greatlearning
Learning for Life

Business report

EASY VISA

DATA SCIENCE AND BUSINESS ANALYTICS

greatlearning
Learning for Life

Content:

1 .Exploratory Data Analysis

- 1.1. Problem definition
- 1.2. Univariate analysis
- 1.3. Bivariate analysis
- 1.4. Use appropriate visualizations to identify the patterns and insights
- 1.5. Answers to the EDA key questions provided
- 1.6. Key meaningful observations on individual variables and the relationship between variables.

2. Data preprocessing

- 2.1. Prepare the data for analysis
- 2.2. Feature Engineering
- 2.3. Missing value Treatment
- 2.4. Outlier Treatment
- 2.5. Ensure no data leakage among train-test and validation sets

3. Model building –Original data

- 3.1. Choose the appropriate metric for model evaluation
- 3.2. Build 5 models (from decision trees, bagging and boosting methods):
- 3.3. Comment on the model performance * You can choose NOT to build XGBoost if you are facing issues with the installation

4. Model Building - Oversampled Data

- 4.1. Oversample the train data
- 4.2. Build 5 models (from decision trees, bagging and boosting methods)
- 4.3. Comment on the model performance * You can choose NOT to build XGBoost if you are facing issues with the installation.

5. Model Building – Under sampled Data

5.1. Under sample the train data

5.2. Build 5 models (from decision trees, bagging and boosting methods)

5.3. Comment on the model performance * You can choose NOT to build XG Boost if you are facing issues with the installation

6. Model Performance Improvement using Hyper parameter Tuning

6.1 Choose 3 models (at least) that might perform better after tuning with proper reasoning

6.2. Tune the 3 models (at least) obtained above using randomized search and metric of interest

6.3. Comment on the performance of 3 tuned models * You can choose NOT to tune XG Boost if you experience long runtimes

7. Model Performance Comparison and Final Model Selection

7.1. Compare the performance of tuned models

7.2. Choose the best model

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

8. Actionable Insights & Recommendations

8.1. Write down insights from the analysis conducted

8.2. Provide actionable business recommendations

greatlearning
Learning for Life

1. Exploratory Data Analysis

1.1. Problem definition:

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 Easy Visa for data-driven solutions. You as a data scientist at Easy Visa 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.

1.2. Data background and content.

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.

Data Description

The data contains the different attributes of customers' booking details. The detailed data dictionary is given below.

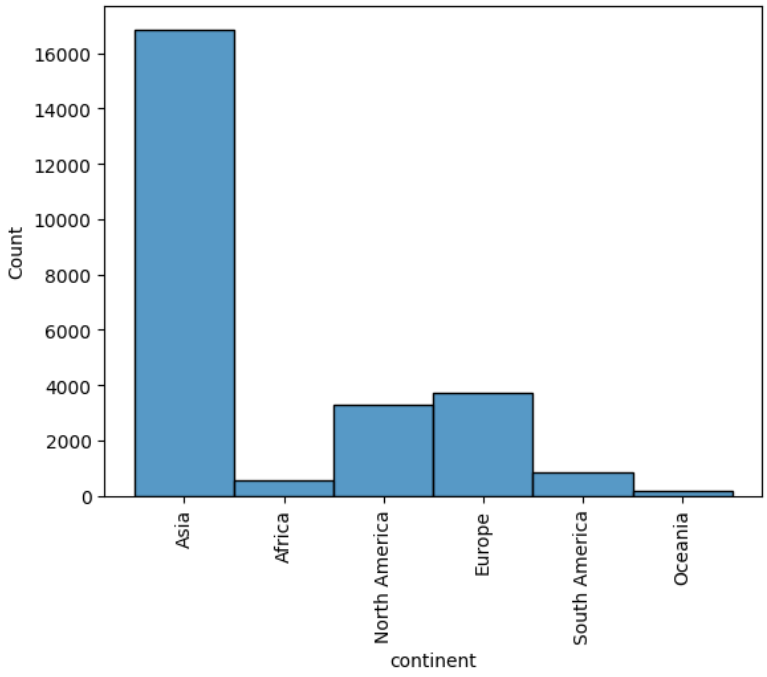
Data Dictionary:

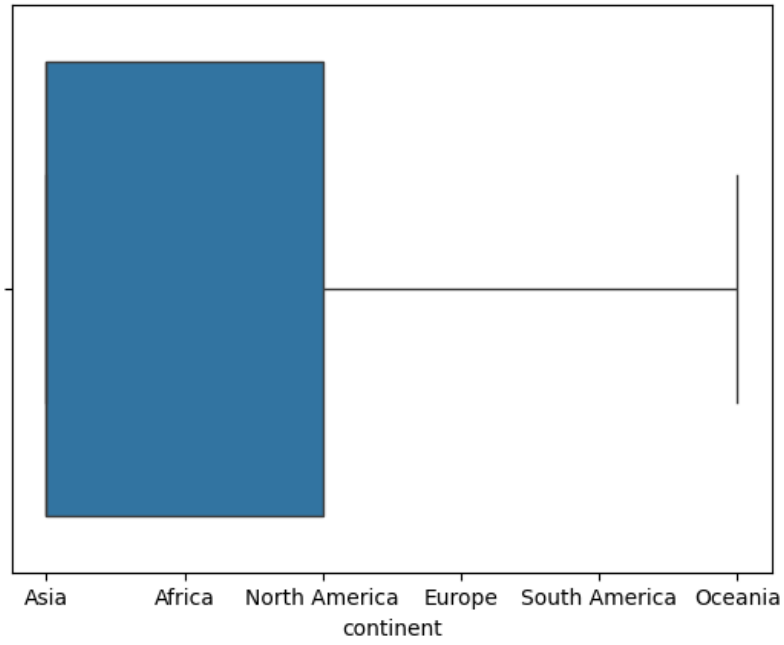
- case_id: ID of each visa application
- continent: Information of continent the employee
- education_of_employee: Information of education of the employee
- has_job_experience: Does the employee has any job experience? Y= Yes; N = No
- requires_job_training: Does the employee require any job training? 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

greatlearning
Learning for Life

1.3. Univariate analysis.

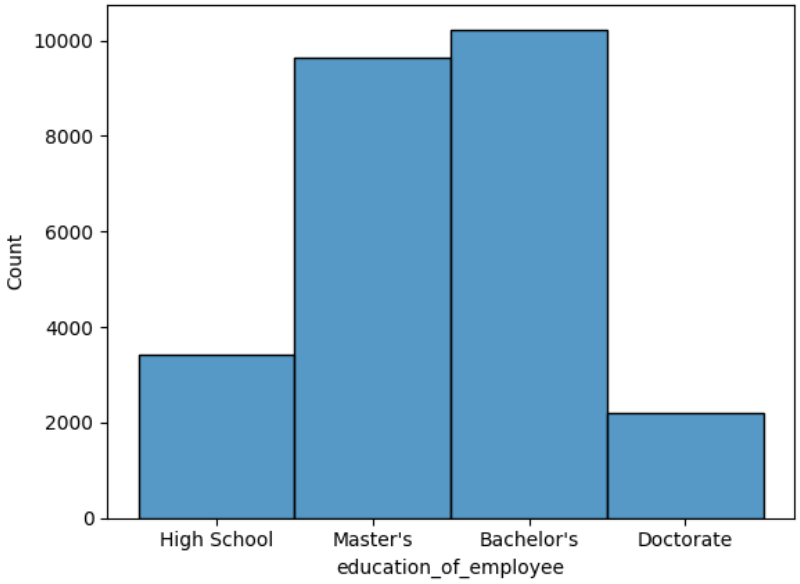
- **Continent :**

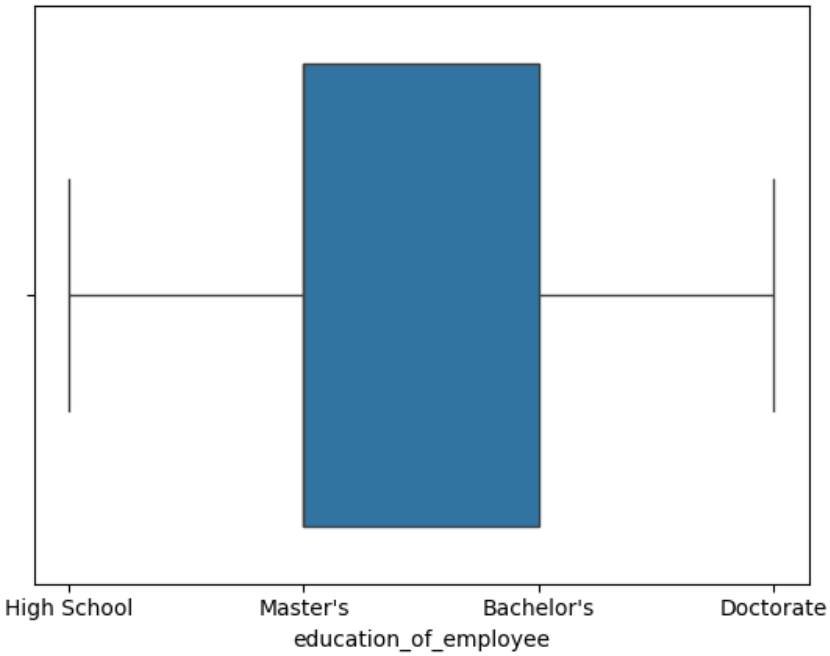




- Most of the employees are from Asia continent
- Less number of employees are from oceania.

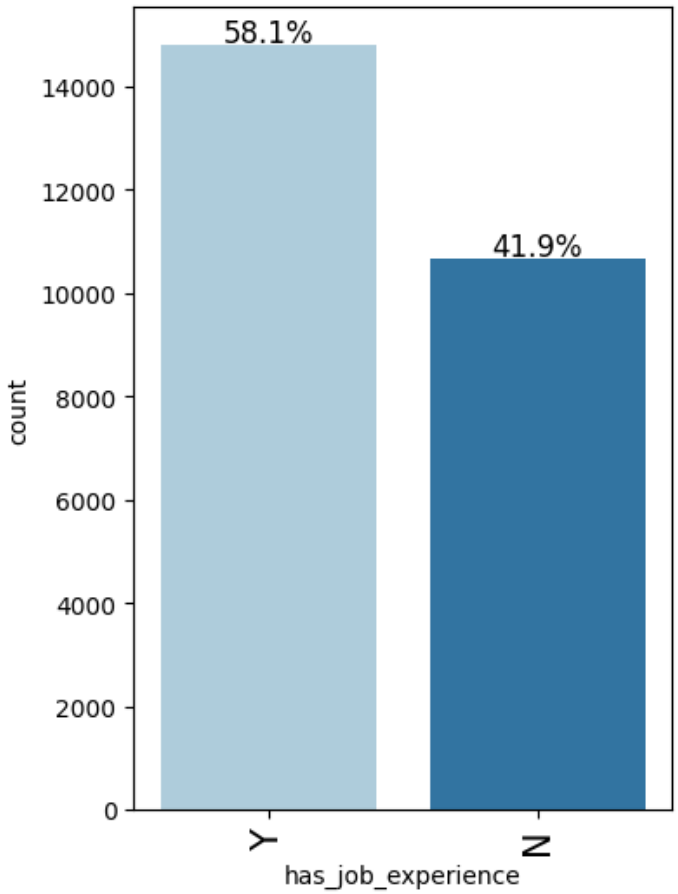
- **Education of employees :**





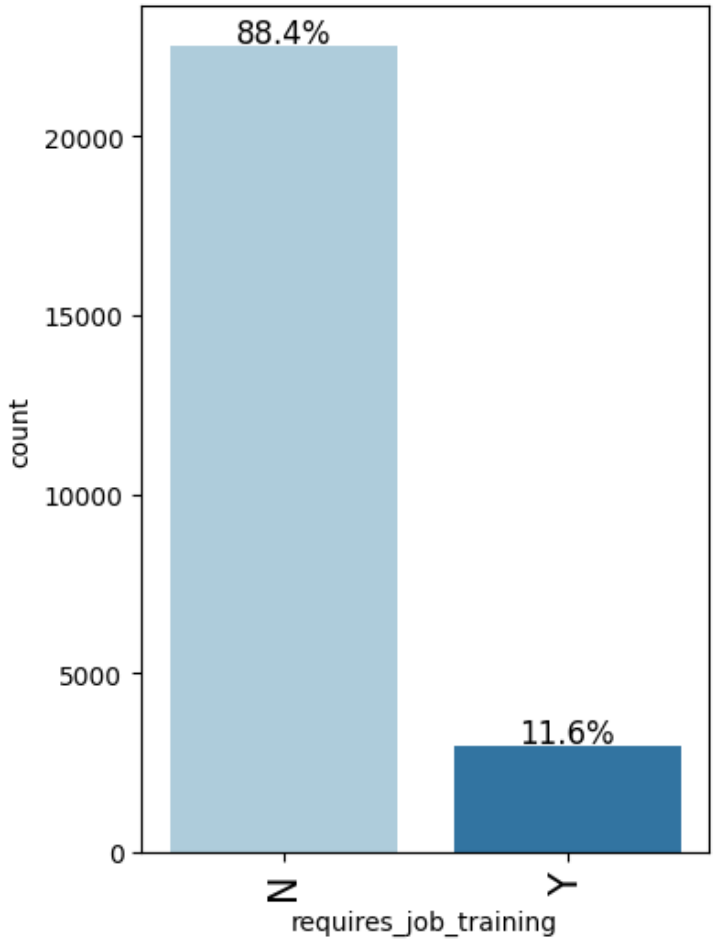
- Number of Bachelor's are more than others.
- Number of Master's are lil less than bachelor's.
- Employees who has doctorate are very less even compared to high school.

- **Has job experience :**



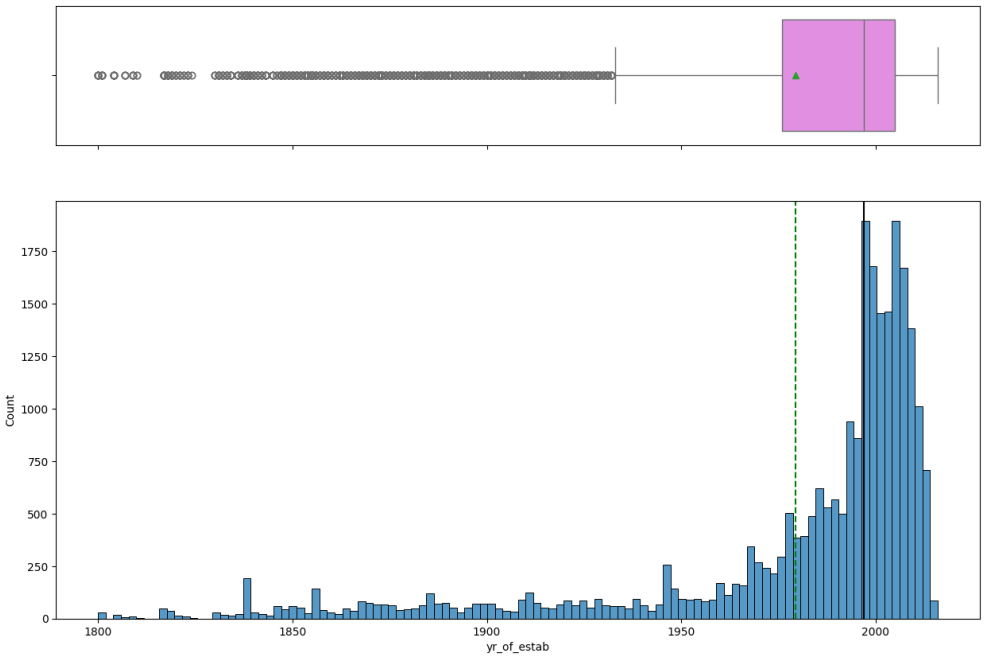
- Employees with experience are more than the employees without experience.
- Employess with experience has 58%.
- Whereas employees without experience has 41.9%.

- **Requires job training :**



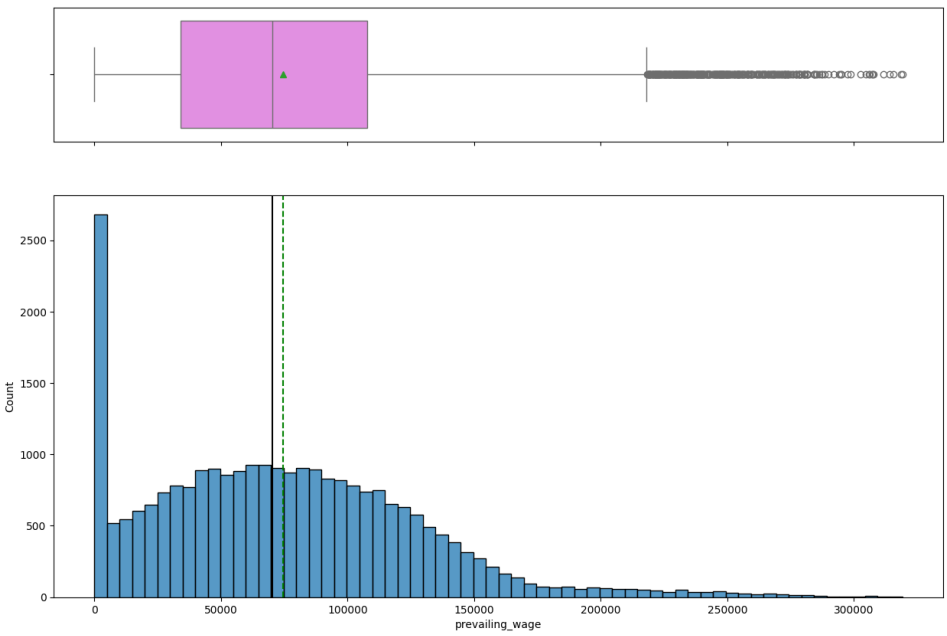
- Employees who doesn't requires job training are more than the employees who requires job training.
- Those who dont need job training are 88.4%
- Whereas who need job training are 11.6%.

- **Year of establishment :**



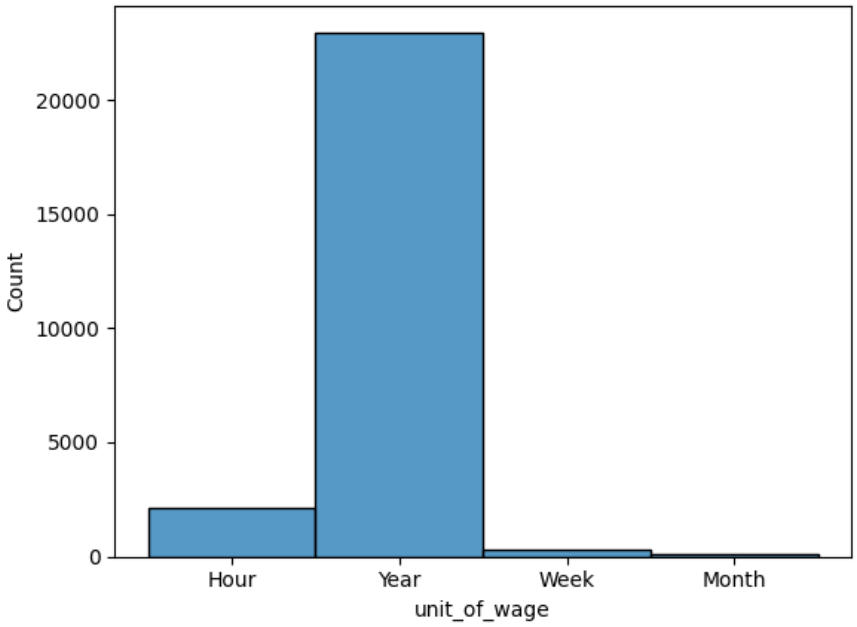
- Year of establishment is from 1800 to 2000.

- **Prevailing wages :**



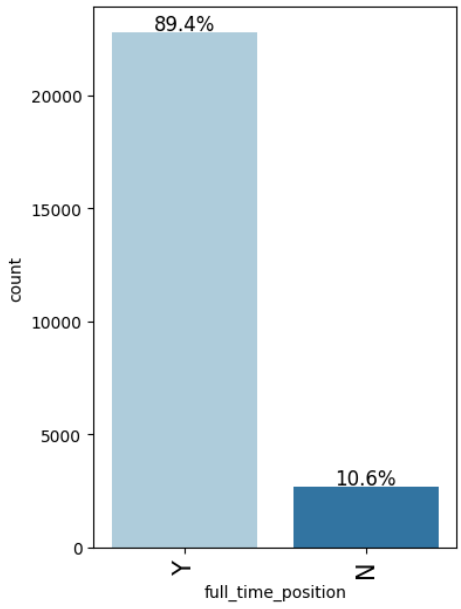
- Prevailing wages have outliers.
- Its right skewed.

- **Unit of wage:**



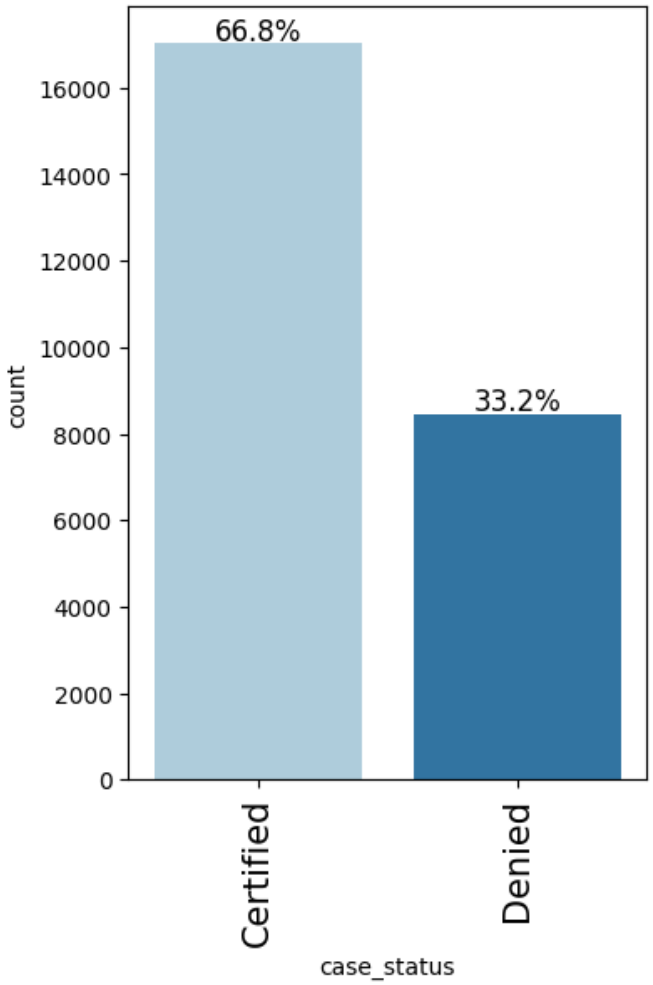
- Wages based on year is much more than others.
- Wages based on month is much lesser than others.

- **Full time position:**



- Employees with full time are much more than employees with part time.
- Employees with full time have 89.4%.
- Employees with part time have 10.64%.

- **Case status :**

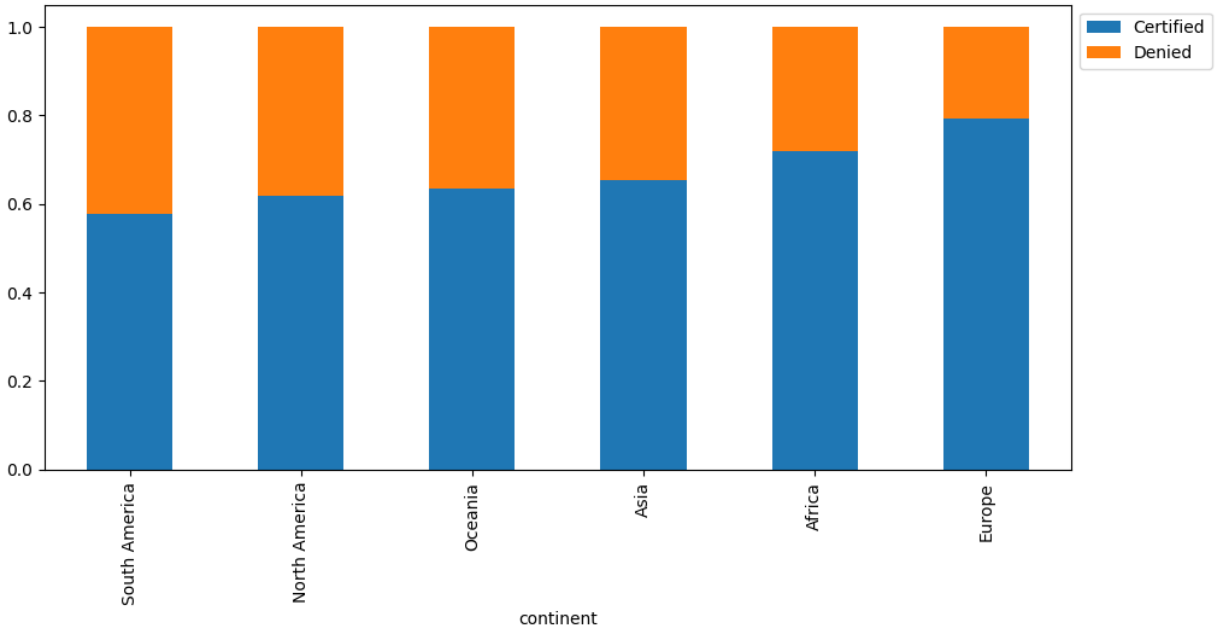


- Employees with certified visa are much more than denied.
- Employees with certified visa are 66.8%.
- Employees with denied visa are 33.2%.

greatlearning
Learning for Life

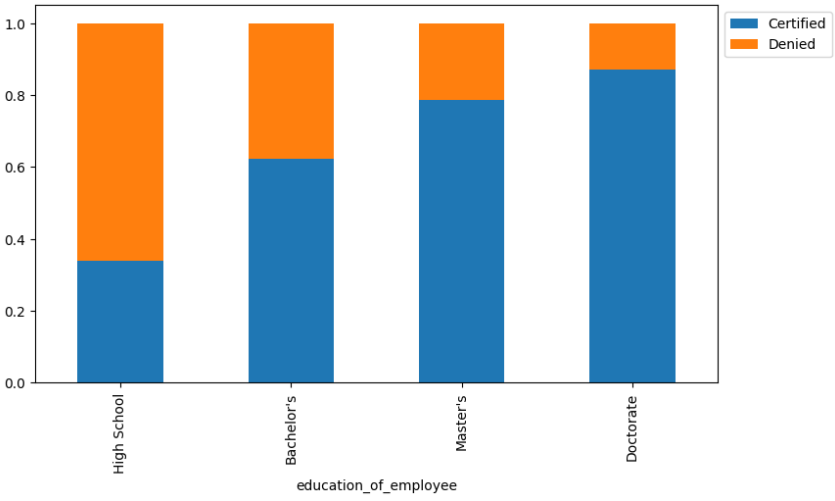
1.4. BIVARIATE ANALYSIS

- **Continent vs case status:**



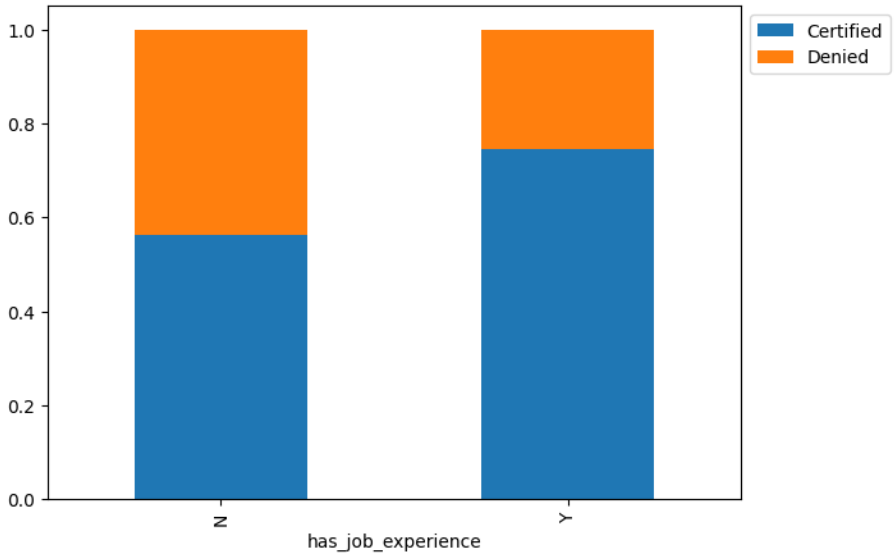
- Employees from Europe are Certified visa more.
- And the employees from South America are Certified less.

- **Education of employee vs case status :**



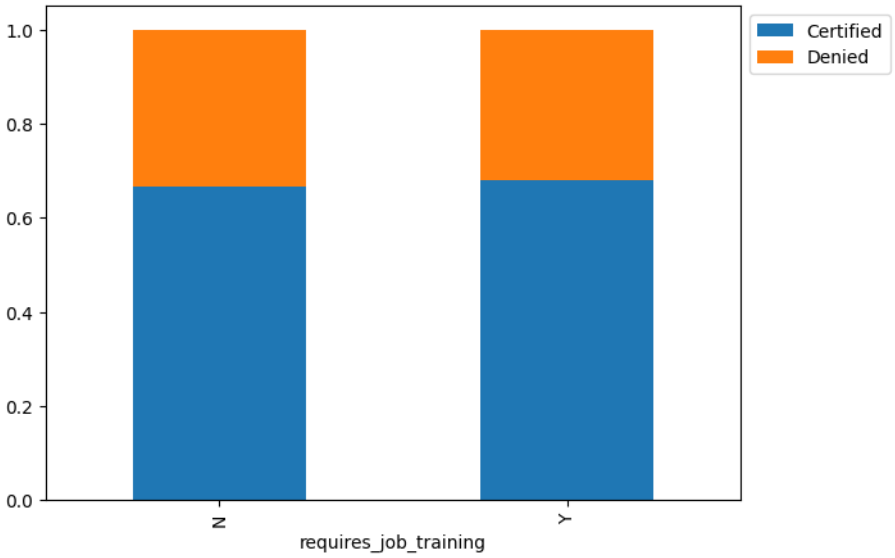
- Employees with doctorate are more Certified than the others.
- Employees with High school education are more denied.

- **Has job experience vs case status:**



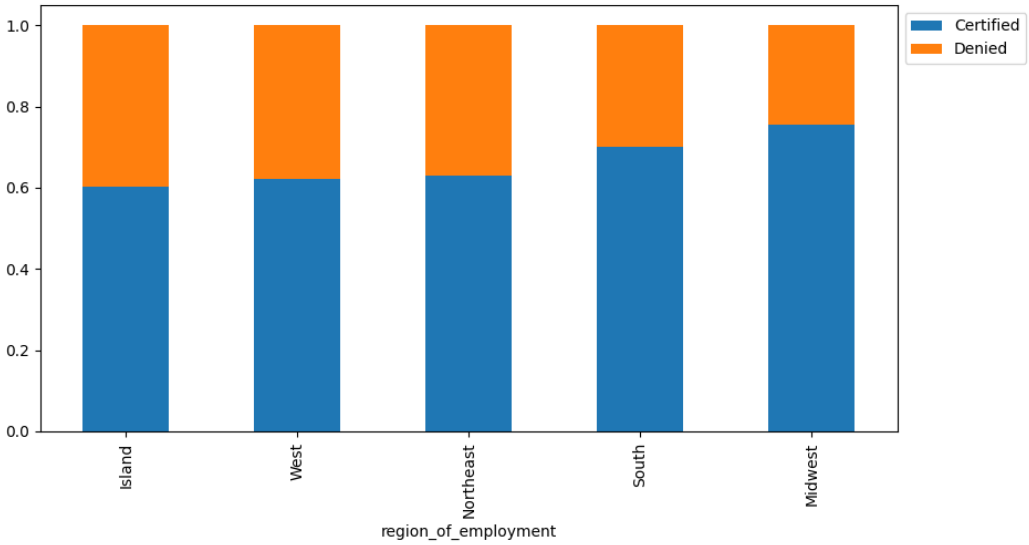
  - Employees with job experience are more Certified than the others.
  - Employees with less job experience are more denied.

- **Requires job training vs case status:**



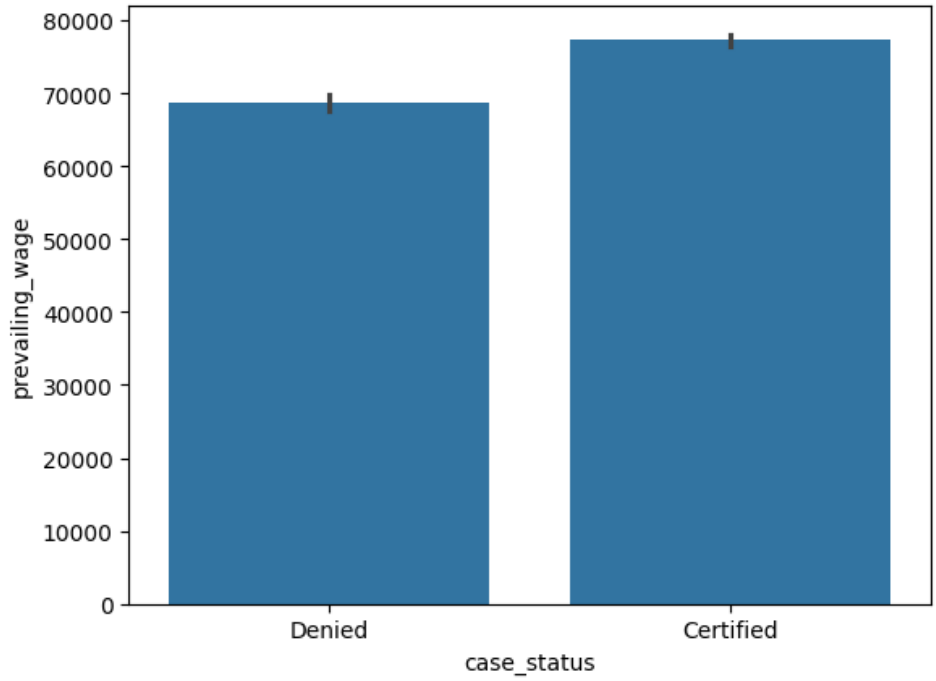
  - Requires job training and not requires job training are not much impacting to the certifiying or deniying visa.

- **Region of Employees vs case status:**



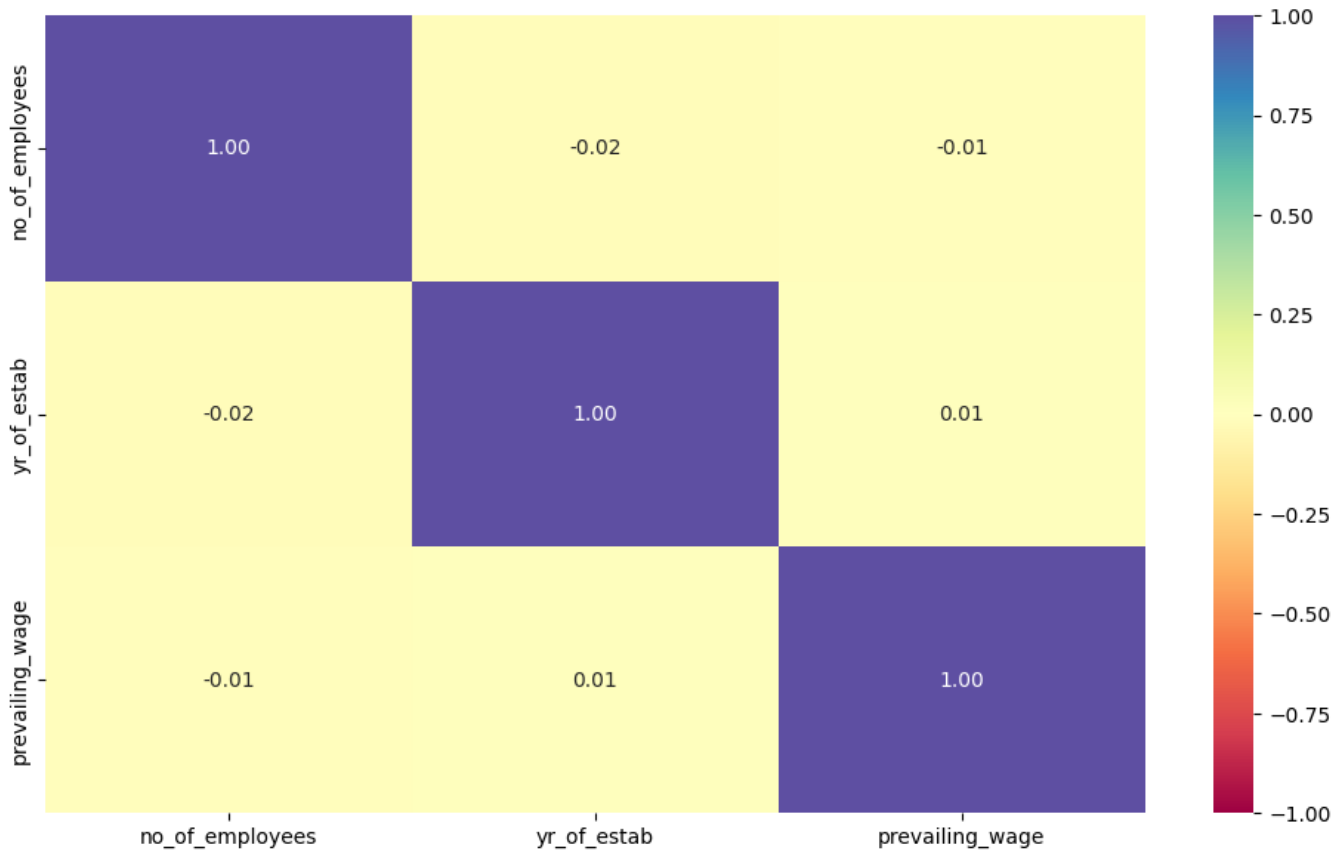
- Mild west region of employment are more certification than other region.
- Island region has less certification.

- **Prevailing wages vs case status:**



- With highest wages has certified more.

- **Correlation matrix:**



- Number of employees has negative correlation with year of establishment and previaling wages.
- year of establishment has postive correlation with prevailing wages.

Summary of EDA

- The dataset has 25480 rows and 12 columns
- Only 3 variables are numerical rest all are object types.
- Should work on no of employees column.

Data cleaning

- Drop "case_id" as "case_id" is unique for each candidate and might not add value to modeling.

greatlearning
Learning for Life

Observations from EDA

Univariate Analysis

-Continent: Most of the employees are from Asia continent. Less number of employees are from oceania.

-Education of employees: Number of Bachelor's are more than others. Number of Master's are lil less than bachelor's. Employees who has doctorate are very less even compared to high school.

-Has job experience: Employees with experience are more than the employees without experience. Employees with experience has 58%.Whereas employees without experience has 41.9%.

-Requires job training: Employees who doesn't requires job training are more than the employees who requires job training. Those who don’t need job training are 88.4%.Whereas who need job training are 11.6%.

-Region of employment: Employees who are from Northeast are more than other regions. Employees from Island are very less compared to others.

-Unit of wages: Wages paid based on year is much more than others. Wages paid based on month is much lesser than others.

-Full time position: Employees with full time are much more than employees with part time. Employees with full time have 89.4%.Employees with part time have 10.64%.

Case study: Employees with certified visa are much more than denied. Employees with certified visa are 66.8%.Employees with denied visa are 33.2%.

Bivariate Analysis

Continent: Employees from Europe are Certified visa more. And the employees from South America are Certified less.

-Education of employees: Employees with doctorate are more Certified than the others. Employees with High school education are more denied.

-Has job experience: Employees with job experience are more Certified than the others. Employees with less job experience are more denied.

-Requires job training: Requires job training and not requires job training are not much impacting to the certifying or denying visa.

-Year of establishment: Certifiying Visa is improving year by year. In 1824 ,not certified a single visa.

-Region of employment: Mild west region of employment are more certification than other region. Island region has less certification.

-full time position: Full time position or not is not effecting that much to visa certification.

2. DATA PREPROCESSING

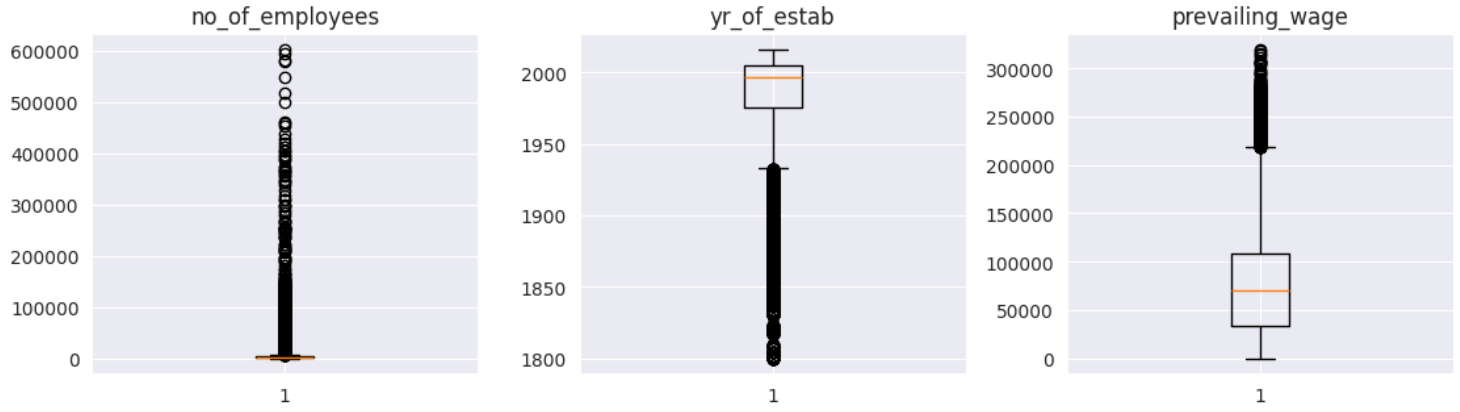
2.1. Checking for duplicate entries in the dataset

- There are no duplicate entries in the dataset.

Checking for missing values in the dataset.

- There are no null values in the dataset.

2.2. Outlier checking and treatment



Observations

- There are quite a few outliers in the data.
- However, we won't treat any of the outliers.

- And Here is a year of establishment, year of establishment, prevailing wages column which looks genuine..
- Treating outliers which looks genuine will lead to loss in data.

2.3. Feature engineering

- In feature engineering we dropped unwanted columns.
- Some of the number of employees were not correctly given, multiplied negative values with -1 not make negative values to positive values.

2.4. Data preparation for modeling

- We want to predict the case status.
- Before we proceed to build a model, we'll have to encode categorical features
- We'll split the data into train, validation and test to be able to evaluate the model that we build on the train data
- We will build a classification models using the train data, validate it with validation data and then check it's performance using test data.

greatlearning
Learning for Life

3. Model building

Model evaluation criterion

Model can make wrong predictions as:

- Predicting as visa denied but in real visa is approved - Loss of resources.
- Predicting as visa approved but in reality visa denied - Loss of opportunity.

Which case is more important?

Both the cases are important as:

- If we predicting as visa denied but in real visa is approved - Loss of resources.
- If we predicting as visa approved but in reality visa denied - Loss of opportunity.

How to reduce this loss?

We need to reduce both False Negatives and False Positives

- f1_score should be maximized as the greater the f1_score, the higher the chances of reducing both False Negatives and False Positives and identifying both the classes correctly
- fi_score is computed as

$$f1\_score = 2 * Precision * Recall / Precision + Recall$$

greatlearning
Learning for Life

Model building – Original data

Training Performance:

```
Bagging: 0. 9888344760039177
Random forest: 1.0
GBM: 0.8768854064642507
Adaboost: 0.886973555337904
dtree: 1.0
```

Validation Performance:

```
Bagging: 0.7746768507638073
Random forest: 0.7955346650998825
GBM: 0.8707403055229143
Adaboost: 0.8786721504112809
dtree: 0.7397179788484136
```

Training and Validation Performance Difference:

Bagging: Training Score: 0.9845, Validation Score: 0.7747, Difference: 0.2098
Random forest: Training Score: 1.0000, Validation Score: 0.7955, Difference: 0.2045
GBM: Training Score: 0.8769, Validation Score: 0.8707, Difference: 0.0061
Adaboost: Training Score: 0.8870, Validation Score: 0.8787, Difference: 0.0083
dtree: Training Score: 1.0000, Validation Score: 0.7397, Difference: 0.2603

- GBM has the best performance followed by AdaBoost model as per the validation performance

greatlearning
Learning for Life

4. Model Building – oversampled data :

Over sample the train data:

```
Before Oversampling, counts of label 'Yes': 10210
Before Oversampling, counts of label 'No': 5078

After Oversampling, counts of label 'Yes': 10210
After Oversampling, counts of label 'No': 10210

After Oversampling, the shape of train_X: (20420, 18)
After Oversampling, the shape of train_y: (20420,)
```

Model performance on training and validation dataset:

```
Training Performance:

Bagging: 0.9794319294809011
Random forest: 0.9999020568070519
GBM: 0.8442703232125367
Adaboost: 0.8590597453476984
dtree: 1.0

Validation Performance:

Bagging: 0.7403055229142186
Random forest: 0.7790834312573442
GBM: 0.836662749706228
Adaboost: 0.8569330199764983
dtree: 0.7162162162162162
```

Training and Validation Performance Difference:

Bagging: Training Score: 0.9794, Validation Score: 0.7403, Difference: 0.2391
Random forest: Training Score: 0.9999, Validation Score: 0.7791, Difference: 0.2208
GBM: Training Score: 0.8443, Validation Score: 0.8367, Difference: 0.0076
Adaboost: Training Score: 0.8591, Validation Score: 0.8569, Difference: 0.0021
dtree: Training Score: 1.0000, Validation Score: 0.7162, Difference: 0.2838

- Adaboost has the best performance on oversampled data followed by GBM

greatlearning
Learning for Life

5. Model building – under sampled data

Undersample the train data:

```
Before Under Sampling, counts of label 'Yes': 10210
Before Under Sampling, counts of label 'No': 5078

After Under Sampling, counts of label 'Yes': 5078
After Under Sampling, counts of label 'No': 5078

After Under Sampling, the shape of train_X: (10156, 18)
After Under Sampling, the shape of train_y: (10156,)
```

Performance on training and validation dataset:

Training Performance:

```
Bagging: 0.9629775502166207
Random forest: 0.9998030720756204
GBM: 0.7512800315084679
Adaboost: 0.7026388341866877
dtree: 1.0
```

Validation Performance:

```
Bagging: 0.61427732079906
Random forest: 0.6551116333725029
GBM: 0.736192714453584
Adaboost: 0.7029964747356052
dtree: 0.6239717978848414
```

Training and Validation Performance Difference:

Bagging: Training Score: 0.9630, Validation Score: 0.6143, Difference: 0.3487
Random forest: Training Score: 0.9998, Validation Score: 0.6551, Difference: 0.3447
GBM: Training Score: 0.7513, Validation Score: 0.7362, Difference: 0.0151
Adaboost: Training Score: 0.7026, Validation Score: 0.7030, Difference: -0.0004
dtree: Training Score: 1.0000, Validation Score: 0.6240, Difference: 0.3760

- GBM has the best performance followed by AdaBoost model as per the validation performance
- After building models, it was observed that both the GBM and Adaboost models, trained on an undersampled dataset, as well as the Adaboost and GBM model trained on an oversampled dataset, exhibited strong performance on both the training and validation datasets.
- Sometimes models might overfit after undersampling and oversampling, so it's better to tune the models to get a generalized performance
- We will tune these 3 models using the same data (undersampled or oversampled) as we trained them on before

6. Model performance improvement using Hyper parameter tuning

6.1. Tuning AdaBoost Classifier model with undersampled data:

Checking model performance using adaboost on undersampled train data:

	Accuracy	Recall	Precision	F1
0	0.687	0.654	0.700	0.676

- In adaboost undersampled accuracy is 68% , recall is 65.

Checking model performance using adaboost on oversampled validation data:

	Accuracy	Recall	Precision	F1
0	0.684	0.659	0.833	0.736

In adaboost oversampled accuracy is 68% , recall is 65.

6.2. Tuning Gradient boosting model with undersampled data:

Checking model performance using gradient boosting on undersampled train data:

	Accuracy	Recall	Precision	F1
0	0.705	0.745	0.689	0.716

- In gradient boosting on undersampled data accuracy is 70% and recall is 74%.

Checking model performance using gradient boosting on undersampled validation data

	Accuracy	Recall	Precision	F1
0	0.720	0.741	0.823	0.780

- In gradient boosting on oversampled data accuracy is 72% and recall is 74%.

6.3. Tuning Gradient boosting model with oversampled data:

Checking model performance using gradient boosting on oversampled train data:

	Accuracy	Recall	Precision	F1
0	0.690	0.748	0.671	0.707

- In gradient boosting oversampled data accuracy is 69% and recall is 74%

Checking model performance using gradient boosting on oversampled validation data:

Accuracy	Recall	Precision	F1
0	0.720	0.741	0.823	0.780

- In gradient boosting oversampled model on validation dataset accuracy is 72% and recall is 74%.

greatlearning
Learning for Life

7 Model performance comparison and final model selection:

Compare the performance of tuned models:

Training performance comparison:

	Gradient boosting trained with Undersampled data	Gradient boosting trained with Oversampled data	AdaBoost trained with Undersampled data
Accuracy	0.705	0.690	0.687
Recall	0.745	0.748	0.654
Precision	0.689	0.671	0.700
F1	0.716	0.707	0.676

Validation performance comparison:

	Gradient boosting trained with Undersampled data	Gradient boosting trained with Oversampled data	AdaBoost trained with Undersampled data
Accuracy	0.720	0.720	0.684
Recall	0.741	0.741	0.659
Precision	0.823	0.823	0.833
F1	0.780	0.780	0.736

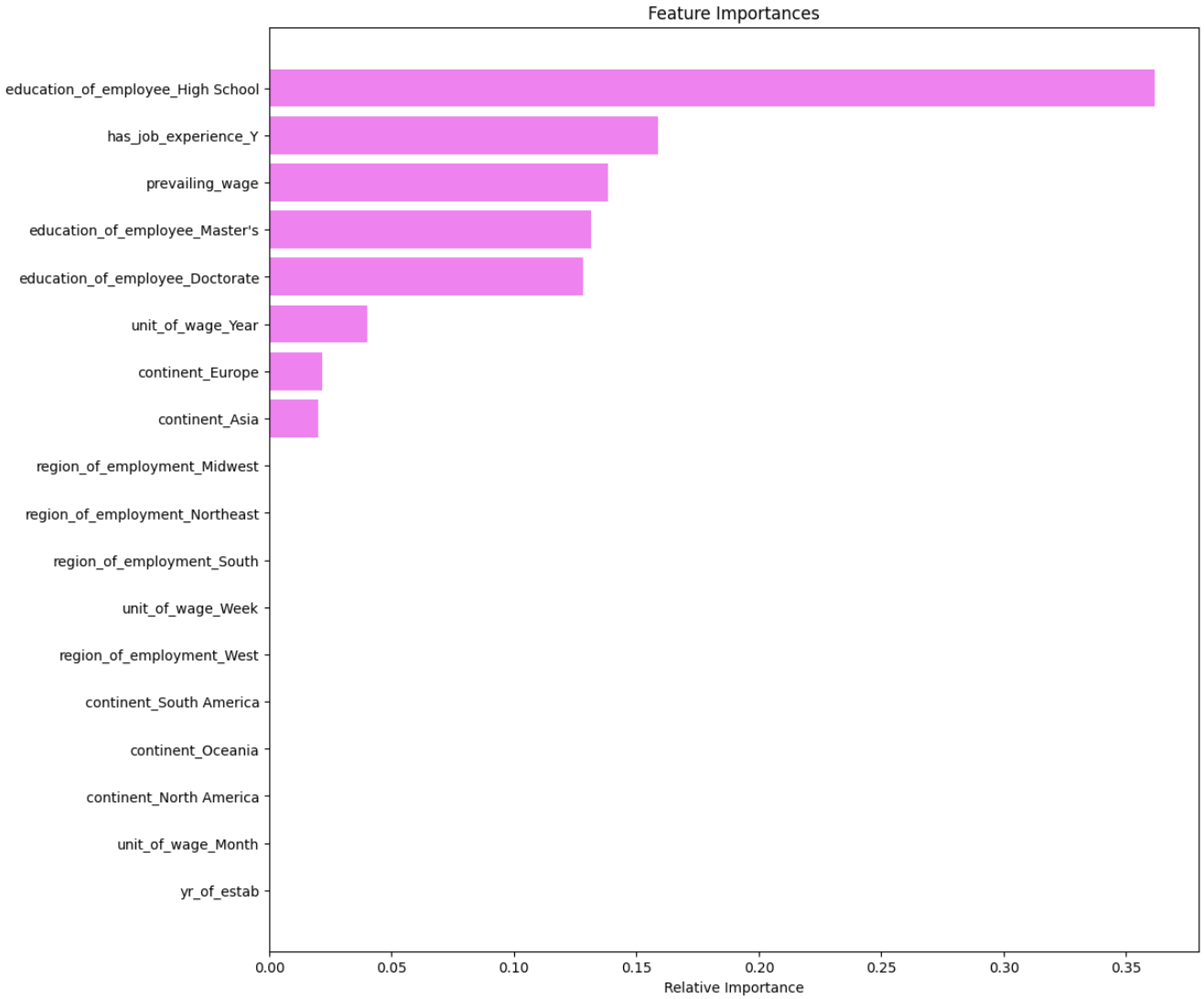
- AdaBoost model trained with undersampled data has generalised performance, so let's consider it as the best model.

Checking the performance on test set:

	Accuracy	Recall	Precision	F1
0	0.669	0.658	0.811	0.727

- The Gradient boosting model trained on undersampled data has given ~65% recall on the test set
- This performance is in line with what we achieved with this model on the train and validation sets
- So, this is a generalized model

Feature engineering:



- We can see that education of employee-high school is effecting the most.
- Has job experience yes, prevailing wages is influencing the visa certification.

greatlearning
Learning for Life

8. Actionable insights and recommendations:

8.1. Business insights

- Employees from Europe are certified visa more. And the employees from South America are certified less.
- Employees with doctorate are more certified than the others. Employees with High school education are more denied.
- Employees with job experience are more certified than the others. Employees with less job experience are more denied.
- Mild west region of employment are more certification than other region. Island region has less certification.
- That employee who has high school education level is affecting a lot. Later on Those employees who have job experience might have higher chances to get certified but high school education level have higher chances to get denied

8.2. Recommendations

Feature Engineering: Explore creating new features based on existing ones.

Focus on the employees who have more education: Those who have high school education are affected.

Focus on the employees who has job experience - yes: Those who have job experience are certified more.
