Office of Foreign Labour Certification

Data based prediction model : Predict if a prospective applicant will be granted visa

- by EasyVisa

Background

1. In FY 2016, the Office of Foreign Labor Certification (OFLC) processed 775,979 employer applications for 1,699,957 positions for temporary and permanent labor certifications.

2. This was a nine percent increase in the overall number of processed applications from the previous year. The process of reviewing every case is becoming a tedious task as the number of applicants is increasing every year.

3. 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

U.S. Citizenship and Immigration Services
U.S. DEPARTMENT OF HOMELAND SECURITY

Business Problem Overview and Solution Approach

Business communities in the United States are facing high demand for human resources, but one of the constant challenges is identifying and attracting the right talent, which is perhaps the most important element in remaining competitive

- OFLC processes job certification applications for employers seeking to bring foreign workers into the United States and grants certifications in those cases where employers can demonstrate that there are not sufficient US workers available to perform the work at wages that meet or exceed the wage paid for the occupation in the area of intended employment.

- **Objective**

- The increasing number of applicants every year calls for a Machine Learning based solution that can help in shortlisting the candidates having higher chances of VISA approval.
- OFLC is interested in a data driven solution to analyze the data provided and, with the help of a classification model:

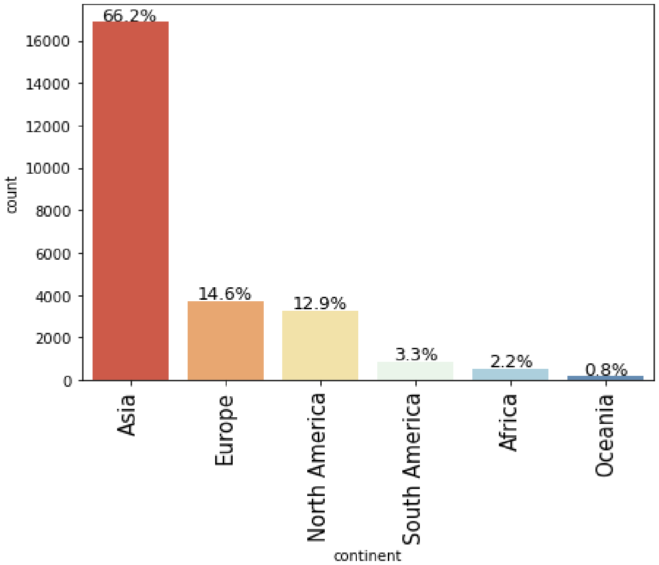
Data Overview

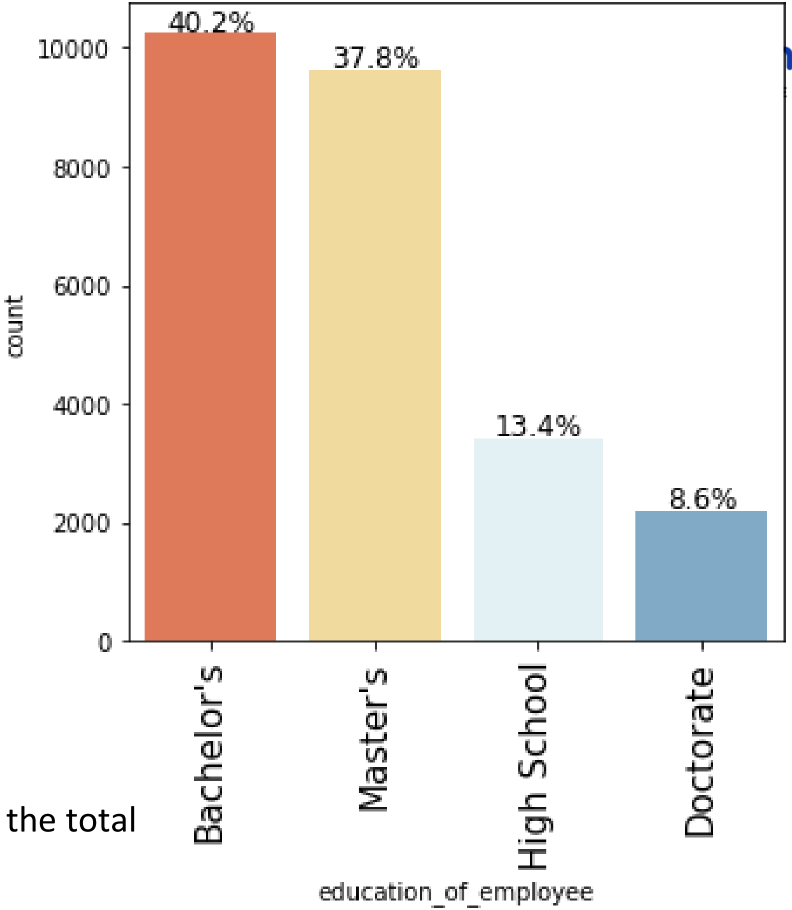
Variable	Description
case_id	ID of each visa application
continent	Information of continent the employee
education_of_employee	Information of education of the employee
has_job_experience	Does the employee has any job experience? Y= Yes; N = No
requires_job_training	Does the employee require any job training? 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.

Variable	Description
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

Observations	Variables
25480	12

Exploratory Data Analysis

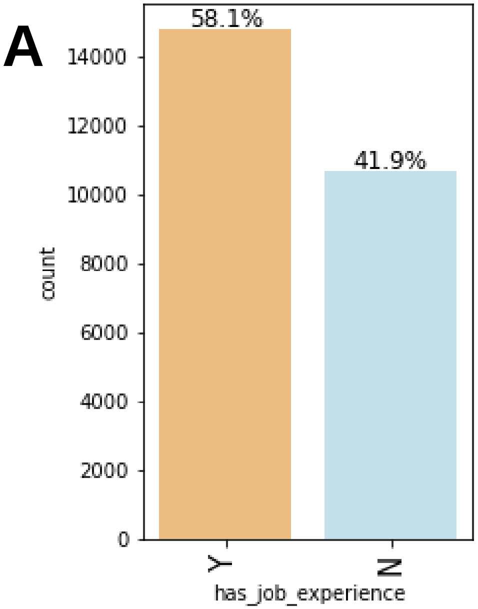


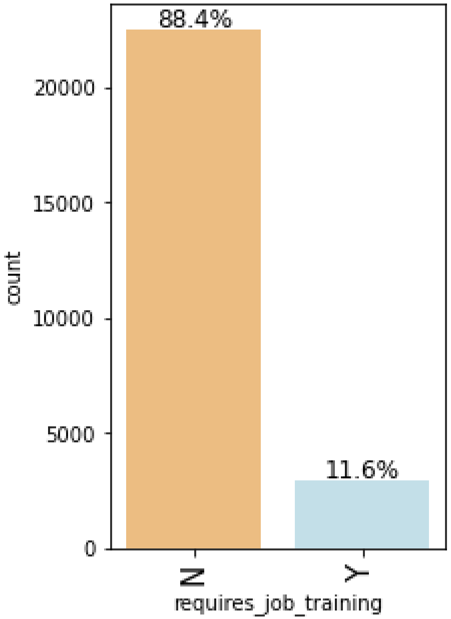


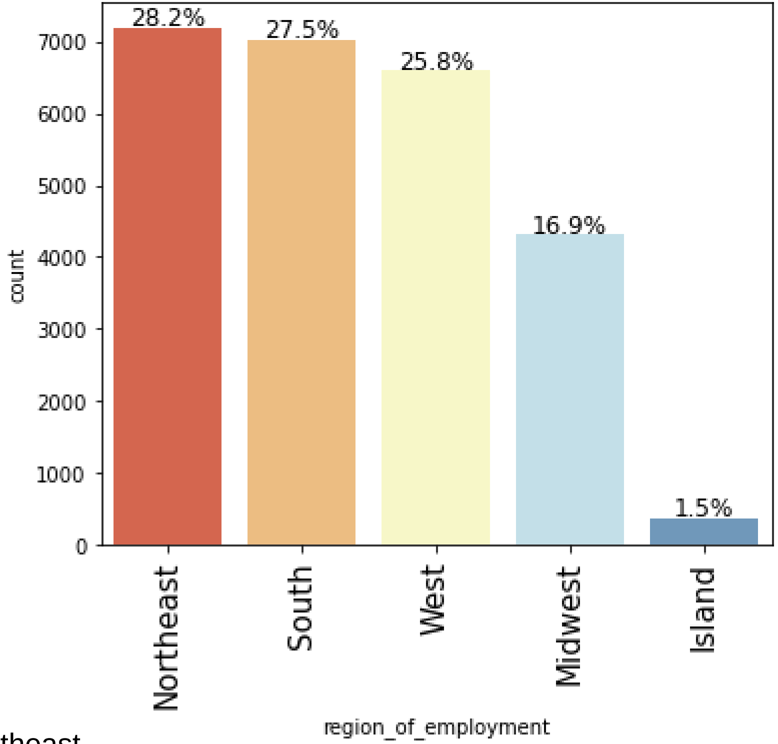
- Asia has about 66 percent of total applications.
- Europe and North America have about 14.6 and 13 percent of the total applications.

•Nearly 40% of the applicants have Bachelor's and about 39% of them have Masters.

EDA





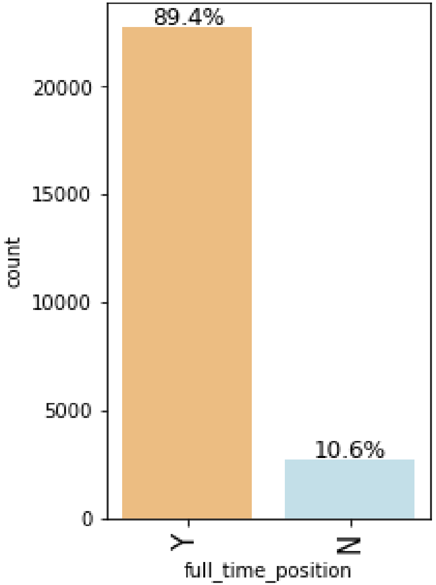


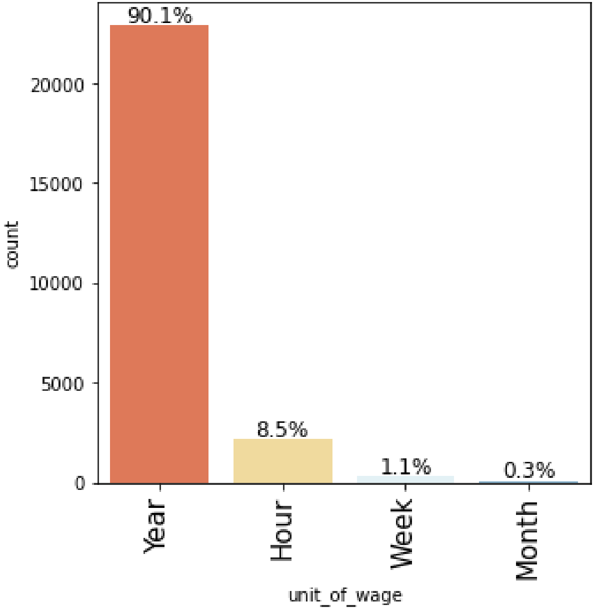
- 58% of the applicants for visa have job experience.

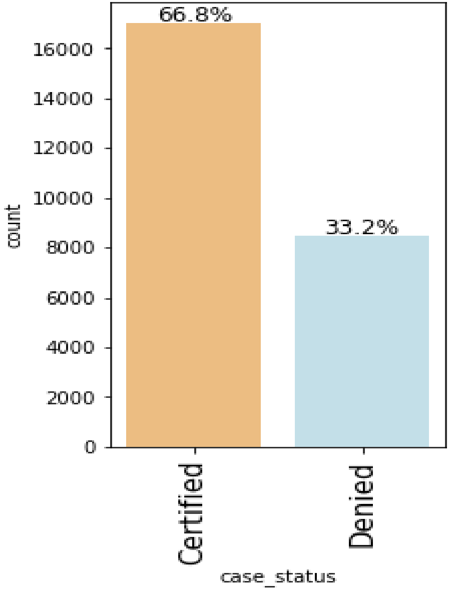
- 88% of then don't need job training.

- The intended region of employment is almost equally distributed for Northeast, South, and West with 28%, 27.5% and 25.8%.

EDA





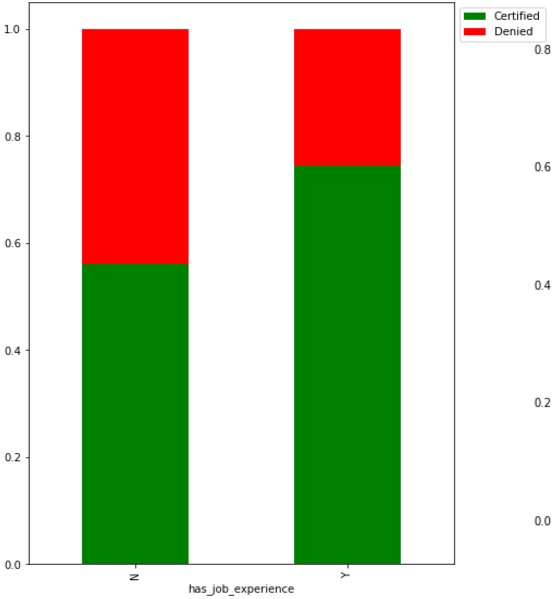


- Almost 89% of the applicants, apply for full time positions.

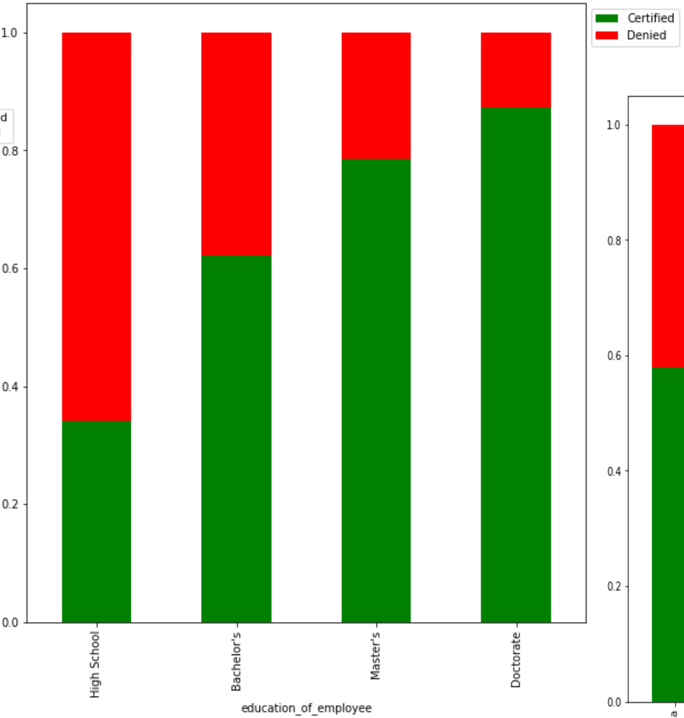
- Most of the job positions prevailing_wage are expressed in years.

- 66.8% of the applicants are successful in getting the Visa.

EDA

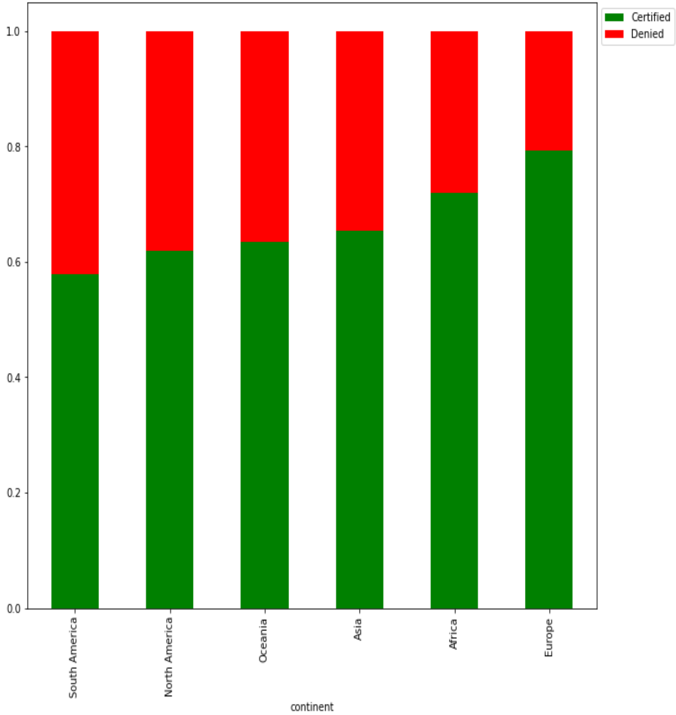


- Employees having prior job experience have higher rate of approval.



•Higher the education of the applicant the higher the chance of approval.

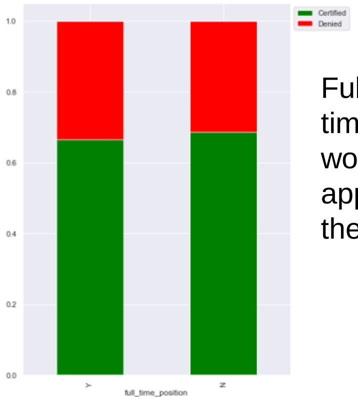
•Doctorate's have almost 90% approval and Master's have roughly 80% approval rate.



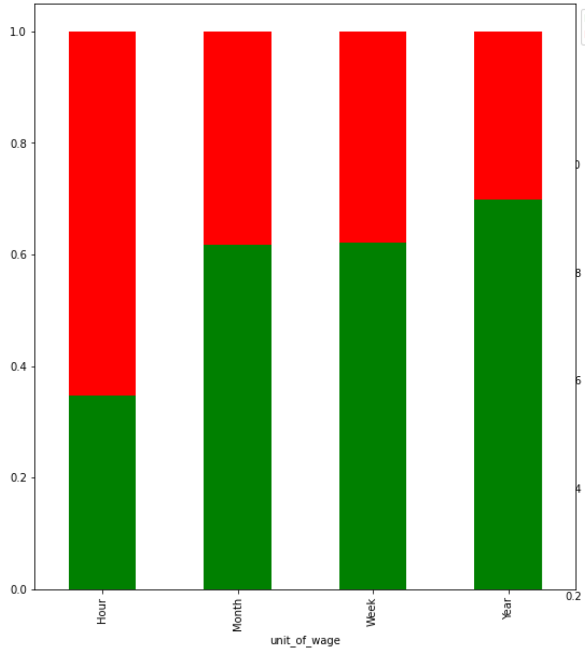
• Applicants from Europe have the highest rate of approval for Visa.

•South America have the least rate of approval.

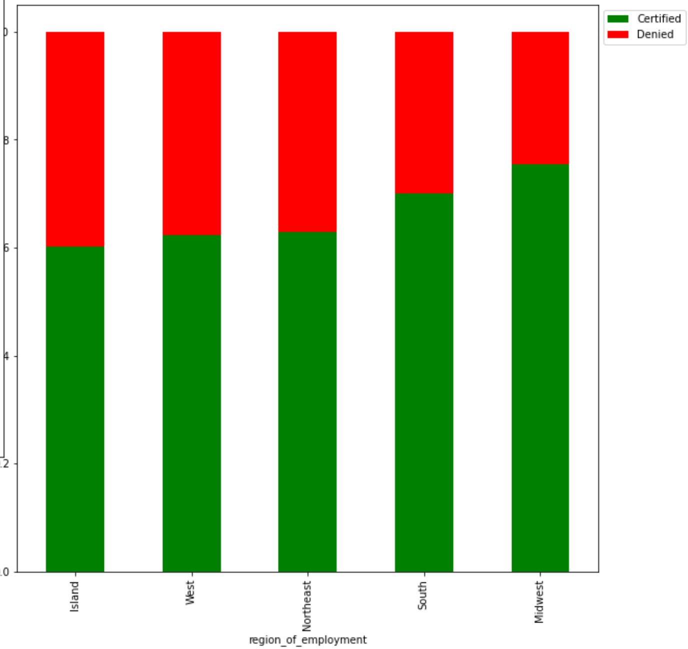
EDA – Contd.



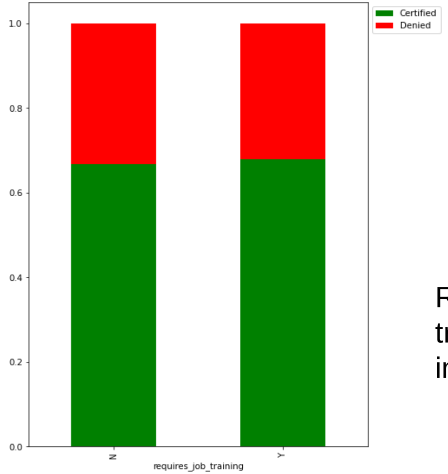
Full time or part time nature of the work does not appear to impact the case status.



Applicants who have the unit_of_wage for prevailing_wage as year have the better chance of approval.

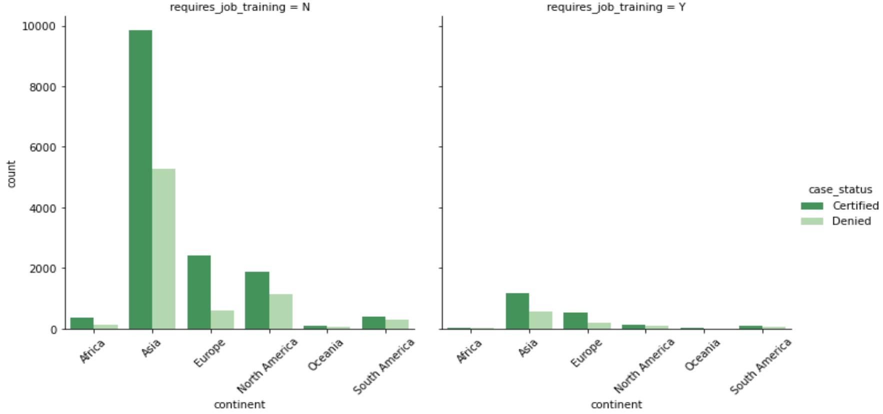


- Midwest has the higher rate of Visa approval.
- The region of intended employment does not seem to impact much, all regions have almost similar rate of approval.

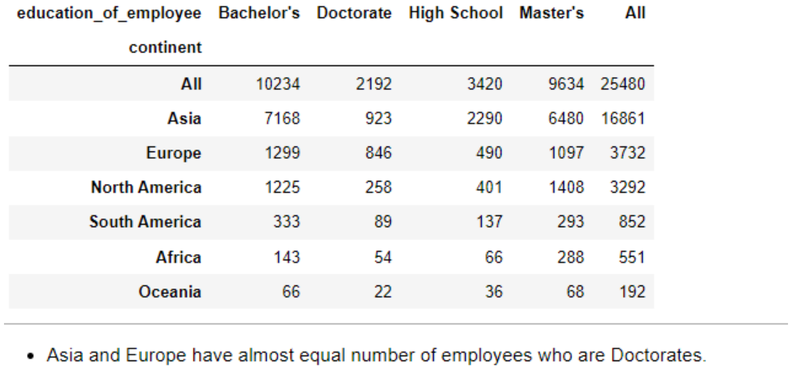


Requirement for employee's Job training does not seem to be impacting the Visa application status.

EDA – Contd.



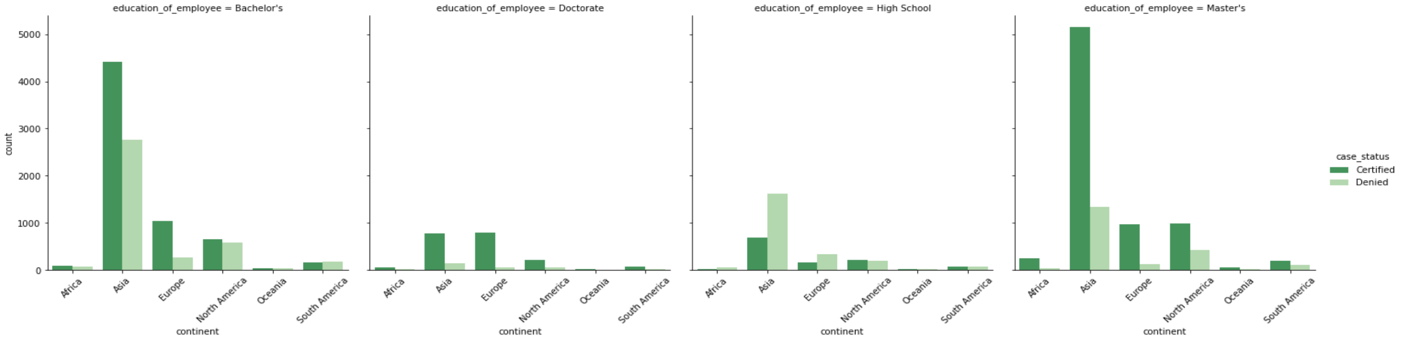
- Asia appears to widely approved Visa category, but there could be other factors like education



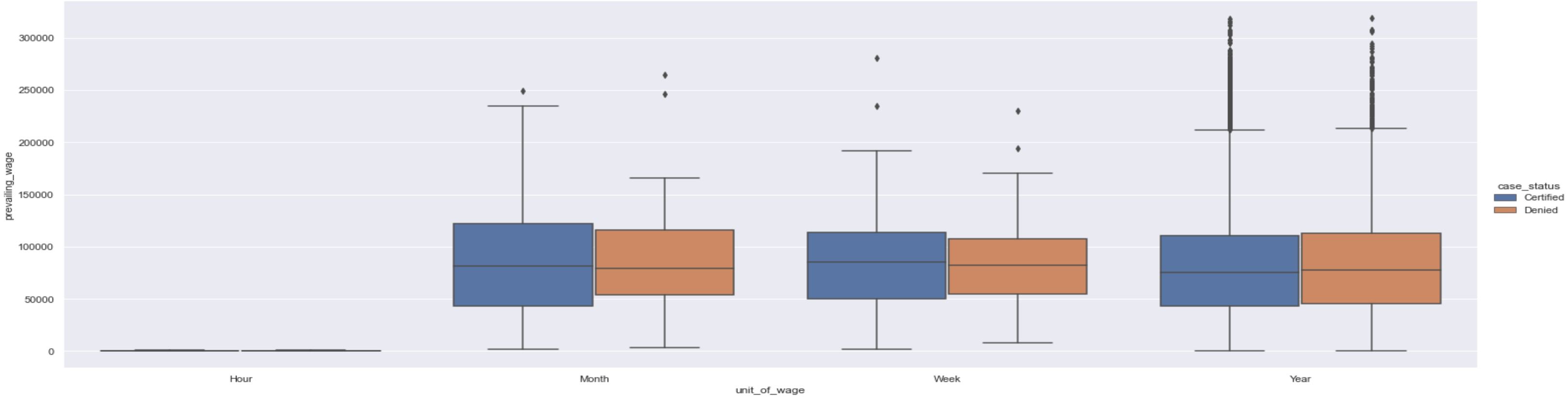
education_of_employee	Bachelor's	Doctorate	High School	Master's	All
continent					
All	10234	2192	3420	9634	25480
Asia	7168	923	2290	6480	16861
Europe	1299	846	490	1097	3732
North America	1225	258	401	1408	3292
South America	333	89	137	293	852
Africa	143	54	66	288	551
Oceania	66	22	36	68	192

- Asia and Europe have almost equal number of employees who are Doctorates.

Applicants with High School have the lesser rate of approval

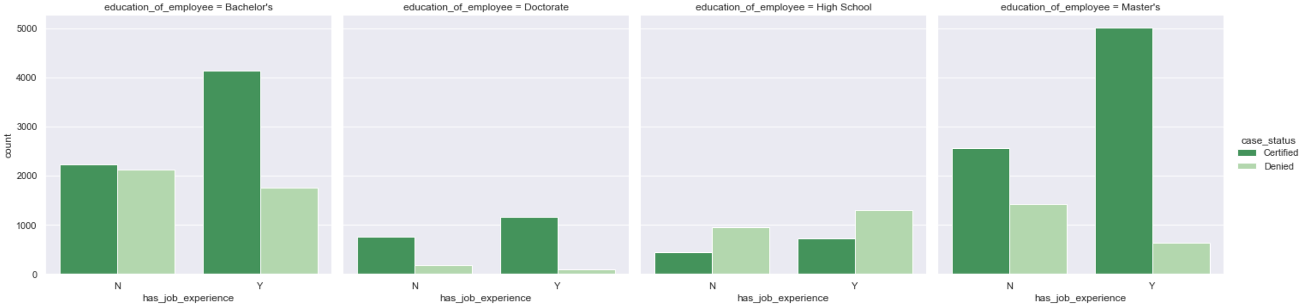


EDA – Contd.



Among the wage groups the wage does not seem to be deciding factor

Job experience seems to be a deciding factor for all education level except for Doctorate.



Key Questions:

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

- Yes, employees with higher education have better rate of approval. Especially with Doctorate.

2. How does the visa status vary across different continents?

- Yes, Europe seems to have higher level of approvals, this is also due to higher level of education among Europe applicants.

3. Experienced professionals might look abroad for opportunities to improve their lifestyles and career development. Does work experience influence visa status?

- Yes, to some extent job experience influences visa status but not at a very high level as education.

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

- Year to be the most approved, with hourly being least approved, but weekly and monthly too have good rates of approval.

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

- Prevailing wage does not seem to appear impacting the visa status

Model evaluation criterion

- **Model can make wrong predictions as:**

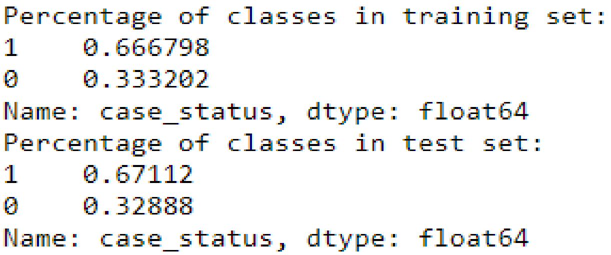
1. Predicting an employee's Visa will *not* be Certified but in reality will be Certified- Loss of potential talent (opportunity cost)
2. Predicting an employee's Visa will *be* Certified but in reality will *not* be Certified - Loss of resources

- **Which Loss is greater ?**

- Loss of potential talent (opportunity cost) is nigher a skilled talent is hard to find.
- Loss of resources is not a heavy burden as the application can be filtered in further processing.

- **How to reduce this loss i.e need to reduce False Negatives ?**
- Model should reduce false negatives, this can be done by maximizing the Recall. Greater the recall lesser the chances of false negatives.

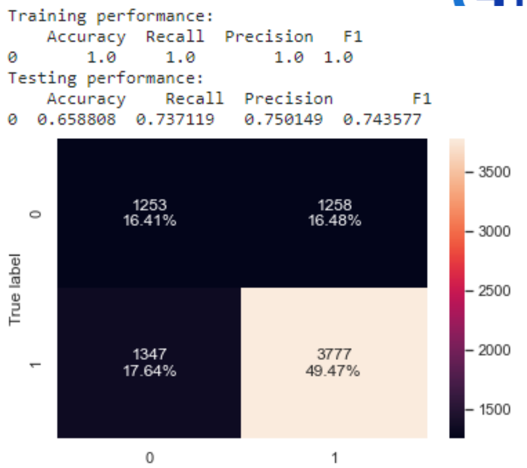
- **Positive event and Negative Event**
- Positive event is Visa certified
- Negative event is Visa is rejected



```
Percentage of classes in training set:
1    0.666798
0    0.333202
Name: case_status, dtype: float64
Percentage of classes in test set:
1    0.67112
0    0.32888
Name: case_status, dtype: float64
```

Total observations taken for model	**25,447**
Training	17,812
Test	7635

Model Building - Bagging



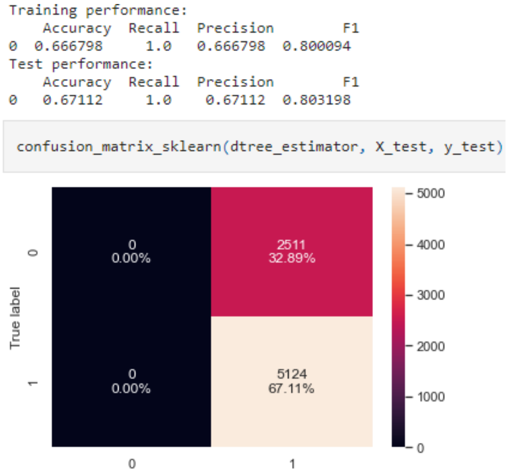
```
Training performance:
    Accuracy  Recall  Precision   F1
0        1.0     1.0        1.0  1.0
Testing performance:
    Accuracy    Recall  Precision        F1
0  0.658808  0.737119   0.750149  0.743577
```

Decision Tree

- The decision tree is overfitting the training data as there is a considerable difference between training and test scores for all the metrics.
- The test recall is 73%. We can see if other models have better recall

Decision Tree - Tuned

DecisionTreeClassifier(max_depth=2, max_leaf_nodes=2, min_impurity_decrease=0.1, min_samples_leaf=5, random_state=1)



```
Training performance:
    Accuracy  Recall  Precision        F1
0  0.666798     1.0   0.666798  0.800094
Test performance:
    Accuracy  Recall  Precision        F1
0    0.67112     1.0    0.67112  0.803198
```

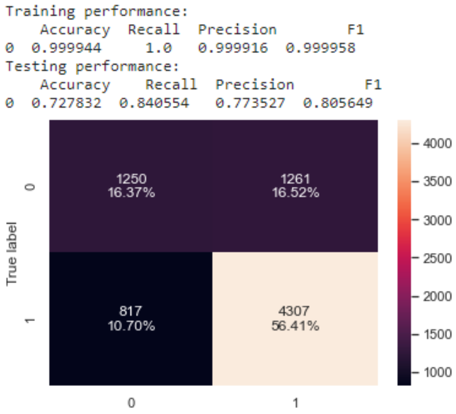
```
confusion_matrix_sklearn(dtree_estimator, X_test, y_test)
```

- The decision tree is now more generalized model with consistent results for testing and training data
- The model has a very high recall score, but the precision is only about 66 to 67%.
- The model is biased and one side.

Bagging

Random Forest

- The decision tree is overfitting the training data as there is a considerable difference between training and test scores for all the metrics.
- The test recall is 84%. Its a good improvement from decision tree

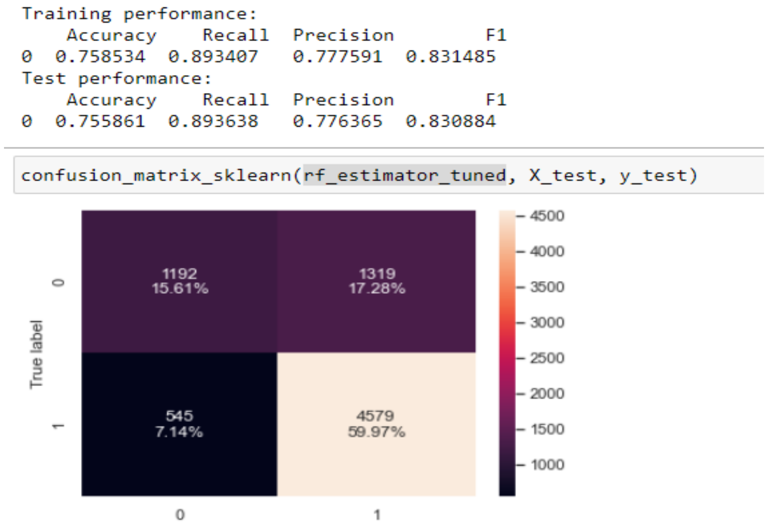


```
Training performance:
   Accuracy  Recall  Precision        F1
0  0.999944     1.0   0.999916  0.999958
Testing performance:
   Accuracy    Recall  Precision        F1
0  0.727832  0.840554   0.773527  0.805649
```

Random Forest - Tuned

RandomForestClassifier(max_features=0.2, max_samples=0.3, min_samples_leaf=9,
n_estimators=150,
random_state=1)

- The Tuned Random Forest model is giving generalized results between training and test
- The model has acceptable recall and precision scores

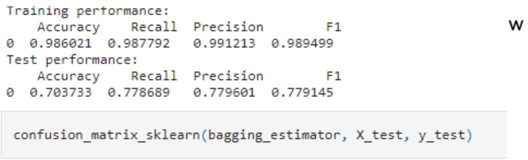


```
Training performance:
   Accuracy    Recall  Precision        F1
0  0.758534  0.893407   0.777591  0.831485
Test performance:
   Accuracy    Recall  Precision        F1
0  0.755861  0.893638   0.776365  0.830884
```

```
confusion_matrix_sklearn(rf_estimator_tuned, X_test, y_test)
```

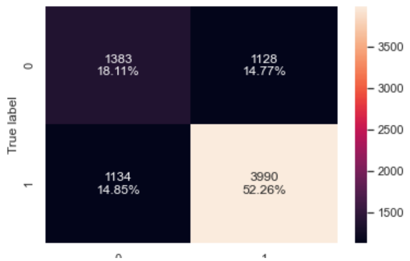
BaggingClassifier(max_features=0.7, max_samples=1, random_state=1)

Bagging



```
Training performance:
    Accuracy    Recall   Precision         F1
0  0.986021  0.987792   0.991213  0.989499
Test performance:
    Accuracy    Recall   Precision         F1
0  0.703733  0.778689   0.779601  0.779145
```

```
confusion_matrix_sklearn(bagging_estimator, X_test, y_test)
```

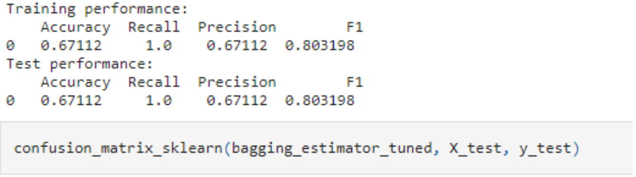


Bagging classifier

- The Bagging classifier is overfitting the training data as there is a considerable difference between training and test scores for all the metrics.
- The test recall is 77%. A slight improvement form the decision tree

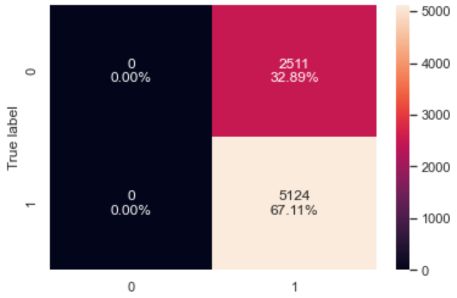
Bagging classifier - Tuned

RandomForestClassifier(max_features=0.2, max_samples=0.3, min_samples_leaf=9, n_estimators=150, random_state=1)



```
Training performance:
   Accuracy  Recall  Precision        F1
0   0.67112     1.0    0.67112  0.803198
Test performance:
   Accuracy  Recall  Precision        F1
0   0.67112     1.0    0.67112  0.803198
```

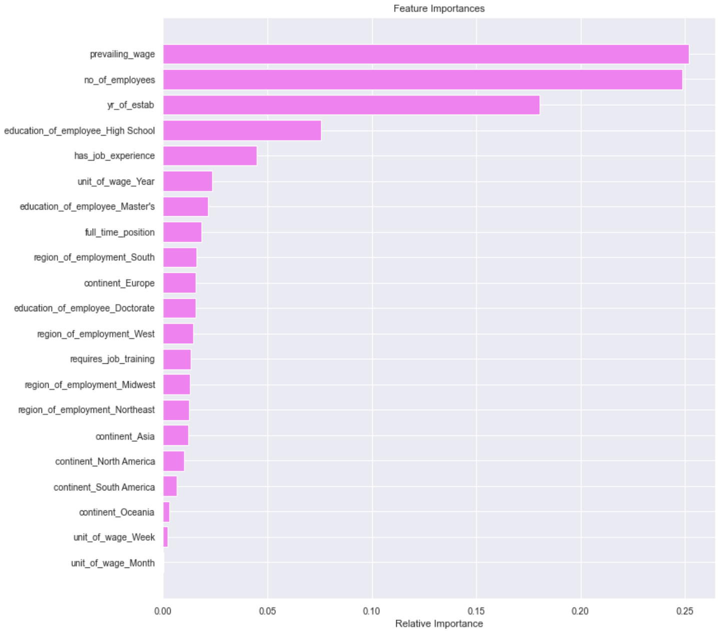
```
confusion_matrix_sklearn(bagging_estimator_tuned, X_test, y_test)
```



- The Bagging classifier is giving a generalized results now
- The model appears to predict one sided, the default bagging classifier appears to be more robust.

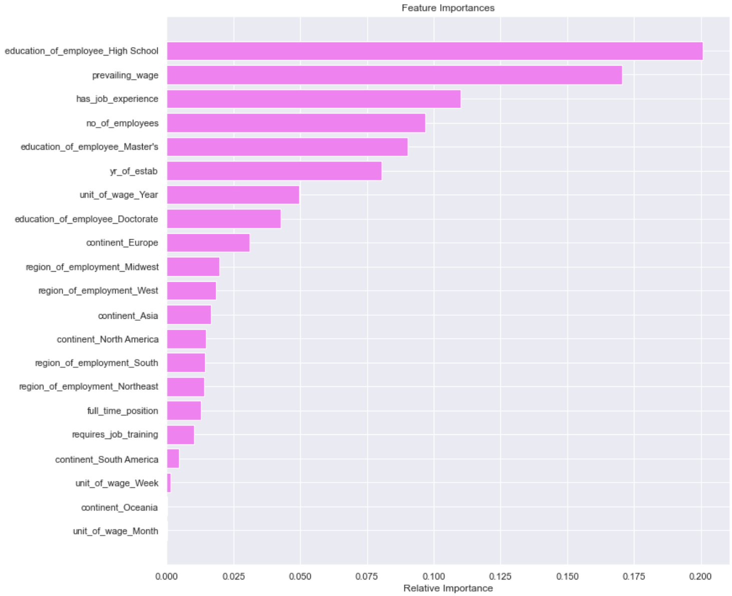
Bagging Model
Decision Tree & Tuned Random Forest – Feature of Importance

Of the bagging models, the Tuned Random Forest without specific class weights assigned is the better model so far
The features of importance from best model so far is education_of_employee_High School prevailing_wage has_job_experience



<- Decision Tree

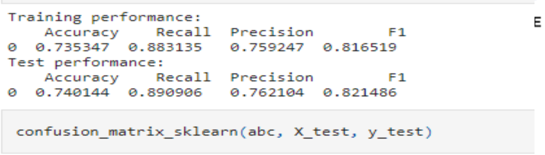
Random Forest – Tuned ->



Model Building - Boosting

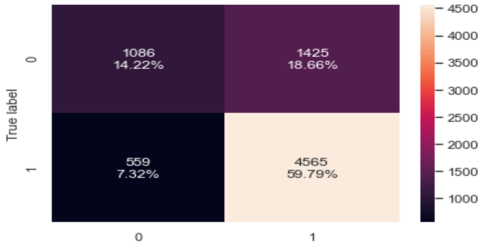
AdaBoost Classifier

- Ada boost classifier is giving comparable results between training and test.
- Model performance is comparable to Tuned Random Forest.
- The importance feature for this model is prevailing_wage and no_of_employees.



```
Training performance:
   Accuracy    Recall  Precision        F1
0  0.735347  0.883135   0.759247  0.816519
Test performance:
   Accuracy    Recall  Precision        F1
0  0.740144  0.890906   0.762104  0.821486
```

```
confusion_matrix_sklearn(abc, X_test, y_test)
```

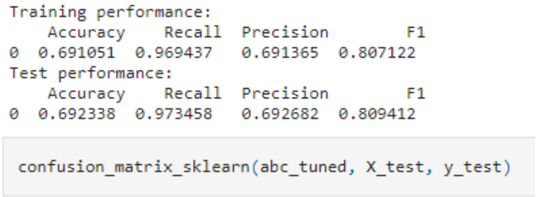


AdaBoost Classifier - Tuned

AdaBoostClassifier(base_estimator=DecisionTreeClassifier(max_depth=1,

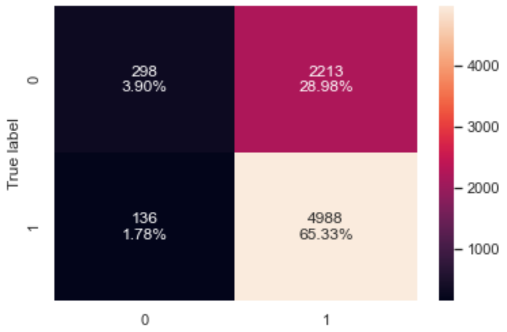
random_state=1),
learning_rate=0.2,
n_estimators=10, random_state=1)

- Tuned Ada boost classifier is giving comparable results between training and test.
- Model performance is improved in terms of recall after tuning, but dropped for precision.
- The importance feature for this model is education_of_employee_High Schoo' and has_job_experience.



```
Training performance:
   Accuracy    Recall  Precision        F1
0  0.691051  0.969437   0.691365  0.807122
Test performance:
   Accuracy    Recall  Precision        F1
0  0.692338  0.973458   0.692682  0.809412
```

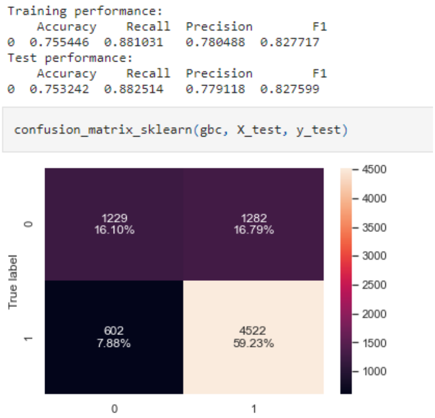
```
confusion_matrix_sklearn(abc_tuned, X_test, y_test)
```



Model Building - Boosting

Gradient Boosting Classifier

- Gradient boost classifier is giving comparable results between training and test.
- Model performance is comparable to Tuned Random Forest.
- The importance feature for this model is education_of_employee_High School and has_job_experience.

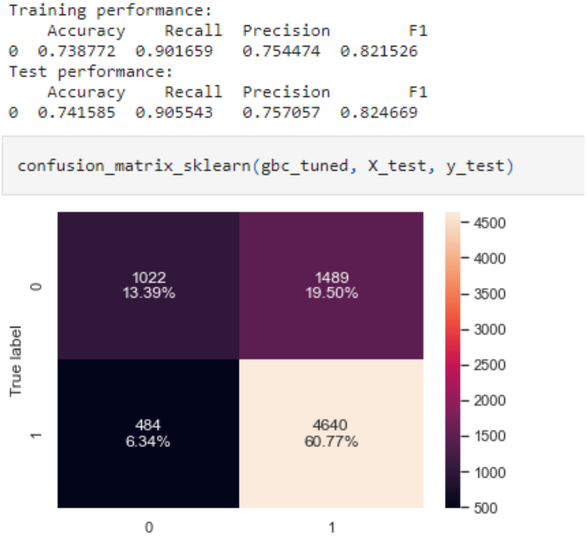


```
Training performance:
   Accuracy    Recall  Precision        F1
0  0.755446  0.881031   0.780488  0.827717
Test performance:
   Accuracy    Recall  Precision        F1
0  0.753242  0.882514   0.779118  0.827599
```

```
confusion_matrix_sklearn(gbc, X_test, y_test)
```

Gradient Boosting Classifier - Tuned

GradientBoostingClassifier(init=AdaBoostClassifier(random_state=1),
                           max_features=1,
random_state=1, subsample=1)



```
Training performance:
   Accuracy    Recall  Precision        F1
0  0.738772  0.901659   0.754474  0.821526
Test performance:
   Accuracy    Recall  Precision        F1
0  0.741585  0.905543   0.757057  0.824669
```

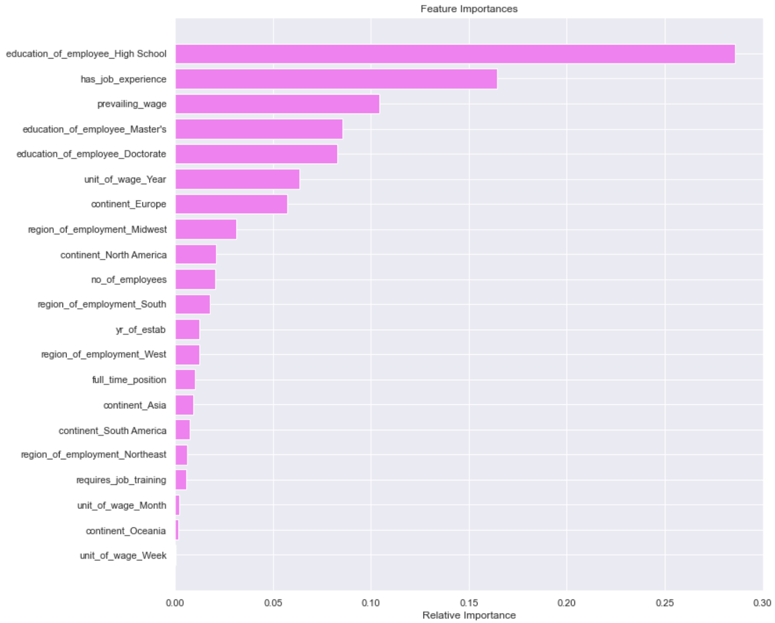
```
confusion_matrix_sklearn(gbc_tuned, X_test, y_test)
```

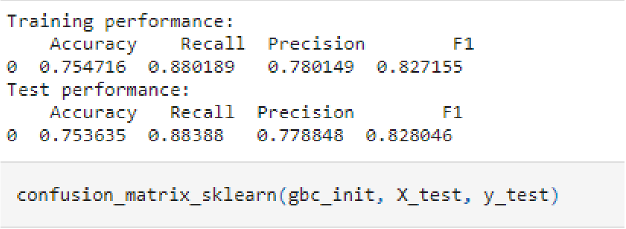
- Tuned Ada boost classifier is giving comparable results between training and test.
- Model performance is improved in terms of recall after tuning, but dropped for precision.
- The importance feature for this model is education_of_employee_High Schoo' and has_job_experience.

Model Building - Boosting

Gradient Boosting Classifier – With AdaBoost classifier as the estimator for initial predictions

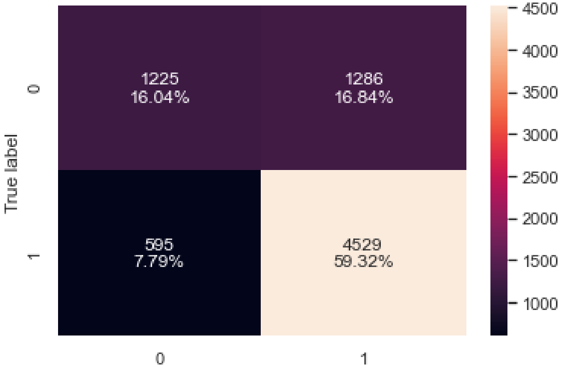
•Gradient boost classifier with Ada boost is giving comparable results between training and test.
•Model performance is comparable to Gradient boost classifier with default.
•The importance feature for this model is education_of_employee_High School and has_job_experience.





```
Training performance:
    Accuracy    Recall  Precision        F1
0  0.754716  0.880189   0.780149  0.827155
Test performance:
    Accuracy   Recall  Precision        F1
0  0.753635  0.88388  0.778848  0.828046
```

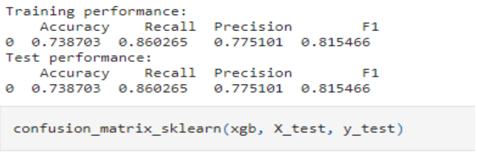
```
confusion_matrix_sklearn(gbc_init, X_test, y_test)
```



Model Building - Boosting

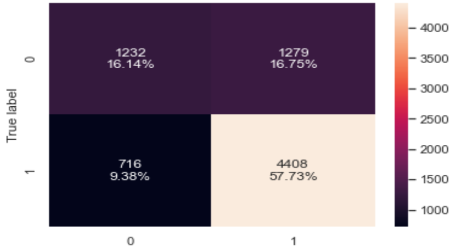
XGBoost Classifier

- XG boost classifier is giving comparable results between training and test.
- Model performance is comparable to Ada boost, Gradient boost and Tuned Random Forest.
- The importance feature for this model is education_of_employee_High School and education_of_employee_Doctorate.



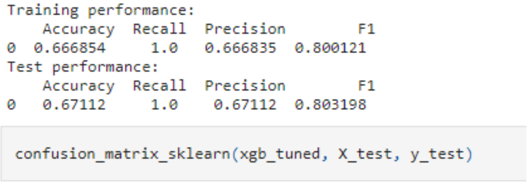
```
Training performance:
    Accuracy    Recall  Precision        F1
0  0.738703  0.860265   0.775101  0.815466
Test performance:
    Accuracy    Recall  Precision        F1
0  0.738703  0.860265   0.775101  0.815466
```

```
confusion_matrix_sklearn(xgb, X_test, y_test)
```



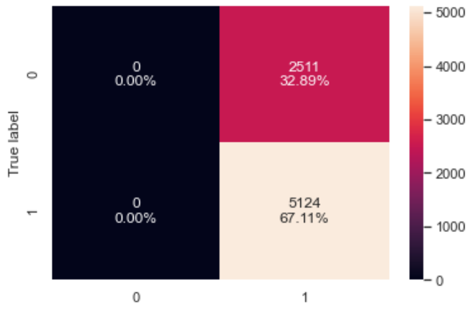
XGBoost Classifier - Tuned

XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=0.5,
colsample_bynode=1, colsample_bytree=0.5, eval_metric='logloss',
gamma=0, gpu_id=-1, importance_type='gain',
interaction_constraints='', learning_rate=0.01, max_delta_step=0,
max_depth=6, min_child_weight=1, missing=nan,
monotone_constraints='()', n_estimators=30, n_jobs=8,
num_parallel_tree=1, random_state=1, reg_alpha=0, reg_lambda=1,
scale_pos_weight=5, subsample=0.5, tree_method='exact',
validate_parameters=1, verbosity=None)



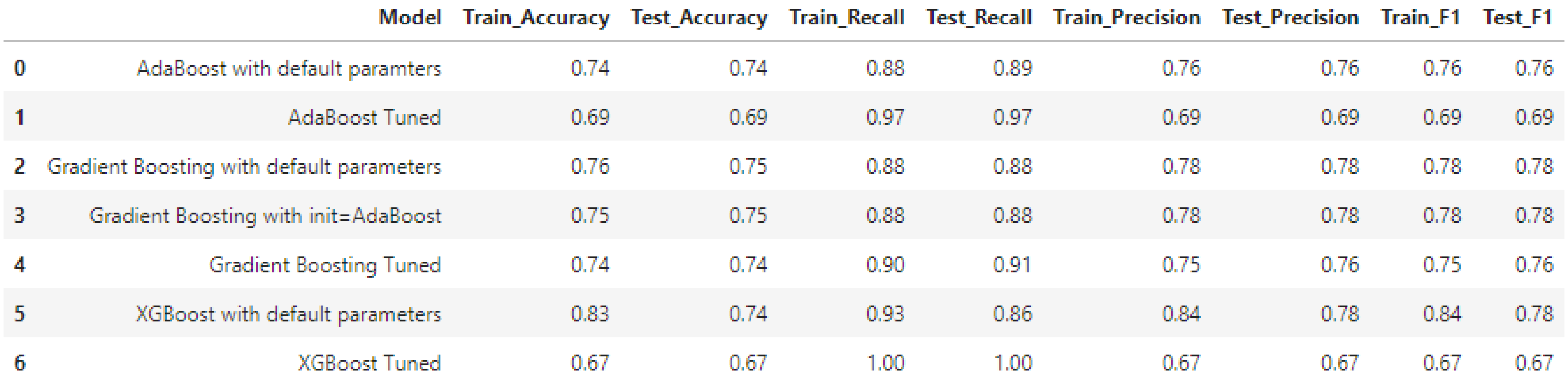
```
Training performance:
    Accuracy  Recall  Precision        F1
0   0.666854     1.0   0.666835  0.800121
Test performance:
    Accuracy  Recall  Precision        F1
0    0.67112     1.0    0.67112  0.803198
```

```
confusion_matrix_sklearn(xgb_tuned, X_test, y_test)
```



- Tuned XG boost classifier is giving comparable results between training and test.
- Model performance is has not improved on tuning and the model is biased.
- The importance feature for this model is education_of_employee_High School and education_of_employee_Master's.

Boosting - Comparing all models



	Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision	Train_F1	Test_F1
0	AdaBoost with default paramters	0.74	0.74	0.88	0.89	0.76	0.76	0.76	0.76
1	AdaBoost Tuned	0.69	0.69	0.97	0.97	0.69	0.69	0.69	0.69
2	Gradient Boosting with default parameters	0.76	0.75	0.88	0.88	0.78	0.78	0.78	0.78
3	Gradient Boosting with init=AdaBoost	0.75	0.75	0.88	0.88	0.78	0.78	0.78	0.78
4	Gradient Boosting Tuned	0.74	0.74	0.90	0.91	0.75	0.76	0.75	0.76
5	XGBoost with default parameters	0.83	0.74	0.93	0.86	0.84	0.78	0.84	0.78
6	XGBoost Tuned	0.67	0.67	1.00	1.00	0.67	0.67	0.67	0.67

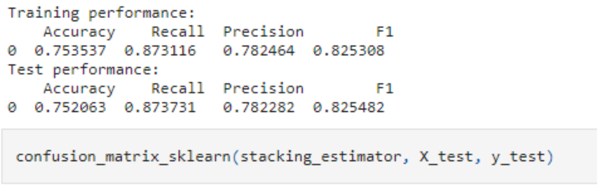
- Gradient Boosting with default parameters and Gradient Boosting with init=AdaBoost are more banaced models

Stacking Model

Stacking Classifier

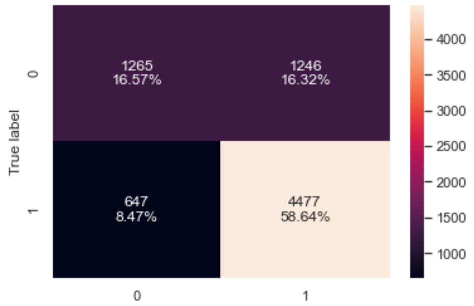
Stacking model with decision tree and tuned random forest, and gradient boosting, then use XGBoost to get the final prediction.

- The model appears to be generalized with comparable results in training and test.



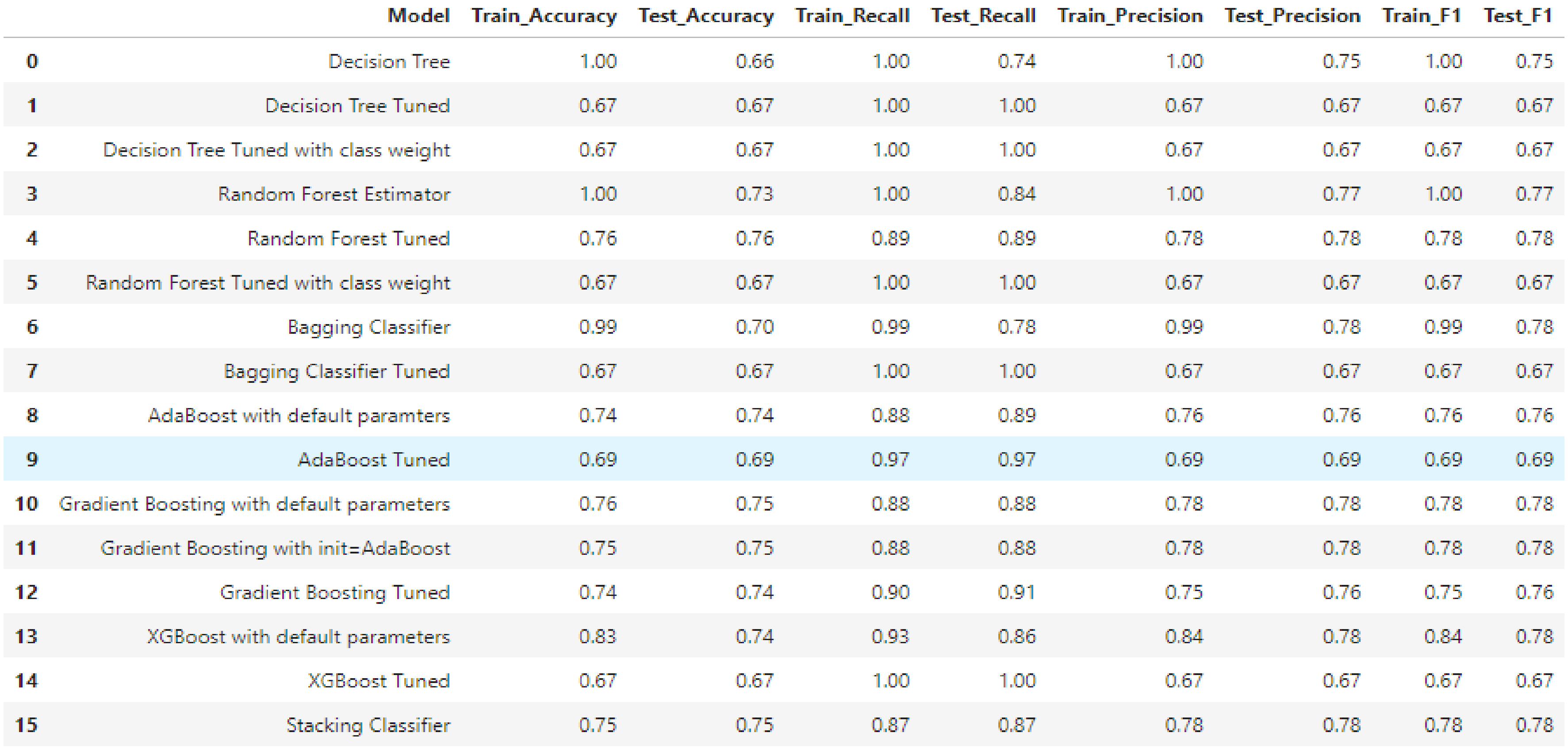
```
Training performance:
    Accuracy    Recall  Precision        F1
0  0.753537  0.873116   0.782464  0.825308
Test performance:
    Accuracy    Recall  Precision        F1
0  0.752063  0.873731   0.782282  0.825482
```

```
confusion_matrix_sklearn(stacking_estimator, X_test, y_test)
```



```
StackingClassifier(cv=5,
            estimators=[('Decision Tree',
                    DecisionTreeClassifier(class_weight={0: 0.665,
                                                    1: 0.335},
                            max_depth=2,
                            max_leaf_nodes=2,
                            min_impurity_decrease=0.1,
                            min_samples_leaf=5,
                            random_state=1)),
                   ('Random Forest',
                    RandomForestClassifier(max_features=0.2,
                            max_samples=0.3,
                            min_samples_leaf=9,
                            n_estimators=150,
                            random_state=1)),
                   ('Gradient Boosting',
                    Gradient...
                        importance_type='gain',
                        interaction_constraints=None,
                        learning_rate=None,
                        max_delta_step=None,
                        max_depth=None,
                        min_child_weight=None,
                        missing=nan,
                        monotone_constraints=None,
                        n_estimators=100, n_jobs=None,
                        num_parallel_tree=None,
                        random_state=1, reg_alpha=None,
                        reg_lambda=None,
                        scale_pos_weight=None,
                        subsample=None,
                        tree_method=None,
                        validate_parameters=None,
                        verbosity=None))
```

Bagging, Boosting & Stacking - Comparing all models



	Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision	Train_F1	Test_F1
0	Decision Tree	1.00	0.66	1.00	0.74	1.00	0.75	1.00	0.75
1	Decision Tree Tuned	0.67	0.67	1.00	1.00	0.67	0.67	0.67	0.67
2	Decision Tree Tuned with class weight	0.67	0.67	1.00	1.00	0.67	0.67	0.67	0.67
3	Random Forest Estimator	1.00	0.73	1.00	0.84	1.00	0.77	1.00	0.77
4	Random Forest Tuned	0.76	0.76	0.89	0.89	0.78	0.78	0.78	0.78
5	Random Forest Tuned with class weight	0.67	0.67	1.00	1.00	0.67	0.67	0.67	0.67
6	Bagging Classifier	0.99	0.70	0.99	0.78	0.99	0.78	0.99	0.78
7	Bagging Classifier Tuned	0.67	0.67	1.00	1.00	0.67	0.67	0.67	0.67
8	AdaBoost with default paramters	0.74	0.74	0.88	0.89	0.76	0.76	0.76	0.76
9	AdaBoost Tuned	0.69	0.69	0.97	0.97	0.69	0.69	0.69	0.69
10	Gradient Boosting with default parameters	0.76	0.75	0.88	0.88	0.78	0.78	0.78	0.78
11	Gradient Boosting with init=AdaBoost	0.75	0.75	0.88	0.88	0.78	0.78	0.78	0.78
12	Gradient Boosting Tuned	0.74	0.74	0.90	0.91	0.75	0.76	0.75	0.76
13	XGBoost with default parameters	0.83	0.74	0.93	0.86	0.84	0.78	0.84	0.78
14	XGBoost Tuned	0.67	0.67	1.00	1.00	0.67	0.67	0.67	0.67
15	Stacking Classifier	0.75	0.75	0.87	0.87	0.78	0.78	0.78	0.78

Model comparison:

- **Which model is best suited ?**

- The below models have comparable performance and balanced performance
  - Gradient Boosting with default parameters
  - Gradient Boosting with init=AdaBoost
  - Stacking Classifier
  - Random Forest Tuned

- The Stacking Classifier might be the best model

Business Insights and Recommendations

- Aim would be to balance the trade off between losing an opportunity (Deny Visa to skilled and qualified individual) in case of FN and losing the resources of office of foreign labour certification in case of False positive.

- We emphasized that recall is the metric of interest here and we tuned our model on recall. But this does not mean that other metrics should be ignored completely.

- Different model indicate that the main feature is education_of_employee and has_job_experience.

- Employees with higher education have better rate of approval. Especially with Doctorate.

- Europe seems to have higher level of approvals.

- Pay unit of Yearly appears to be the most approved.

Observations:

- The region of intended employment does not seem to impact much, all regions have almost similar rate of approval.
- Full time or part time nature of the work does not appear to impact the case status.
- Applicants who have the unit_of_wage for prevailing_wage as year have the better chance of approval.
- Asia and Europe have almost equal number of employees who are Doctorates.
- Job experience seems to be a deciding factor for all education level except for Doctorate.

greatlearning

Power Ahead

Happy Learning !