EasyVisa Case Study

ET EasyVisa Project
May 2023

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary

What and How

Executive Summary

- The goal was to build a Machine Learning solution that can help shortlisting VISA canidates that have a higher chance of a VISA approval

- The classification model will facilitate the process of visa approvals and recommend a profile of canidates should be certified or denied based on identified factors

- Utlizied data to build classification models that would provide VISA recomendations

- Identify factors that influence VISA approvals and rejections
- Focused on data from
    - Employee attributes
    - Wages
    - Geographic factors
    - Previous jobs

Conclusions and Recomendations

Executive Summary

- Based on our analysis, people who are granted a VISA have the following attributes
  - At least a high school education
  - Higher Education
  - Has job experience
  - Are paid yearly
- OFLC should focus on fastracking people with university level education, who have work experience and are have salaried wages
- Once the desired performance is achieved from the model, the company can use it to utilize the attributes to fast-track people in the VISA application process.

How can we disover the best attributes for VISA approvals

Business Problem Overview and Solution Approach

- Find the best attributes that will lead to fast tracking VISA candidates that are likely to be approved
- What does the data tell us?
- The Approach
  - Developed the questions to explore data with
  - Perform data overview
  - Exploratory Data Analysis
  - Data Preprocessing
  - Model Building – Decision Tree, Bagging, Random Forest, Boosting, XGBoost, Stacking
  - Finalize model summary
  - Developed recomendations

Data Overview

EDA Results

- 25,480 Rows
- 12 Columns
  - Case Id (object)
  - Continent (object)
  - Education of Employee (object)
  - Has Job Experience (object)
  - Requires Job Training (object)
  - No of Employees (int64)
  - Years of Establishment (int64)
  - Region of Employment (object)
  - Prevailing Wage (float64)
  - Unit of Wage (object)
  - Full Time Position (object)
  - Case Status (object)
- Object (9), Int64 (2), Float64 (1)

- No duplicates

Data – Average, Max, Min

EDA Results

Average

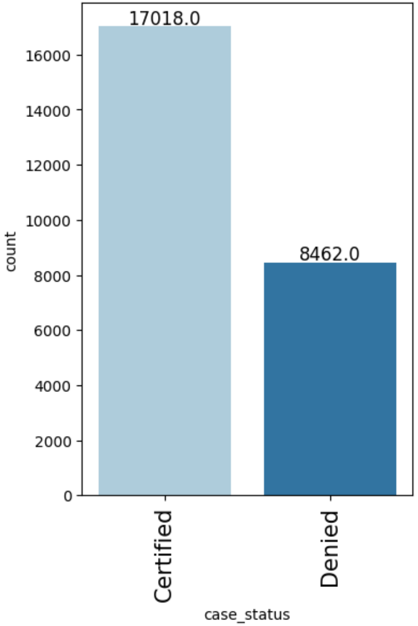
- Number of Employees
  - 5,667
- Year Company Established
  - 1979
- Prevailing Wage
  - 74,456

Max

- Number of Employees
  - 602,069
- Year Company Established
  - 2016
- Prevailing Wage
  - 319,210

Min

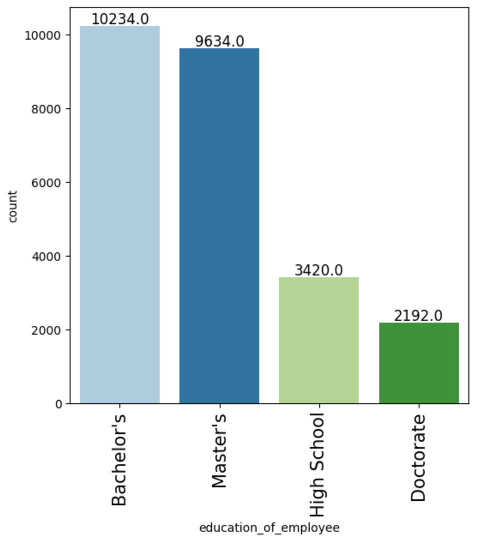
- Number of Employees
  - 11
- Year Company Establis
  - 1800
- Prevailing Wage
  - 2

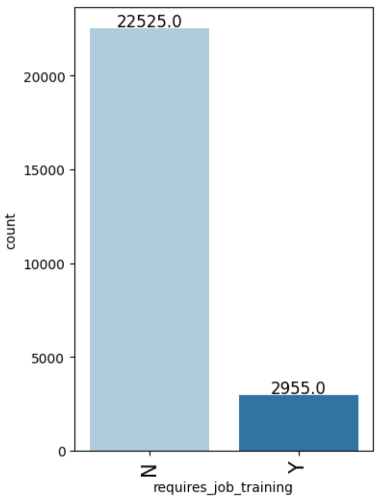


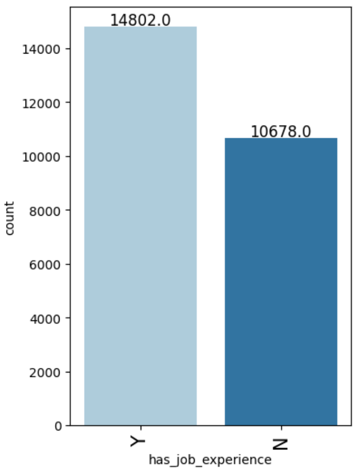
Data – Employee Attributes

EDA Results

- Most applicant have
  - Higher Education
  - Don’t need job training
  - Have worked before



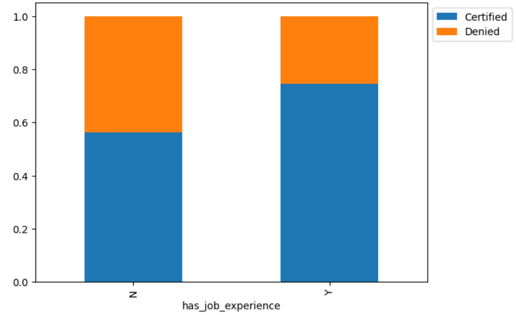


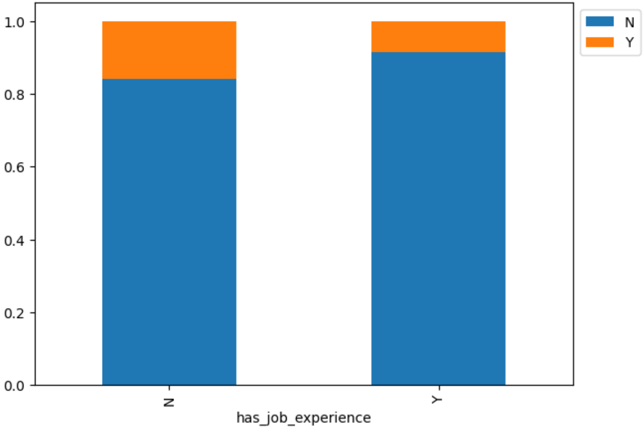


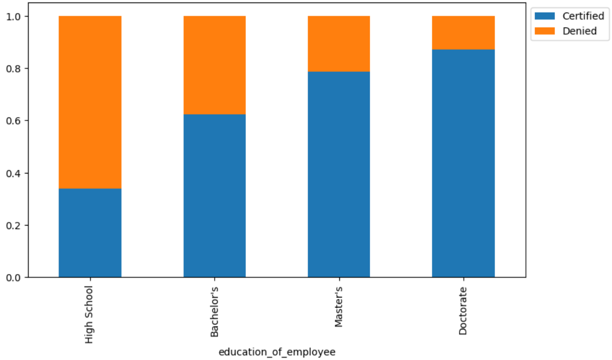
Data – Employee Attributes

EDA Results

- Europe and Africa are the most likely to be approved
- As education level rises so does the likelihood of approval
- Most have job experience and do not need training



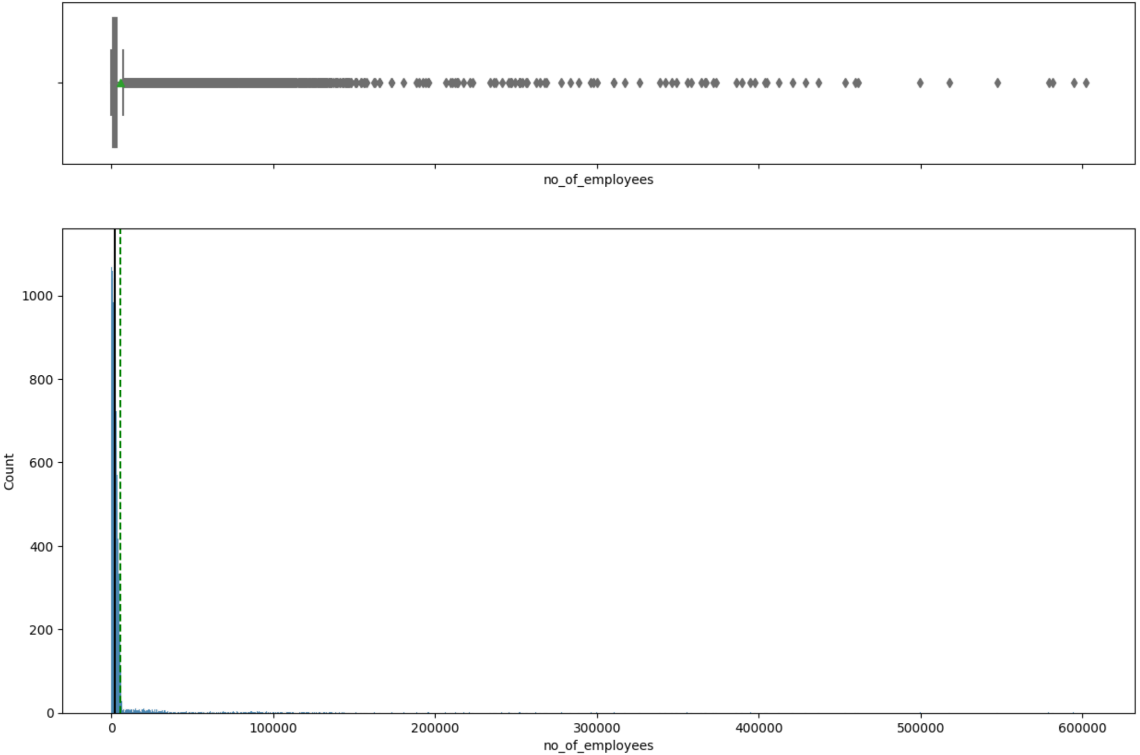




Data – Employer Attributes

EDA Results

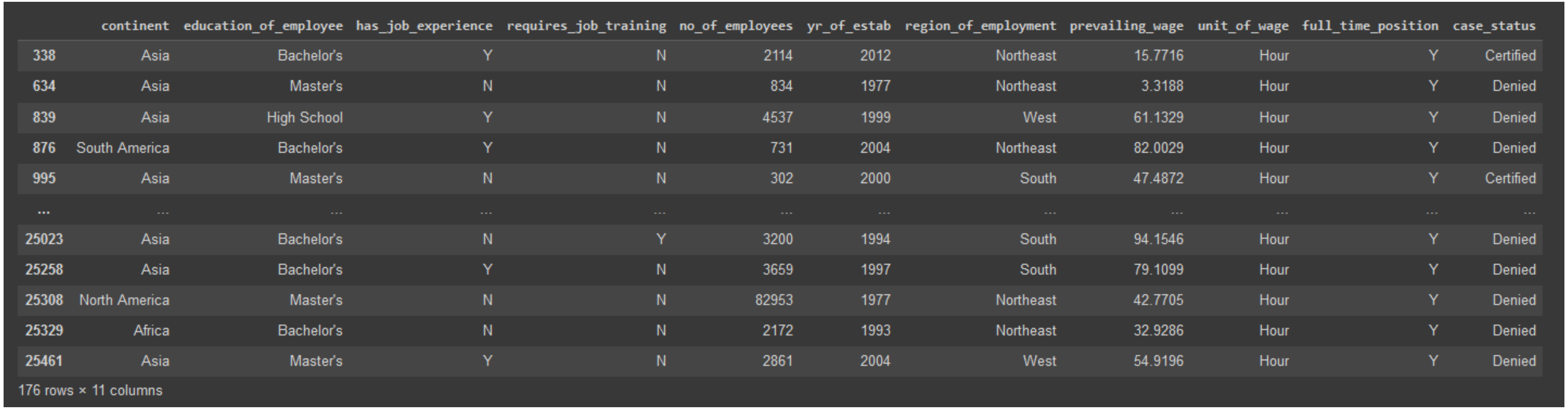
- Most of the employers are small companies



Data – Wage Attributes

EDA Results

Observations wit have less than 100 in prevailing wage



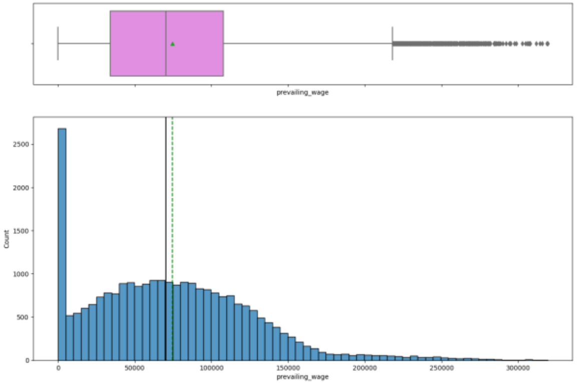
	continent	education_of_employee	has_job_experience	requires_job_training	no_of_employees	yr_of_estab	region_of_employment	prevailing_wage	unit_of_wage	full_time_position	case_status
338	Asia	Bachelor's	Y	N	2114	2012	Northeast	15.7716	Hour	Y	Certified
634	Asia	Master's	N	N	834	1977	Northeast	3.3188	Hour	Y	Denied
839	Asia	High School	Y	N	4537	1999	West	61.1329	Hour	Y	Denied
876	South America	Bachelor's	Y	N	731	2004	Northeast	82.0029	Hour	Y	Denied
995	Asia	Master's	N	N	302	2000	South	47.4872	Hour	Y	Certified
...	...	...	...	...	...	...	...	...	...	...	...
25023	Asia	Bachelor's	N	Y	3200	1994	South	94.1546	Hour	Y	Denied
25258	Asia	Bachelor's	Y	N	3659	1997	South	79.1099	Hour	Y	Denied
25308	North America	Master's	N	N	82953	1977	Northeast	42.7705	Hour	Y	Denied
25329	Africa	Bachelor's	N	N	2172	1993	Northeast	32.9286	Hour	Y	Denied
25461	Asia	Master's	Y	N	2861	2004	West	54.9196	Hour	Y	Denied

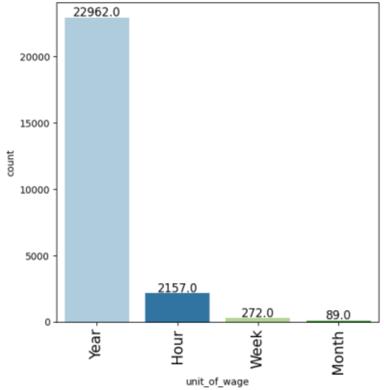
176 rows × 11 columns

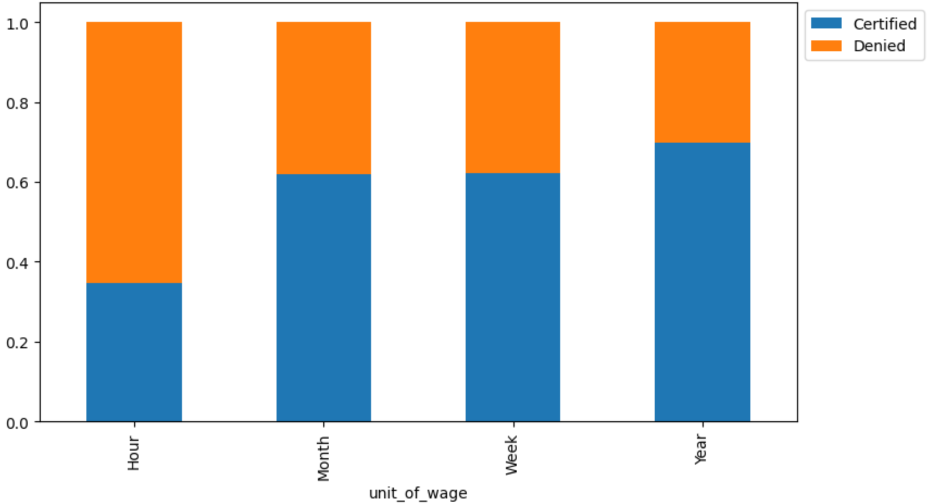
Data – Wage Attributes

EDA Results

- Wages have a right skew
- Most applicants are salary
- If an applicant is paid hourly they are the most likely to be denied



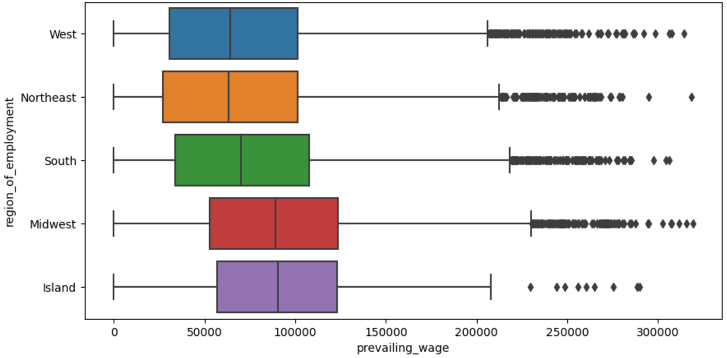


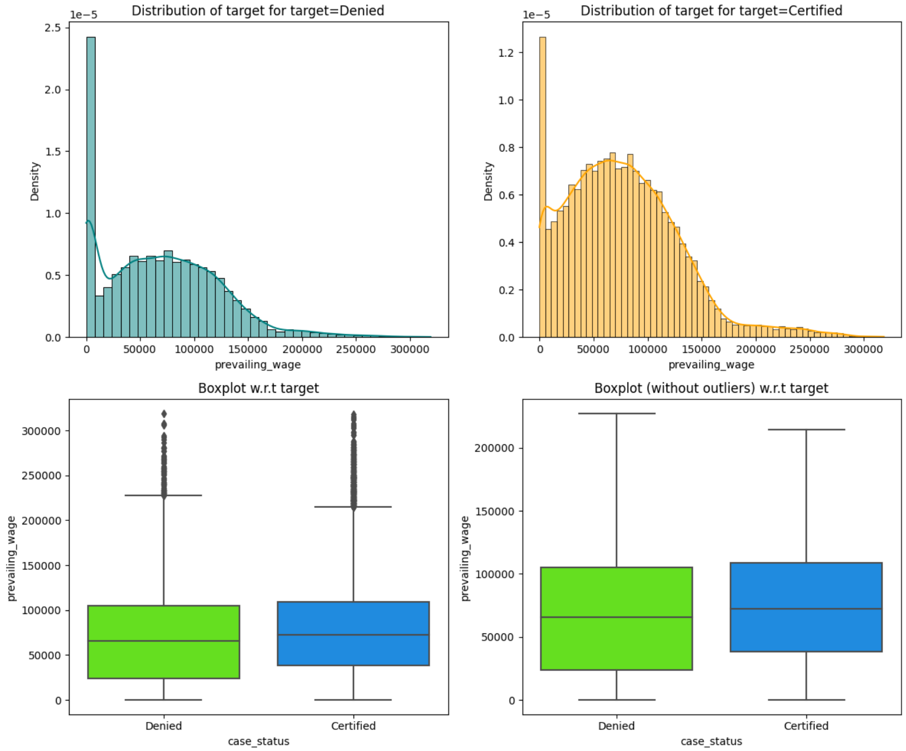


Data – Wage Attributes

EDA Results

- Wages are highest in islands and the Midwest
- Prevailing wages are right skewed

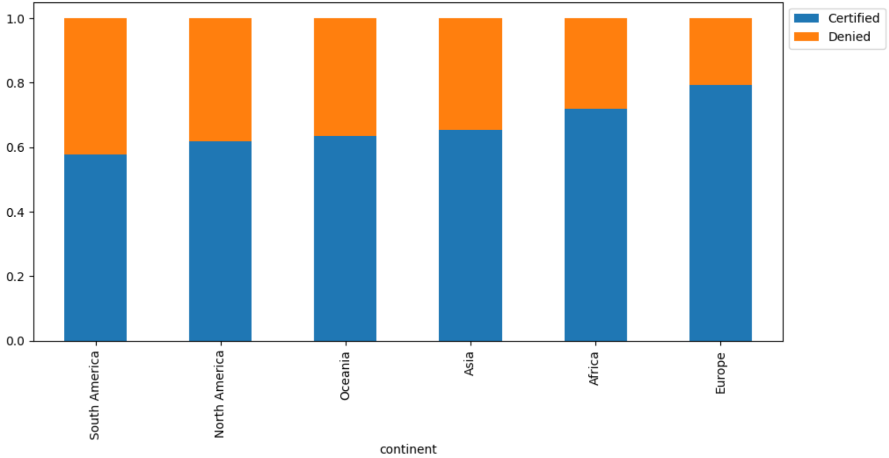


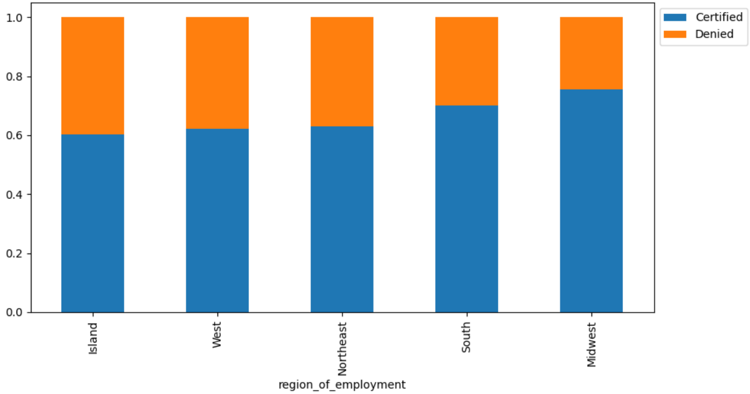


Data – Geograpical Attributes

EDA Results

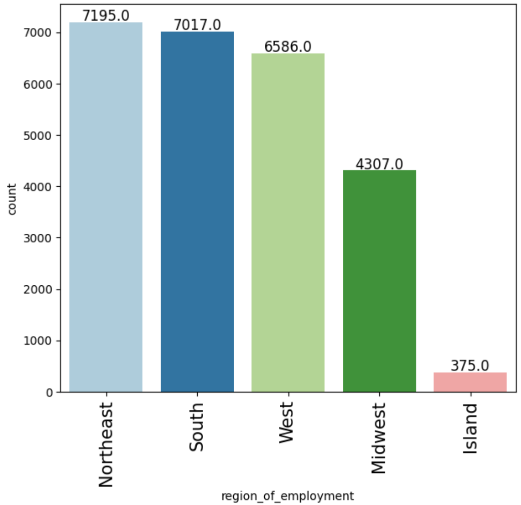
- The Midwest and South have the highest approval
- Europe and Africa have the most approvals



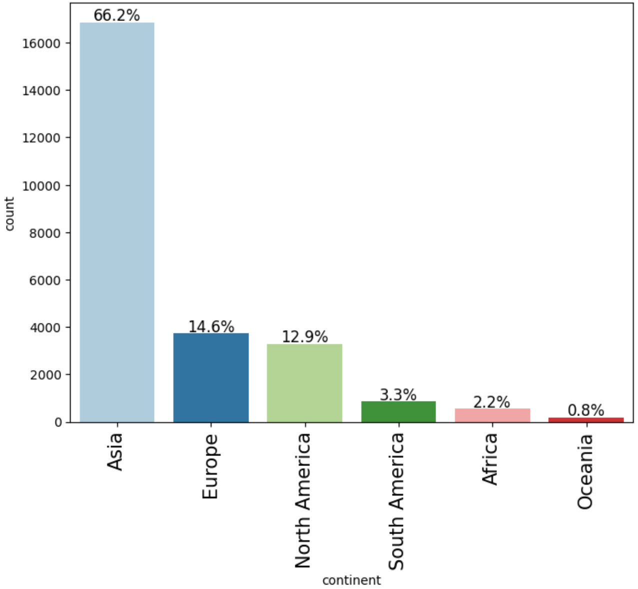


Data – Geograpical Attributes

EDA Results



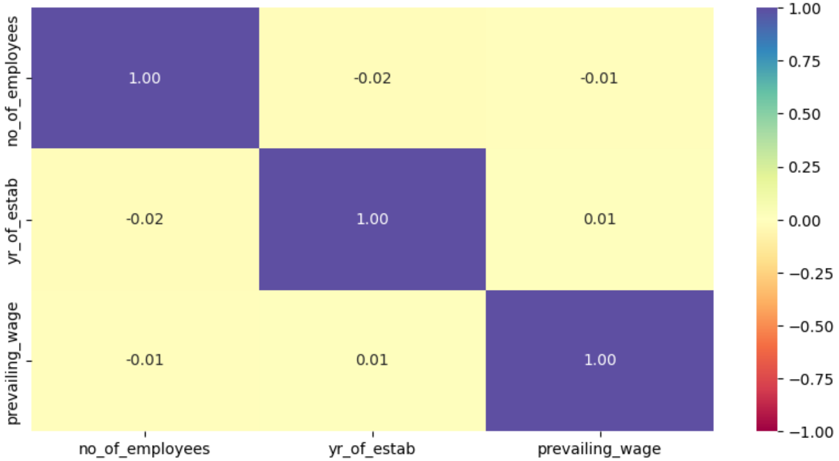
- Most applicant have
  - Are from Asia
  - Work in the Northeast, South and West
  - Don’t need job training
  - Have worked before

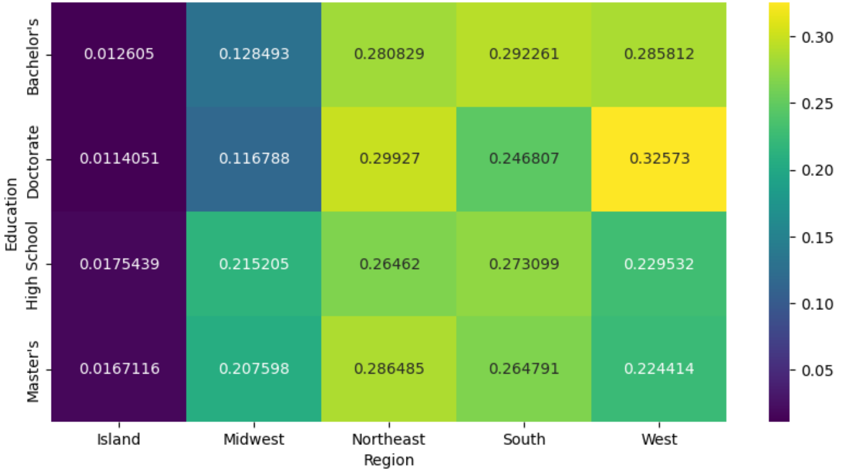


Data – Correlation

EDA Results

- Little correlation between
  - No of employees
  - Year company established
  - Prevailing wage

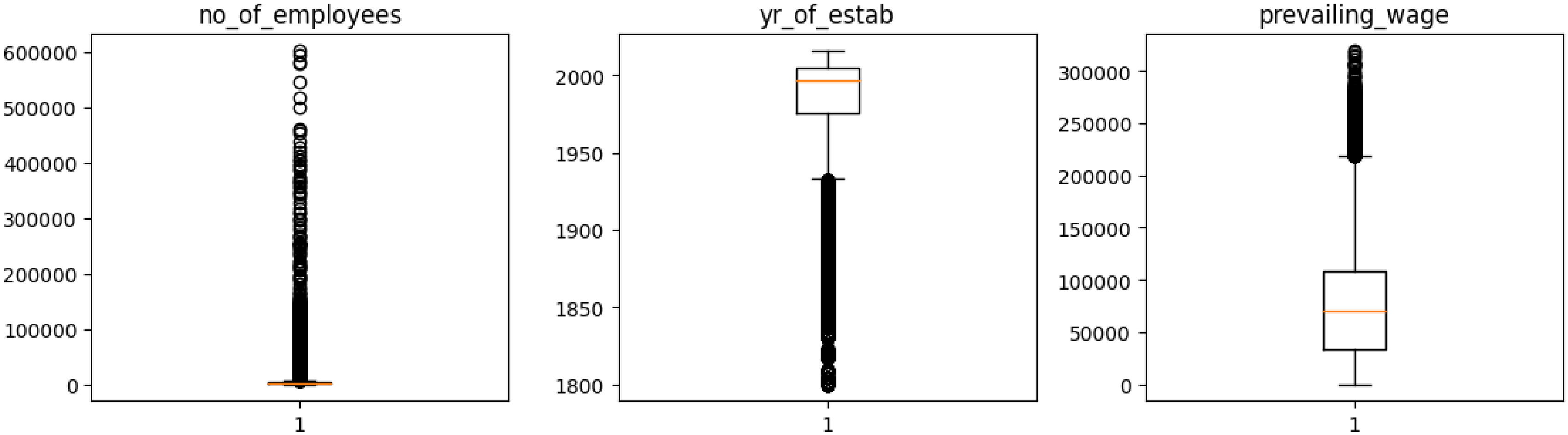




Data Prep

Data Preprocessing

- Made all negative values positive
- There are a lot of outliers
- Dropped Case Status Column
- Created dummy variables
- Split data into training and testing sets (70/30)
- Both training and test sets are 66% (train) and 33% (test)

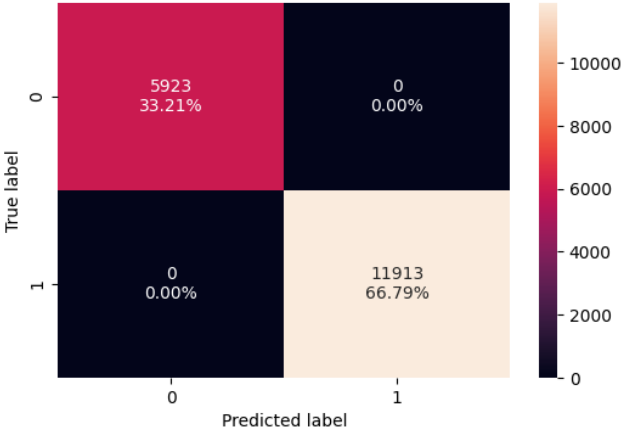


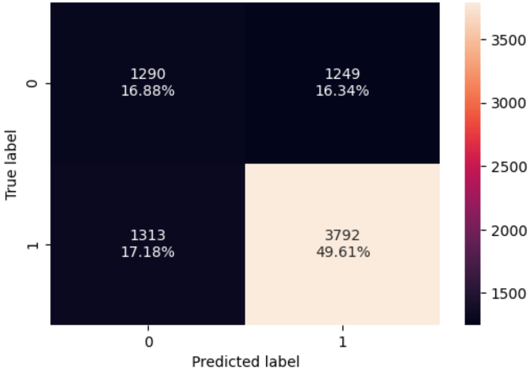
Decision Tree

Model Building

Training	Accuracy	Recall	Precision	F1
D-Tree	1.0	1.0	1.0	1.0

Testing	Accuracy	Recall	Precision	F1
D-Tree	0.66	0.74	0.75	0.75





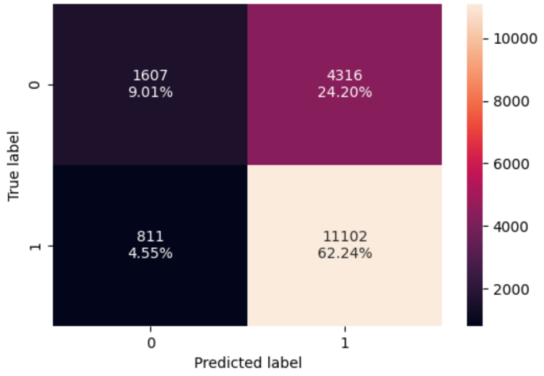
- DecisionTreeClassifier (Random_State1)
- Overfitted

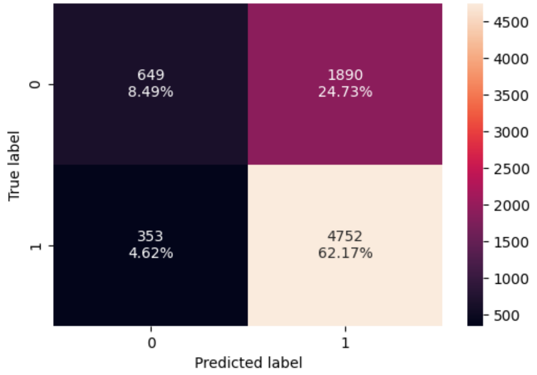
Decision Tree with Hyperparameter Tuning

Model Building

Training	Accuracy	Recall	Precision	F1
D-Tree Hyper	0.71	0.93	0.72	0.81

Testing	Accuracy	Recall	Precision	F1
D-Tree Hyper	0.71	0.93	0.72	0.81





- DecisionTreeClassifier (class_weight='balanced', max_depth=5, max_leaf_nodes=2, min_impurity_decrease=0.0001, min_samples_leaf=3, Random_state=1)

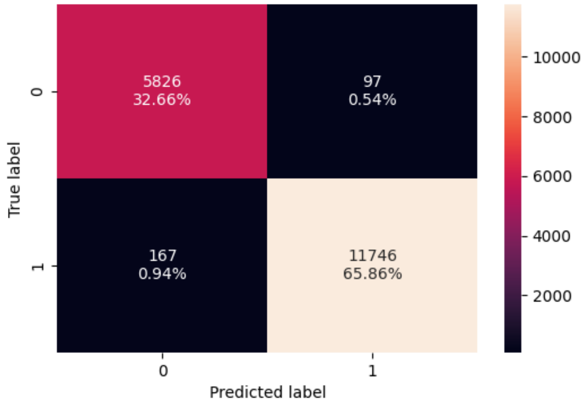
- Not overfit
- All measures match

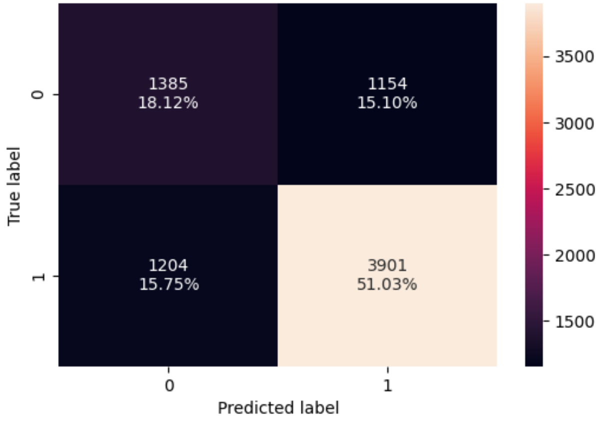
Bagging

Model Building

Training	Accuracy	Recall	Precision	F1
Bagging	0.99	0.99	0.99	0.99

Testing	Accuracy	Recall	Precision	F1
Bagging	0.69	0.76	0.77	0.77





- BaggingClassifier(Random_State=1

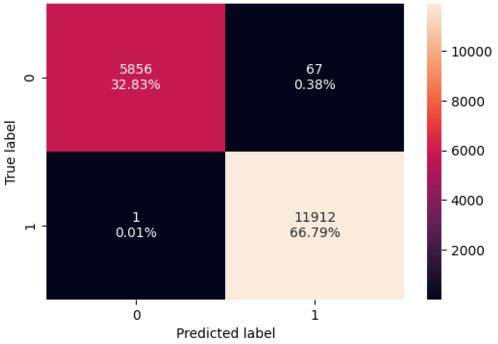
- Overfit

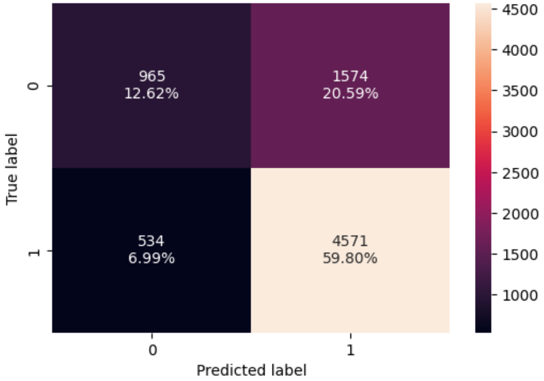
Bagging with Hyperparameter Tuning

Model Building

Training	Accuracy	Recall	Precision	F1
Bagging Hyper	1.0	1.0	0.99	1.0

Testing	Accuracy	Recall	Precision	F1
Bagging Hyper	0.73	0.90	0.74	0.81





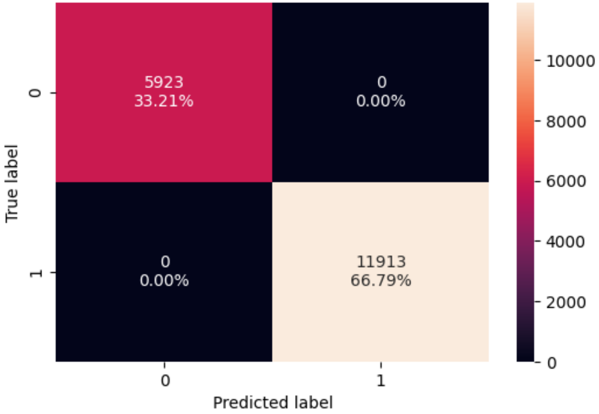
- BaggingClassifier (max_features=0.7, max_samples=0.7, n_estimators=100, Random_state=1)
- Overfit

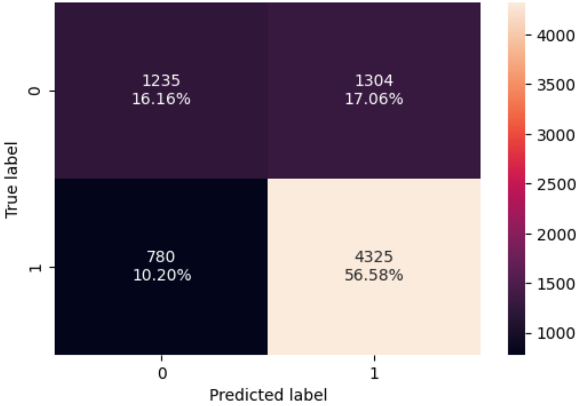
Random Forest

Model Building

Training	Accuracy	Recall	Precision	F1
Random Forest	1.0	1.0	1.0	1.0

Testing	Accuracy	Recall	Precision	F1
Random Forest	0.73	0.85	0.77	0.81





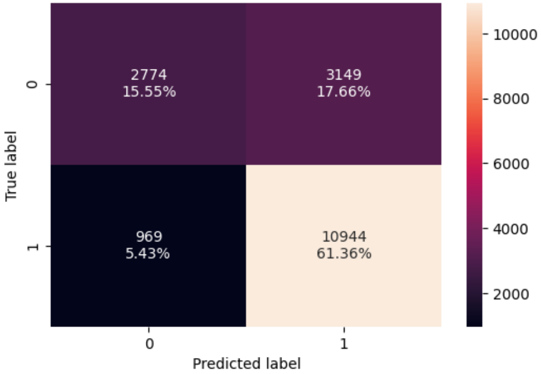
- RandomForestClassifier(class_weight=‘balanced’, Random_State=1)
- Overfit

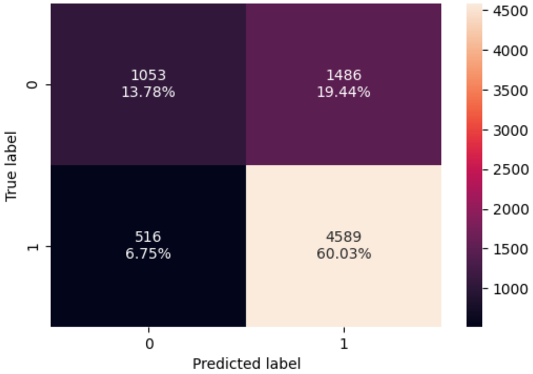
Random Forest with Hyperparameter Tuning

Model Building

Training	Accuracy	Recall	Precision	F1
Random Forest Hyper	0.77	0.92	0.78	0.84

Testing	Accuracy	Recall	Precision	F1
Random Forest Hyper	0.74	0.90	0.76	0.82





- RandomForestClassifier (max_depth=10, min_samples=7, n_estimators=20, oob_score=True, Random_state=1)

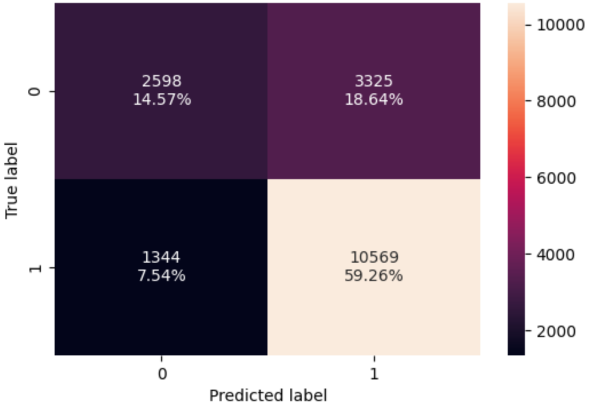
- Not overfit
- All measures are close except Accuracy is out of the 2% threshold

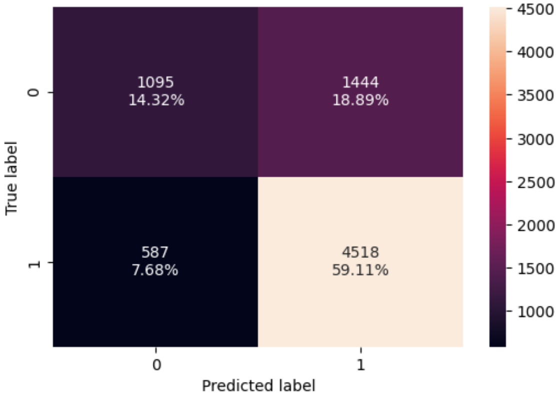
Boosting - AdaBoost

Model Building

Training	Accuracy	Recall	Precision	F1
AdaBoost	0.74	0.89	0.76	0.82

Testing	Accuracy	Recall	Precision	F1
AdaBoost	0.73	0.89	0.76	0.82





- AdaBoostClassifier(Random_State=1)

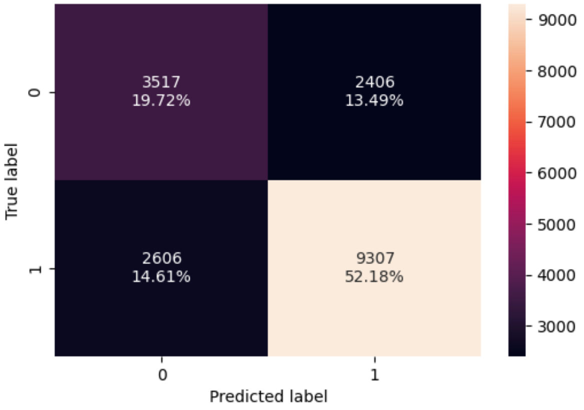
- Measures have a good fit

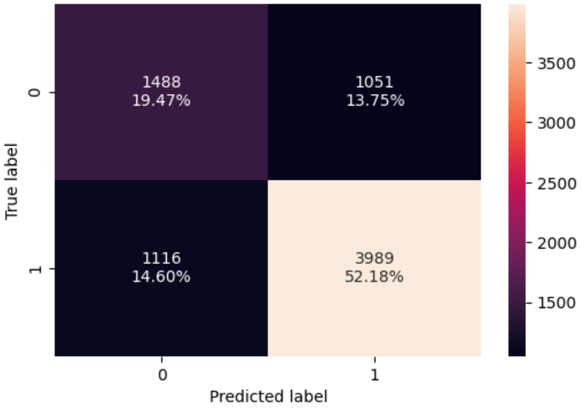
Boosting – ADABoost with Hyperparameter Tuning

Model Building

Training	Accuracy	Recall	Precision	F1
AdaBoost DTree	0.72	0.78	0.79	0.79

Testing	Accuracy	Recall	Precision	F1
AdaBoost DTree	0.71	0.78	0.79	0.79





- AdaBoostClassifier
- Base_estimator: DecisionTreeClassifier

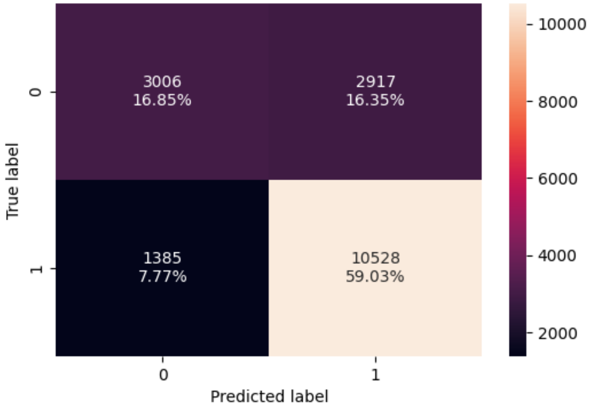
- Not overfit
- All measures are within the 2% threshold

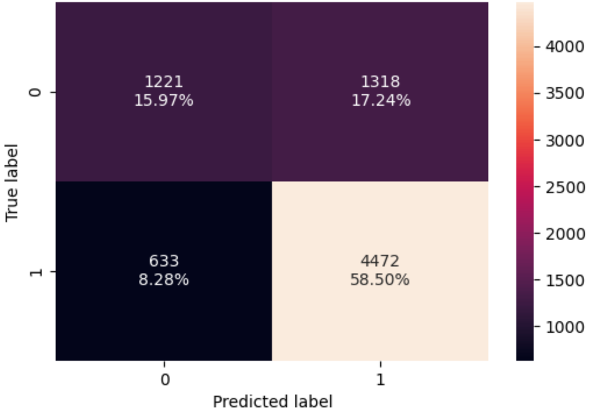
Boosting - Gradient

Model Building

Training	Accuracy	Recall	Precision	F1
Gradient Boosting	0.75	0.88	0.78	0.83

Testing	Accuracy	Recall	Precision	F1
Gradient Boosting	0.74	0.88	0.77	0.82





- GradientBoostingClassifier (random_state=1)

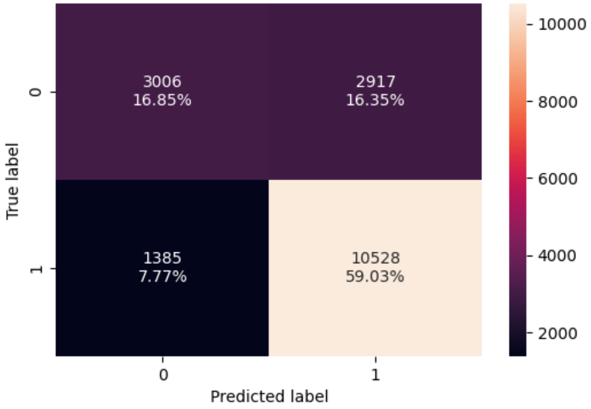
- Not overfit
- All measures are within the 2% threshold

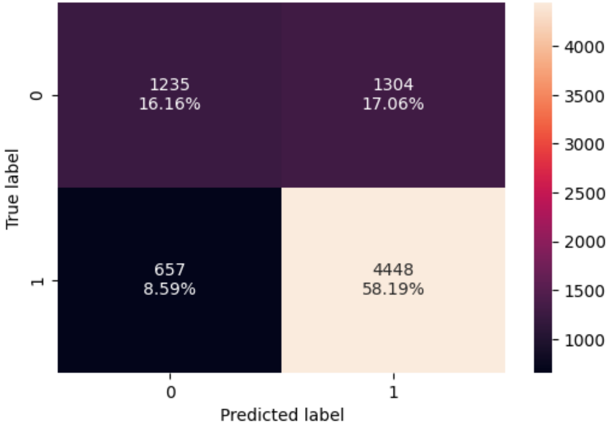
Boosting - Gradient with Hyperparameter Tuning

Model Building

Training	Accuracy	Recall	Precision	F1
Gradient Boosting Hyper	0.76	0.88	0.79	0.83

Testing	Accuracy	Recall	Precision	F1
Gradient Boosting Hyper	0.74	0.87	0.77	0.82





- GradientBoostingClassifier
- Init: AdaBoost Classifier

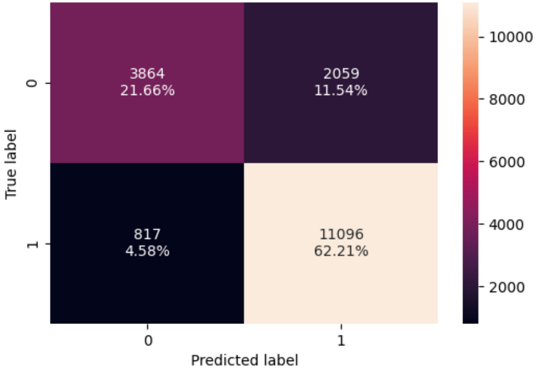
- Not overfit
- All measures are within the 2% threshold

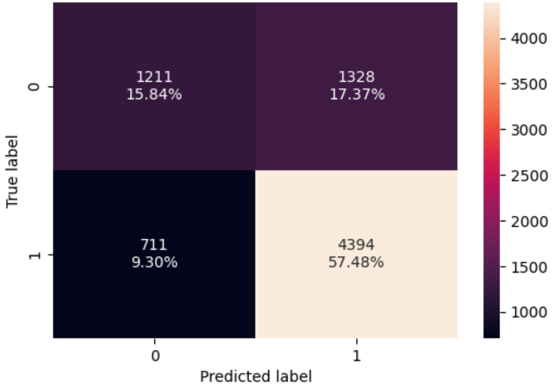
Boosting - XGBoost

Model Building

Training	Accuracy	Recall	Precision	F1
XGBoost	0.84	0.93	0.84	0.89

Testing	Accuracy	Recall	Precision	F1
XGBoost	0.73	0.86	0.77	0.81





- XGBClassifier(base_score=none, booster=none, callbacks=none, colsample_bylevel=none, colsample_bynode=none, colsample_bytree=none, early_stopping_rounds=none, enable_categorical=false, eval_metrics='logloss', feature_types=none, interaction_constraints=none, learning_rate=none, max_bin=none, max_cat_threshold=none, max_cat_to_onehot=none, max_delta_step=none, max_depth=none, max_leaves=none, min_child_weight=none, missing=nan, monotone_constraints=none, n_estimators=100, n_jobs=none, num_parallel_tree=none, predictor=none, random_state=1, ...)

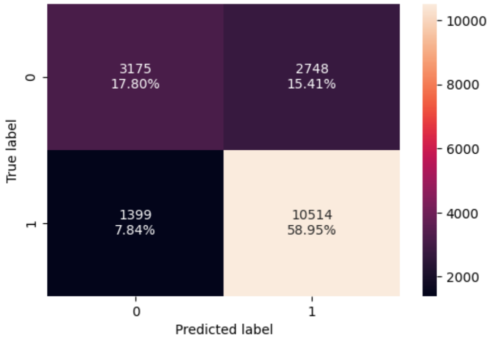
- All measures are out of the 2% threshold

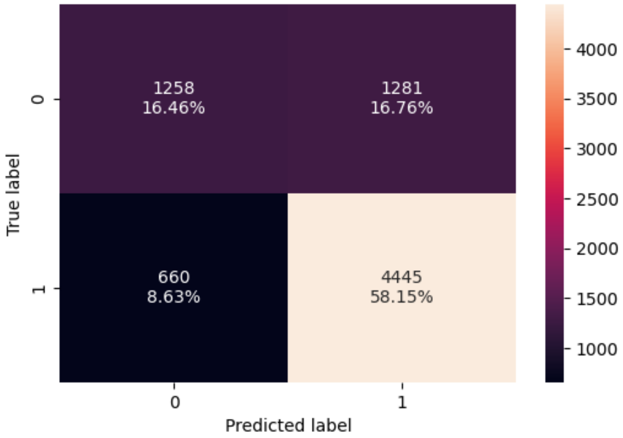
Boosting – XGBoost with Hyperparameter Tuning

Model Building

Training	Accuracy	Recall	Precision	F1
XGBoost Hyper	0.77	0.88	0.79	0.84

Testing	Accuracy	Recall	Precision	F1
XGBoost Hyper	0.75	0.87	0.78	0.82





- XGBClassifier(base_score=none, booster=none, callbacks=none, colsample_bylevel=0.9, colsample_bynode=none, colsample_bytree=0.9, early_stopping_rounds=none, enable_categorical=false, eval_metrics='logloss', feature_types=none, gamma=5, gpu_id=none, grow_policy=none, importance_type=none, interaction_constraints=none, learning_rate=0.1, max_bin=none, max_cat_threshold=none, max_cat_to_onehot=none, max_delta_step=none, max_depth=none, max_leaves=none, min_child_weight=none, missing=nan, monotone_constraints=none, n_estimators=150, n_jobs=none, num_parallel_tree=none, predictor=none, random_state=1, ...)

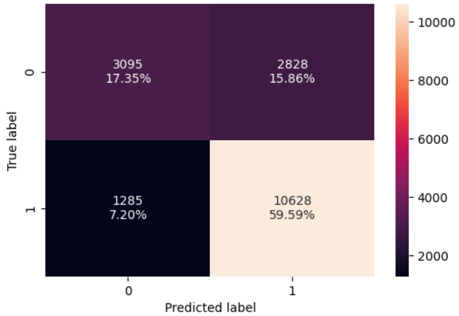
- Not overfit
- All measures are close except Accuracy is out of the 2% threshold

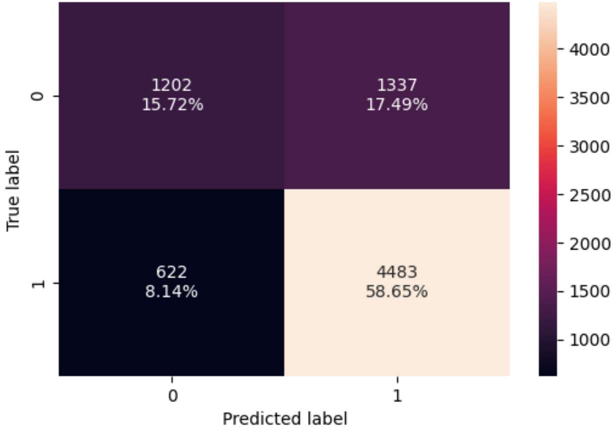
Stacking

Model Building

Training	Accuracy	Recall	Precision	F1
Stacking	0.77	0.89	0.79	0.84

Testing	Accuracy	Recall	Precision	F1
Stacking	0.74	0.88	0.77	0.82



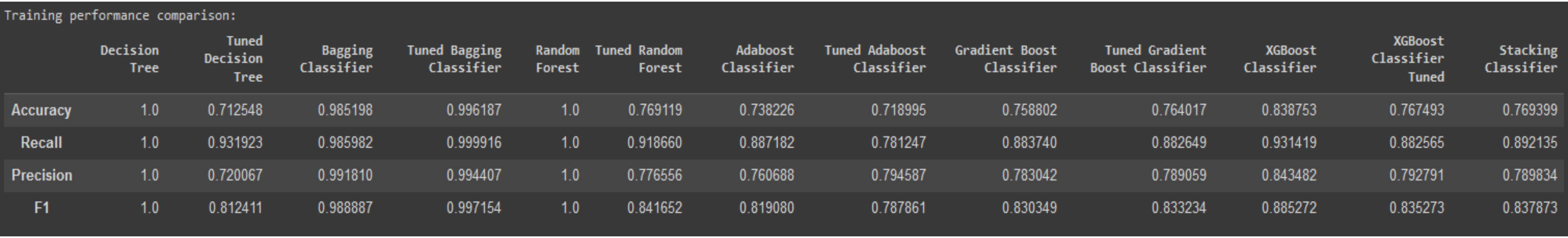


- AdaBoostClassifier
- Gradient Boosting
  - Init: AdaBoostClassifier
- RandomForestClassifier
- Final_Estimator
  - XGBClassifier

- All measures are out of the 2% threshold

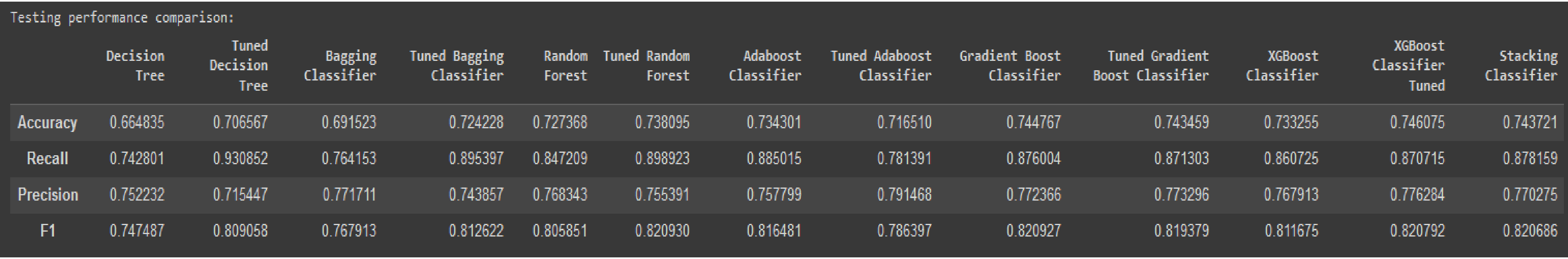
Machine Learning Summary

Model Performance Summary



Training performance comparison:

	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	1.0	0.769119	0.738226	0.718995	0.758802	0.764017	0.838753	0.767493	0.769399
Recall	1.0	0.931923	0.985982	0.999916	1.0	0.918660	0.887182	0.781247	0.883740	0.882649	0.931419	0.882565	0.892135
Precision	1.0	0.720067	0.991810	0.994407	1.0	0.776556	0.760688	0.794587	0.783042	0.789059	0.843482	0.792791	0.789834
F1	1.0	0.812411	0.988887	0.997154	1.0	0.841652	0.819080	0.787861	0.830349	0.833234	0.885272	0.835273	0.837873



Testing performance comparison:

	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	0.727368	0.738095	0.734301	0.716510	0.744767	0.743459	0.733255	0.746075	0.743721
Recall	0.742801	0.930852	0.764153	0.895397	0.847209	0.898923	0.885015	0.781391	0.876004	0.871303	0.860725	0.870715	0.878159
Precision	0.752232	0.715447	0.771711	0.743857	0.768343	0.755391	0.757799	0.791468	0.772366	0.773296	0.767913	0.776284	0.770275
F1	0.747487	0.809058	0.767913	0.812622	0.805851	0.820930	0.816481	0.786397	0.820927	0.819379	0.811675	0.820792	0.820686

Machine Learning Summary

Model Performance Summary

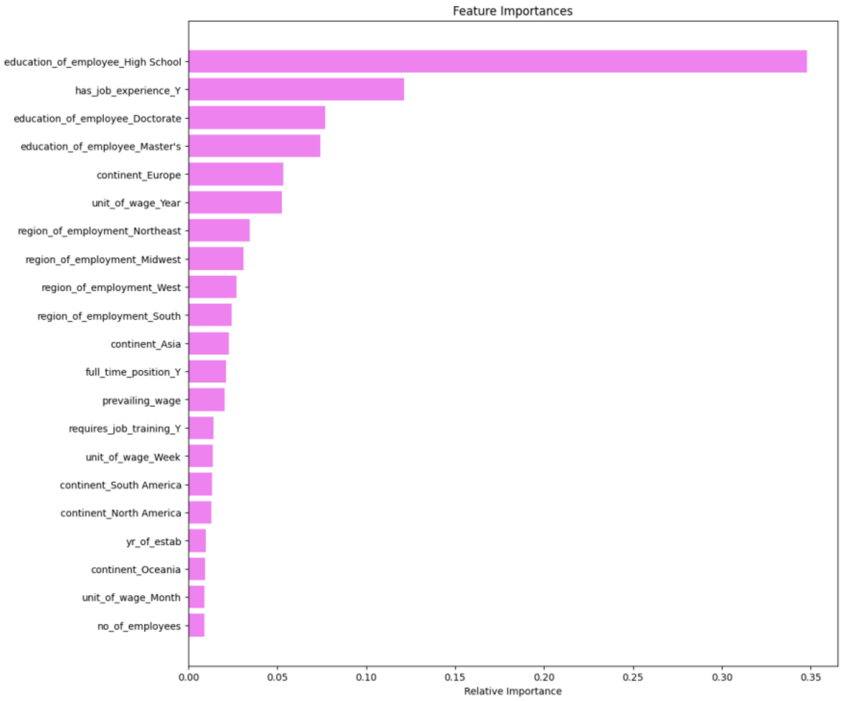
- Decision Tree: Overfit
- Tuned Decision Tree: All measures matched
- Bagging: Overfit
- Tuned Bagging: Overfit
- Random Forest: Overfit
- Tuned Random Forest: All measures are within 2% threshold, expect Accuracy
- AdaBoost: Measures have a good fit
- Tuned AdaBoost: Measures have a good fit, but not as good as AdaBoost
- Gradient: All measures are within 2% threshold
- Tuned Gradient: All measures are within 2% threshold, but not as good as Gradient
- XGBoost: All measures are out of the 2% threshold
- Tuned XGBoost: All measures are within 2% threshold
- Stacking: All measures are within 2% threshold, expect Accuracy

Machine Learning Summary

Model Performance Summary

Tuned XGBoost Classifier has the best fit Machine Learning Model

- Does not overfit
- Has the best Accuracy, Precision and F1 out of the models that did not overfit
- Education, job experience, prevailing wage are the three most important factors



Great Learning

Happy Learning !