Easy Visa (Case Study)

Project 5: Ensemble Techniques

Date: January 11th, 2022.

By: Ijeoma Ejem

Contents / Agenda

EXECUTIVE SUMMARY

Executive Summary

	Conclusion	Recommendation
1	We were able to collect helpful insights that influence the case status based on an applicant's profile but there could also be factors in the application process that have previously affected case status due to factors on the employer's side.	We can to collect and analyze more data pertaining to the employer and employee that could help employer's avoid costly mistakes and make the process easier to understand and plan for for both parties.
2	The data shows that the higher the applicant's level of education, the more likely they are to get certified. Doctorate (87%), Master's (79%), Bachelor's (62%), High School Cert (34%). Education has also been identified as the most important feature in the data.	OFLC is about helping US employer's fill jobs while protecting both **U.S. and foreign workers**. There should be more resources and information on the OFLC website for education (study) grants for both classes of applicants. This will help them prepare and increase their chances for success.
3	We found that the Northeast, South and West **regions** have the highest number of hires across all regions. In these regions, they have similar counts of hires based on different levels of **education**.	These **two attributes** should be studied some more to discover insights that will be helpful to both U.S. and foreign applicants. We can explore geographical data that can help applicants better assess different regions based on their industry.

	Conclusion	Recommendation
4	Unit of wage data shows that the least likely applications to get approved are those whose units are hourly, as opposed to, weekly/monthly/annually.	Applicants who are compensated in hourly units should be made aware of the factors that make their application weaker or stronger. Based on data from previous applicants, we can determine these contributing factors.
5	From the EDA, we've identified some attributes that influence the case status, like; highest level of education, job experience, required training, and unit of wages.	Applicants will usually want to know how to increase their odds when applying. OFLC can make graphs and insights available on their website to guide applicants. While, independently some factors only provide a slight advantage, collectively they can make a huge difference for an applicant.
6	We tested several different tuned and untuned classification models that attempted to correctly classify candidates, and review their cases, with an emphasis on minimizing false positives and false negatives.	We recommend using the model with the highest F1 score. The stacking classifier proved to be the best performing model on the test set based on this metric.

PROBLEM AND SOLUTION OVERVIEW

Business Problem and Solution Overview

Problem:

- OFLC (the Client) experienced 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 for the company 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.

Solution:

- Using exploratory data analysis (EDA) to explore significant influences and patterns that determine whether the profile of a candidate/applicant will be certified or denied.
- Building and testing a classification model to help facilitate the process of visa approvals by recommending 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.
- Analyzing results from the data and providing business insights and recommendations to help OFLC accomplish its mission of helping U.S. employers, while protecting employees.

DATA OVERVIEW

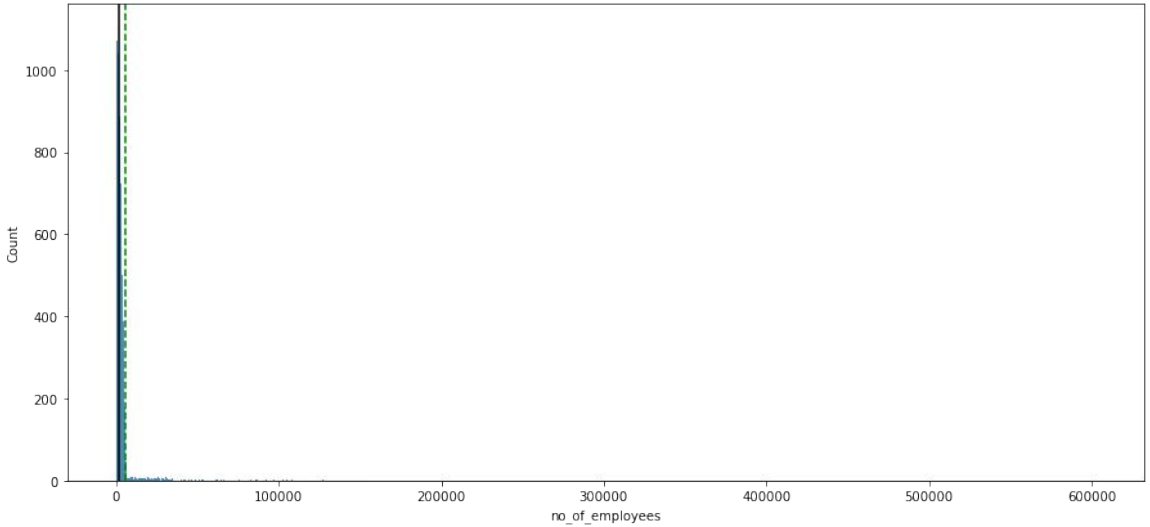
Data Overview

- There are 25,480 rows (observations) and 12 columns (attributes) in total.
- All datatypes are accurately represented; there are 9 objects, 2 integers and 1 float in the dataset. The attributes we are analyzing and building a model around have 9 categorical variables and 3 numerical variables.
- From the statistical summary of the numerical data, the following is deduced;
  - Number of employees appears to have a negative minimum value, which shows that the data has some inaccuracies.
  - The average number of employees in the employer's company is 5,667, while the Median value is 2,109 and maximum is 602,069.
  - The average year of establishment of the employer's company is 1979, while the most recurring (median) year (value) is 1997.
  - Prevailing wage has a minimum value of $2, an mean of $74,455, a median of $70,308 and a maximum value of $319,210.
- There are 0 missing values and duplicate values in the data.
- There are 33 rows in the dataset where the number of employees is less than 0.

EDA: UNIVARIATE ANALYSIS

EDA Results

UNIVARIATE ANALYSIS

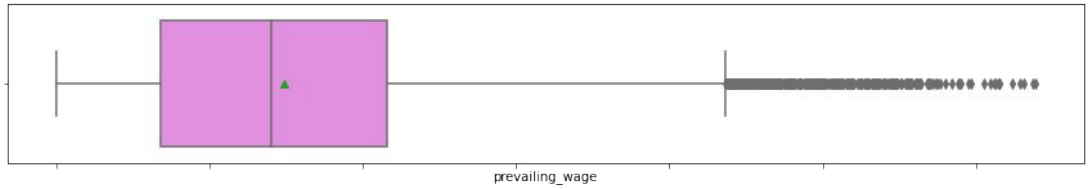


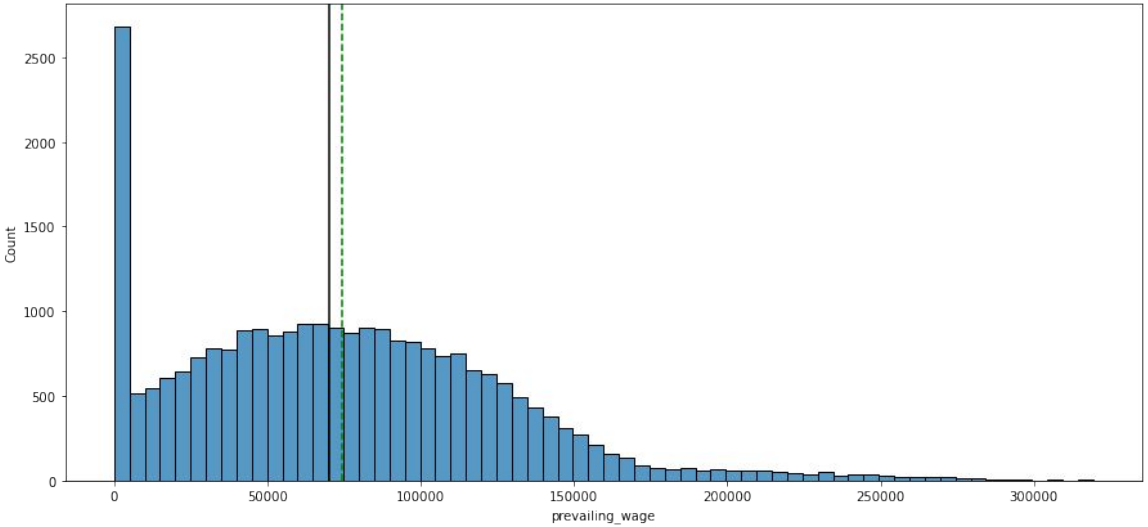
Number of Employees

The histogram confirms what was highlighted in the statistical summary. The average number of employees in the employer's company is just above 5,500 with a maximum value of just over 600,000 employees. There are numerous outliers in the data.

EDA Results

UNIVARIATE ANALYSIS



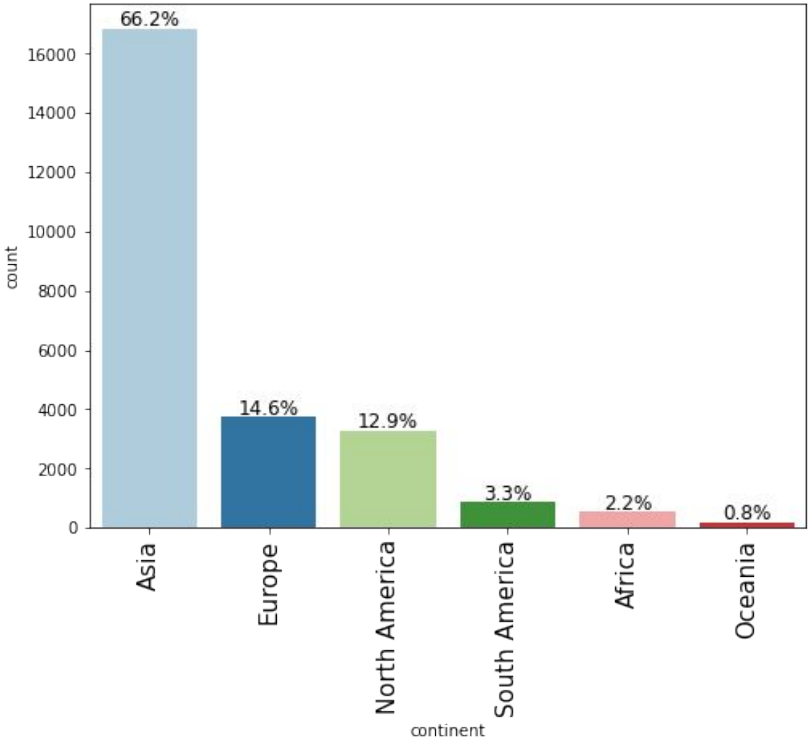


Prevailing Wage

The distribution of prevailing wages is right-skewed. The plot shows that the mean and median values are both within the range of $70,000 - $80,000.
We also checked prevailing wage for observations with values less than $100 and found that all prevailing wages below $100 are in hourly units. This means that similarly employed workers who earned less than $100, in their area of intended employment, did so at an hourly rate.

EDA Results

UNIVARIATE ANALYSIS

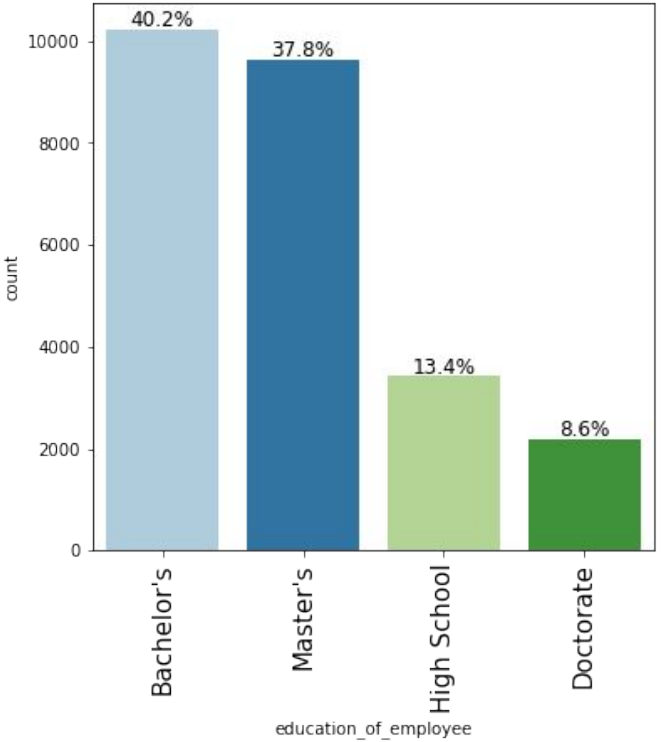


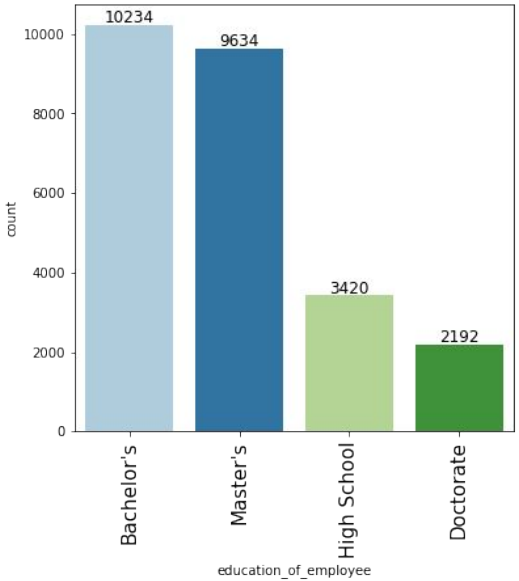
Continent

Asia has by far the highest percentage of applicants (66.2%) of any other continent in the data, followed by Europe (14.6%) and North America (12.9%).

EDA Results

UNIVARIATE ANALYSIS



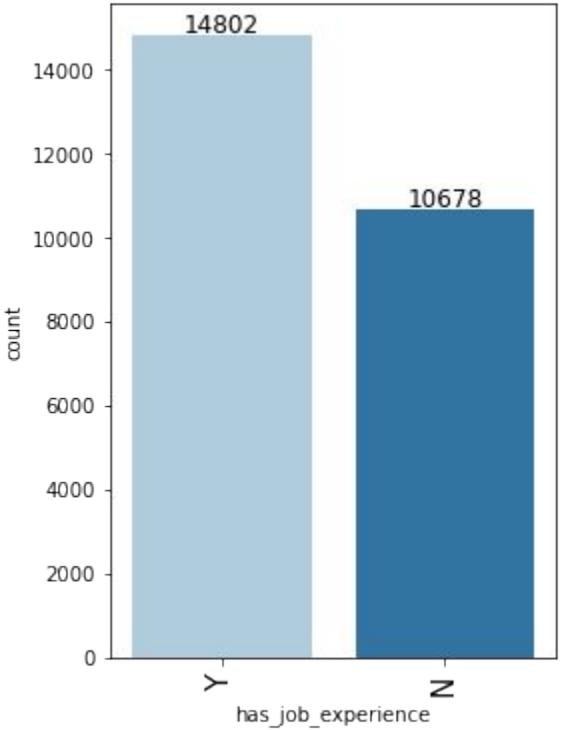


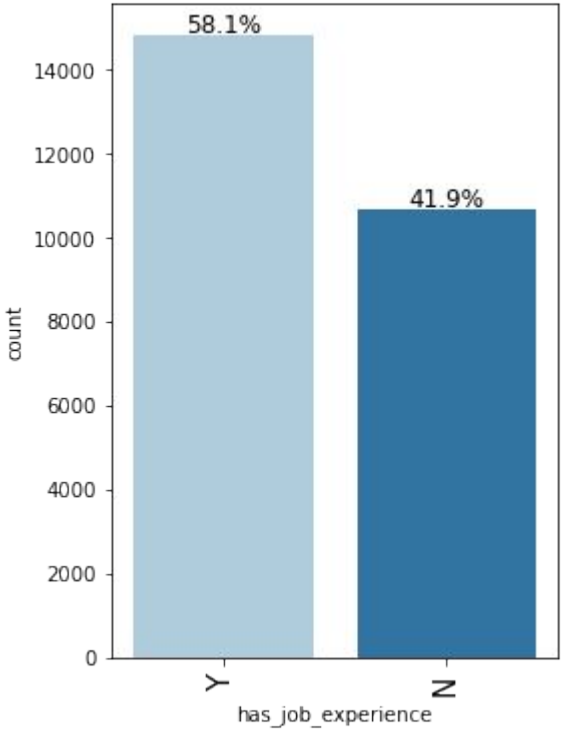
Education of Employee

40.2% (10,234) of applicants hold a Bachelor's as their highest level of education, while 37.8% (9,634) hold a Master's. Doctorate earners make up 8.6% (2,192) of the applicants in the data and the remaining 13.4% (3,420) are applicants with high school level education only.

EDA Results

UNIVARIATE ANALYSIS



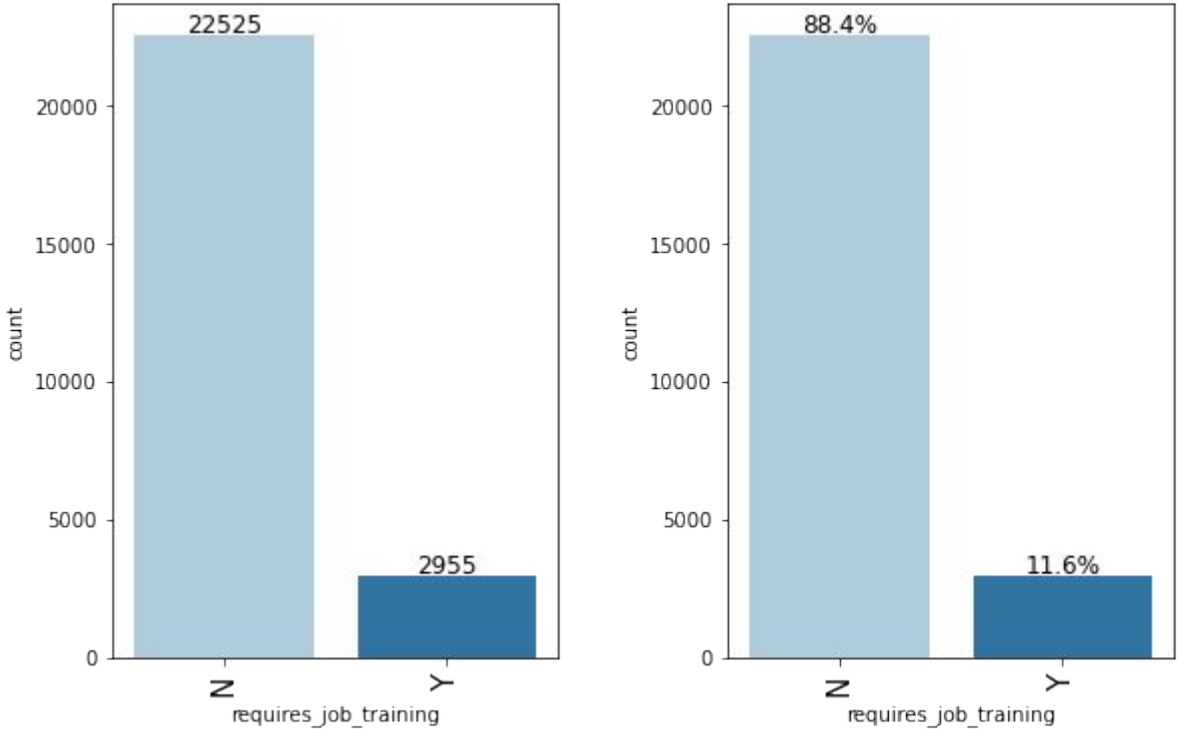


Job Experience

There are more applicants with job experience than without. 58.1% (14,802) of applicants answered yes to having job experience while 41.9% (10,678) answered no.

EDA Results

UNIVARIATE ANALYSIS

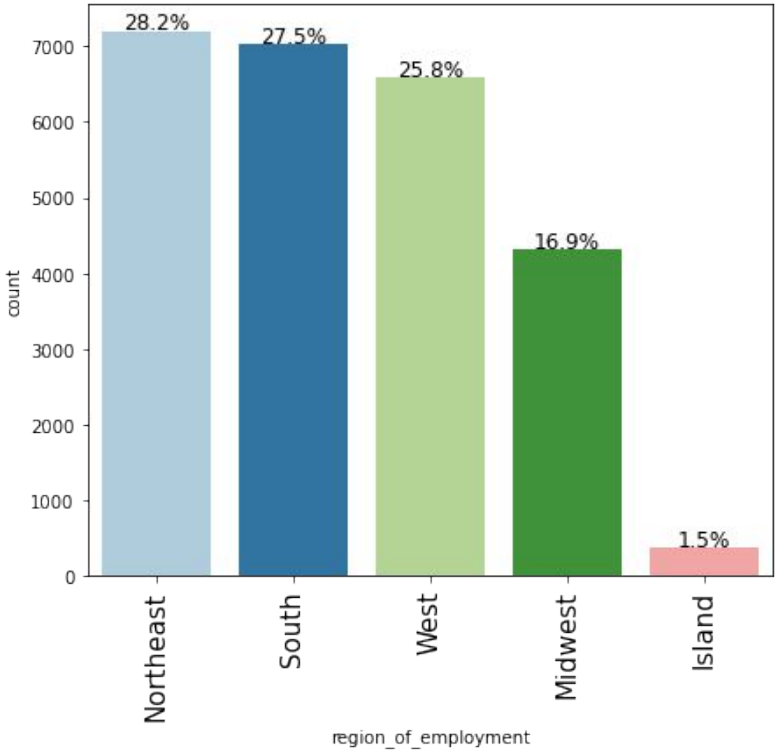


Job Training

88.4% (22,525) applicants will not require job training, while 11.6% (2,955) will need to be trained.

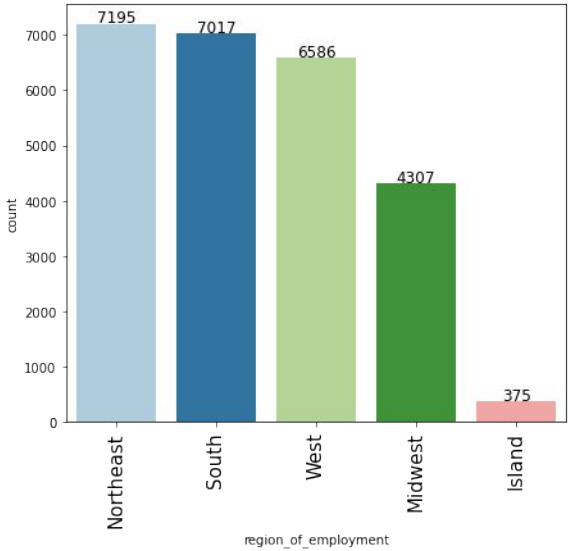
EDA Results

UNIVARIATE ANALYSIS



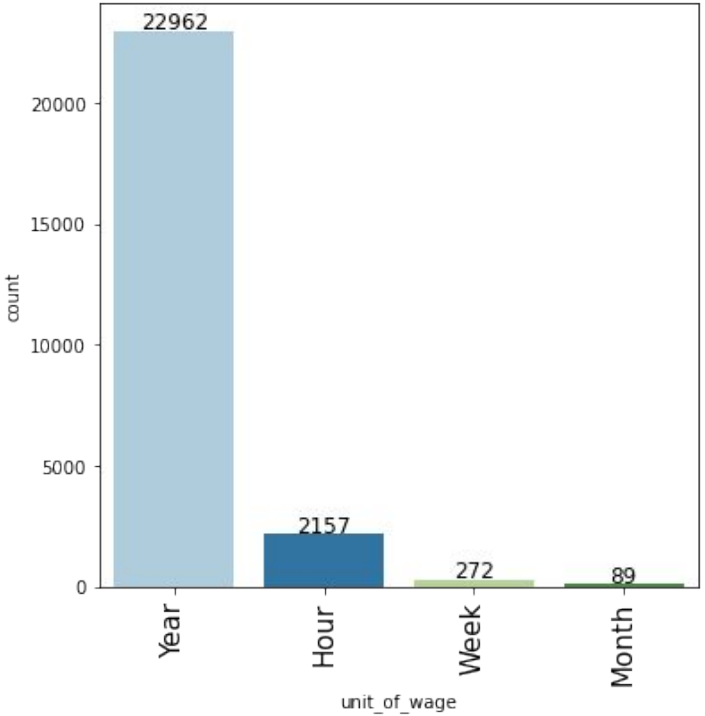
Region of Employment

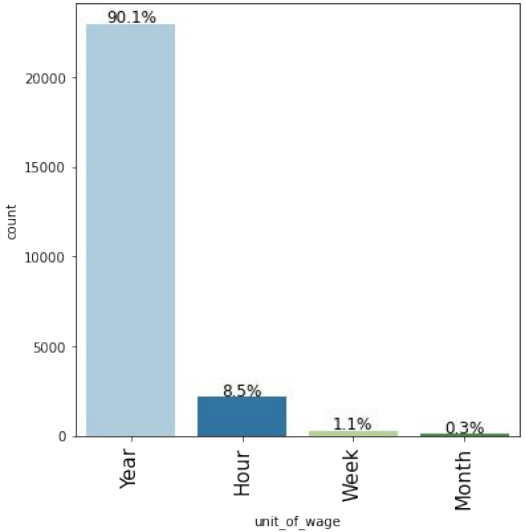
The highest regions of employment for applicants in the data are Northeast, South and West. The mentioned regions collectively (and almost uniformly) make up 81.5% of applicants.



EDA Results

UNIVARIATE ANALYSIS



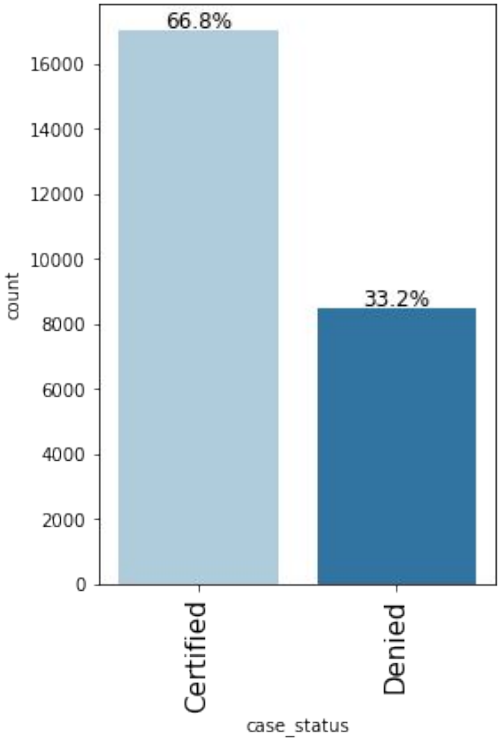


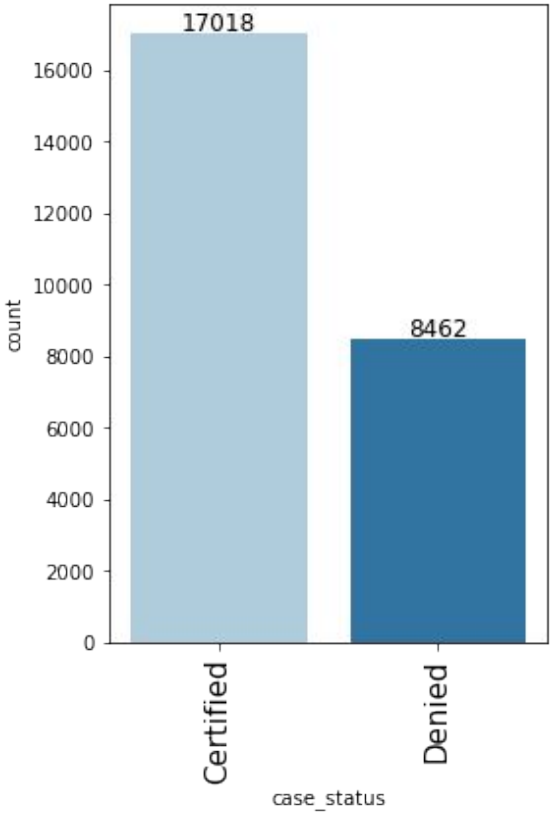
Unit of Wage

Most employees in a similar area of intended employment earn in yearly units. 90.1% of employees received work compensation yearly, 8.5% were compensated hourly, 1.1% were compensated weekly, and 0.3% were compensated monthly.

EDA Results

UNIVARIATE ANALYSIS





Case Status

Most applicants were certified. 66.8%, approximately 2/3 of the applicants in the data, were certified, while the remaining 33.2% were denied.

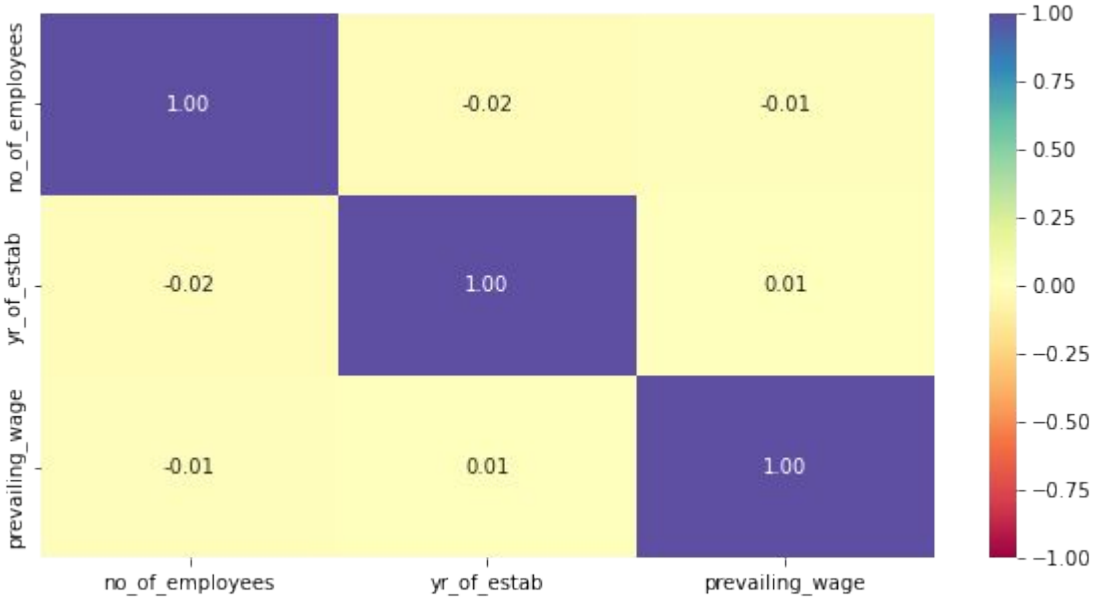
EDA: BIVARIATE ANALYSIS

EDA Results

BIVARIATE ANALYSIS

Heatmap

We plotted a heatmap and found no correlation between the numerical attributes in the data.

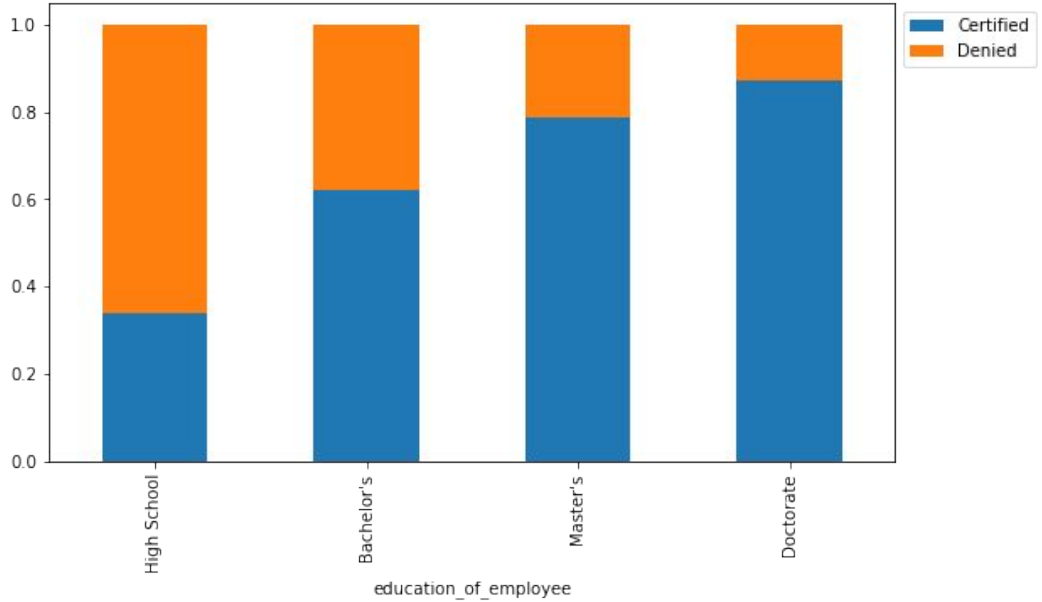


EDA Results

BIVARIATE ANALYSIS

Those with a higher level of education may want to travel abroad for a well-paid job. We checked to see if education has any impact on visa certification.

The bar plot shows that applicants with the highest levels of education are likely to have their applications certified than others. Doctorate holders have 80% more certified cases than denied cases, while high school level education holders had less than 40% certified cases. 60% - 80% of Bachelor's and Master's holders had their applications certified.

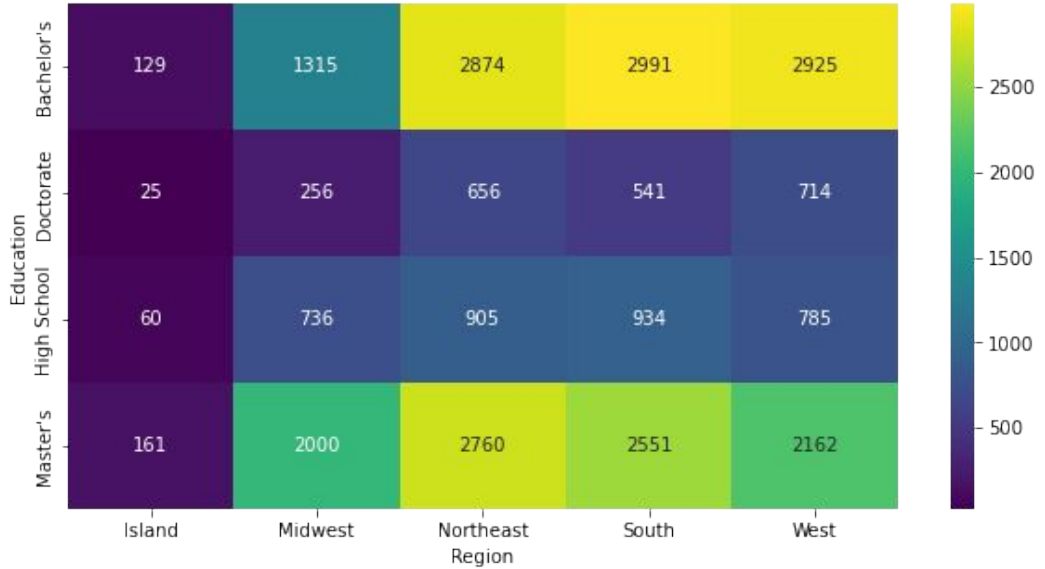


EDA Results

BIVARIATE ANALYSIS

We analyzed different regions and their requirements of talent having diverse educational backgrounds.

The heatmap shows that the highest correlation exists between Bachelor's and Master's levels of education and the Northeast, South and West Regions. While these regions employ Doctorate and High School Cert holders, all three regions hired more Bachelor’s and Master’s holders than any other education level.



EDA Results

BIVARIATE ANALYSIS

We analyzed case status based on the different levels of education.

Approximately, 87% (1,912) of Doctorate holders got certified.

Approximately, 79% (7,575) of Master's holders got certified.

Approximately, 62% (6,367) of Bachelor's holders got certified.

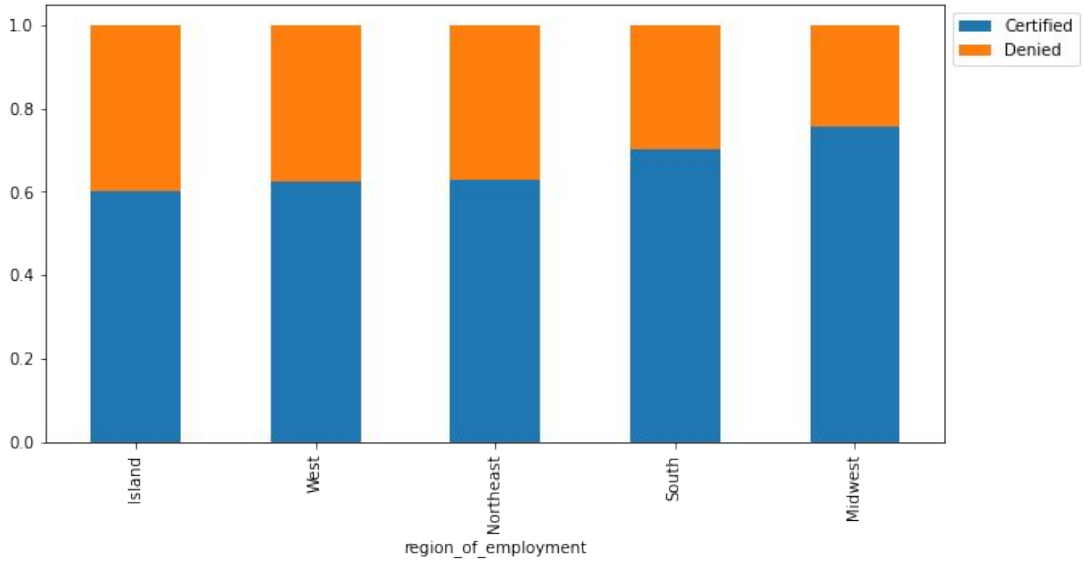
Approximately, 34% (1,164) of High School certificate holders got certified.

EDA Results

BIVARIATE ANALYSIS

We looked at the percentage of visa certifications across each region.

All regions seem to have similar (somewhat uniform) rates of certification across all applicants (between 60% - 80%), with the highest percentage of certifications being in the Midwest, followed by the South region. The South region had the highest count of certifications (4,913 applicants) and the Northeast had the highest count of denials (2,669 applicants).

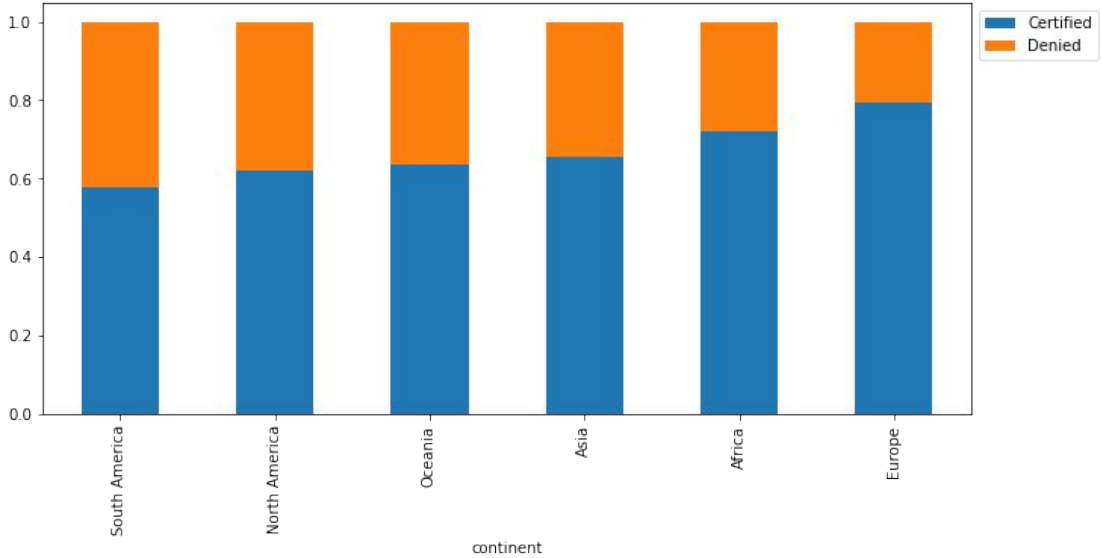


EDA Results

BIVARIATE ANALYSIS

We checked the continents to determine how the visa status varies across different continents.

Europe has the highest percentage of certified applicants (at a count of 2,957), followed by the African continent (397). Asia has a lower percentage of certified applicants than the two but also has the highest count of certifications (11,012 applicants) and denials (5,849 applicants) across all continents. This is due to very high numbers of applications from the Asian continent.

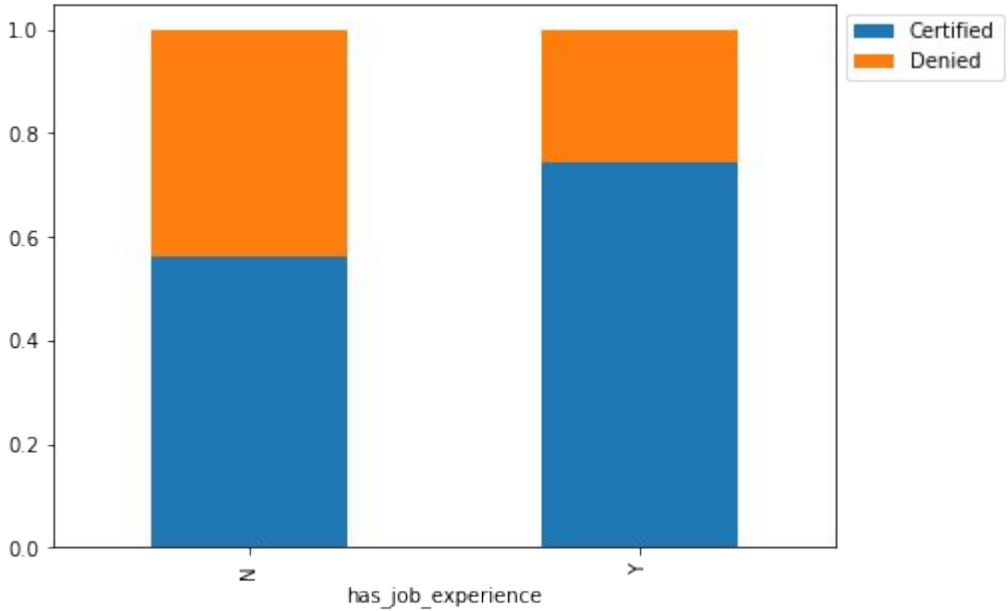


EDA Results

BIVARIATE ANALYSIS

Experienced professionals might look abroad for opportunities to improve their lifestyles and career development. We checked if having work experience has any influence over visa certification

Over 70% of applicants with job experience were certified, while under 60% of applicants without job experience were certified. The highest count of certifications were from applicants (11,024) with job experience and the highest count of denials were from applicants (4,684) without job experience.

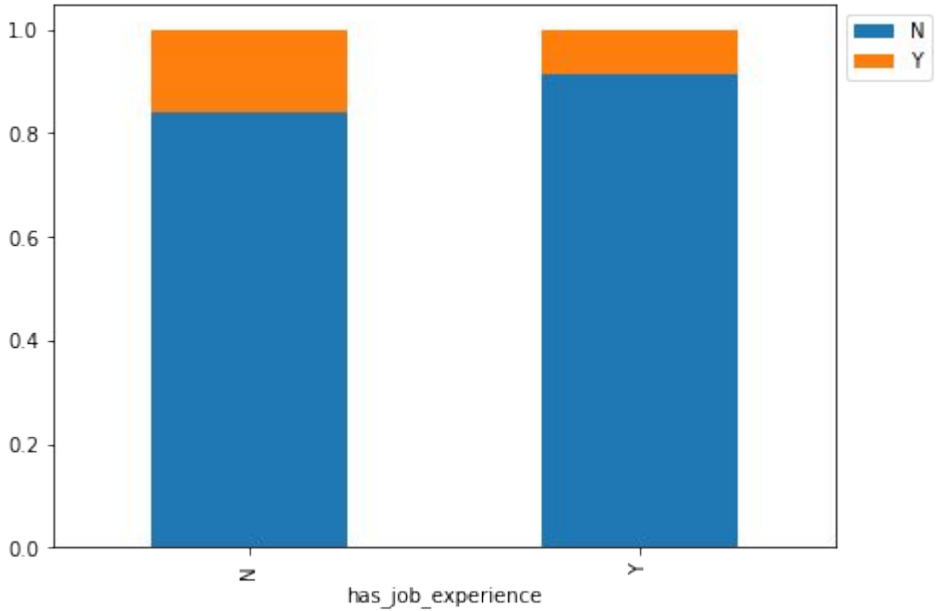


EDA Results

BIVARIATE ANALYSIS

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

The data shows that most employees with prior job experience will not require any additional job training, while most employees without any prior job experience also tend to not require job training. However, the data shows that those employees without prior experience tend to require job training more frequently than those with prior experience.

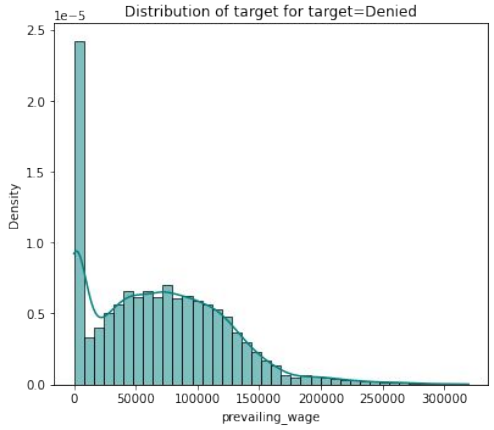


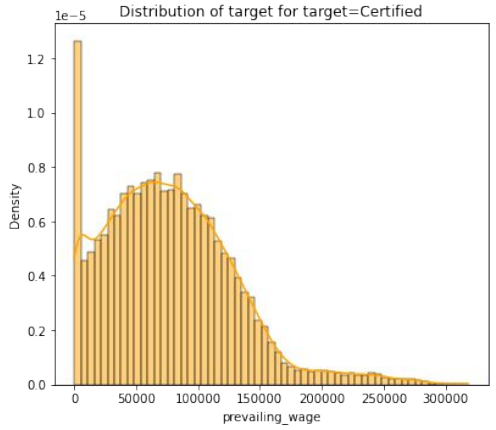
EDA Results

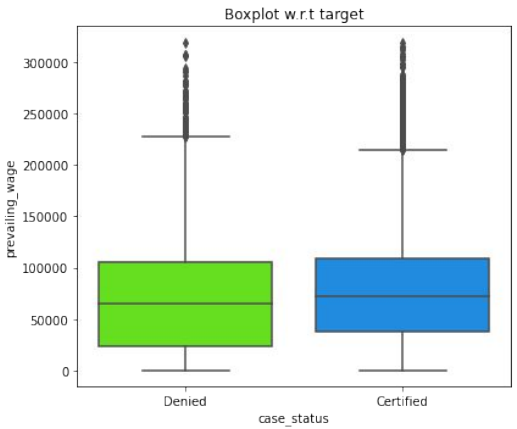
BIVARIATE ANALYSIS

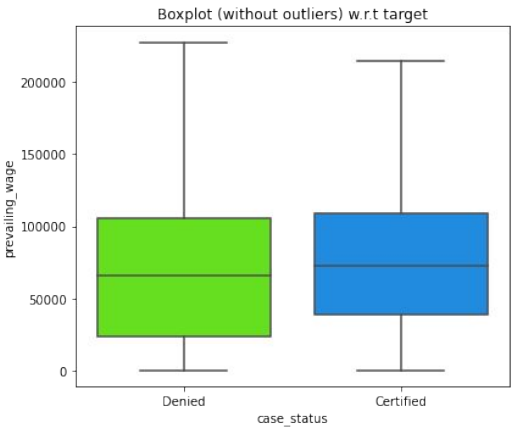
Prevailing wage was established to protect local and foreign workers. We analyzed the data and see if the visa status changes with the prevailing wage

The distribution of the combined attributes (prevailing wage and case status) are positively skewed, separately the distribution for both certified and denied profiles appear to be quite similar. The box plot shows the median prevailing wage for applicants, certified and denied, to be between 50,000 - 100,000 dollars respectively. The conclusion here is that the visa status does not change significantly based on the prevailing wage.







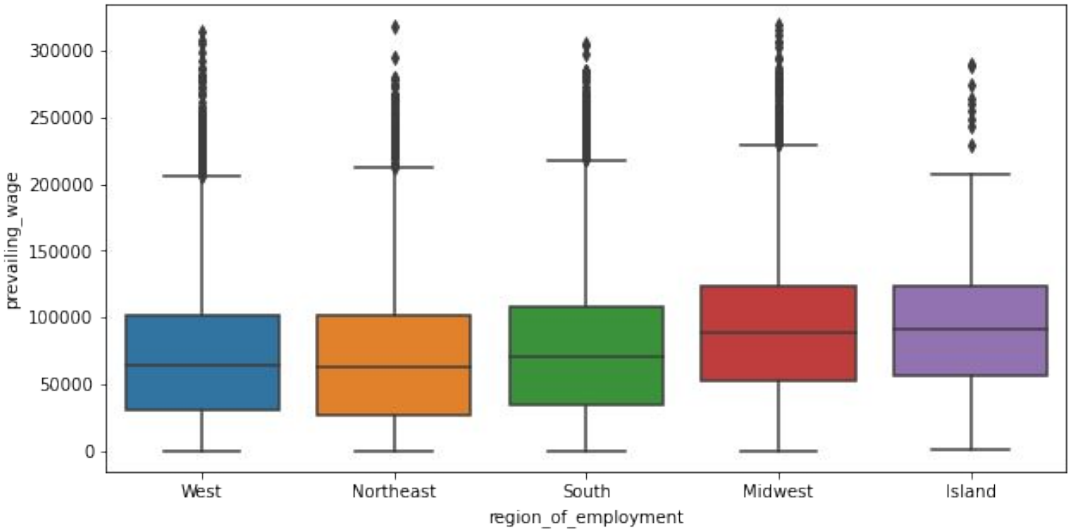


EDA Results

BIVARIATE ANALYSIS

We checked the data to determine whether prevailing wage is similar across all the regions of the US.

The prevailing wage appears to be similar across all regions in the US. Some have slightly higher and lower values than others but the differences are not so stark. For instance, Midwest and Island regions both have a higher median value of prevailing wage than the other three regions, yet all regions have median values between 60,000 and 90,000. Island has the lowest maximum prevailing wage than other regions. All regions have numerous outliers in the data.

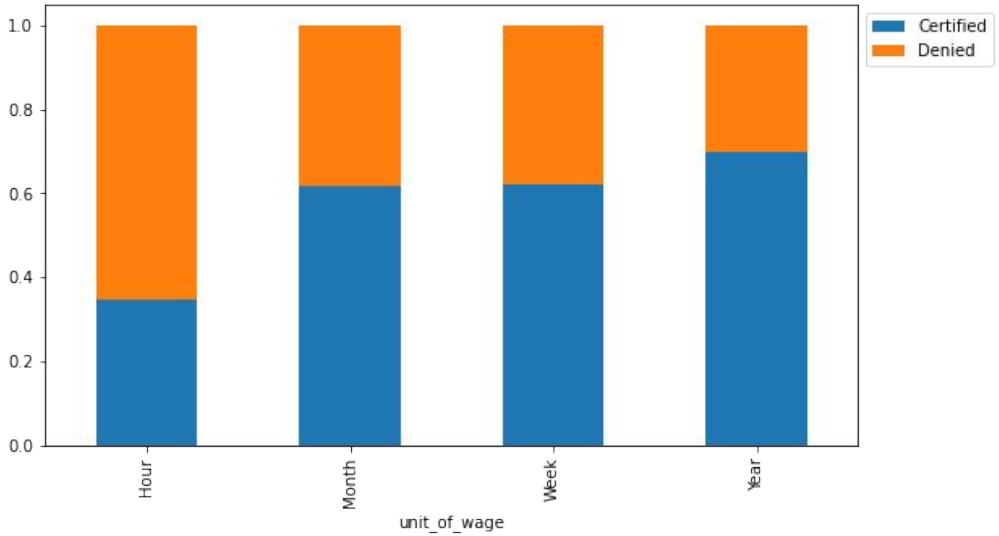


EDA Results

BIVARIATE ANALYSIS

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

Employees who received yearly, monthly and weekly pay were certified at higher rates than hourly workers. Employees who are paid an annual salary have the highest certified (16,047) and denied (6,915) counts, this makes sense since the majority of candidates in the data are paid annually.



DATA PROCESSING & MODELING CRITERIA

Data Processing Overview

Data Preprocessing:

Since there are no missing or duplicate values, we do not have to treat the data for this.

We have made necessary adjustments to observations and attributes in the data. For instance, the case ID column was dropped and negative values in the "no_of_employees" column were fixed by converting them into positive values.

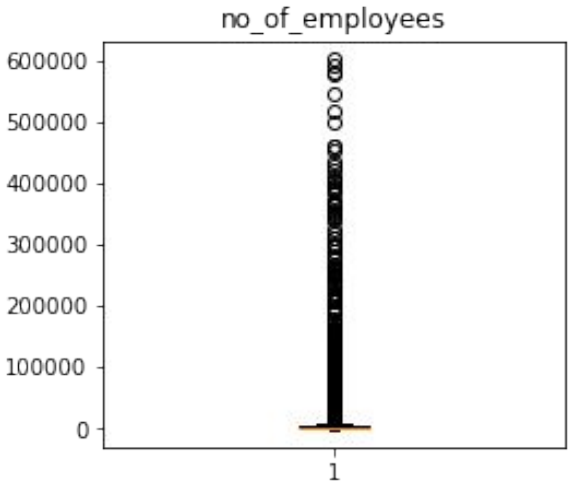
Data Preparation for Modeling:

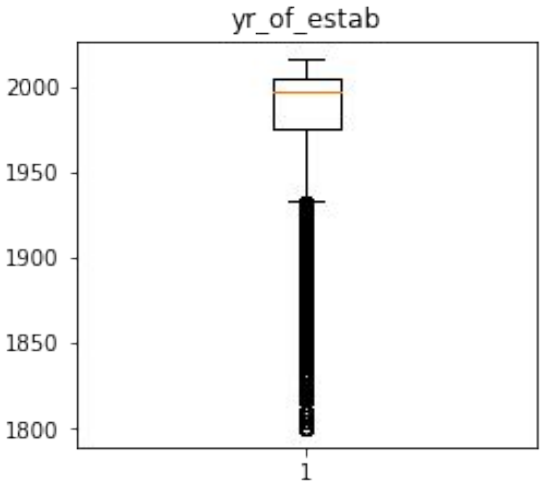
We encoded categorical features and split the data into train and test (70:30). After these steps, there were 17,836 observations in the training set and 7,644 observations in the test set, and 21 attributes in both training and test sets.

Data Processing: Outlier Detection and Treatment

Outlier Check:

There are a lot of outliers in the data. However, we will not treat them as they are proper values.





Model Building Criterion

Model can make wrong predictions as:

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

Which case is more important?

Both cases are as important because:

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

MODEL BUILDING & HYPERPARAMETER TUNING

Confusion Matrix Overview

True Positives (TP): The visa should get certified and the model predicted that it will get certified.

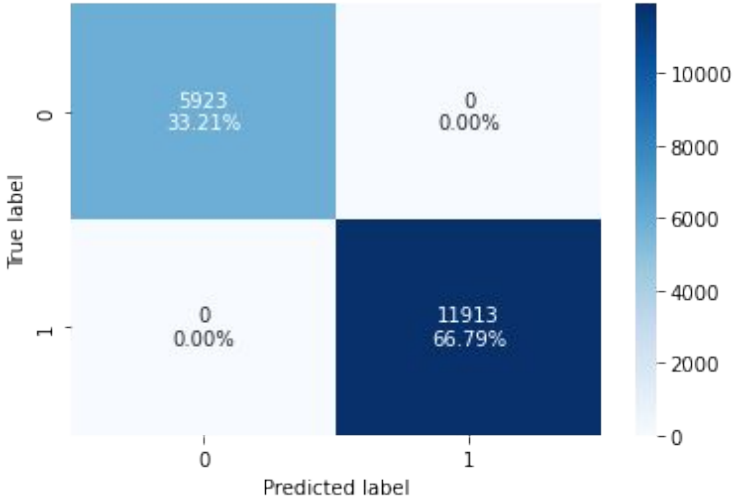
True Negatives (TN): The visa should get denied and the model predicted that it will not get certified.

False Positives (FP): The visa should get denied and the model predicted that it will get certified.

False Negatives (FN): The visa should get certified and the model predicted that it will not get certified.

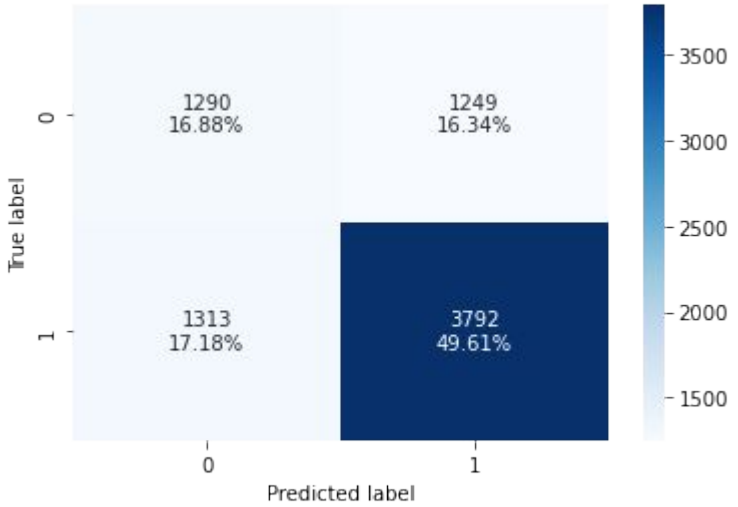
Decision Tree Model - No Pruning

TRAINING SET:



	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

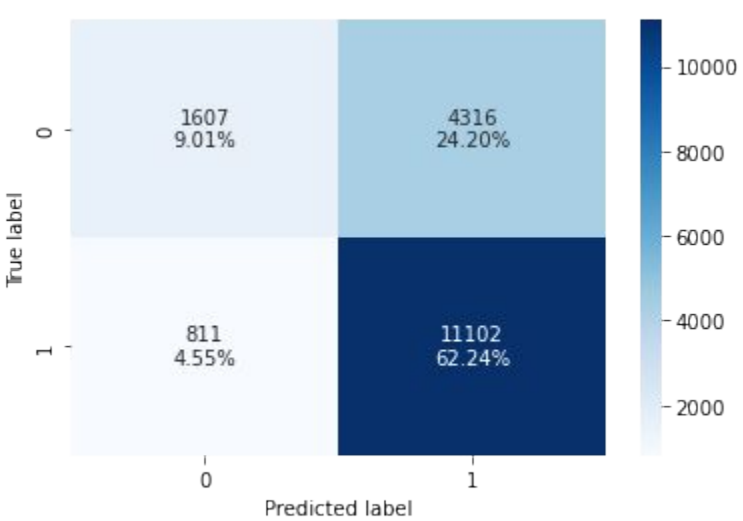
TEST SET:



	Accuracy	Recall	Precision	F1
0	0.664835	0.742801	0.752232	0.747487

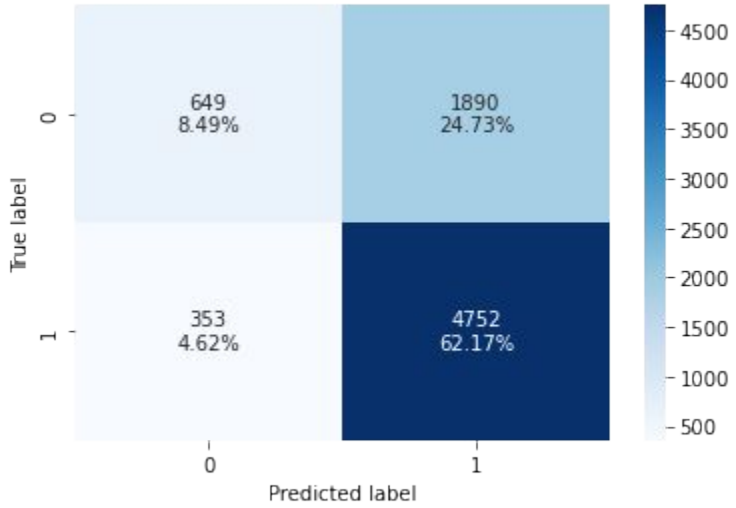
Decision Tree Model - Hyperparameter Tuning

TRAINING SET:



	Accuracy	Recall	Precision	F1
0	0.712548	0.931923	0.720067	0.812411

TEST SET:



	Accuracy	Recall	Precision	F1
0	0.706567	0.930852	0.715447	0.809058

Model Performance Evaluation

UN-PRUNED DECISION TREE:

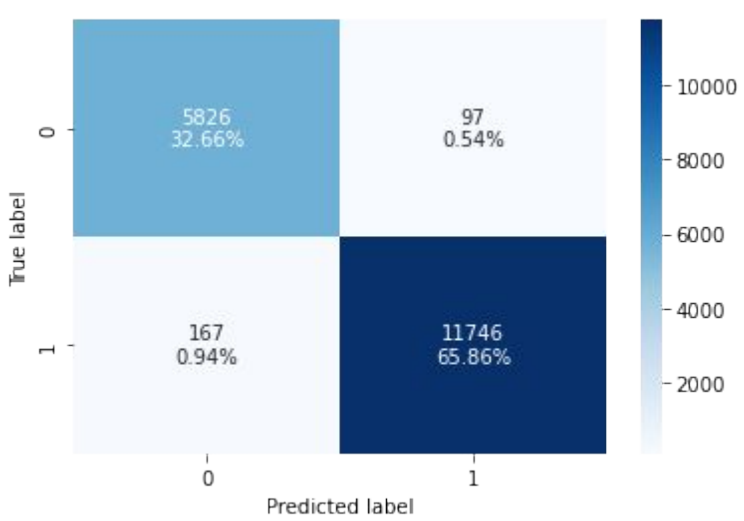
This model was able to perfectly classify all the data points on the training set. Decision trees, without restrictions, will continue to grow until all data points are correctly classified and the trees will learn all the patterns in the training set.
The training set has performed excellently, while the test set has not performed well when both are compared. This model performance check shows us that there is overfitting in this model.

TUNED DECISION TREE:

This model has performed better than the unpruned decision tree, the overfitting of the previous model has been corrected in this model. The training and test set produced similar results. The test set produced an F1 score of 0.8091 and has not performed poorly but we still tested other models to get the best F1 score possible.

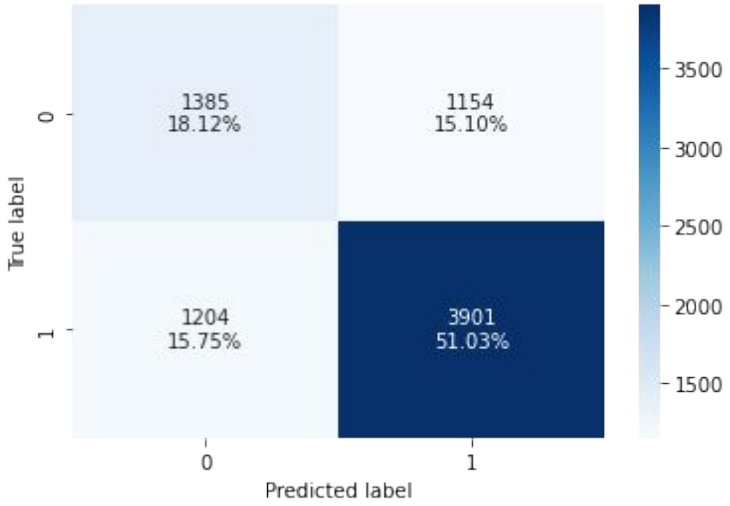
Bagging Classifier - No Tuning

TRAINING SET:



	Accuracy	Recall	Precision	F1
0	0.985198	0.985982	0.99181	0.988887

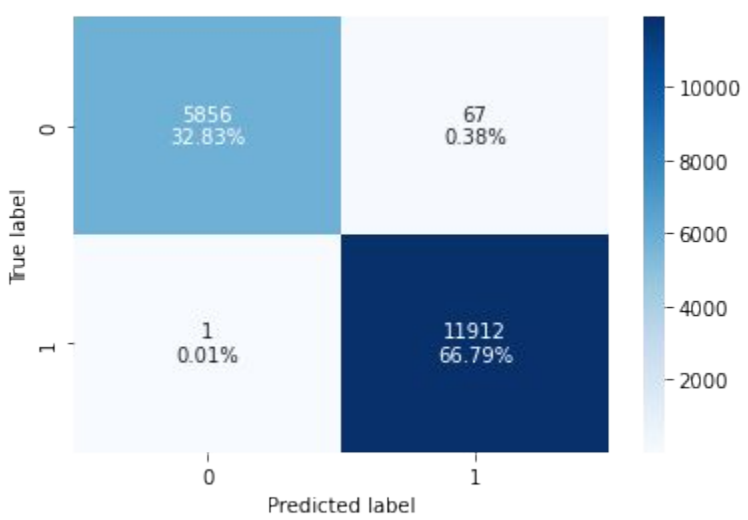
TEST SET:



	Accuracy	Recall	Precision	F1
0	0.691523	0.764153	0.771711	0.767913

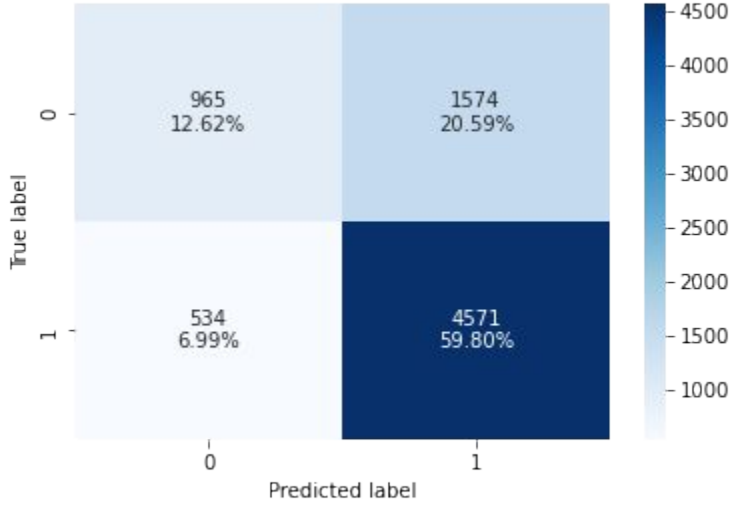
Bagging Classifier - Hyperparameter Tuning

TRAINING SET:



	Accuracy	Recall	Precision	F1
0	0.996187	0.999916	0.994407	0.997154

TEST SET:



	Accuracy	Recall	Precision	F1
0	0.724228	0.895397	0.743857	0.812622

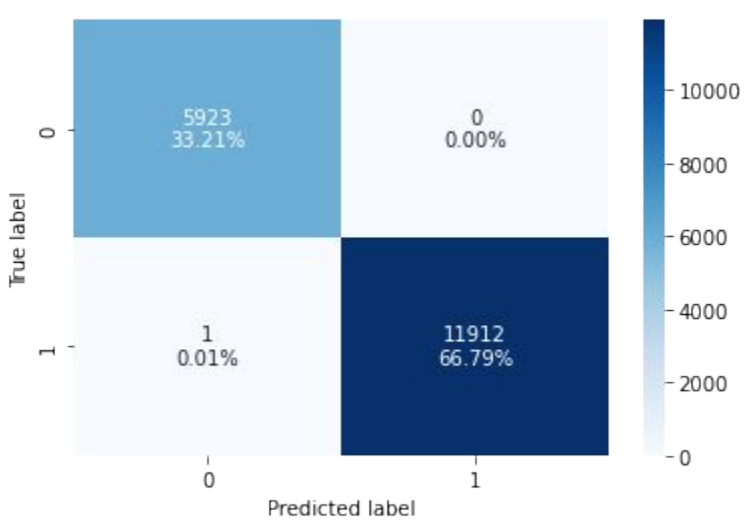
Model Performance Evaluation

BAGGING CLASSIFIER:

These models were able to almost perfectly classify all the data points on the training set but did not do as well on the test set. Based on our chosen metric (F1 score), our test sets performed better than the previous decision tree models we built, on both un-tuned (0.7679) models and tuned (0.8126) models, respectively. We tested more models to get the best F1 score possible.

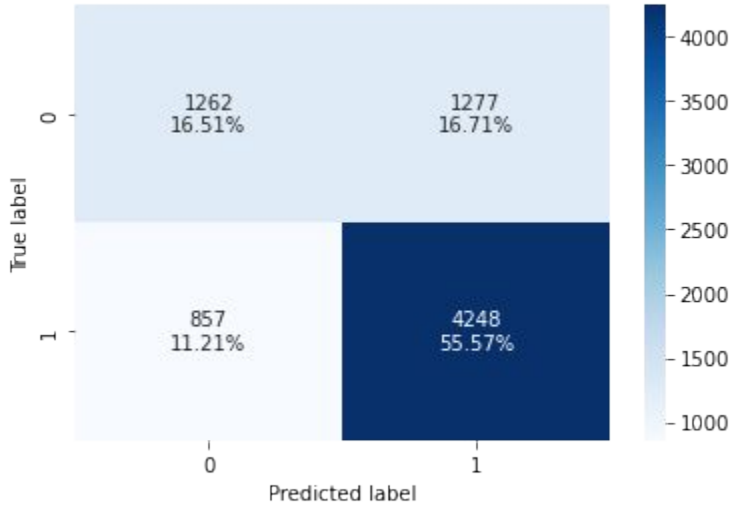
Random Forest - No Tuning

TRAINING SET:



	Accuracy	Recall	Precision	F1
0	0.999944	0.999916	1.0	0.999958

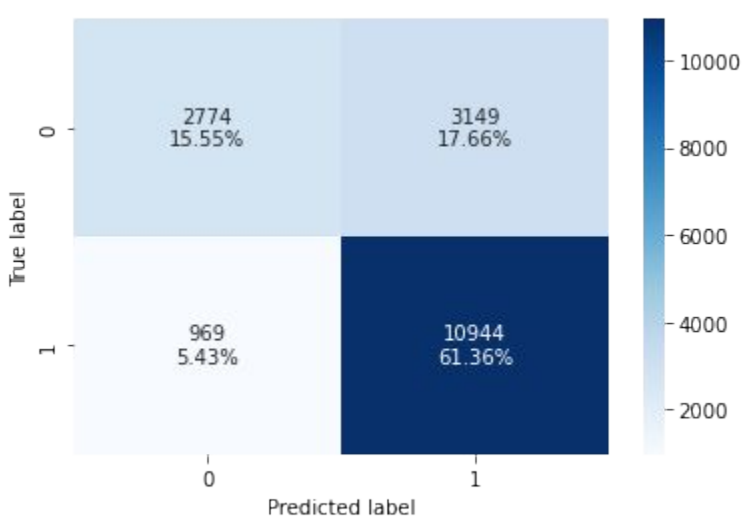
TEST SET:



	Accuracy	Recall	Precision	F1
0	0.720827	0.832125	0.768869	0.799247

Random Forest - Hyperparameter Tuning

TRAINING SET:



	Accuracy	Recall	Precision	F1
0	0.769119	0.91866	0.776556	0.841652

TEST SET:

	Accuracy	Recall	Precision	F1
0	0.738095	0.898923	0.755391	0.82093

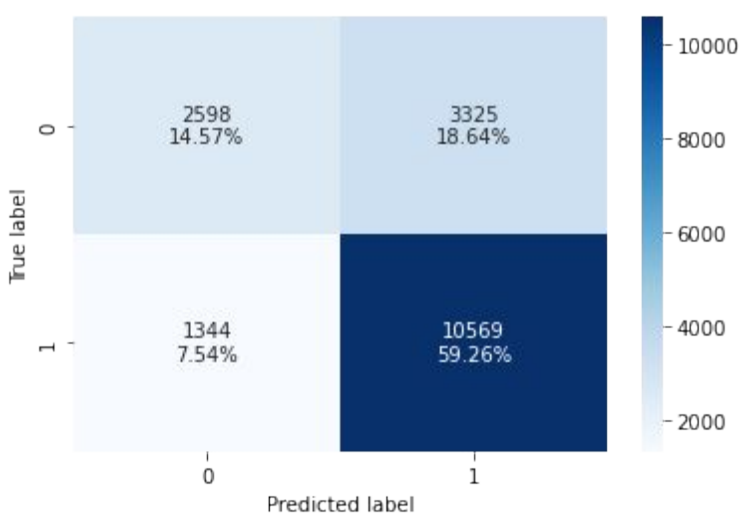
Model Performance Evaluation

RANDOM FOREST:

Our un-tuned random forest models was able to capture most of the data in the training set in an attempt to classify all the data points. Our tuned random forest performed well on the test set and was able to produce an F1 score of 0.82093. This is the highest F1 score we have received so far from all models in the test set, so we continued testing more models to get the best F1 score possible.

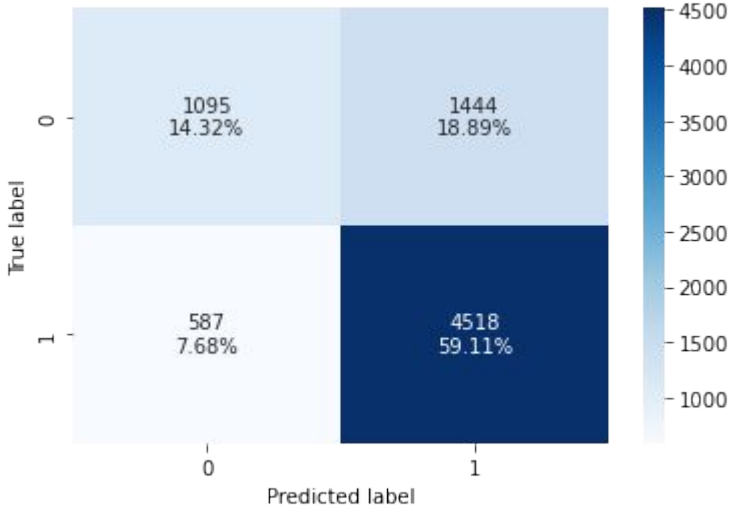
AdaBoost Classifier - No Tuning

TRAINING SET:



	Accuracy	Recall	Precision	F1
0	0.738226	0.887182	0.760688	0.81908

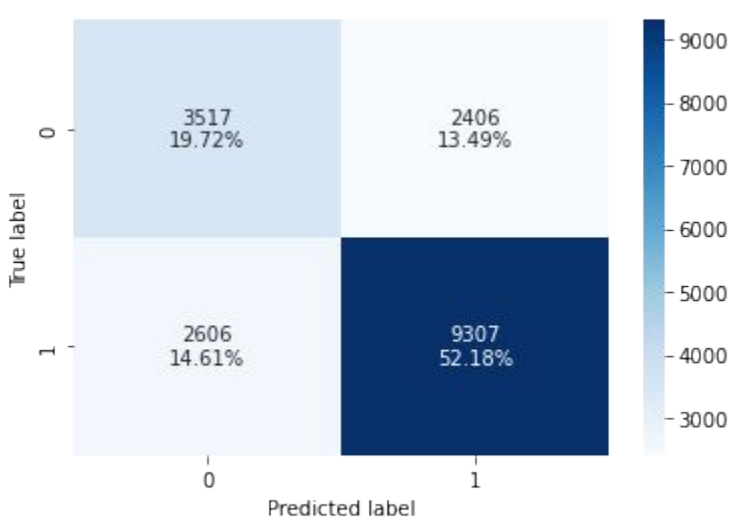
TEST SET:



	Accuracy	Recall	Precision	F1
0	0.734301	0.885015	0.757799	0.816481

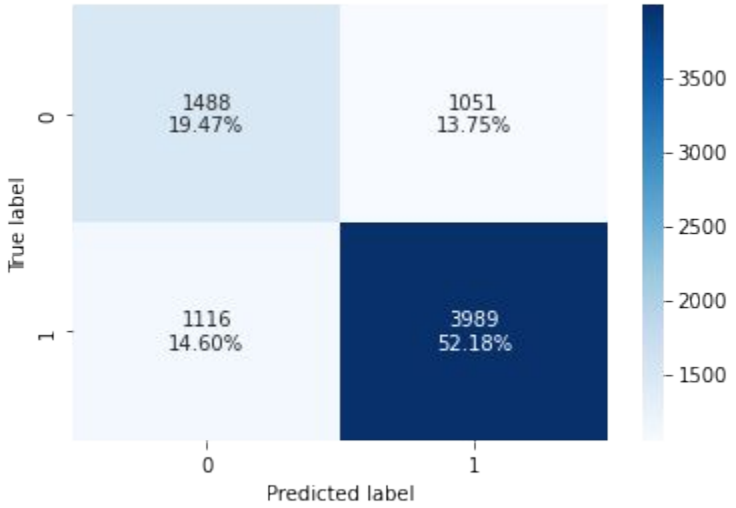
AdaBoost Classifier - Hyperparameter Tuning

TRAINING SET:



	Accuracy	Recall	Precision	F1
0	0.718995	0.781247	0.794587	0.787861

TEST SET:



	Accuracy	Recall	Precision	F1
0	0.71651	0.781391	0.791468	0.786397

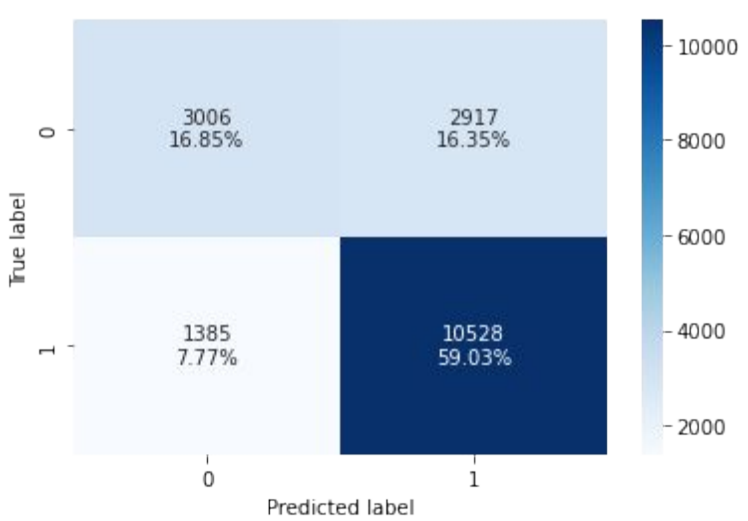
Model Performance Evaluation

ADABOOST CLASSIFIER:

These models were not able to classify all the data points on the training set as well as the previous models but there appears to be much less overfitting on the training data in comparison. Training and test sets have performed similarly, but compared to the results from our previous models not as well. Our un-tuned adaboost classifier model performed slightly worse than our tuned random forest so we did not choose this model, instead, we continued to test other models to get the best F1 score possible.

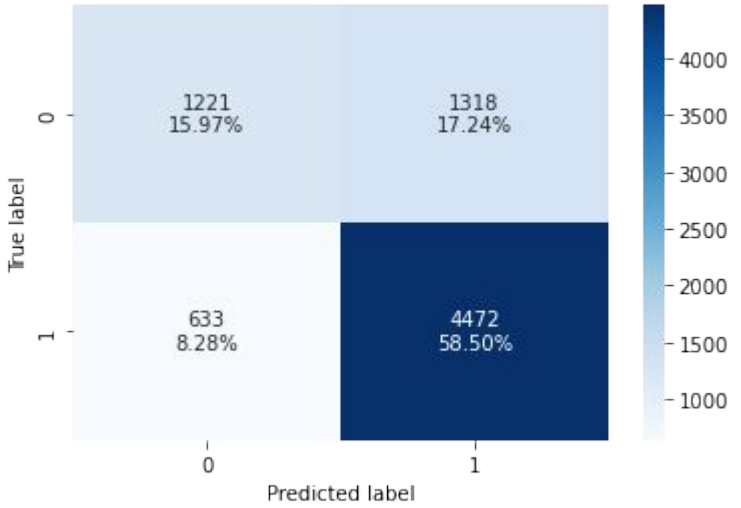
Gradient Boosting Classifier - No Tuning

TRAINING SET:



	Accuracy	Recall	Precision	F1
0	0.758802	0.88374	0.783042	0.830349

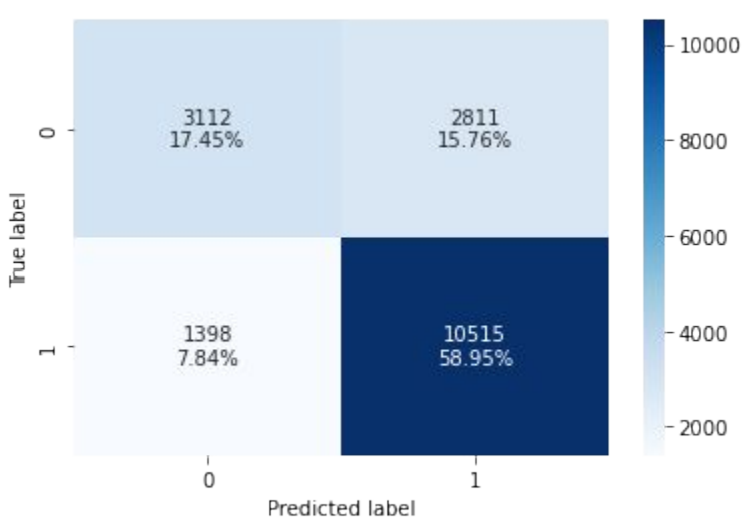
TEST SET:



	Accuracy	Recall	Precision	F1
0	0.744767	0.876004	0.772366	0.820927

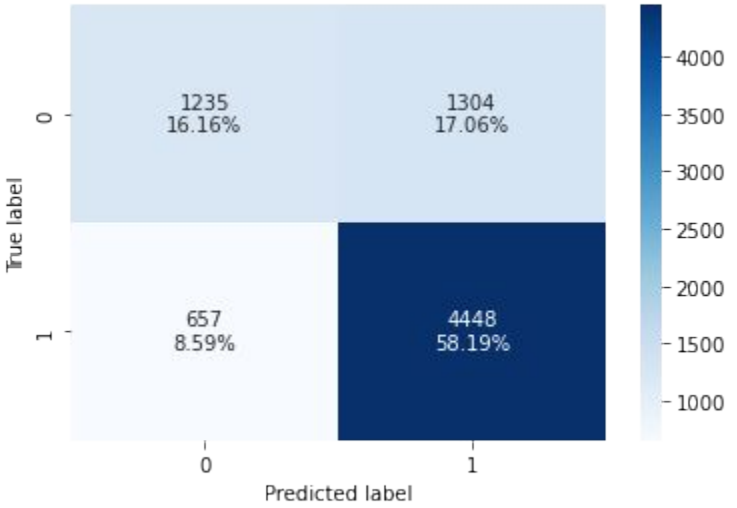
Gradient Boosting Classifier - Hyperparameter Tuning

TRAINING SET:



	Accuracy	Recall	Precision	F1
0	0.764017	0.882649	0.789059	0.833234

TEST SET:



	Accuracy	Recall	Precision	F1
0	0.743459	0.871303	0.773296	0.819379

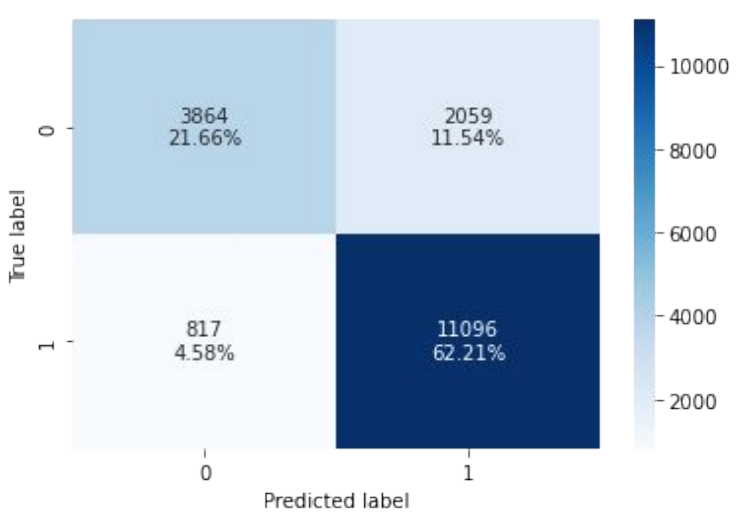
Model Performance Evaluation

GRADIENT BOOSTING CLASSIFIER:

Both models performed well on the test set. However, the untuned gradient boosting classifier performed better than the tuned gradient boosting classifier model on the test set. We received the same F1 score from our untuned gradient boosting classifier model and our tuned random forest. These test sets have performed well but we can still test other models to get the best F1 score possible.

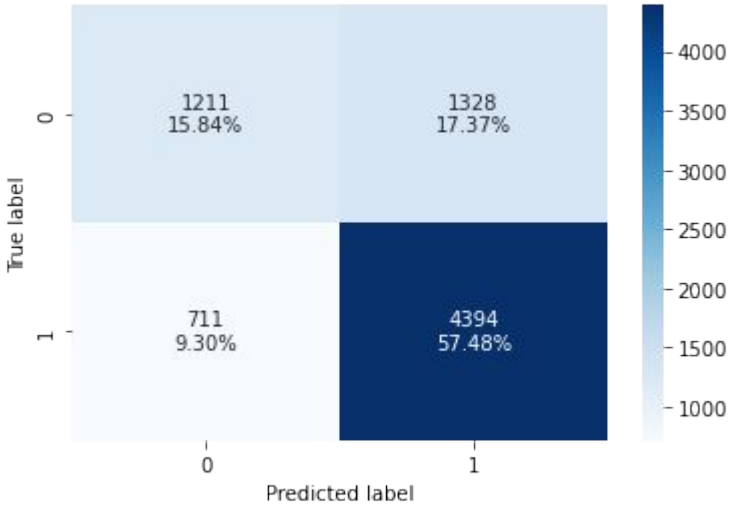
XGBoost Classifier - No Tuning

TRAINING SET:



	Accuracy	Recall	Precision	F1
0	0.838753	0.931419	0.843482	0.885272

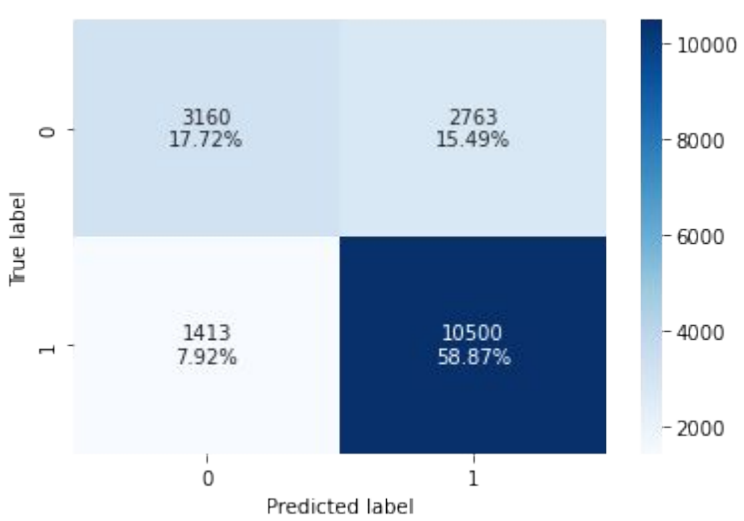
TEST SET:



	Accuracy	Recall	Precision	F1
0	0.733255	0.860725	0.767913	0.811675

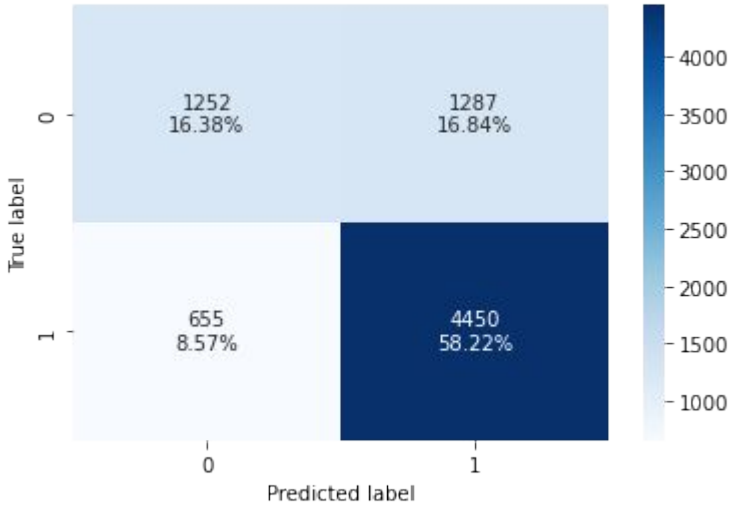
XGBoost Classifier - Hyperparameter Tuning

TRAINING SET:



	Accuracy	Recall	Precision	F1
0	0.765867	0.88139	0.791676	0.834128

TEST SET:



	Accuracy	Recall	Precision	F1
0	0.745945	0.871694	0.775667	0.820882

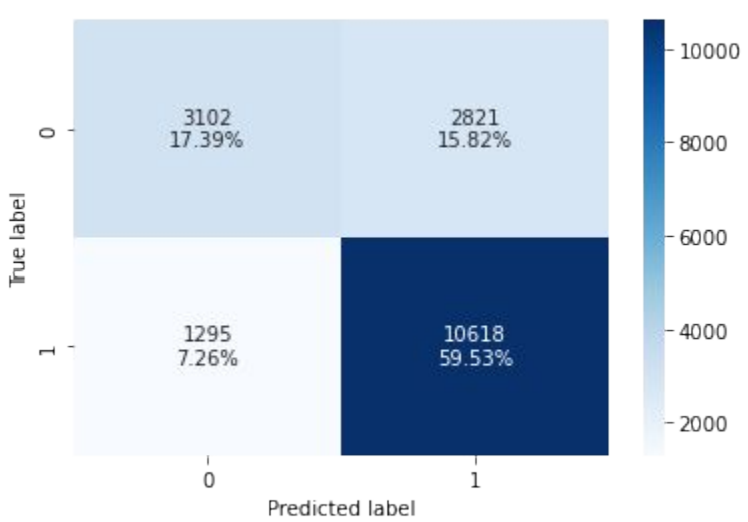
Model Performance Evaluation

XGBOOST CLASSIFIER:

Both xgboost models did not perform as well as our previous models. While the test set of the tuned xgboost classifier model gave us an F1 score of 0.8209, it is still performing only marginally worse than our tuned random forest and untuned gradient boosting classifier. As a result, we did not select xgboost classifier and we tested one final classifier model to see if we can get our best F1 score yet.

Stacking Classifier

TRAINING SET:



	Accuracy	Recall	Precision	F1
0	0.769231	0.891295	0.790089	0.837646

TEST SET:

	Accuracy	Recall	Precision	F1
0	0.744113	0.878942	0.7703	0.821043

Model Performance Evaluation

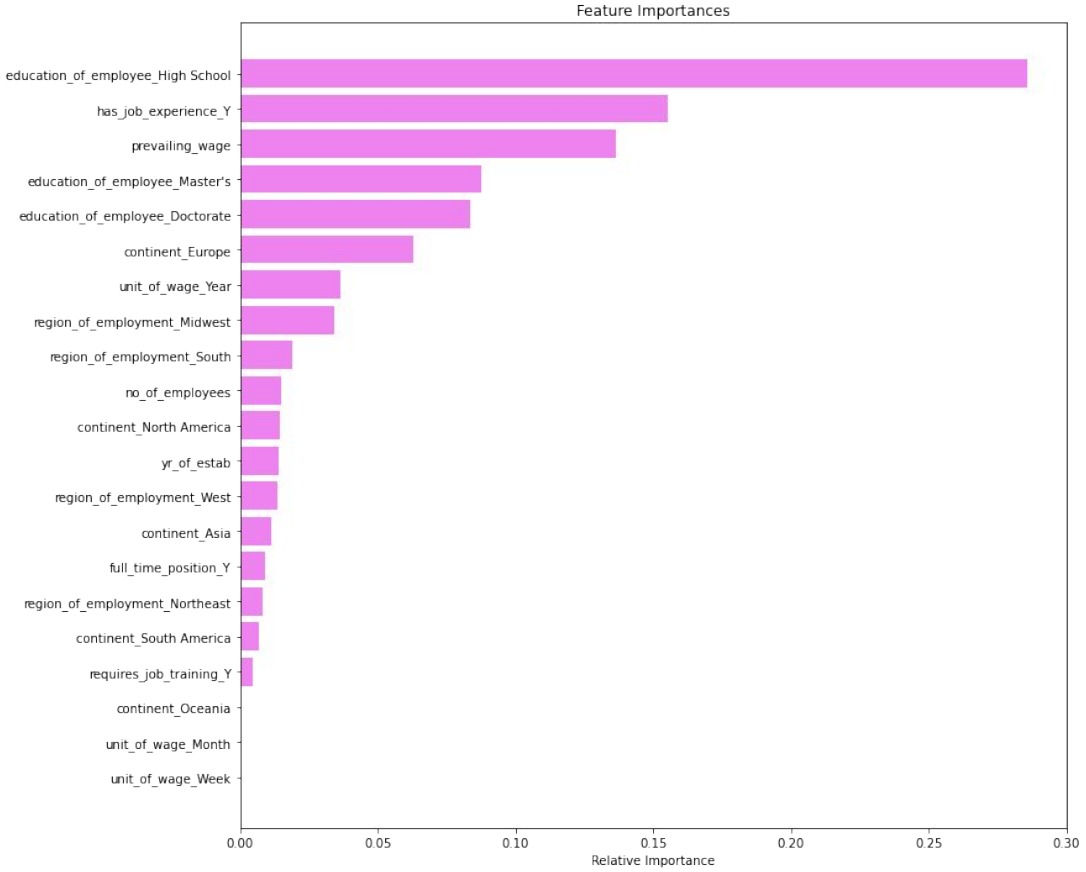
STACKING CLASSIFIER:

This model has given us our best model performance so far, with an F1 score of 0.821043, on the test set. We know that the stacking classifier has produced the best result of all models based on our chosen metric. This means we have a 82.1% chance of minimizing false negatives and false positives during profile examinations by using this stacking classifier model.

PERFORMANCE EVALUATION & MODEL SELECTION

Feature Importance Evaluation

The most important feature in our final model is the high school value in the education of employee column. The next most important feature is having job experience, followed by prevailing wage.



Model Performance Summary:

TRAINING SET:

	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	XGBoost Classifier Tuned	Stacking Classifier
Accuracy	1.0	0.712548	0.985198	0.996187	0.999944	0.769119	0.738226	0.718995	0.758802	0.764017	0.838753	0.765867	0.769231
Recall	1.0	0.931923	0.985982	0.999916	0.999916	0.918660	0.887182	0.781247	0.883740	0.882649	0.931419	0.881390	0.891295
Precision	1.0	0.720067	0.991810	0.994407	1.000000	0.776556	0.760688	0.794587	0.783042	0.789059	0.843482	0.791676	0.790089
F1	1.0	0.812411	0.988887	0.997154	0.999958	0.841652	0.819080	0.787861	0.830349	0.833234	0.885272	0.834128	0.837646

TEST SET:

	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	XGBoost Classifier Tuned	Stacking Classifier
Accuracy	0.664835	0.706567	0.691523	0.724228	0.720827	0.738095	0.734301	0.716510	0.744767	0.743459	0.733255	0.745945	0.744113
Recall	0.742801	0.930852	0.764153	0.895397	0.832125	0.898923	0.885015	0.781391	0.876004	0.871303	0.860725	0.871694	0.878942
Precision	0.752232	0.715447	0.771711	0.743857	0.768869	0.755391	0.757799	0.791468	0.772366	0.773296	0.767913	0.775667	0.770300
F1	0.747487	0.809058	0.767913	0.812622	0.799247	0.820930	0.816481	0.786397	0.820927	0.819379	0.811675	0.820882	0.821043

Result: Our chosen metric is the F1 score, so we chose the stacking classifier model, since it produced the highest F1 score on test set.

THE END :)