CODED PROJECT: MACHINE LEARNING – 2

EASY VISA

Data Science and Business Analytics

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.

Objective

In FY 2016, the OFLC processed 775,979 employer applications for 1,699,957 positions for temporary and permanent labour 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 analyse 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.

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.

Contents

List of Figures

1 Exploratory Data Analysis

1.1 Problem definition

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.

1.2 Data Overview

- Data has 25480 rows and 12 columns
- Data information shows that 9 string type columns and two numerical columns.

```
#   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
0   case_id                25480 non-null  object
1   continent              25480 non-null  object
2   education_of_employee  25480 non-null  object
3   has_job_experience     25480 non-null  object
4   requires_job_training  25480 non-null  object
5   no_of_employees        25480 non-null  int64
6   yr_of_estab            25480 non-null  int64
7   region_of_employment   25480 non-null  object
8   prevailing_wage        25480 non-null  float64
9   unit_of_wage           25480 non-null  object
10  full_time_position     25480 non-null  object
11  case_status            25480 non-null  object
dtypes: float64(1), int64(2), object(9)
```

Figure 1 Data information

- There are duplicate values checked with help of duplicated function
- There are no missing values found with help of isnull function.
- Statistical summary

	no_of_employees	yr_of_estab	prevailing_wage
count	25480.000	25480.000	25480.000
mean	5667.043	1979.410	74455.815
std	22877.929	42.367	52815.942
min	-26.000	1800.000	2.137
25%	1022.000	1976.000	34015.480
50%	2109.000	1997.000	70308.210
75%	3504.000	2005.000	107735.513
max	602069.000	2016.000	319210.270

Figure 2 Statistical Summary of numerical Variables

	case_id	continent	education_of_employee	has_job_experience	requires_job_training	region_of_employment	unit_of_wage	full_time_position	case_status
count	25480	25480	25480	25480	25480	25480	25480	25480	25480
unique	25480	6	4	2	2	5	4	2	2
top	EZYV25480	Asia	Bachelor's	Y	N	Northeast	Year	Y	Certified
freq	1	16861	10234	14802	22525	7195	22962	22773	17018

Figure 3 Statistical Summary of categorical variables

Observations

- No of employees : minimum is -26 which looks incorrect and maximum is very high which may be outlier or data inaccuracy.
- Year estd : min is1800 which goes up to 2016

- Prevailing wage = 2.137 is min which very low and max very high indicating outliers.
- continent = There are 6 continents highest is from Asia.
- Job experience : Most have job experience
- Job training - most require job training
- Employment region : dataset has 5 region and most from Northeast
- Wage unit : 4 unique values most is yearly wages
- Full time : Most of them are full time
- case status: Most get visa certified

1.3 Univariate analysis

1.3.1 Observation on No of employees

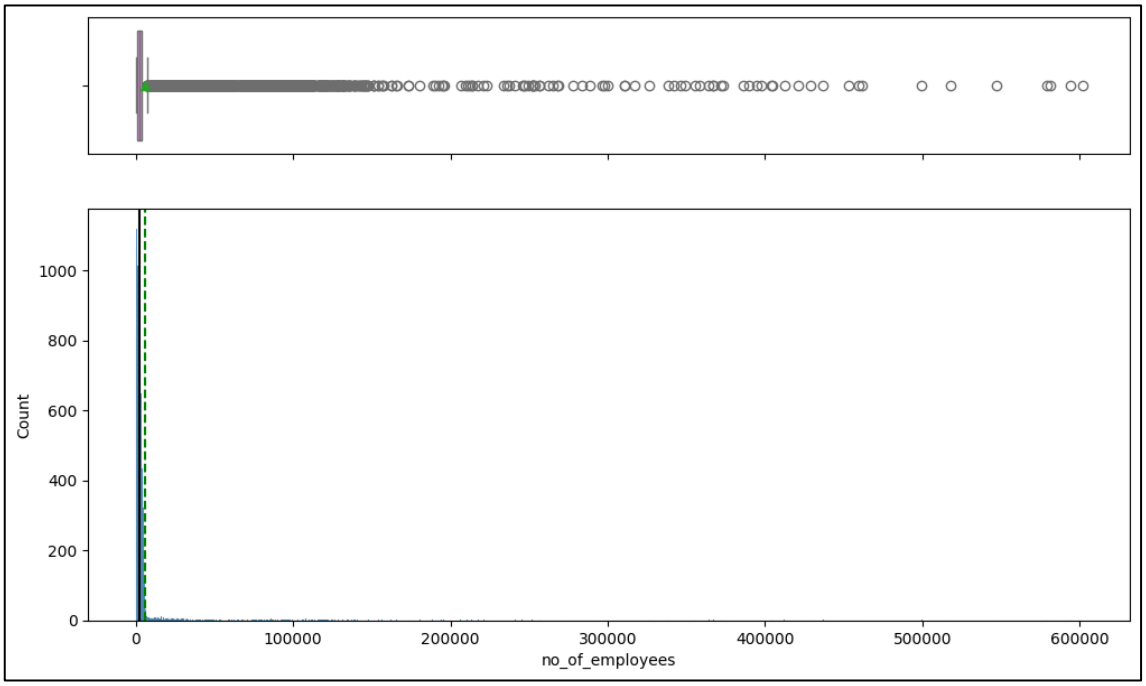

Figure 4 Distribution of count of employees

Observation

- Mean is close to zero and there are many outliers.
- Employees count has negative values in most cases which should be investigated.

1.3.2 Observation on year established

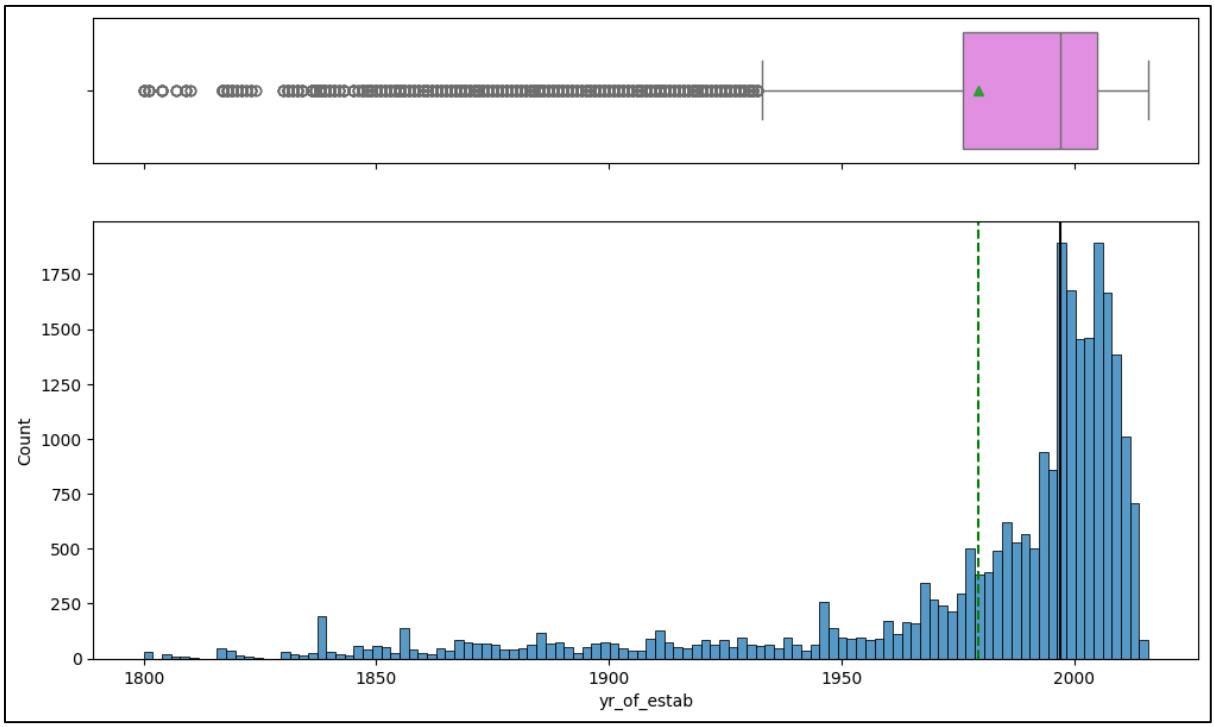

Figure 5 Year established

Observation

- Left skewed hence most companies established recently.
- Median year is close to 1997
- Many outliers in left side meaning there many older companies

1.3.3 Observation on prevailing wage

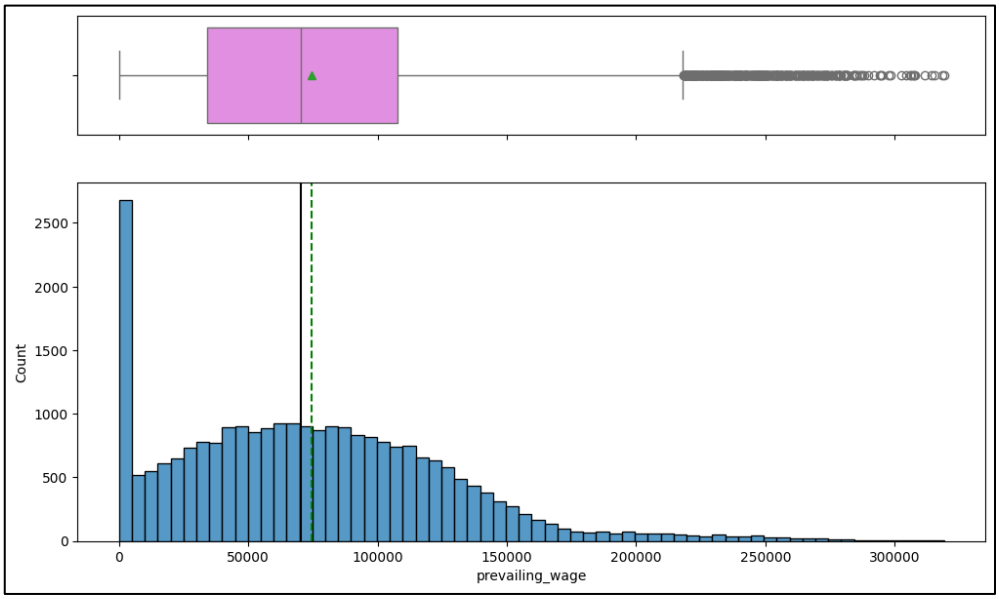

Figure 6 Distribution of prevailing wage

Observation

- Mean and median are almost closer hence normally distributed curve
- High count is observed in Zero values which needs to investigated.

1.3.4 Observation on continent

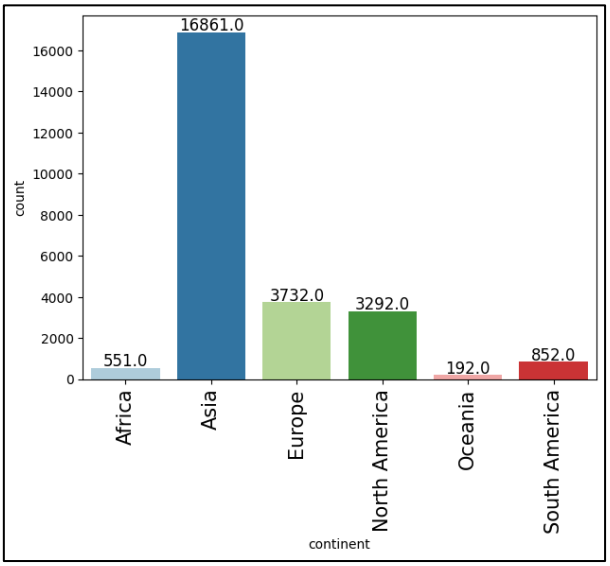

Figure 7 Distribution of continents

Observation

- Asia received highest Visa application followed by Europe.

1.3.5 Observation on Education of Employee

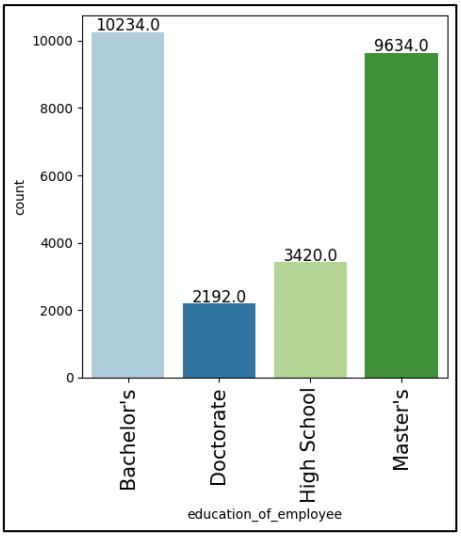

Figure 8 Distribution of Education of employee

Observation

- Employees with Bachelor’s degree is highest followed Master’s degree

1.3.6 Observation on Job Experience

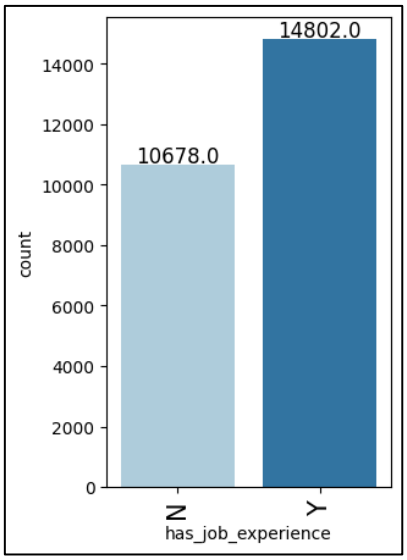

Figure 9 Distribution of Job Experience

Observation

- Employees with Job experience are highest hence skilled workforce mostly applies.

1.3.7 Observation on Region of Employment

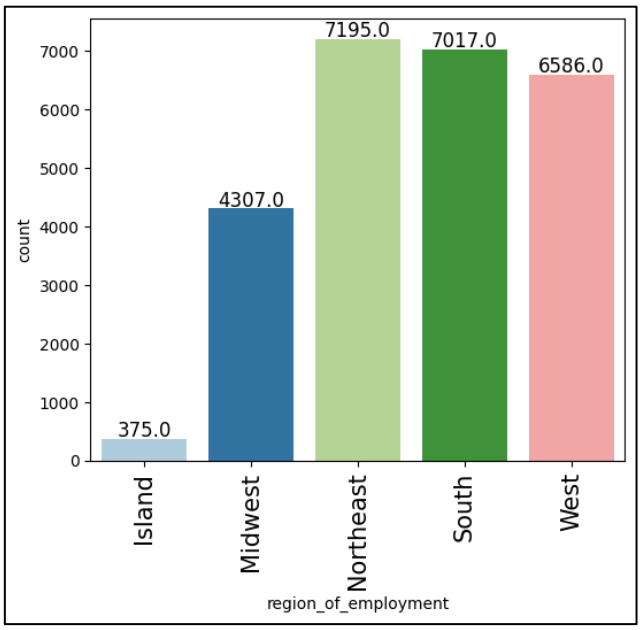

Figure 10 Region of employment

Observation

- Northeast has highest Visa application followed by South and west indicating high demand for labour.

1.3.8 Observation on Unit of wage

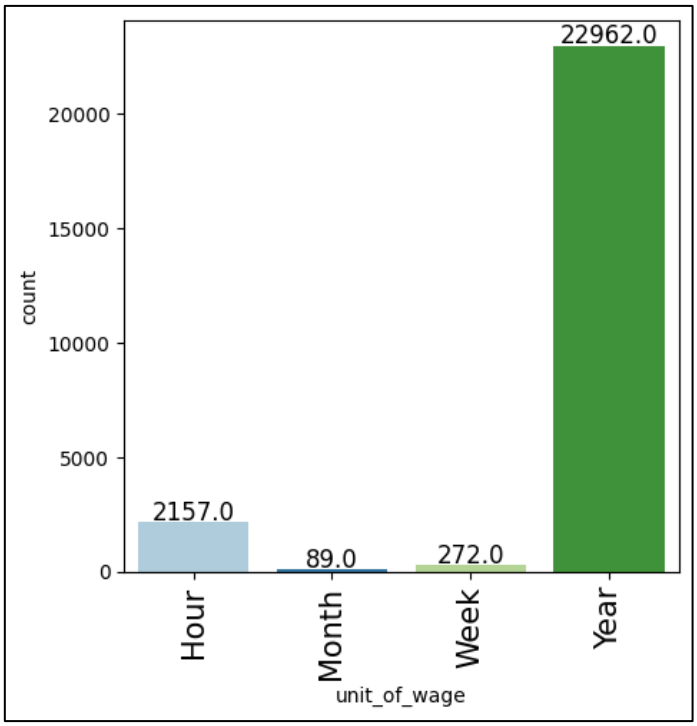

Figure 11 Distribution wage unit

Observation

- unit of wage is mostly in years and followed by Hours meaning organised sector has highest visa application and labour demand followed by unskilled workers mostly paid in hours.

1.3.9 Observation on full time position

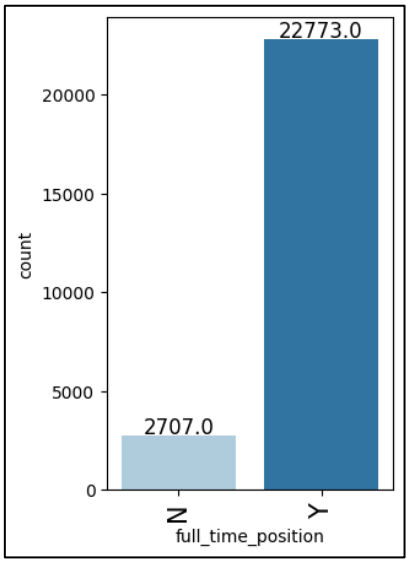

Figure 12 Distribution of full time position

Observation

- Most application are for full time position.

1.3.10 Observation on case status

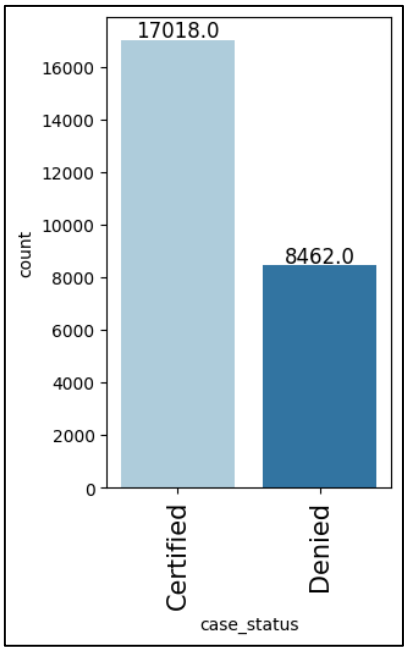

Figure 13 Distribution of case status

Observation

- Most employees get Visa certified and 8462 got denied.

1.4 Bivariate analysis

1.4.1 Heat Map

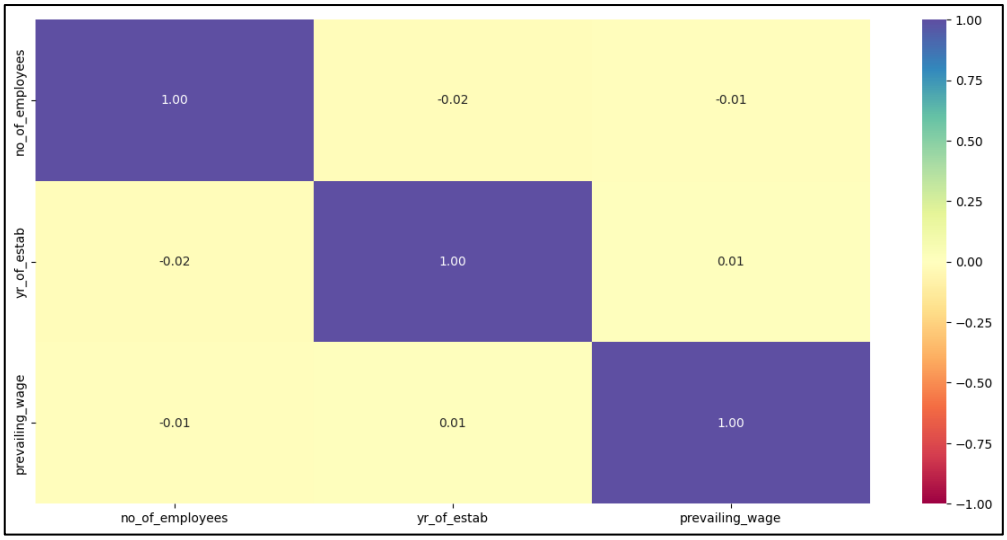

Figure 14 Heat Map for Numerical variables

Observation

- No Meaningful correlations exist between variables

1.4.2 Continent Vs Case status

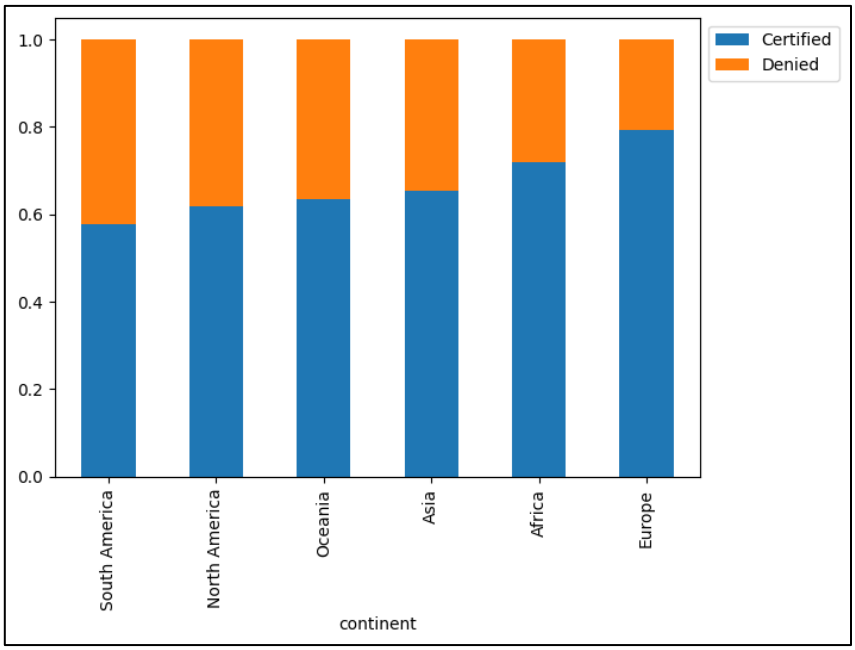

Figure 15 Continent Vs Case status

Observations

- South America has highest denial followed by North America and Europe has highest approval rate.

1.4.3 Education Vs Case status

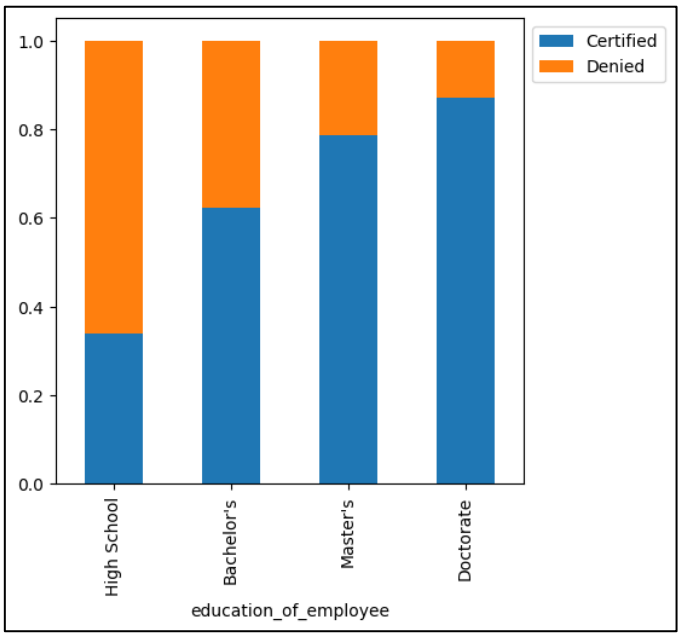

Figure 16 Education Vs Case status

Observation

- High School education is mostly denied indicating low skill labour poor demand. Higher the education higher chance of getting visa approved.

1.4.4 Job experience Vs Case status

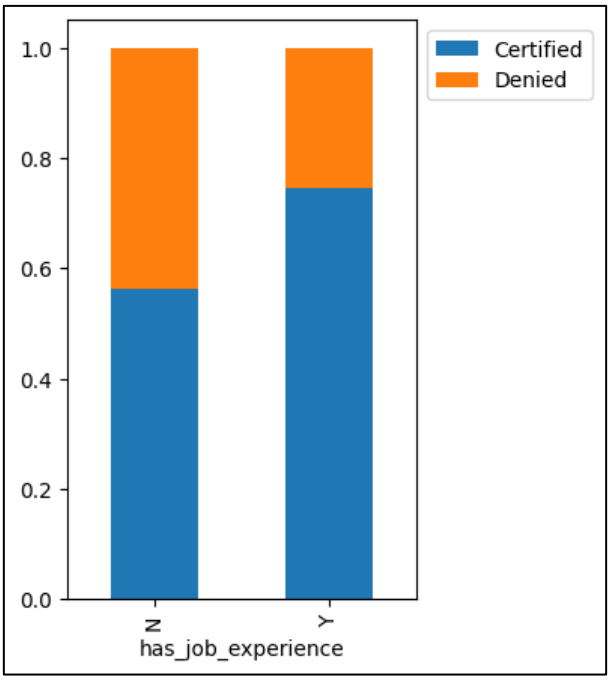

Figure 17 Job experience Vs Case status

Observation

- Employees with no job experience has lower Visa approval

1.4.5 Job training Vs Case status

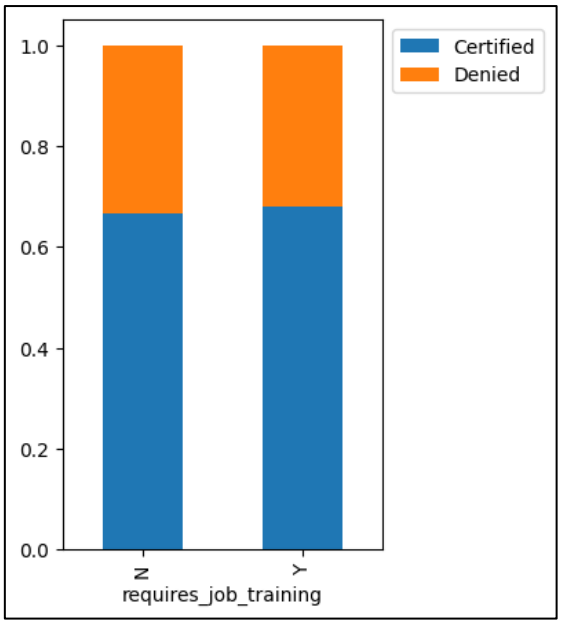

Figure 18 Job Training Vs Case status

Observation

- Job training has less effect on Visa approval.

1.4.6 Region of Employment Vs Case status

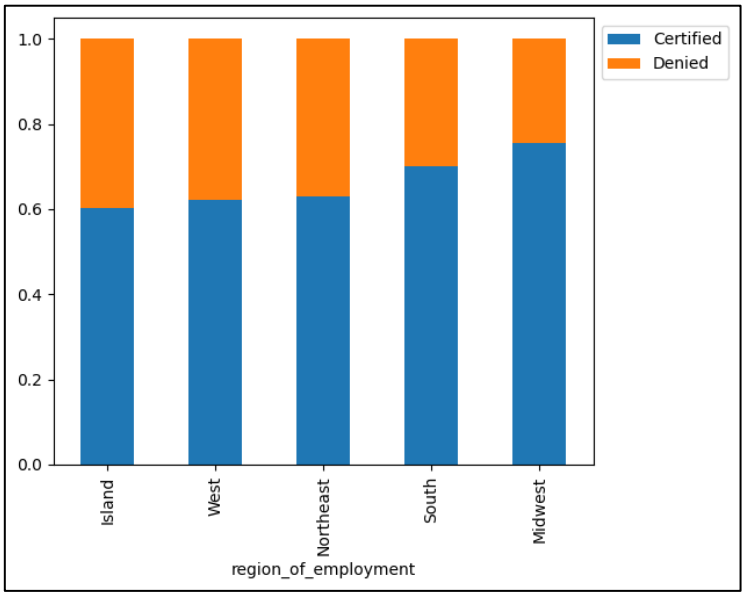

Figure 19 Region of employment Vs case status

Observation

- Employees to Midwest and South has highest visa approval rate and Island has lowest.

1.4.7 Unit of wage Vs Case Status

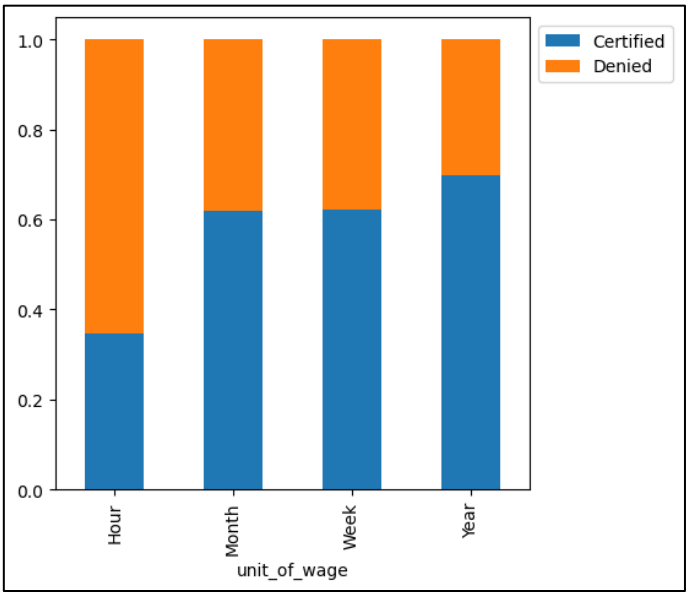

Figure 20 Unit of wage Vs case status

Observation

- Hourly wages are denied the most may be less skilled worker are given hourly wages and Yearly wages has highest visa approval indicating skilled workers.

1.4.8 Full time position Vs Case Status

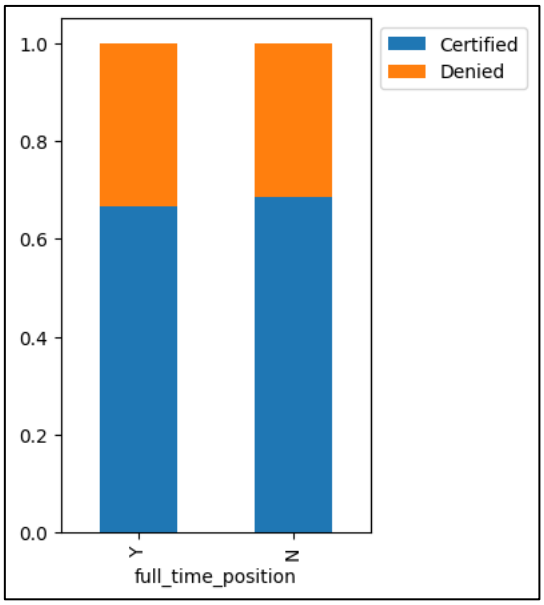

Figure 21 Full time position Vs case status

Observation

Full time position and part time position has equal denial and approval rate.

1.4.9 No. of employees Vs Case status

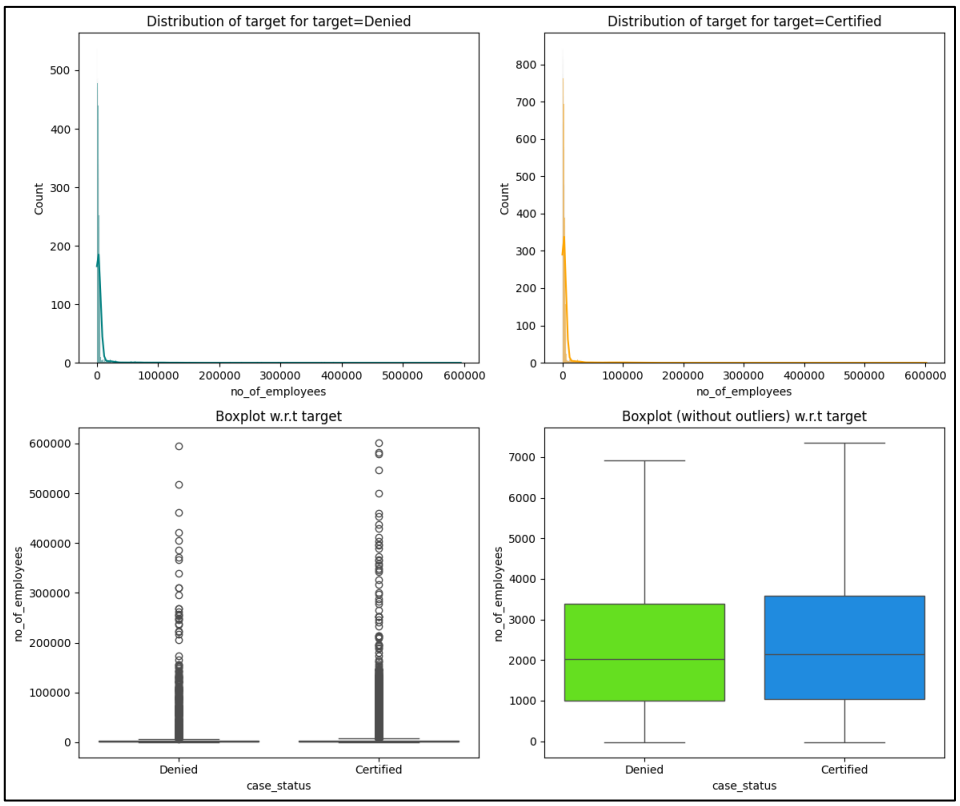

Figure 22 No of employees and case status

Observation

- The dataset contains extreme outliers in the number of employees.
- The majority of companies applying for visas have small or medium-sized workforces.
- There is no significant difference in company size between certified and denied visa cases.

1.4.10 Prevailing Wage Vs case status

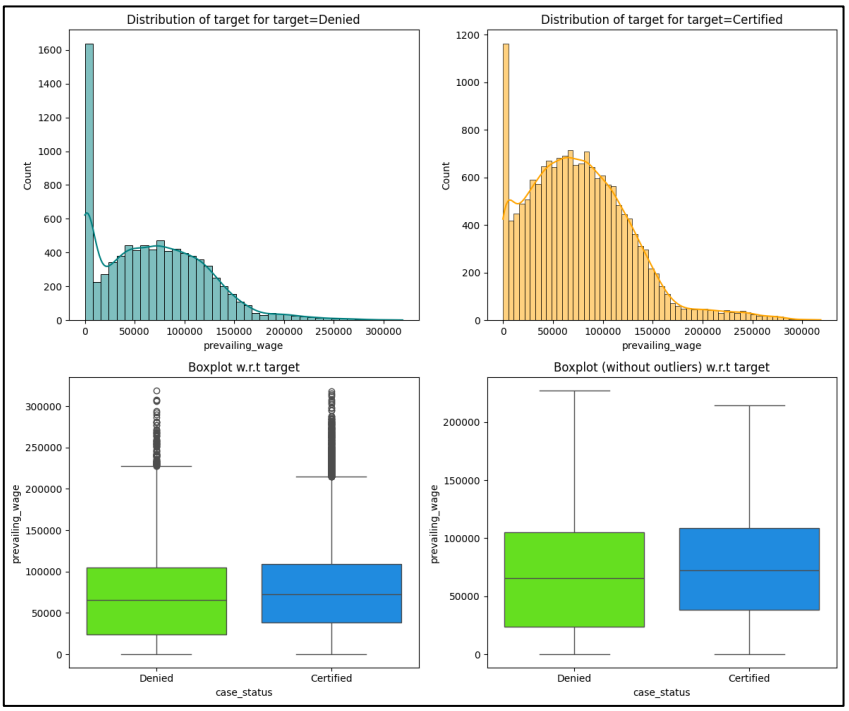

Figure 23 Prevailing wage Vs case status

Observation

- Median wage of certified is higher than denied implying lower wages are denied visa compared to higher wages.

1.4.11 Year of established Vs case status

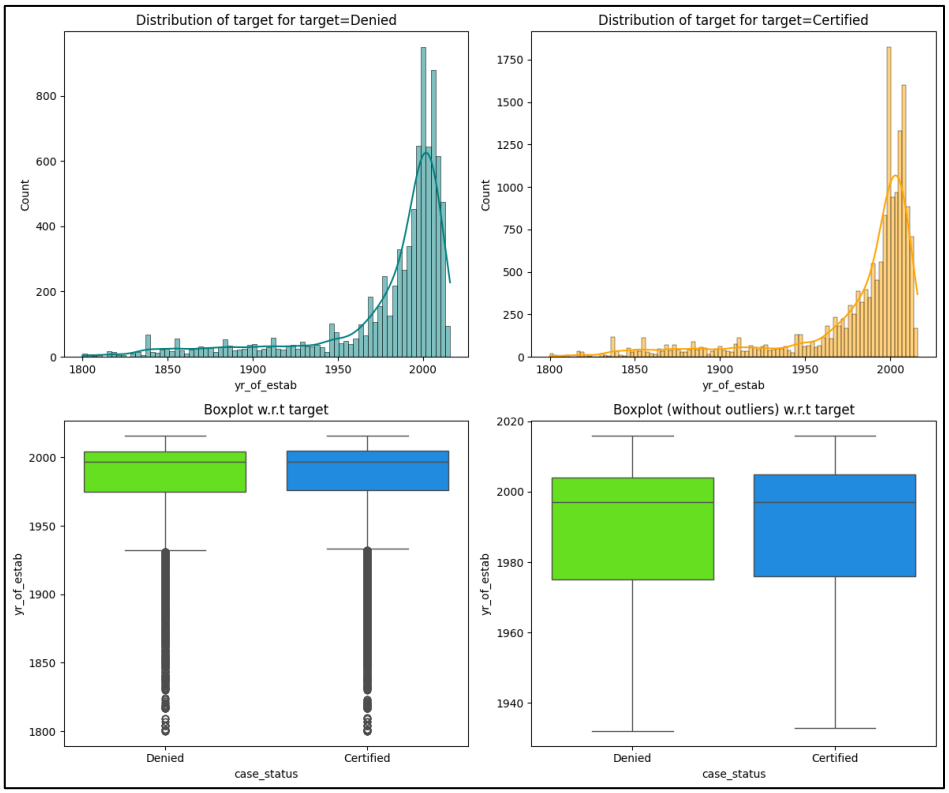

Figure 24 year established Vs Case status

Observation

- Year established has no effect on Visa approval.

1.5 Key meaningful observations on individual variables and the relationship between variables

1.5.1 Key meaningful observations on individual variables

1. **Number of Employees**
    - The presence of negative values suggests data inaccuracies that need investigation.
    - Many outliers indicate potential reporting issues or extreme variations in company sizes.
2. **Year Established**
    - The data is left-skewed, meaning most companies were established recently.
    - The median establishment year is around 1997.
    - The presence of older companies (outliers) suggests a mix of legacy and new firms.
3. **Prevailing Wage**
    - The distribution appears normal, with mean and median being close.

- A high number of zero values suggests potential data entry issues or unpaid internships.

4. **Continent**
    - Asia has the highest number of visa applications, followed by Europe.
    - This indicates a significant demand for skilled labour from Asian countries.
5. **Education Level**
    - Most applicants hold a Bachelor's degree, followed by a Master's degree.
    - This suggests that skilled and highly educated individuals are applying for visas.
6. **Job Experience**
    - The majority of applicants have job experience, indicating a demand for skilled professionals.
7. **Employment Region**
    - The Northeast region has the highest number of visa applications, followed by the South and West.
    - This aligns with economic activity and job concentration in these regions.
8. **Wage Unit**
    - Most applicants are paid yearly, indicating a dominance of professional and salaried roles.
    - Hourly wages are also significant, reflecting demand for hourly or contract-based workers.
9. **Full-Time Position**
    - Most applications are for full-time positions, reinforcing the trend of professional employment.
10. **Case Status**
    - Most visa applications get certified.
    - A significant number (8,462) were denied, which could be analysed further to understand rejection patterns.

1.5.2 Key meaningful observations on relationship between variables

- Higher education and job experience significantly improve visa approval chances.
- Higher wages correlate with higher approval rates, while lower wages face more denials.
- Continent and region of employment play a role, with Europe and the Midwest/South regions showing higher approval rates.
- Job training and company establishment year have little impact on visa decisions.

2 Data Preprocessing

2.1 Prepare the data for analysis

We shall copy the data into new data frame called df to avoid altering the original data.

2.2 Feature Engineering

From EDA, we noted employee count has negative values. Upon reviewing it, these are data entry errors as negative sign is added inadvertently to few cases. Hence, let us replace those negative values with positives and check if no of employees has any negative values.

We shall replace denied for 0 and certified for 1 for case status and also noted that case status has imbalanced dataset.

Next , we shall split the dataset per below

- Train data with 16307 Rows, 10 columns – 80% of temp data
- Validation data 4077 Rows, 10 columns – 20% of temp data
- Test data with 5096 Rows, 10 Columns – 20% of whole data

Further, we shall create dummy variables for all train , validation and test data having categorical variables.

- Train data with 16307 Rows, 21 columns
- Validation data 4077 Rows, 21 columns
- Test data with 5096 Rows, 21 Columns

2.3 Missing value Treatment

We do not have any missing values hence no requirement for treatment.

2.4 Outlier Treatment

We are not treating outliers as they are original values.

2.5 Ensure no data leakage among train-test and validation sets

We ensured there is no data leakage as missing value treatment is no required and split the data first and created dummy variables to ensure no data leakage.

3 Model Building Original Data

3.1 Choose the appropriate metric for model evaluation

Model can make wrong predictions as:

- Predicting an employee will be denied Visa and the he gets it
- Predicting an employee will be certified Visa and the he is denied

Which case is more important?

Predicting an employee will be certified Visa and he is denied. As business are in high demand of human resources it is important to certify all the employees correctly and denying Visa to deserving employees will negatively affect the business and economy.

How to reduce this loss i.e need to reduce False Negatives?

We want Recall to be maximized, greater the Recall higher the chances of minimizing false negatives. Hence, the focus should be on increasing Recall or minimizing the false negatives or in other words identifying the true positives(i.e. Class 1) so that Visa is provided to correct employees.

3.2 Build 5 models

We shall build below six models

- Bagging Classifier
- Random Forest Classifier
- Gradient Boosting Classifier
- AdaBoost Classifier
- XGB Classifier
- Decision Tree Classifier

We shall first check cross validation scores of each model with train data and test it on validation data

```
Cross-Validation Performance:

Bagging: 78.22967882336233
Random forest: 84.24377278563936
GBM: 87.34720057178234
Adaboost: 88.95412888111791
Xgboost: 85.2813630063411
dtree: 74.56621789676986

validation Performance:

Bagging: 0.7840616966580977
Random forest: 0.8431876606683805
GBM: 0.8626514873301506
Adaboost: 0.884318766066838
Xgboost: 0.8464928387807565
dtree: 0.7495409474843923
```

Figure 25 Cross Validation score

Training and Validation Performance Difference for Original date:

	Model	Training Score	Validation Score	Difference
0	Bagging	0.986	0.784	0.202
1	Random forest	1.000	0.843	0.157
2	GBM	0.876	0.863	0.013
3	Adaboost	0.890	0.884	0.005
4	Xgboost	0.937	0.846	0.091
5	dtree	1.000	0.750	0.250

Figure 26 Training and Validation Performance Difference for Original data

3.3 Comment on the model performance

Bagging: Cross validation score shows moderate performance, and recall score is Overfitting when compared to train and test data.

Random forest: Cross validation score shows good performance, and recall score Overfits when compared to train and test data.

Gradient Boosting Classifier: Cross validation score shows good performance, and recall score **have moderate differences 0.013** indicating a reasonable balance.

AdaBoost Classifier: Cross validation score shows excellent performance, and recall score **have less differences 0.005 i**ndicating it's generalizing very well

XGB Classifier: Cross validation score shows good performance, and recall score **have moderate differences 0.091** indicating a reasonable balance.

Decision Tree Classifier: Cross validation score shows poor performance, and recall score **is overfitting.**

4 Model Building - Oversampled Data-

4.1 Oversample the train data

We shall use SMOTE to perform oversampling of the data with KNN as 5 to reduce the effect the class imbalance

Below is result of oversampled data

- Before Oversampling, counts of label 'Yes': 10891
- Before Oversampling, counts of label 'No': 5416
- After Oversampling, counts of label 'Yes': 10891
- After Oversampling, counts of label 'No': 10891
- After Oversampling, the shape of train_X: (21782, 21)
- After Oversampling, the shape of train: (21782,)

4.2 Build 5 models

We shall build below six models

- Bagging Classifier
- Random Forest Classifier
- Gradient Boosting Classifier
- AdaBoost Classifier
- XGB Classifier
- Decision Tree Classifier

We shall first check cross validation scores of each model with train data and test it on validation data

```
Cross-Validation Performance:

Bagging: 74.66721956938486
Random forest: 81.0760784869008
GBM: 84.9141378320735
Adaboost: 85.53846277030391
Xgboost: 83.81240331050502
dtree: 71.69227002386499

Validation Performance:

Bagging: 0.7488064634594197
Random forest: 0.8116048475945649
GBM: 0.8329048843187661
Adaboost: 0.8479618068307014
Xgboost: 0.8395152405435182
dtree: 0.7168564083731179
```

Figure 27 Cross Validation score Oversampled data

Training and Validation Performance Difference for Oversampled:

	Model	Training Score	Validation Score	Difference
0	Bagging	0.985	0.749	0.236
1	Random forest	1.000	0.812	0.188
2	GBM	0.846	0.833	0.013
3	Adaboost	0.859	0.848	0.011
4	Xgboost	0.914	0.840	0.075
5	dtree	1.000	0.717	0.283

Figure 28 Training and Validation difference for recall score of oversampled data

4.3 Comment on the model performance

Bagging: Cross validation score shows moderate performance, and recall score is Overfitting when compared to train and test data.

Random forest: Cross validation score shows good performance, and recall score Overfits when compared to train and test data.

Gradient Boosting Classifier: Cross validation score shows good performance, and recall score have moderate differences 0.013 indicating it's generalizing very well

AdaBoost Classifier: Cross validation score shows excellent performance, and recall score have less differences 0.011 indicating it's generalizing very well

XGB Classifier: Cross validation score shows good performance, and recall score have moderate differences 0.075 indicating a reasonable balance.

Decision Tree Classifier: Cross validation score shows poor performance, and recall score is overfitting.

5 Model Building – Under sampled Data

5.1 Under sample the train data

With help RandomUnderSampler, we shall under sample the data

- Before Under Sampling, counts of label 'Yes': 10891
- Before Under Sampling, counts of label 'No': 5416
- After Under Sampling, counts of label 'Yes': 5416
- After Under Sampling, counts of label 'No': 5416
- After Under Sampling, the shape of train_X: (10832, 21)
- After Under Sampling, the shape of train: (10832,)

5.2 Build 5 models

We shall build below six models

- Bagging Classifier
- Random Forest Classifier
- Gradient Boosting Classifier
- AdaBoost Classifier
- XGB Classifier
- Decision Tree Classifier

We shall first check cross validation scores and of each model with train data and test it on validation data and get recall for comparing train and validation data.

```
Cross-Validation Performance:

Bagging: 60.0078025711005
Random forest: 66.59944189469596
GBM: 71.84283781895992
Adaboost: 70.25445240601991
Xgboost: 68.05772198996229
dtree: 63.49739176062121

Validation Performance:

Bagging: 0.6107234667645979
Random forest: 0.6709511568123393
GBM: 0.7161219243481455
Adaboost: 0.689680499449137
Xgboost: 0.6845391112743298
dtree: 0.6228424531766434
```

Figure 29 Cross validation under sampled data

Training and Validation Performance Difference undersampled data

	Model	Training Score	Validation Score	Difference
0	Bagging	0.970	0.611	0.359
1	Random forest	1.000	0.671	0.329
2	GBM	0.730	0.716	0.014
3	Adaboost	0.697	0.690	0.008
4	Xgboost	0.873	0.685	0.189
5	dtree	1.000	0.623	0.377

Figure 30 Recall score Train and validation for under sampled data

5.3 Comment on the model performance

Bagging: Cross validation score shows poor performance, and recall score is Overfitting when compared to train and test data.

Random forest: Cross validation score shows poor performance, and recall score Overfits when compared to train and test data.

Gradient Boosting Classifier: Cross validation score shows good performance, and recall score have moderate differences 0.014 indicating it's generalizing very well

AdaBoost Classifier: Cross validation score shows good performance, and recall score have less differences 0.008 indicating it's generalizing very well

XGB Classifier: Cross validation score shows good performance, and recall score have moderate differences 0.189 indicating a reasonable balance.

Decision Tree Classifier: Cross validation score shows poor performance, and recall score is overfitting.

6 Model Performance Improvement using Hyperparameter Tuning

6.1 Choose 3 models

Based on recall score, we shall choose GBM , Adaboost and Xgboost to tune further to improve the results.

6.2 Tune the 3 models and comment on performance

6.2.1 Tuning AdaBoost using original data

With help of randomized_cv, below are best parameters for model

- n_estimators= 30,
- learning_rate= 0.1,
- estimator= DecisionTreeClassifier(max_depth=1, random_state=1))

	Accuracy	Recall	Precision	F1
0	0.692	0.972	0.692	0.808

Figure 31 Train with original data

	Accuracy	Recall	Precision	F1
0	0.693	0.971	0.693	0.809

Figure 32 Validation with original data

Observation

- Recall : Good performance and generalised well without overfitting
- Precision: Moderate performance with good generalisation
- F1 : Good performance with excellent generalisation

6.2.2 Tuning GBM using original data

With help of randomized_cv, below are best parameters for model

- max_features=1,
- init=AdaBoostClassifier(random_state=1),
- random_state=1,
- learning_rate=0.01,
- n_estimators=100,
- subsample=0.9.

	Accuracy	Recall	Precision	F1
0	0.719	0.940	0.723	0.817

Figure 33 Train performance GBM original data

	Accuracy	Recall	Precision	F1
0	0.720	0.939	0.724	0.818

Figure 34 Validation performance GBM with Original data

Observation

- Recall : Good performance and generalised well without overfitting
- Precision: Moderate performance with good generalisation
- F1 : Good performance with excellent generalisation

6.2.3 Tuning XG boost using original data

With help of randomized_cv, below are best parameters for model

random_state=1,

eval_metric="logloss",

subsample=0.7,

scale_pos_weight=5,

n_estimators=np.int64(50),

learning_rate=0.01,

gamma=3.

	Accuracy	Recall	Precision	F1
0	0.668	1.000	0.668	0.801

Figure 35 Train performance XGboost with original data

	Accuracy	Recall	Precision	F1
0	0.668	1.000	0.668	0.801

Figure 36 Validation performance XGboost with original data

Observation

- Recall : Model has generalised well without overfitting
- Precision: Moderate performance with good generalisation
- F1 : Good performance with good generalisation

6.2.4 Tuning AdaBoost using Oversampled data

Model performance is obtained by fitting with best parameter from randomised CV.

	Accuracy	Recall	Precision	F1
0	0.712	0.934	0.719	0.813

Figure 37 Train performance Adaboost oversampled

	Accuracy	Recall	Precision	F1
0	0.710	0.928	0.719	0.811

Figure 38 Validation performance Adaboost oversampled

Observation

- Recall : Model has generalised well without overfitting
- Precision: Moderate performance with good generalisation
- F1 : Good performance with good generalisation

6.2.5 Tuning GBM using Oversampled data

Model performance is obtained by fitting with best parameter from randomised CV.

	Accuracy	Recall	Precision	F1
0	0.736	0.884	0.760	0.818

Figure 39 Train performance with GBM Oversampled

	Accuracy	Recall	Precision	F1
0	0.737	0.876	0.765	0.816

Figure 40 Validation performance GBM oversampled

Observation

- Recall : Model has generalised well without overfitting
- Precision: Moderate performance with good generalisation
- F1 : Good performance with good generalisation

6.2.6 Tuning XGBoost using Oversampled data

Model performance is obtained by fitting with best parameter from randomised CV.

	Accuracy	Recall	Precision	F1
0	0.668	1.000	0.668	0.801

Figure 41 Train performance with XGBoost Oversampled

	Accuracy	Recall	Precision	F1
0	0.668	1.000	0.668	0.801

Figure 42 Validation performance with XGBoost Oversampled

Observation

- Recall : Model has generalised well without overfitting

- Precision: Moderate performance with good generalisation
- F1 : Good performance with good generalisation

6.2.7 Tuning AdaBoost using Under sampled data

Model performance is obtained by fitting with best parameter from randomised CV.

	Accuracy	Recall	Precision	F1
0	0.599	0.933	0.559	0.700

Figure 43 Train Performance Adaboost under sampled

	Accuracy	Recall	Precision	F1
0	0.710	0.928	0.719	0.811

Figure 44 Validation Performance Adaboost under sampled

Observation

- Recall : Model has generalised well without overfitting
- Precision: Poor performance with underfitting
- F1 : Good performance with underfitting

6.2.8 Tuning GBM using Under sampled data

Model performance is obtained by fitting with best parameter from randomised CV.

	Accuracy	Recall	Precision	F1
0	0.705	0.747	0.689	0.717

Figure 45 Train Performance GBM under sampled

	Accuracy	Recall	Precision	F1
0	0.720	0.939	0.724	0.818

Figure 46 Validation Performance GBM under sampled

Observation

- Recall : Moderate performance with underfitting
- Precision: Poor performance with underfitting
- F1 : Poor performance with underfitting

6.2.9 Tuning XGBoost using Under sampled data

Model performance is obtained by fitting with best parameter from randomised CV.

	Accuracy	Recall	Precision	F1
0	0.500	1.000	0.500	0.667

Figure 47 Train Performance XGBoost under sampled

	Accuracy	Recall	Precision	F1
0	0.668	1.000	0.668	0.801

Figure 48 Validation Performance XGBoost under sampled

Observation

- Recall : Good performance with generalisation
- Precision: Poor performance with underfitting
- F1 : Poor performance with underfitting

7 Model Performance Comparison and Final Model Selection

7.1 Compare the performance of tuned models

We shall compare training and validation performance of models

Training Performance

Training performance comparison:

	Adaboost trained with Original data	Gradient boosting trained with Original data	Xg boosting trained with Original dataa	Adaboost trained with oversample data	Gradient boosting trained with oversample data
Accuracy	0.692	0.719	0.668	0.712	0.736
Recall	0.972	0.940	1.000	0.934	0.884
Precision	0.692	0.723	0.668	0.719	0.760
F1	0.808	0.817	0.801	0.813	0.818

Xg boosting trained with oversample dataa	Adaboost trained with undersample data	Gradient boosting trained with undersample data	Xg boosting trained with undersample data
0.668	0.599	0.705	0.500
1.000	0.933	0.747	1.000
0.668	0.559	0.689	0.500
0.801	0.700	0.717	0.667

Figure 49 Training performance of tuned models

Validation performance comparison

Validation performance comparison:

	Adaboost trained with Original data	Gradient boosting trained with Original data	Xg boosting trained with Original dataa	Adaboost trained with oversample data	Gradient boosting trained with oversample data
Accuracy	0.693	0.720	0.668	0.710	0.737
Recall	0.971	0.939	1.000	0.928	0.876
Precision	0.693	0.724	0.668	0.719	0.765
F1	0.809	0.818	0.801	0.811	0.816

Xg boosting trained with oversample dataa	Adaboost trained with undersample data	Gradient boosting trained with undersample data	Xg boosting trained with undersample dataa
0.668	0.710	0.720	0.668
1.000	0.928	0.939	1.000
0.668	0.719	0.724	0.668
0.801	0.811	0.818	0.801

Figure 50 Validation Performance of tuned models

7.2 Choose the best model

Gradient Boosting with Original Data:

- This model has the best overall F1 scores on both training and validation.
- It shows minimal discrepancy between training and validation (excellent generalization).
- It provides good balance between Precision and Recall.

7.3 Comment on the performance of the best model on the test set

	Accuracy	Recall	Precision	F1
0	0.719	0.930	0.726	0.816

Figure 51 Test performance GBM Model

Observation:

The test data performance confirms that the Gradient Boosting model trained with original data achieves a high recall (0.930) while maintaining a reasonable precision (0.726) and a good overall F1-score (0.816). This aligns with our goal of prioritizing recall and indicates that the model is performing consistently well on unseen data.

8 Actionable Insights & Recommendations

8.1 Write down insights from the analysis conducted

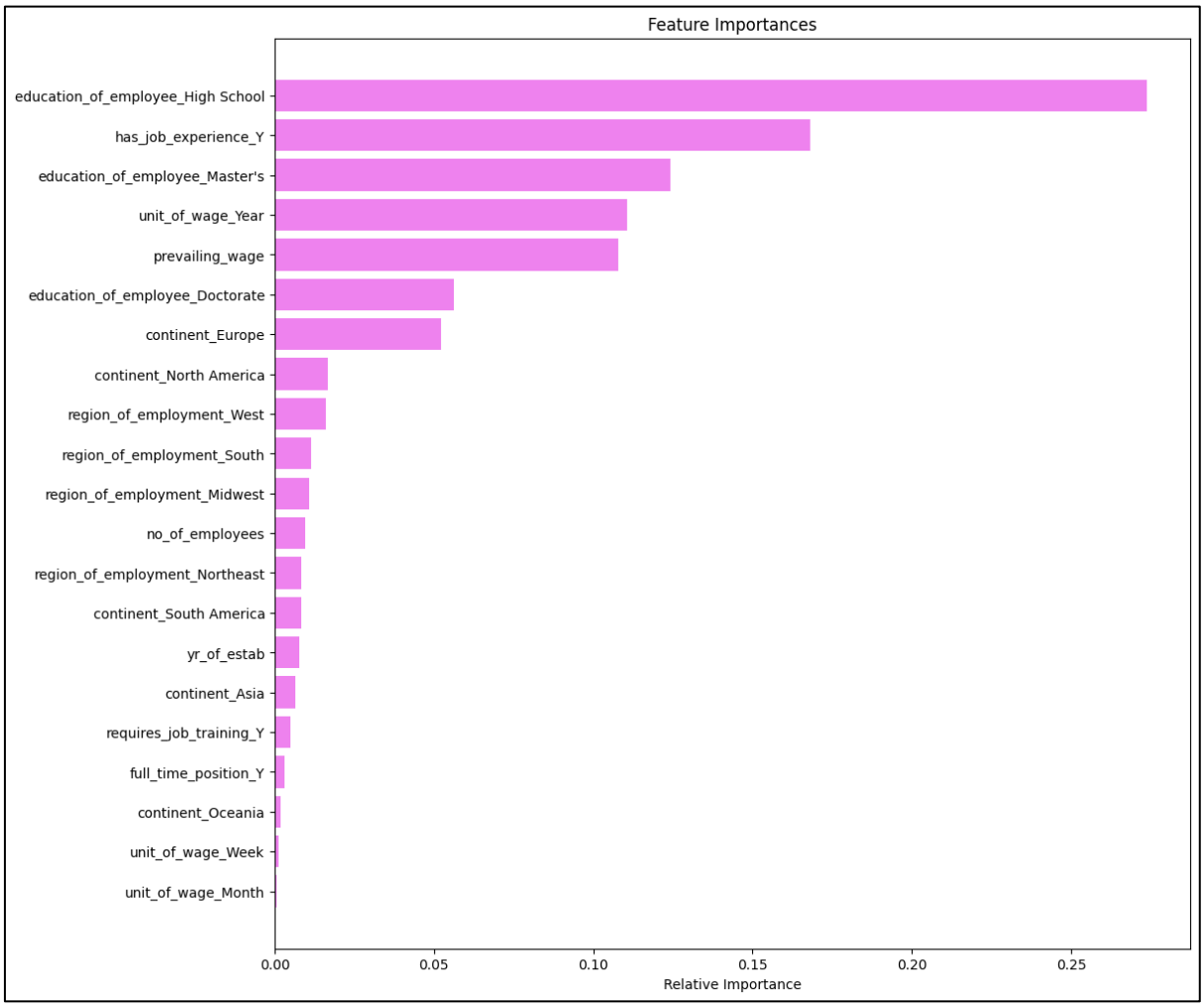

Figure 52 Feature importance

Insights

Dominant feature : High school are dominant feature that are strong predictors for Visa approval rate.

Significant feature: Job experience, Masters education, wage unit as year, prevailing wage are important predictors for Visa approval.

Important feature: Doctorate degree and European application plays role in predicting the visa approval.

Other features have low importance in predicting the visa approval.

Key Factors for Visa Approval – Higher education, job experience, and higher wages significantly improve visa approval chances, while lower wages face more denials.

Regional and Continental Trends – Asia has the highest number of visa applications, while Europe and the Midwest/South regions show higher approval rates.

Employment Characteristics – Most applicants hold at least a Bachelor's degree, have job experience, and apply for full-time, salaried positions.

Data Anomalies – Issues like negative employee counts, suggest potential data inaccuracies.

Significant Predictors – High school education is a dominant predictor, while job experience, Master's education, prevailing wage, and yearly wage unit play a significant role in visa approval.

8.2 Provide actionable business recommendations

Enhance Recruitment Strategies

- Focus on attracting candidates with higher education (Master's, Doctorate) and significant job experience to improve visa approval rates.
- Prioritize skilled professionals from Asia and Europe, given high application volumes and approval rates.

Optimize Wage Structures

- Offer competitive wages, as higher wages correlate with higher visa approval rates.
- Reduce reliance on unpaid internships or zero-wage positions to avoid application rejections.

Regional Expansion & Workforce Planning

- Target recruitment efforts in high-approval regions like the Midwest and South to increase hiring success rates.
- Strengthen business presence in the Northeast, where visa applications are highest, to meet workforce demands.