Project 5

EasyVisa

Ensemble Techniques

Date: 28 – Jan -2023
Name: Ann Mariya Jomon

Contents / Agenda

01

Executive Summary

02

Business problem overview & solution approach

03

EDA Results

04

Data Preprocessing

05

Model performance summary

06

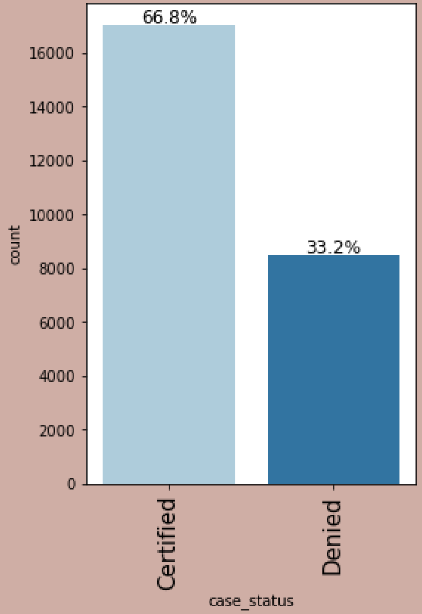
Appendix

01

Executive Summary

Insights

- Majority of the candidates are from Asia and have completed their education till Bachelor's. This is followed by Master's.
- Northeast and South are the most common intended region of employment. However, Midwest has more visas being certified than the other regions.
- 90% of the observations have prevailing wages that are paid yearly.
- Europe has the highest number of visas certified.
- Gradient Boosting Classifier is the best model since it gives the highest F1 score with minimum disparities in train and test set. The top three features according to this model are, Education of employee (High School), Job experience (Yes), and Prevailing wage.
- Candidates with higher level of education have greater chances of getting the visa certified.
- Candidates with job experience have more chances of their visa getting certified than candidates without any job experience.
- Approximately 67% of the cases were Certified while 33% were Denied.



Recommendations

- The key drivers for the visas being certified are education, job experience, and the prevailing wage rates. OFLC must consider these while shortlisting the candidates for visa approvals.

- Job experience can be further analyzed by looking into the relevant years of experience of the employees and comparing with the requirement of the employers in US.

- Segregating the data further into different categories based on job position or industry can help to look at the job-specific requirements. This will improve the predictability of the ML model which will then improve the shortlisting process.

Candidates with the following profile tend to have higher chances of the visa getting approved and shortlisting can be done based on this:

Education: Master's / Doctorate

Region of employment: Midwest / South

Continent: Europe

Job experience: Yes

02
Business problem overview

& solution approach

Problem Overview

The Office of Foreign Labor Certification (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.

The process of reviewing every case is becoming a tedious task as the number of applicants is increasing every year.

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.

Solution Approach

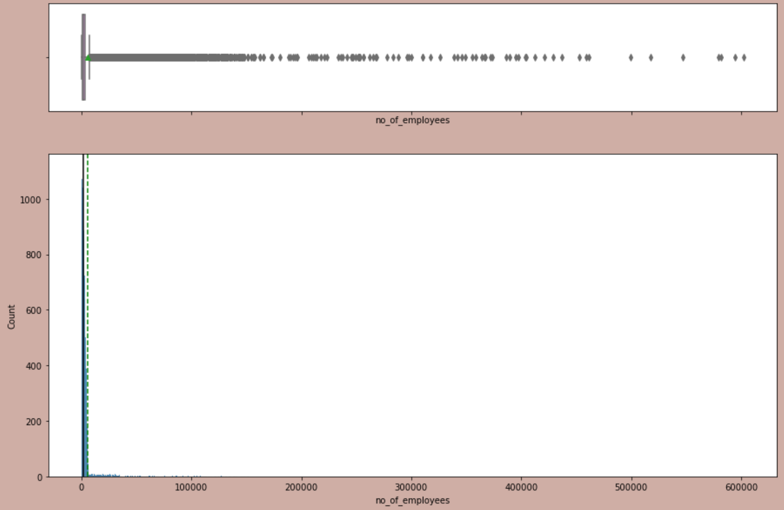
1. EDA (Univariate & Bivariate analysis)
2. Data preprocessing (missing value check, duplicate check, feature engineering, outlier check, data split)
3. Model Building & Improvement – Bagging (Decision Tree, Bagging Classifier, Random Forest)
4. Model Building & Improvement – Boosting (AdaBoost, Gradient Boosting, XGBoost)
5. Stacking Classifier
6. Selecting final model and identifying important features – Comparing performance metrics
7. Insights & Business Recommendations

03

EDA Results

Univariate Analysis

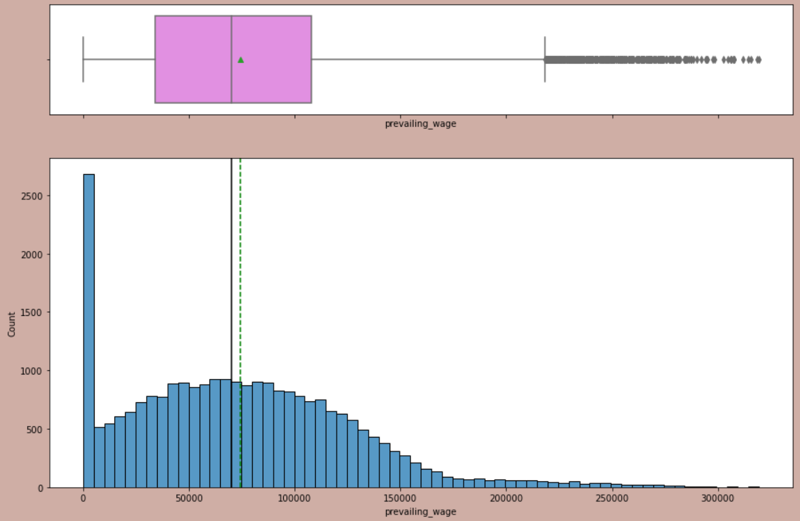
Number of employees



- The distribution is heavily right skewed.
- There are a lot of outliers in the boxplot.
- It ranges between 0 – 602069.
- The mean number of employees is at 5667.

Univariate Analysis

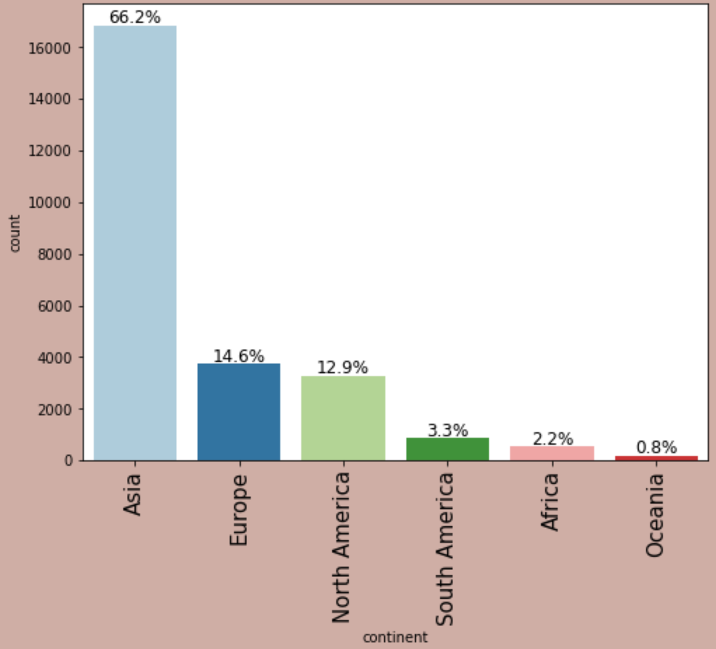
Prevailing wage



- The distribution is skewed to the right.
- There are 176 observations with prevailing wages of less than 100. The unit for all the 176 observations is hourly.
- The average rate is between 50,000 – 100,000

Univariate Analysis

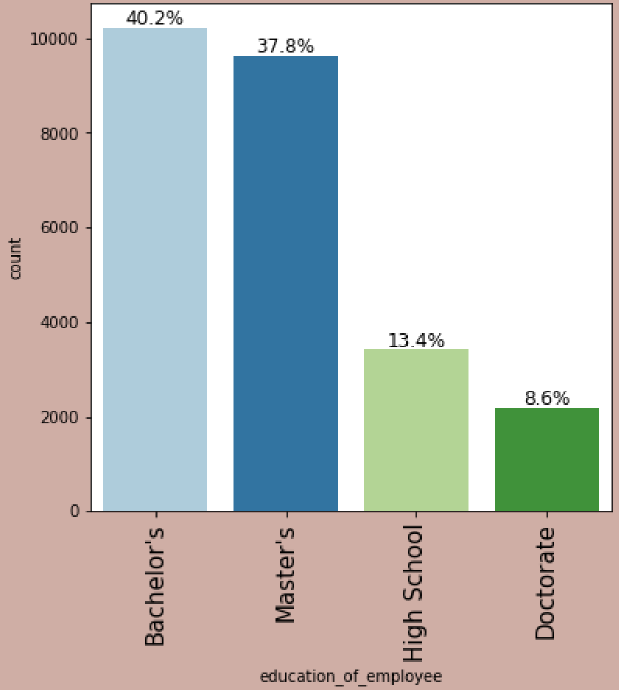
Continent



- Majority of the candidates are from Asia.
- Europe and North America has an almost equal proportion of candidates.
- The most uncommon Continent is Oceania.

Univariate Analysis

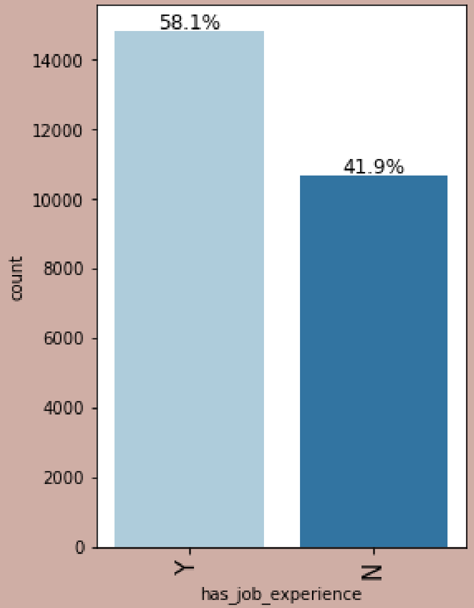
Education of employee



- Majority of the candidates have completed their education till Bachelor’s.
- This is followed by Master’s.
- High School and Doctorate is the least common.

Univariate Analysis

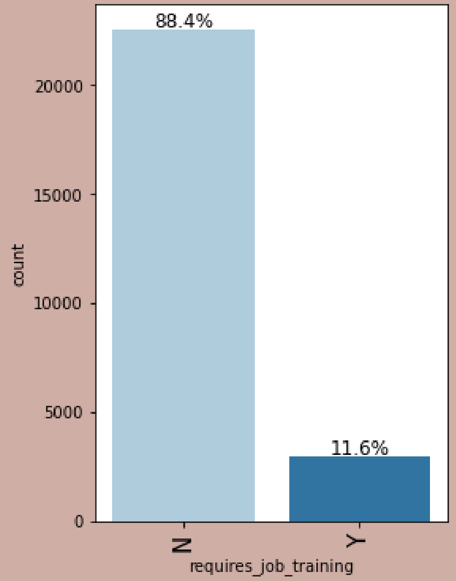
Job experience



- ○ More than 50% of candidates have job experience.
- ○ Around 42% do not have a job experience.

Univariate Analysis

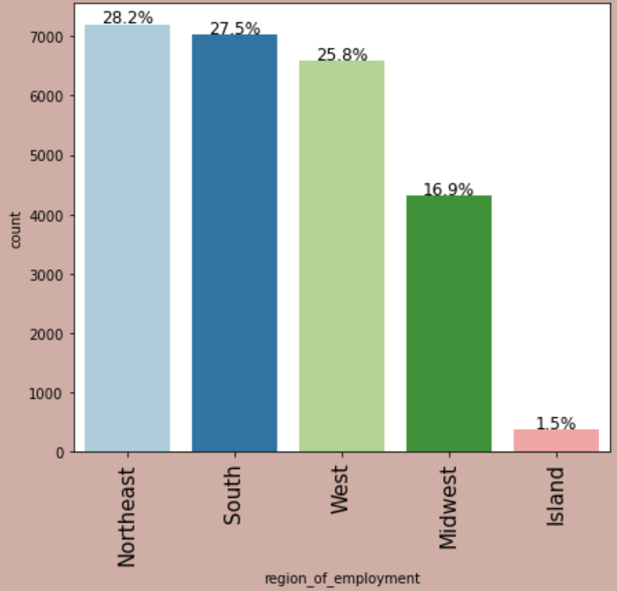
Job training



- Almost 90% of the candidates do not require any job training.

Univariate Analysis

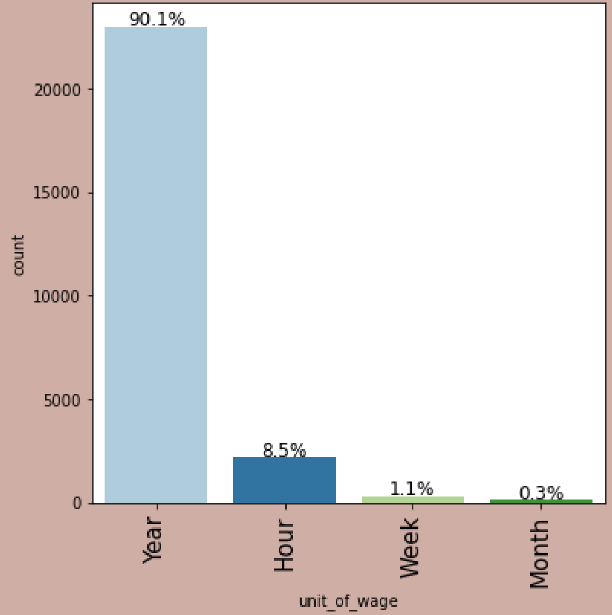
Region of employment



- Northeast and South are the most common intended region of employment.
- This is followed by West, Midwest, and Island.

Univariate Analysis

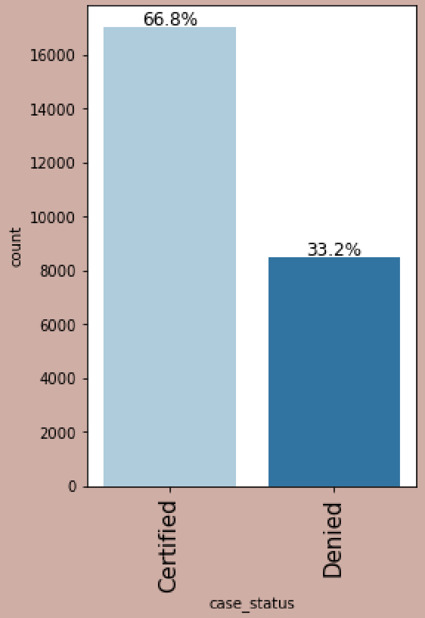
Unit of wage



- 90% of the observations have prevailing wages that are paid yearly.
- Only 0.3% are monthly units.

Univariate Analysis

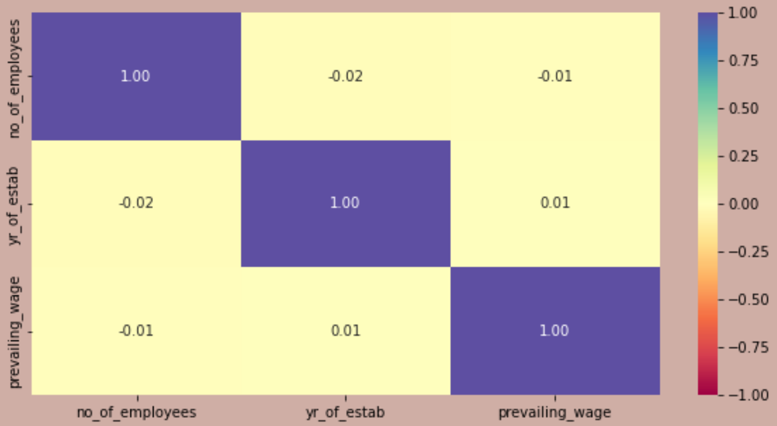
Case status



- Approximately 67% of the cases were Certified while 33% were Denied.

Bivariate Analysis

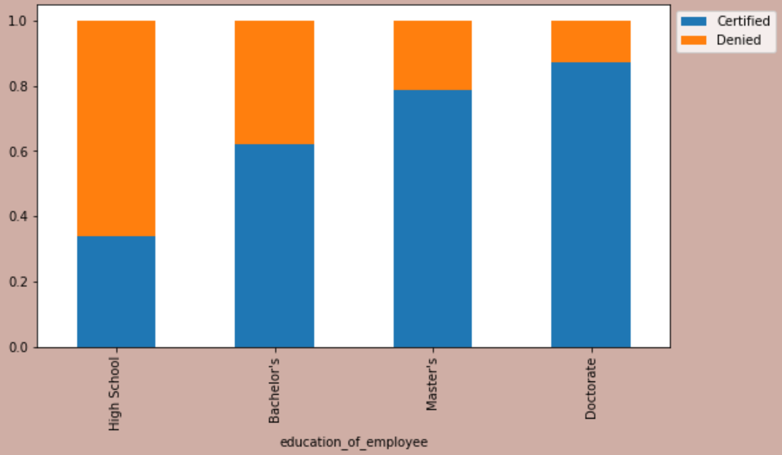
Correlation



- There is no high correlation between the variables.
- There is a very slight negative correlation between prevailing wage and number of employees, and year of establishment and number of employees.

Bivariate Analysis

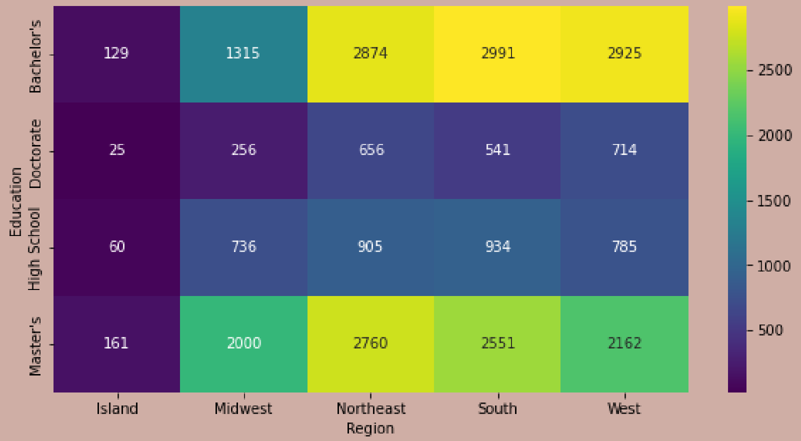
Education of employee vs Case status



- Candidates with higher level of education have greater chances of getting the visa certified.
- High school graduates have very less chances of getting certified.

Bivariate Analysis

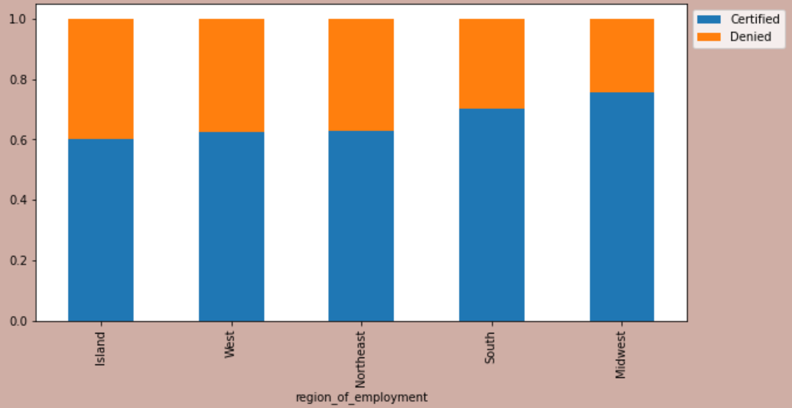
Region of employment vs Education



- Island and Midwest have more candidates with Master's compared to other education.
- Northeast, South, and West have more candidates with Bachelor's.
- More number of candidates with Bachelor's are in South.
- More number of candidates with Doctorate are in West.
- More number of candidates who are just high school graduates are in South.
- More number of candidates with Master's are in Northeast region.

Bivariate Analysis

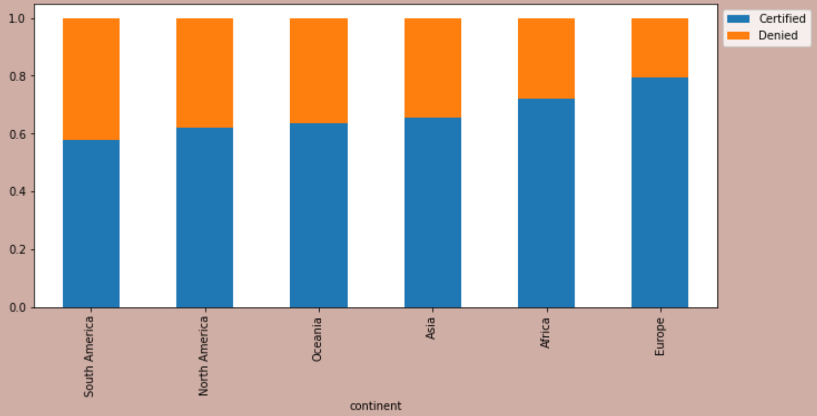
Region of employment vs Case status



- Island, West, and Northwest have lesser number of visas being certified compared to South and Midwest.
- Midwest has more visas being certified than the other regions.

Bivariate Analysis

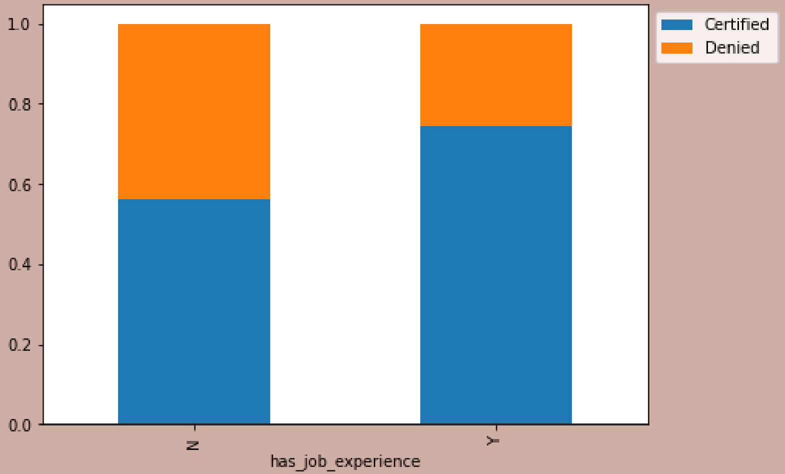
Continent vs Case status



- Europe has the highest number of visas certified.
- South America has the lowest number of visas certified.
- North America and Oceania has an almost equal number of visas certified.

Bivariate Analysis

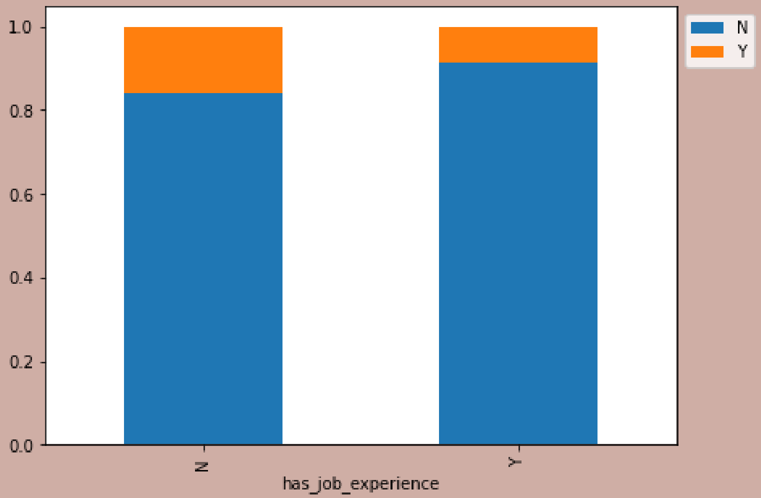
Job experience vs Case status



- Candidates with job experience have more chances of their visa getting certified than candidates without any job experience.

Bivariate Analysis

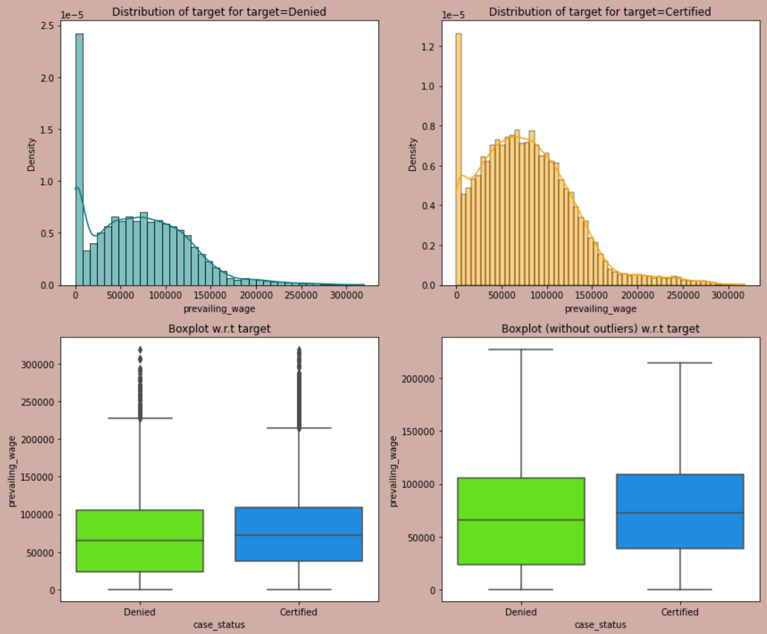
Job experience vs Job training



- Candidates with no job experience have more chances of receiving job training than candidates without any experience.
- However, there is no significant correlation between the two variables as per the graphs.

Bivariate Analysis

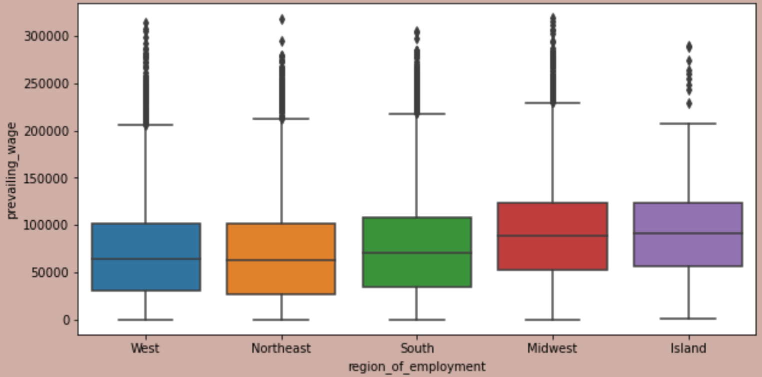
Prevailing wage vs Case status



- The distribution of prevailing wage in case of visas that were certified have a higher density than visas that were denied.
- The visas that were certified have a lower upper whisker than visas that were denied.
- The median prevailing wage is higher for visas that were certified.

Bivariate Analysis

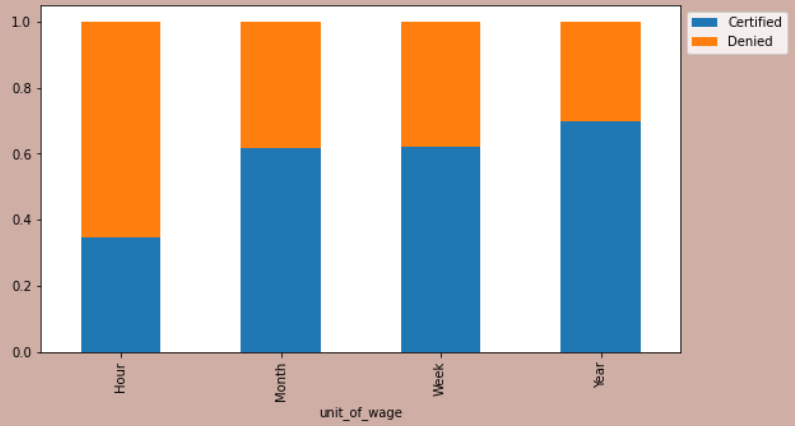
Region of employment vs Prevailing wage



- The prevailing wage range is higher for the Midwest region while it is lower for Island.
- West and Northeast regions have the same median prevailing wage.
- Island and Midwest regions also have the same median prevailing wage.

Bivariate Analysis

Unit of wage vs Case status



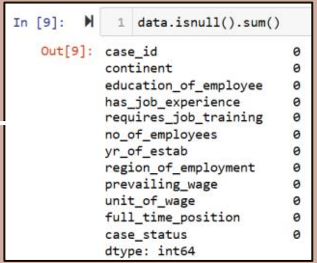
- Unit of wages that are yearly have higher number of visas that are certified.
- Hourly wages have the lowest number of visas that were certified.
- Both monthly and weekly wages seem to have an equal proportion of visas that were certified and denied.

04

Data preprocessing

Missing Values

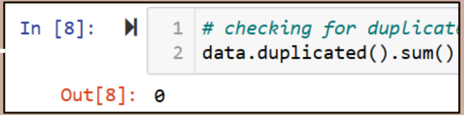
No missing values



```
In [9]:  data.isnull().sum()

Out[9]:  case_id                  0
         continent                0
         education_of_employee    0
         has_job_experience       0
         requires_job_training    0
         no_of_employees          0
         yr_of_estab              0
         region_of_employment     0
         prevailing_wage          0
         unit_of_wage             0
         full_time_position       0
         case_status              0
         dtype: int64
```

Duplicates

No duplicates



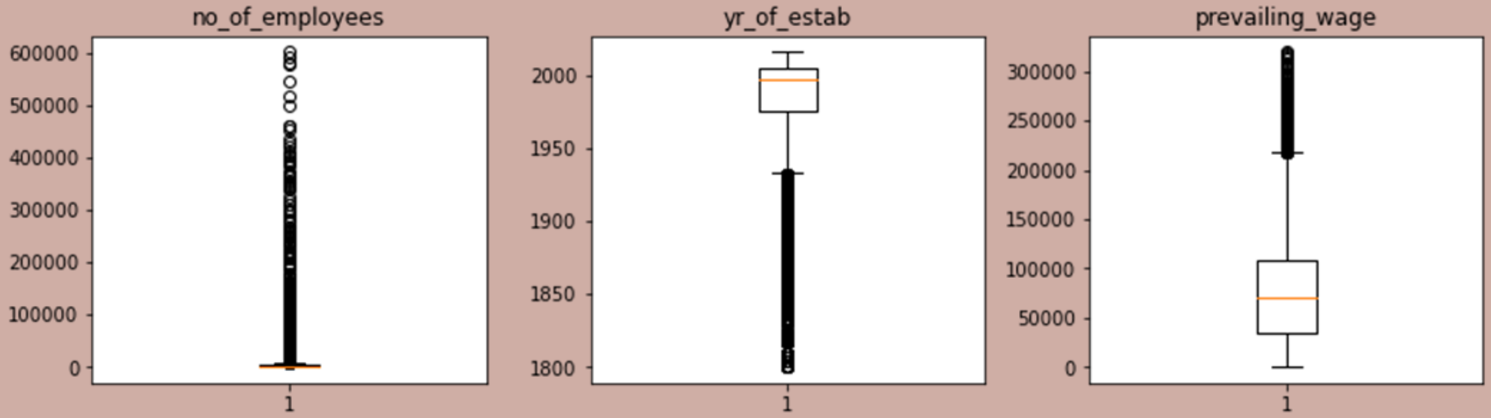
```
In [8]:  # checking for duplicat
         data.duplicated().sum()

Out[8]:  0
```

Feature Engineering

Fixed negative values of number of employees column

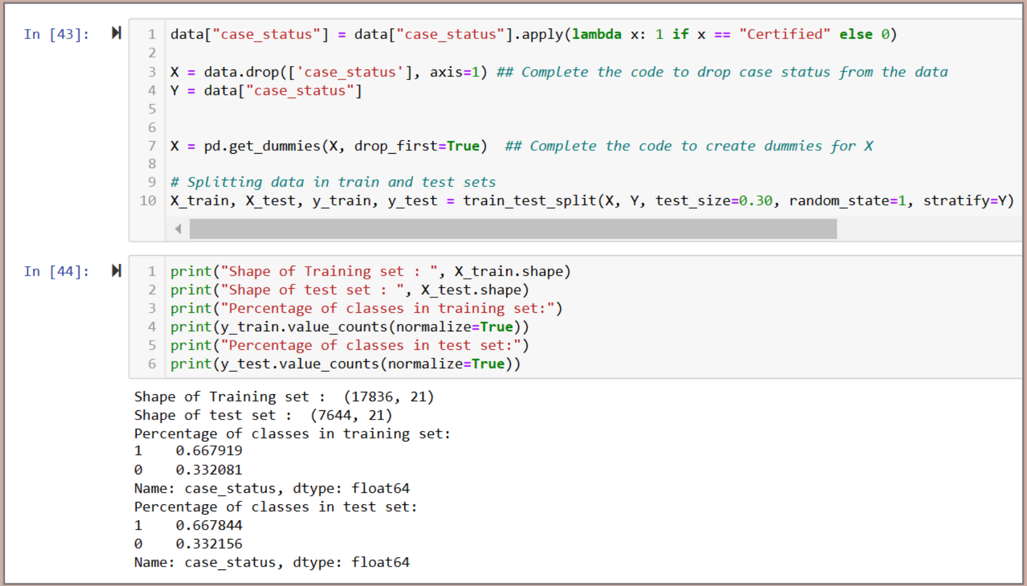
Dropped ‘case_id’ column since it has all unique values

Outlier Check



All the three variables have lot of outliers. However, since all of them are proper values, there is no need for treatment.

Data Preparation for modeling



In [43]:

```
data["case_status"] = data["case_status"].apply(lambda x: 1 if x == "Certified" else 0)

X = data.drop(['case_status'], axis=1) ## Complete the code to drop case status from the data
Y = data["case_status"]


X = pd.get_dummies(X, drop_first=True)  ## Complete the code to create dummies for X

# Splitting data in train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.30, random_state=1, stratify=Y)
```

In [44]:

```
print("Shape of Training set : ", X_train.shape)
print("Shape of test set : ", X_test.shape)
print("Percentage of classes in training set:")
print(y_train.value_counts(normalize=True))
print("Percentage of classes in test set:")
print(y_test.value_counts(normalize=True))
```

```
Shape of Training set :  (17836, 21)
Shape of test set :  (7644, 21)
Percentage of classes in training set:
1    0.667919
0    0.332081
Name: case_status, dtype: float64
Percentage of classes in test set:
1    0.667844
0    0.332156
Name: case_status, dtype: float64
```

- It was observed from the EDA results that 67% of the observations were Certified for Visa while 33% were Not Certified.
- After splitting the data into train and test, the same percentages have been maintained in both train and test data.

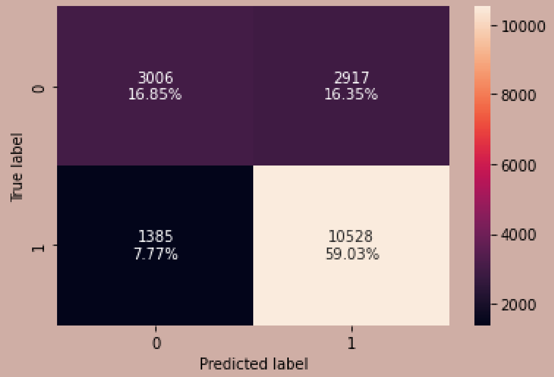
05

Model performance summary

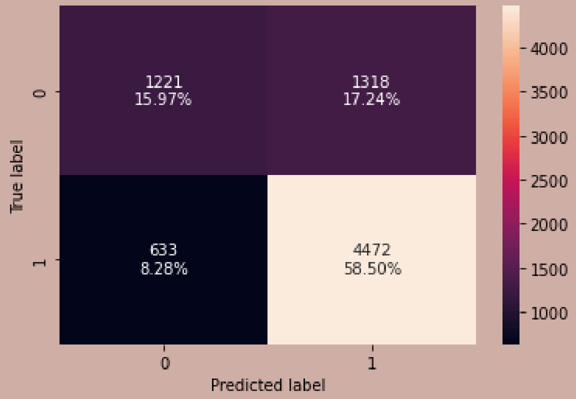
Final Model –

Gradient Boosting Classifier

Training data

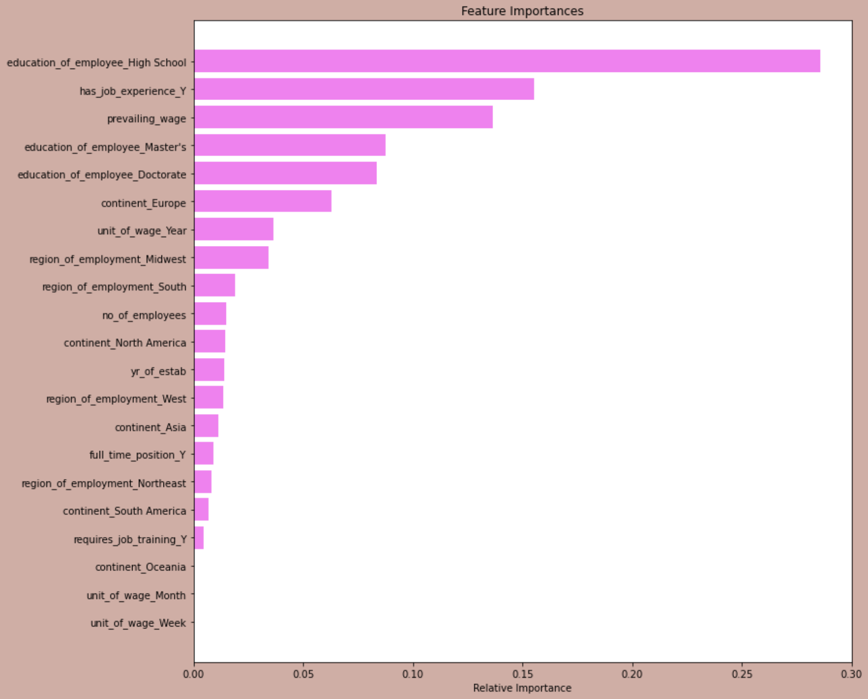


Testing data



- It can be observed that the Gradient Boosting Classifier is able to generalize the data well. The model does not indicate signs of overfitting.

Most important features



The top three most important features as per Gradient Boosting Classifier are:

Education of employee : High School

Job Experience : Yes

Prevailing wage

Key performance metrics

	Train Performance	Test Performance
Accuracy	0.758802	0.744767
Recall	0.88374	0.876004
Precision	0.783042	0.772366
F1 score	0.830349	0.820927

- The train and test performance does not indicate overfitting. This means the model is able to generalize well on both the data sets.
- F1 score is used for model evaluation as the aim is to minimize both False Negatives and False Positives
- This model gives the highest F1 score with minimum disparities in train and test set.

06

Appendix

Data Background and Content

Data Dictionary

Sl.No.:	Variable	Description
1	case_id	ID of each visa application
2	continent	Information of continent the employee
3	education_of_employee	Information of education of the employee
4	has_job_experience	Does the employee has any job experience? Y= Yes; N = No
5	requires_job_training	Does the employee require any job training? Y = Yes; N = No
6	no_of_employees	Number of employees in the employer's company
7	yr_of_estab	Year in which the employer's company was established
8	region_of_employment	Information of foreign worker's intended region of employment in the US.
9	prevailing_wage	Average wage paid to similarly employed workers in a specific occupation in the area of intended employment.
10	unit_of_wage	Unit of prevailing wage. Values include Hourly, Weekly, Monthly, and Yearly.
11	full_time_position	Is the position of work full-time? Y = Full-Time Position; N = Part-Time Position
12	case_status	Flag indicating if the Visa was certified or denied

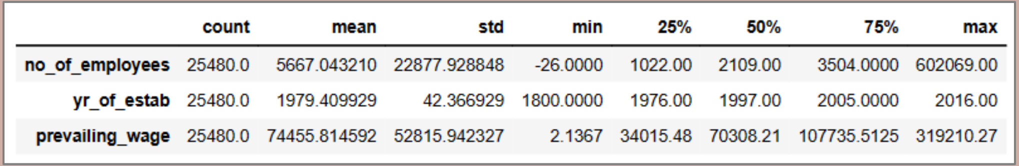
Data Background and Content

Datatypes & Shape

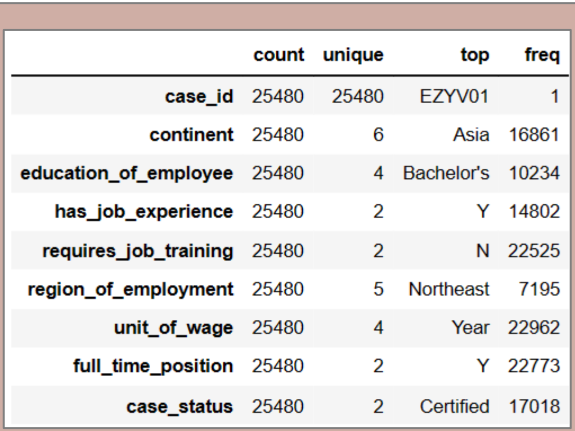
Datatypes	
integer	2
float	1
object	9

Shape of dataset	
Rows	25480
Columns	12

Statistical Summary



	count	mean	std	min	25%	50%	75%	max
no_of_employees	25480.0	5667.043210	22877.928848	-26.0000	1022.00	2109.00	3504.0000	602069.00
yr_of_estab	25480.0	1979.409929	42.366929	1800.0000	1976.00	1997.00	2005.0000	2016.00
prevailing_wage	25480.0	74455.814592	52815.942327	2.1367	34015.48	70308.21	107735.5125	319210.27

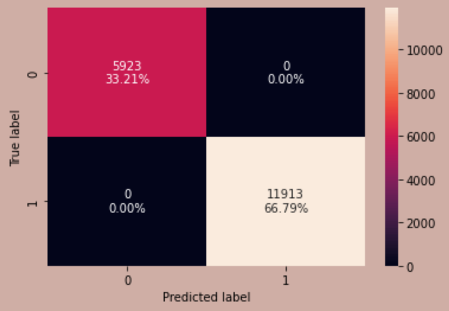


	count	unique	top	freq
case_id	25480	25480	EZYV01	1
continent	25480	6	Asia	16861
education_of_employee	25480	4	Bachelor's	10234
has_job_experience	25480	2	Y	14802
requires_job_training	25480	2	N	22525
region_of_employment	25480	5	Northeast	7195
unit_of_wage	25480	4	Year	22962
full_time_position	25480	2	Y	22773
case_status	25480	2	Certified	17018

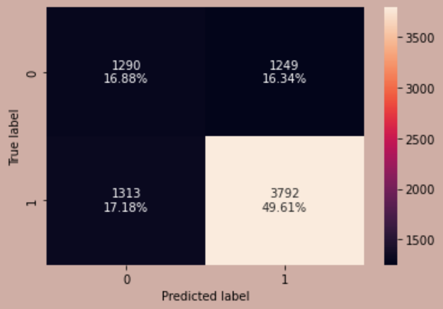
Model Building

Bagging – Decision Tree

Train



Test



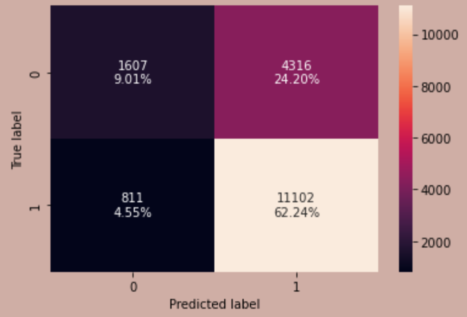
	Accuracy	Recall	Precision	F1
Train	1	1	1	1
Test	0.66	0.74	0.75	0.75

- DecisionTreeClassifier() was used to build the model.
- The model is overfitting on train set.
- There is high disparity between all performance metrics of train and test set.

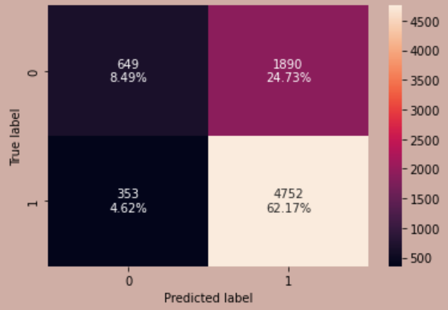
Model Improvement

Bagging – Tuned Decision Tree

Train



Test



	Accuracy	Recall	Precision	F1
Train	0.712548	0.931923	0.720067	0.812411
Test	0.706567	0.930852	0.715447	0.809058

max_depth = 5

max_leaf_nodes = 2

min_impurity_decrease = 0.0001

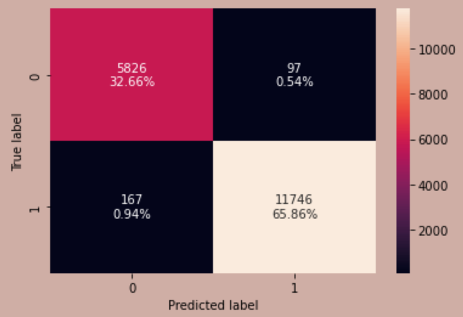
min_samples_leaf = 3

- The model is not overfitting.
- The accuracy has dropped but the **F1 score has improved** in test set.
- The recall score has improved.
- Precision has dropped.

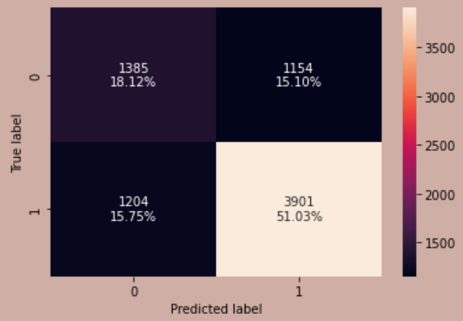
Model Building

Bagging – Bagging Classifier

Train



Test

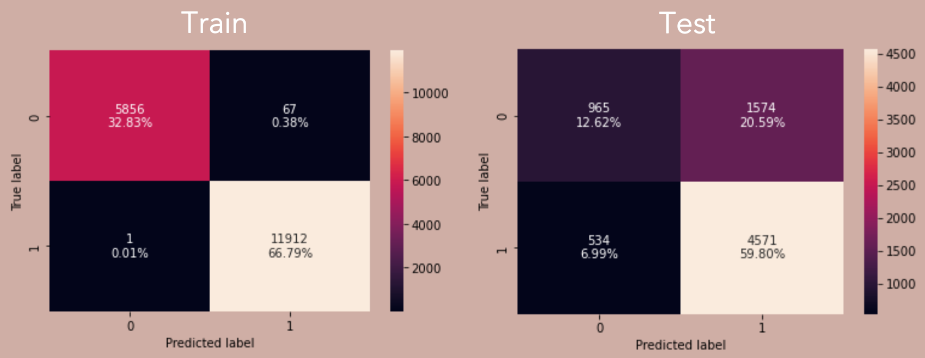


	Accuracy	Recall	Precision	F1
Train	0.985198	0.985982	0.99181	0.988887
Test	0.691523	0.764153	0.771711	0.767913

- BaggingClassifier() was used to build the model.
- The model tends to be overfitting on train data.
- There is high disparity between the train and test performance.
- The **F1 score has dropped** in test data compared to the tuned decision tree.

Model Improvement

Bagging – Tuned Bagging Classifier



	Accuracy	Recall	Precision	F1
Train	0.996187	0.999916	0.994407	0.997154
Test	0.724228	0.895397	0.743857	0.812622

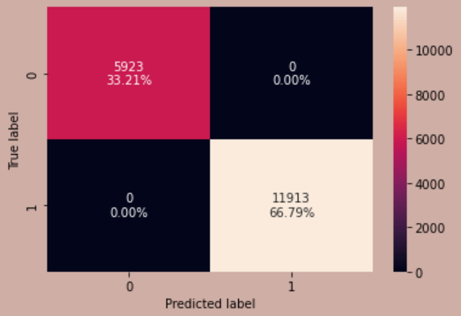
max_features = 0.7 max_samples = 0.7 n_estimators = 100

- The model tends to overfit on train data.
- There is disparity between train and test performances.
- The **F1 score has improved**.

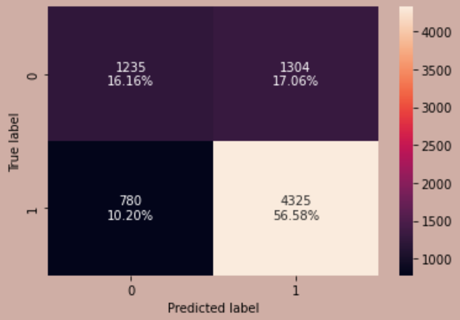
Model Building

Bagging – Random Forest

Train



Test



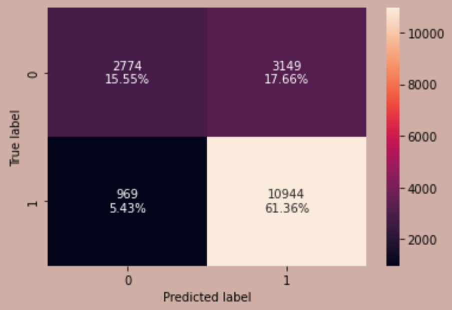
	Accuracy	Recall	Precision	F1
Train	1.0	1.0	1.0	1.0
Test	0.727368	0.847209	0.768343	0.805851

- RandomForestClassifier() with class_weight = 'balanced' was used to build the model.
- The model is overfitting on train data.
- There is disparity between train and test performance.
- The **F1 score has dropped** slightly.

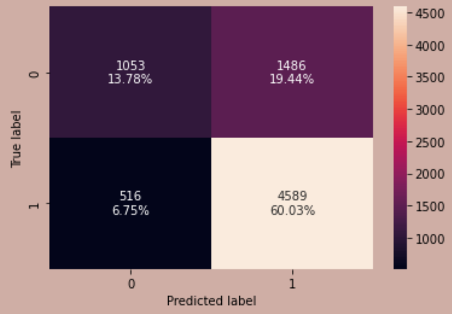
Model Improvement

Bagging – Tuned Random Forest

Train



Test



	Accuracy	Recall	Precision	F1
Train	0.769119	0.91866	0.776556	0.841652
Test	0.738095	0.898923	0.755391	0.82093

max_depth = 10 max_features = ‘sqrt’ n_estimators = 20

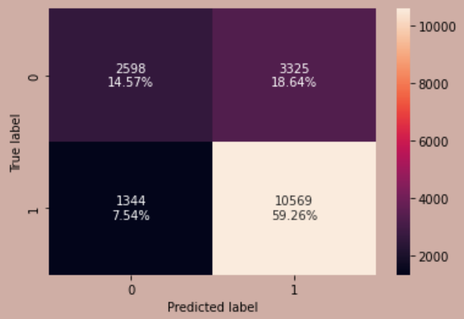
min_sample_split = 7 oob_score = True

- The model is not overfitting.
- There is slight disparity between train and test performances.
- The **F1 score has improved**.

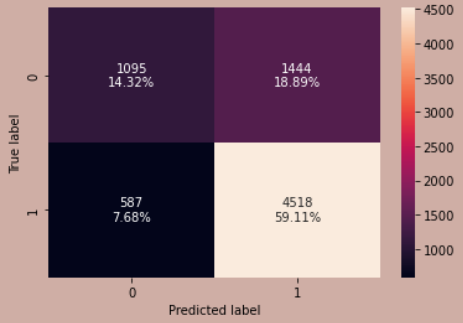
Model Building

Boosting – AdaBoost

Train



Test



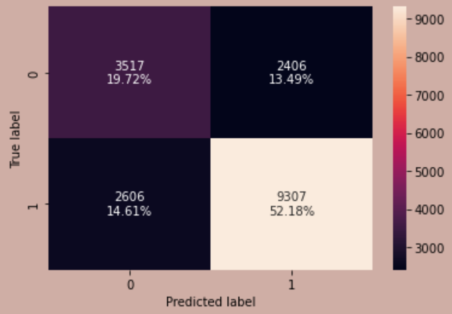
	Accuracy	Recall	Precision	F1
Train	0.738226	0.887182	0.760688	0.81908
Test	0.734301	0.885015	0.757799	0.816481

- AdaBoostClassifier() was used to build the model.
- The model is not overfitting.
- The model is able to generalize well on test data.
- The **F1 score has slightly dropped** when compared to the tuned Random forest. But it is performing better than the tuned Decision tree.

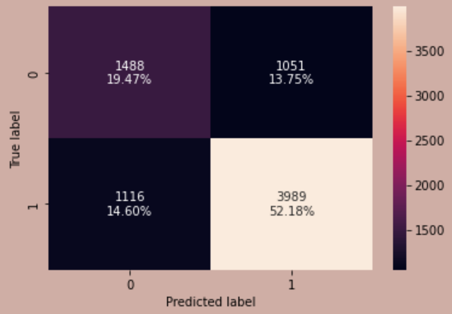
Model Improvement

Boosting – Tuned AdaBoost

Train



Test



	Accuracy	Recall	Precision	F1
Train	0.718995	0.781247	0.794587	0.787861
Test	0.71651	0.781391	0.791468	0.786397

base_estimator = DecisionTreeClassifier

max_depth = 1

class_weight = 'balanced'

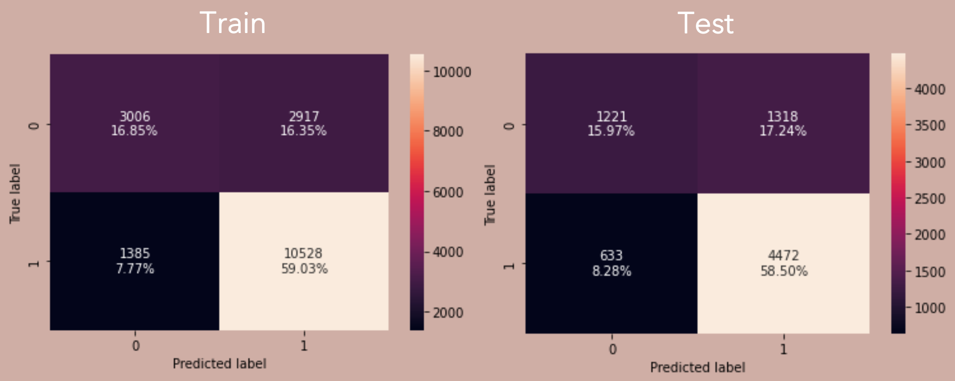
n_estimators = 100

learning_rate = 0.1

- The model is not overfitting.
- The model is able to generalize well on the test data.
- However, the **F1 score has dropped**.
- The model has not improved after hyperparameter tuning.

Model Building

Boosting – Gradient Boosting



Train

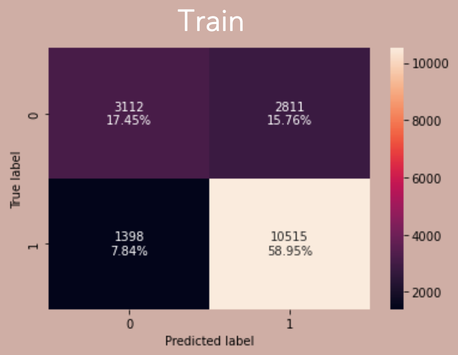
Test

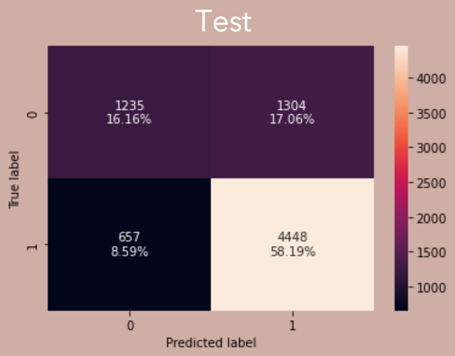
	Accuracy	Recall	Precision	F1
Train	0.758802	0.88374	0.783042	0.830349
Test	0.744767	0.876004	0.772366	0.820927

- GradientBoostingClassifier() was used to build the model.
- The model is not overfitting.
- There is slight disparity between the train and test performance.
- The **F1 score has improved.**

Model Improvement

Boosting – Tuned Gradient Boosting





	Accuracy	Recall	Precision	F1
Train	0.764017	0.882649	0.789059	0.833234
Test	0.743459	0.871303	0.773296	0.819379

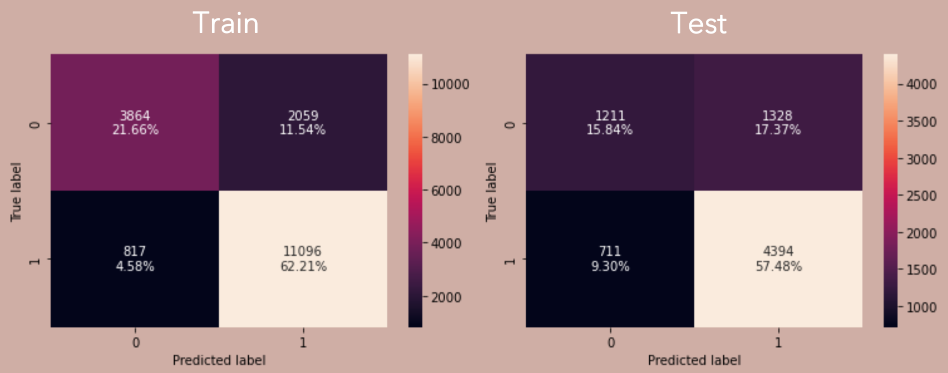
Init = AdaBoostClassifier   max_features = 0.8

n_estimators = 200   subsample = 1

- The model is not overfitting.
- There is slight disparity between the train and test performance.
- The **F1 score has dropped slightly** in test set but has improved in train set.

Model Building

Boosting – XGBoost



Train

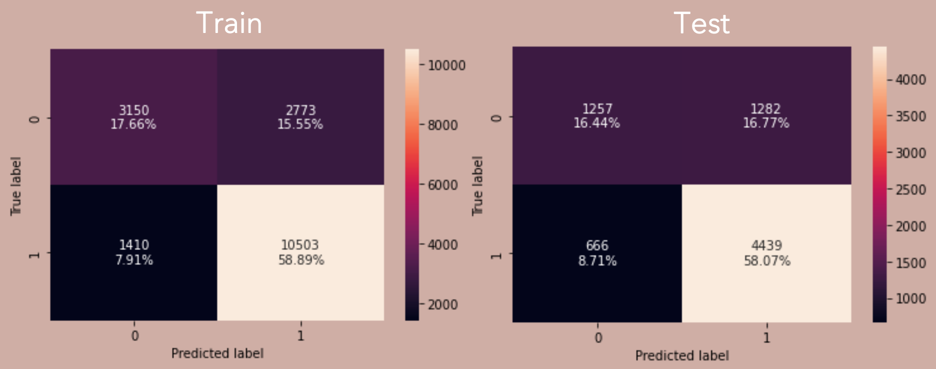
Test

	Accuracy	Recall	Precision	F1
Train	0.838753	0.931419	0.843482	0.885272
Test	0.733255	0.860725	0.767913	0.811675

- XGBClassifier() was used to build the model.
- The model seems to be overfitting.
- There is high disparity between the train and test performance.
- The **F1 score has dropped in test set** but has improved in train set.

Model Improvement

Boosting – Tuned XGBoost


Train

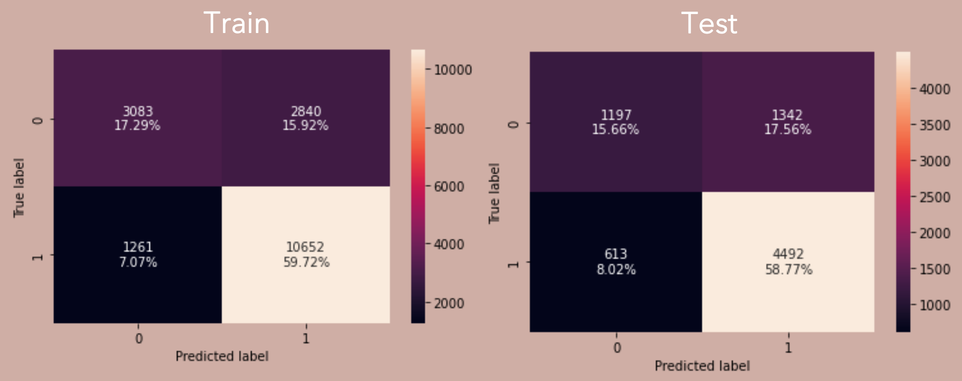
Test

	Accuracy	Recall	Precision	F1
Train	0.765474	0.881642	0.791127	0.833935
Test	0.74516	0.86954	0.775913	0.820063

- The model is not overfitting.
- There is slight disparity between the train and test performance.
- The **F1 score has improved on the test set.**

Model Building

Stacking Classifier



	Accuracy	Recall	Precision	F1
Train	0.770072	0.894149	0.789505	0.838575
Test	0.744244	0.879922	0.769969	0.821282

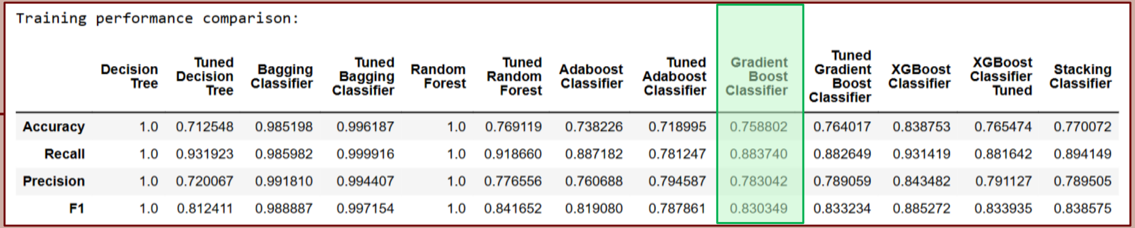
Estimators = AdaBoost Classifier, Tuned Gradient Boost, Tuned Random Forest

Final estimator = Tuned XGBoost

- StackingClassifier() was used to build the model.
- The model is not overfitting.
- There is some disparity between train and test performances.
- However, the **F1 score has improved**.

Final Model selection

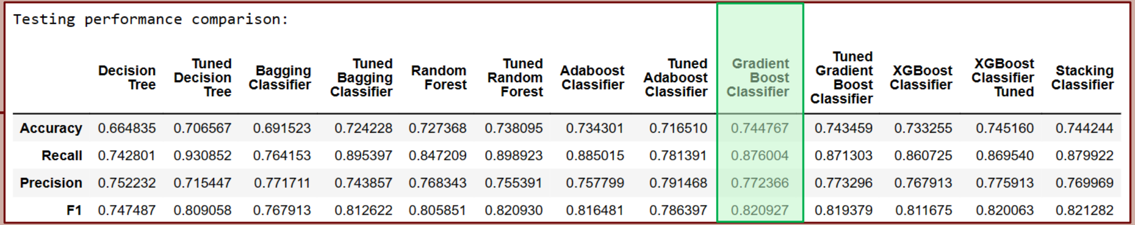
Train data



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	0.765474	0.770072
Recall	1.0	0.931923	0.985982	0.999916	1.0	0.918660	0.887182	0.781247	0.883740	0.882649	0.931419	0.881642	0.894149
Precision	1.0	0.720067	0.991810	0.994407	1.0	0.776556	0.760688	0.794587	0.783042	0.789059	0.843482	0.791127	0.789505
F1	1.0	0.812411	0.988887	0.997154	1.0	0.841652	0.819080	0.787861	0.830349	0.833234	0.885272	0.833935	0.838575

Test data



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	0.745160	0.744244
Recall	0.742801	0.930852	0.764153	0.895397	0.847209	0.898923	0.885015	0.781391	0.876004	0.871303	0.860725	0.869540	0.879922
Precision	0.752232	0.715447	0.771711	0.743857	0.768343	0.755391	0.757799	0.791468	0.772366	0.773296	0.767913	0.775913	0.769969
F1	0.747487	0.809058	0.767913	0.812622	0.805851	0.820930	0.816481	0.786397	0.820927	0.819379	0.811675	0.820063	0.821282

- From the above results, it can be observed that the '**Gradient Boosting Classifier**' gives the highest F1 score with least differences between train and test performance and no overfitting.
- Therefore, this model can be selected as the final model which can be used to assist in shortlisting the candidates with higher chances of Visa approval.

THANK YOU!