EasyVisa Project Business Presentation

Prepared by Dheeraj Mishra

As of February 5, 2022

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Contents

- Business Problem Overview
- Data Overview
- Exploratory Data Analysis (EDA)
- Model Evaluation Criterion
  - What's important?
  - How to reduce losses?
- Model Performance Summary and Comparison
  - Decision Tree, Bagging Classifier, Random Forest, Adaboost Classifier, Gradient Boost Classifier, XGBoost Classifier, and Stacking Classifier
- Business Insights and Recommendations

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Business Problem Overview and Solution Approach

- Office of Foreign Labor Certification (OFLC) processes visa certification applications for employers seeking to bring foreign workers into the United States.
- OFLC grants visa 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.
- The process of reviewing every case is becoming a tedious task as the number of applicants is increasing every year.
- EasyVisa firm has been tasked to analyze the data provided and, with the help of a classification model:
  - Facilitate the process of visa approvals.
  - Recommend a suitable profile for the applicants for whom the visa should be certified or denied based on the drivers that significantly influence the case status.
- The task at hand is to analyze the data provided and find which factors have a high influence on visa certifications. Build a predictive classification model that can predict which visa application will get certified or denied.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

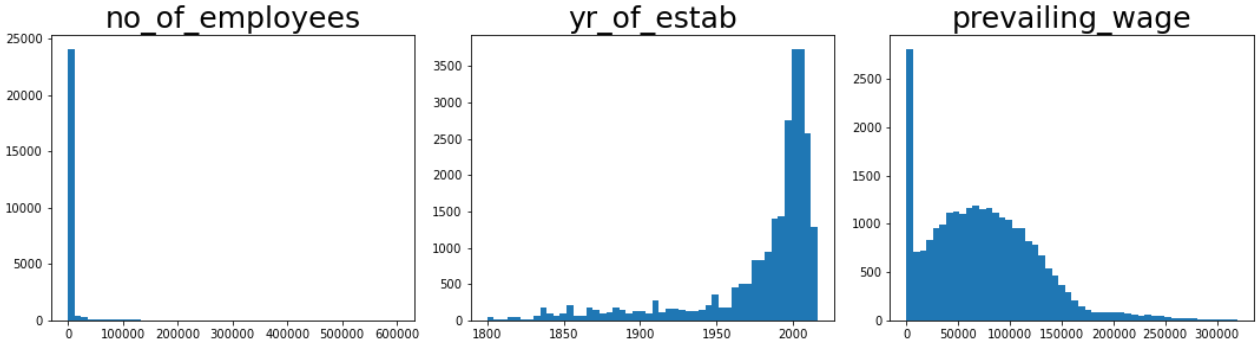
Data Overview

- The data contains information about 25,480 visa applications bookings and their following characteristics.
    - case_id: ID of each visa application
    - continent: continent the employee
    - education_of_employee: education of the employee
    - has_job_experience: Does the employee has any job experience?
    - requires_job_training: Does the employee require any job training?
    - 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: intended region of employment in the US.
    - prevailing_wage: Average wage paid for similar service.
    - unit_of_wage: Hourly, Weekly, Monthly, and Yearly.
    - full_time_position: Full Time Position or Part Time Position
    - case_status: Flag indicating if the Visa was certified or denied
- Most features are categorical and will be encoded using the dummy one-hot value for model building
- No Missing values and No Duplicate rows found in data set.
- Negative values are present in the No of Employees. Absolute value will be used to handle negatives.
- Extreme Outliers present in No of Employees, Prevailing Wages and Year of Establishment
    - We will apply the log transformation to deal with skewness in the data.

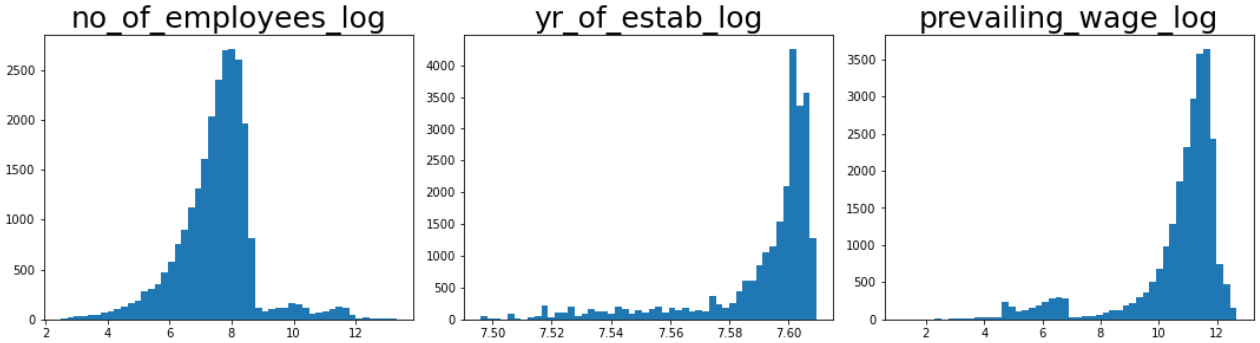
Exploratory Data Analysis

- The No of Employees, Year of Establishment and Prevailing Wage are heavily skewed.
- Log Transformation will be applied to reduce the extreme skewness.

Before Treatment

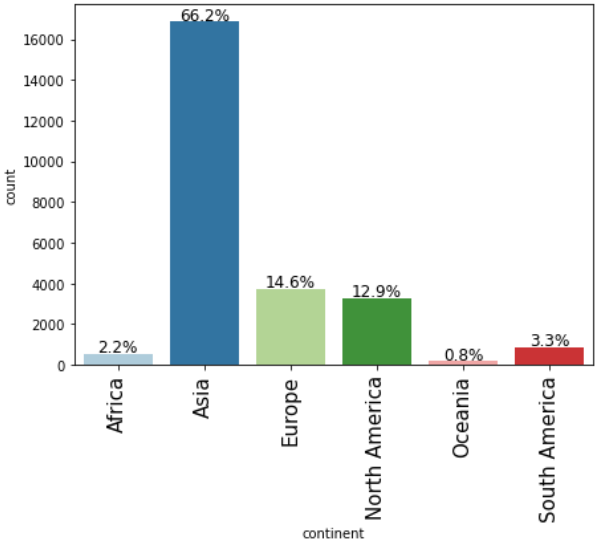


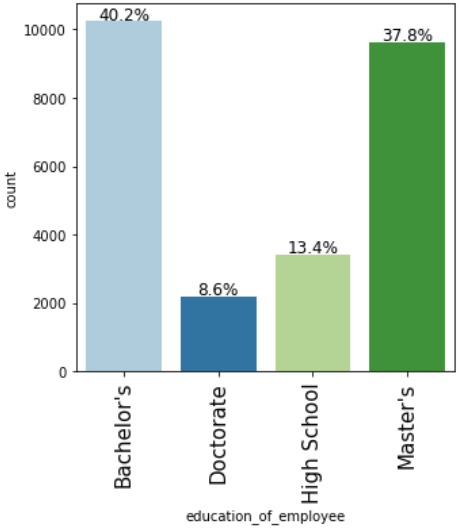
After Treatment

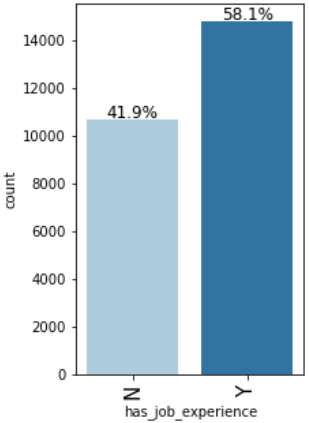


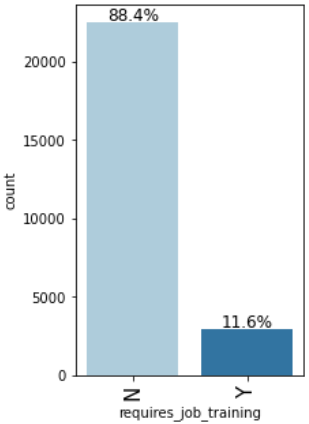
Exploratory Data Analysis

- Visa applications from Asia dominates the pool by 66%
- Visa candidates with either Bachelors or Masters degree represents 78% of all application.
- 88% applicants don't require job training





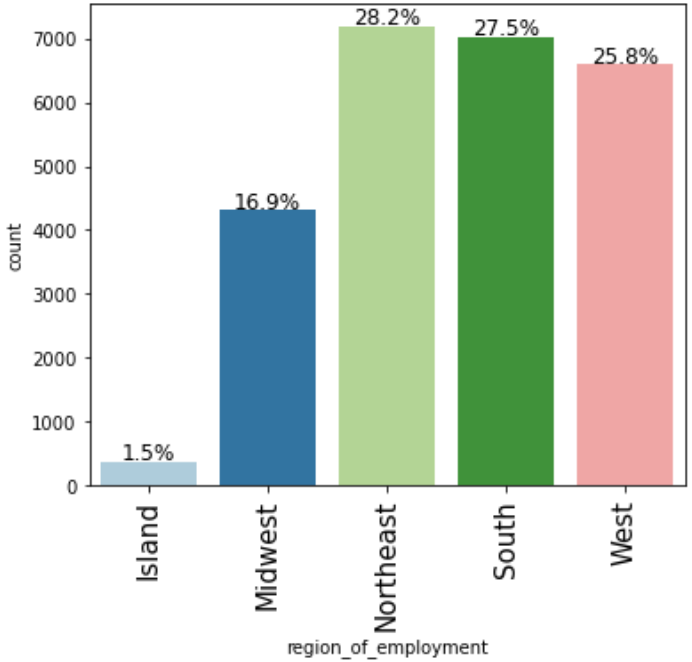


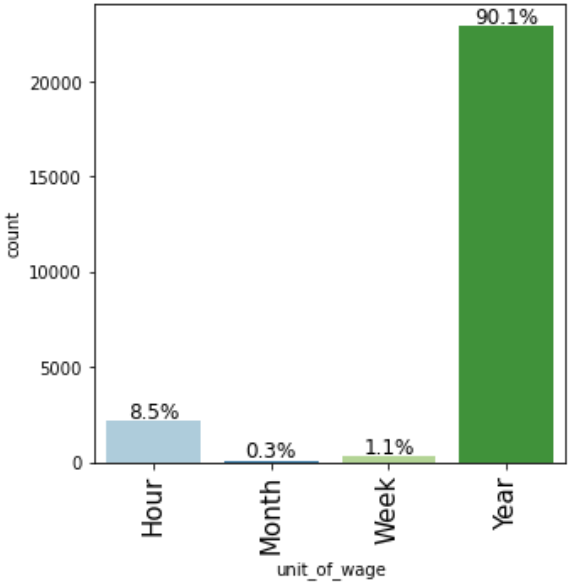


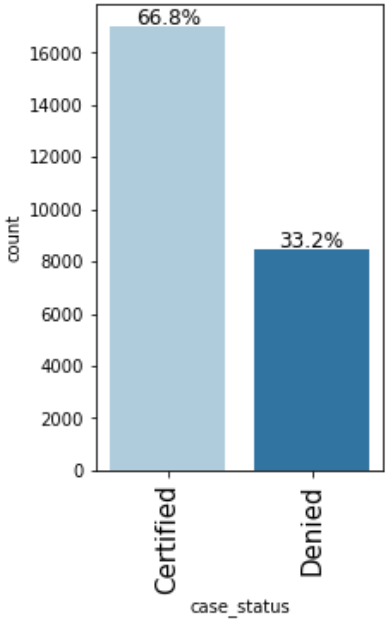
Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Exploratory Data Analysis

- Northeast region has the highest number of visa application followed by South and West
- Prevailing wage are reported in Yearly unit for 90% of the visa applications. Wages reported in Hourly introduce skewness of data.
- 2/3 of all visa applications are approved

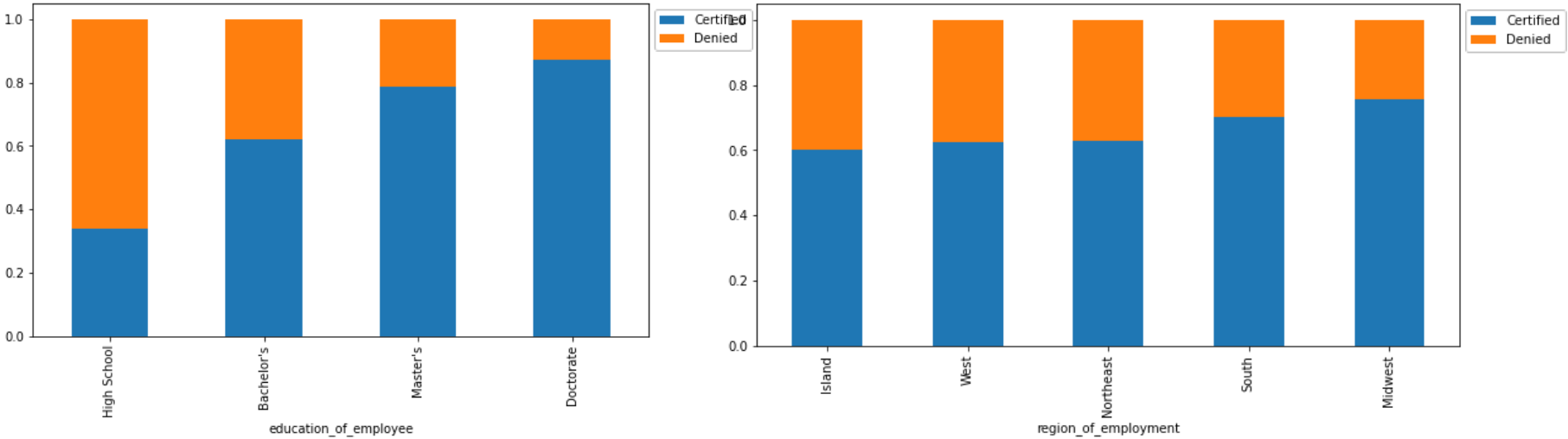






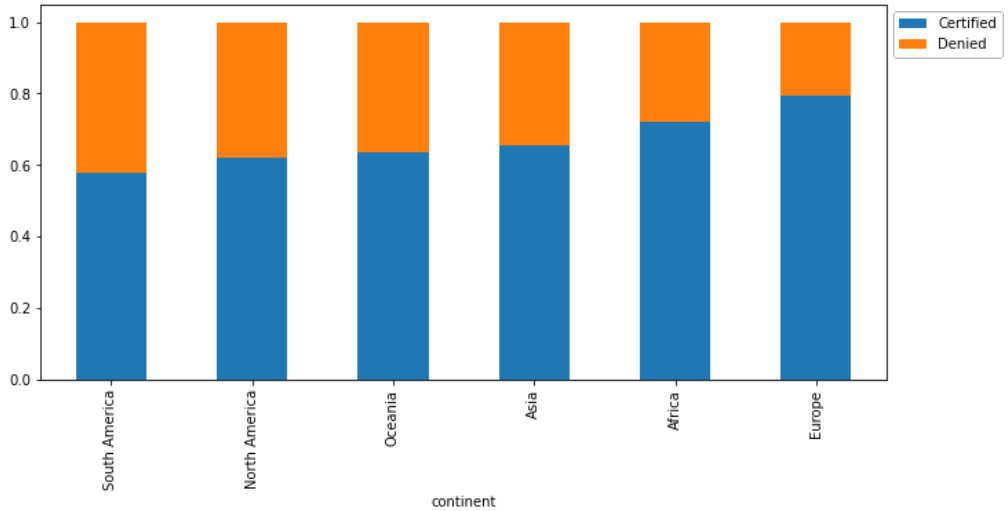
Exploratory Data Analysis

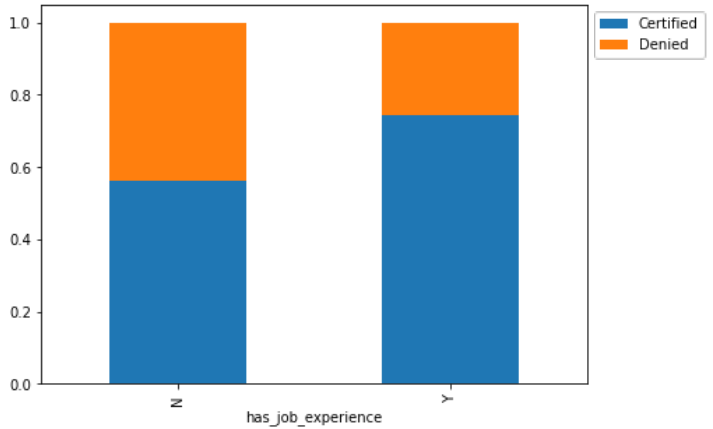
- Higher education has positive impact on visa certification
- Visa application with Doctorate have the highest approval rate of 87%, followed by 78% with Master's degree and 62% with Bachelor's degree
- Visa certification has higher percentage (~76%) for Midwest region
- Only 2/3 of all visa applications are approved



Exploratory Data Analysis

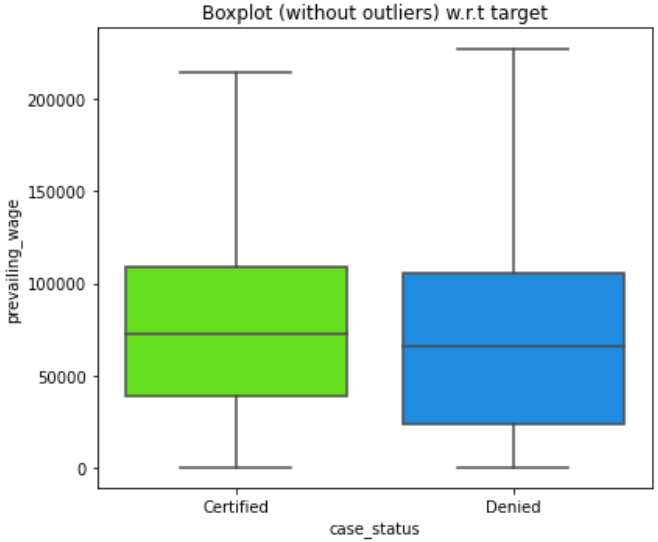
- Visa status vary across different continents. Visa Denied is lowest (20%) for applicants from Europe region
- Prior Job Experience plays a significant factor in visa certification.
- 65% of visa certification is awarded to applications with prior job experience.
- Only 8.5% of applicants with prior job experience require some sort of training, compared to 16% of Non-Experience job applicants that require training.

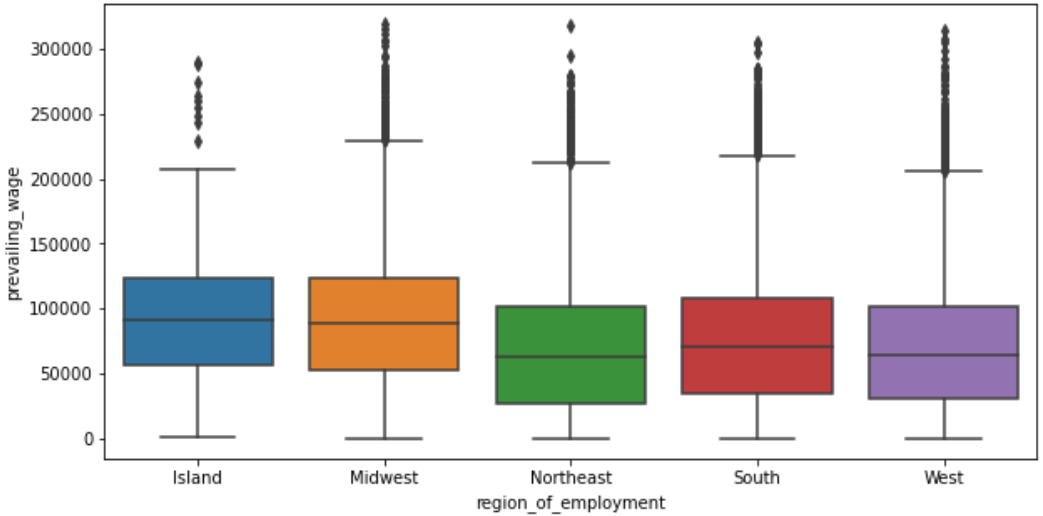




Exploratory Data Analysis

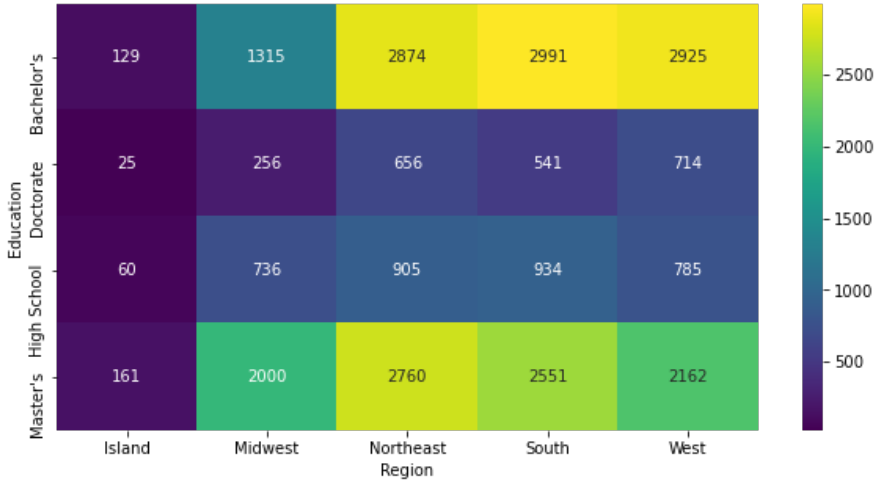
- Jobs for lower prevailing wage have higher chance of visa denial. It make sense as US visa process wants to bring talented and qualified individuals from outside to augment it's workforce to remain competitive.
- Median prevailing wage is around $70,000 yearly.
- Prevailing wage is similar across Northeast, South and West regions. Wages are higher in Island and Midwest regions.





Exploratory Data Analysis

- There is week correlation between prevailing wages, years of establishment, and no of employees.
- Different regions have different requirements of talent having diverse educational backgrounds
  - West region has more requirement for workers with Doctorate degree compare to other regions
  - Midwest region has less requirements for workers with Bachelors degree compare to other regions
  - Northeast region has highest requirement for Masters degree compare to other regions



Model Evaluation Criterion

- **Model can make wrong predictions as:**
  - **False Positive:** Model predicts that the visa application will get certified but in reality, the visa should get denied
  - **False Negative:** Model predicts that the visa application will not get certified but in reality, the visa should get certified.

- **Which case is more important?**
  - Both the cases are important as:
  - If a visa is certified when it had to be denied a wrong employee will get the job position while US citizens will miss the opportunity to work on that position.
  - If a visa is denied when it had to be certified the U.S. will lose a suitable human resource that can contribute to the economy.

- **How to reduce the losses?**
  - Hotel would want `F1 Score` to be maximized, greater the F1 score higher are the chances of minimizing False Negatives and False Positives.
  - F1 score is defined as the harmonic mean between precision and recall
  - We will use balanced class weights so that model focuses equally on both classes.

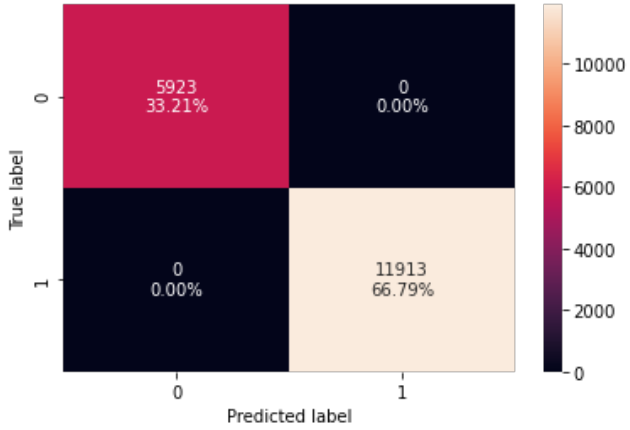
Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Model Performance Summary

- We want to predict which visa application is going to certified (approved), based on the characteristics provided to us.
- Used different ensemble techniques to build models
  - Bagging
  - Boosting
  - Stacking
- Built several models using various different techniques:
  - Decision Tree, Bagging Classifier, Random Forest, Adaboost Classifier, Gradient Boost Classifier, XGBoost Classifier, and Stacking Classifier
- Built using the train data and checked the performance on test data to understand the predictive power of our models.
- Models fill be further improved using the hyper parameters tuning to find best Classification model with highest F1 score
- The most significant predictors of the booking cancellation are:
  - Education of Employee (high-school).
  - Has job experience (Y)
  - Prevailing wage
  - Education of Employee (Masters-degree)
  - No of employee
  - Year of Establishment
  - Unit of Wage (in Year)
  - Education of Employee (Doctorate-degree)
  - Continent (Europe)

Model Performance Summary – Decision Tree Model

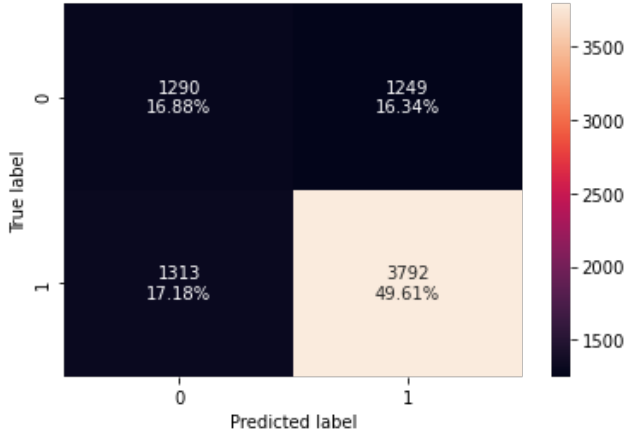
Training-Data Performance



Accuracy	Recall	Precision	F1
1.0	1.0	1.0	1.0

Overfitting

Testing-Data Performance



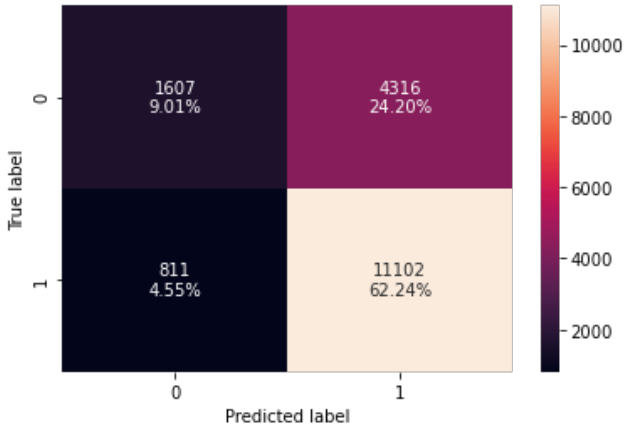
Accuracy	Recall	Precision	F1
0.664835	0.742801	0.752232	0.747487

Hyper Parameter Tuned

Model Performance Summary – Decision Tree Model

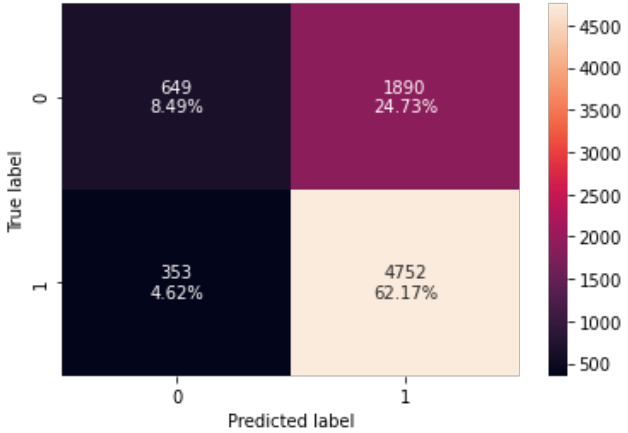
Great Learning
POWER AHEAD

Training-Data Performance



Accuracy	Recall	Precision	F1
0.712548	0.931923	0.720067	0.812411

Testing-Data Performance



Accuracy	Recall	Precision	F1
0.706567	0.930852	0.715447	0.809058

Generalized Performance

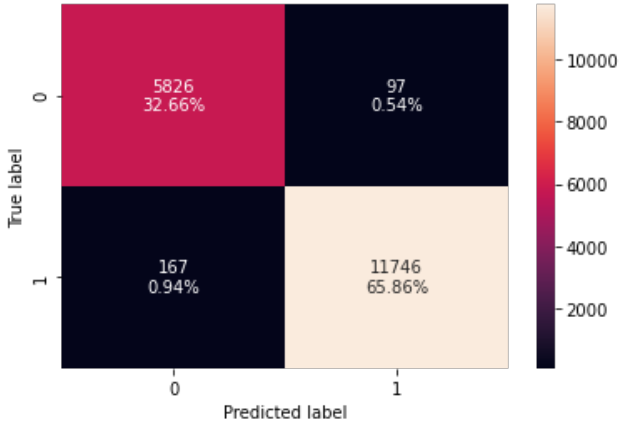
Hyper Parameters

max_depth : np.arange(10, 30, 5),
min_samples_leaf: [3, 5, 7],
max_leaf_nodes: [2, 3, 5],
min_impurity_decrease: [0.0001, 0.001]
grid search with n_jobs = -1

```
DecisionTreeClassifier(class_weight='balanced', max_depth=10,
max_leaf_nodes=2, min_impurity_decrease=0.0001,
min_samples_leaf=3, random_state=1)
```

Model Performance Summary – Bagging Classifier

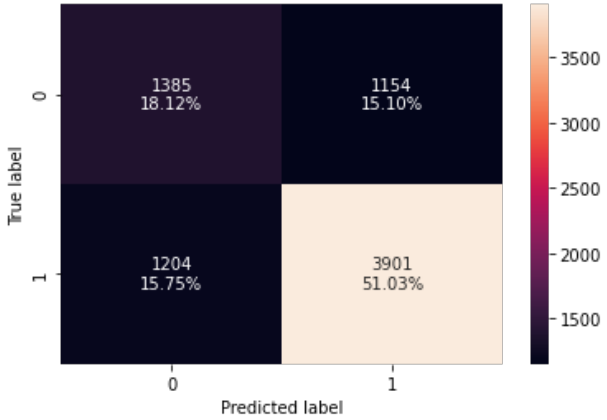
Training-Data Performance



Accuracy	Recall	Precision	F1
0.985198	0.985982	0.99181	0.988887

Overfitting

Testing-Data Performance



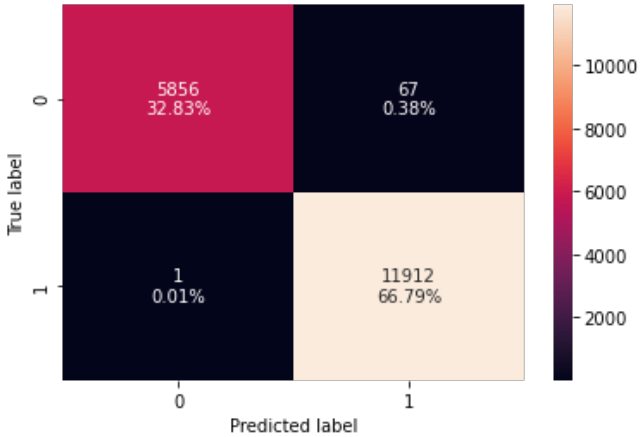
Accuracy	Recall	Precision	F1
0.691523	0.764153	0.771711	0.767913

Generalized Performance

Hyper Parameter Tuned

Model Performance Summary – Bagging Classifier

Training-Data Performance



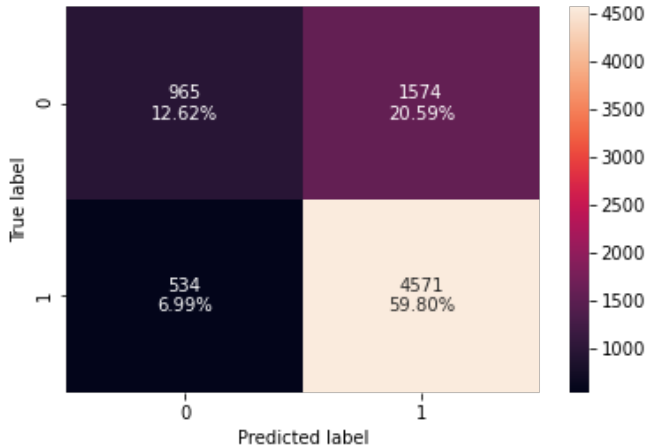
Accuracy	Recall	Precision	F1
0.996187	0.999916	0.994407	0.997154

Overfitting

Hyper Parameters

max_samples: [0.7, 0.8, 0.9],
max_features: [0.7, 0.8, 0.9],
n_estimators: np.arange(90, 120, 10)
cv =5

Testing-Data Performance



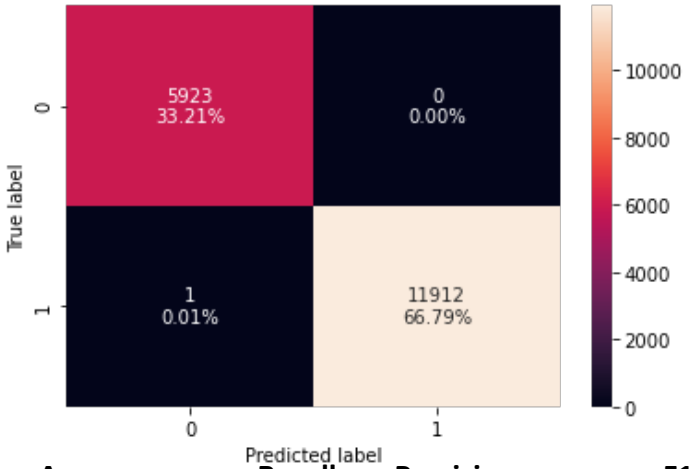
Accuracy	Recall	Precision	F1
0.724228	0.895397	0.743857	0.812622

Generalized Performance

```
BaggingClassifier(max_features=0.7, max_samples=0.7,
n_estimators=100, random_state=1)
```

Model Performance Summary – Random Forest Classifier

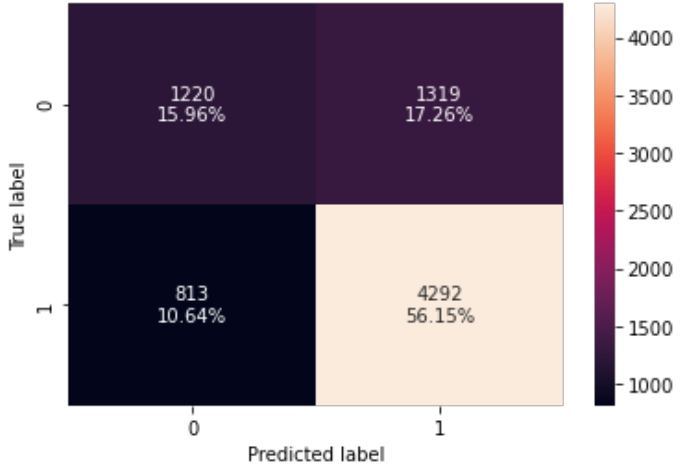
Training-Data Performance



Accuracy	Recall	Precision	F1
0.999944	0.999916	1.0	0.999958

Overfitting

Testing-Data Performance



Accuracy	Recall	Precision	F1
0.721088	0.840744	0.764926	0.801045

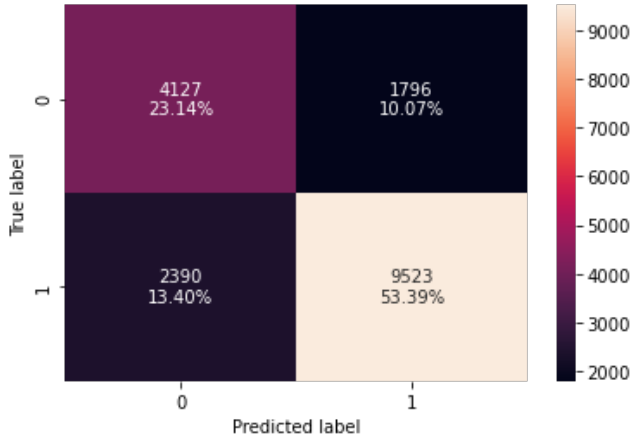
Generalized Performance

Hyper Parameter Tuned

Model Performance Summary – Random Forest Classifier

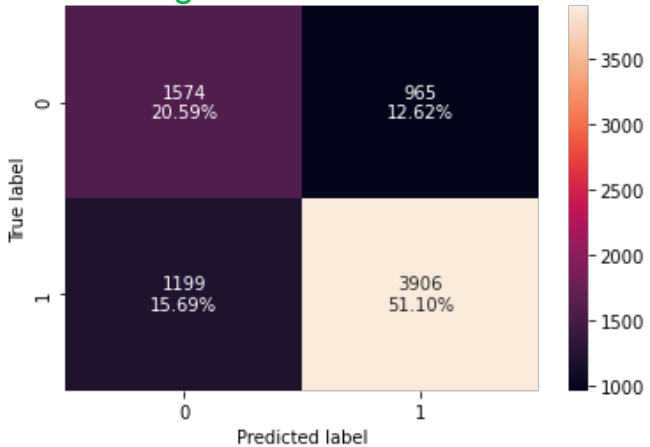
Great Learning
POWER AHEAD

Training-Data Performance



Accuracy	Recall	Precision	F1
0.765306	0.799379	0.841329	0.819817

Testing-Data Performance



Accuracy	Recall	Precision	F1
0.716902	0.765132	0.801889	0.783079

Generalized Performance

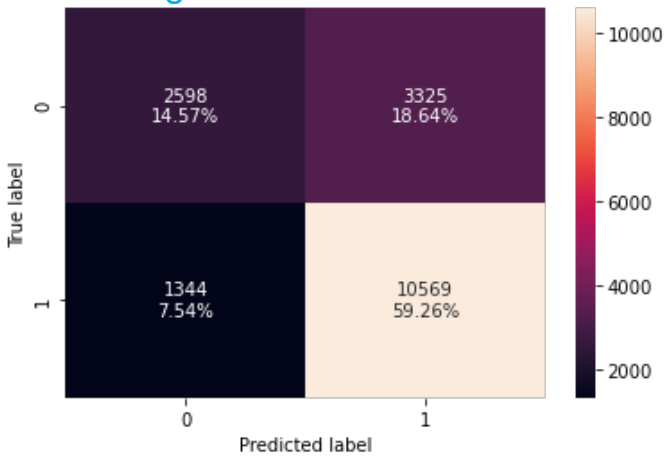
Hyper Parameters

max_depth: list(np.arange(5, 15, 5))
max_features: ["sqrt", "log2"]
min_samples_split: [3, 5, 7]
n_estimators: np.arange(10, 40, 10)
cv=5 . n_jobs=-1

```
RandomForestClassifier(class_weight='balanced', max_depth=10,
max_features='sqrt', min_samples_split=3,
n_estimators=30, random_state=1)
```

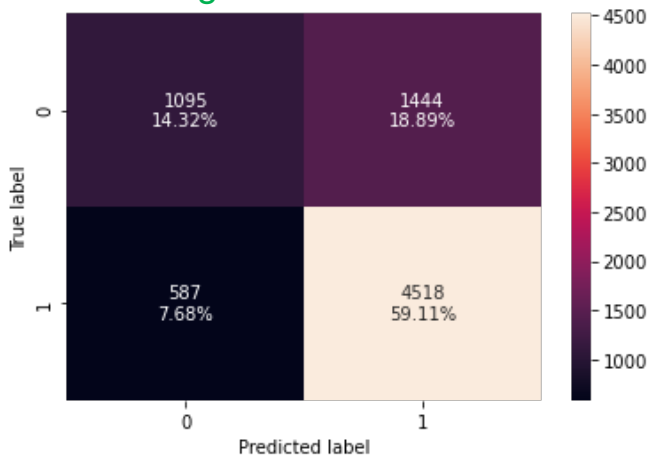
Model Performance Summary – AdaBoost Classifier

Training-Data Performance



Accuracy	Recall	Precision	F1
0.738226	0.887182	0.760688	0.81908

Testing-Data Performance



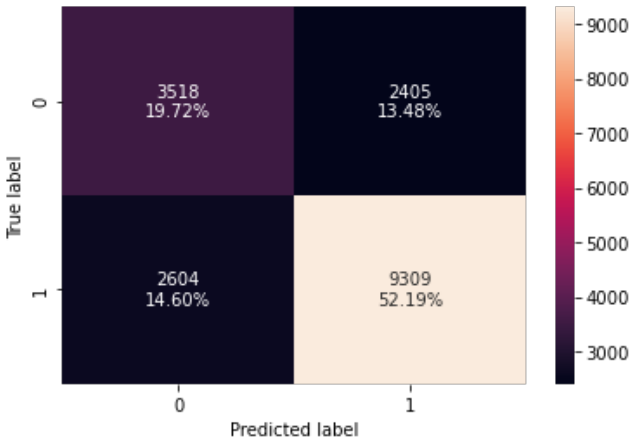
Accuracy	Recall	Precision	F1
0.734301	0.885015	0.757799	0.816481

Generalized Performance

Hyper Parameter Tuned
Model Performance Summary – AdaBoost Classifier

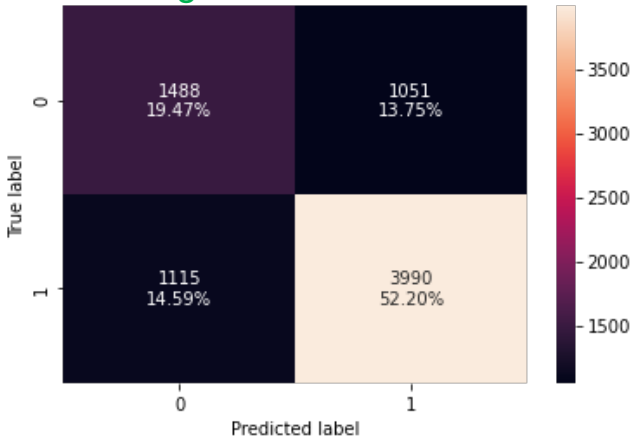
Great Learning
POWER AHEAD

Training-Data Performance



Accuracy	Recall	Precision	F1
0.719163	0.781415	0.79469	0.787997

Testing-Data Performance



Accuracy	Recall	Precision	F1
0.716641	0.781587	0.79151	0.786517

Generalized Performance

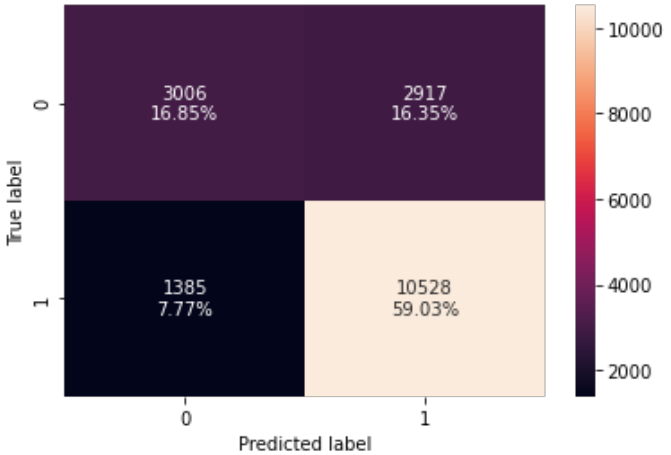
Hyper Parameters

base_estimator:
DecisionTreeClassifier, max_depth=1, class_weight=balanced
DecisionTreeClassifier, max_depth=2, class_weight=balanced
DecisionTreeClassifier, max_depth=3, class_weight=balanced
n_estimators: np.arange(60, 100, 10
learning_rate": np.arange(0.1, 0.4, 0.1), cv=5

```
AdaBoostClassifier(base_estimator=
DecisionTreeClassifier(class_weight='balanced',
max_depth=1, random_state=1), learning_rate=0.1,
n_estimators=90, random_state=1)
```

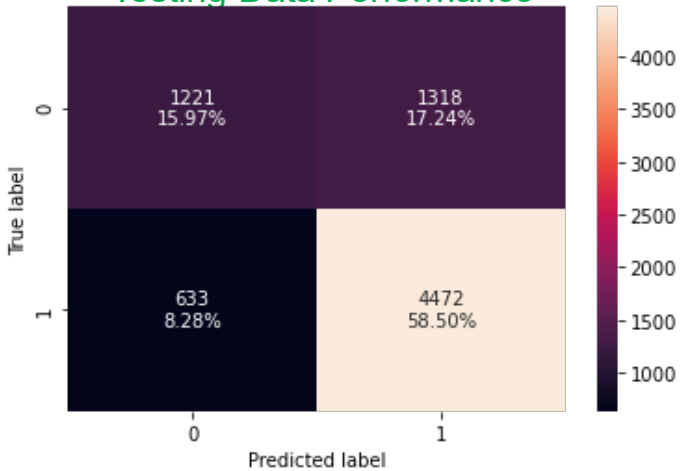
Model Performance Summary – Gradient Boosting Classifier

Training-Data Performance



Accuracy	Recall	Precision	F1
0.758802	0.88374	0.783042	0.830349

Testing-Data Performance



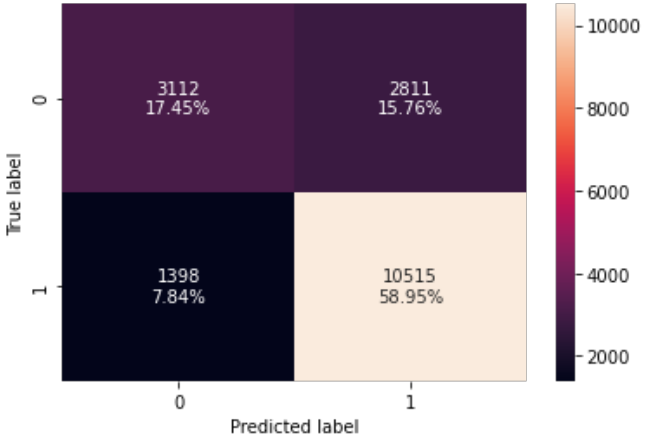
Accuracy	Recall	Precision	F1
0.744767	0.876004	0.772366	0.820927

Generalized Performance

Hyper Parameter Tuned

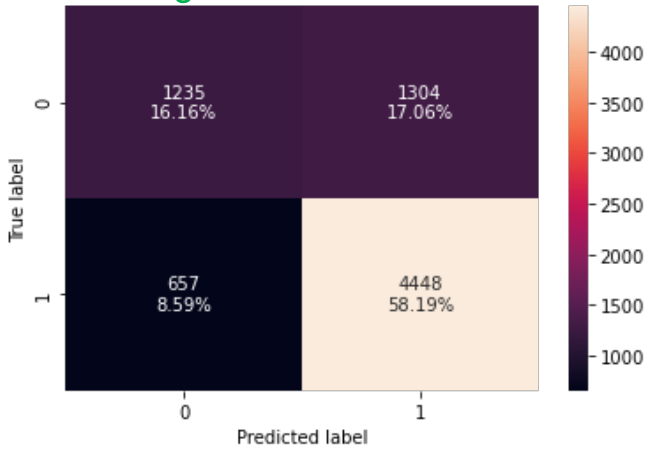
Model Performance Summary – Gradient Boosting Classifier`

Training-Data Performance



Accuracy	Recall	Precision	F1
0.764017	0.882649	0.789059	0.833234

Testing-Data Performance



Accuracy	Recall	Precision	F1
0.743459	0.871303	0.773296	0.819379

Generalized Performance

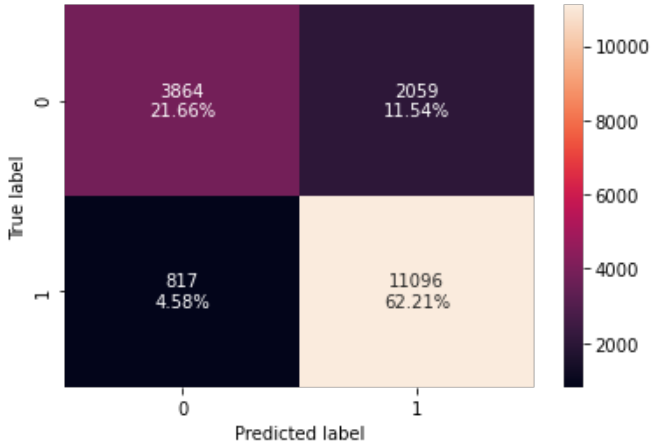
Hyper Parameters

n_estimators: [200, 250, 300
subsample": [0.8, 0.9, 1]
max_features": [0.7, 0.8, 0.9, 1]
learning_rate": np.arange(0.1, 0.4, 0.1)
cv = 5

```
GradientBoostingClassifier
(init=AdaBoostClassifier(random_state=1),
max_features=0.8, n_estimators=200, random_state=1,
subsample=1)
```

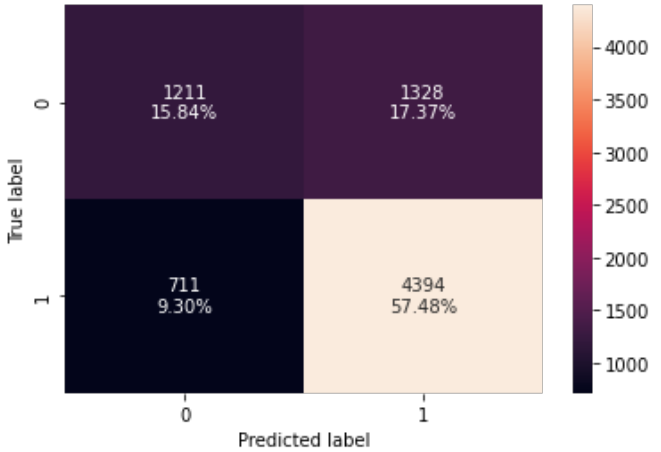
Model Performance Summary – XGBoost Classifier

Training-Data Performance



Accuracy	Recall	Precision	F1
0.838753	0.931419	0.843482	0.885272

Testing-Data Performance



Accuracy	Recall	Precision	F1
0.733255	0.860725	0.767913	0.811675

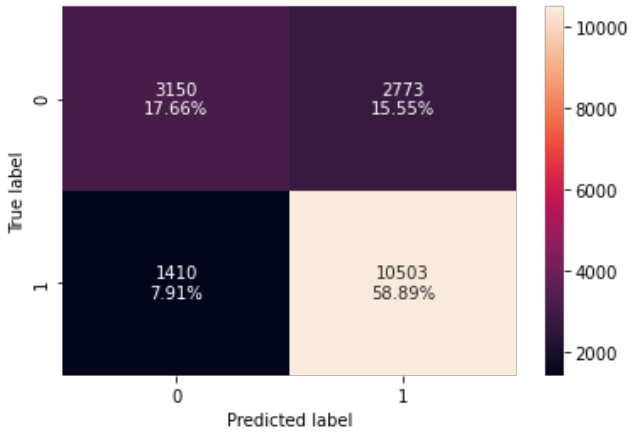
Generalized Performance

```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,colsample_bynode=1,colsample_
enable_categorical=False, eval_metric='logloss', gamma=0, gpu_id=-1, importance_type=None,
interaction_constraints='', learning_rate=0.300000012, max_delta_step=0, max_depth=6, min_child_weight=1,
missing=nan, monotone_constraints='()',n_estimators=100, n_jobs=8, num_parallel_tree=1, predictor='auto',
random_state=1, reg_alpha=0, reg_lambda=1, scale_pos_weight=1,subsample=1, tree_method='exact',
validate_parameters=1, verbosity=None)
```

Hyper Parameter Tuned

Model Performance Summary – XGBoost Classifier

Training-Data Performance

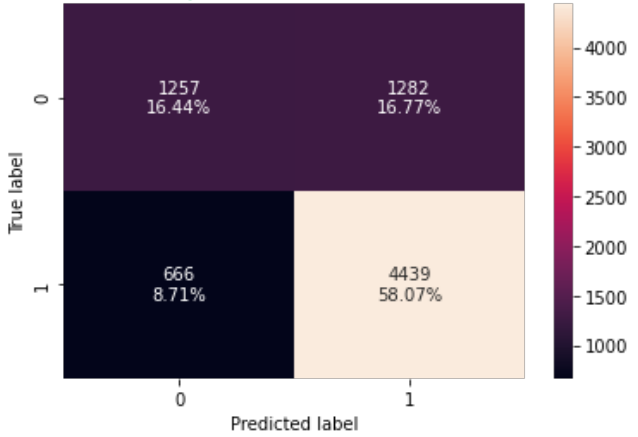


Accuracy	Recall	Precision	F1
0.765474	0.881642	0.791127	0.833935

Hyper Parameters

n_estimators: np.arange(150, 250, 50)
scale_pos_weight: [1, 2]
subsample: [0.7, 0.9, 1]
learning_rate: np.arange(0.1, 0.4, 0.1)
gamma: [1, 3, 5]
colsample_bytree: [0.7, 0.8, 0.9]
colsample_bylevel: [0.8, 0.9, 1]

Testing-Data Performance



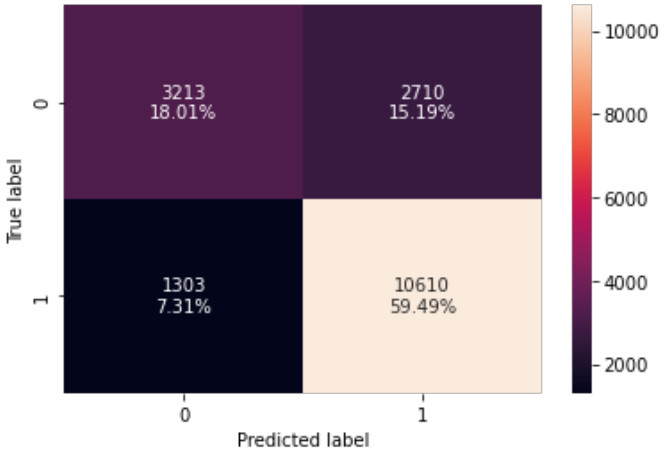
Accuracy	Recall	Precision	F1
0.74516	0.86954	0.775913	0.820063

Generalized Performance

```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1, colsample_bynode=1,
colsample_bytree=0.9, enable_categorical=False, eval_metric='logloss', gamma=5, gpu_id=-1,
importance_type=None, interaction_constraints='',learning_rate=0.1, max_delta_step=0, max_depth=6,
min_child_weight=1, missing=nan, monotone_constraints='()',n_estimators=200, n_jobs=8,
num_parallel_tree=1, predictor='auto', random_state=1, reg_alpha=0, reg_lambda=1,
scale_pos_weight=1, subsample=1, tree_method='exact', validate_parameters=1, verbosity=None)
```

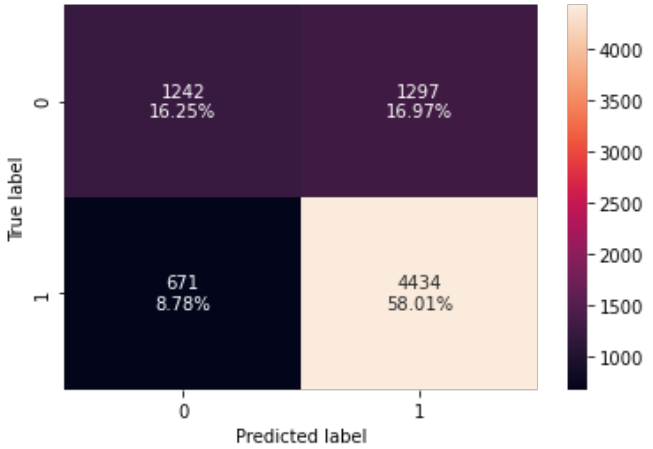
Model Performance Summary – Stacking Classifier

Training-Data Performance



Accuracy	Recall	Precision	F1
0.775006	0.890624	0.796547	0.840962

Testing-Data Performance



Accuracy	Recall	Precision	F1
0.742543	0.86856	0.773687	0.818383

Generalized Performance

Model Performance Summary – XGBoost Classifier

```
StackingClassifier(estimators=[('AdaBoost', AdaBoostClassifier(random_state=1)),
                               ('Gradient Boosting',
                                GradientBoostingClassifier(init=AdaBoostClassifier(random_state=1),
                                                           max_features=0.8,
                                                           n_estimators=200,
                                                           random_state=1,
                                                           subsample=1)),
                               ('Random Forest',
                                RandomForestClassifier(class_weight='balanced',
                                                       max_depth=10,
                                                       max_features='sqrt',
                                                       min_samples_split=3,
                                                       n_est...
                                                  gpu_id=-1,
                                                  importance_type=None,
                                                  interaction_constraints='',
                                                  learning_rate=0.1,
                                                  max_delta_step=0, max_depth=6,
                                                  min_child_weight=1,
                                                  missing=nan,
                                                  monotone_constraints='()',
                                                  n_estimators=200, n_jobs=8,
                                                  num_parallel_tree=1,
                                                  predictor='auto',
                                                  random_state=1, reg_alpha=0,
                                                  reg_lambda=1,
                                                  scale_pos_weight=1,
                                                  subsample=1,
                                                  tree_method='exact',
                                                  validate_parameters=1,
                                                  verbosity=None))
```

Models Performance Comparison

Great Learning
POWER AHEAD

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	0.999944	0.765306	0.738226	0.719163	0.758802	0.764017	0.838753	0.765474	0.775006
Recall	1.0	0.931923	0.985982	0.999916	0.999916	0.799379	0.887182	0.781415	0.883740	0.882649	0.931419	0.881642	0.890624
Precision	1.0	0.720067	0.991810	0.994407	1.000000	0.841329	0.760688	0.794690	0.783042	0.789059	0.843482	0.791127	0.796547
F1	1.0	0.812411	0.988887	0.997154	0.999958	0.819817	0.819080	0.787997	0.830349	0.833234	0.885272	0.833935	0.840962

The Gradient Boost is the best model here, has the highest f1 score of approx. 82% on test and 83% on training data. It has the highest Recall, Accuracy, Precision combination too.

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	0.721088	0.716902	0.734301	0.716641	0.744767	0.743459	0.733255	0.745160	0.742543
Recall	0.742801	0.930852	0.764153	0.895397	0.840744	0.765132	0.885015	0.781587	0.876004	0.871303	0.860725	0.869540	0.868560
Precision	0.752232	0.715447	0.771711	0.743857	0.764926	0.801889	0.757799	0.791510	0.772366	0.773296	0.767913	0.775913	0.773687
F1	0.747487	0.809058	0.767913	0.812622	0.801045	0.783079	0.816481	0.786517	0.820927	0.819379	0.811675	0.820063	0.818383

Models Performance Summary & the best model

- Decision tree overfitted on training but gave generalized performance on test set.
- Tuned Decision tree performed well on both training and test set.
- Bagging classifier overfitted the data but performed well on test data set.
- Tuned Bagging classifier overfitted the training data but performed well on test data set.
- Random Forest with default parameters overfitted the test data but performed well on test data
- Tuned Random Forest performed well on both training and test data
- The majority of the models are overfitting the training data in terms of f1-score.

- The Gradient Boost model is giving the highest f1-score on the test data without overfitting on the training data.

- Gradient Boost, Tuned Gradient Boost, XGBoost, Tuned XGBoost and Stacking Classifier are the top 5 models. They are all giving a similar performance.

- **The Gradient Boost model is the best model here. It has the highest f1 score of approx. 82% on test and 83% on training data.**

Business Insights

- Prevailing Wages under $100 are hourly based
- Visa applications from Asia dominates the pool by 66%
- Visa candidates with either Bachelors or Masters degree represents 78% of all application.
- 88% of all submitted applications are for candidates that don't require job training
- 2/3 of all visa applications are approved
- Those with higher education may want to travel abroad for a well-paid job.
- Higher education has positive impact on visa certification
- Visa application with Doctorate have the highest approval rate of 87%, followed by 78% with Master's degree and 62% with Bachelor's degree, in their respective degree pools
- Different regions have different requirements of talent having diverse educational backgrounds
- Visa certification has higher percentage (~76%) for Midwest US region
- Visa status vary across different continents. Visa Denied rate is lowest (20%) for applicants from Europe region
- 65% of visa certification is awarded to applications with prior job experience.
- Only 8.5% of experienced Job applicant requires some sort of training, compared to 16% of Non-Experience job applicants that require training.
- Visa status for jobs for lower prevailing wage have higher chance of denial.
- Median prevailing wage across US is $70,000 yearly

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Business Recommendations

- Based on our analysis, we can say that the suitable profile for the applicants for whom the visa should be certified has the following significant features in comparison to the less suitable applicants:
    - Education of Employee (high-school).
    - Has job experience (Y)
    - Prevailing wage
    - Education of Employee (Masters-degree)
    - Education of Employee (Doctorate-degree)
    - Continent (Europe)
-
- The US employer should submit visa application for hard-working, talented, and qualified individuals from abroad to supplement the shortage of workers.
- Our analysis show there is low possibility of visa approval for applicant who doesn't have high-school diploma or job related prior experience
- The employee has higher chance of visa approval if he/she has job related experience and doesn't require job training.
- Analysis show that prevailing wages have significant influential factor in the visa certification
- Visa application for employee based in Europe has increasing chance of visa-certification in comparison to non-European employees.

- **We recommend Gradient Boosting Classifier model to be used for the predicting the visa certifications**