Easy Visa Machine Learning Solution

Business Analysis for Ensemble Techniques
November 1, 2021
James Watt

Easy Visa

Contents

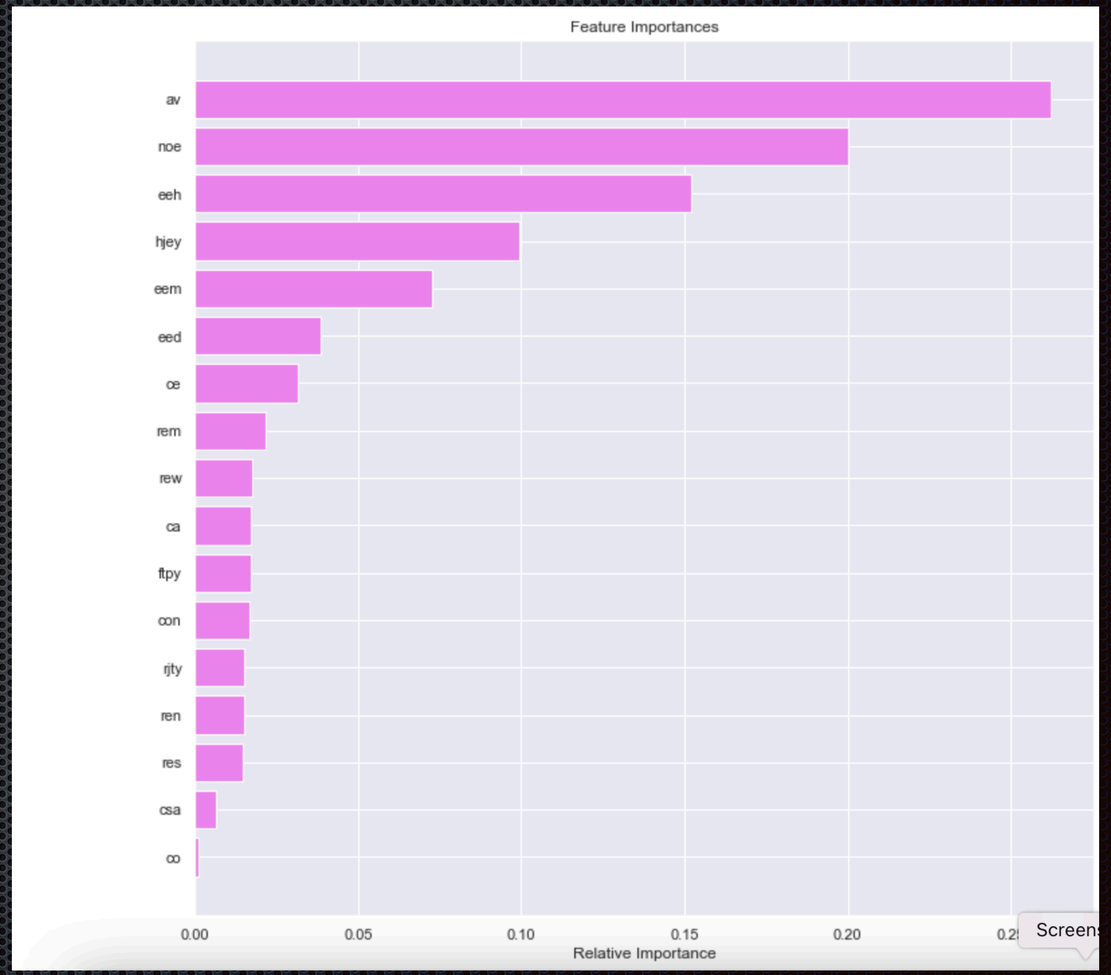
- Opportunity
- Data Overview
- Analysis
- Insight and Recommendations

Opportunity

The Office of Foreign Labor Certification requires a Machine Learning capability to expedite VISA approval processes.

There are two expected outcomes:

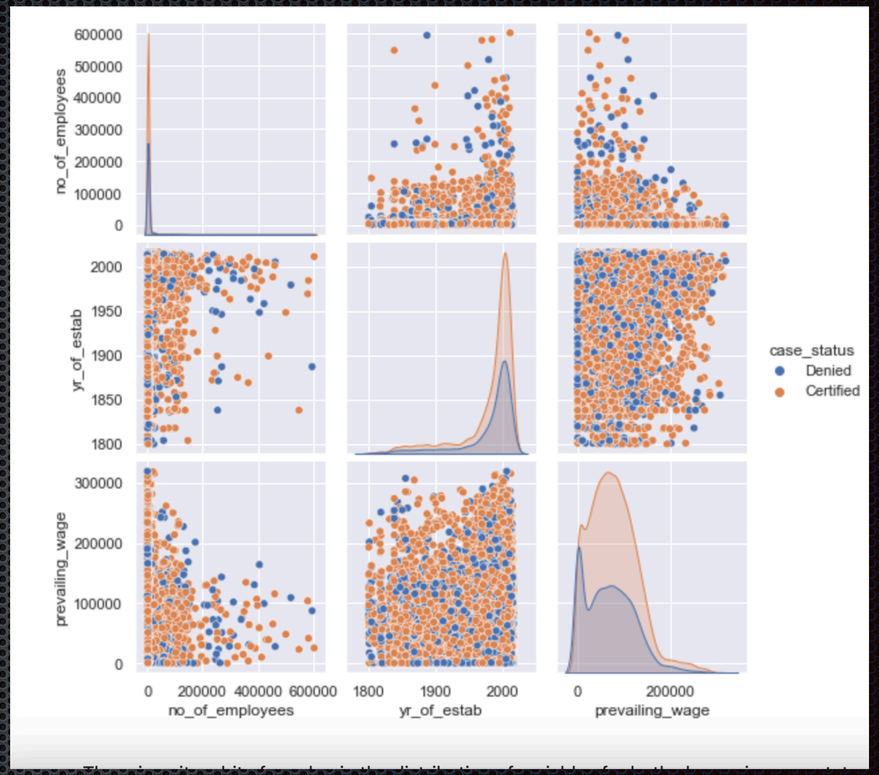
- 1. Facilitate the process of visa approvals.
- 2. Recommend a suitable profile for the applicants for whom the visa should be certified or denied.



Data Overview

Observations:

- There are 25480 records in 12 columns; This will expand to over 25 columns after data processing
- There are nine object type rows of data, one float and two integers.
- One category (no_of_employees) has negative numbers. This is treated as a typo.
- There are no missing values
- Not all values are numeric; All values will be transmuted to numeric for analysis
- Unit_of_wage and prevailing_wage must be combined to have a standard measure of annual pay;
- Our goal is that data is normalized to ensure same scale of comparison

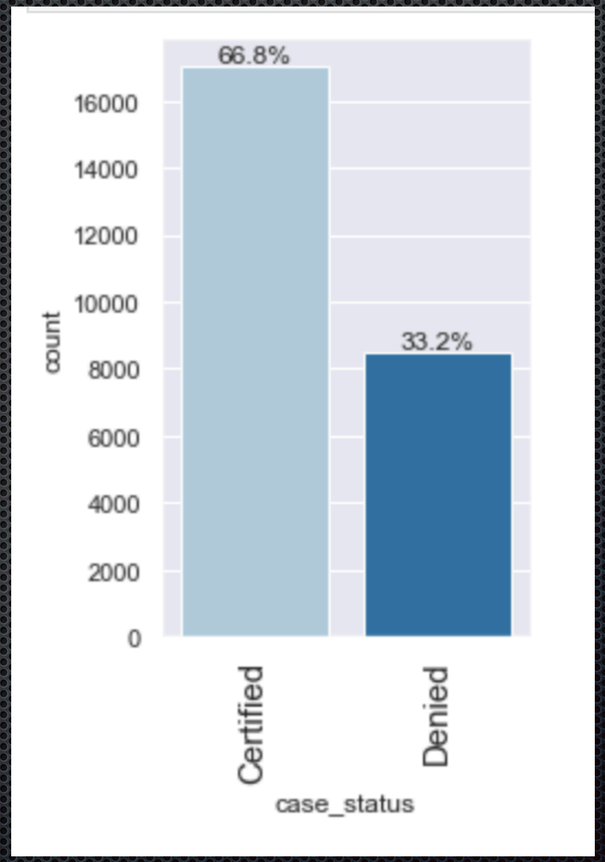


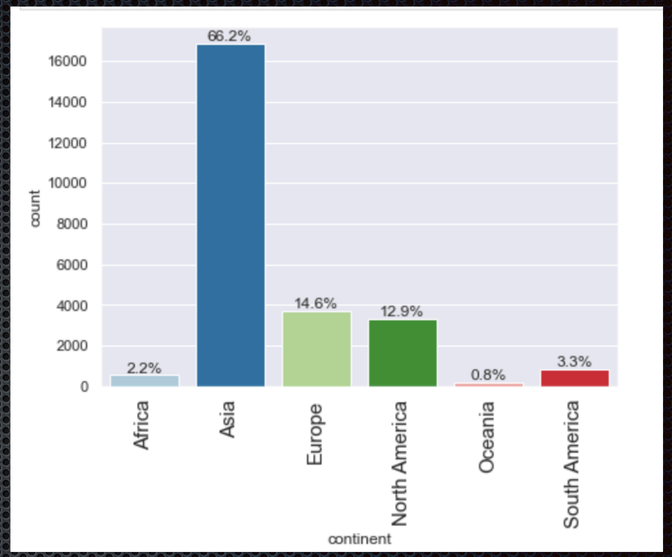
Data Dictionary:

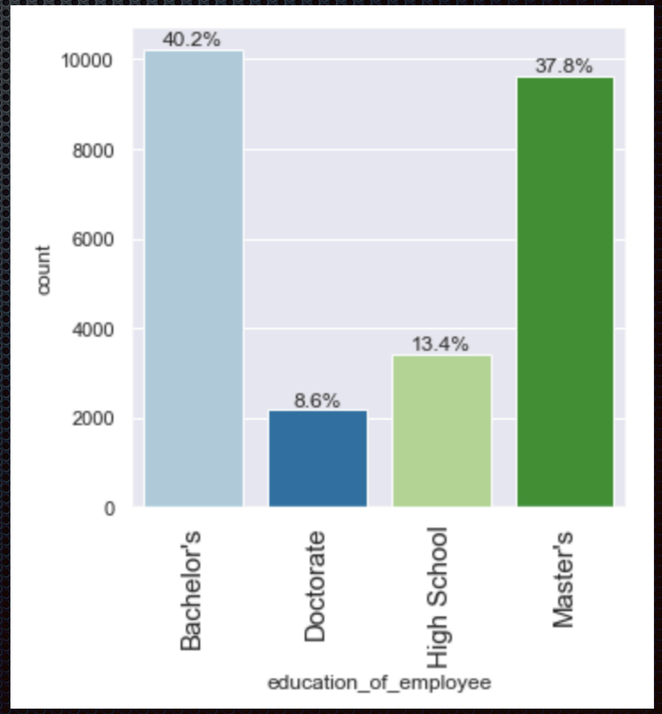
case_id: ID of each visa application
continent: Information of continent the employee
education_of_employee: Information of education of the employee
has_job_experience: Does the employee has any job experience? Y= Yes; N = No
requires_job_training: Does the employee require any job training? Y = Yes; 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
Annual Value: The calculated value per annum to the employer

Exploratory Data Analysis

- Each data field was examined using univariate and bivariate analysis.
- Results from this analysis help shape understanding of feature importance.
- Topical analysis is necessary to answer OFLC questions.

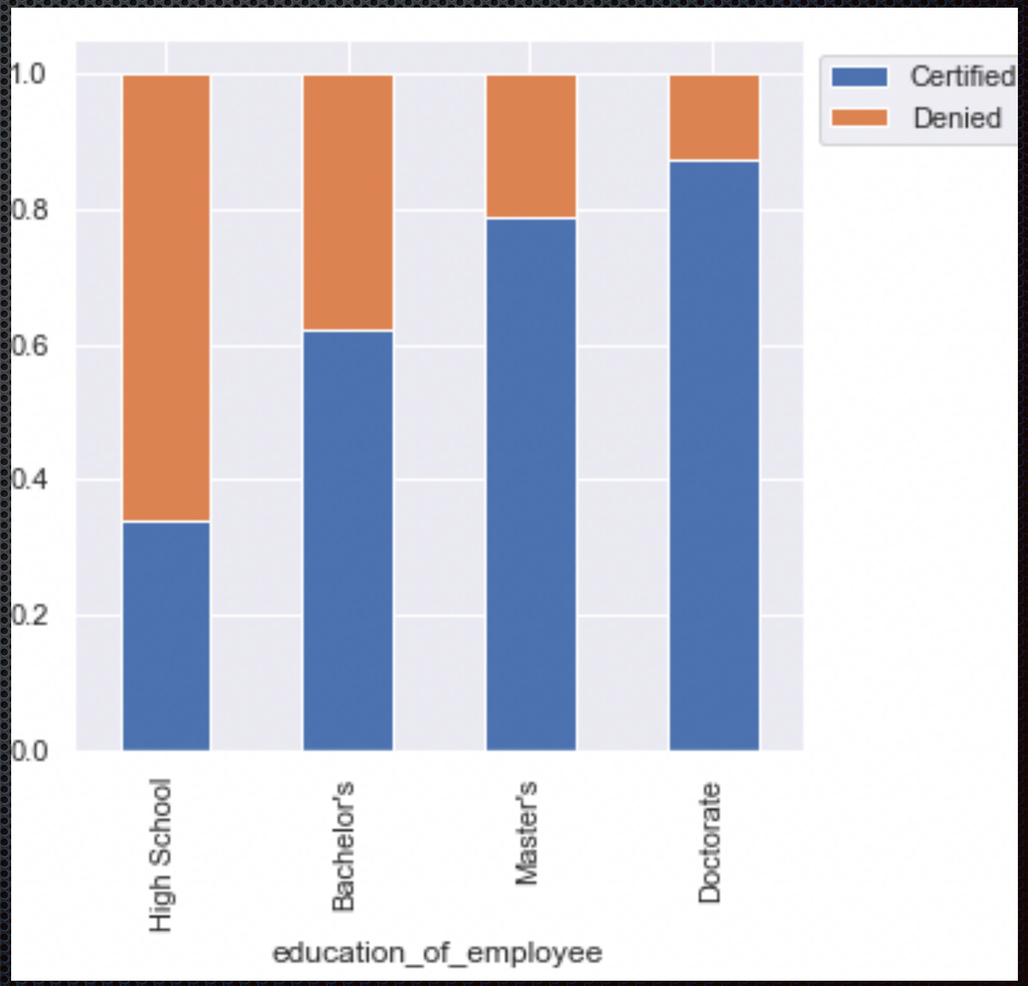






Those with higher education may want to travel abroad for a well-paid job. Does education play a role in Visa certification?

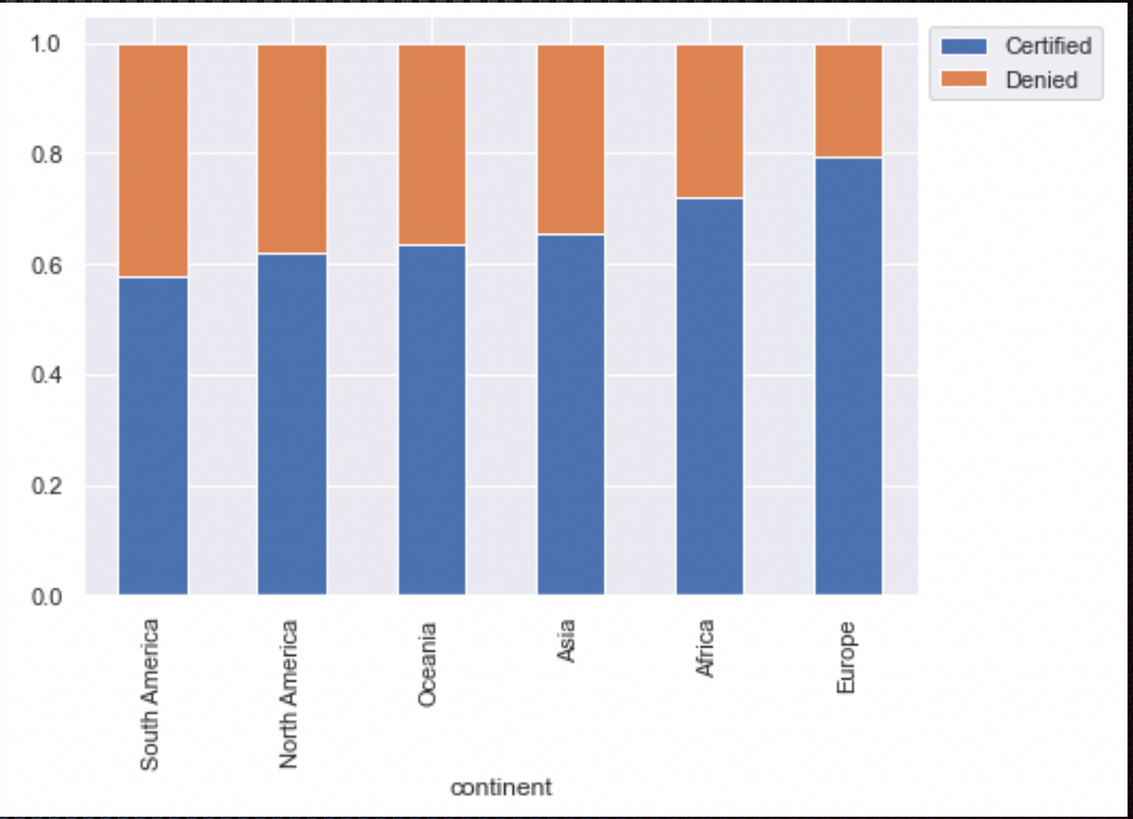
- There is a general trend in the data that shows greater certification for higher education attained.
- One can conclude that education favorably contributes to feature selection of fast-tracking VISA candidates.



How does the visa status vary across different continents?

- There is less than 20% difference between the least approved and most approved continents.
- This does indicate a preference in approvals for European origin.
- Note: The number of applicant per continent varied widely.

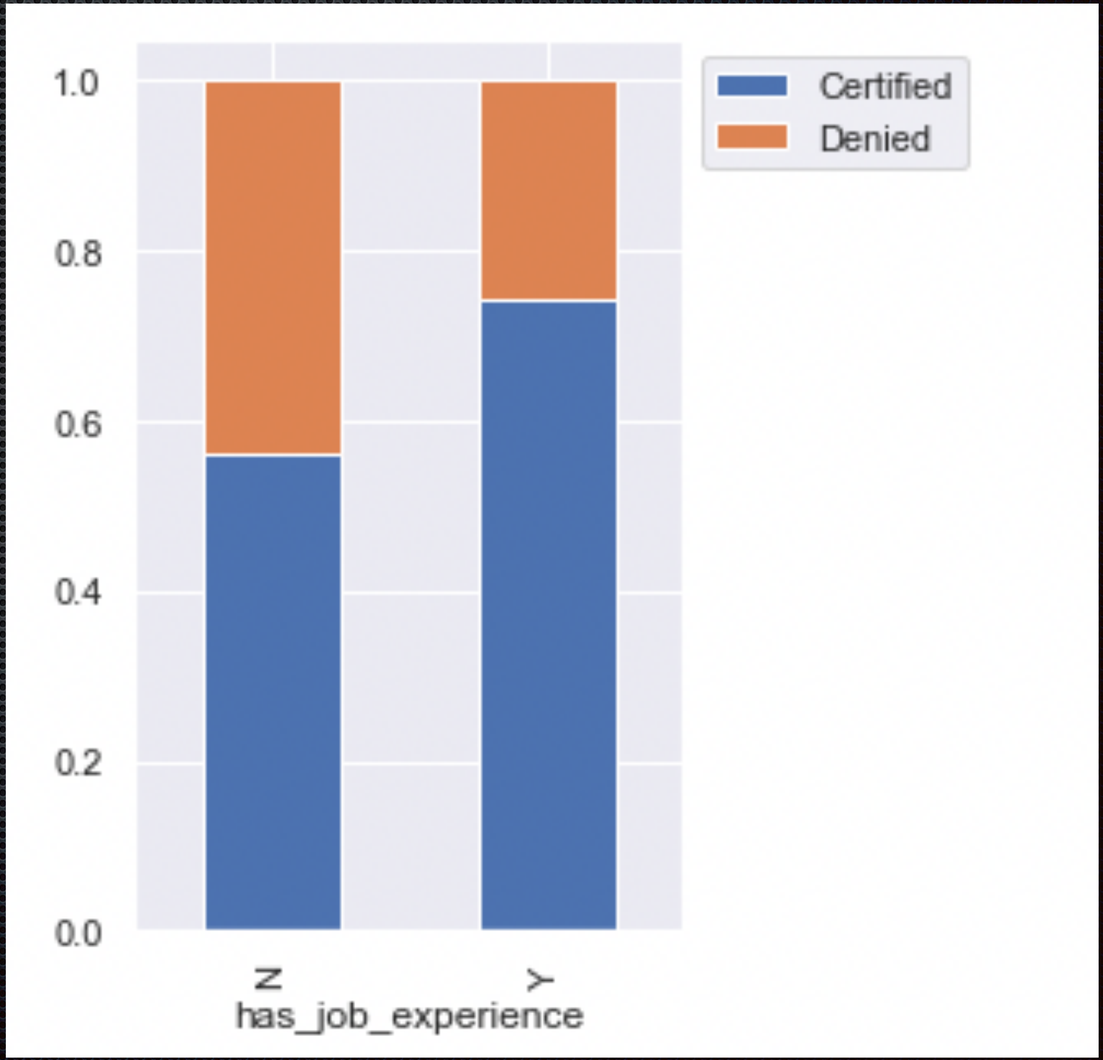
case_status	Certified	Denied	All
continent			
All	17018	8462	25480
Asia	11012	5849	16861
North America	2037	1255	3292
Europe	2957	775	3732
South America	493	359	852
Africa	397	154	551
Oceania	122	70	192



Does Work Experience influence Visa satus?

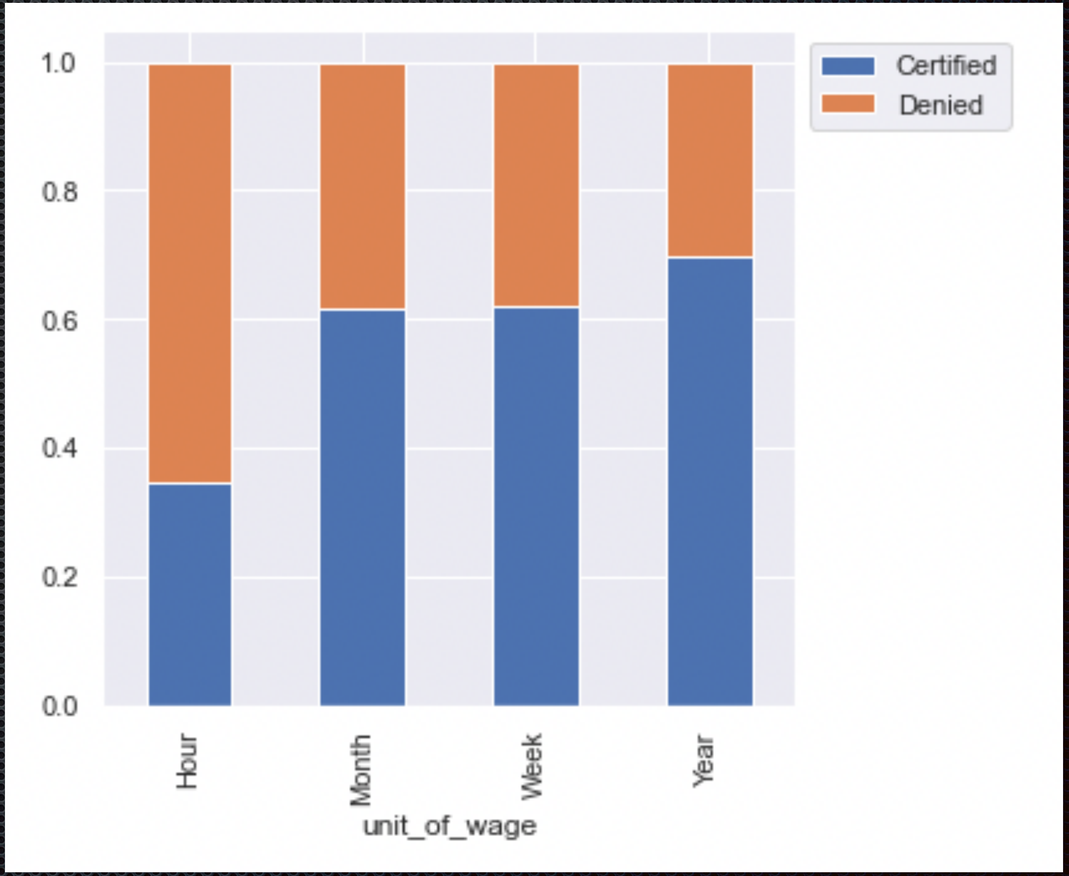
- Work experience directly influences Visa approval.
- 11,024 candidates were certified who had work experience.

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



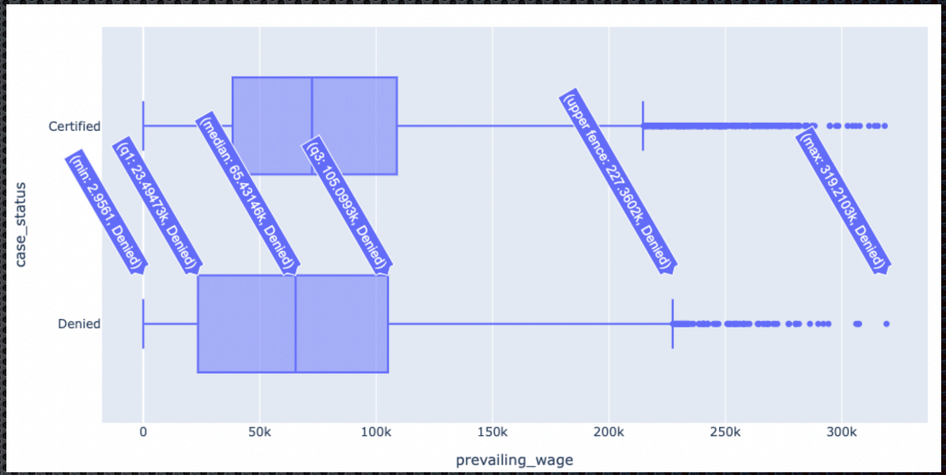
In the United States, employees are paid at different intervals. Which pay unit is most likely to be certified for a Visa?

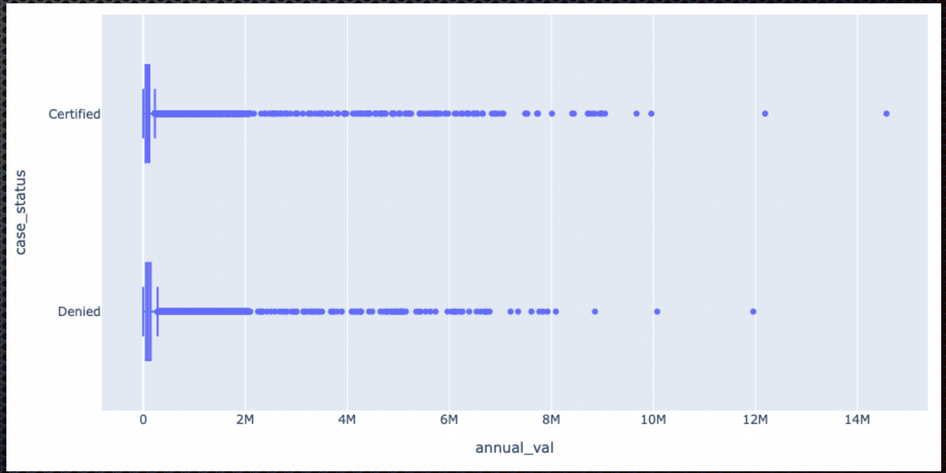
- Employees who are report receiving an annual salary are significantly more likely to be approved for Visas.
- Comparison of unit of wage complicates the definition of value to an employer.
- A better measure is to compare annual value to employer.
- This would be the comparison of wages earned on an annual basis.



The US government has established a prevailing wage to protect workers. How does the visa status change with the prevailing wage?

- Generally candidates who earn more are more likely to be granted a visa.
- The median for candidates who certify is higher for prevailing wage.
- In addition, the outlying candidates who earn Visas tend to have higher annual value.





Insights on Model Creation and Comparison

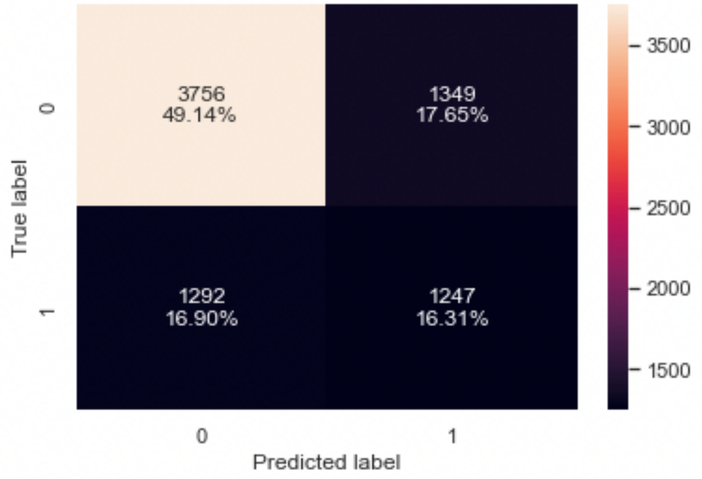
- A Visa model can make wrong predictions:
- It can predict a certification when an employee is denied.
- It can predict a denial when an employee should be given a Visa.
- F-score is the important factor for overall comparison of model accuracy and it combines both precision and recall.
- F-score will be the criteria to judge this model.

$$F_1 = \frac{2}{\frac{1}{\text{recall}} \times \frac{1}{\text{precision}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$
$$= \frac{\text{tp}}{\text{tp} + \frac{1}{2}(\text{fp} + \text{fn})}$$

Decision Tree Classifier

Before Tuning

```
Training performance:
    Accuracy  Recall  Precision   F1
0       1.0     1.0        1.0  1.0
Testing performance:
    Accuracy    Recall  Precision        F1
0    0.6545  0.491138   0.480354  0.485686
```

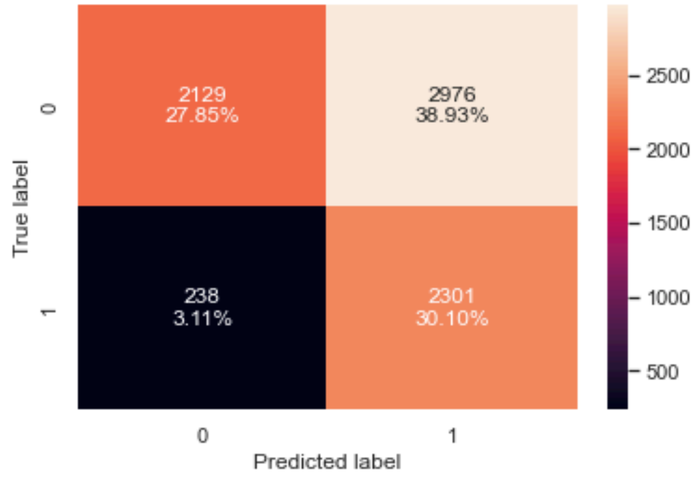


- The decision tree is entirely overfitting the training data (all '1's).

- F-score on training is 1
- F-1 score on testing is .486

Tuning Complete

```
Training performance:
    Accuracy  Recall  Precision   F1
0       1.0     1.0        1.0  1.0
Testing performance:
    Accuracy    Recall  Precision        F1
0    0.6545  0.491138   0.480354  0.485686
```



- The overfitting on training data remains; The model reduces false negatives to 3%.

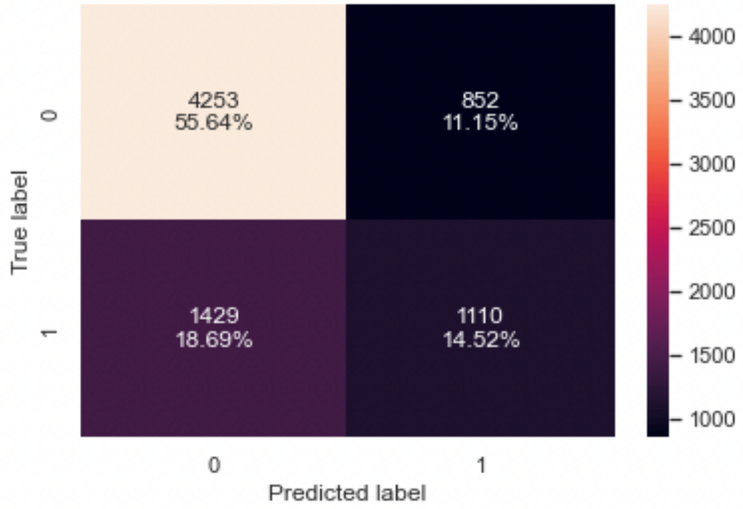
- F-score on training is 1
- F-1 score on testing is .486
- No noticiable improvement

Bagging Classifier

Before Tuning

	Accuracy	Recall	Precision	F1
0	0.974714	0.936519	0.98666	0.960935

	Accuracy	Recall	Precision	F1
0	0.701596	0.43718	0.565749	0.493224

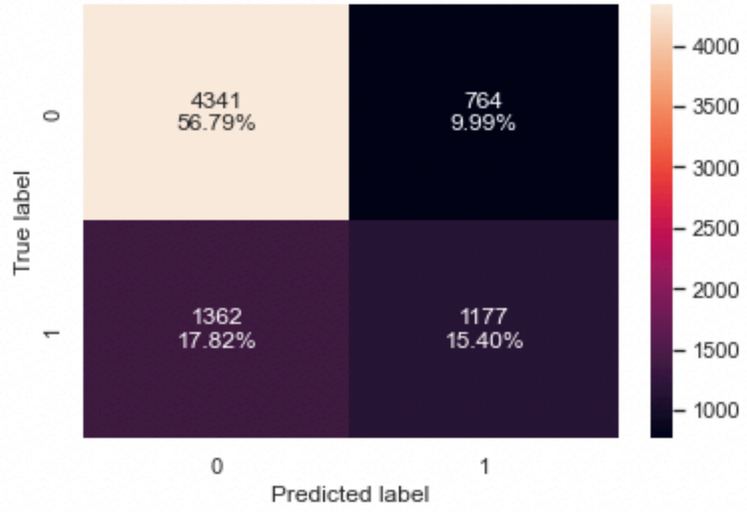


- Bagging classifier is still overfitting the training data.

Tuning Complete

	Accuracy	Recall	Precision	F1
0	0.994281	0.984974	0.997777	0.991334

	Accuracy	Recall	Precision	F1
0	0.721873	0.463568	0.606388	0.525446



- In this bagging model, hyperparameter tuning shows increases in f1 for both training and testing.

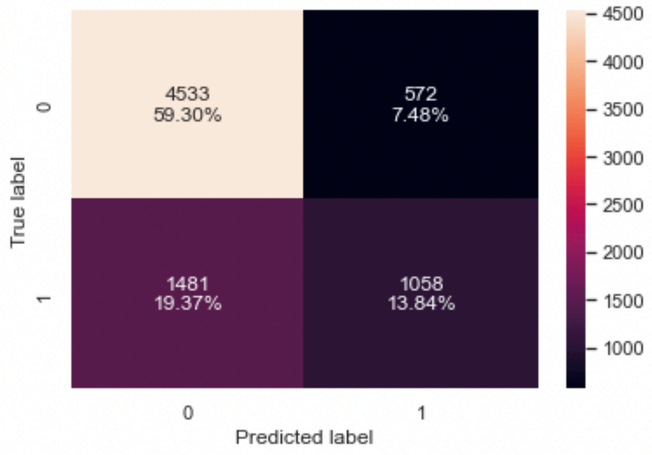
- F-1 scores on tuned model are improved compared to before tuning.
- Both these models appear to overfit the data.

Ada Boost Classifier

Before Tuning

	Accuracy	Recall	Precision	F1
0	0.733292	0.425967	0.650258	0.51474

	Accuracy	Recall	Precision	F1
0	0.731423	0.416699	0.64908	0.507556

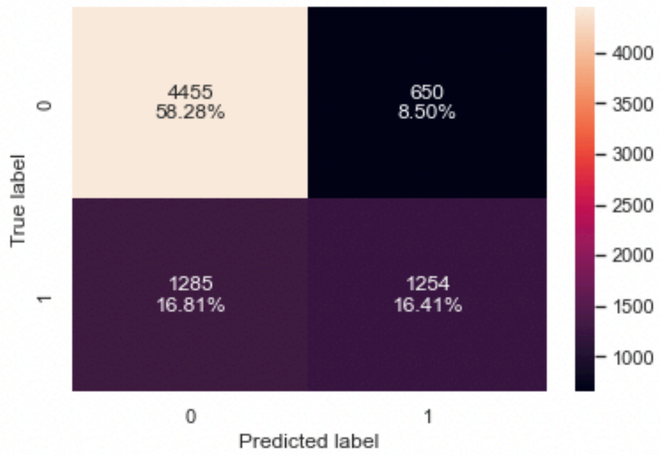


- Adaboost is giving more generalized performance than previous models but the test f1-score is too low.

Tuning Complete

	Accuracy	Recall	Precision	F1
0	0.75	0.500253	0.664052	0.570631

	Accuracy	Recall	Precision	F1
0	0.74686	0.493895	0.658613	0.564483



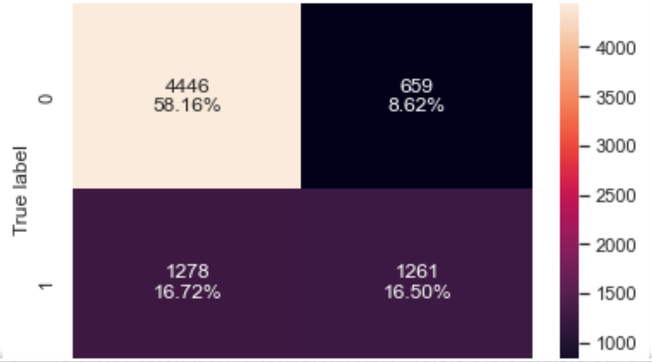
- The model performance has increased slightly but is only slightly better than random guessing.

ar quite low; both before tuning and after tuning, the Ada Boost model is
an random guessing

Gradient Boost

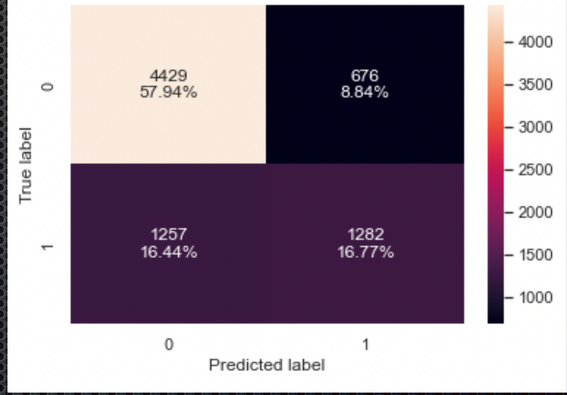
Before Tuning

```
Training performance:
    Accuracy    Recall  Precision        F1
0  0.750336  0.500929   0.664651  0.571291
Testing performance:
    Accuracy    Recall  Precision        F1
0  0.746599  0.496652   0.656771  0.565598
```



Tuning Complete

```
Training performance:
    Accuracy    Recall  Precision        F1
0  0.753869  0.513929   0.668277  0.581027
Testing performance:
    Accuracy    Recall  Precision        F1
0  0.747122  0.504923    0.65475  0.570158
```

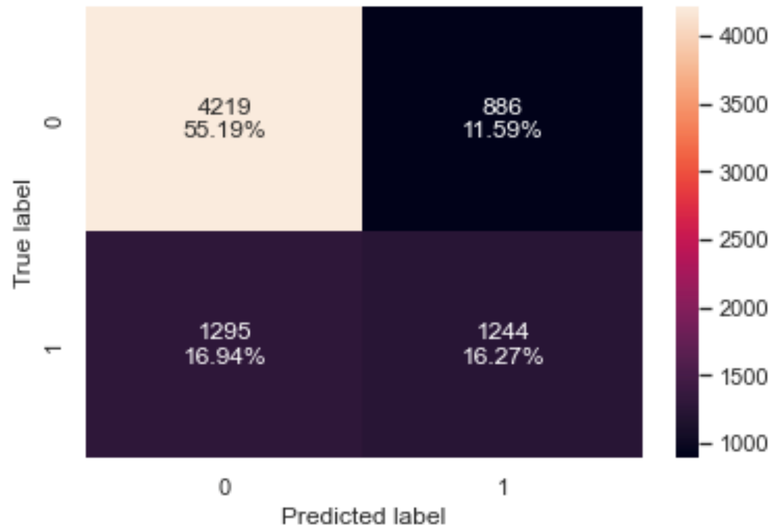


- There is one hundredth of improvement against training data (.01)
- There is .005 of improvement on testing data for F-score.
- Gradient Boost is better than all preceding models because it does not appear to overfit, but scores most highly on F-score.

Random Forest Classifier

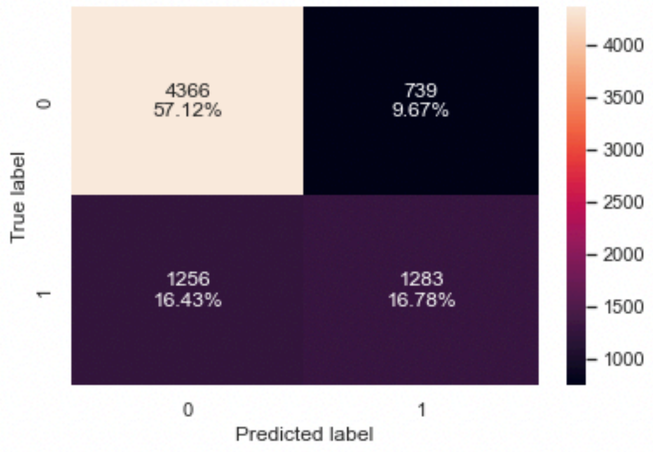
Before Tuning

```
Training performance:
    Accuracy    Recall  Precision        F1
0  0.999888  0.999662        1.0  0.999831
Testing performance:
    Accuracy    Recall  Precision        F1
0  0.714678  0.489957   0.584038  0.532876
```



Tuning Complete

```
Training performance:
    Accuracy    Recall  Precision        F1
0  0.841612  0.660476   0.827761  0.734717
Testing performance:
    Accuracy    Recall  Precision        F1
0  0.739011  0.505317    0.63452  0.562596
```



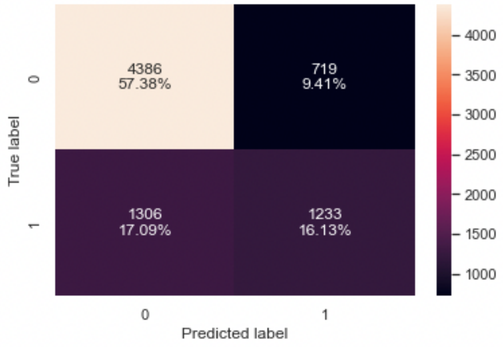
- The overfitting has reduced significantly and the model performance has improved.
- The test recall and test f1-score have increased.

- Random forest is giving a slightly higher test f1-score than decision trees but it is overfitting the training data (F1=.99916 and .734).

XGBoost

Before Tuning

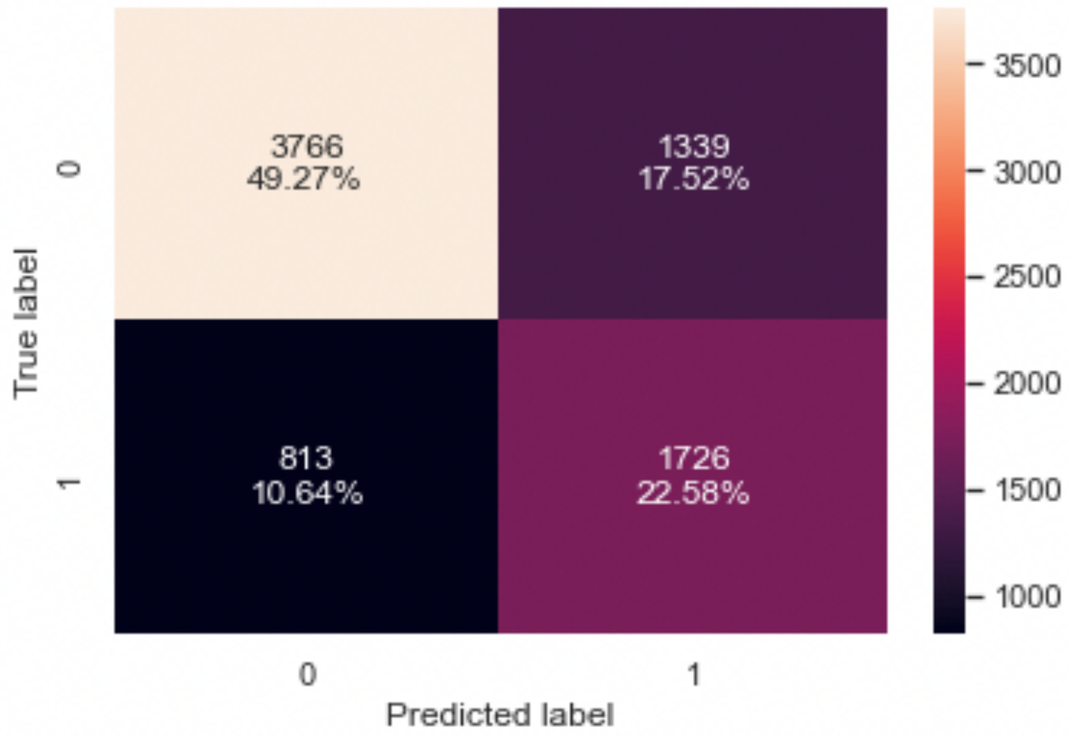
```
Training performance:
    Accuracy    Recall  Precision       F1
0  0.817896  0.613203   0.791458  0.69102
Testing performance:
    Accuracy    Recall  Precision        F1
0  0.735086  0.485624    0.63166  0.549098
```



- xgboost classifier is overfitting the training data.

Tuning Complete

```
Training performance:
    Accuracy    Recall  Precision        F1
0   0.73817  0.706568   0.588029  0.641871
Testing performance:
    Accuracy    Recall  Precision        F1
0  0.718472  0.679795   0.563132  0.615989
```

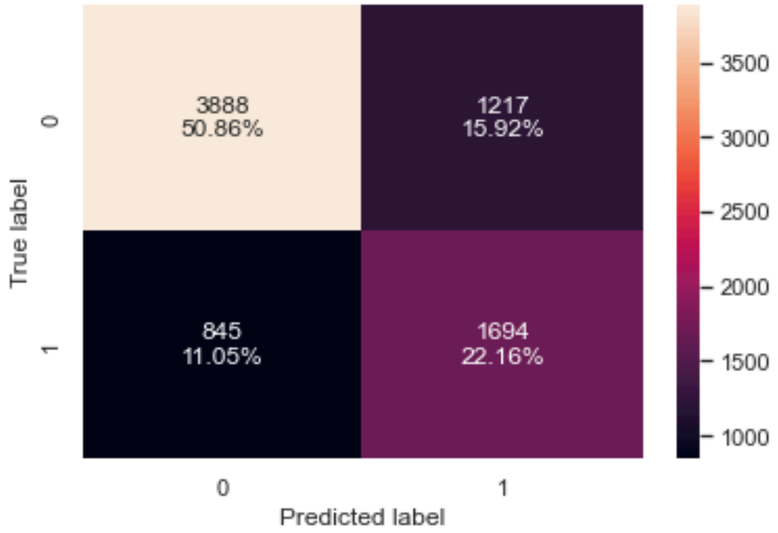


- Even before tuning, XGBoost is overfitting the training data.
- Before tuning, XGBoost does not have a remarkable F-score (.549).
- After tuning, XGBoost is the best model, with .615 F-score.

Stacking Classifier

- Stacking classifier is a combination of the three strongest weak learners.
- Random Forest
- Gradient Boost
- Ada Boost
- It rendered an F-score of .62.
- This is the best overall predictor of Visa certification.

```
Training performance:
    Accuracy    Recall  Precision        F1
0  0.768894  0.730542   0.631402  0.677364
Testing performance:
    Accuracy    Recall  Precision        F1
0  0.730246  0.667192   0.581931  0.621651
```



- The stacking classifier has the highest overall F-score (.62).
- The stacking classifier is giving a similar performance as compared to XGBoost with slightly less overfitting.

Stacking and Feature Selection

The best weak learners are: Ada Boost Tuned, Gradient Boost Tuned, XG Boost Tuned (F-score: .61), and Random Forest Tuned.

Stacking of best weak learners may improve results. We used Ada Boost Tuned, Gradient Boost Tuned and Random Forest Tuned to train XG Boost Tuned for the final model selection.

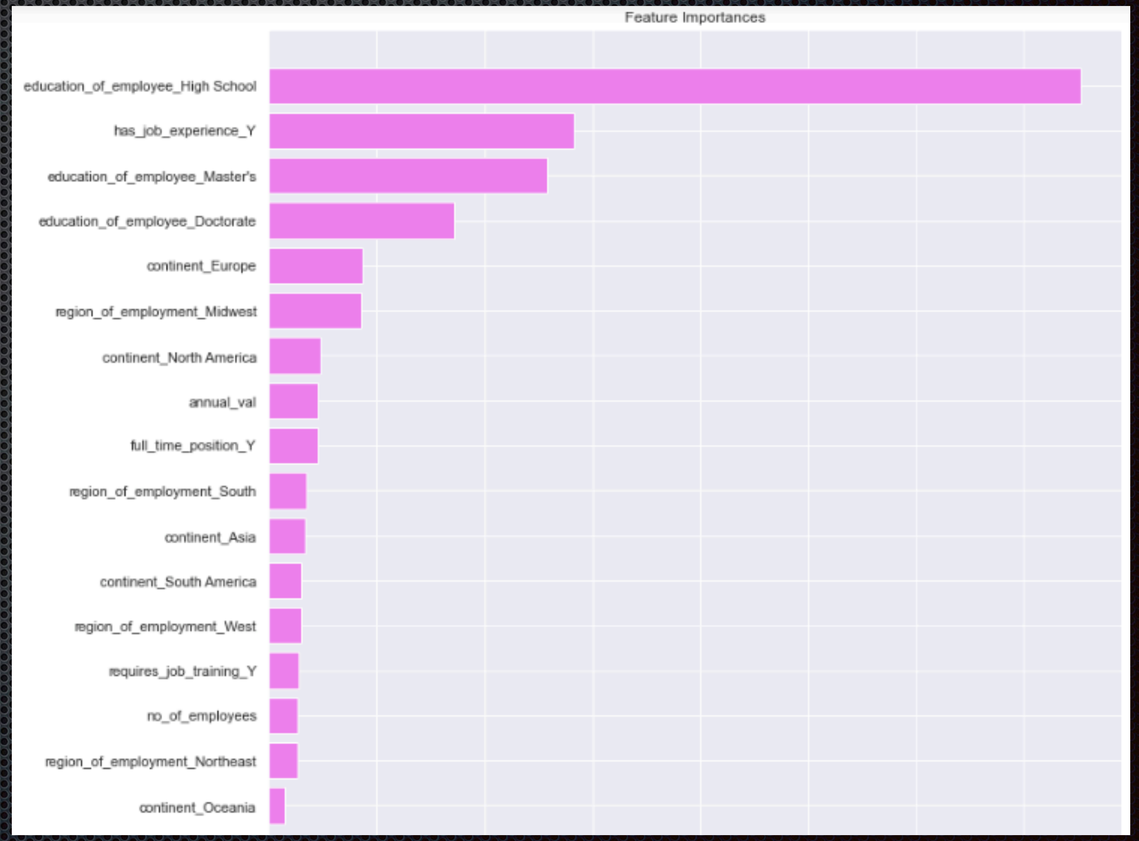
Stacking the best models renders the best F-score (.62).

By using the final model, OFLC can expedite selection of likely candidates to certify and receive Visas.

This is objective 1: to develop a machine learning model to facilitate the process of Visa approval.

Objective 2: the profile for applicants consists of the following features:

- Employee will be educated; Doctorate>Master's>Bachelor's>High School.
- Employee will have job experience.



Recommendations

Based on our analysis, we can say that the employee most likely to be certified has:

- Has higher education. The education measures of highschool, Bachelor's, Master's and Doctorate are all significant features and all influence selectivity.
- Has job experience.

- This model reinforces continued selection of the same type of candidates that were previously selected; it would be better if there was some measure of employer satisfaction or successful completion of one year of employment. Instead this model reinforces selection criteria that led to previous years selection, but does not measure value as a visa holder.
- Although the model can offer some guidance on what features are significant in predicting certification of a Visa for an employee desiring to work, the overall labor pool was short by 923,978 people. Thus, a better model would be to approve all (100%) appliants unless they fail a background check for felony in their country of origin.