Easy Visa

Ensemble Techniques

3/1/2024

Contents / Agenda

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

Executive Summary

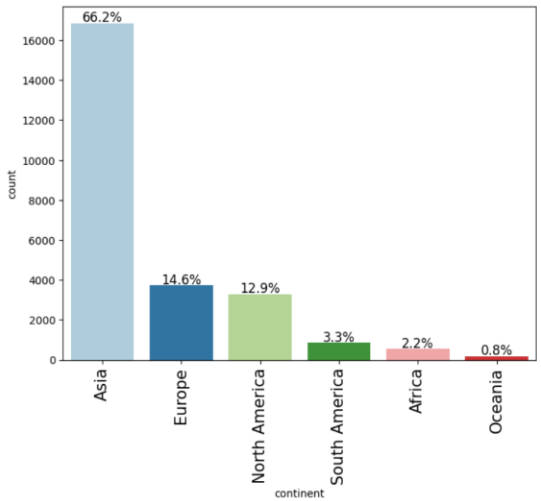
- **Insights**
  - There are 3 significant features for pre-screening an applicant
    - Education Level: the higher the level of education, more likely they will be certified.
    - Prior Job Experience: having prior job experiecne increases the chance of an approved visa.
    - Prevailing wage: the higher the prevailing wage, the more likely the application will be certified.
- **Recommendations**
  - Focus on Key Features: Prioritize visa applicants with higher education, relevant job experience, and wages at or above the prevailing rate for better decision-making and faster screening.
  - Continuous Model Evaluation: Regularly update and adjust the model with new data to keep it accurate and relevant to current visa and labor market trends.
  - Policy Recommendations: Update visa application guidelines to emphasize the importance of education, experience, and wage levels based on model insights, helping applicants understand how to showcase their qualifications more effectively.
  - Hybrid Review: Incorporate a manual review stage post-machine learning analysis to capture nuanced details about an applicant's experience not detected by the model, ensuring a comprehensive evaluation.

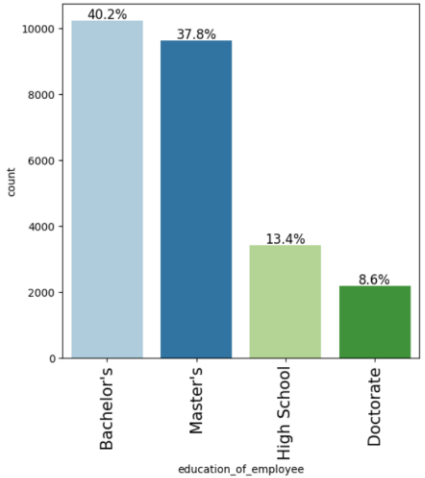
Business Problem Overview and Solution Approach

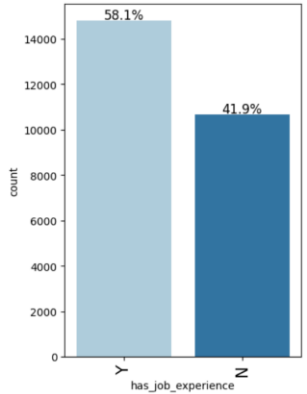
- **Business Challenge:** Efficiently processing an increasing number of visa applications and ensuring the selection of qualified foreign workers.
- **Solution Goal:** Implement a Machine Learning model to predict visa approval probabilities, aiding in the prioritization and decision-making process.
- **Data Utilization:** Analyze past application data to identify patterns and factors that significantly influence visa approvals.
- **Methodology:** Design and train a predictive classification model to evaluate new applications and suggest approvals or denials.
- **Impact:** Streamline OFLC's certification process, safeguard U.S. labor conditions, and maintain the competitive edge of U.S. businesses.

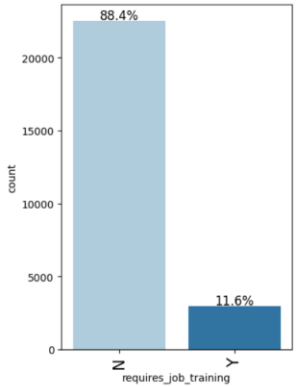
EDA Results (Univariate Analysis)

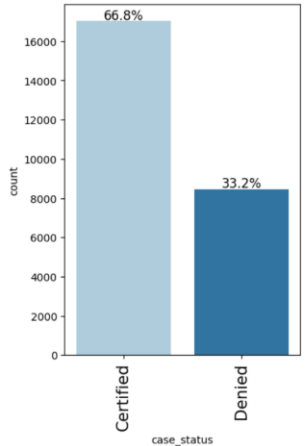
- Asia has the highest amount of applications
- Most applicants have a Bachelor's Degree followed by a Master's
- Over half the applicants have job experience and majority do not require job training
- Over half the cases are certified.









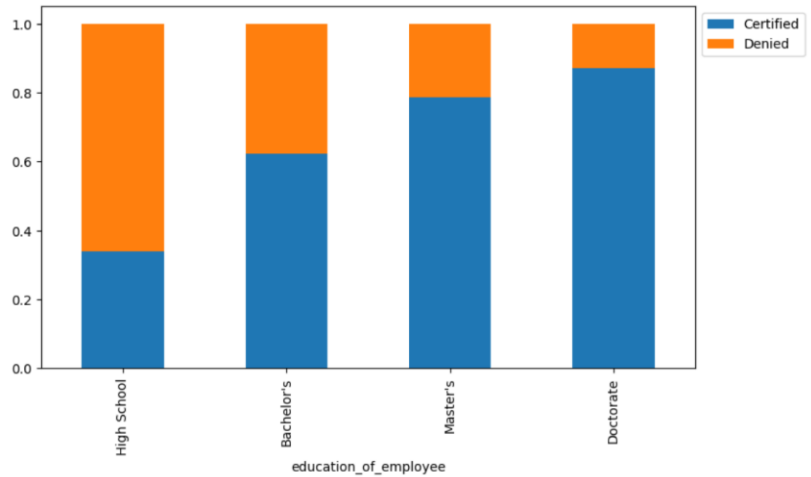


[Link to Appendix slide on data background check]

EDA Results (Bi-Variate Analysis)

- As education level increases, the ratio of certified to denied applications increases as well
  - High school: 1|2
  - Doctorate: 7|1

case_status	Certified	Denied	All
education_of_employee			
All	17018	8462	25480
Bachelor's	6367	3867	10234
High School	1164	2256	3420
Master's	7575	2059	9634
Doctorate	1912	280	2192

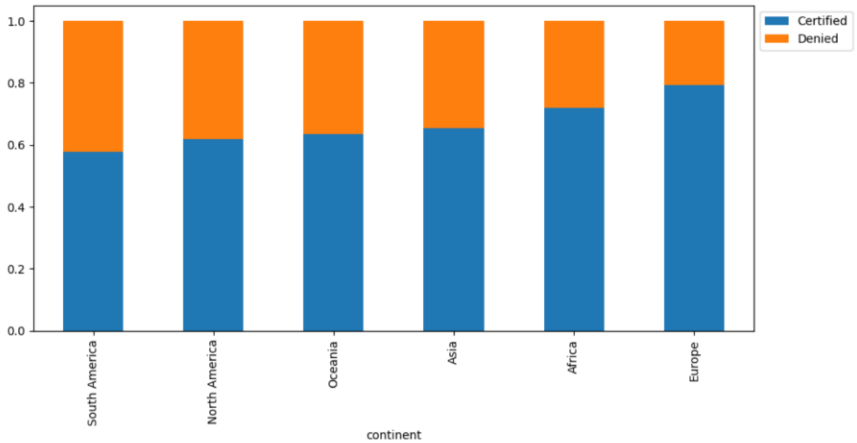


Link to Appendix slide on data background check

EDA Results (Bi-Variate Analysis)

- Cases from Europe have the highest ratio of certified to denied applications
- South America has the lowest ratio
- Oceania has the least amount of cases

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

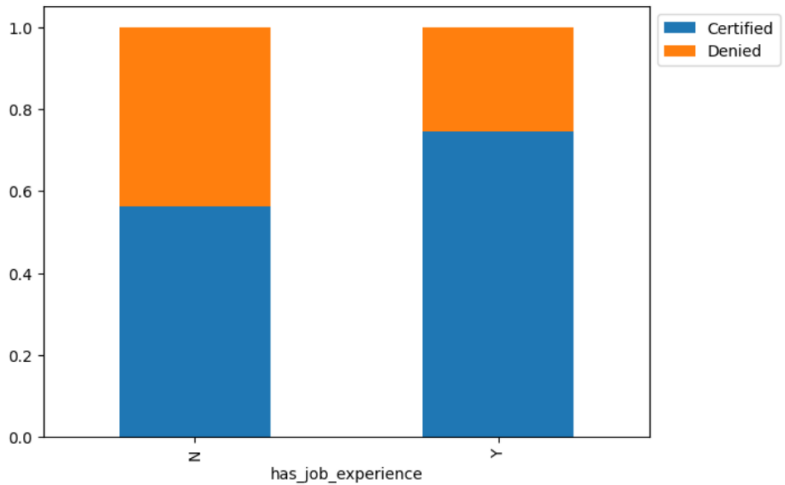


Link to Appendix slide on data background check

EDA Results (Bi-Variate Analysis)

- At first glance, having job experience seems to have a positive influence on visa certification with more applications being certified if prior work experience is present

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

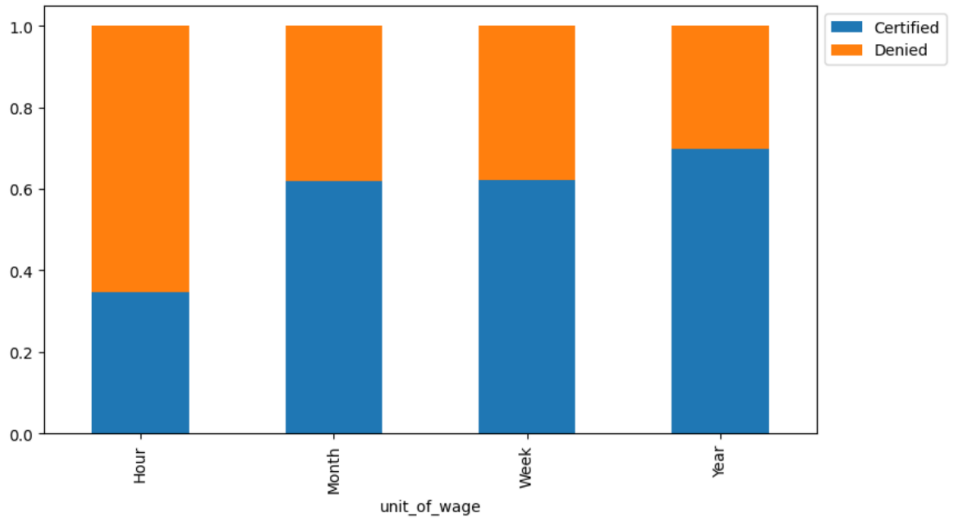


Link to Appendix slide on data background check

EDA Results (Bi-Variate Analysis)

- The pay unit with the most certifications is Yearly.

case_status	Certified	Denied	All
unit_of_wage			
All	17018	8462	25480
Year	16047	6915	22962
Hour	747	1410	2157
Week	169	103	272
Month	55	34	89



Link to Appendix slide on data background check

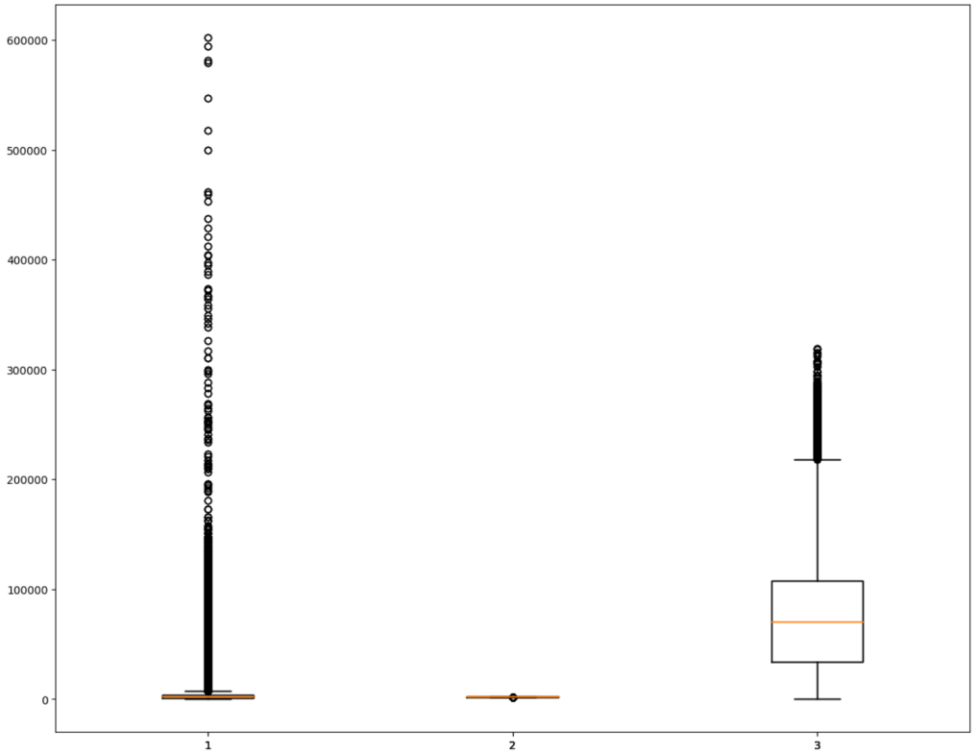
EDA Results (Bi-Variate Analysis)

- Correlation Matrix indicates the numerical features have little to no correlation to each other.

Link to Appendix slide on data background check

Data Preprocessing

- Outlier Check
  - Group 1: No. Employees
    - Large number of outliers
    - Substantial spread
  - Group 2: Year Est.
    - No visible outliers
    - Year is more of a categorical variable
  - Group 3: Prevailing Wage
    - Variation in the upper range but not as extreme as Group 1

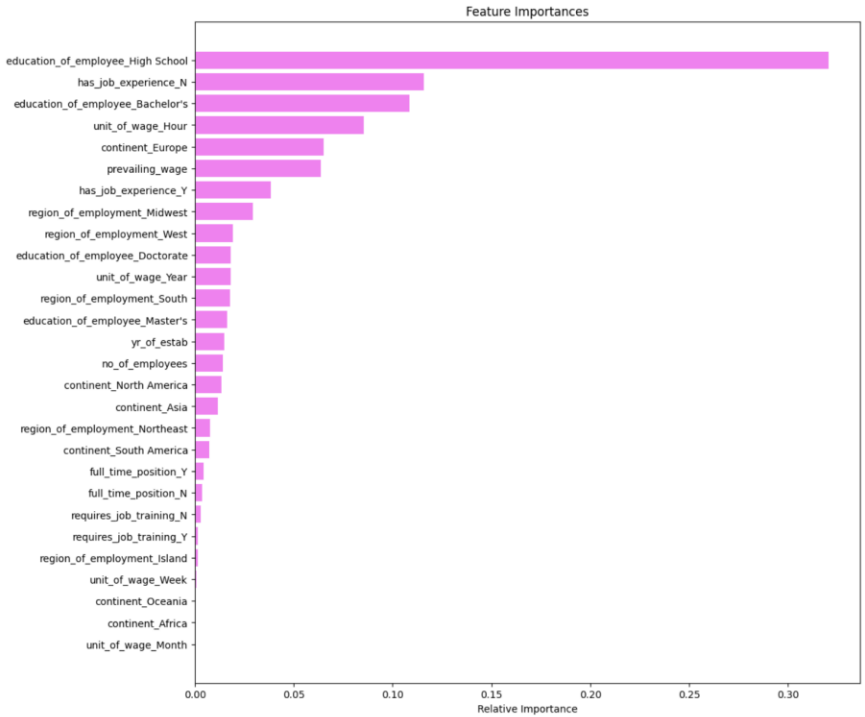


Data Preprocessing

- Encode categorical features
  - Case status was converted to a binary value (Certified: 1, Denied: 0)
- Create dummies for X
- Convert negative values to positive values in the no. of employees column

Model Performance Summary

- Data was split 70% Train, 30% Test, stratified, then fitted to the model
- A grid search was performed to find the best combination of hyperparameters.
- Final model was selected based on recall
  - High recall ensures all potential candidates who should be approved are indeed approved.
- The tuned decision tree classifier performed the best with a recall of 0.93 on the test set.



Model Performance Summary

- Summary of key performance metrics for training and test data of final model.
- **Training metrics** (Tuned Decision Tree)

	Accuracy	Recall	Precision	F1
0	0.712548	0.931923	0.720067	0.812411

- **Testing metrics** (Tuned Decision Tree)

	Accuracy	Recall	Precision	F1
0	0.706567	0.930852	0.715447	0.809058

Model Performance Summary

- Summary of key performance metrics for training and test data of all models.

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	0.983797	0.995234	1.0	0.772090	0.738058	0.755270	0.757849	0.753756	0.840884	0.758971	0.766764
Recall	1.0	0.931923	0.984639	0.999496	1.0	0.900865	0.886259	0.887770	0.883657	0.885671	0.930664	0.889532	0.873248
Precision	1.0	0.720067	0.991044	0.993409	1.0	0.788190	0.760937	0.777418	0.782095	0.776894	0.846400	0.780339	0.796982
F1	1.0	0.812411	0.987831	0.996443	1.0	0.840769	0.818830	0.828938	0.829780	0.827724	0.886534	0.831365	0.833373

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	0.660387	0.706567	0.704212	0.745029	0.719911	0.743590	0.734301	0.742282	0.745814	0.745029	0.726583	0.745160	0.748038
Recall	0.739275	0.930852	0.777081	0.881489	0.835651	0.881685	0.883252	0.880705	0.878355	0.881489	0.850735	0.879138	0.862880
Precision	0.748958	0.715447	0.779371	0.770021	0.766164	0.768482	0.758580	0.767628	0.772305	0.770021	0.765826	0.771267	0.782277
F1	0.744085	0.809058	0.778225	0.821993	0.799400	0.821201	0.816182	0.820288	0.821923	0.821993	0.806050	0.821677	0.820604

APPENDIX

Data Background and Contents

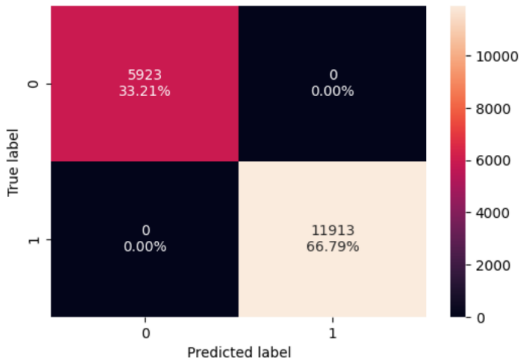
- Dataset Overview: The dataset contains attributes related to visa applications for employees and their potential employers, used for analyzing and predicting visa approval outcomes.

- **Identifiers:**
  - case_id: Unique identifier for each visa application.
- **Employee Information:**
  - continent: The applicant's continent of origin.
  - education_of_employee: The educational qualifications of the applicant.
  - has_job_experience: Indicates if the applicant has prior job experience (Y for yes, N for no).
  - requires_job_training: Specifies if the applicant needs job training (Y for yes, N for no).
- **Employer Details:**
  - no_of_employees: The total number of employees working for the employer.
  - yr_of_estab: The establishment year of the employer's company.
- **Employment Specifics:**
  - region_of_employment: The intended region of employment within the US for the foreign worker.
  - prevailing_wage: The standard wage for similar positions in the intended area of employment.
  - unit_of_wage: The time unit for the prevailing wage (Hourly, Weekly, Monthly, Yearly).
  - full_time_position: Whether the job is full-time (Y) or part-time (N).
- **Outcome Variable:**
  - case_status: The result of the visa application, indicating certification or denial.

Model Building - Bagging

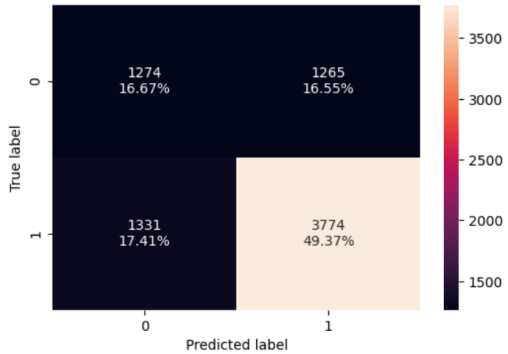
- All models were built with:
  - A 70/30 split of training and test sets
  - Stratified data
  - Random State = 1
- Decision Tree Metrics
  - All instances in train were correctly predicted
  - Possible overfitting
  - Considerable number of false positives and false negatives in test metrics
  - Model is not reliable

Training



	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Test

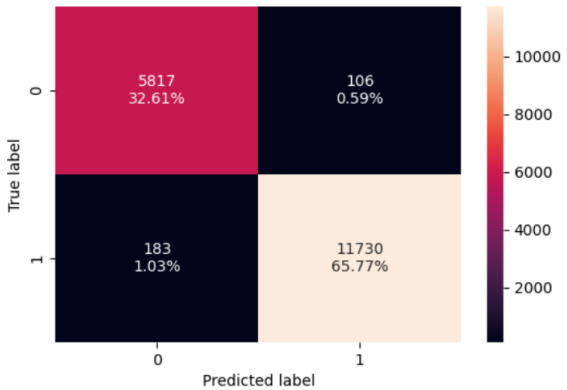


	Accuracy	Recall	Precision	F1
0	0.660387	0.739275	0.748958	0.744085

Model Building - Bagging

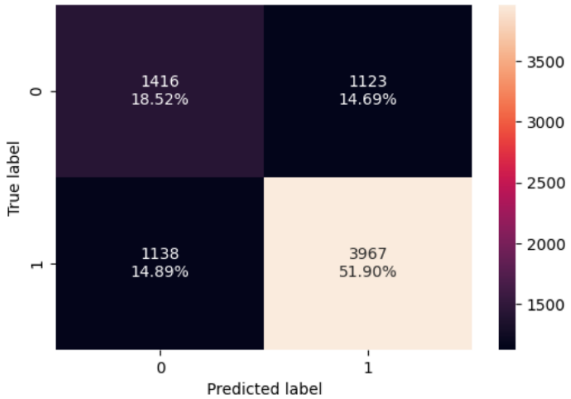
- Bagging Classifier Metrics
  - High accuracy, recall and precision in train set.
  - Significant drops in metrics in test set, indicating overfitting issues
  - Needs tuning

Training



	Accuracy	Recall	Precision	F1
0	0.983797	0.984639	0.991044	0.987831

Test

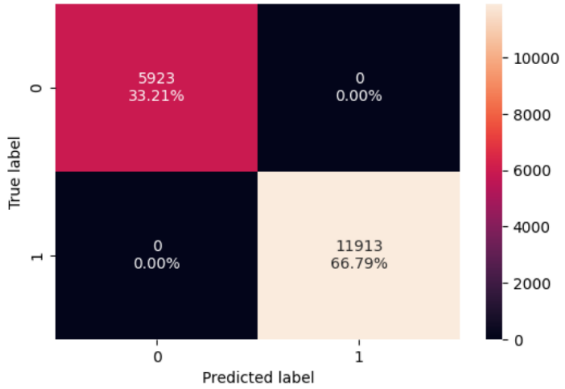


	Accuracy	Recall	Precision	F1
0	0.704212	0.777081	0.779371	0.778225

Model Building - Bagging

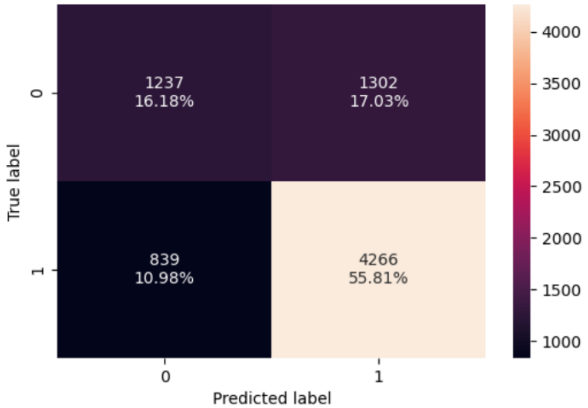
- Random Forest Metrics
  - No misclassifications in train set
  - Significant drop in performance in test set
  - Similar results to decision tree model
  - Unreliable

Training



	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Test

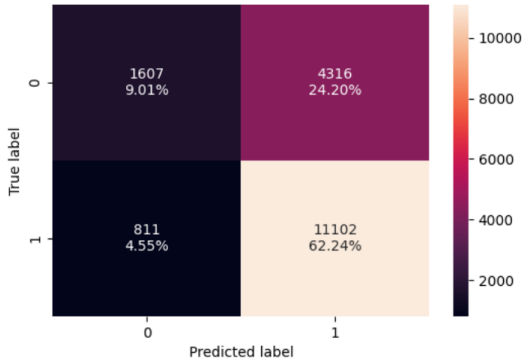


	Accuracy	Recall	Precision	F1
0	0.719911	0.835651	0.766164	0.7994

Model Improvement - Bagging

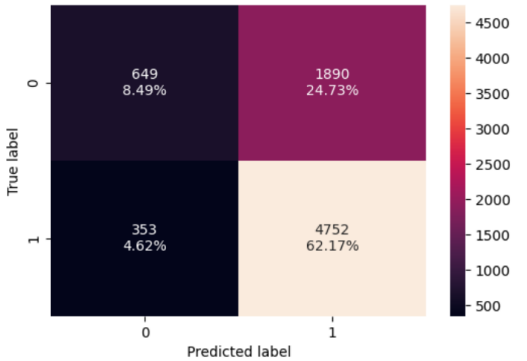
- Decision Tree Tuned Metrics
  - Tuned using GridSearch
  - Parameters tuned:
    - max_depth = 10
    - min_samples_leaf = 3
    - max_leaf_nodes = 2
    - min_impurity_decrease = 0.001
  - Bias towards class 1
  - Model ability to generalize is consistent
  - Performance increase of 0.064 (+8.7%) in F1-Score after tuning

Training



	Accuracy	Recall	Precision	F1
0	0.712548	0.931923	0.720067	0.812411

Test

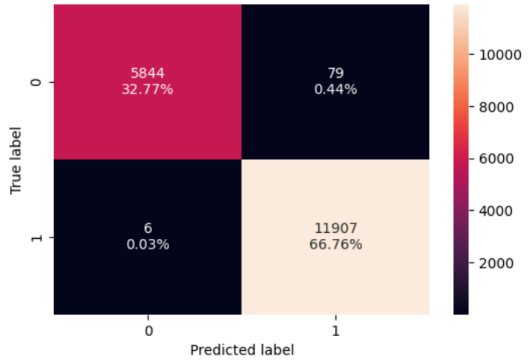


	Accuracy	Recall	Precision	F1
0	0.706567	0.930852	0.715447	0.809058

Model Improvement - Bagging

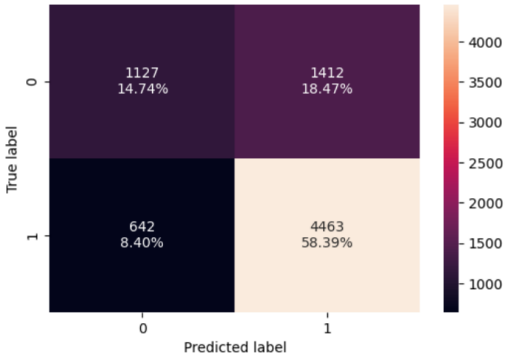
- Bagging Classifier Tuned Metrics
  - Tuned using GridSearch
  - Parameters tuned:
    - max_samples = 0.7
    - max_features = 0.7
    - n_estimators = 100
  - Higher misclassification rate on test data
  - Discrepancy between train metrics and test metrics could indicate overfitting
  - Performance increase of 0.044(+5.32%) in F1-Score after tuning

Training



	Accuracy	Recall	Precision	F1
0	0.995234	0.999496	0.993409	0.996443

Test

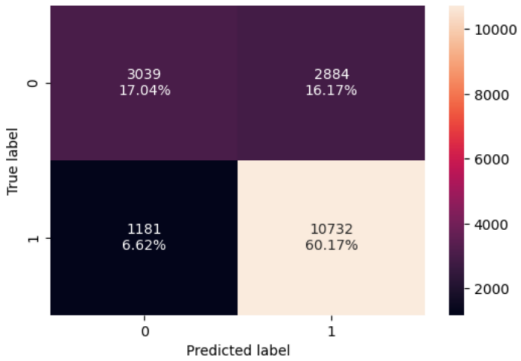


	Accuracy	Recall	Precision	F1
0	0.745029	0.881489	0.770021	0.821993

Model Improvement - Bagging

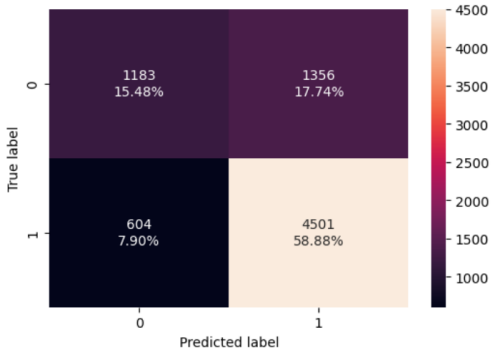
- Random Forest Tuned Metrics
  - Tuned using GridSearch
  - Parameters tuned:
    - max_depth = 10
    - max_features = log2
    - min_samples_split = 5
    - n_estimators = 30
  - Performance increase of 0.021(+2.65%) in F1-Score after tuning

Training



	Accuracy	Recall	Precision	F1
0	0.77209	0.900865	0.78819	0.840769

Test

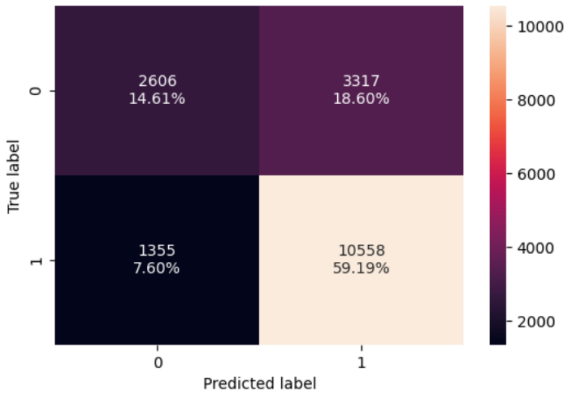


	Accuracy	Recall	Precision	F1
0	0.74359	0.881685	0.768482	0.821201

Model Building - Boosting

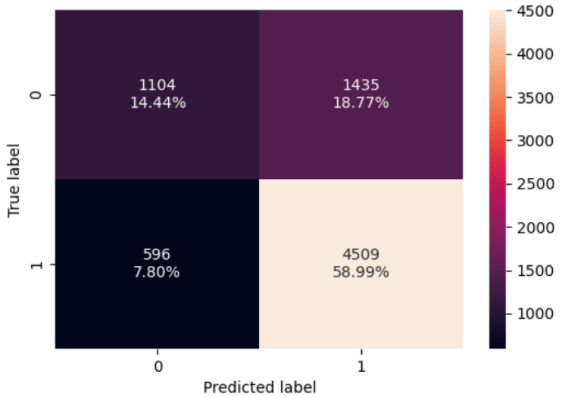
- AdaBoost Classifier Metrics
  - Model is better at predicting positive class in both train and test data
  - Good generalization

Training



	Accuracy	Recall	Precision	F1
0	0.738058	0.886259	0.760937	0.81883

Test

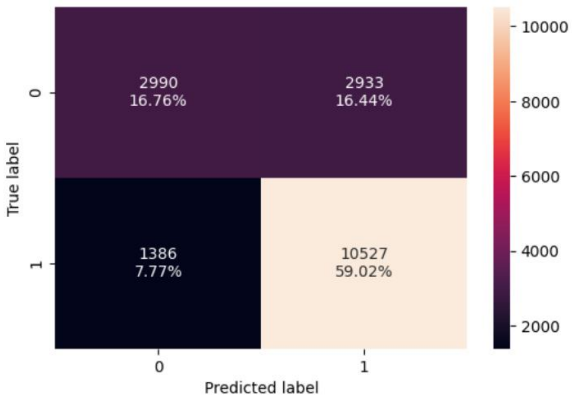


	Accuracy	Recall	Precision	F1
0	0.734301	0.883252	0.75858	0.816182

Model Building - Boosting

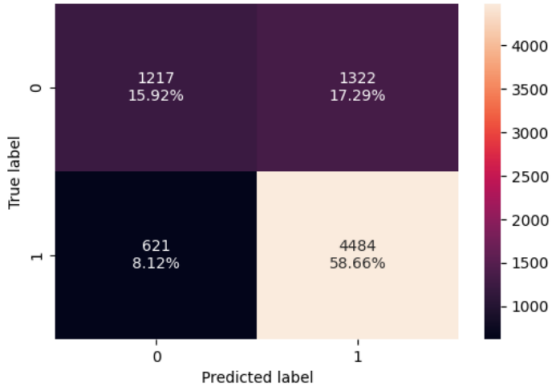
- Gradient Boosting Classifier Metrics
  - Reasonable generalization
  - Higher false positive rate

Training



	Accuracy	Recall	Precision	F1
0	0.757849	0.883657	0.782095	0.82978

Test

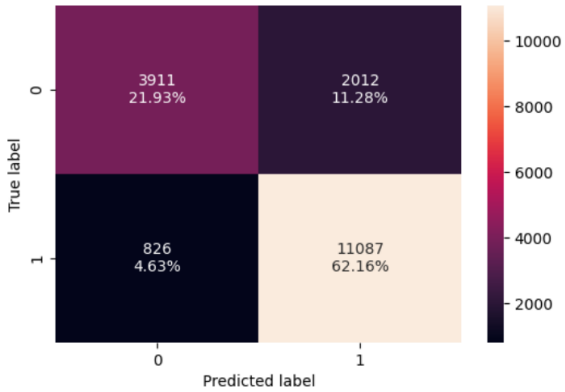


	Accuracy	Recall	Precision	F1
0	0.745814	0.878355	0.772305	0.821923

Model Building - Boosting

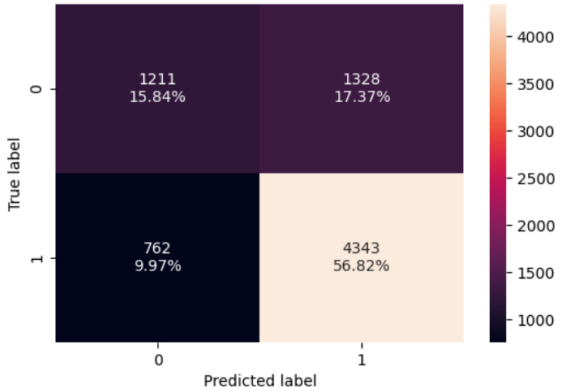
- XGBoost Classifier Metrics
  - High rate of correct predictions for class 1, but struggles with class 0
  - Noticeable difference between training and testing accuracy
  - Potential overfitting

Training



	Accuracy	Recall	Precision	F1
0	0.840884	0.930664	0.8464	0.886534

Test

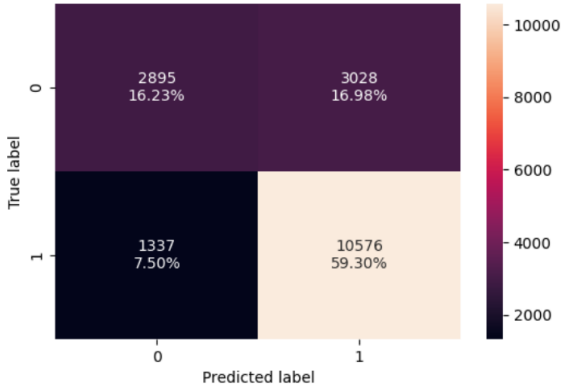


	Accuracy	Recall	Precision	F1
0	0.726583	0.850735	0.765826	0.80605

Model Improvement - Boosting

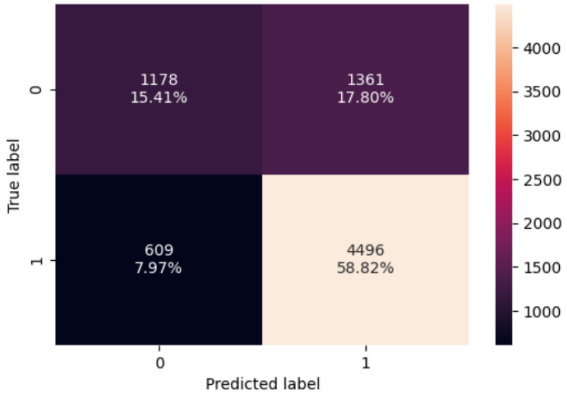
- AdaBoost Classifier Tuned Metrics
  - Tuned using GridSearch
  - Parameters tuned:
    - base_estimator = DecisionTreeClassifier
    - n_estimators = 100
    - learning_rate = 0.06
  - Higher FP rate than FN
  - Good generalization
  - Performance increase of 0.004 (+0.5%) in F1-Score after tuning

Training



	Accuracy	Recall	Precision	F1
0	0.75527	0.88777	0.777418	0.828938

Test

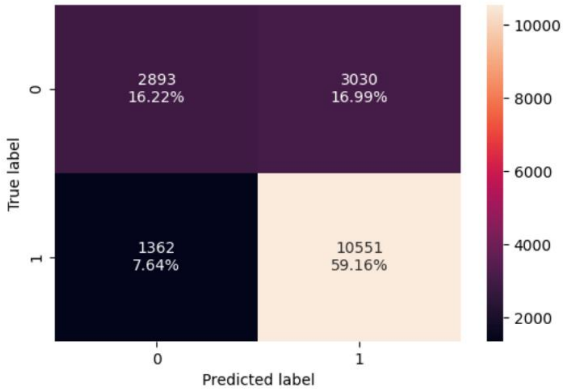


	Accuracy	Recall	Precision	F1
0	0.742282	0.880705	0.767628	0.820288

Model Improvement - Boosting

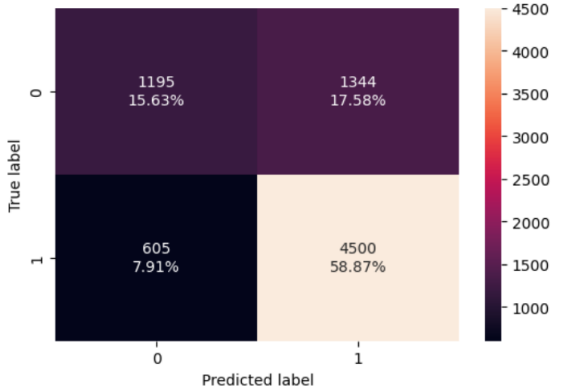
- GradientBoosting Classifier Tuned Metrics
  - Tuned using GridSearch
  - Parameters tuned:
    - n_estimators = 50
    - subsample = 0.9
    - max_features = 0.7
  - Reasonable generalization
  - False positive rate is higher than false negative rate
  - No noteable performance increase after tuning

Training



	Accuracy	Recall	Precision	F1
0	0.753756	0.885671	0.776894	0.827724

Test

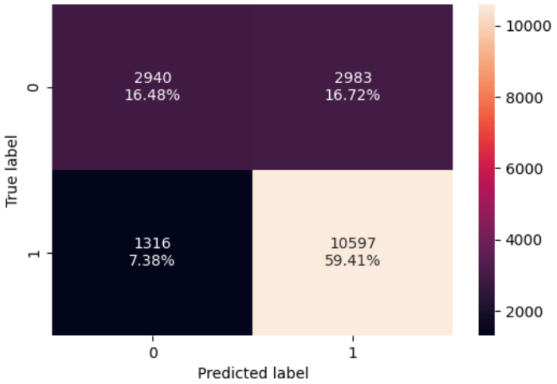


	Accuracy	Recall	Precision	F1
0	0.745029	0.881489	0.770021	0.821993

Model Improvement - Boosting

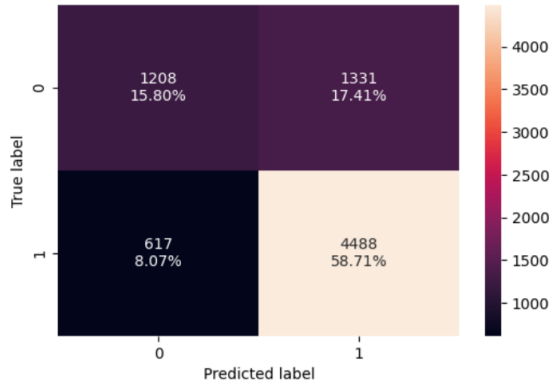
- XGBoost Classifier Tuned Metrics
  - Tuned using GridSearch
  - Parameters tuned:
    - n_estimators = 50
    - learning_rate = 0.05
    - gamma = 3
    - scale_pos_weight
    - subsample
  - Biased towards predicting class 1
  - Reasonable generalization
  - Higher FP rate compared to FN
  - Performance increase of 0.0156 (+1.9%) in F1-Score after tuning

Training



	Accuracy	Recall	Precision	F1
0	0.758971	0.889532	0.780339	0.831365

Test



	Accuracy	Recall	Precision	F1
0	0.74516	0.879138	0.771267	0.821677

Great Learning

Happy Learning !