MACHINE LEARNING 2
BUSINESS REPORT

Krishnan CS

GREAT LEARNING DSBA

Contents

1. Exploratory Data Analysis

1.1 Problem Definition

1.1.1 Context

Business communities in the United States are facing high demand for human resources, but one of the constant challenges is identifying and attracting the right talent, which is perhaps the most important element in remaining competitive. Companies in the United States look for hard-working, talented, and qualified individuals both locally as well as abroad.

The Immigration and Nationality Act (INA) of the US permits foreign workers to come to the United States to work on either a temporary or permanent basis. The act also protects US workers against adverse impacts on their wages or working conditions by ensuring US employers' compliance with statutory requirements when they hire foreign workers to fill workforce shortages. The immigration programs are administered by the Office of Foreign Labor Certification (OFLC).

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.

1.1.2 Objective

In FY 2016, the OFLC processed 775,979 employer applications for 1,699,957 positions for temporary and permanent labor certifications. This was a nine percent increase in the overall number of processed applications from the previous year. The process of reviewing every case is becoming a tedious task as the number of applicants is increasing every year.

The increasing number of applicants every year calls for a Machine Learning based solution that can help in shortlisting the candidates having higher chances of VISA approval. OFLC has hired the firm EasyVisa for data-driven solutions. You as a data scientist at EasyVisa have to analyze the data provided and, with the help of a classification model:

1. Facilitate the process of visa approvals.
2. 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.

1.1.3 Data Description

The data contains the different attributes of the employee and the employer. The detailed data dictionary is given below.

- case_id: ID of each visa application
- continent: Information of continent the employee
- education_of_employee: Information of education of the employee
- has_job_experience: Does the employee have any job experience? Y= Yes; N = No
- requires_job_training: Does the employee require any job training? Y = Yes; N = No
- no_of_employees: Number of employees in the employer's company
- yr_of_estab: Year in which the employer's company was established
- region_of_employment: Information of foreign worker's intended region of employment in the US.

- prevailing_wage: Average wage paid to similarly employed workers in a specific occupation in the area of intended employment. The purpose of the prevailing wage is to ensure that the foreign worker is not underpaid compared to other workers offering the same or similar service in the same area of employment.
- unit_of_wage: Unit of prevailing wage. Values include Hourly, Weekly, Monthly, and Yearly.
- full_time_position: Is the position of work full-time? Y = Full-Time Position; N = Part-Time Position
- case_status: Flag indicating if the Visa was certified or denied.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
case_id	25480	25480	EZYV25480	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
continent	25480	6	Asia	16861	NaN	NaN	NaN	NaN	NaN	NaN	NaN
education_of_employee	25480	4	Bachelor's	10234	NaN	NaN	NaN	NaN	NaN	NaN	NaN
has_job_experience	25480	2	Y	14802	NaN	NaN	NaN	NaN	NaN	NaN	NaN
requires_job_training	25480	2	N	22525	NaN	NaN	NaN	NaN	NaN	NaN	NaN
no_of_employees	25480.0	NaN	NaN	NaN	5667.04321	22877.928848	-26.0	1022.0	2109.0	3504.0	602069.0
yr_of_estab	25480.0	NaN	NaN	NaN	1979.409929	42.366929	1800.0	1976.0	1997.0	2005.0	2016.0
region_of_employment	25480	5	Northeast	7195	NaN	NaN	NaN	NaN	NaN	NaN	NaN
prevailing_wage	25480.0	NaN	NaN	NaN	74455.814592	52815.942327	2.1367	34015.48	70308.21	107735.5125	319210.27
unit_of_wage	25480	4	Year	22962	NaN	NaN	NaN	NaN	NaN	NaN	NaN
full_time_position	25480	2	Y	22773	NaN	NaN	NaN	NaN	NaN	NaN	NaN
case_status	25480	2	Certified	17018	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 1. Description of the columns in the data

1.2 Univariate Analysis

Case_id: This column contains a unique identification number for the records. It is not very significant to the analysis and will be dropped.

Continent: Asia accounts for 66% of the total number of employees. On the other hand, South America, Africa, and Oceania have a combined representation of a little over 5%.

continent	Count	Percentage
Asia	16861	66.17
Europe	3732	14.65
North America	3292	12.92
South America	852	3.34
Africa	551	2.16
Oceania	192	0.75

Figure 2. Count and percentages for continents

education_of_employee: About 78% of the employees have a Bachelor's or Master's degree, while the remaining population is spread between High School and Doctorate. This feature will be converted to a numeric level with High School = 1, Bachelor's = 2, Master's = 3 and Doctorate = 4.

education_of_employee	Count	Percentage
Bachelor's	10234	40.16
Master's	9634	37.81
High School	3420	13.42
Doctorate	2192	8.60

Figure 3. Count and percentages for education level

has_job_experience: About 58% have job experience and 42% do not.

has_job_experience	Count	Percentage
Y	14802	58.09
N	10678	41.91

Figure 4. Count and percentages for job experience

requires_job_training: A significant proportion of the employees do not require job training.

requires_job_training	Count	Percentage
N	22525	88.4
Y	2955	11.6

Figure 5. Count and percentages for job training

region_of_employment: Most of the employees are employed in the Northeast, South or West.

region_of_employment	Count	Percentage
Northeast	7195	28.24
South	7017	27.54
West	6586	25.85
Midwest	4307	16.90
Island	375	1.47

Figure 6. Count and percentages for region of employment

unit_of_wage: 90% of the employees are under the annual payroll. About 8.5 employees are paid by the hour, while the rest have a weekly or monthly payment schedule.

unit_of_wage	Count	Percentage
Year	22962	90.12
Hour	2157	8.47
Week	272	1.07
Month	89	0.35

Figure 7. Count and percentages for unit of wage

full_time_position: The number of employees with and without a full-time position is roughly in the ratio of 9:1.

full_time_position	Count	Percentage
Y	22773	89.38
N	2707	10.62

Figure 8. Count and percentages for full-time position

case_status: This is our target variable. About 2/3 of the cases are certified, while the rest are denied.

case_status	Count	Percentage
Certified	17018	66.79
Denied	8462	33.21

Figure 9. Count and percentages for case status

prevailing_wage: The prevailing_wage is shown as a box plot, coloured by the unit. As expected, the hourly wage values are very low. The yearly, weekly and monthly wages have very similar values! This looks like an error in the data. This would be handled after studying the column in detail.

The boxplot shows a lot of outliers. Even though it would be good to keep the numerical values as they are, the data contains values in different units. This problem would be solved by creating a new column for wage, with values ranging from 1 to 5 (lowest to highest), such that each level contains roughly 20% of the data, for each unit of wage. The *prevailing_wage* column will be dropped after creating the *wage* column.

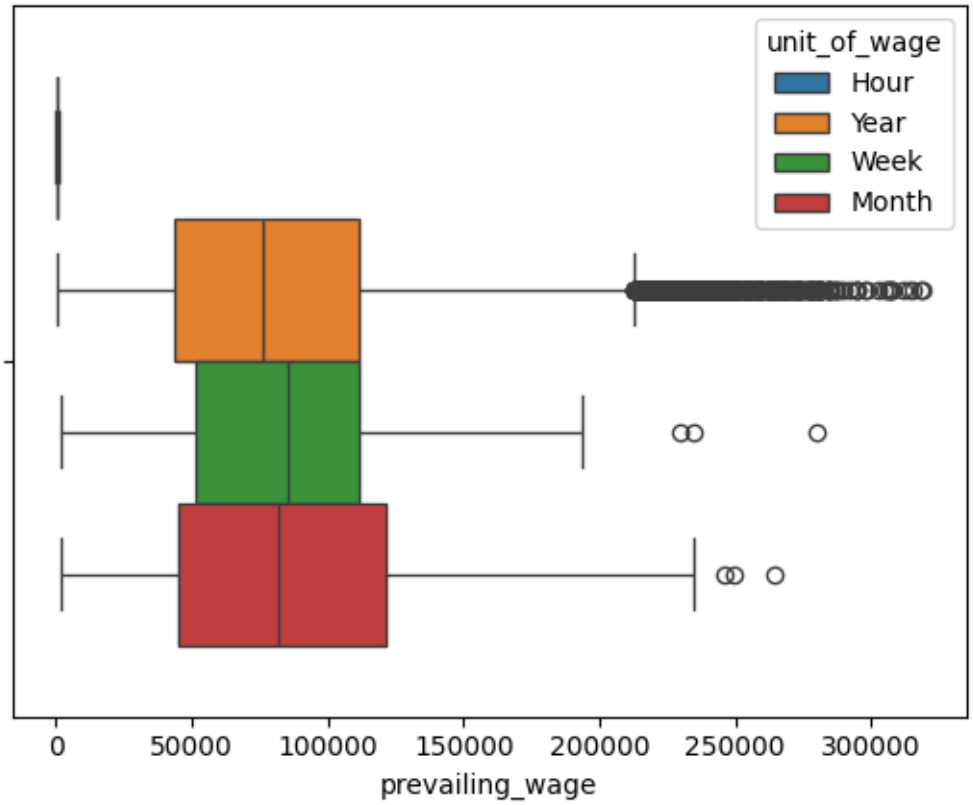

Figure 10. Box plot for prevailing wage

no_of_employees: The box plot of the *no_of_employees* suggests that it is right skewed. The maximum value for this field goes to over 600,000. The very large values have been hidden from this plot for better clarity.

There are about 30 records with values between -1 and -26. This would be assumed as a typo and the negative sign will be removed.

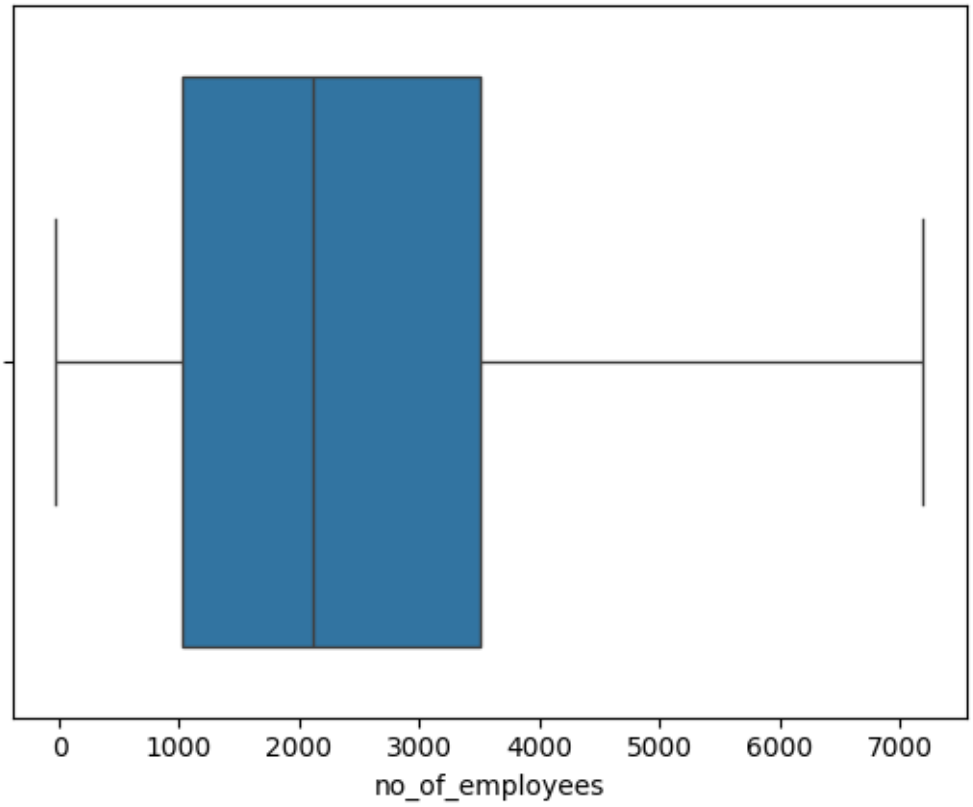

Figure 11. Box plot for number of employees

yr_of_estab: There are companies which have been operational since 1800. 75% of the companies were established after 1975. This is a left-skewed distribution.

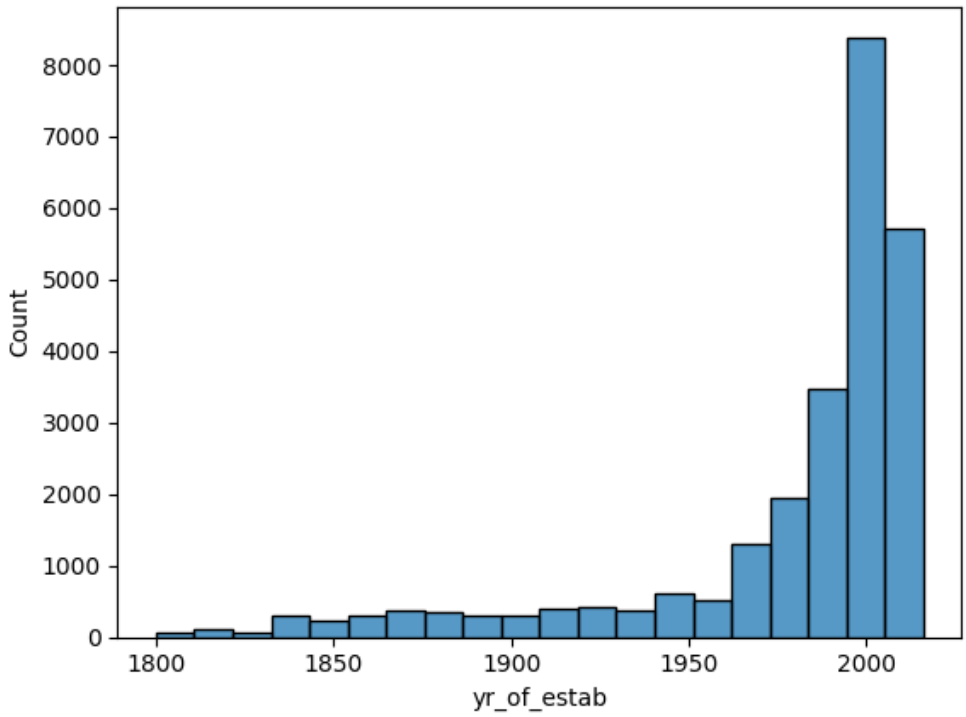

Figure 12. Histogram for year of establishment

1.3 Bivariate analysis

For bivariate analysis, the correlation between the numeric columns can be studied with a heatmap. The wage column will be used instead of prevailing_wage as mentioned in prevailing_wage univariate analysis. The encoded education_level is also included as a numeric column. The heatmap shows very little correlation between the numeric columns.

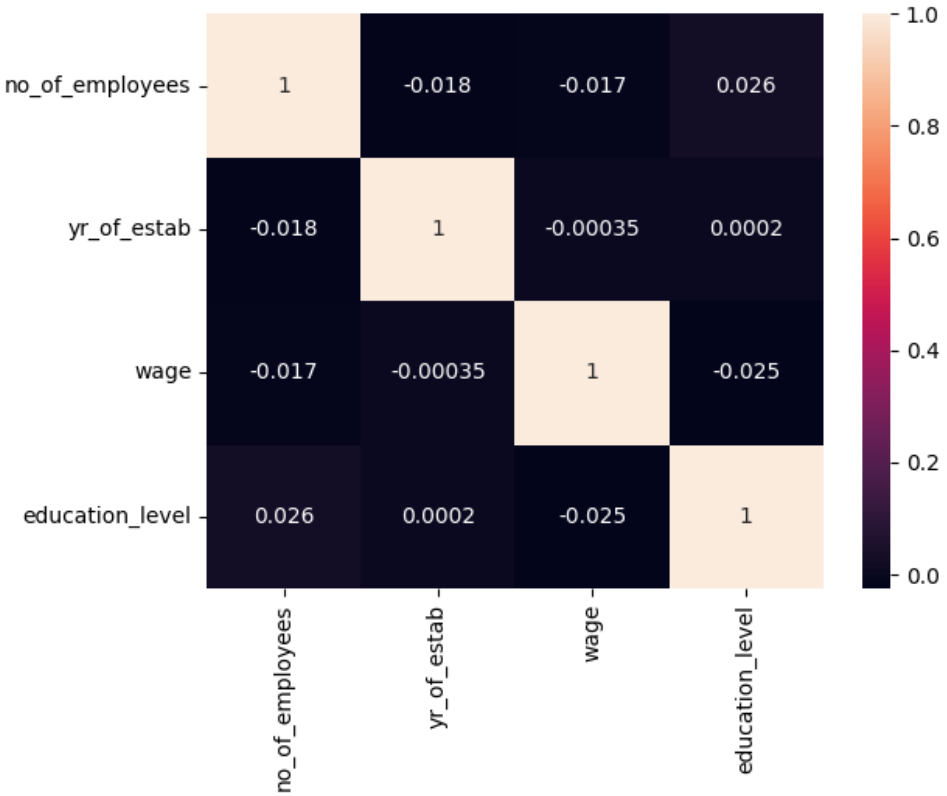

Figure 13. Heat map for numerical columns

Continent vs Case_status: From the bar chart, continents like Europe have a low percentage of Denial while South America has a higher ratio of Denied to Certified. The model will combine other features to create a classification algorithm for the case_status.

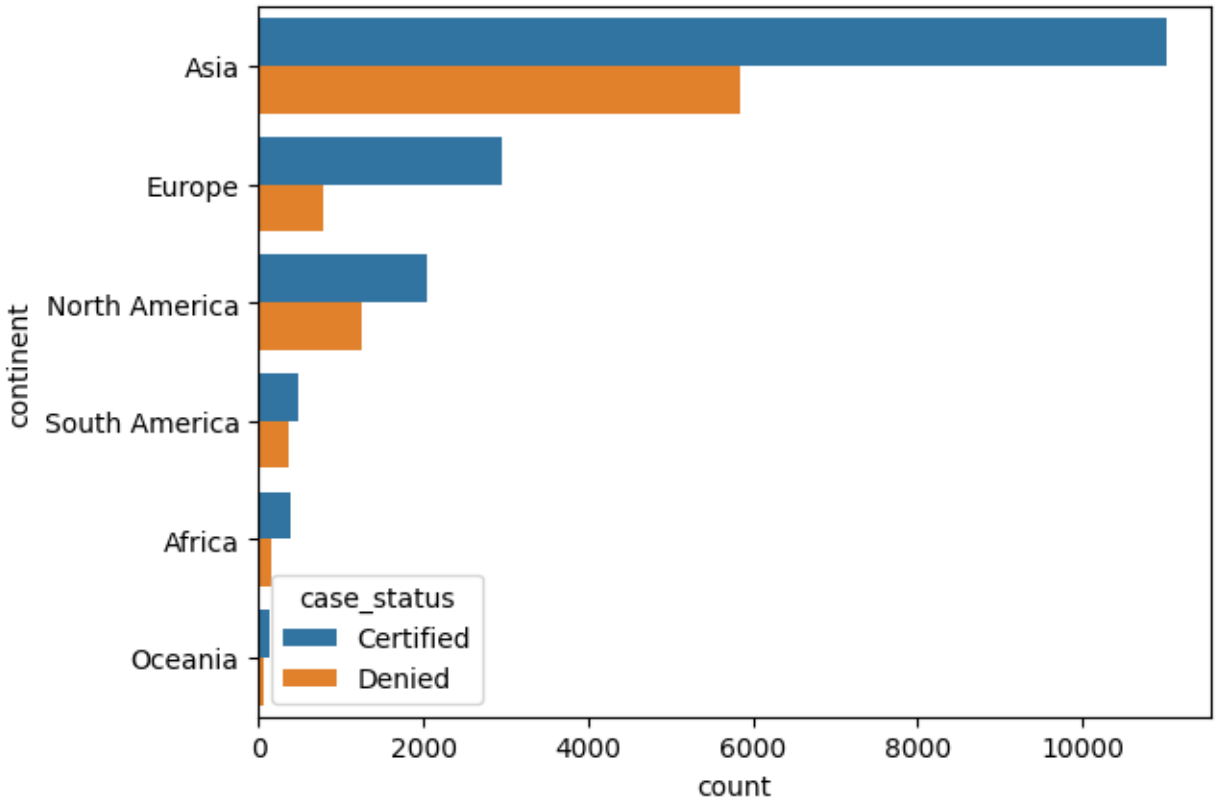

Figure 14 Bar chart for case status and continent

Education of employee vs Case status: It is quite evident from the graph that high education level increases the chance of the visa getting certified.

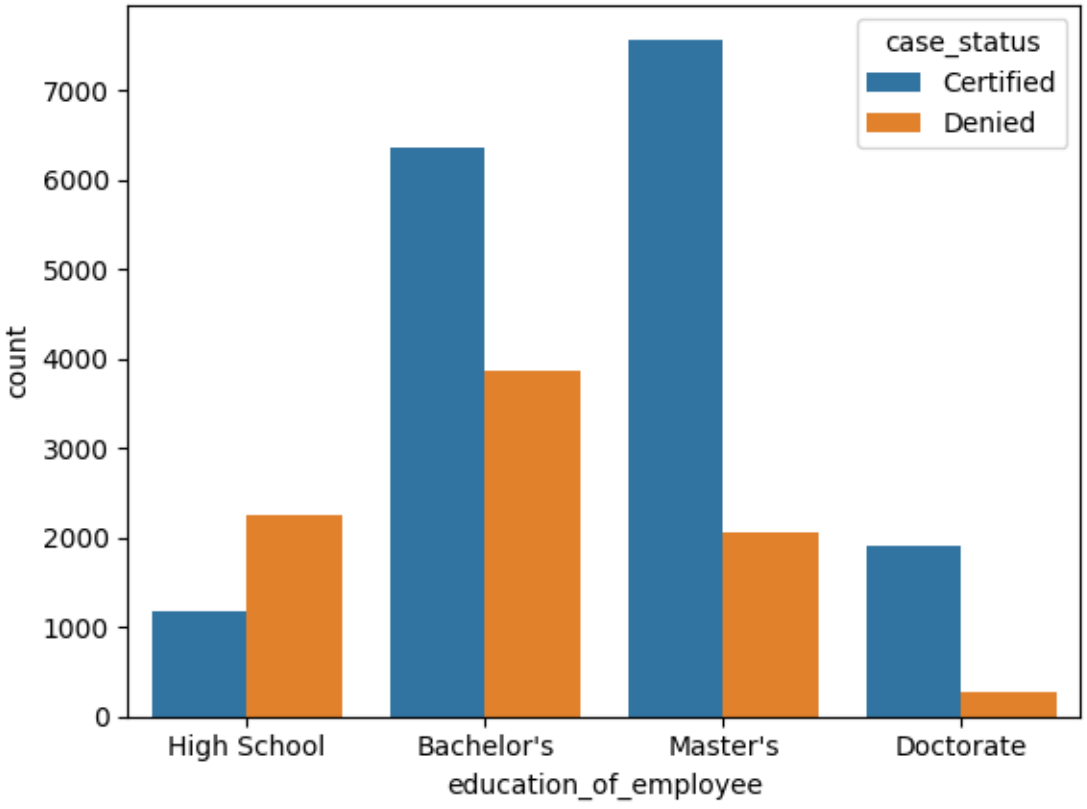

Figure 15 Bar chart for case status and education of employee

Job experience vs Case status: The bar chart below shows that, for people with job experience, there is a better chance for the visa to be certified.

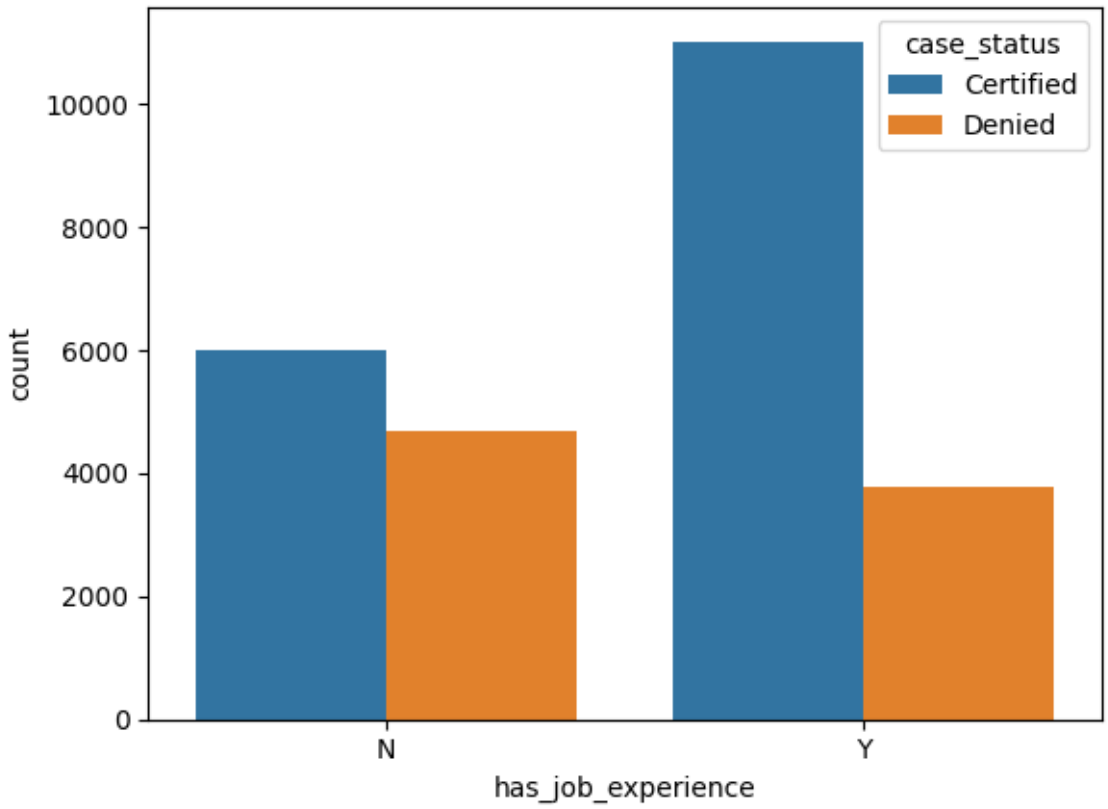

Figure 16 Bar chart for case status and job experience

Requires job training vs Case status: It is not very clear if the job training requirement plays a big role when it comes to visa certification. The decision tree-based model will provide more insights by considering other features.

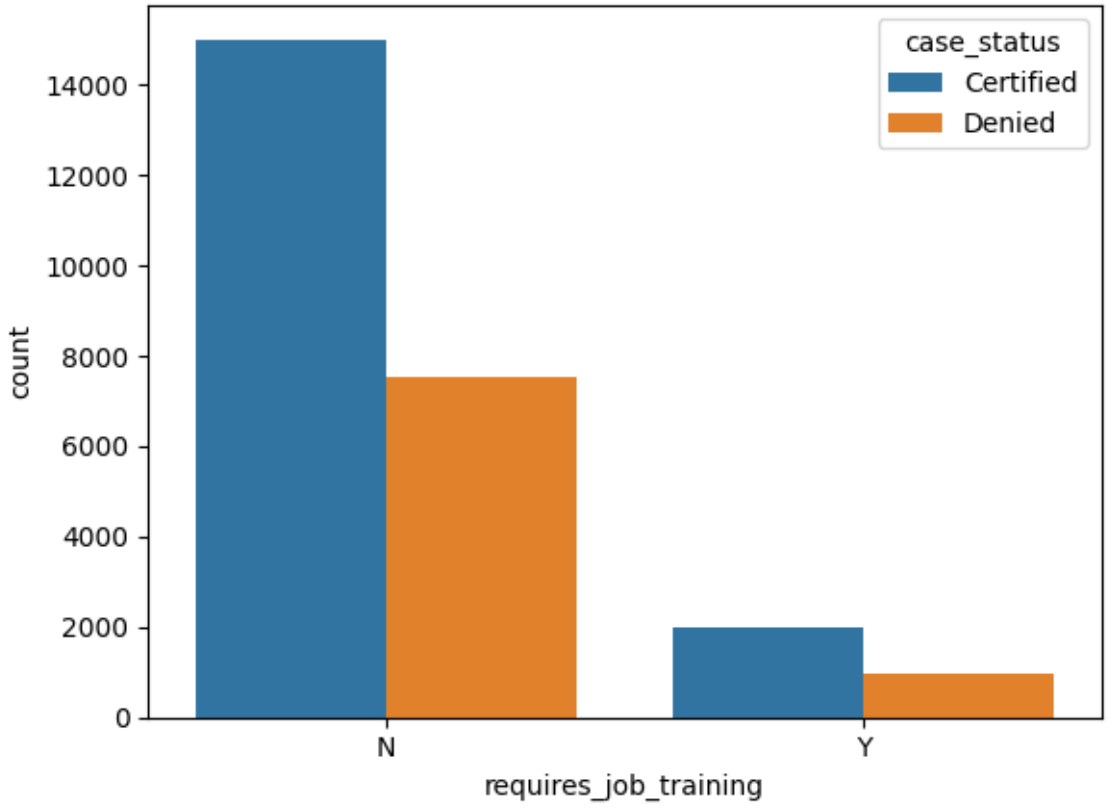

Figure 17 Bar chart for case status and job training

No_of_employees vs Case status: There is very little difference in the box plots for Certified and Denied based on the number of employees in the company. The outliers have been hidden from the graph to provide better clarity.

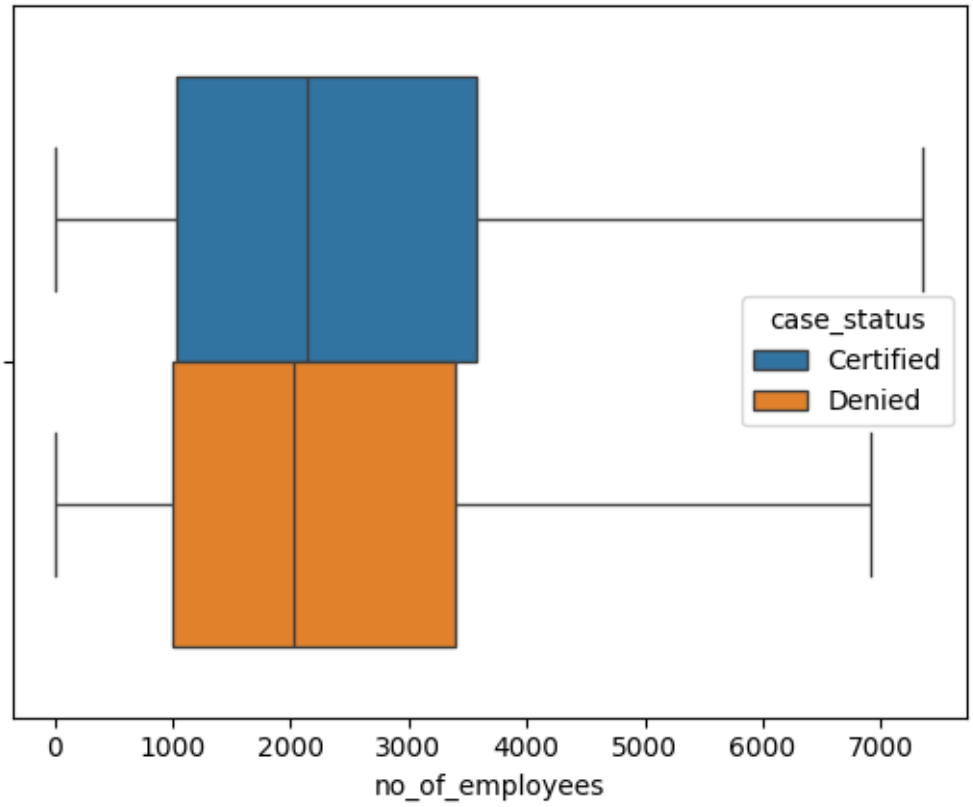

Figure 18 Box plot for number of employees grouped by case status

Year of establishment vs Case status: There is no evidence that the year of establishment plays an important role in the case_status.

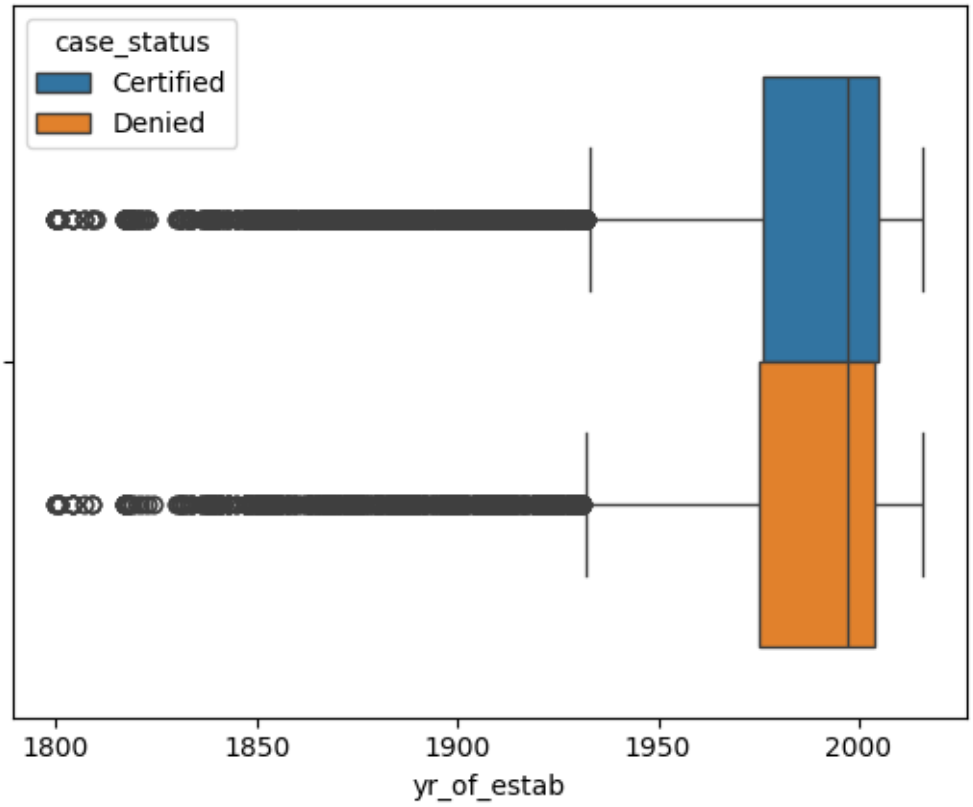

Figure 19 Box plot for year of establishment grouped by case status

Region of employment vs Case status: From the bar chart, regions like South have a low percentage of Denial while Island has a higher ratio of Denied to Certified. The model will combine other features to create a classification algorithm for the case_status.

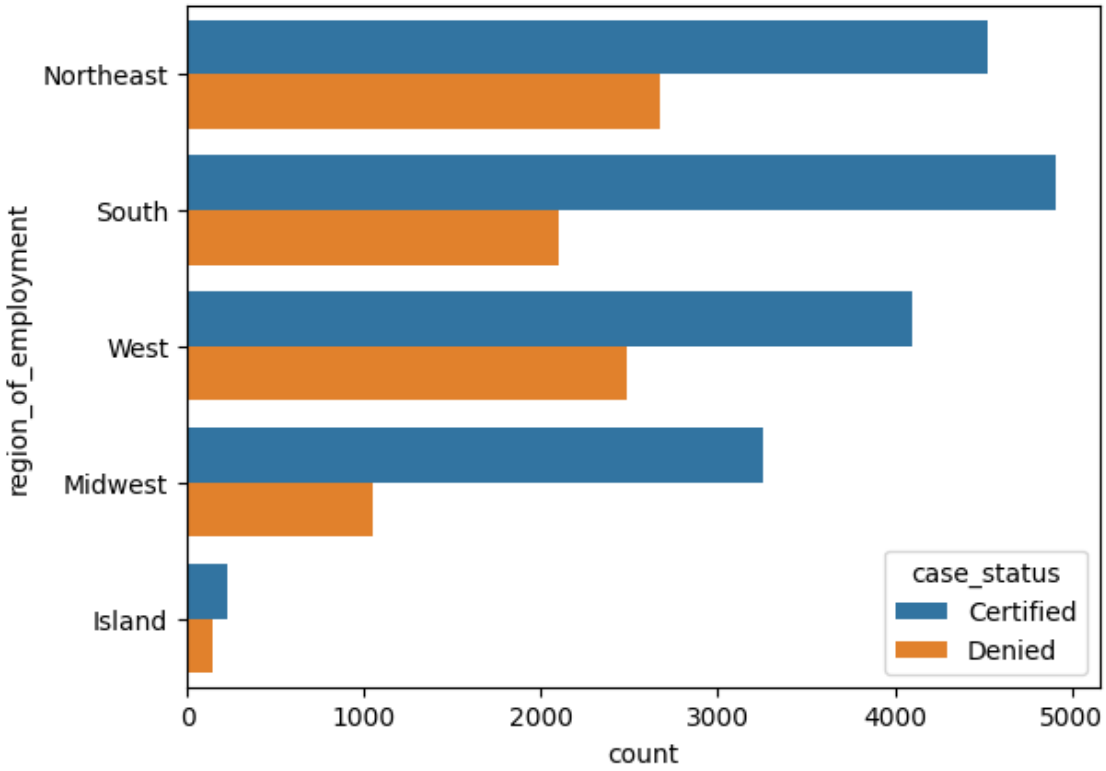

Figure 20 Bar chart for case status and region of employment

Unit of wage vs Case status: There is definitely a greater percentage of the visa getting denied for employees who earn hourly wages compared to those who earn their wages annually. There is very little data in the weekly and monthly categories.

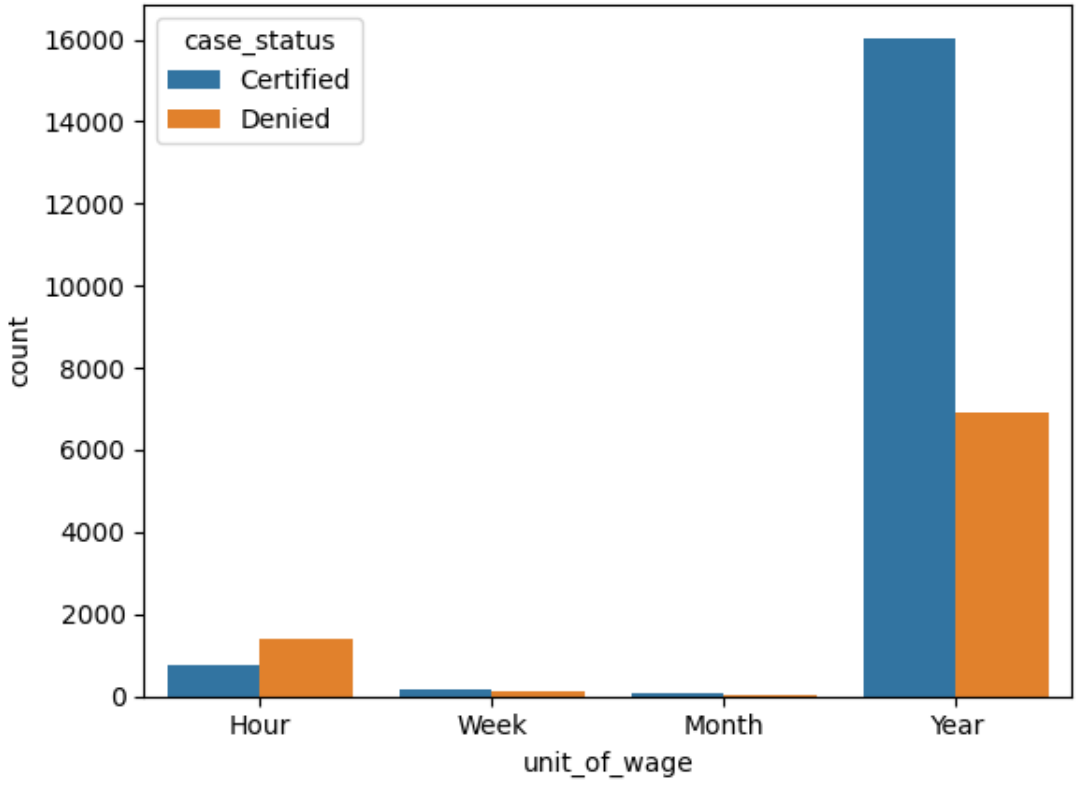

Figure 21 Bar chart for case status and unit of wage

Full time position vs Case status: It is not very clear if a full-time position plays a big role when it comes to visa certification. The decision tree-based model will provide more insights by considering other features.

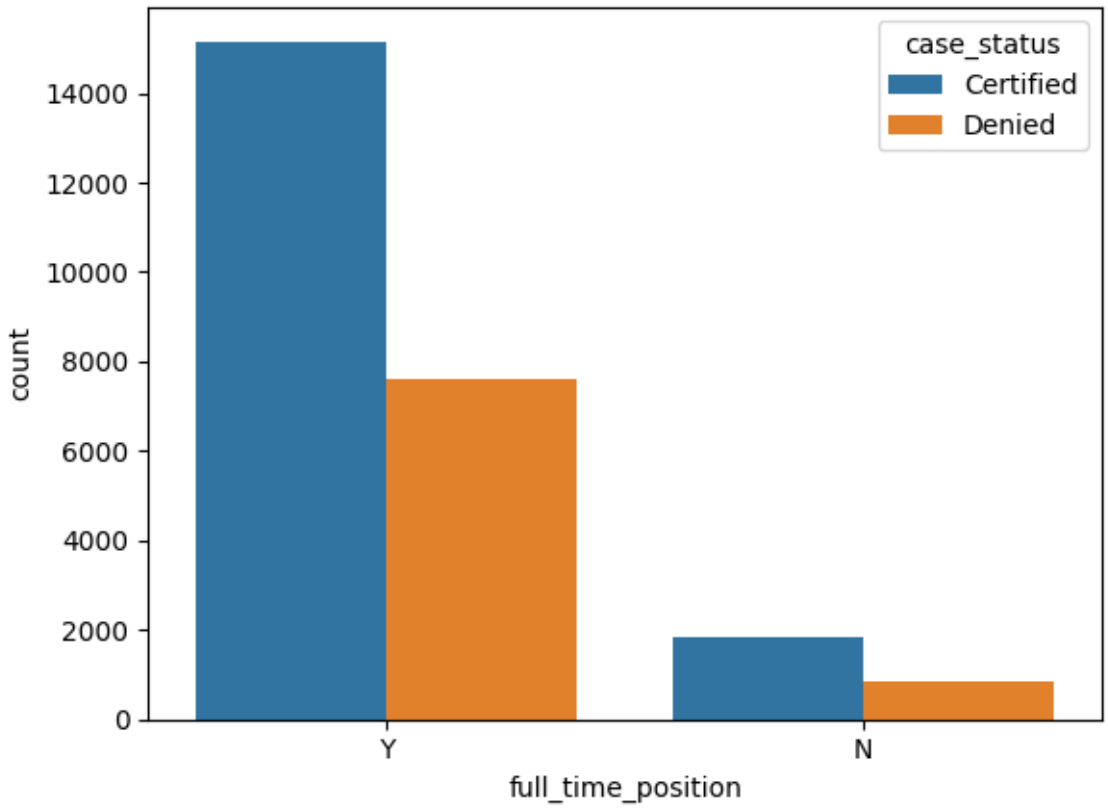

Figure 22 Bar chart for case status and full time position

Wage level vs Case status: There is no evidence that the wage level plays an important role in the case_status.

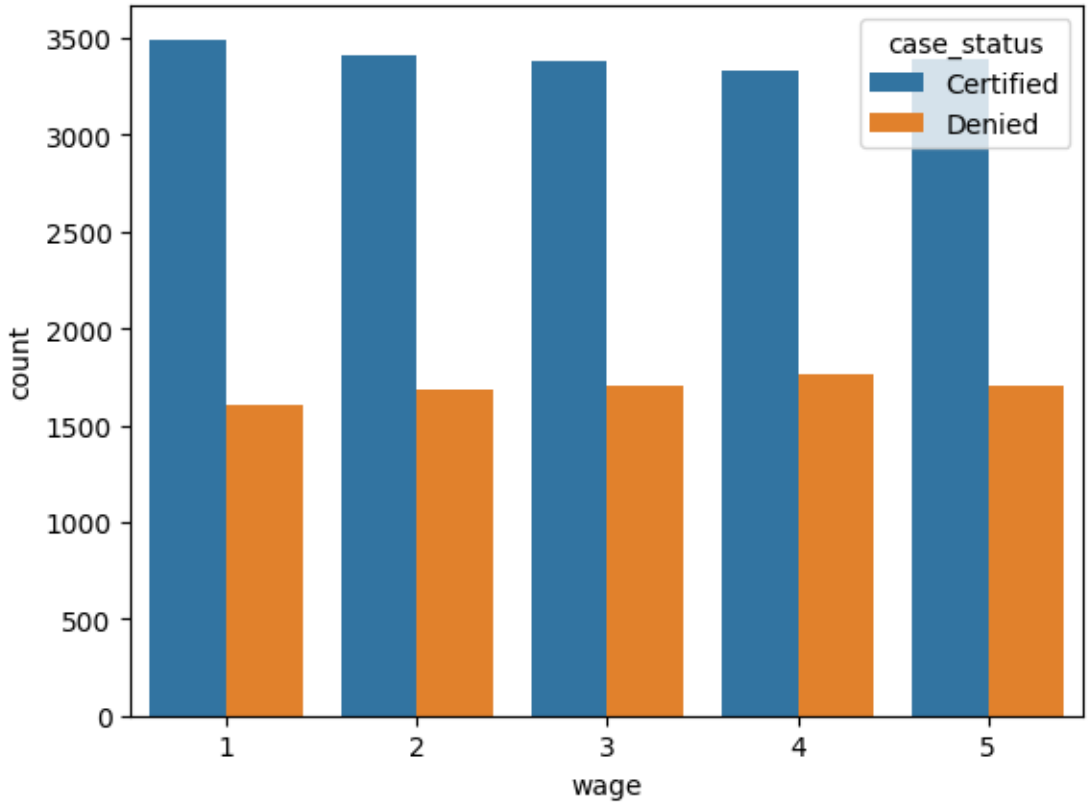

Figure 23 Bar chart for case status and wage

1.4 Key observations

- About 2/3 of the cases are certified, while the rest are denied. The dataset can be considered as balanced.
- 66% of the employees come from Asia. Among the continents, Europe has the lowest percentage of visa denial.
- Like Europe, the South region has a lower percentage of visa denial.
- The prevailing_wage column seems to contain incorrect weekly and monthly wage data. This has been treated by creating a new wage column which indicates a level from 1 to 5.
- Higher education level increases the chance of visa getting certified.
- Employees with job experience and yearly payroll also have a better chance of the visa getting certified.

2 Data Preparation

The bagging and boosting models require all data to be numeric. Along with the conversion of data into the numeric form, missing values and outliers will be treated.

2.1 Preparing the data for analysis

- Convert 'Y' and 'N' in *has_job_experience*, *requires_job_training*, and *full_time_position* to 1 and 0.
- In *no_of_employees*, convert negative values to positive since it is mostly likely a typo.
- In *case_status*, set 'Certified' to 1 and 'Denied' to 0, since 'Certified' is our class of interest.

2.2 Feature Engineering

- In *education_of_employee*, set High School to 1, Bachelor's to 2, Master's to 3 and Doctorate to 4.
- In *unit_of_wage*, set Hour to 1, Week to 2, Month to 3 and Year to 4.
- Apply one-hot encoding to *continent* and *region_of_employment*.
- Convert *prevailing_wage* to 5 levels based on quantiles.

2.3 Missing Value Treatment

There were no missing values in the dataset.

2.4 Outlier Treatment

The outliers in *prevailing_wage* were treated by converting them to 5 levels based on quantiles.

2.5 Ensure No Data Leakage among train, test, and validation sets

- First split the data into a temp set and a test set.
- Then split temp into a train set and a validation set.
- Apply additional treatments on the three sets separately.
- Use validation set for hyperparameter tuning.
- Test set is used only at the end, once the model is finalized.

```
Number of records in training set: 16307
Number of records in validation set: 4077
Number of records in test set: 5096
```

Figure 24 Number of records in training, validation, and test datasets

3 Model Building- Original Data

5 different machine learning models will be built. The techniques followed would be Bagging, Random Forest, Adaptive Boost, Gradient Boost and Extreme Gradient Boost.

3.1 Metric for model evaluation

There are two cases where the model could go wrong.

- Case 1: Model predicts that visa gets certified for a bad candidate.
- Case 2: Model predicts that visa gets denied for good candidate.

Since both cases need to be minimized, the metric for evaluation will be f1 score, since it balances precision and recall.

3.2 Model 1 – Bagging

In bagging, an ensemble of estimators, which is created by sampling the data with replacement, is used for prediction. The confusion matrix for the training and validation sets is shown below.

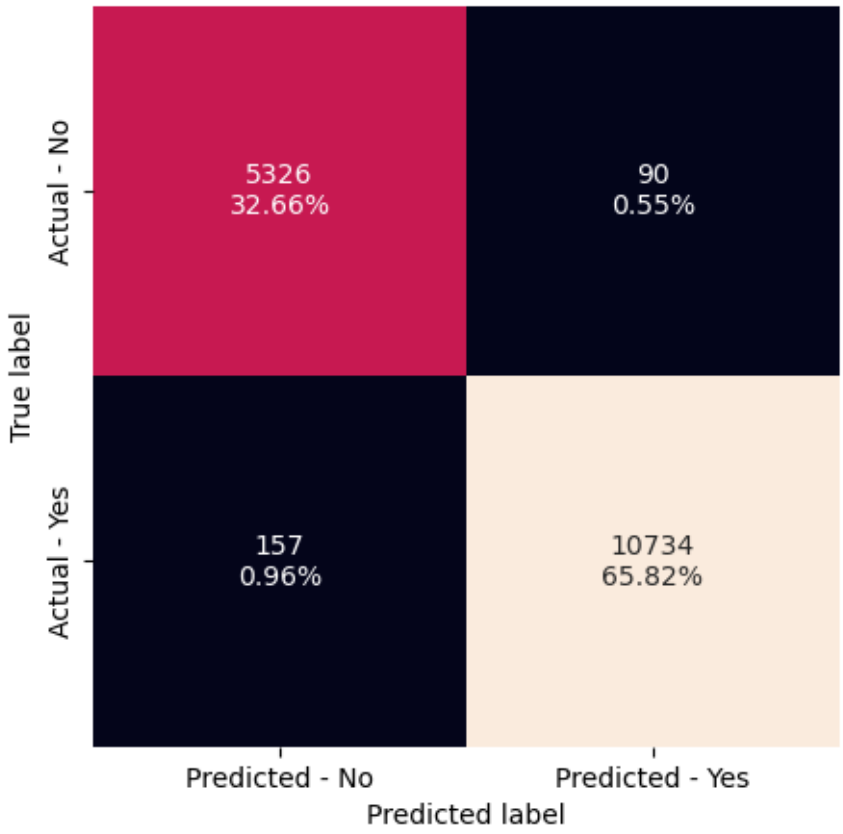

Figure 25 Confusion matrix for bagging model - Training data

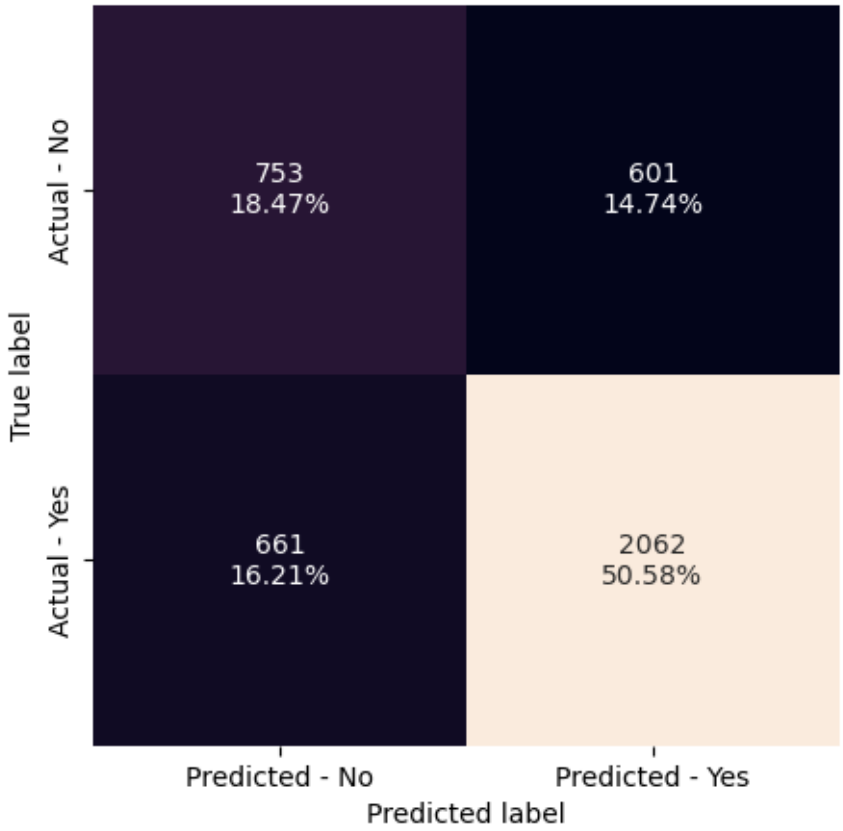

Figure 26 Confusion matrix for Bagging model - Validation data

3.3 Model 2 – Random Forest

Random forest is similar to bagging. The key difference is that, the features that are used in the estimators, are also selected randomly. The confusion matrix for the training and validation sets is shown below.

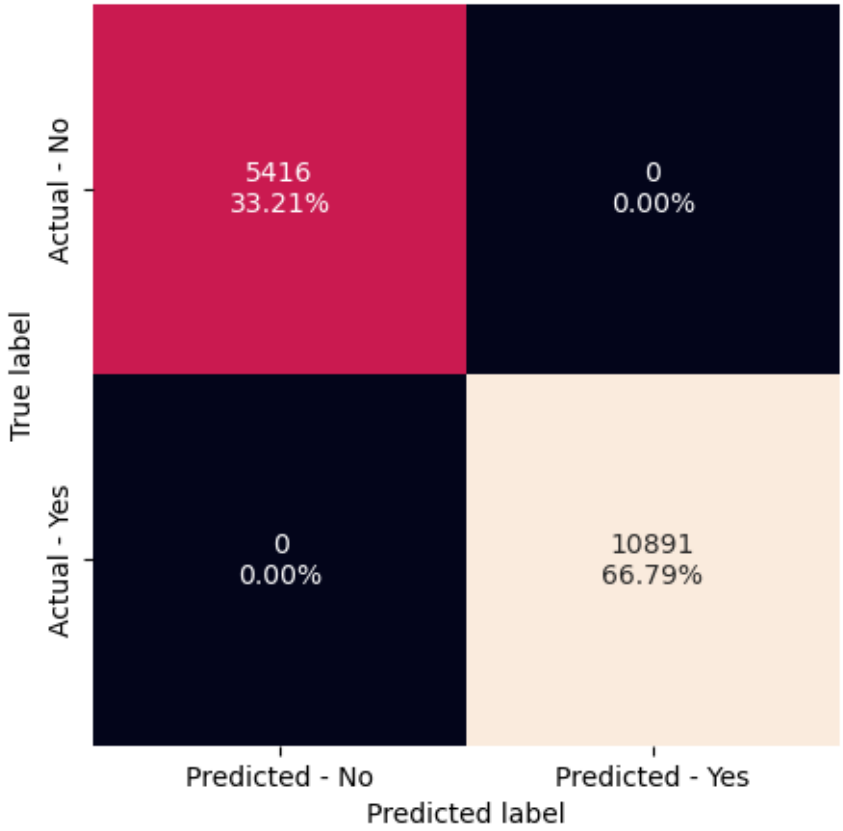

Figure 27 Confusion matrix for Random Forest model - Training data

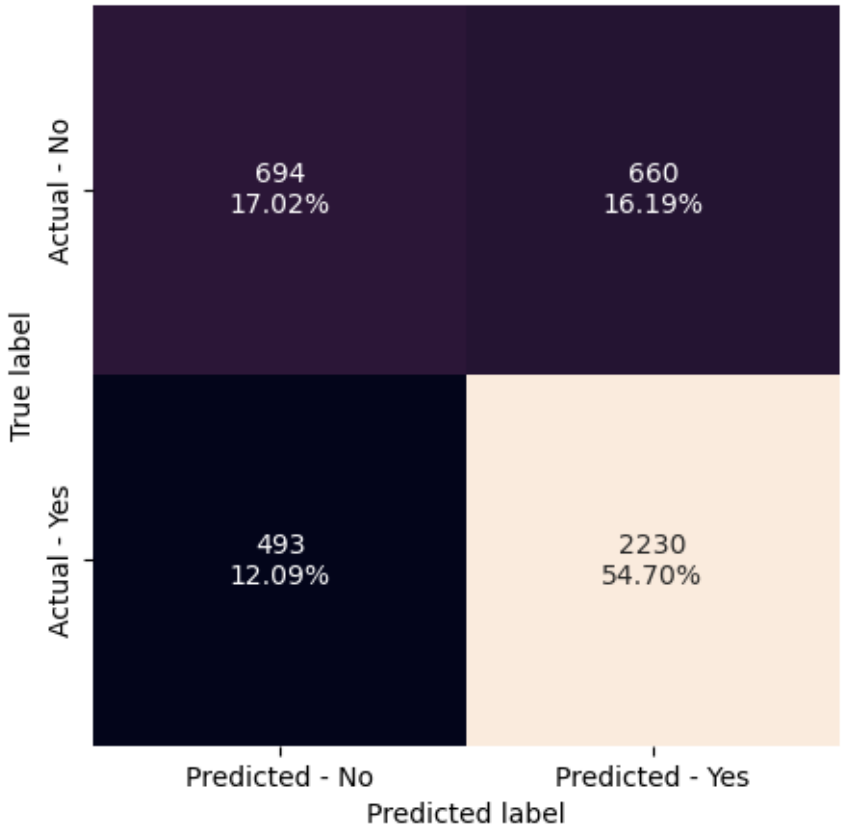

Figure 28 Confusion matrix for Random Forest model - Validation data

3.4 Model 3 – Adaptive Boost

In adaptive boosting, the samples are trained sequentially. In each iteration, the errors from the previous sample are reduced adjusting the weights. The confusion matrix for the training and validation sets is shown below.

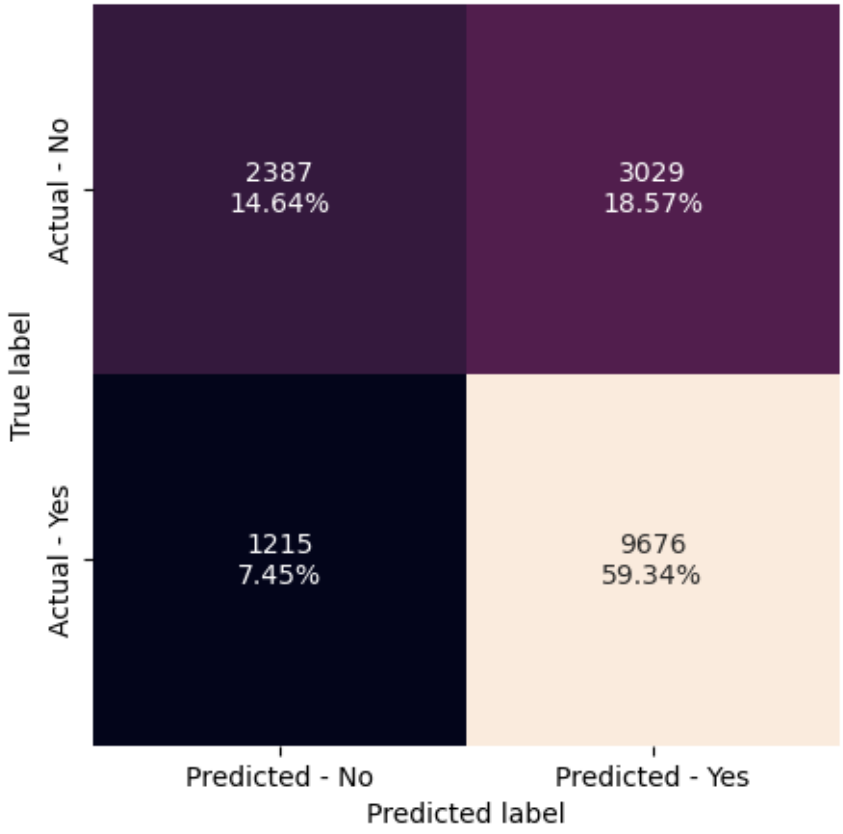

Figure 29 Confusion matrix for Adaptive Boost model - Training data

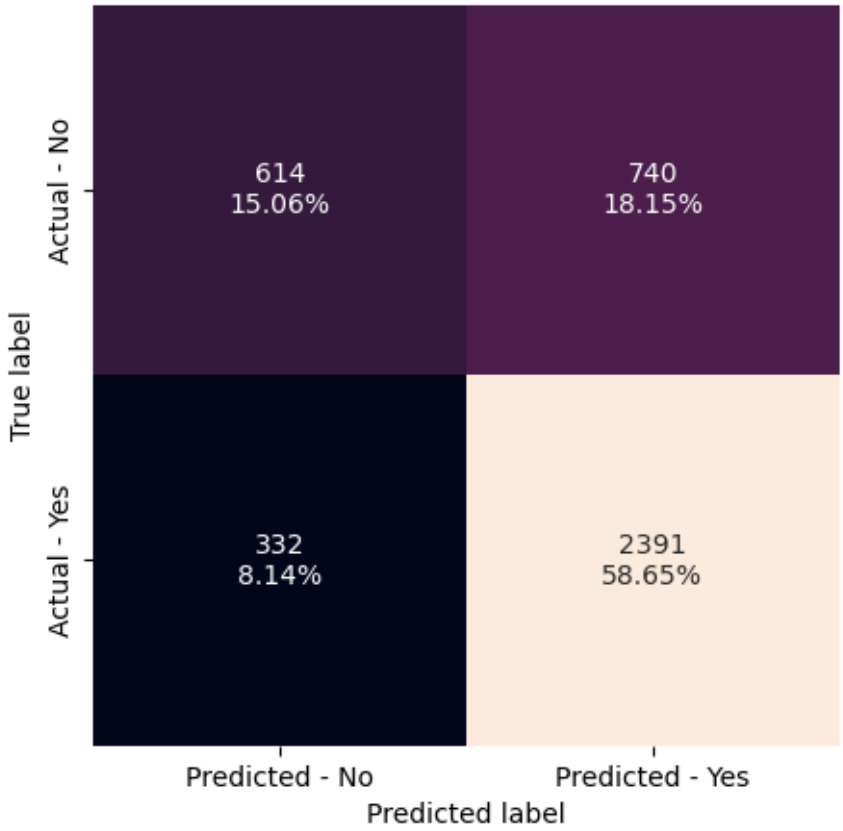

Figure 30 Confusion matrix for Adaptive Boost model - Validation data

3.5 Model 4 – Gradient Boost

Gradient boost also trains the models sequentially like Adaptive Boost. The main difference is that, instead of adjusting the sample weights, Gradient Boost works on the residuals of the previous learner. The confusion matrix for the training and validation sets is shown below.

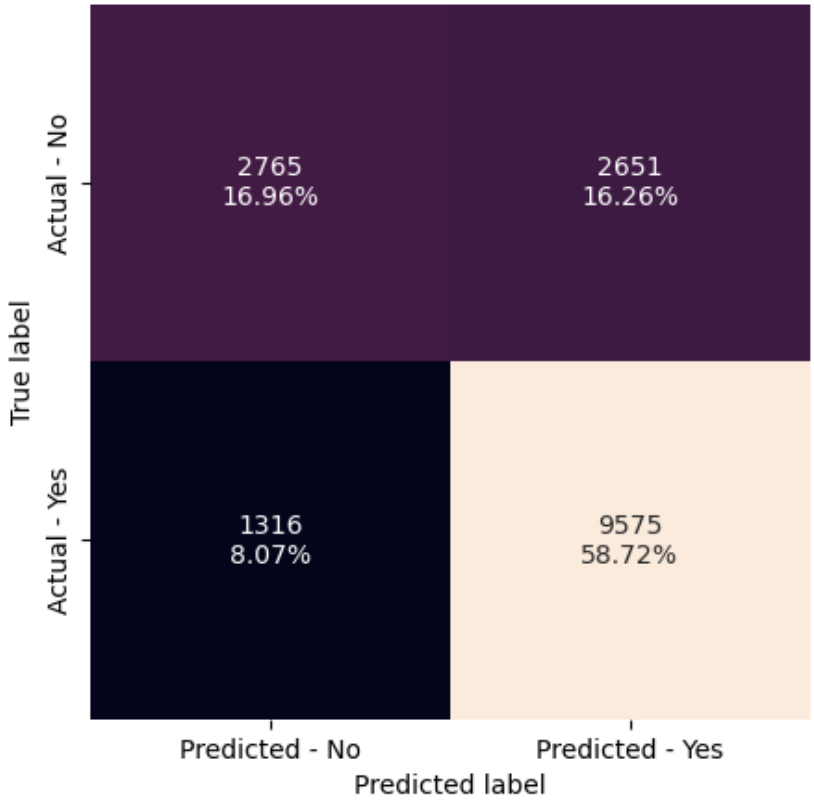

Figure 31 Confusion matrix for Gradient Boost model - Training data

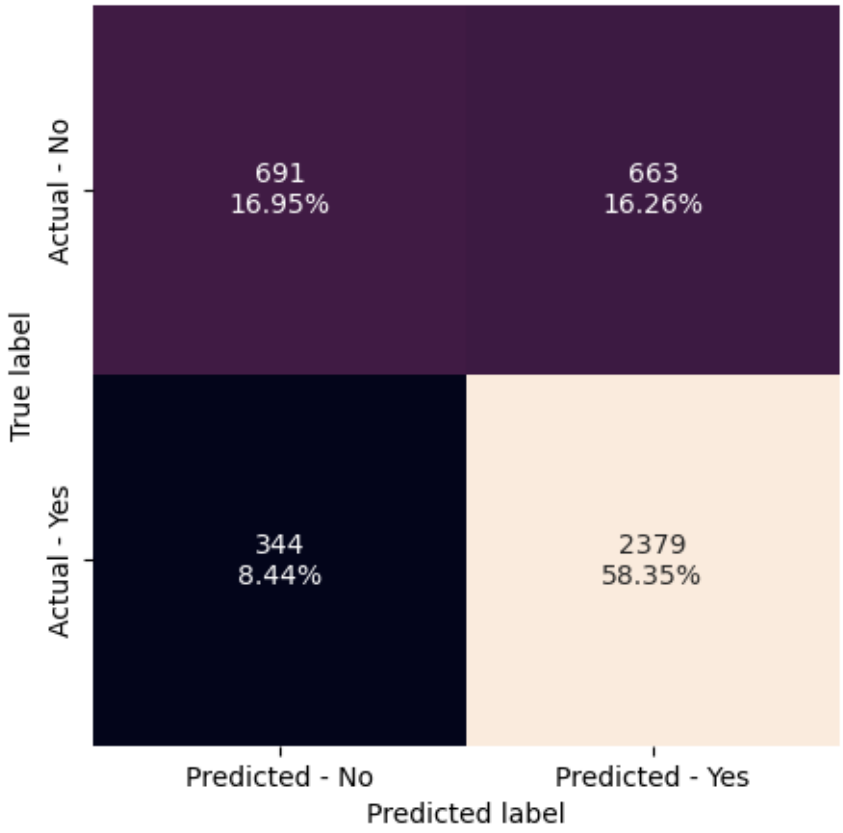

Figure 32 Confusion matrix for Gradient Boost model - Validation data

3.6 Model 5 – Extreme Gradient Boost

Extreme Gradient Boost is a variation of Gradient Boost that is tuned for high performance and faster implementation. The confusion matrix for the training and validation sets is shown below.

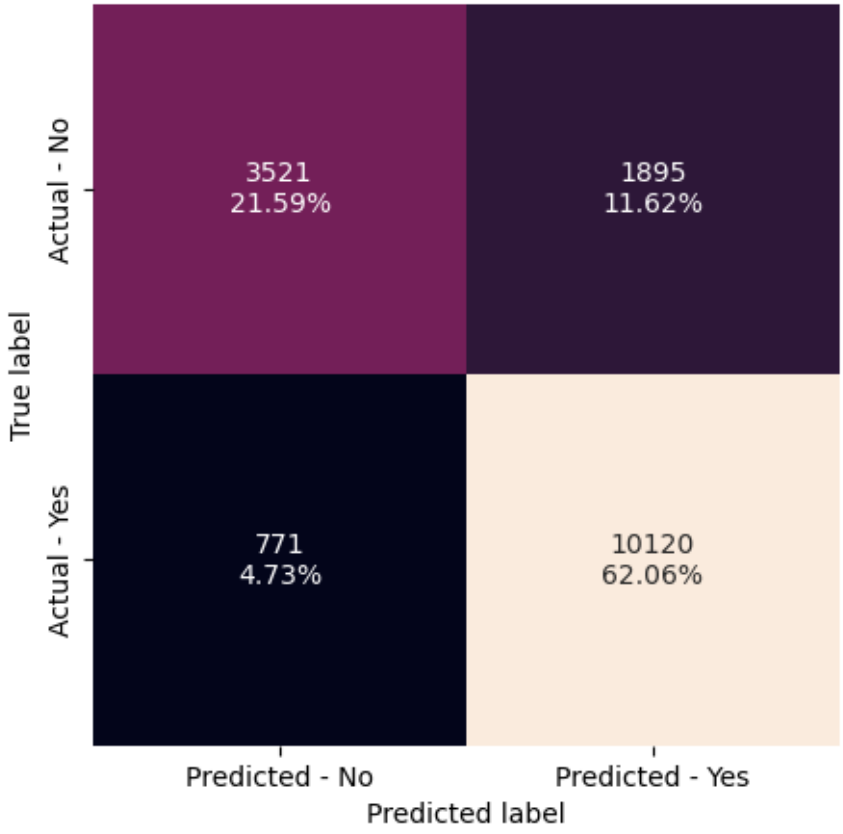

Figure 33 Confusion matrix for Extreme Gradient Boost model - Training data

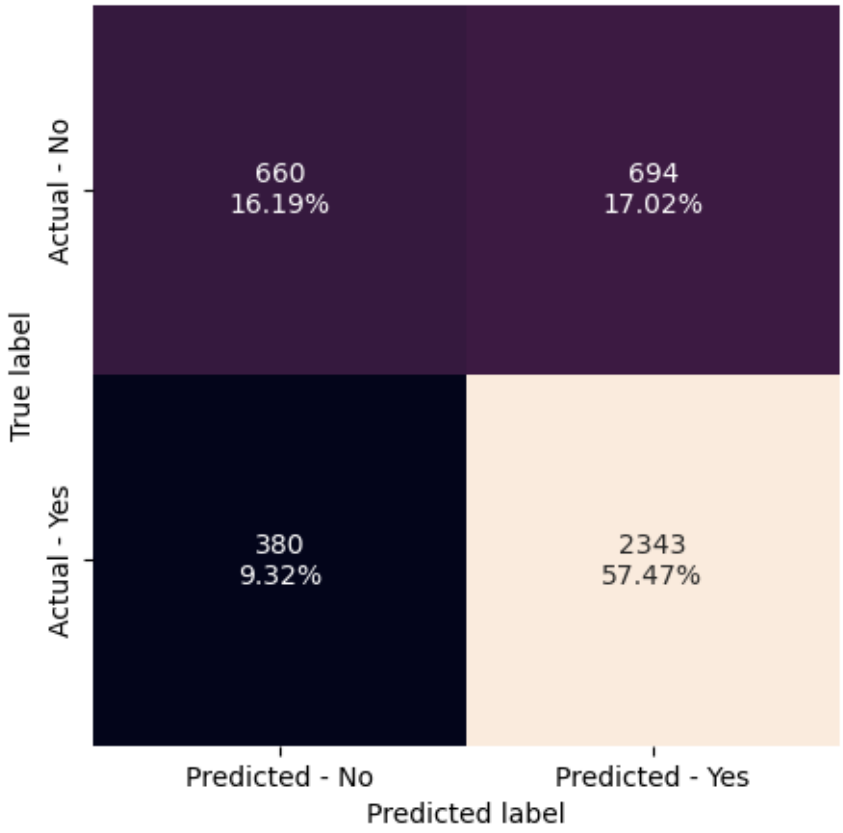

Figure 34 Confusion matrix for Extreme Gradient Boost model - Validation data

3.7 Comments on Model Performance

- The Bagging model has performed very well on the training set, but poorly on the validation sets. This indicates that the models are overfitting.
- The Random Forest model has a better score on the validation data. However, this also overfits.
- Adaptive Boosting and Gradient Boosting models perform similarly on the training and validation datasets.
- The Extreme Gradient Boosting model shows a larger difference between training and validation datasets.

4 Model Building- Oversampled Data

Oversampling and undersampling are used, when datasets are imbalanced. In this case, there are two-thirds of the records are 'Certified' and the remaining one-third is 'Denied'.

4.1 Oversample the train data

Oversampling is performed on the training data, by artificially creating additional data for the minority class using a technique called SMOTE (Synthetic Minority Oversampling Technique).

```
Original size of majority class: 10891
Original size of minority class: 5416
Size of majority class after oversampling: 10891
Size of minority class after oversampling: 10891
```

Figure 35 Size of datasets after oversampling

4.2 Model 1 – Bagging Oversampled

The confusion matrix for the training and validation sets is shown below.

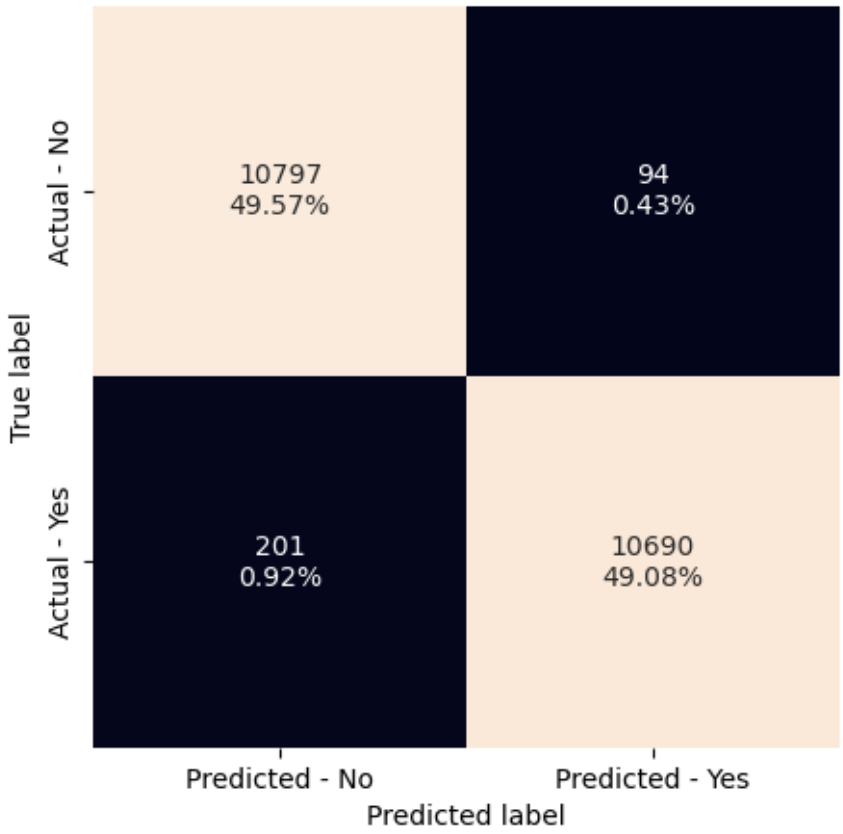

Figure 36 Confusion matrix for oversampled Bagging model - Training data

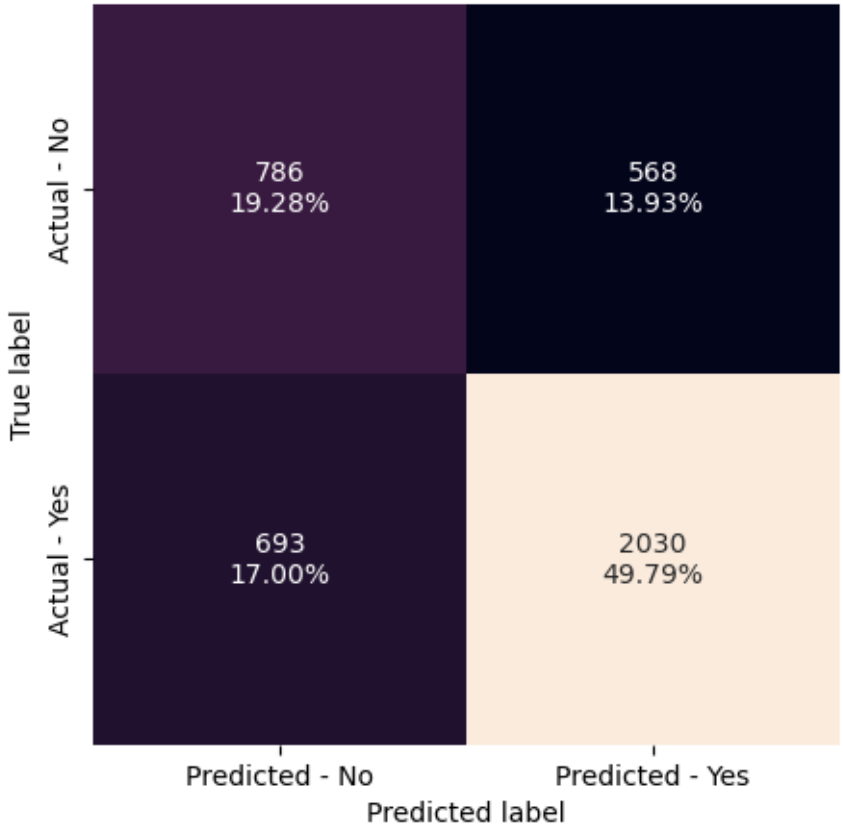

Figure 37 Confusion matrix for oversampled Bagging model - Validation data

4.3 Model 2 – Random Forest Oversampled

The confusion matrix for the training and validation sets is shown below.

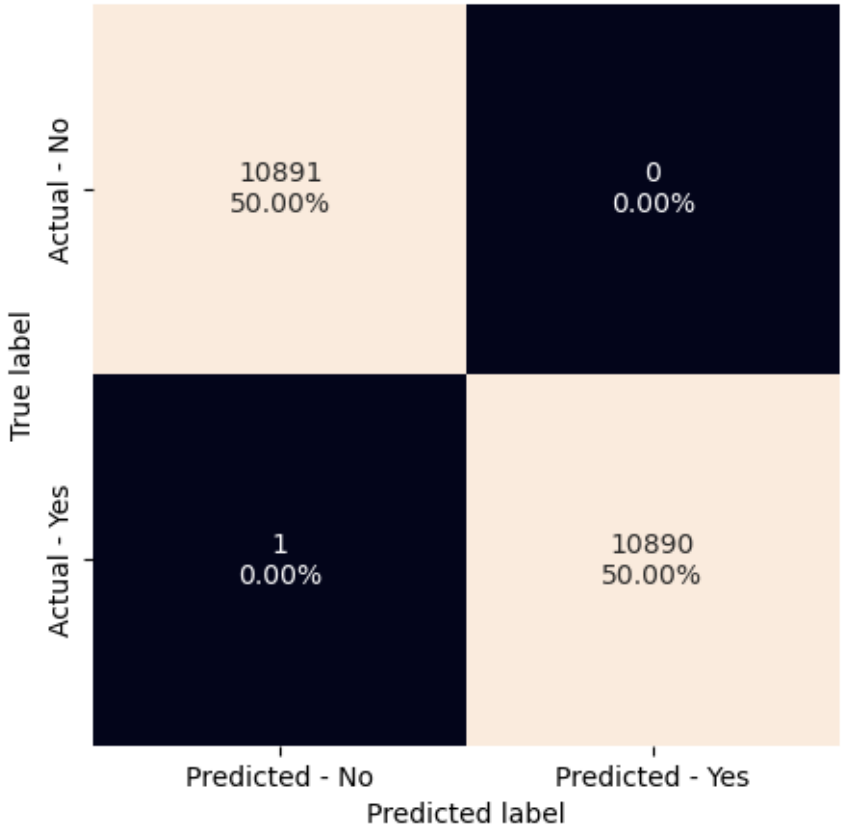

Figure 38 Confusion matrix for oversampled Random Forest model - Training data

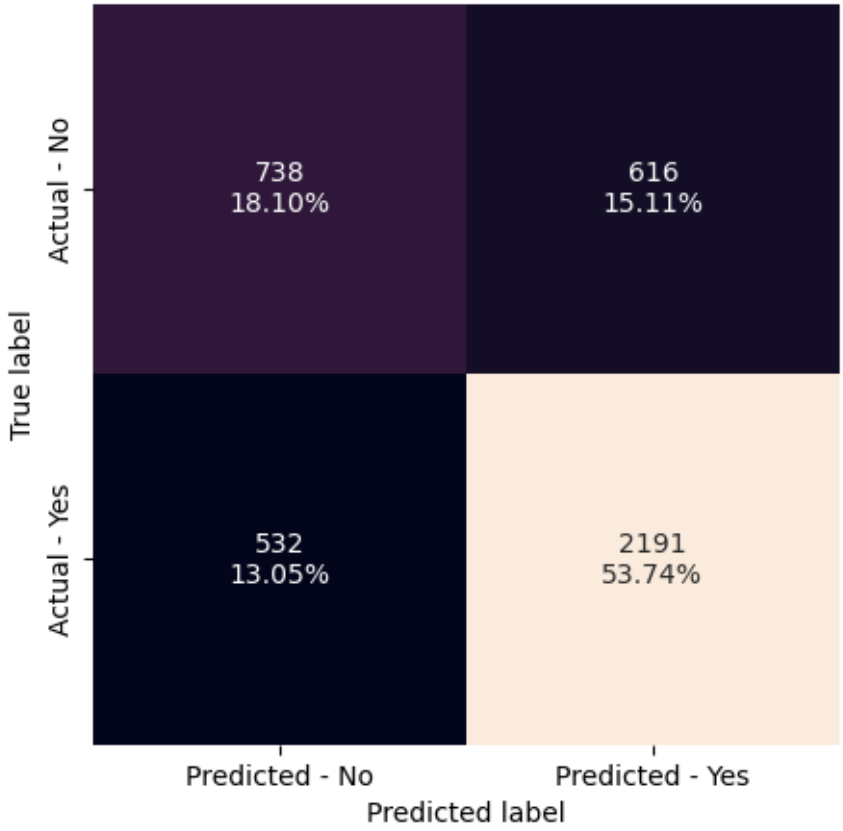

Figure 39 Confusion matrix for oversampled Random Forest model - Validation data

4.4 Model 3 – Adaptive Boost Oversampled

The confusion matrix for the training and validation sets is shown below.

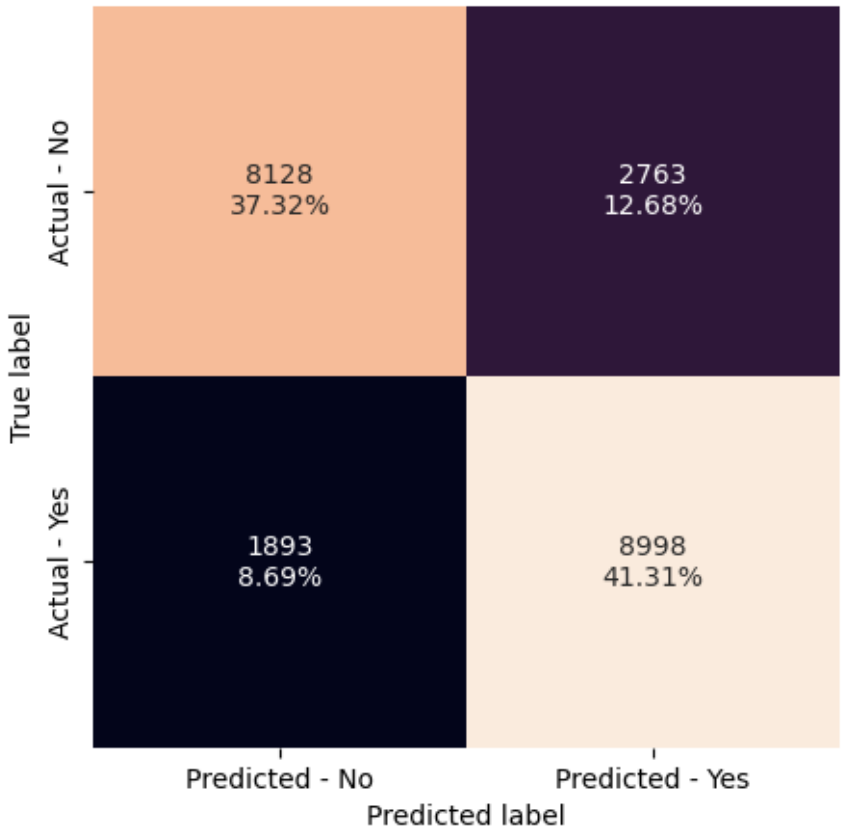

Figure 40 Confusion matrix for oversampled Adaptive Boost model - Training data

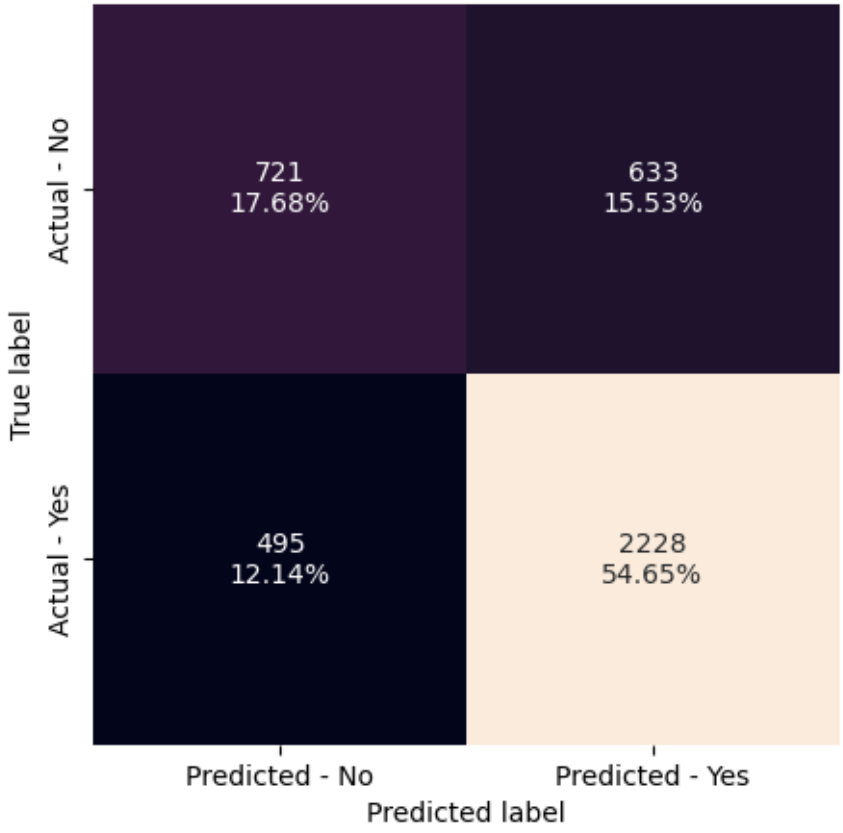

Figure 41 Confusion matrix for oversampled Adaptive Boost model - Validation data

4.5 Model 4 – Gradient Boost Oversampled

The confusion matrix for the training and validation sets is shown below.

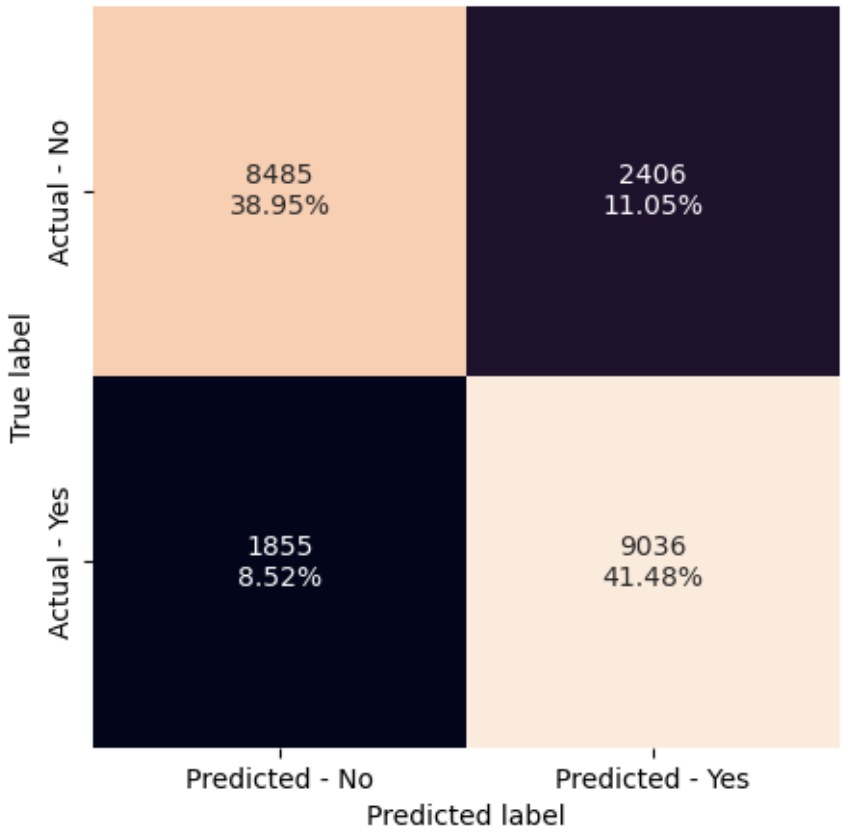

Figure 42 Confusion matrix for oversampled Gradient Boost model - Training data

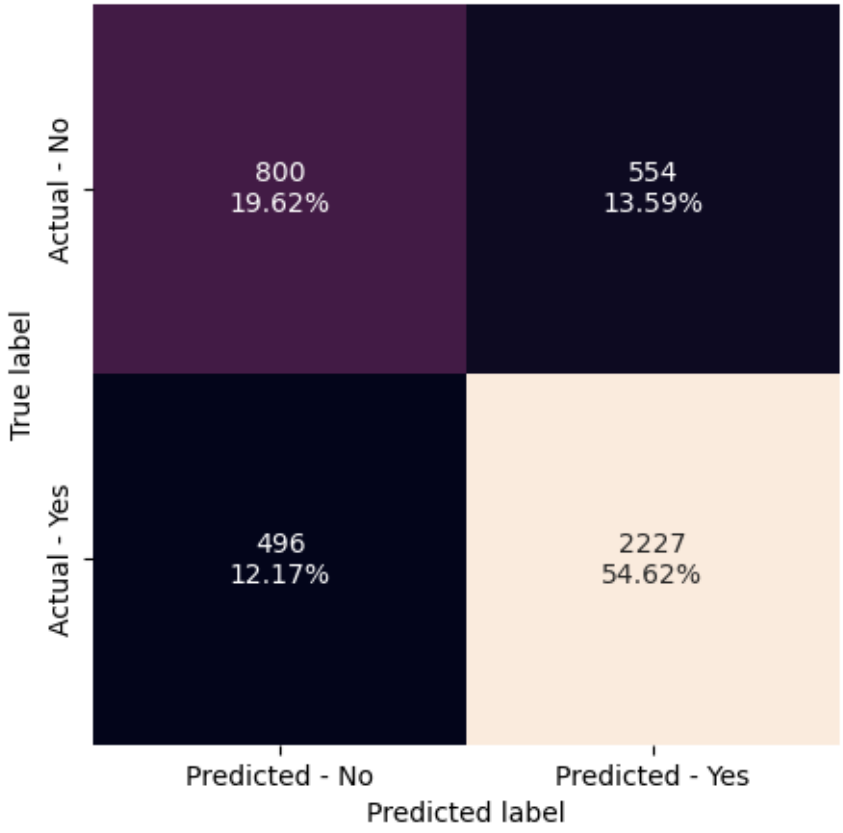

Figure 43 Confusion matrix for oversampled Gradient Boost model - Validation data

4.6 Model 5 – Extreme Gradient Boost Oversampled

The confusion matrix for the training and validation sets is shown below.

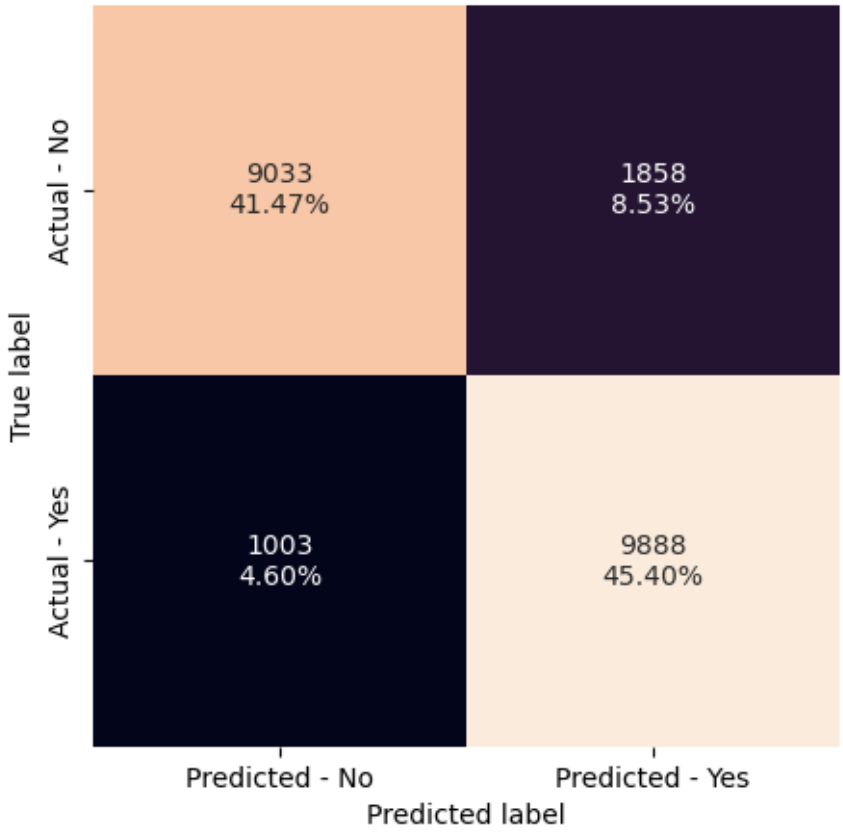

Figure 44 Confusion matrix for oversampled Extreme Gradient Boost model - Training data

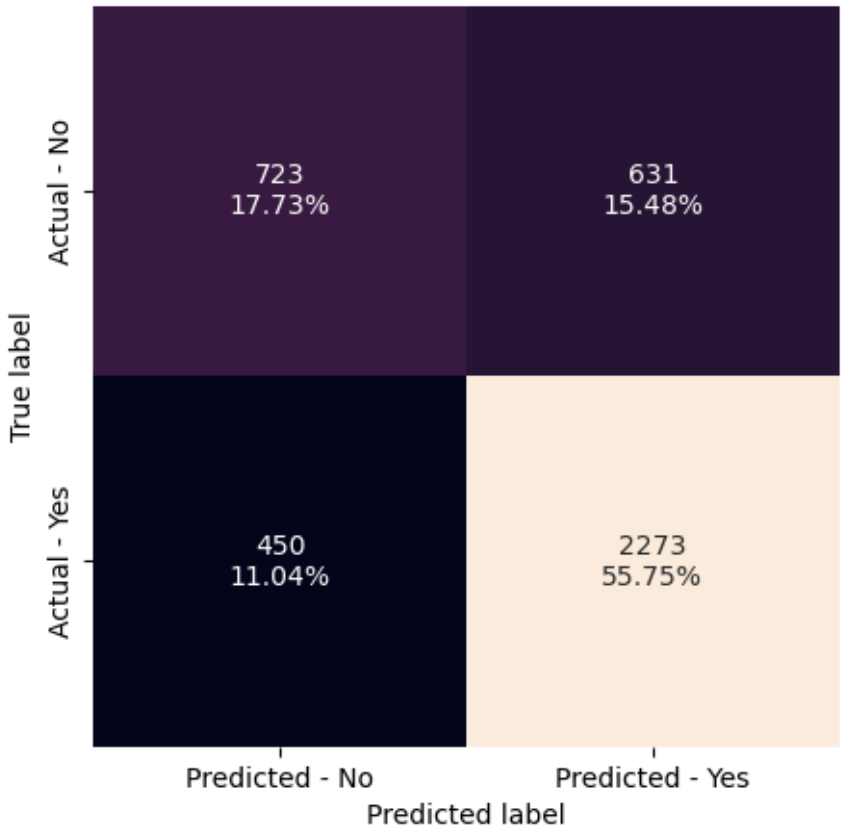

Figure 45 Confusion matrix for oversampled Extreme Gradient Boost model - Validation data

4.7 Comments on Model Performance

- The performance of the models built on the oversampled data is similar to that of the models built on the original data.
- The dataset is not highly imbalanced. As a result, the difference between the results is not significant.
- Oversampling increases the size of the dataset. It is more computationally expensive. This data will not be used further in the analysis.

5 Model Building- Undersampled Data

In undersampling, data is removed from the majority class to reduce the imbalance.

5.1 Undersample the training data

Undersampling is performed by a technique called Tomek links, where the data from the majority class is removed in such a way that there is an increased separation in the majority and minority class clusters.

```
Original size of majority class: 10891
Original size of minority class: 5416
Size of majority class after undersampling: 5416
Size of minority class after undersampling: 5416
```

Figure 46 Size of datasets after undersampling

5.2 Model 1 – Bagging Undersampled

The confusion matrix for the training and validation sets is shown below.

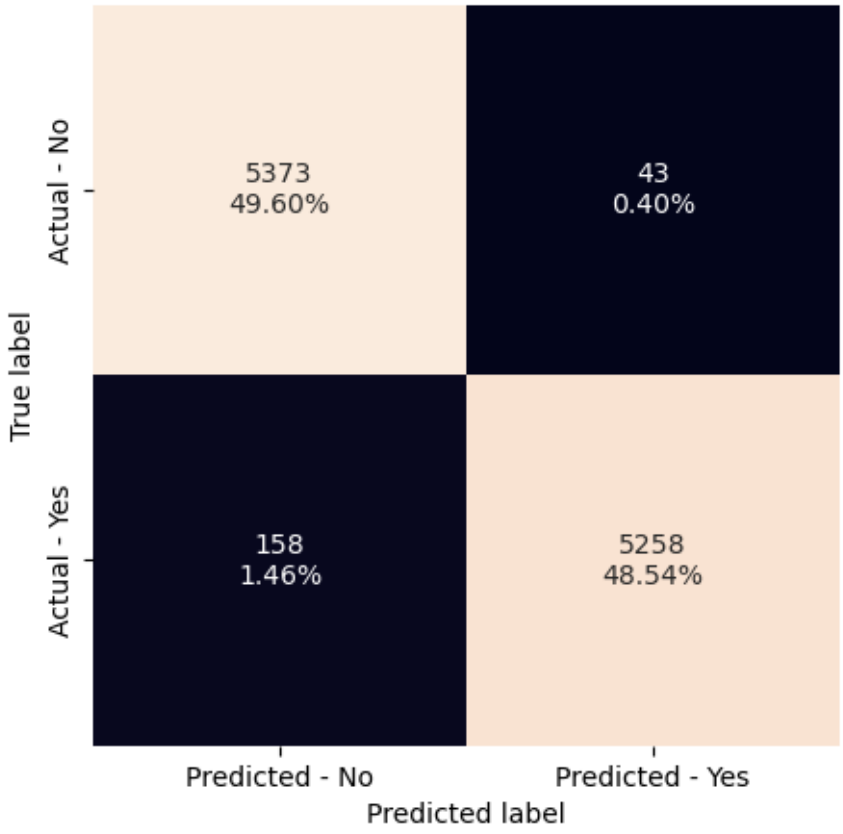

Figure 47 Confusion matrix for undersampled Bagging model - Training data

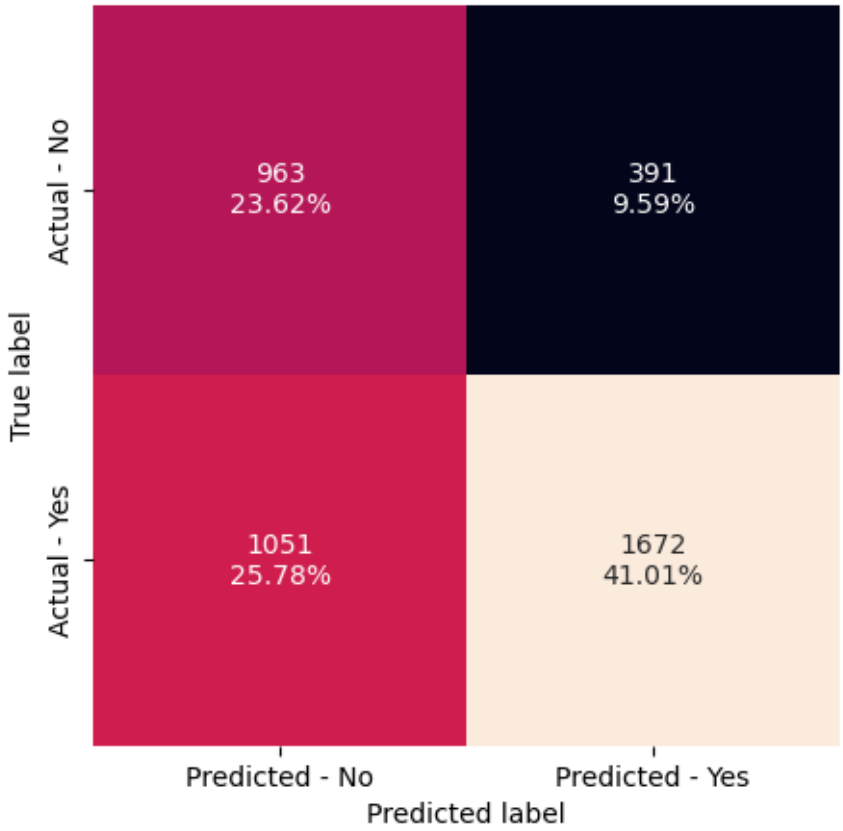

Figure 48 Confusion matrix for undersampled Bagging model - Validation data

5.3 Model 2 – Random Forest Undersampled

The confusion matrix for the training and validation sets is shown below.

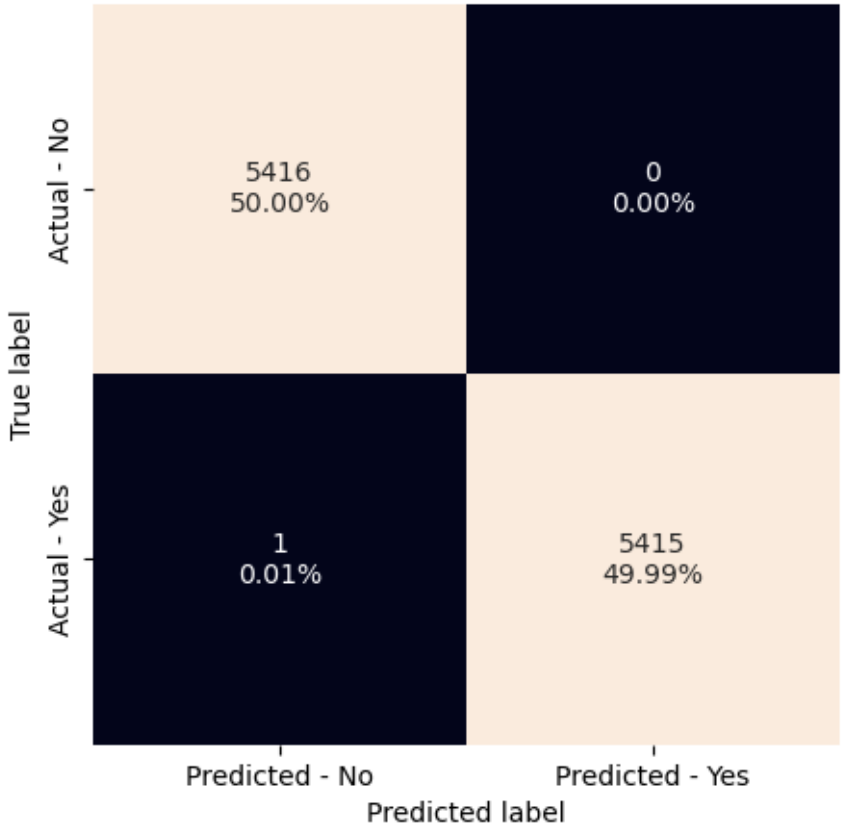

Figure 49 Confusion matrix for undersampled Random Forest model - Training data

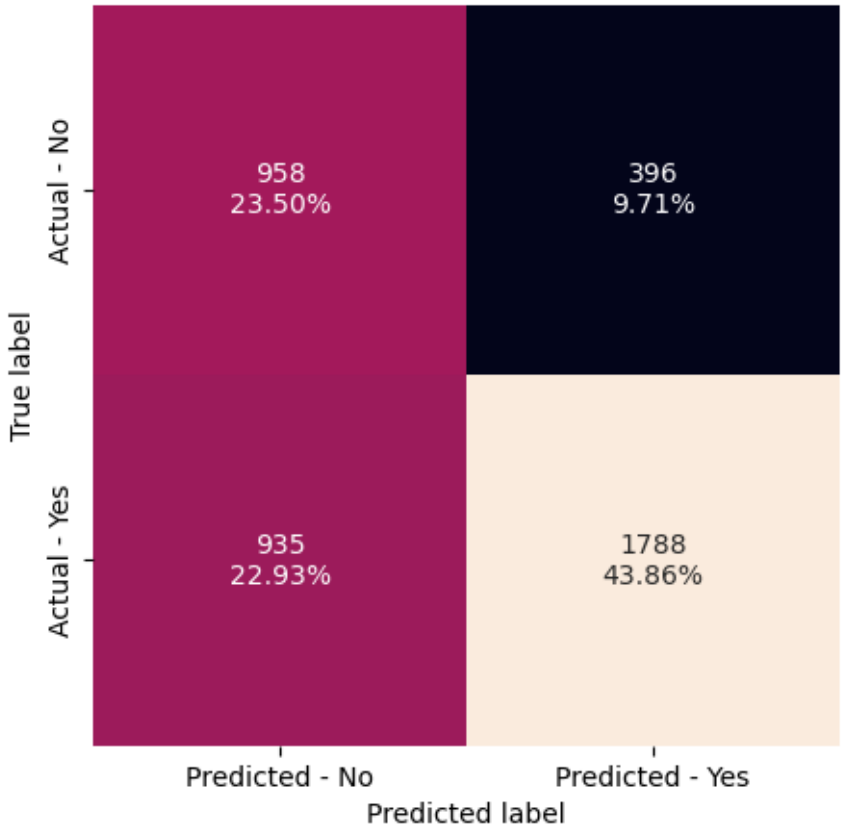

Figure 50 Confusion matrix for undersampled Random Forest model - Validation data

5.4 Model 3 – Adaptive Boost Undersampled

The confusion matrix for the training and validation sets is shown below.

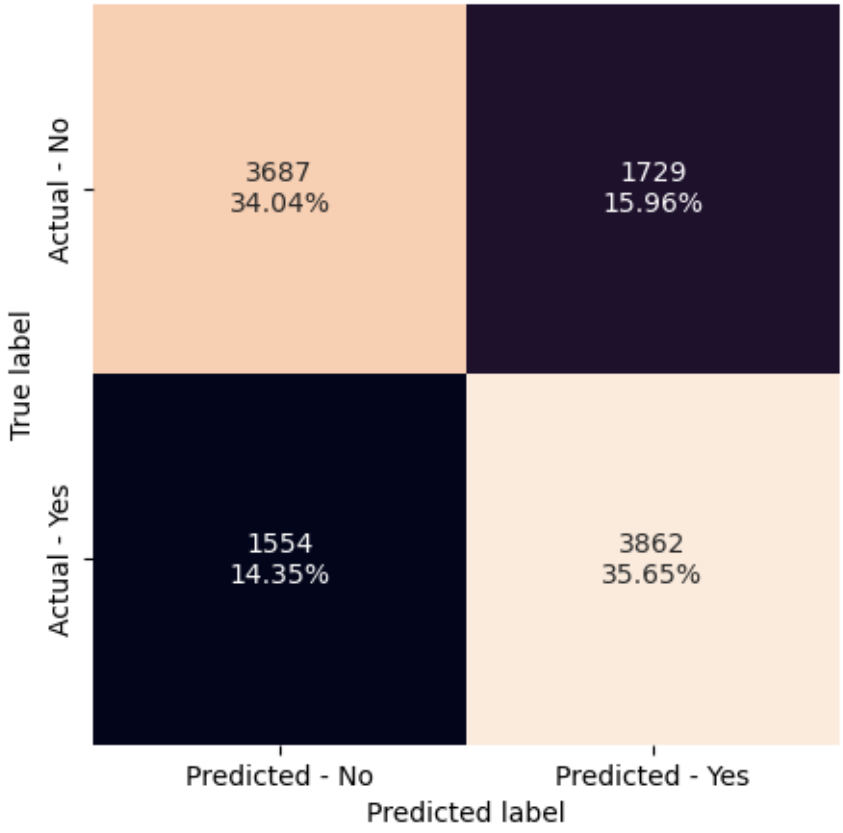

Figure 51 Confusion matrix for undersampled Adaptive Boost model - Training data

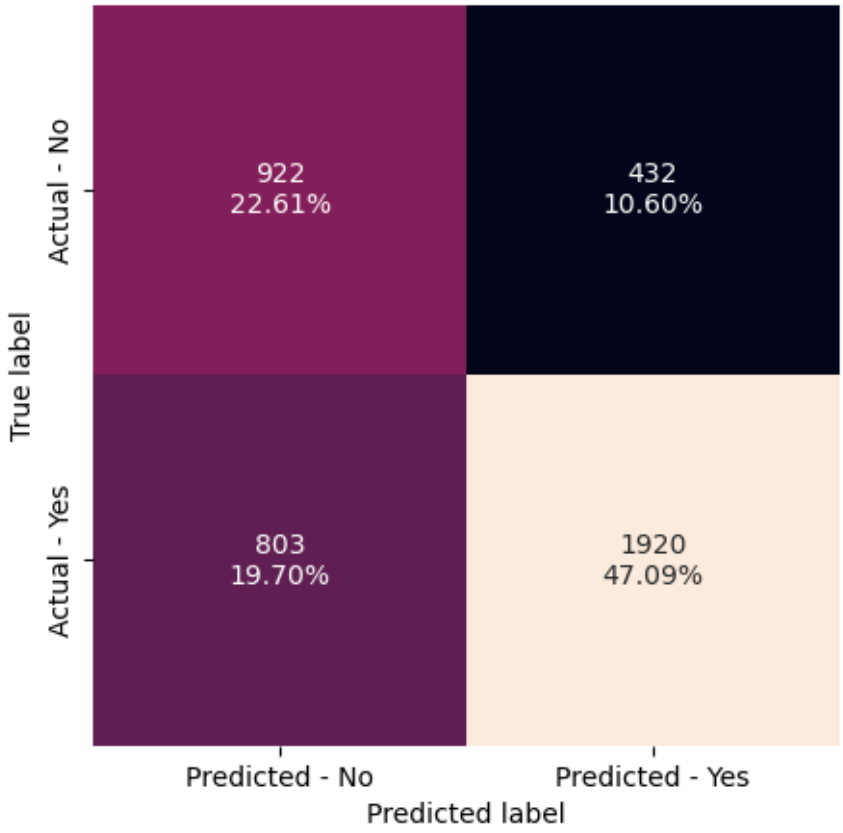

Figure 52 Confusion matrix for undersampled Adaptive Boost model - Validation data

5.5 Model 4 – Gradient Boost Undersampled

The confusion matrix for the training and validation sets is shown below.

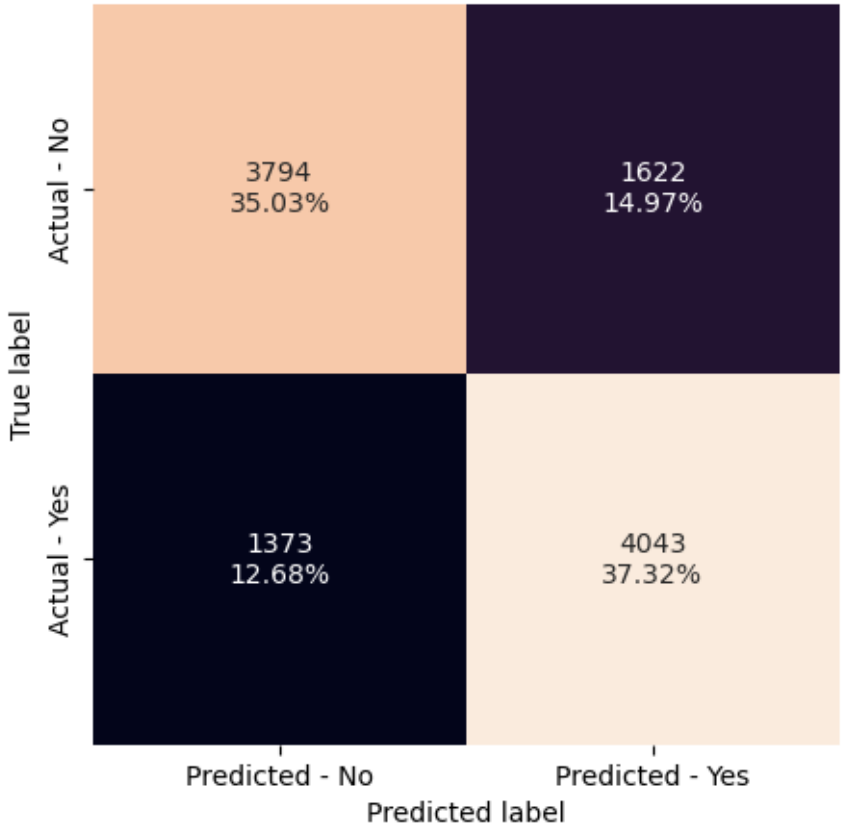

Figure 53 Confusion matrix for undersampled Gradient Boost model - Training data

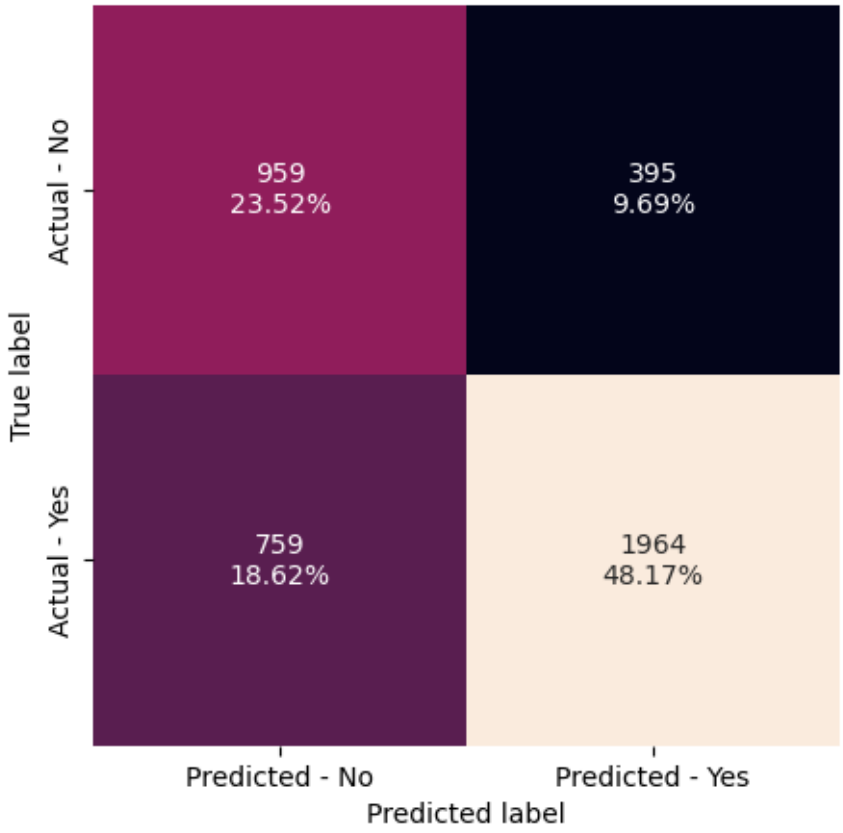

Figure 54 Confusion matrix for undersampled Gradient Boost model - Validation data

5.6 Model 5 – Extreme Gradient Boost Undersampled

The confusion matrix for the training and validation sets is shown below.

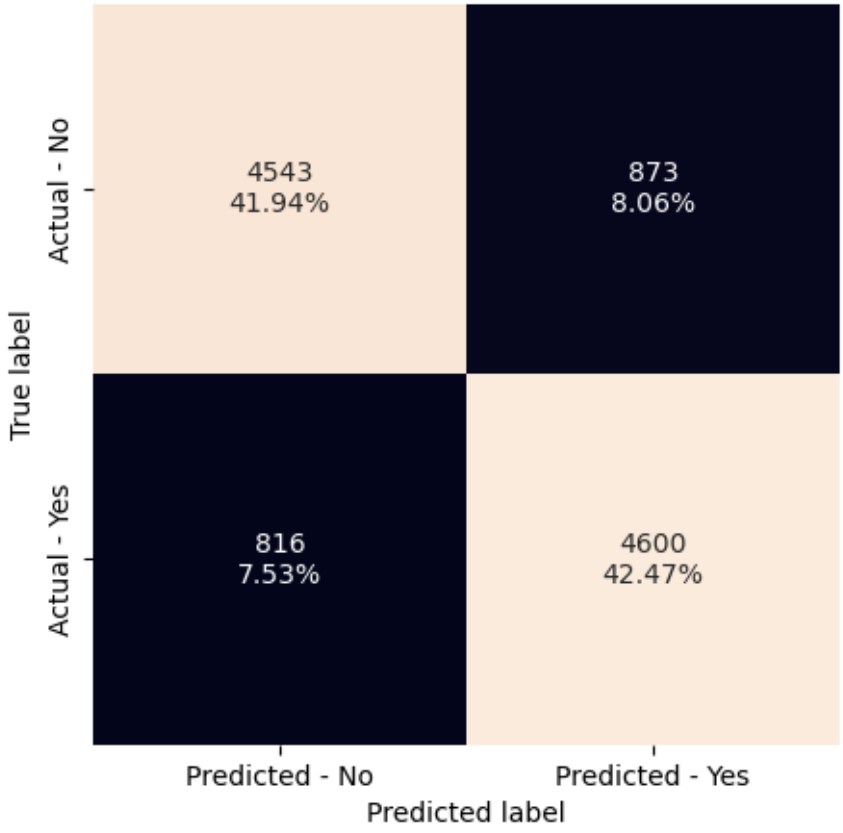

Figure 55 Confusion matrix for undersampled Extreme Gradient Boost model - Training data

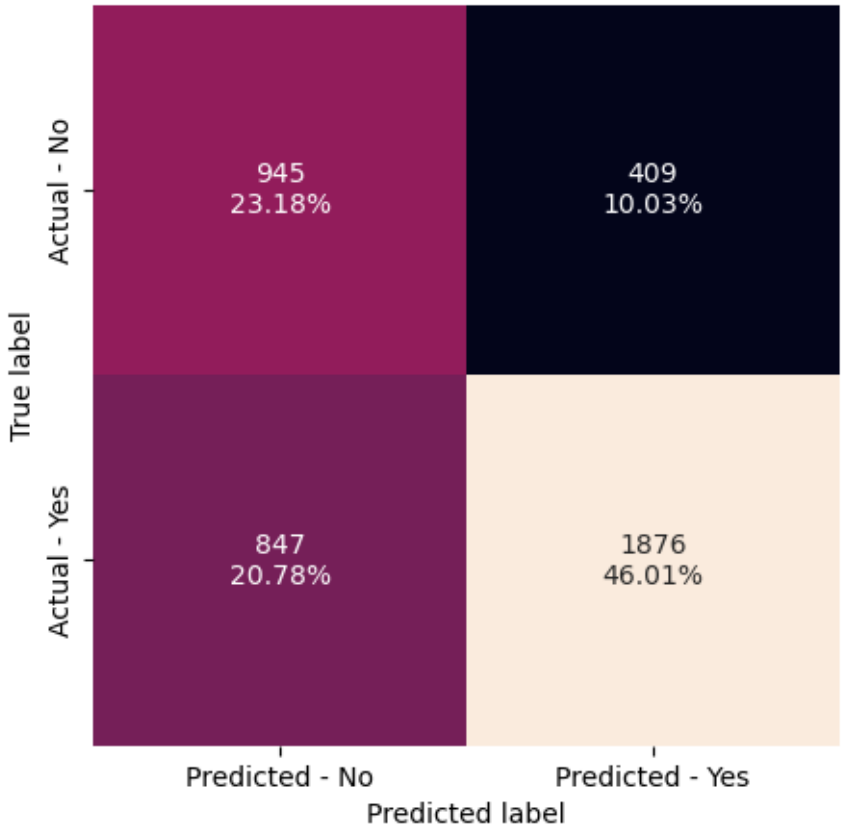

Figure 56 Confusion matrix for undersampled Extreme Gradient Boost model - Validation data

5.7 Comments on Model Performance

- The performance of the models built on the undersampled data is low.
- Data is lost in undersampling.
- It can be faster, but since the performance is not good enough, this data will not be used further.

6 Model Performance Improvement- Hyperparameter Tuning

The accuracy, recall, precision and f1 score for all the models is shown below.

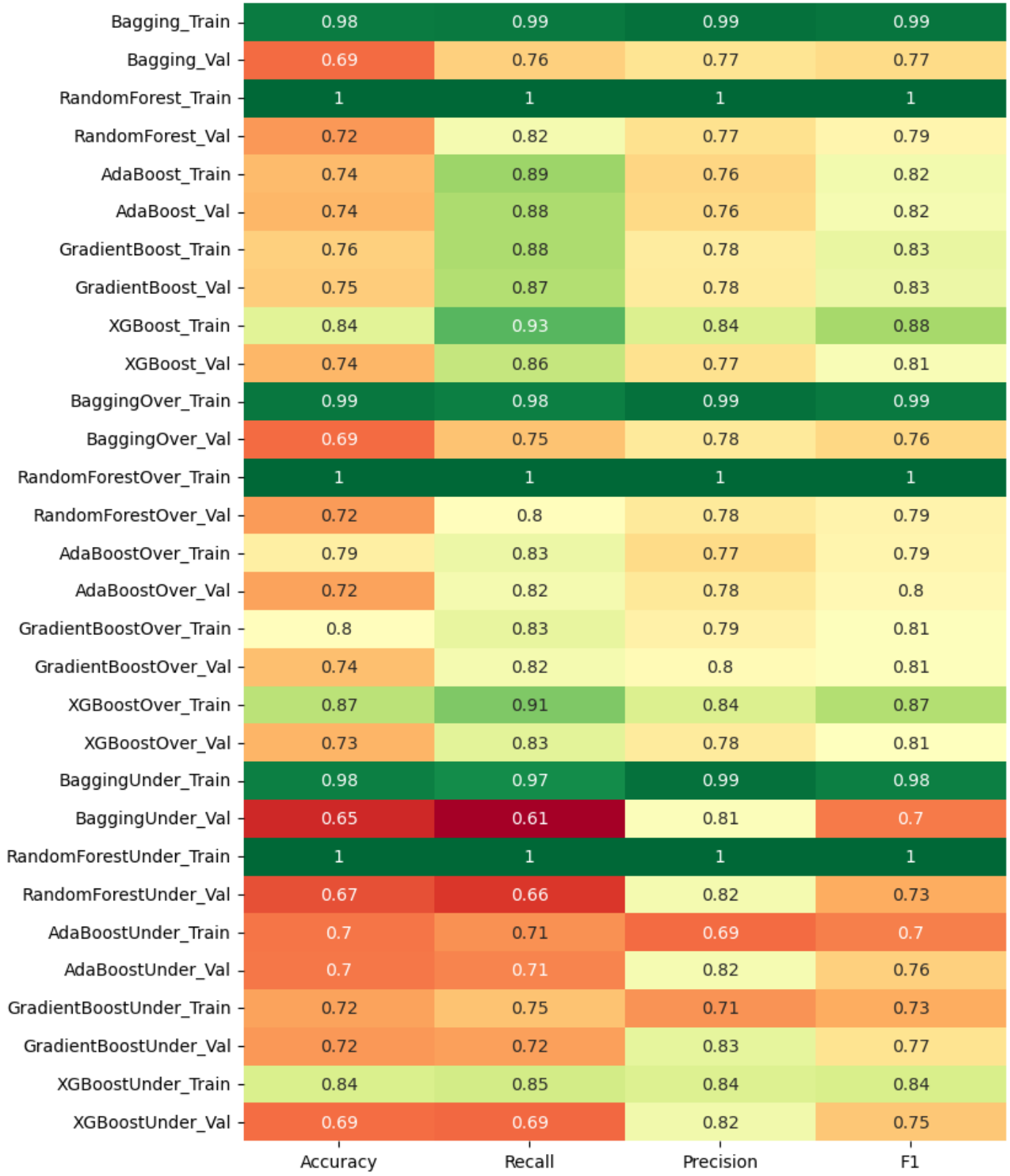

Figure 57 Accuracy, Recall, Precision, and F1 score for all models

6.1 Choose 3 models for tuning

- **Model 1**: Bagging and Random Forest models are similar. Among them, Random Forest is chosen because it samples features as well. This increases the level of independence between the estimators. In addition, random forests are robust and handle variances well.
- **Model 2**: AdaBoost is good for simple models, while Gradient Boost can handle complex models better. The chart also shows slightly better values for Gradient Boost. Gradient Boost will be chosen as the second model for hyperparameter tuning.
- **Model 3**: Extreme gradient boost has a wide range of parameters for tuning. It has also scored the best among the boosting models. It is an obvious choice for hyperparameter tuning.

6.2 Tune the models

The models will be tuned using the randomized search technique. The metric of interest will be f1 score.

6.2.1 Tuned Random Forest Model

- The random forest model will be tuned based on the following hyperparameters:
  - N_estimators: Number of trees in the forest.
  - Max_features: Number of features to consider.
  - Max_samples: The number of samples to train each estimator.
  - Min_samples_leaf: Minimum samples required at the leaf nodes.
- The best combination of values and the best score are shown in the figure.
- The top 5 features of importance for the tuned random forest are no_of_employees, yr_of_estab, education_level, wage and has_job_experience, as shown below.

```
{'n_estimators': 150, 'min_samples_leaf': 20, 'max_samples': 0.6, 'max_features': 0.2}
Best score: 0.8236121720839705
```

Figure 58 Best parameters for tuned Random Forest model

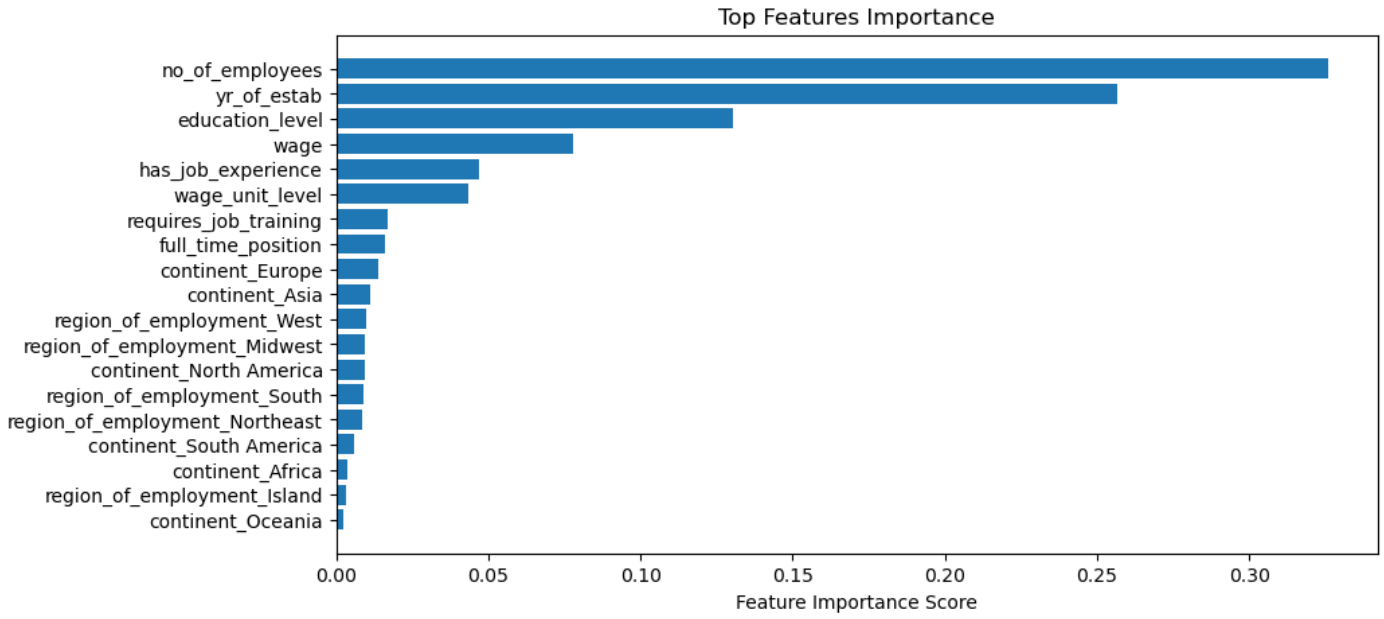

Figure 59 Most important features for tuned Random Forest model

6.2.2 Tuned Gradient Boost Model

- The Gradient Boost model will be tuned based on the following hyperparameters:
  - N_estimators: Number of boosting stages to perform.
  - Learning_rate: Shrinking for the contribution of each tree.
  - Max_depth: Maximum depth for the trees.

- Sub_sample: Fraction of samples to be used for fitting individual base learners.
- The best combination of values and the best score are shown in the figure.
- The top 5 features of importance for the tuned gradient boost model are education_level, no_of_employees, wage_unit_level, has_job_experience and yr_of_estab.

```
Best GB parameters: {'subsample': 0.6, 'n_estimators': 200, 'max_depth': 7, 'learning_rate': 0.01}
Best score: 0.8240724997259411
```

Figure 60 Best parameters for tuned Gradient Boost model

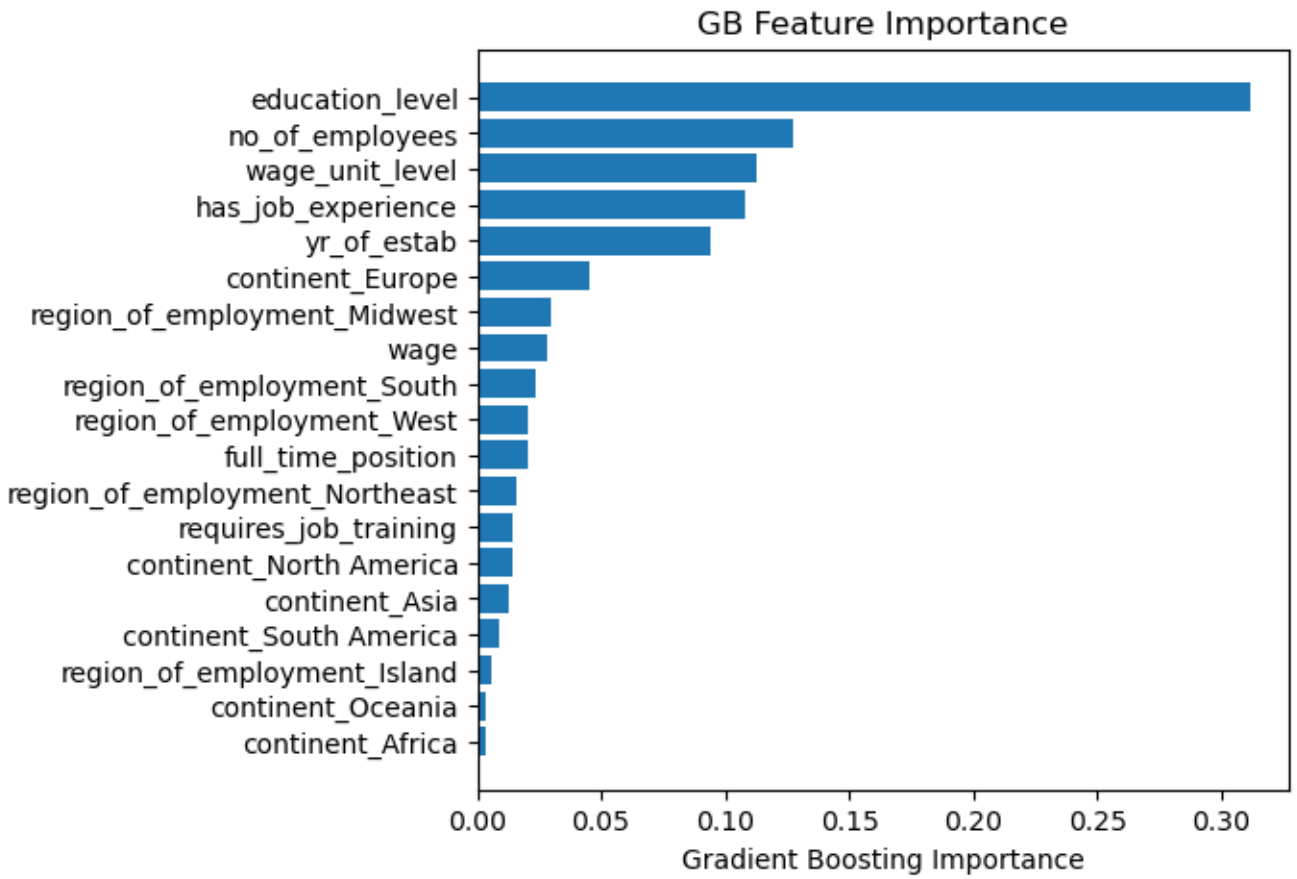

Figure 61 Most important features for tuned Gradient Boost model

6.2.3 Tuned Extreme Gradient Boost Model

- The Extreme Gradient Boost model will be tuned based on the following hyperparameters:
  - N_estimators: Number of boosting stages to perform.
  - Learning_rate: Shrinking for the contribution of each tree.
  - Max_depth: Maximum depth for the trees.
  - Sub_sample: Fraction of samples to be used for fitting individual base learners.
  - Gamma: This ensures that a node is split only when there is a positive reduction in the loss function.
  - colsample_bytree: Ratio of columns to be used when constructing the trees.
- The best combination of values and the best score are shown in the figure.
- The top 5 features of importance for the tuned gradient boost model are education_level, has_job_experience, wage_unit_level, continent_europe and region_of_employment_Midwest.

```
Best XGB parameters: {'subsample': 0.8, 'n_estimators': 300, 'max_depth': 7, 'learning_rate': 0.005, 'gamma': 0.2, 'colsample_bytree': 0.8}
Best score: 0.8245958196524696
```

Figure 62 Best parameters for tuned Extreme Gradient Boost model

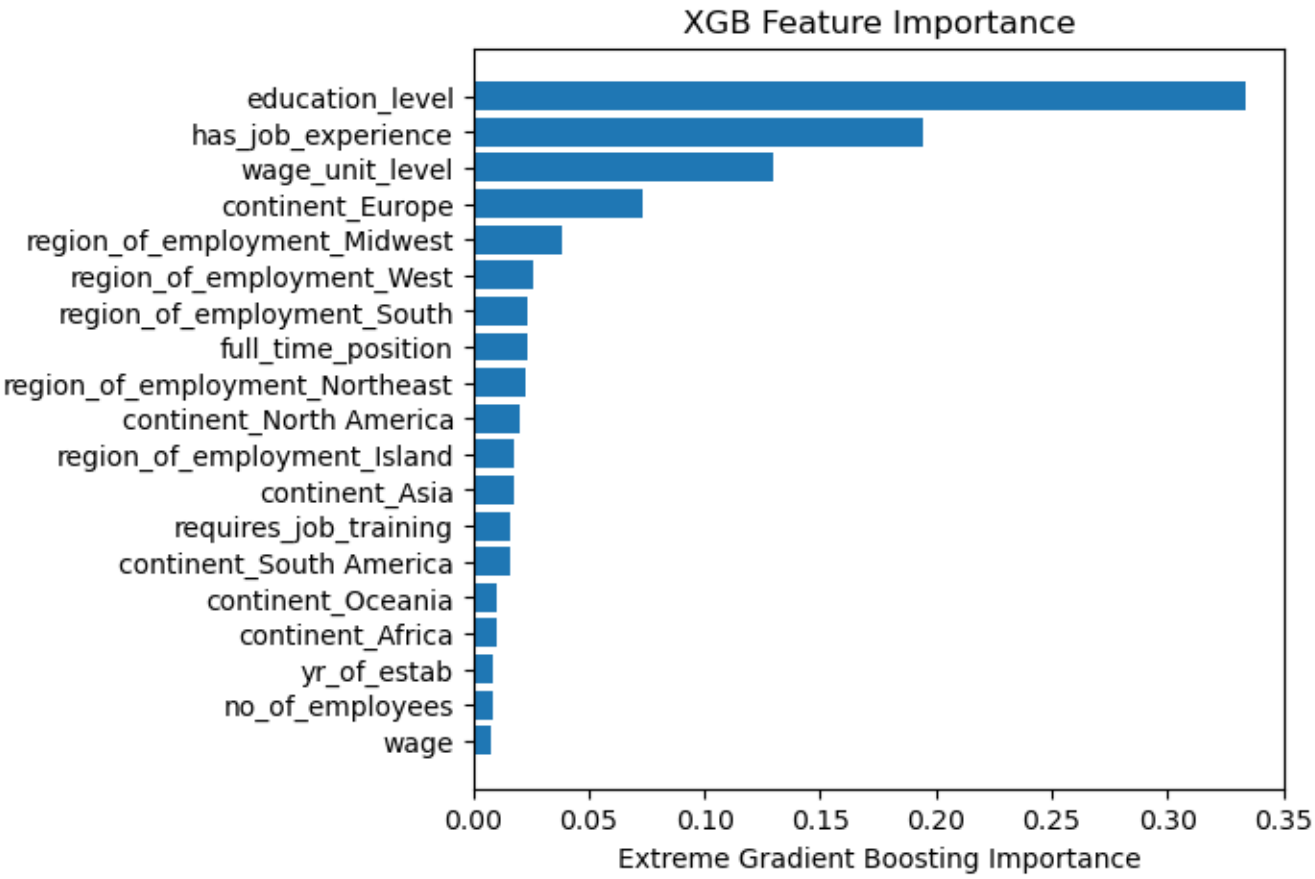

Figure 63 Most important features for tuned Extreme Gradient Boost model

6.3 Performance of the tuned models

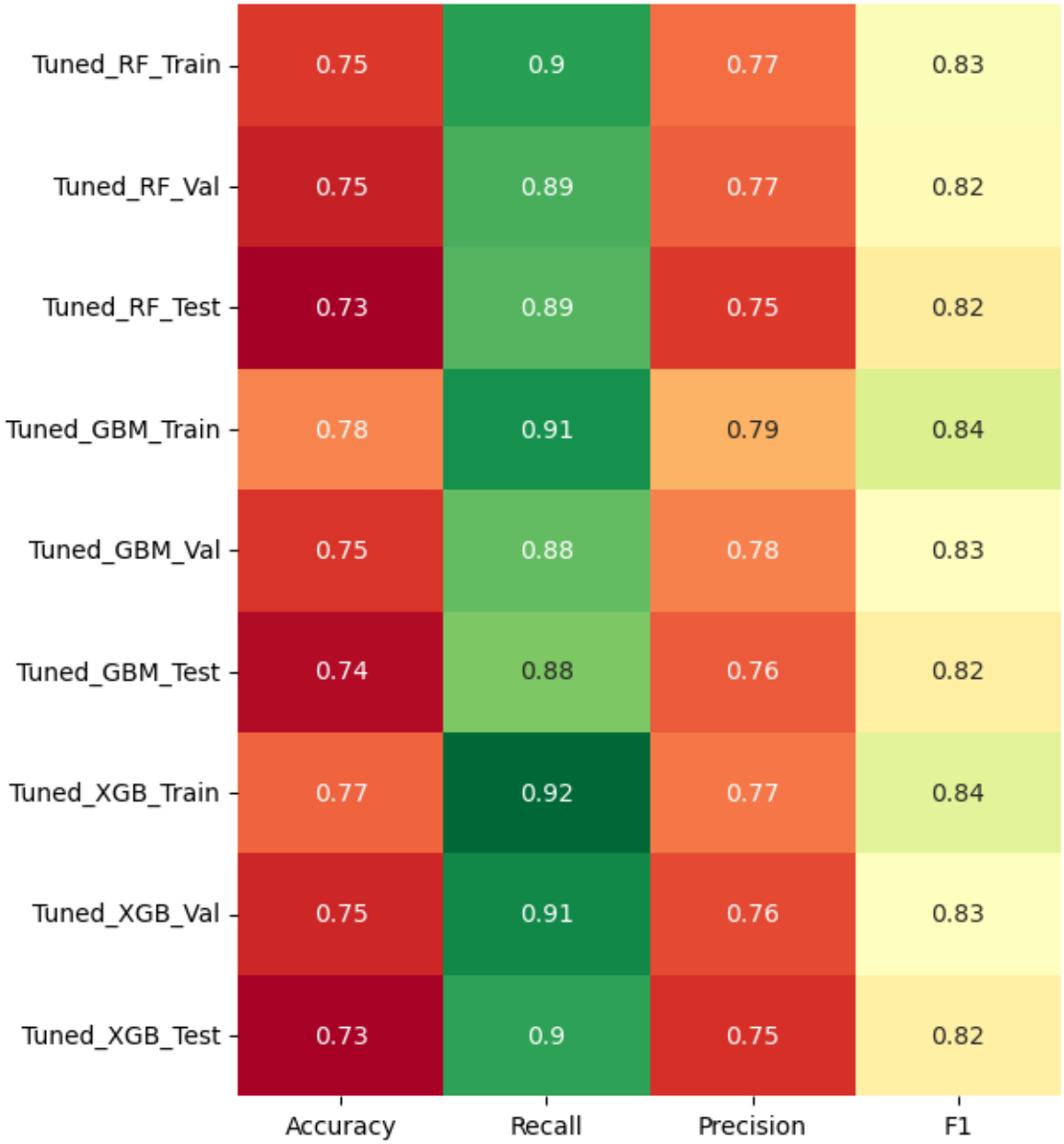

Figure 64 Accuracy, Recall, Precision, and F1 score for tuned models

- There is very little difference in scores across different performance metrics for different models.
- All models have performed best in recall.
- The models have scored low in accuracy and precision.

7 Model Performance Comparison and Final Model Selection

7.1 Compare the performance of tuned models

The metric that was decided for comparison was f1 score. Extreme Gradient Boost and Gradient Boost models have the same f1 score of 0.84. The recall for Extreme Gradient Boost model is slightly higher and this will be chosen as our final model.

7.2 Choose the best model

The chosen model is the Extreme Gradient Boost model.

7.3 Performance of the best model on the test set

The performance of the model on test data is slightly lower than that of the validation data. Since the difference is very small (0.01), the model is said to be a good fit.

8 Actionable Insights & Recommendations

8.1 Write down insights from the analysis conducted

- Models were built using 5 techniques – Bagging, random forest, adaptive boost, gradient boost, and extreme gradient boost.
- Apart from using the original dataset, oversampling and unsampling of the data was performed.
- Models built on the original data performed better.
- Random forest, gradient boost, and extreme gradient boost models were fine-tuned using hyperparameters.
- Extreme gradient boost model showed the best results and has been selected as the final model.

8.2 Provide actionable business recommendations

- The top 5 features of importance for the tuned random forest are no_of_employees, yr_of_estab, education_level, wage and has_job_experience, as shown below.
- The top 5 features of importance for the tuned gradient boost model are education_level, no_of_employees, wage_unit_level, has_job_experience and yr_of_estab.
- The top 5 features of importance for the tuned gradient boost model are education_level, has_job_experience, wage_unit_level, continent_europe and region_of_employment_Midwest.
- Education level has appeared in all 3 models. Employees with Master's and Doctorate degrees are more likely to get the Visa Certified, while employees who have only studied up to High school are likely to be Denied.
- Job experience has also appeared in all 3 models, indicating that Job experience helps in getting a visa.
- Wage is a feature in random forest, while wage_unit_level appears in gradient boost and extreme gradient boost. Higher salaries and higher periods of billing increase the chance of visa being Certified.
- The extreme gradient boost model has also shown interesting trends like 'employees from Europe' and 'employees working in the Midwest' have a higher probability of getting the visa certified.
- The model can be trained and tuned from time to time based on new data.