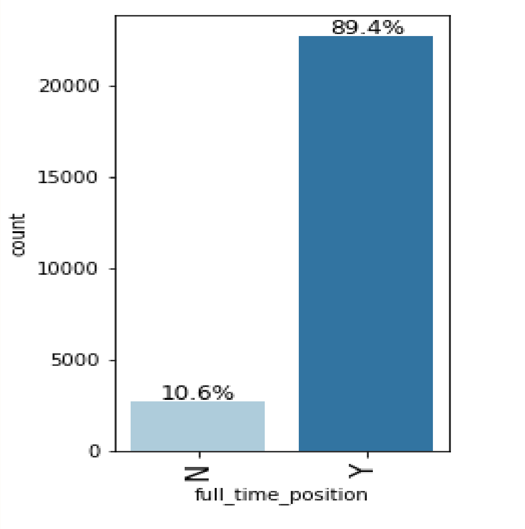


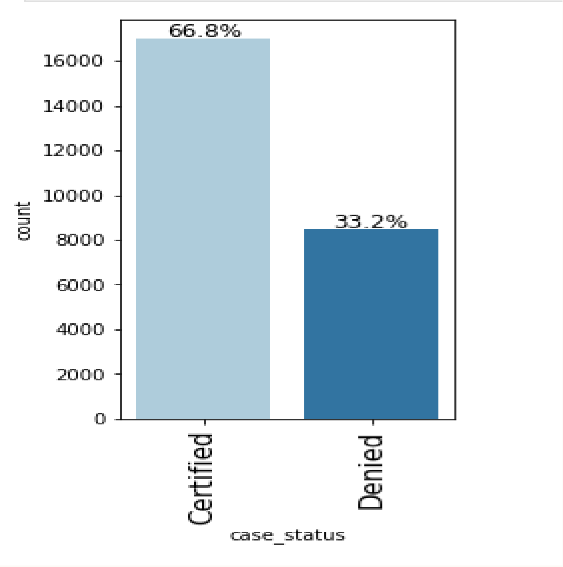








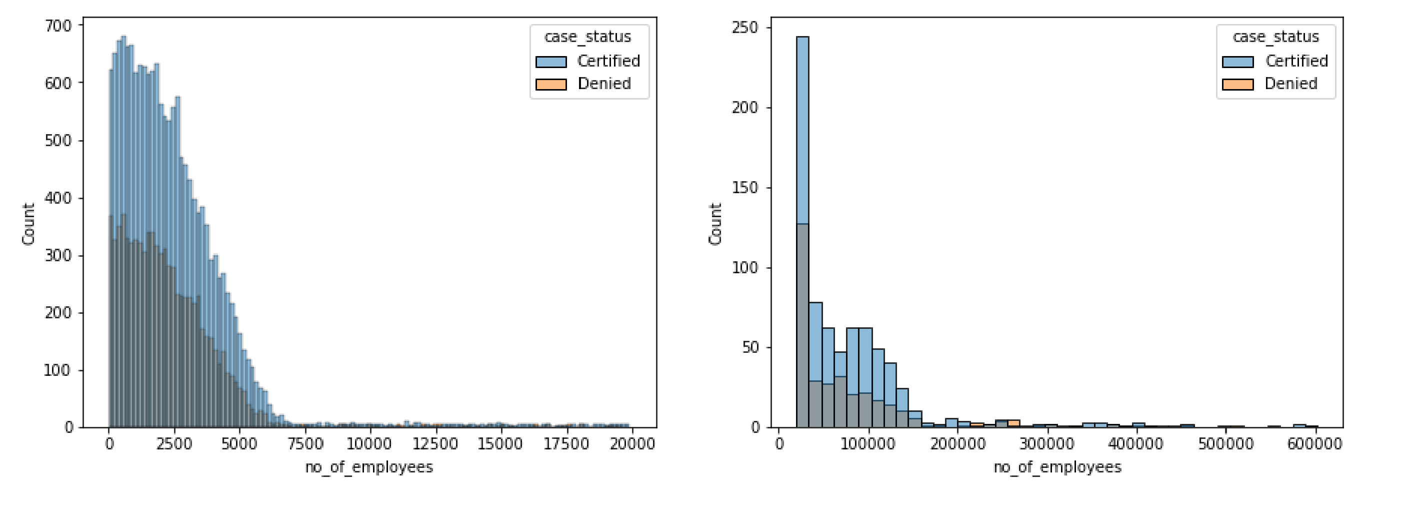


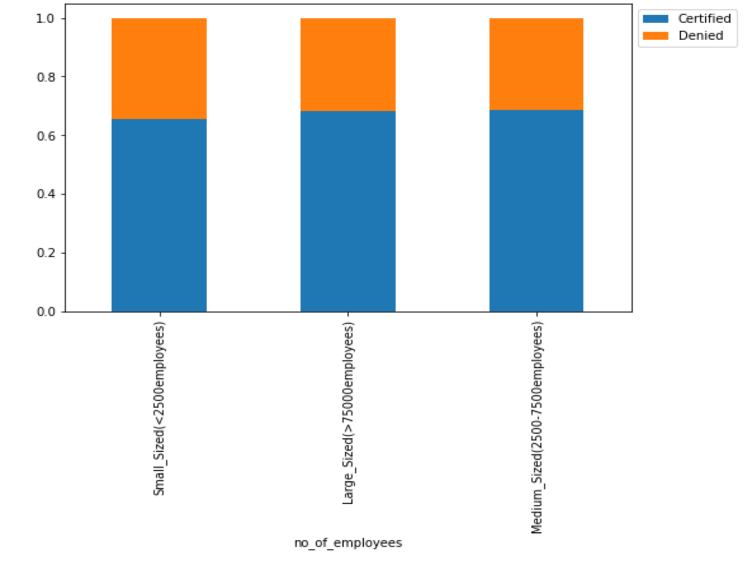


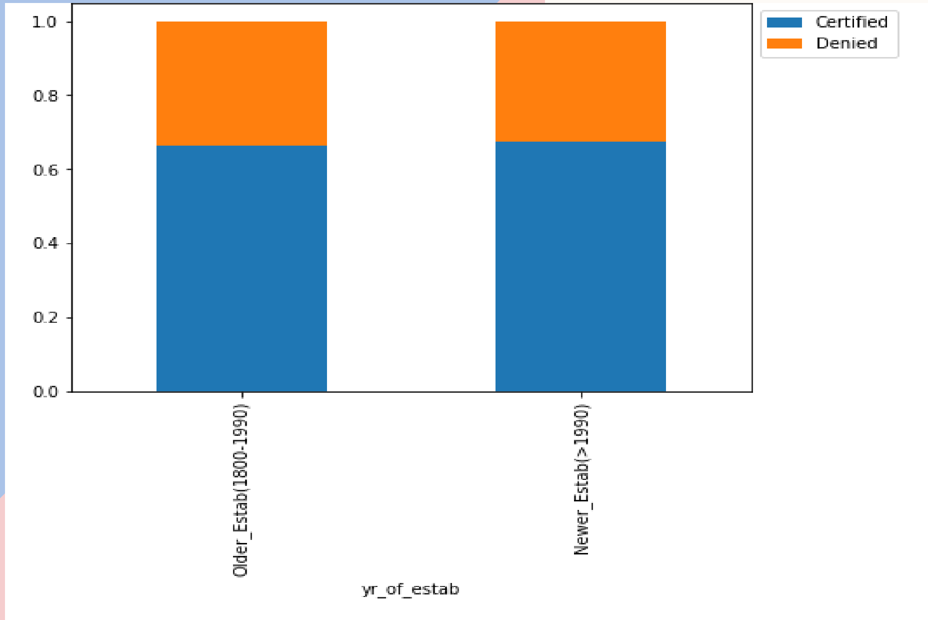
OBSERVATIONS

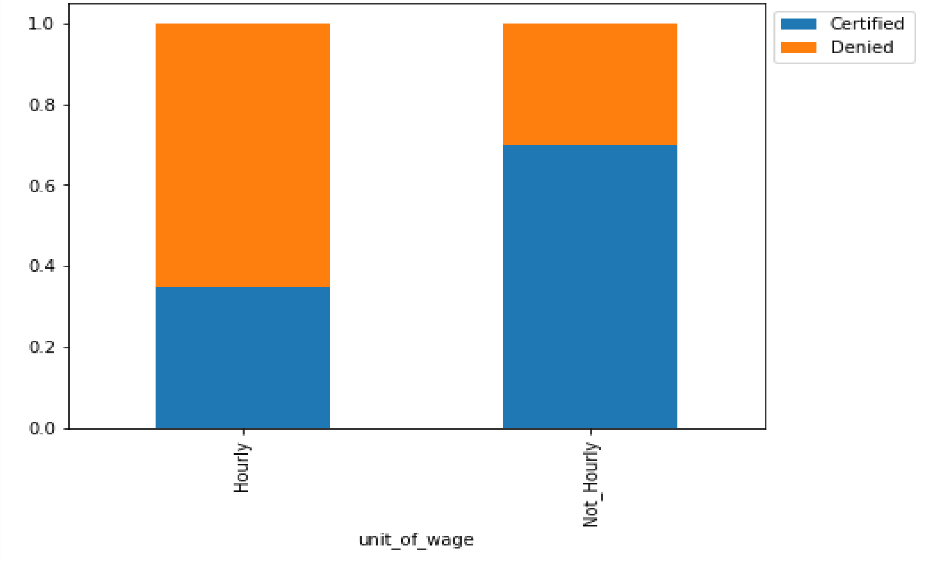
- It is shown that majority of employees are from Asia (>50%); and majority of the employees comprise of either bachelors (40%) or masters (38%) whereas the minority consist of either doctorate (8%) or high school diploma (13%)
- More employees are shown to have prior experience compared to those who do not
- Also, majority of the employees do not require job training
- It is shown that Northeast, South, and West have equally employment opportunities with Human Resource applying for visa approval then followed by Midwest and Island
- However, 88% are full time positions; within case status, 67% cases are approved, and 33% cases are denied

EDA: BIVARIATE ANALYSIS

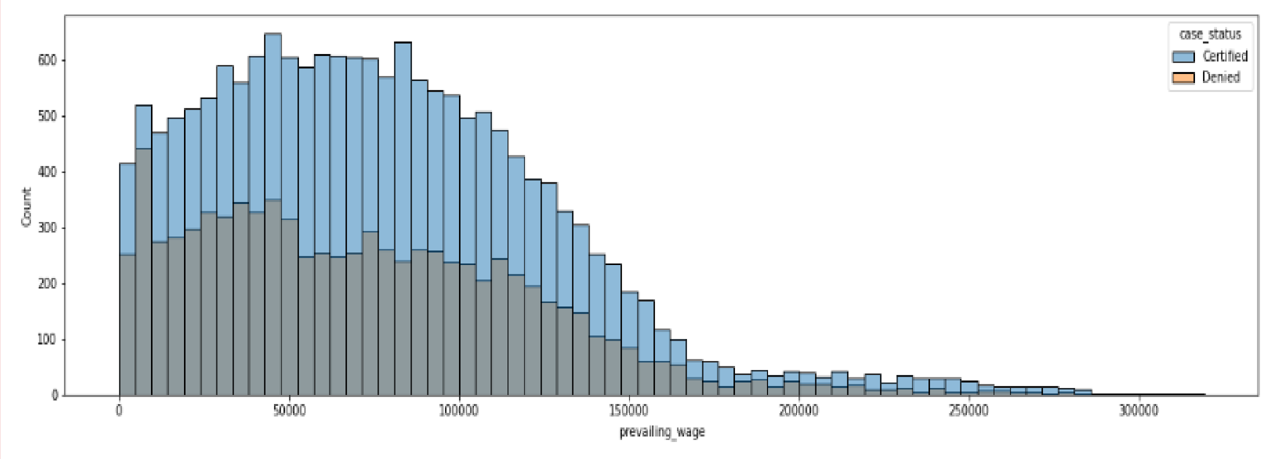


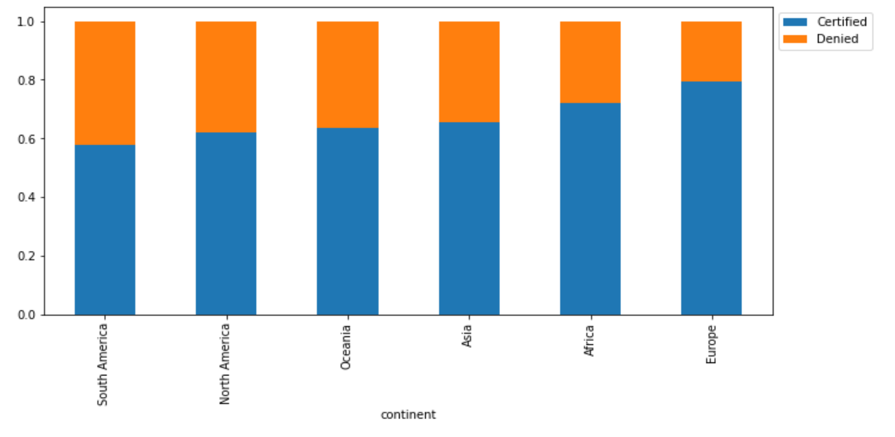


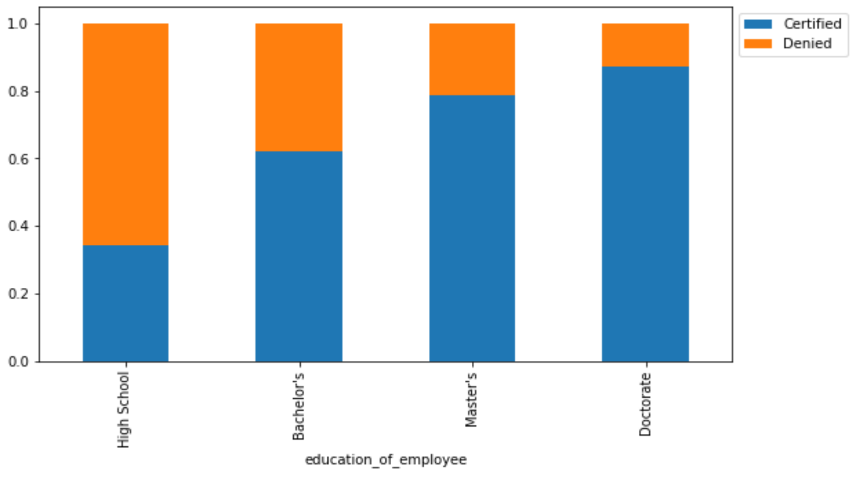


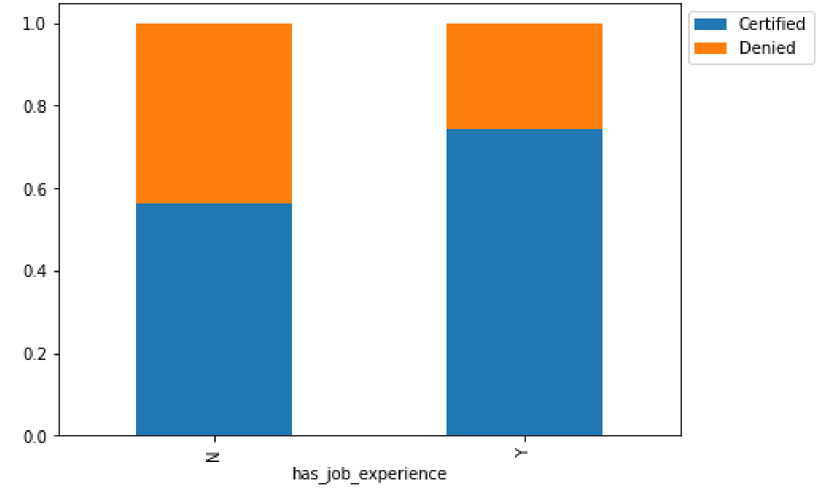


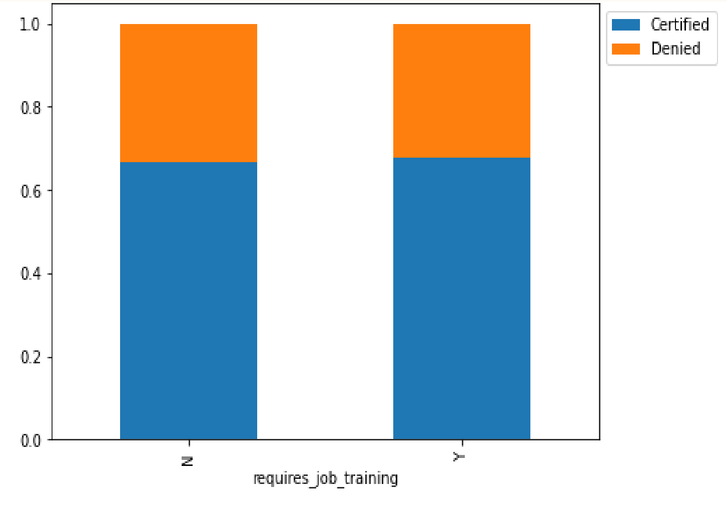
EDA: BIVARIATE ANALYSIS



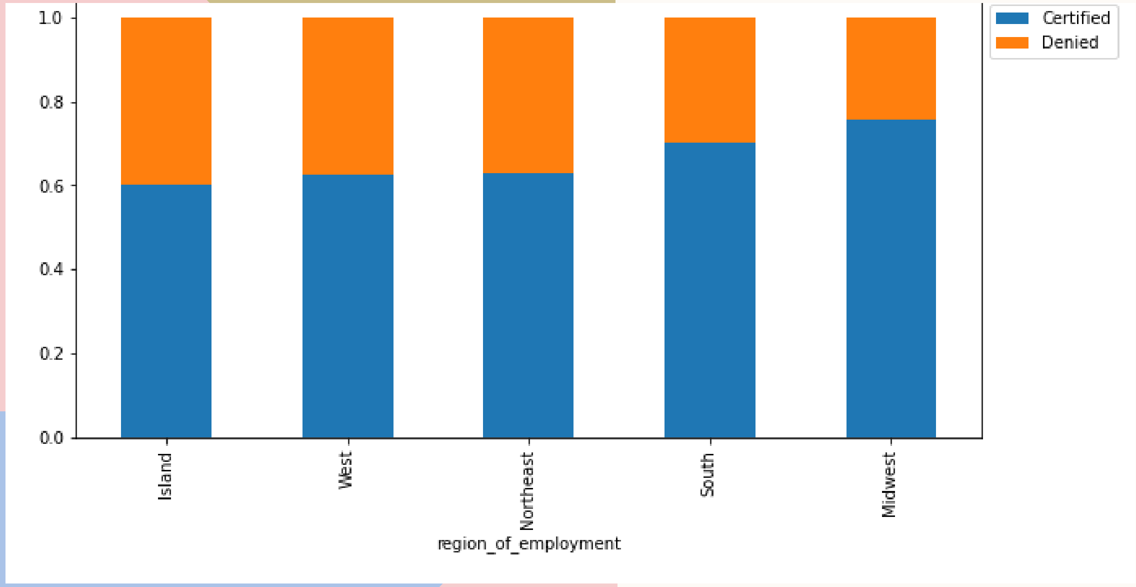


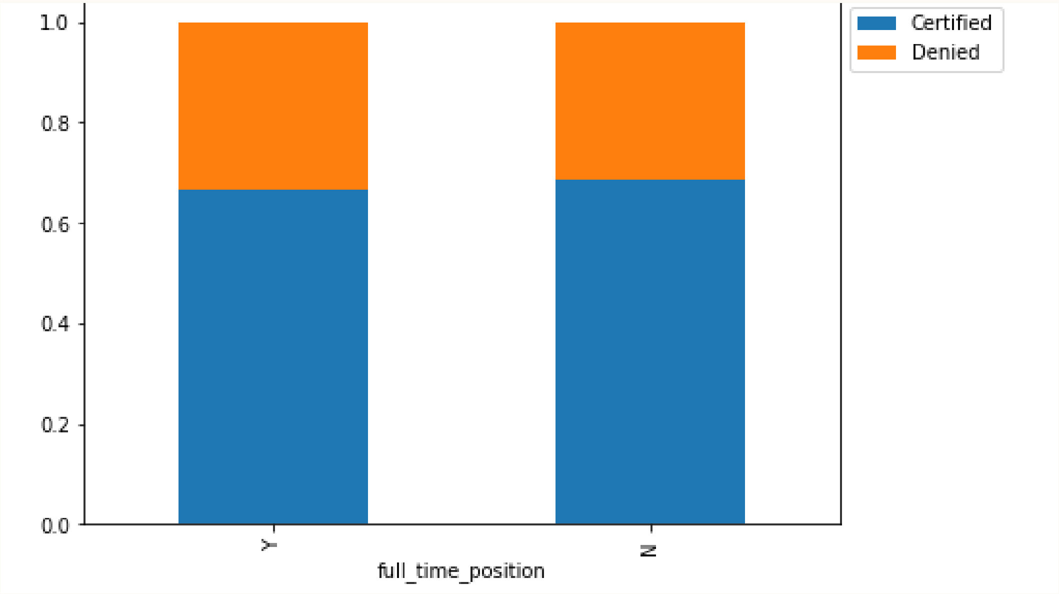






EDA: BIVARIATE ANALYSIS

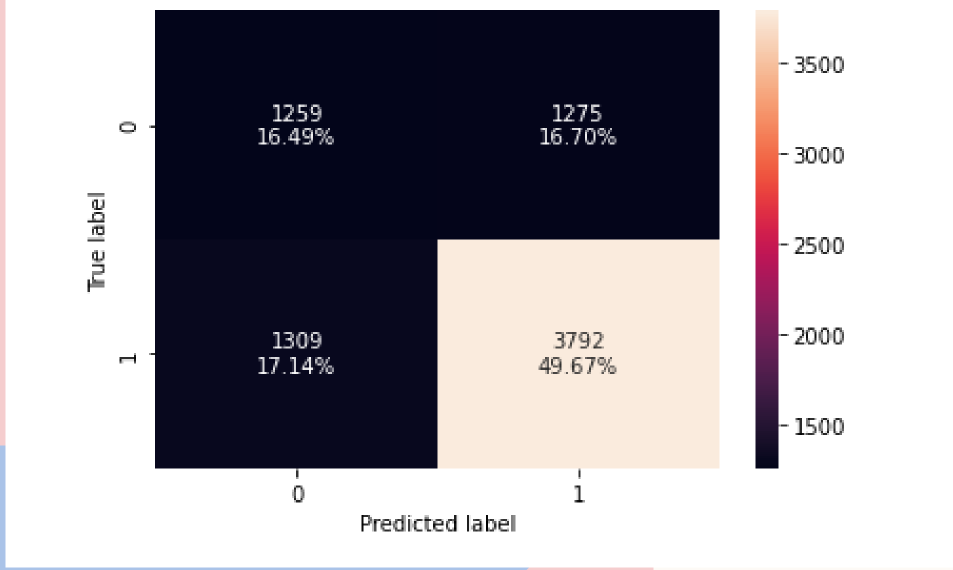


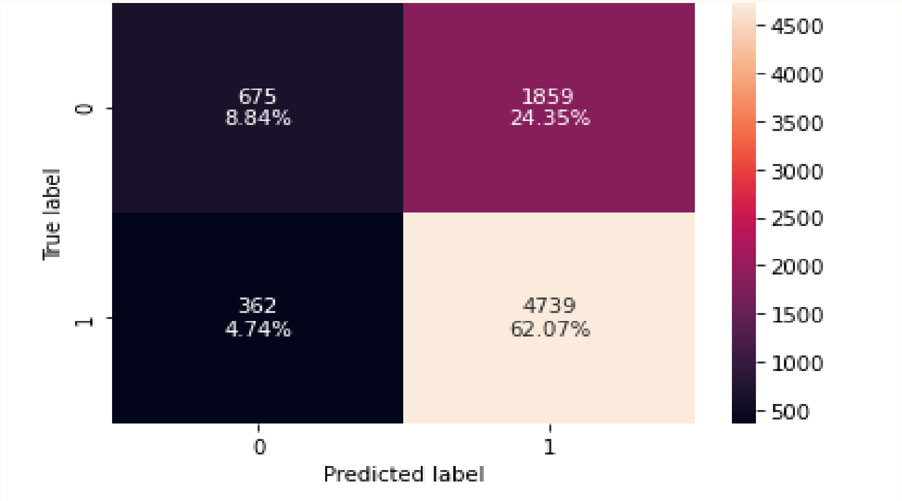


OBSERVATIONS

- It is shown that ># of cases are certified than denied both for employers with more or lesser #of employees
- ~70% of cases are certified when unit of wage is not hourly, and 35% cases certified when unit of wage is hourly
- More cases are seen to be certified than denied irrespective of the continent the employee is from
- On the plot, no_of_employees is right skewed whereas yr_of_estab is left skewed
- ~ 58% of all cases were for smaller organizations and 61% of all cases were for employer's established after 1990

DECISION TREE





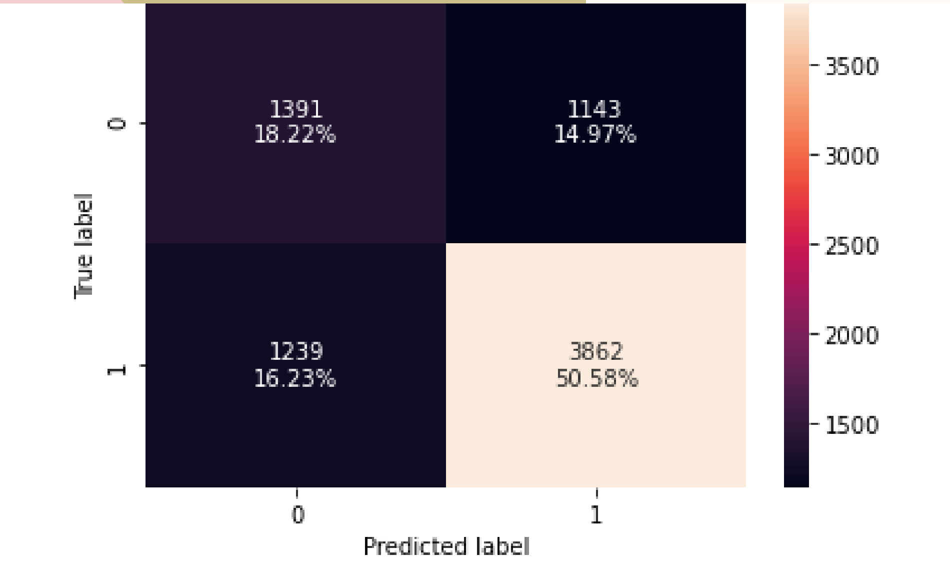
CLASSIFIER

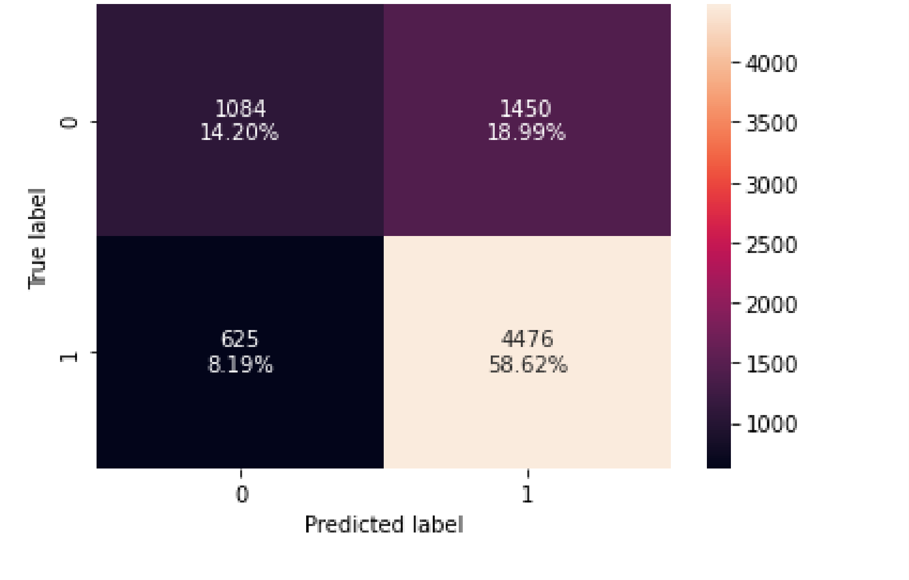
The tree is overfitting the training data. Training metrics are high but testing metrics are not…F1 score is 0.75 and model can be improved by hyperparameter tuning

HYPERPARAMETER TUNING

The tree is not overfitting the dataset and improvement is seen w/n F1 score. F1 score of both train and test datasets are 0.812 & 0.810 respectively

BAGGING





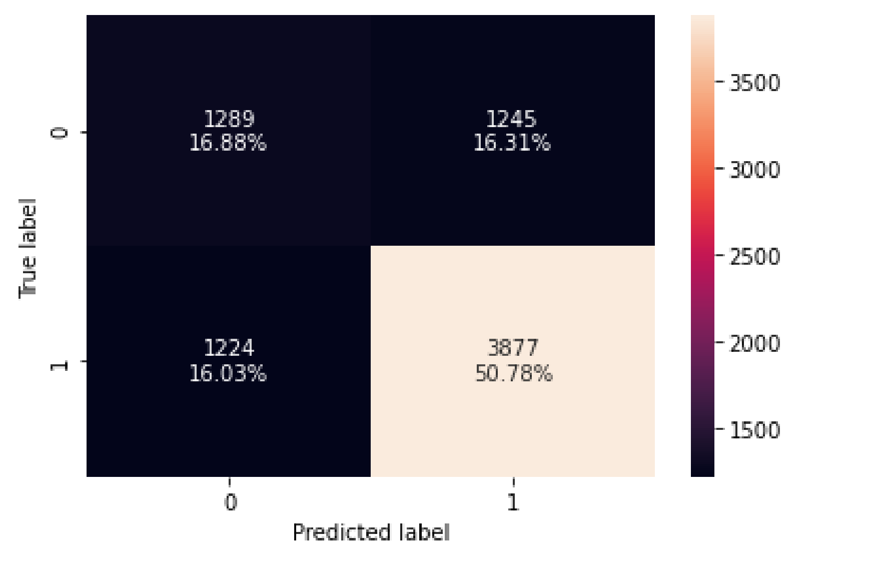
CLASSIFIER

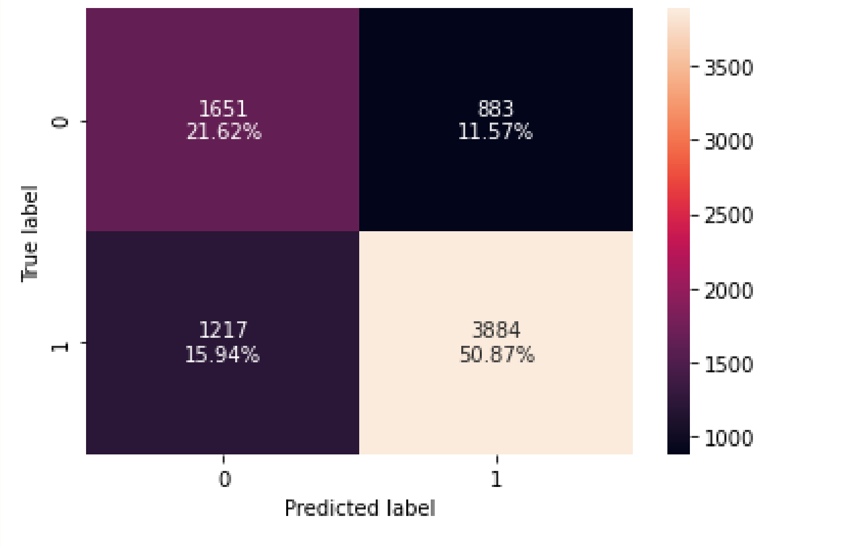
The training data is seen to be overfitting

HYPERPARAMETER TUNING

Model seen to overfit training data…it is shown that training metrics are high but testing metrics are not….

RANDOM FOREST





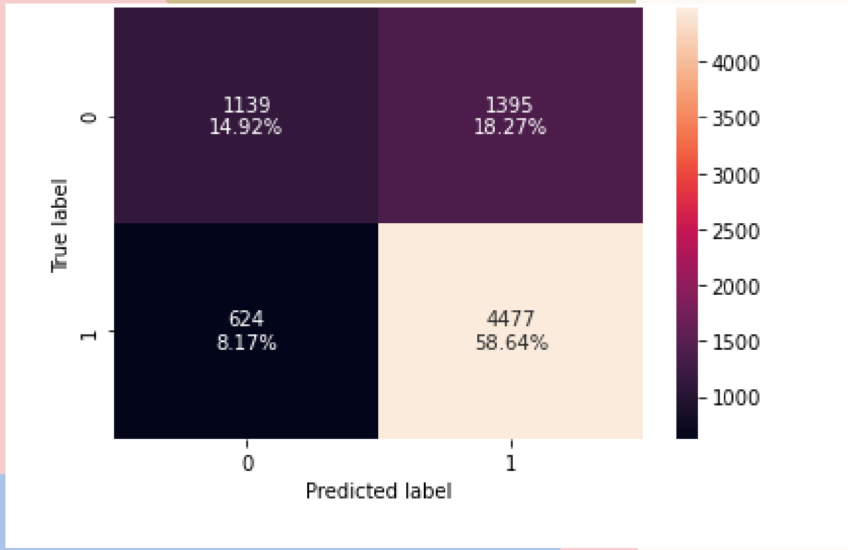
CLASSIFIER

Model is shown to overfit the training data

HYPERPARAMETER TUNING

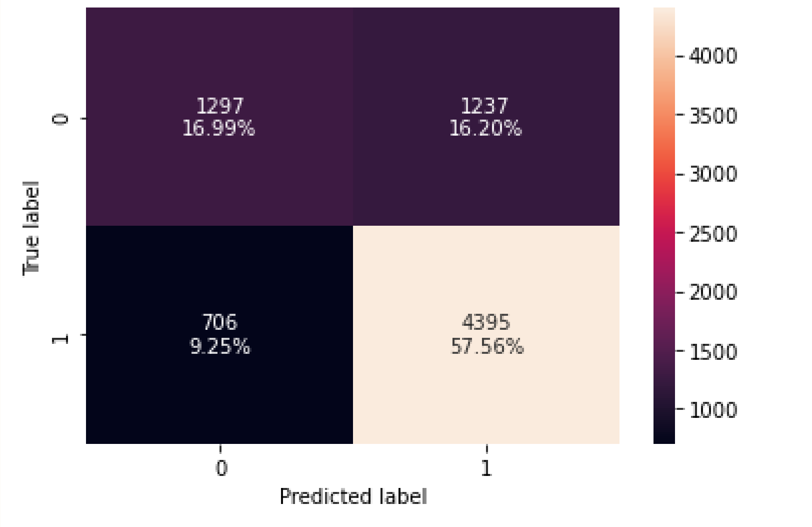
The model reduced the overfit and increased F1 score. Thus, the model was not performed optimally as hyperparameter tuned decision tree

BOOSTING



ADABOOST CLASSIFIER

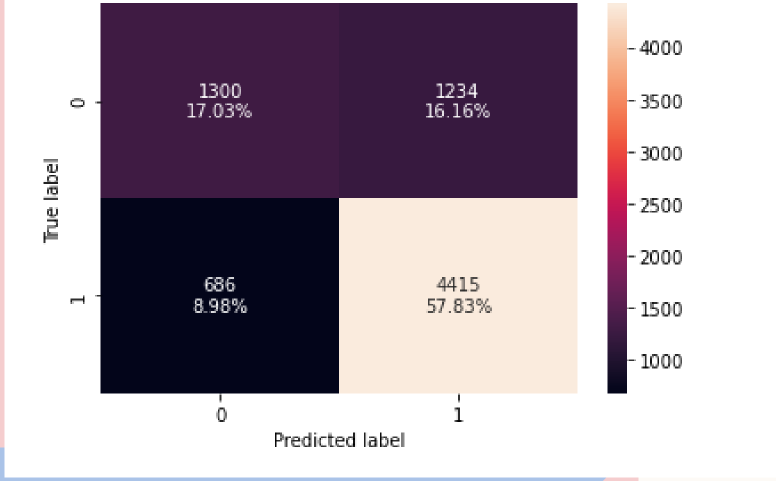
Model not found to overfit the training data. F1 score for training & testing data is 0.819 & 0.816

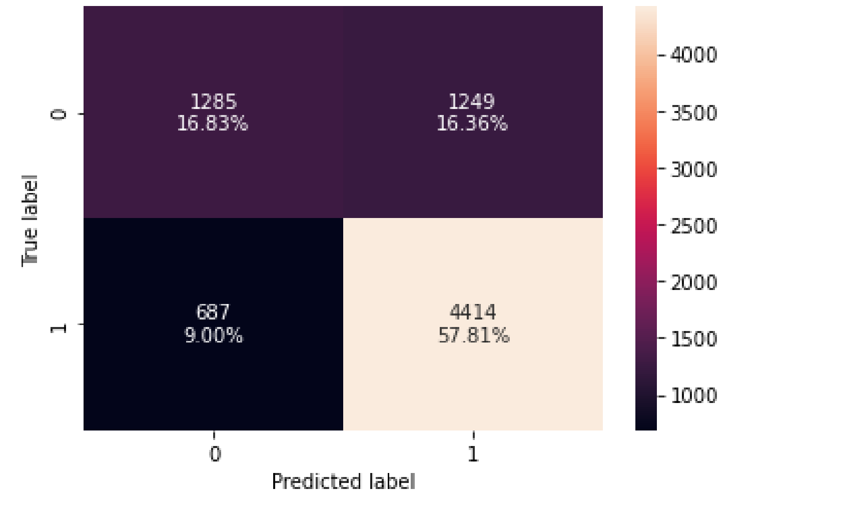


ADABOOST – HYPERPARAMETER TUNING

The model shows similar representation with AdaBoost model

BOOSTING





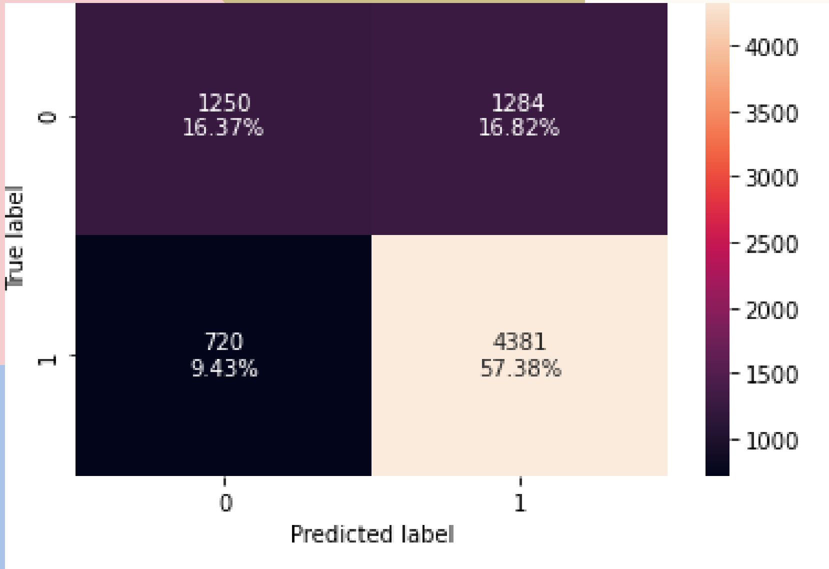
GRADIENT BOOSTING CLASSIFIER

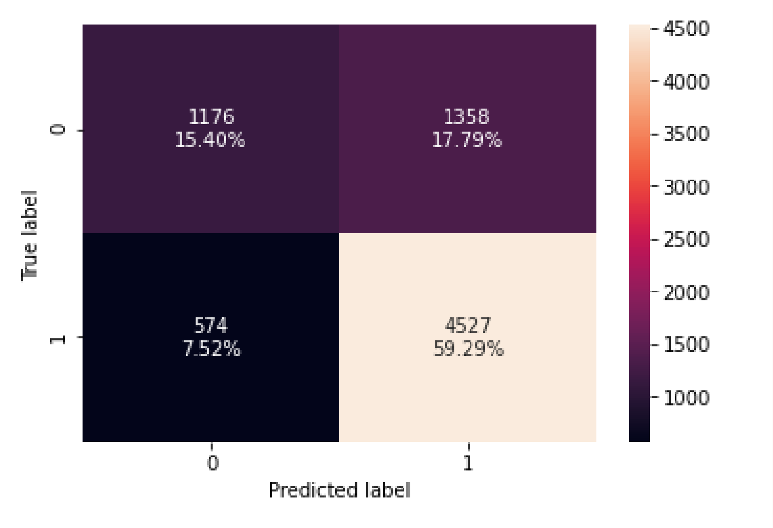
Model depicts high F1 scores on both training and testing data of 0.827 and 0.821 respectively

GRADIENT BOOSTING – HYPERPARAMETER TUNING

Not much of a difference seen in the model performance

BOOSTING





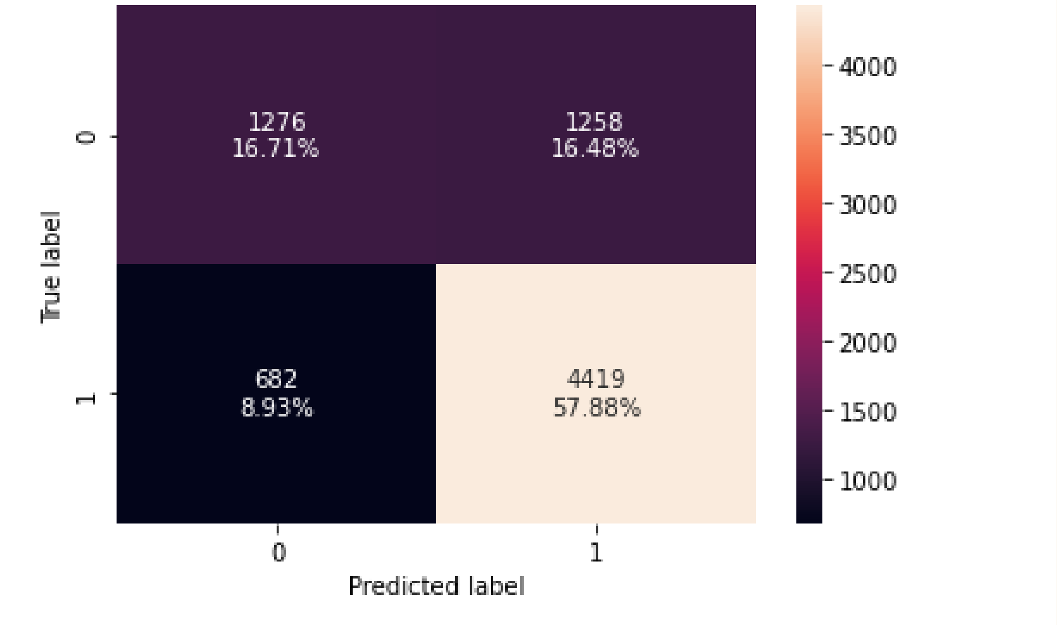
XGBOOST CLASSIFIER

The model is slightly overfitting the training data

XGBOOST – HYPERPARAMETER TUNING

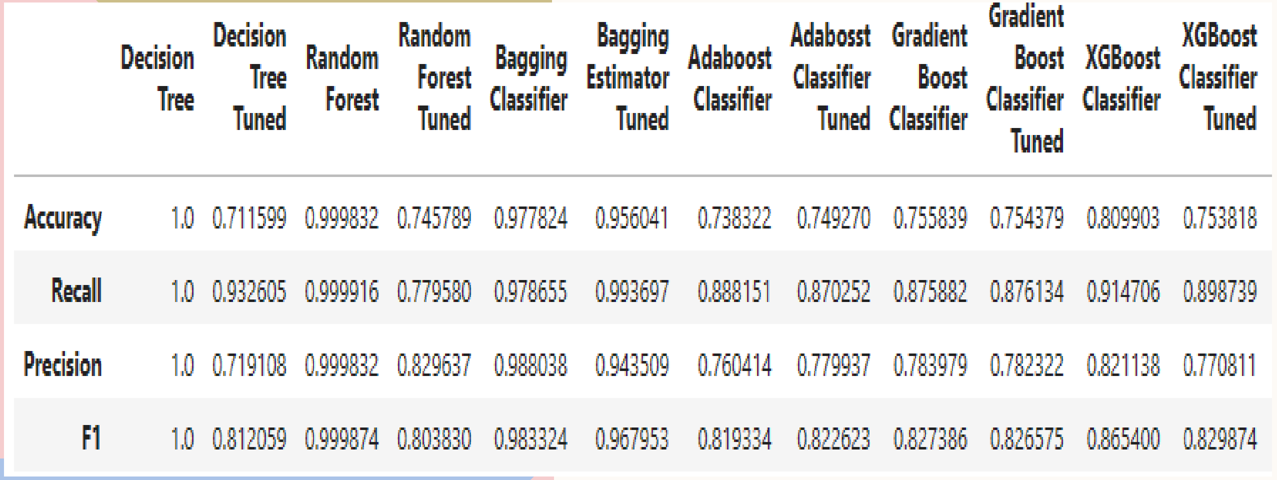
Overfitting of the model is reduced and F1 score for training & testing data are 0.829 and 0.8214 respectively which is high

STACKING CLASSIFIER

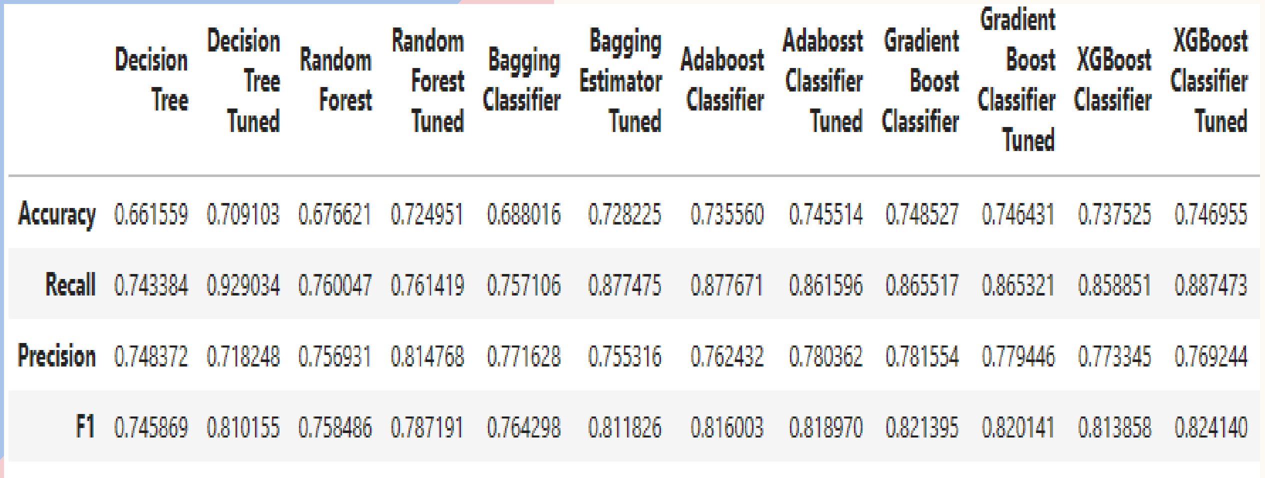


The model is not overfitting and yields generalized performance with training and testing F1 scores 0.826 & 0.820

ALL MODELS COMPARED



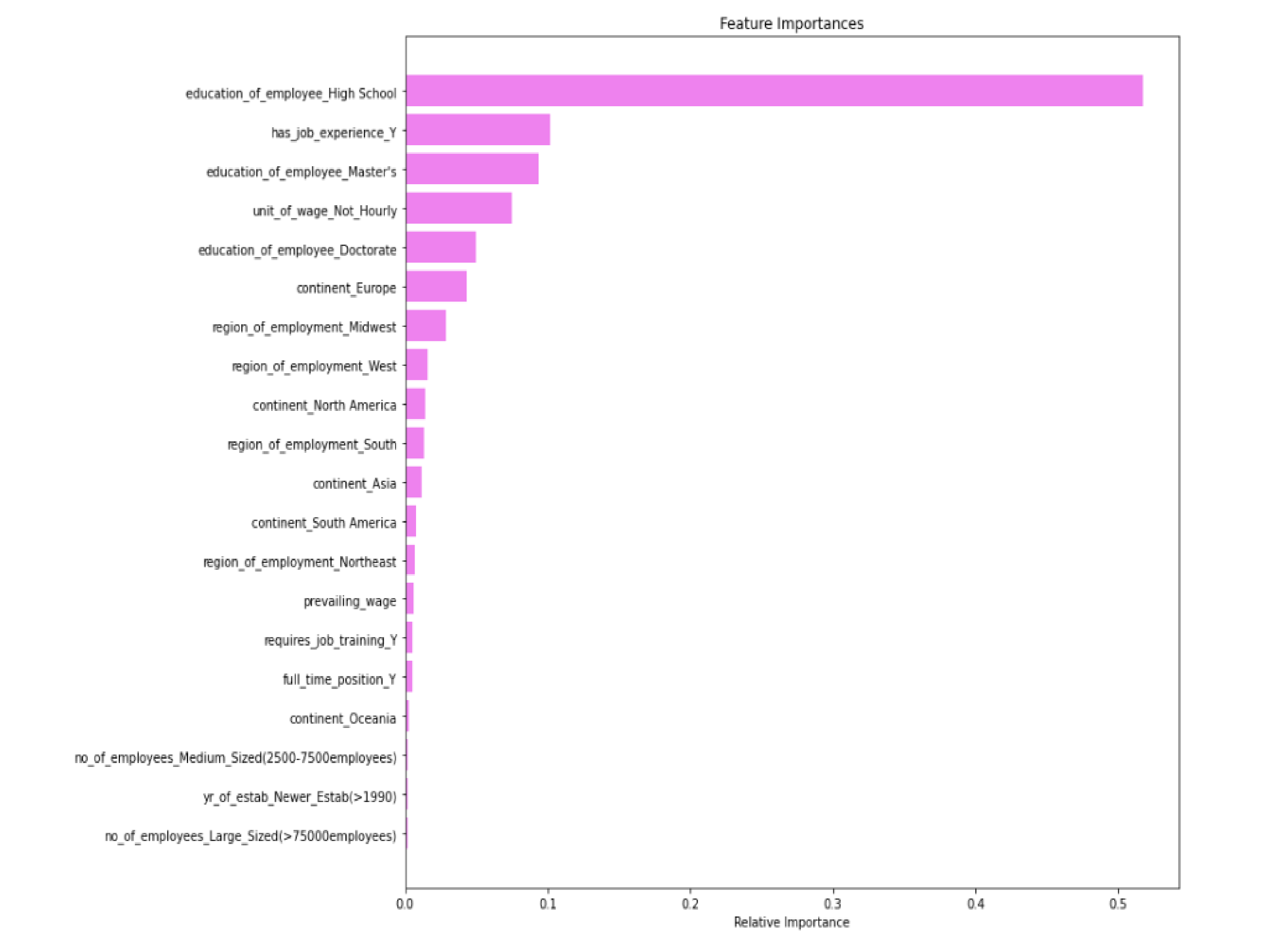
	Decision Tree	Decision Tree Tuned	Random Forest	Random Forest Tuned	Bagging Classifier	Bagging Estimator Tuned	Adaboost Classifier	Adabosst Classifier Tuned	Gradient Boost Classifier	Gradient Boost Classifier Tuned	XGBoost Classifier	XGBoost Classifier Tuned
Accuracy	1.0	0.711599	0.999832	0.745789	0.977824	0.956041	0.738322	0.749270	0.755839	0.754379	0.809903	0.753818
Recall	1.0	0.932605	0.999916	0.779580	0.978655	0.993697	0.888151	0.870252	0.875882	0.876134	0.914706	0.898739
Precision	1.0	0.719108	0.999832	0.829637	0.988038	0.943509	0.760414	0.779937	0.783979	0.782322	0.821138	0.770811
F1	1.0	0.812059	0.999874	0.803830	0.983324	0.967953	0.819334	0.822623	0.827386	0.826575	0.865400	0.829874



	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
Accuracy	0.661559	0.709103	0.676621	0.724951	0.688016	0.728225	0.735560	0.745514	0.748527	0.746431	0.737525	0.746955
Recall	0.743384	0.929034	0.760047	0.761419	0.757106	0.877475	0.877671	0.861596	0.865517	0.865321	0.858851	0.887473
Precision	0.748372	0.718248	0.756931	0.814768	0.771628	0.755316	0.762432	0.780362	0.781554	0.779446	0.773345	0.769244
F1	0.745869	0.810155	0.758486	0.787191	0.764298	0.811826	0.816003	0.818970	0.821395	0.820141	0.813858	0.824140

Decision tree, Random Forest (default & tuned), and Bagging classifier (default & tuned) were shown to overfit the training dataset

Decision tree tuned, Adaboost (default & tuned), Gradient boost (default & tuned) and XGBoost (tuned) were shown to generalized performance on training & testing data sets. Thus, XGBoost (tuned) consist of highest F1 score



Education of the employee was shown to be the most factor having effect on visa certifications

CONCLUSION

- Based on the analysis, it is concluded that:
  - Education of employee --> high school certification employee displays over 65% chance of visa denial when compared to employee with doctorate with 85% of visa certification
  - Unit Wage --> hourlyb pay employee consist of 65% chance of visa denial when compared to non-hourly employee with >70% chance of visa cerification
  - Employee Region --> based on the continent, it is shown that employee with prior work experience has 75% chance of visa approval than those without prior work experience with 50% chance of visa denial. Thus, employees who reside within Europe are shown to have >80% chance of visa certification.
  - It is shown that U.S region if Midwest or South has >70% chance of getting visa certification

RECOMMENDATIONS

- Factors such as 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 were not important factors and do not possess much weigh on case determination whether certified vs denied
- The model was able to display 80% of the information while verifying predictions.
    - %certifications correctly verified is high as per test confusion matris (>88%) while % denied is on lower end (~54%)
        - This demonstrates the importance of re-evaluation of cases being denied which will lead to saving 60% of processing time

THANK YOU