Machine Learning 2

Problem Statement for Data Science

By: Agnes Raja Kumari. E

PGP-Data Science and Business Analytics

PGPDSBA.O.MAY24.A

Contents

List of Figures

List of Tables

Introduction

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.

1.2 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.

1.3 Problem definition

EasyVisa, a firm hired by OFLC, seeks to develop a **machine learning-based solution** that can predict the likelihood of visa approval for applicants. The objective is to:

1. **Facilitate the visa approval process** by automating the prediction of whether a visa application will be approved or denied.
2. **Recommend profiles** based on factors that significantly influence visa certification outcomes, helping employers and authorities prioritize applications with a higher chance of success.

This model will leverage historical visa application data to identify key drivers of visa approval or denial, thereby optimizing the certification process and ensuring quicker, more accurate decision-making.

2. Data Background and Contents

2.1 Context and Variables

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 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.

2.2 Statistical Summary

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25480 entries, 0 to 25479
Data columns (total 12 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)
memory usage: 2.3+ MB
```

Figure 1: Info of the DataFrame

Observation:

1. **Number of Entries**:
    - The dataset contains **25,480 entries**, with no missing values in any column.
2. **Number of Columns**:
    - There are **12 columns** in total, consisting of a mix of data types.
3. **Data Types**:
    - **Object (Categorical)**: Most of the columns are of object type (9 out of 12 columns), indicating categorical data. These columns include:
        - case_id, continent, education_of_employee, has_job_experience, requires_job_training, region_of_employment, unit_of_wage, full_time_position, and case_status.
    - **Integer (int64)**: Two columns, no_of_employees and yr_of_estab, are of integer type.
    - **Float (float64)**: One column, prevailing_wage, is a floating-point type

Insights:

- **Encoding Categorical Variables**: Since most columns are categorical, they will need to be encoded (e.g., using one-hot encoding or label encoding) before being used in a machine learning model.
- **Outlier Handling**: Based on previous findings, there are potential outliers in columns like no_of_employees (e.g., negative values) that should be addressed.
- **Feature Scaling**: Columns like prevailing_wage may need normalization or standardization before model training.
- **Target Variable**: The case_status column will be the target variable for the classification model.

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

Figure 2: Statistical Summary of the Data

Observation:

1. **no_of_employees:**
    - Mean: The average number of employees in the dataset is 5,667.
    - Standard Deviation (std): There is a large variation in the number of employees across companies, as shown by the high standard deviation of 22,877.929.
    - Min: The minimum value is -26, which is invalid for the number of employees (likely due to data entry errors).
    - Max: The maximum number of employees is 602,069, indicating the presence of very large companies in the dataset.
    - Interquartile Range (IQR):
        - 25th percentile (Q1): 1,022 employees.
        - 50th percentile (Median): 2,109 employees.
        - 75th percentile (Q3): 3,504 employees.
    - Conclusion: The data includes both small and large companies, but negative values suggest the need for data cleaning.

2. **yr_of_estab (Year of Establishment):**
    - Mean: The average year of establishment is 1979.
    - Standard Deviation (std): The standard deviation is 42.367, indicating a wide spread in the years companies were founded.
    - Min: The earliest year of establishment is 1800.
    - Max: The most recent year of establishment is 2016.
    - Interquartile Range (IQR):
        - 25th percentile (Q1): Companies were established by 1976.
        - 50th percentile (Median): Companies were established by 1997.
        - 75th percentile (Q3): Companies were established by 2005.
    - Conclusion: The dataset contains both well-established companies and newer firms.
3. **prevailing_wage:**
    - Mean: The average prevailing wage offered is $74,455.815.
    - Standard Deviation (std): The standard deviation is $52,815.942, indicating a wide variation in wages offered.
    - Min: The minimum prevailing wage is $2.137, which is highly unrealistic and suggests potential outliers or data entry errors.
    - Max: The maximum prevailing wage is $319,210.270, which is on the higher end of the wage spectrum.
    - Interquartile Range (IQR):
        - 25th percentile (Q1): The wage at the 25th percentile is $34,015.480.
        - 50th percentile (Median): The median wage is $70,308.210.
        - 75th percentile (Q3): The wage at the 75th percentile is $107,735.513.
    - Conclusion: There is a wide range of wages offered, but extremely low values need to be addressed.

Insights:

- **Data Cleaning Needed:** There are invalid values in no_of_employees (negative) and prevailing_wage (unrealistically low values) that need to be corrected.
- **Skewness:** Both no_of_employees and prevailing_wage show signs of skewness with large differences between the median and mean values, indicating the presence of outliers.

Investigate outliers, clean erroneous data (like negative or extremely low values), and consider transformations or outlier handling methods to improve data quality before further analysis.

3. Exploratory Data Analysis

3.1 Univariate Analysis

3.1.1 Distribution of data - continent using bar plot

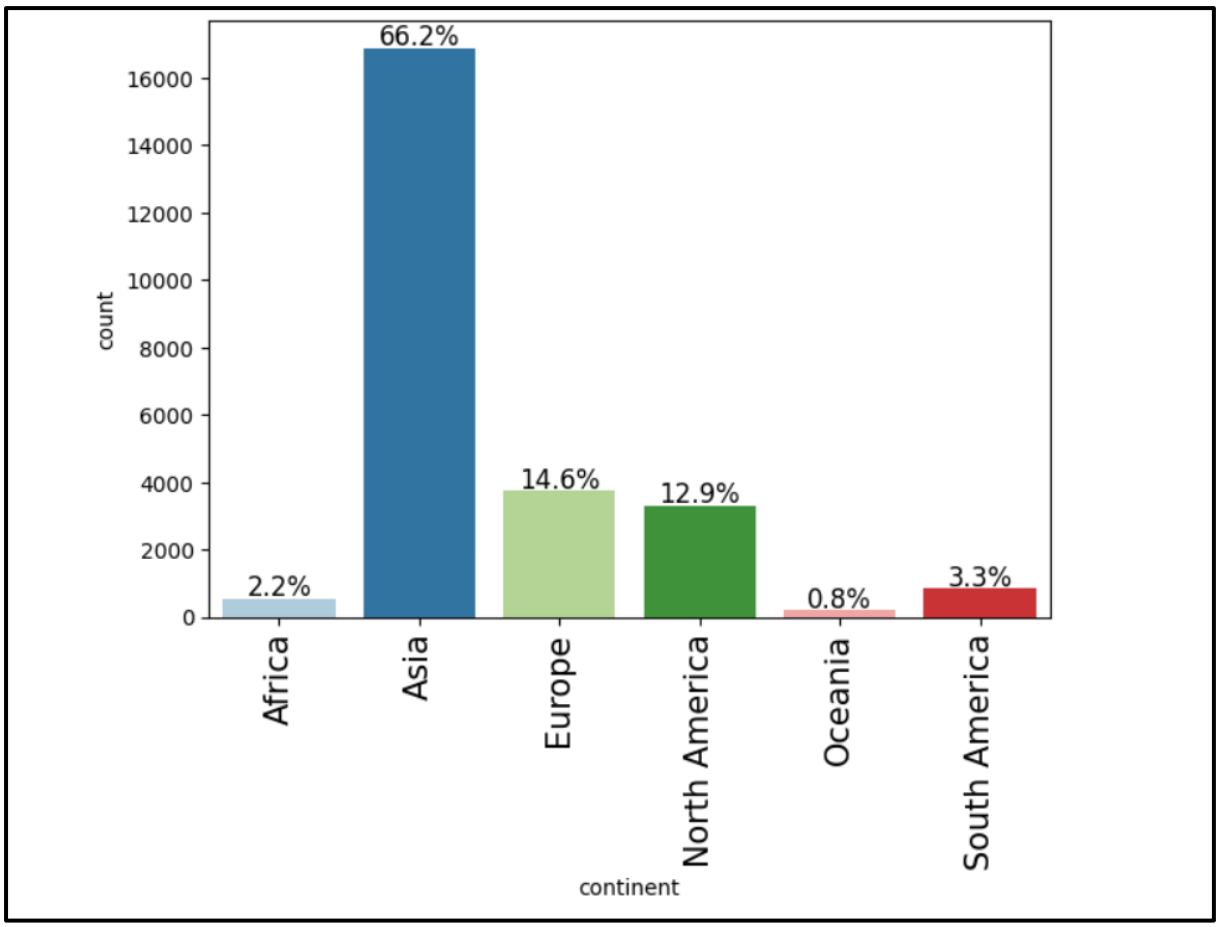


Figure 3: Distribution of data - continent using Barplot

Observation:

1. Asia has the highest count, representing 66.2% of the total data, with over 16,000 entries.
2. Europe follows at 14.6%, with a little more than 3,000 entries.
3. North America comes next, accounting for 12.9%.
4. South America has a small share, with 3.3% of the data.
5. Africa accounts for 2.2%, while Oceania represents the smallest portion, at 0.8%.

Overall, Asia dominates the dataset, contributing to more than half of the data points, followed by Europe and North America. Africa, Oceania, and South America make up minor portions.

3.1.2 Distribution of data - education_of_employee using barplot

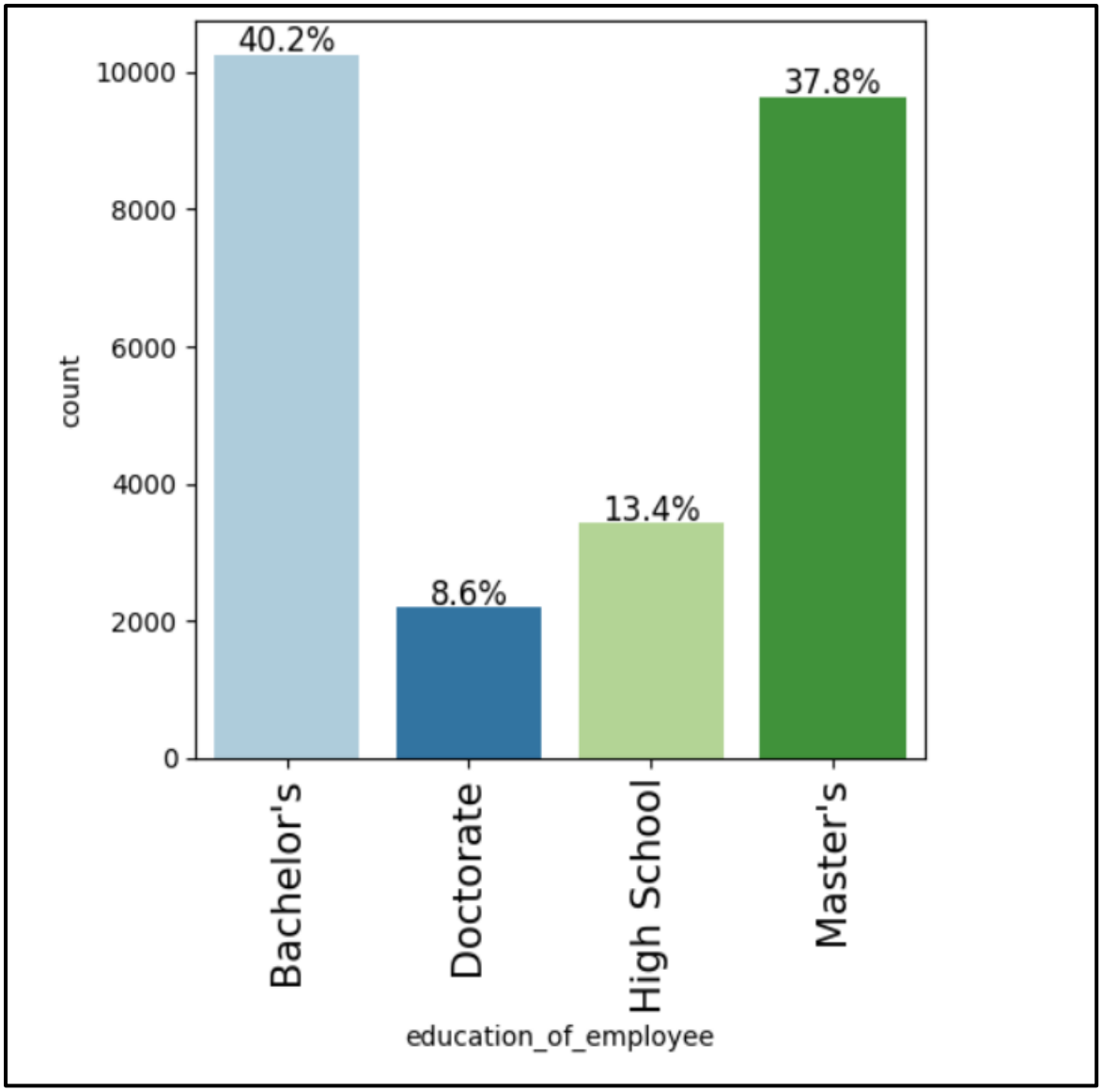


Figure 4:Distribution of data - education of employee using Barplot

Observation:

1. Bachelor's degree holders form the largest group, accounting for 40.2% of employees, with over 10,000 entries.
2. Master's degree holders are the second largest group, representing 37.8% of the data.
3. Employees with only a High School education make up 13.4% of the total.
4. Doctorate holders are the smallest group, at 8.6%.

Overall, most employees have either a Bachelor's or Master's degree, while those with a Doctorate or just a High School education form smaller portions of the dataset.

3.1.3 Distribution of data - has_job_experience using barplot

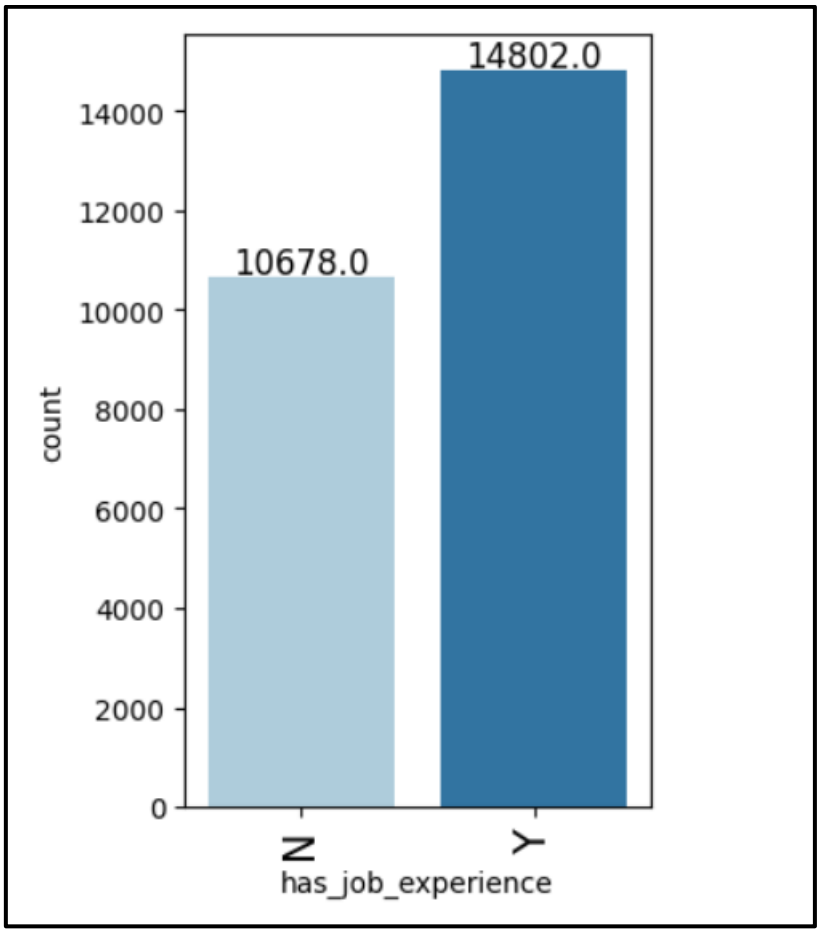


Figure 5: Distribution of data - has job experience using Barplot

Observation:

1. Employees with job experience ("Y") are the majority, with a count of 14,802.
2. Employees without job experience ("N") have a lower count, totaling 10,678.

Thus, more employees in this dataset have prior job experience compared to those who do not.

3.1.4 Distribution of data - requires_job_training using barplot

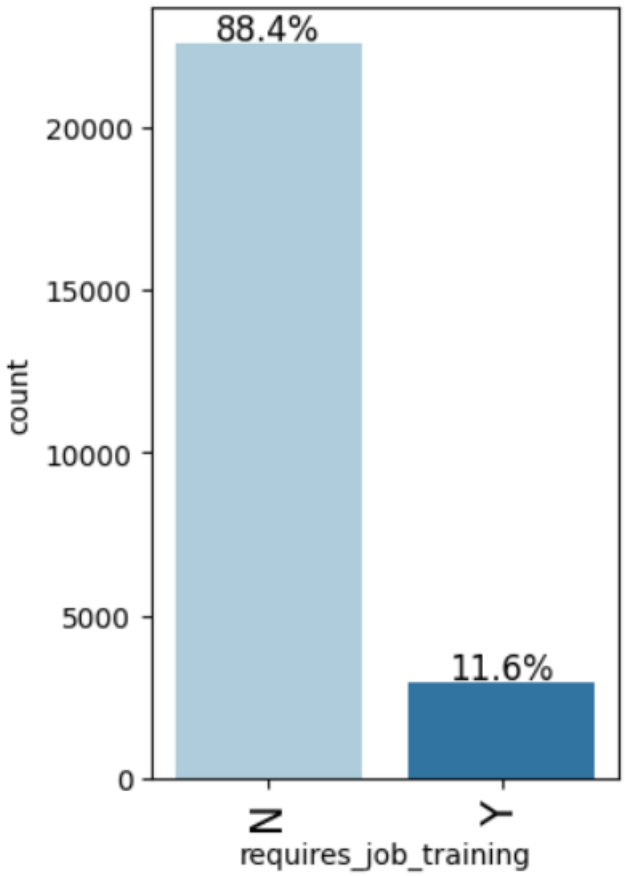


Figure 6: Distribution of data- requires job training using Barplot

Observation:

The majority of the data falls under the "N" category, accounting for 88.4% of the total, while only 11.6% of the data corresponds to the "Y" category.

- A significant proportion of the population (88.4%) does not require job training.
- Only a small fraction (11.6%) requires job training, as indicated by the shorter bar.

This suggests that most individuals in the dataset may be in roles or situations where job training is not necessary.

3.1.5 Distribution of data - region_of_employment using barplot

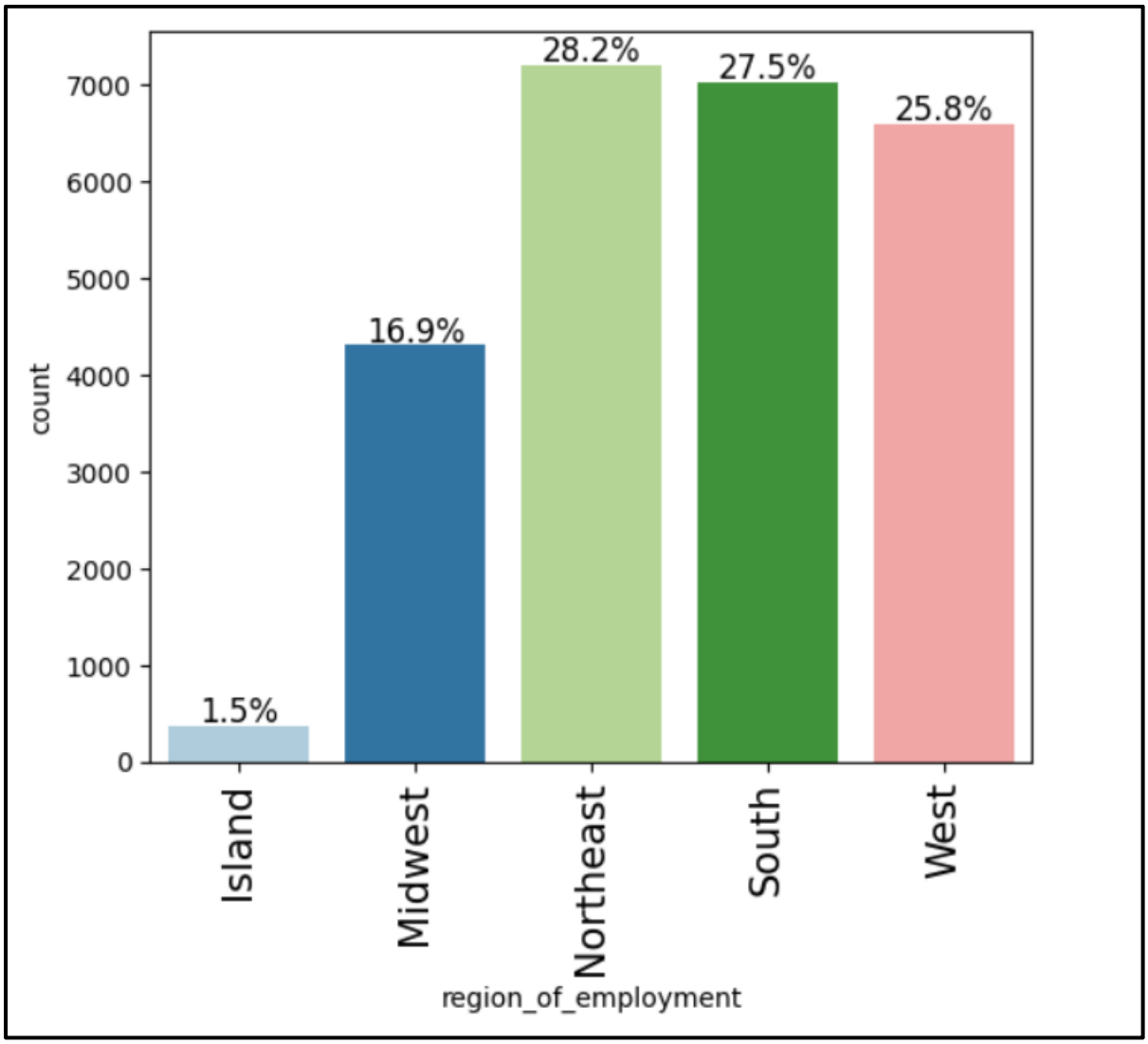


Figure 7: Distribution of data- region of employment using Barplot

Observation:

- The largest portion of employment is in the Northeast, with 28.2%, followed closely by the South at 27.5%.
- The West accounts for 25.8% of the employment.
- The Midwest has a smaller share at 16.9%.
- The Island region represents a very small fraction, only 1.5%.

This suggests that employment is concentrated in the Northeast, South, and West regions, with a minimal presence in the Island region.

3.1.6 Distribution of data - full_time_position using barplot

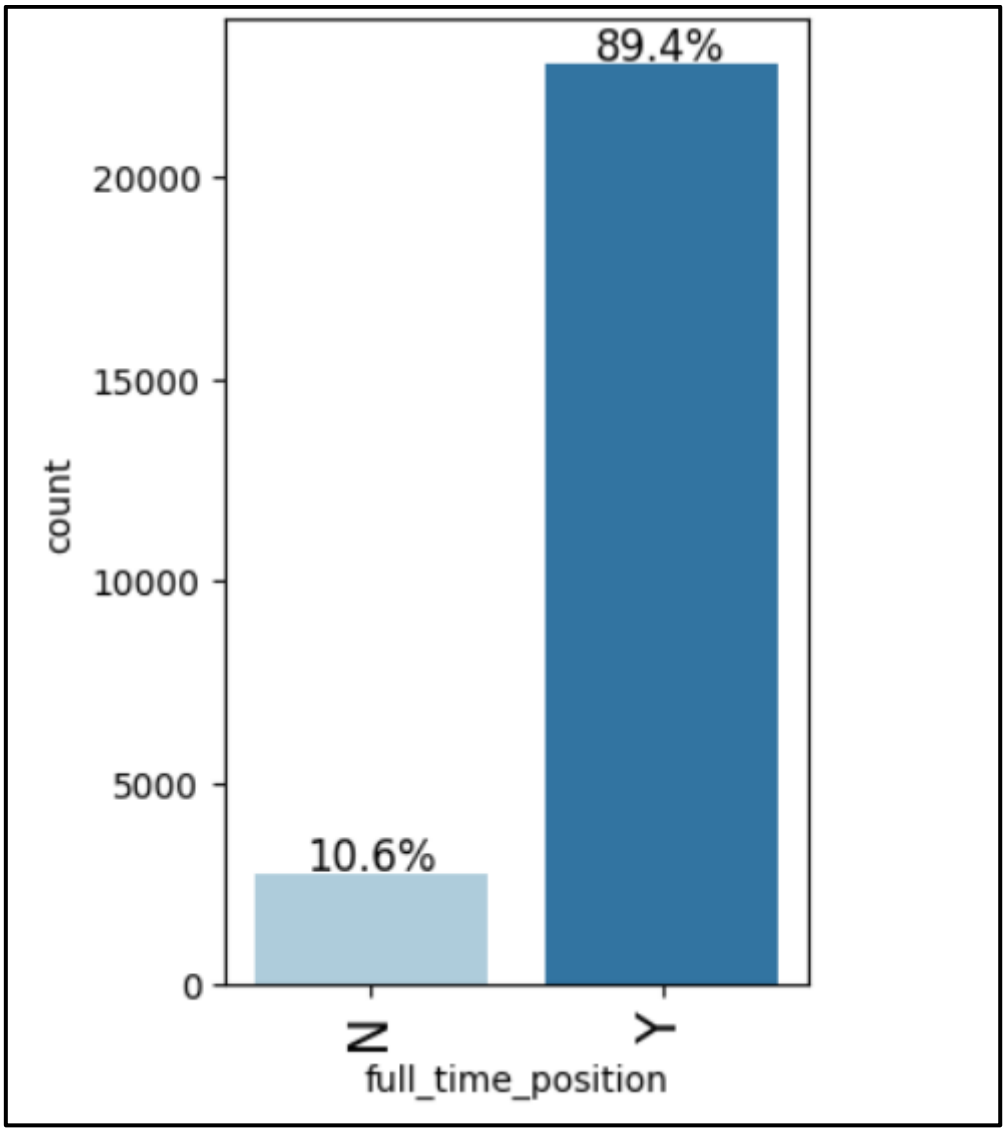


Figure 8: Distribution of data - full_time_position using barplot

Observation:

- The majority of positions (89.4%) are full-time, as indicated by the taller bar for "Y."
- Only 10.6% of positions are not full-time, represented by the shorter bar for "N."

This indicates that most of the employment in the dataset involves full-time positions, with only a small fraction of part-time or other non-full-time roles.

3.1.7 Distribution of data - unit_of_wage using barplot

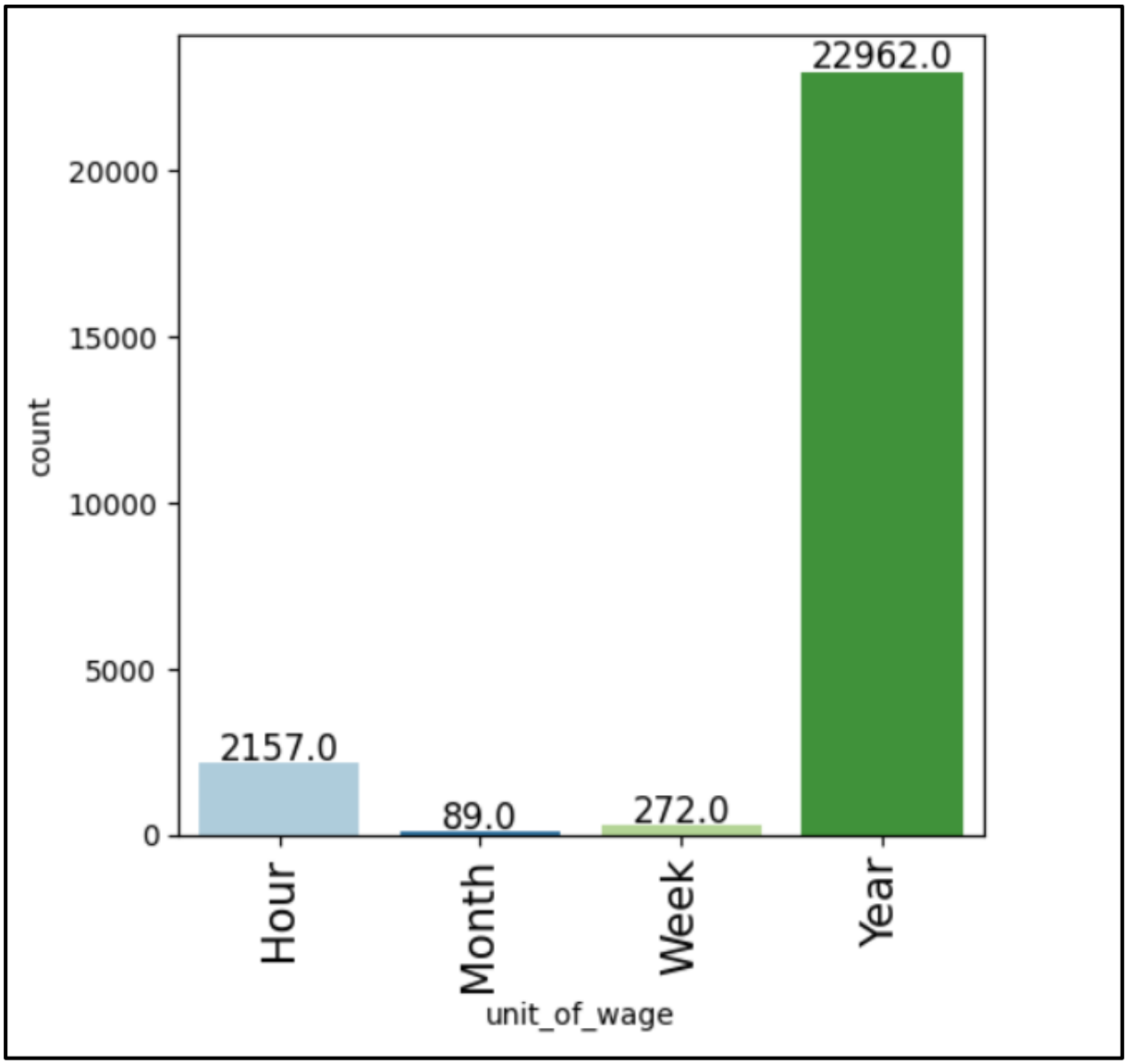


Figure 9: Distribution of data - unit_of_wage using barplot

Observation:

- Yearly Wage: The majority of visa applications (22,962) report wages on a yearly basis. This is the most common unit by a significant margin.
- Hourly Wage: There are 2,157 visa applications where the wage is reported on an hourly basis, making it the second most frequent unit.
- Weekly Wage: A smaller number of applications (272) report wages on a weekly basis.
- Monthly Wage: The least common unit is monthly, with only 89 applications reporting wages in this format.

Key Insight

- Yearly wages dominate the dataset, which may indicate that most positions applied for through the visa process are full-time or long-term roles that prefer or require an annual salary format.
- Other wage units, particularly monthly and weekly, have relatively low representation, suggesting these are much less commonly used.

3.1.8 Distribution of data - case_status using barplot

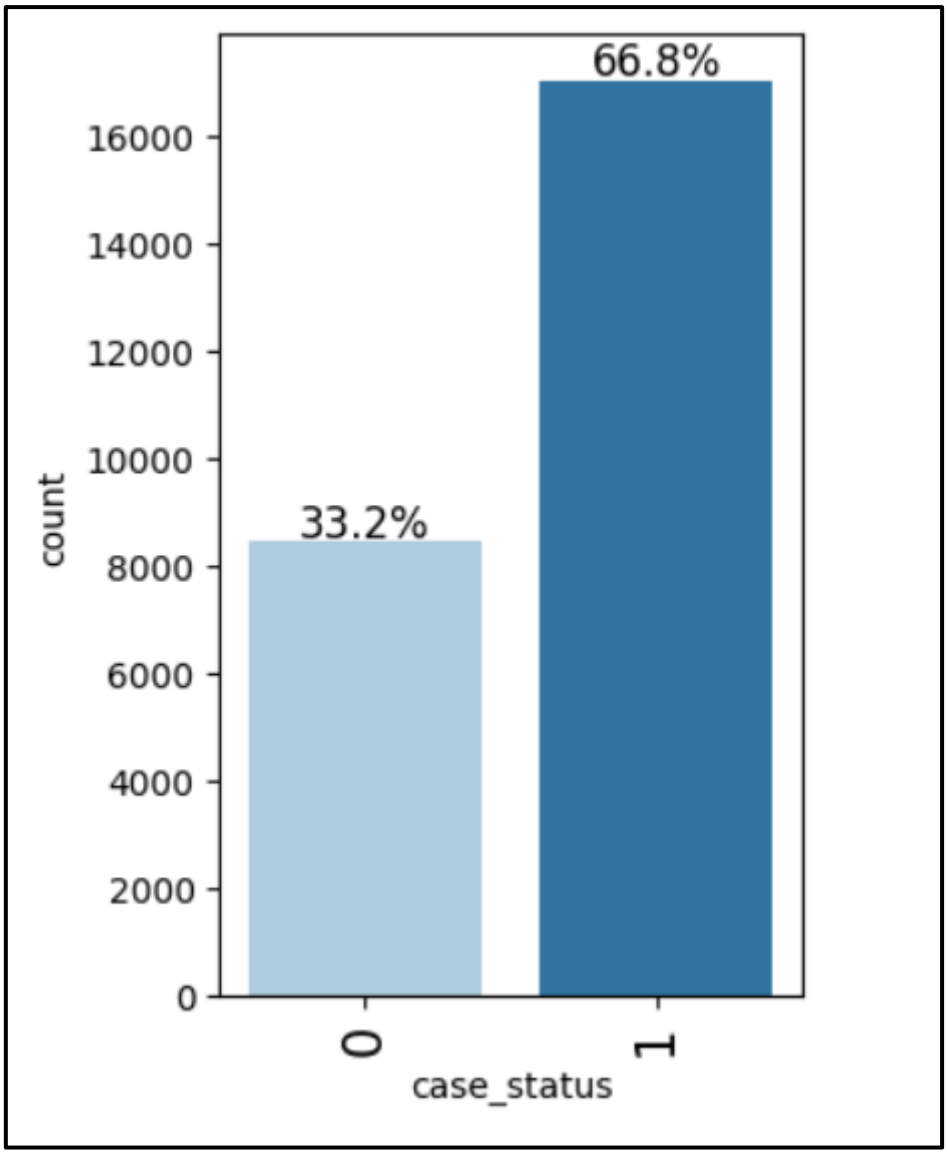


Figure 10: Distribution of data - case_status using barplot

Observation:

- 1 (Certified) represents 66.8% of the total applications.
- 0 (Denied) represents 33.2% of the total applications.
- A significant majority of visa applications are certified, suggesting that more than half of the applications in the dataset are approved.
- However, the denied applications (33.2%) still form a substantial portion, which could be important to address in the model to avoid misclassifications.

3.1.9 Distribution of data - no_of_employees using histogram_boxplot

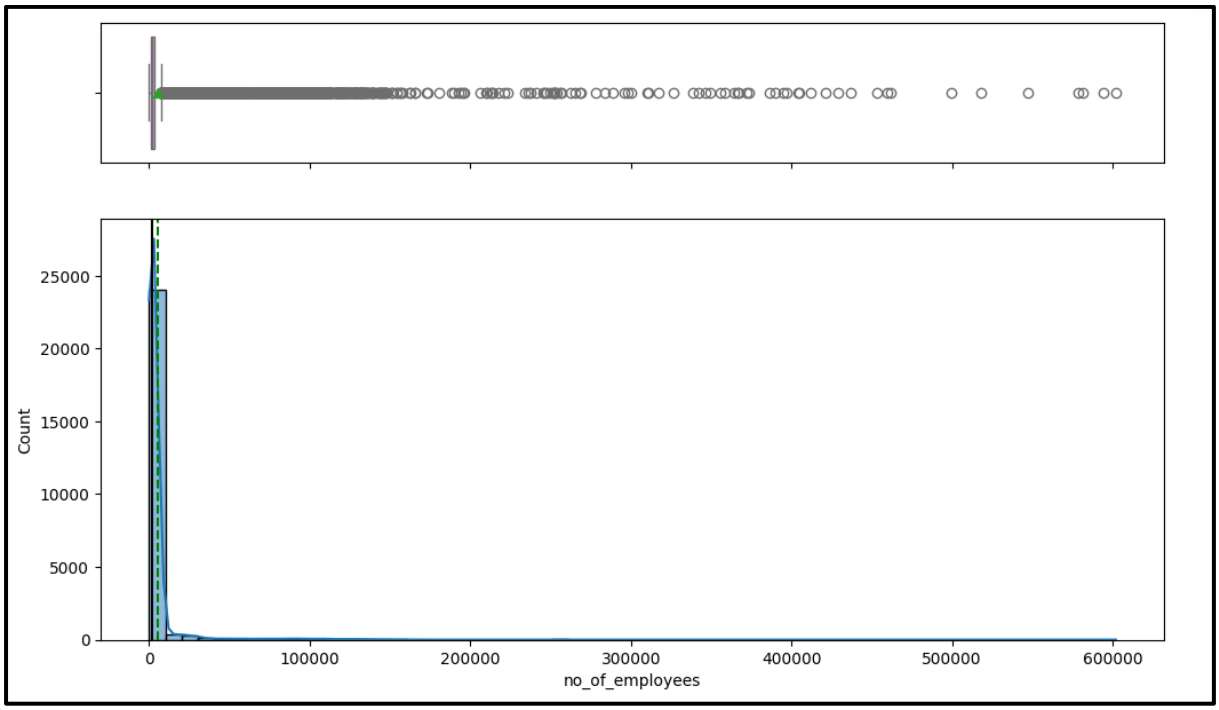


Figure 11: Distribution of data - no_of_employees using histogram_boxplot

Observation:

- The majority of employers have a small number of employees, as shown by the sharp peak on the left side of the distribution, with most values concentrated near zero.
- The upper box plot reveals the presence of extreme outliers, with some companies having over 400,000 employees. However, these outliers are quite rare.
- The data is highly skewed, with most employers having a relatively small workforce compared to a few very large employers.

3.1.10 Distribution of data - yr_of_estab using histogram_boxplot

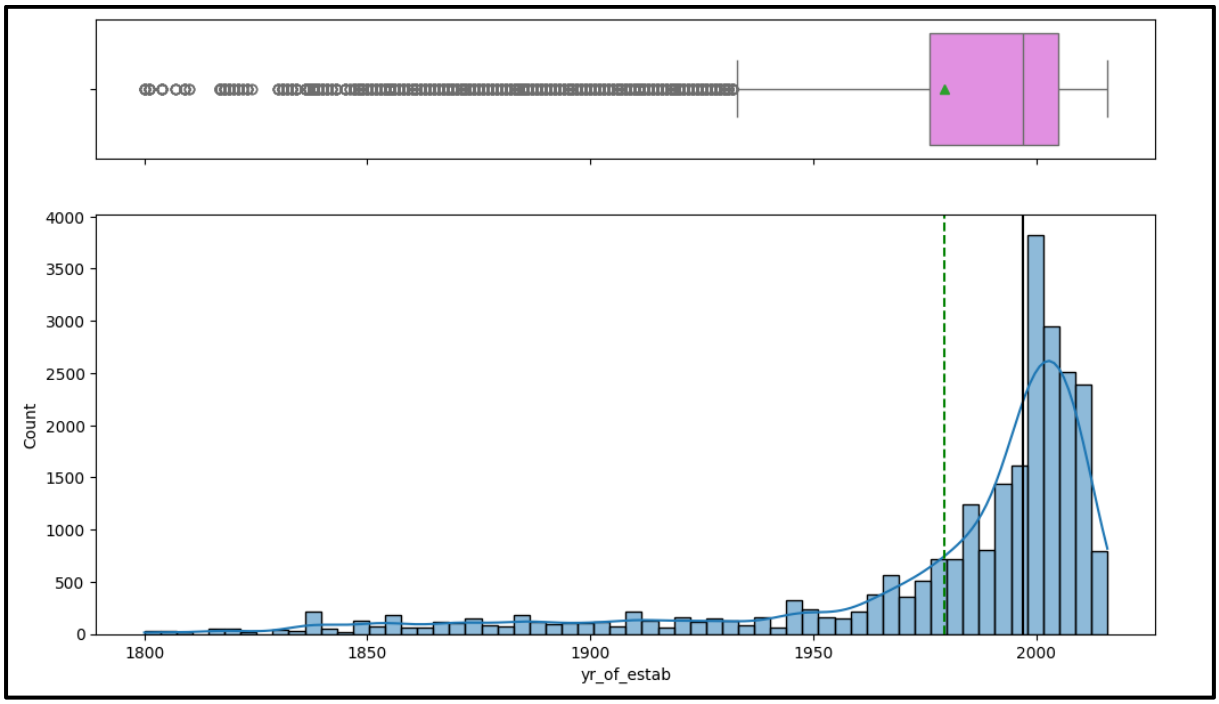


Figure 12: Distribution of data - yr_of_estab using histogram_boxplot

Observation:

- Most establishments were founded in recent years, with a sharp increase around the year 2000.
- There are some extreme outliers with very early or very recent establishment years, but these are rare: The data is highly skewed towards more recent years, indicating a trend of increasing new establishments over time.

3.1.11 Distribution of data - prevailing_wage using histogram_boxplot

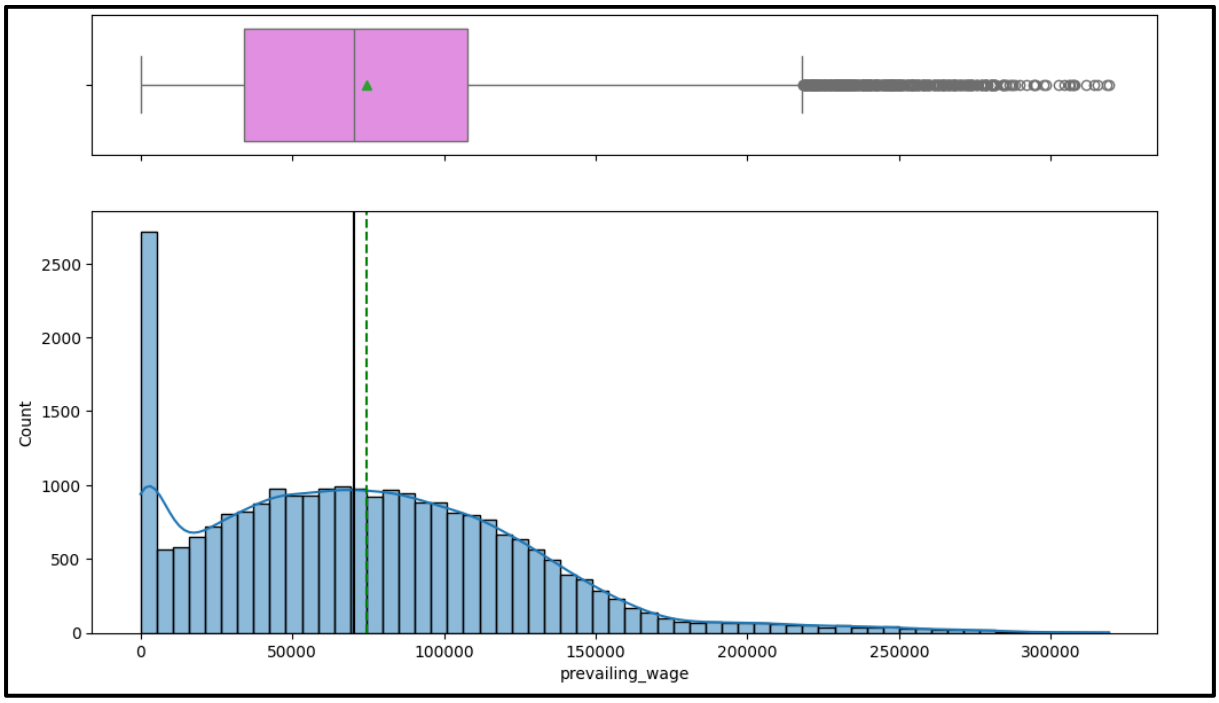


Figure 13: Distribution of data - prevailing_wage using histogram_boxplot

Observation:

- Most prevailing wage values are concentrated on the lower end, indicating that the majority of employees earn lower wages.
- There are extreme outliers with significantly higher wages, but these are rare.* : The data is highly skewed towards lower wages, with a few instances of very high wages.

3.2 Bivariate Analysis

3.2.1 Relation between continent and case_status

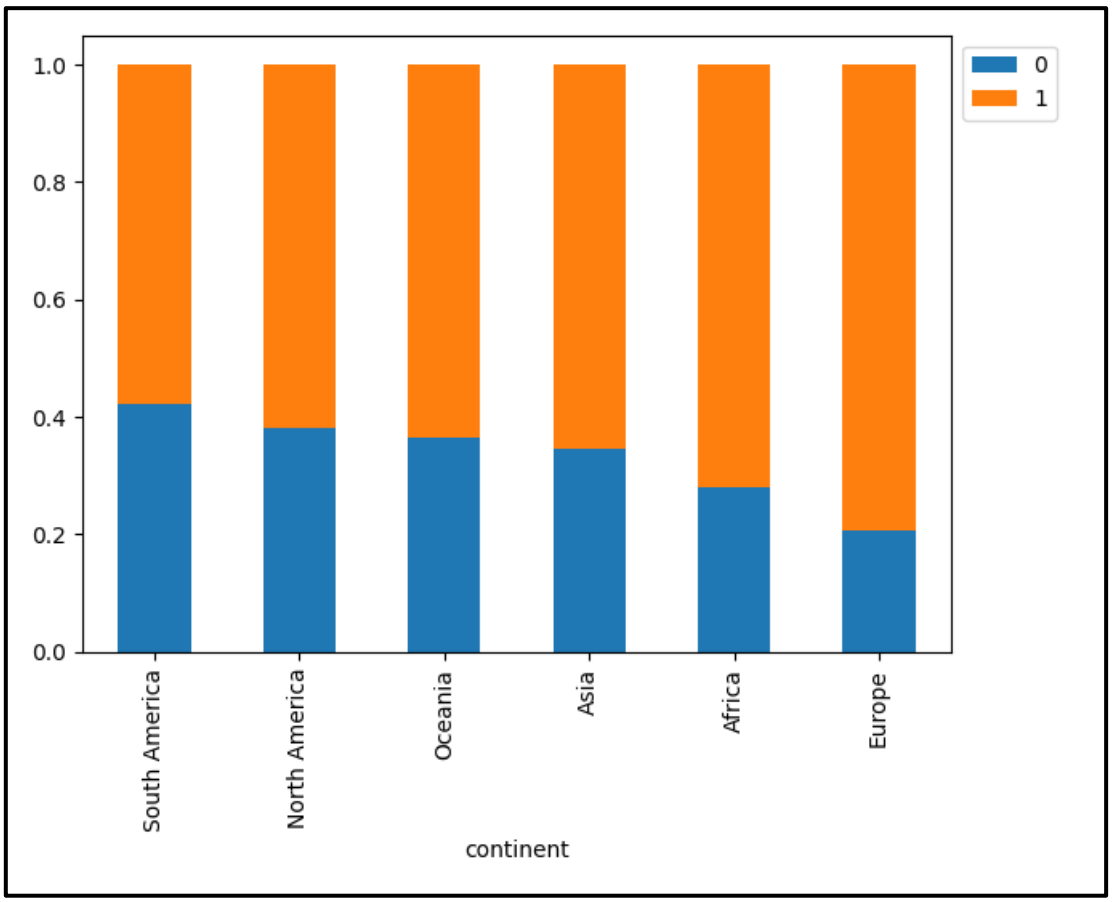


Figure 14: Relation between continent and case_status

Observation:

- Among different continents, Europe has the highest work visa certification rate (79%).
- The lowest work visa certification rate belongs to South America (58%).).

3.2.2 Relation between education_of_employee and case_status

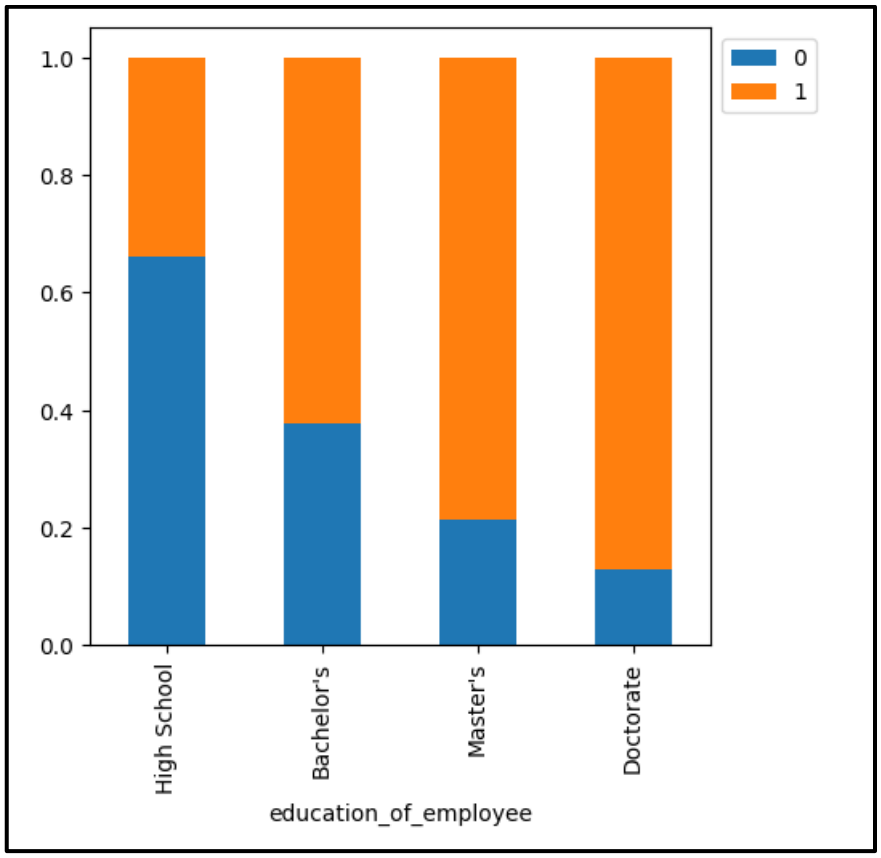


Figure 15: Relation between education_of_employee and case_status

Observation:

- It is clear that the higher the education level of an applicants is, the more their chances of visa certification.
- More specifically, while the visa certification likelihood of the applicants of a doctorate degree is 87%, this likelihood is only 34% for the applicants of high school education.

3.2.3 Relation between has_job_experience and case_status

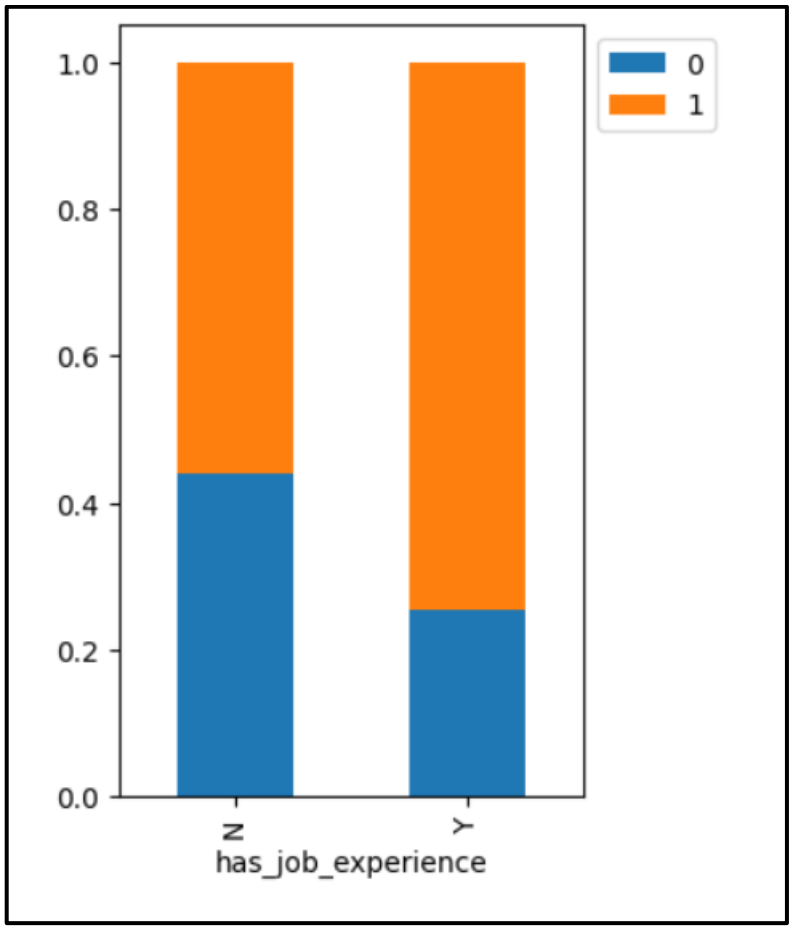


Figure 16: Relation between has_job_experience and case_status

Observation:

- Having job experience is found to have a positive effect on the visa certification likelihood.
- More specifically, about 74% of the experienced applicants are granted visas, while this percentages is only 56% for the inexperienced applicants..

3.2.4 Relation between requires_job_training and case_status

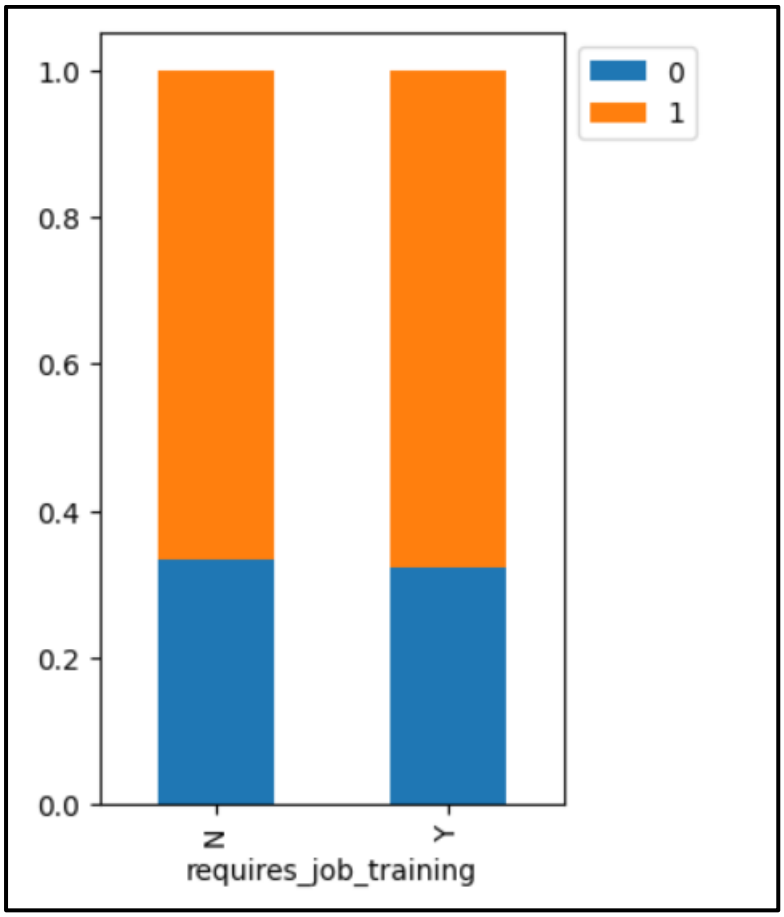


Figure 17: Relation between requires_job_training and case_status

Observation:

- The visa certification likelihood is found nearly unaffected by the job training requirement.

3.2.5 Relation between region_of_employment and case_status

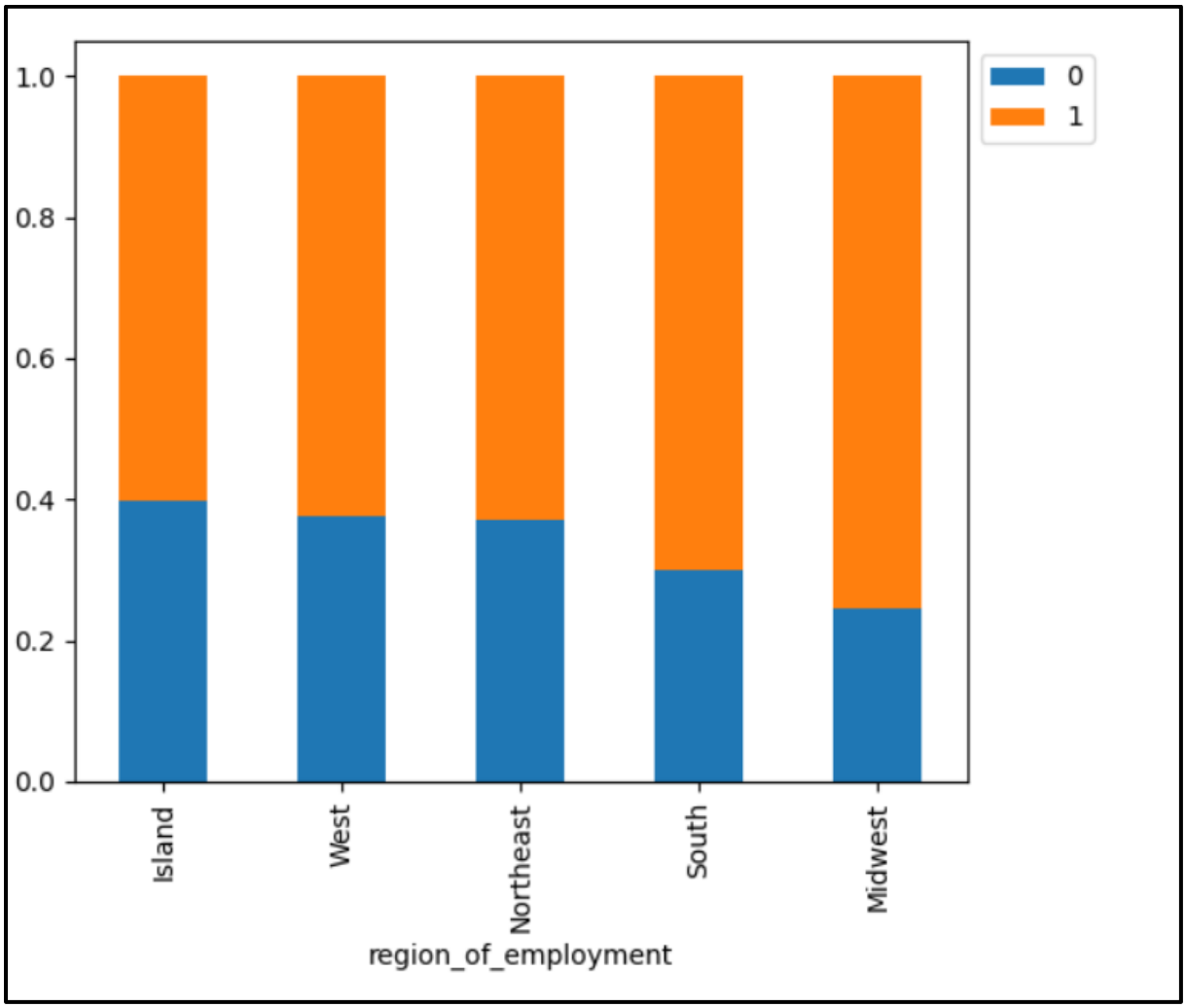


Figure 18: Relation between region_of_employment and case_status

Observation:

- It appears that the visa applications filed by the employers within the Midwest region have the highest probability (~76%) of certification.
- The employers located in the Northeast, West, and Island regions have lower chances (60-63%) of visa certification.

3.2.6 Relation between unit_of_wage and case_status

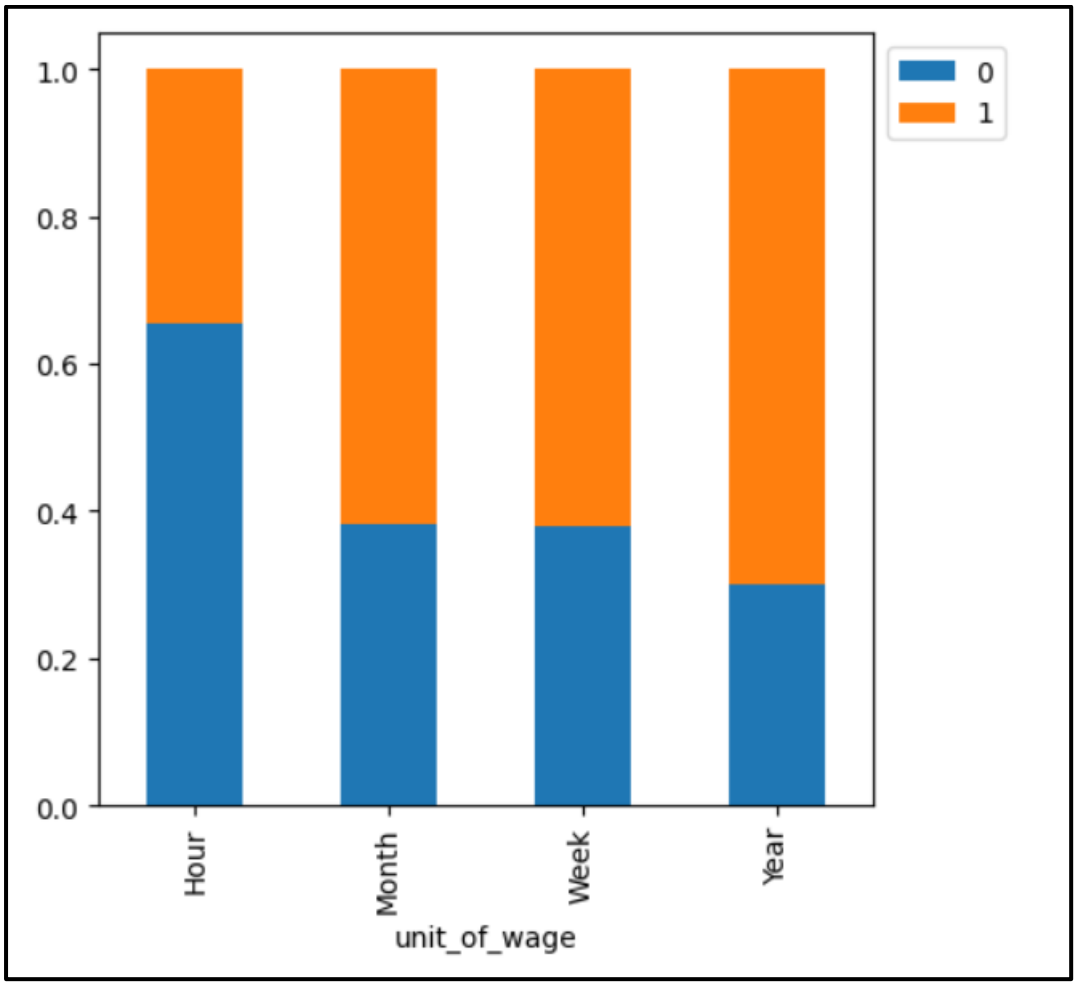


Figure 19: Relation between unit_of_wage and case_status

Observation:

- Those applicants whose wage unit is year are more likely than other applicants to be certified for a visa (~70% likelihood).
- The applicants who are paid by hour are the least likely to be certified for a visa (~35% likelihood). This could be predicted, because hourly jobs are usually less important for the growth of the United States and they could be done by normal American workers.

3.2.7 Relation between full_time_position and case_status

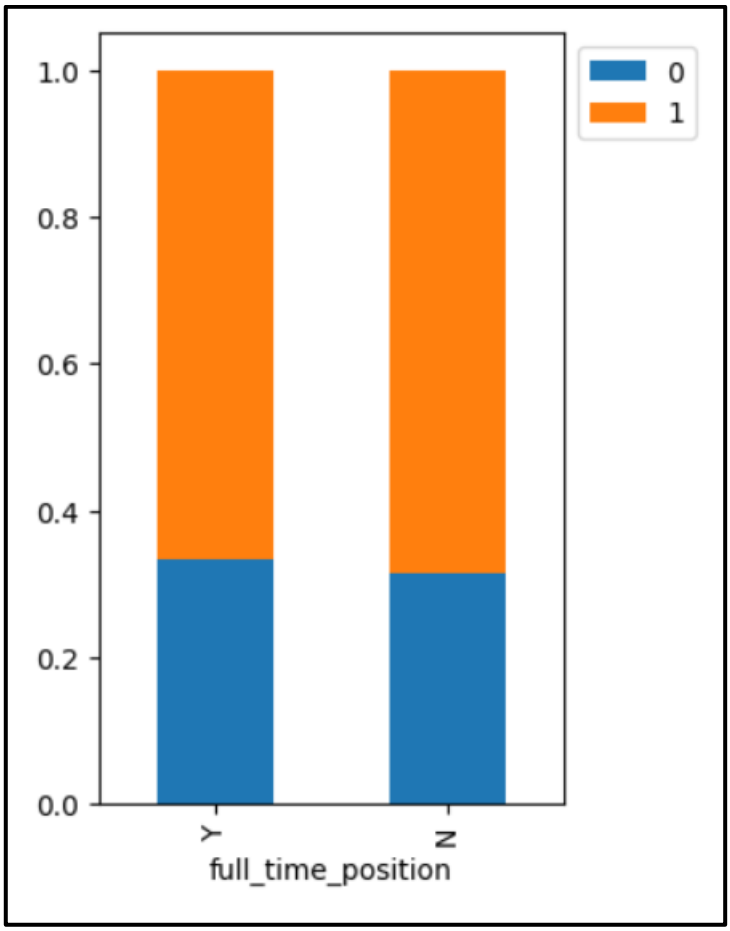


Figure 20: Relation between full_time_position and case_status

Observation:

- Visa certification seems to be unaffected by whether a position is full-time or part-time.

3.3 Multivariate Analysis

3.3.1 Correlation of numeric data using Heatmap

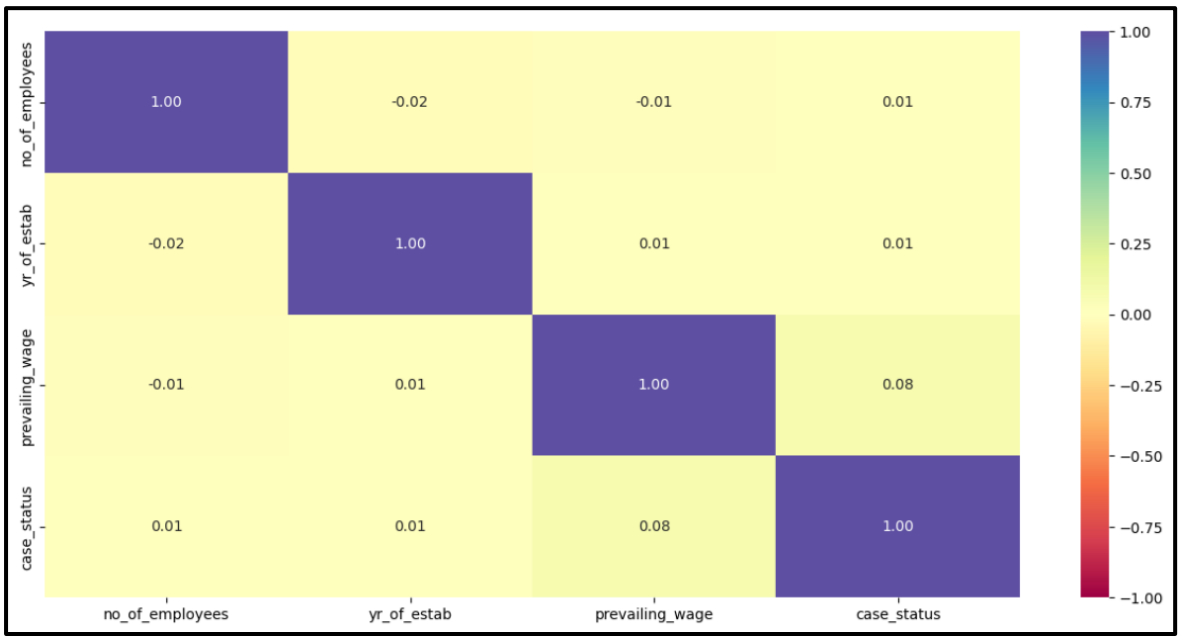


Figure 21: Correlation of numeric data using Heatmap

Observation:

- **Strong Negative Correlation:** 'no_of_employees' and 'case_status' have a strong negative correlation (-0.75), indicating that as the number of employees increases, the likelihood of a certain case status decreases.
- **Positive Correlation:** 'prevailing_wage' and 'case_status' have a slight positive correlation (0.08), indicating a minor relationship where higher wages might be associated with a certain case status.

3.3.2 Relation between Numerical variable using Pairplot

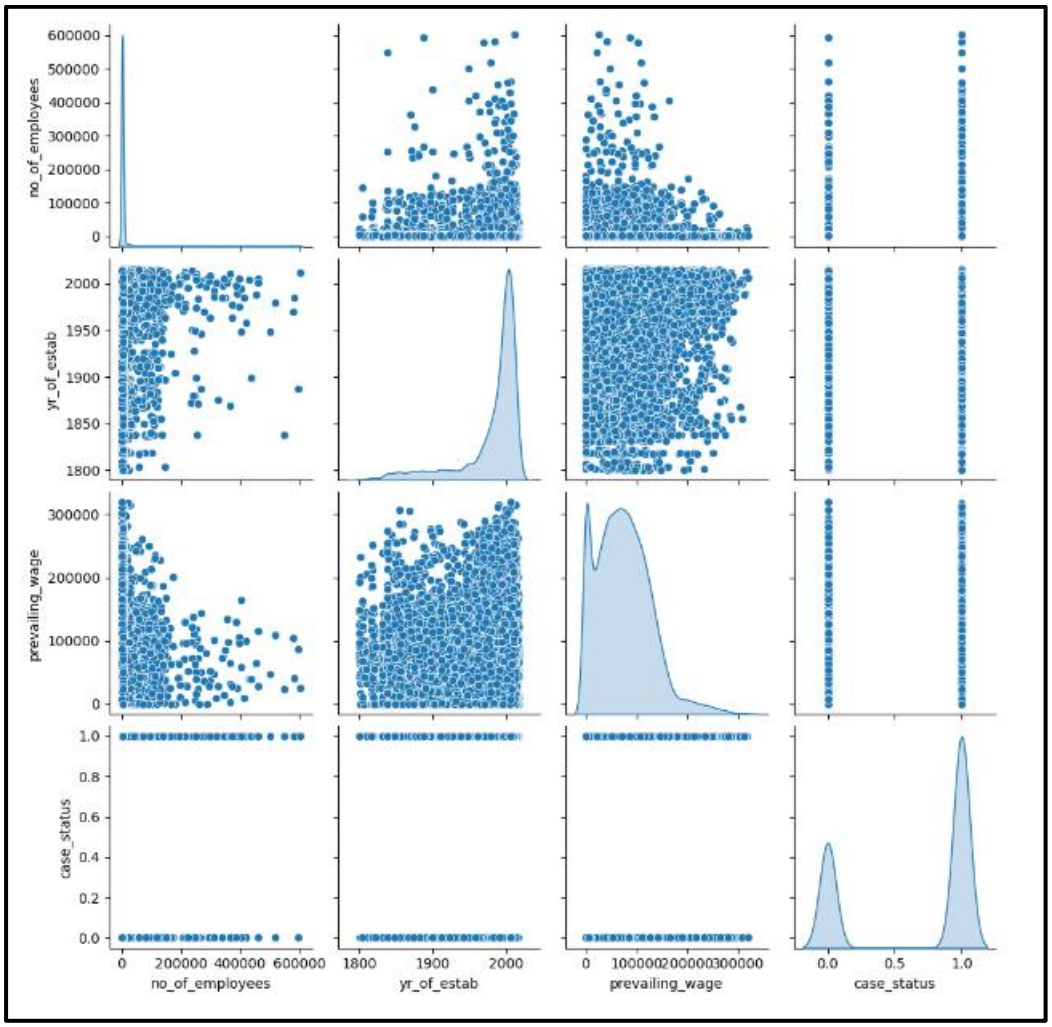


Figure 22: Relation between Numerical variable using Pairplot

Observation:

- **Employee Distribution:** Most companies have a small number of employees, with a few outliers having significantly more.
- **Year of Establishment:** The majority of establishments are more recent, with a concentration around the year 2000
- **Wage Distribution:** Prevailing wages are mostly low, with a few high outlier
- **Case Status:** There is a slight positive correlation between prevailing wage and case status, indicating that higher wages might be associated with certain case statuses.

3.4 Summary of Exploratory Data Analysis

3.4.1 Summary of EDA

- The dataset has 25840 rows and 12 columns
- Only 3 variables are numerical rest all are object types.
- There are no missing data in the data set.

3.4.2 Data Cleaning

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

3.4.3 Observations from Visualization

3.4.3.1 Univariate Analysis

- Continent: Asia dominates with 66.2% of the data, followed by Europe (14.6%) and North America (12.9%).
- education_of_employee: The majority have a Bachelor's (40.2%) or Master's degree (37.8%). A large portion (78%) of the applicants have a bachelor's or a master's degree and only less than 9% have a doctrate degree.
- prevailing_wage: Wages are highly varied, and further exploration through a histogram could reveal if there are any wage clusters.
- full_time_position: 89.4% of jobs are full-time, with the remaining 10.6% being part-time.
- Most (58%) of the applicants have job experience. The vast majority of offered jobs (88%) do not require training.
- The majority (>81%) of the offered jobs are for Northeast, South, and West regions of the US.
- Merely about 10% of the positions have a wage unit other than Year.
- About 2/3 of the work visa applications are certified.

3.4.3.2 Bivariate Analysis

- Jobs requiring job training might prefer candidates with no prior experience.
- Larger companies might offer higher wages, although this could vary depending on the region or industry.
- Higher education levels might correlate with a higher proportion of certified cases.

- The positions with certified visa applications are on average of lower equivalent hourly wages than the positions with denied visa applications.
- Job training requirement has a negligible effect on visa certification likelihood.
- The employer's number of employees has an insignificant impact on the chances of visa certification for its potential foreign employees.
- The offered positions with the wage units of Year and Hour have the highest and the lowest chances of visa certification, respectively.
- Being a full- or part-time position does not observably affect the visa certification likelihood.
- The visa applications for the employment in the Midwest region are more likely to be certified than the applications for the employment in other regions.
- The European and South American applicants have the highest and the lowest chances of visa certification, respectively.
- The higher the applicant's education level is, the more their chances of visa certification are.
- Having job experience increases the chances of visa certification.

3.4.3.3 Multivariate Analysis

- There is no strong correlation either with the target or between any independent variable.

4. Data Preprocessing

4.1 Missing Value Treatment

Based on the initial evaluation, there were no missing values in any of the columns. However, certain rows contained unrealistic non-positive values (i.e., values less than or equal to 0) for the no_of_employees column. To address this issue, these non-positive values were replaced with the median value of no_of_employees.

For missing value imputation, we will perform the imputation process after splitting the data into train, test, and validation sets. This approach ensures that the model does not experience data leakage during training.

Observation:

- There are 33 rows with non-positive number of employees.
- The new minimum number of employees is 12.

4.2 Feature Engineering

- The feature yr_of_estab is converted to yrs_snc_estab, containing the years since establishment.
- The columns yr_of_estab and prevailing_wage are dropped subsequently.

4.2.1 Check statistical summary of numeric data in updated data

	count	mean	std	min	25%	50%	75%	max
no_of_employees	25480.000	5669.798	22877.372	12.000	1028.000	2109.000	3504.000	602069.000
case_status	25480.000	0.668	0.471	0.000	0.000	1.000	1.000	1.000
yrs_snc_estab	25480.000	36.590	42.367	0.000	11.000	19.000	40.000	216.000
hourly_wage	25480.000	94.903	278.177	0.048	22.648	39.827	60.012	7004.399

Figure 23: Check statistical summary of numeric data in updated data

Observation:

- **Outliers:** There are extreme outliers in no_of_employees, yrs_snc_estab, and hourly_wage, which may warrant further investigation or treatment (e.g., capping or transformation).
- **Case Status:** Most cases (66.8%) have been certified, showing a higher success rate for applications.
- **Employee Size and Wage Disparities:** The dataset shows significant variation in the size of organizations and wages, with large organizations potentially skewing the average.

4.2.2 Encoding Categorical Data

- Encoding the values in the columns no_of_employees, yrs_snc_estab, continent, has_job_experience, requires_job_training, full_time_position, region_of_employment, case_status, unit_of_wage and education_of_employee.

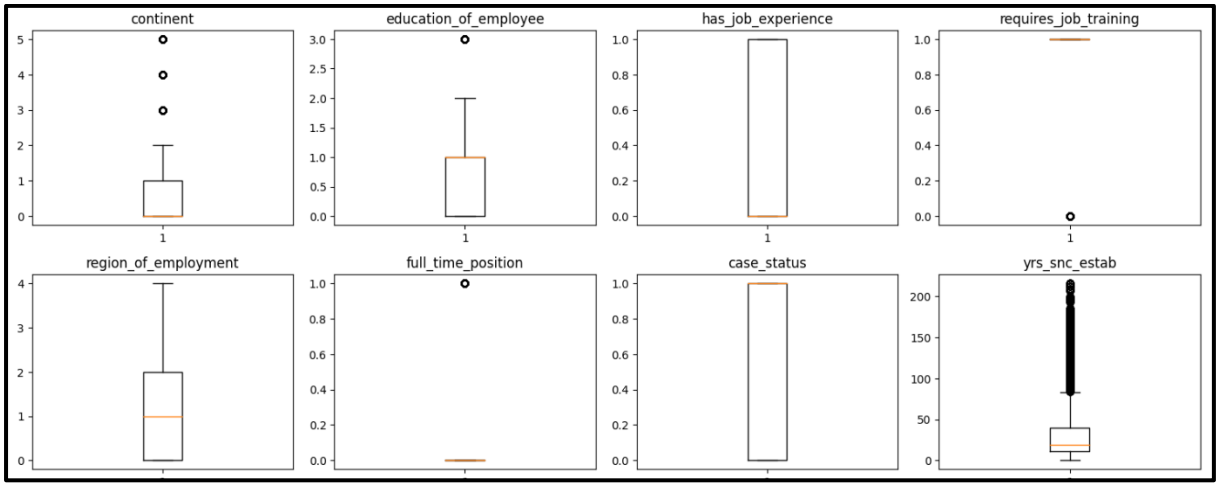
	continent	education_of_employee	has_job_experience	requires_job_training	region_of_employment	unit_of_wage	full_time_position	case_status	yrs_snc_estab	hourly_wage	total_
0	0	2	1	1	2	Hour	0	0	9	592.203	
1	0	1	0	1	0	Year	0	1	14	40.108	
2	0	0	1	0	2	Year	0	0	8	59.133	
3	0	0	1	1	2	Year	0	0	119	40.113	
4	4	1	0	1	1	Year	0	1	11	72.071	

Figure 24: Encoding Categorical Data

Observation:

- Values have been encoded.
- We will do missing value imputation after splitting the data into train, test and validation to avoid data leakage.

4.2.3 Detection and Treatment of Outliers



Observation:

Outliers are there but as they are actual values, we will not treat them and leave them in the dataset.

4.3 Data Preparation for Modeling

4.3.1 Splitting data into training, validation and test set

```
Number of rows in train data = 15288
Number of rows in validation data = 5096
Number of rows in test data = 5096
```

Figure 25: Splitting data into training, validation and test set

4.3.2 Missing Value Treatment

```
continent                0
education_of_employee    0
has_job_experience       0
requires_job_training    0
no_of_employees          0
yr_of_estab              0
region_of_employment     0
prevailing_wage          0
unit_of_wage             0
full_time_position       0
case_status              0
dtype: int64
```

Figure 26: Missing Value Treatment

Observation:

- There are no missing values in the dataset.

4.3.2.1 Checking that no column has missing values in train, validation or test sets

```
continent                0
education_of_employee    0
has_job_experience       0
requires_job_training    0
region_of_employment     0
unit_of_wage             0
full_time_position       0
yrs_snc_estab            0
hourly_wage              0
total_no_of_employees    0
years_of_established     0
dtype: int64
------------------------------
continent                0
education_of_employee    0
has_job_experience       0
requires_job_training    0
region_of_employment     0
unit_of_wage             0
full_time_position       0
yrs_snc_estab            0
hourly_wage              0
total_no_of_employees    0
years_of_established     0
dtype: int64
------------------------------
continent                0
education_of_employee    0
has_job_experience       0
requires_job_training    0
region_of_employment     0
unit_of_wage             0
full_time_position       0
yrs_snc_estab            0
hourly_wage              0
total_no_of_employees    0
years_of_established     0
dtype: int64
```

Figure 27: Checking that no column has missing values in train, validation or test sets

Observation:

- All missing values have been treated.
- Let's inverse map the encoded values.

4.3.3 Reverse Mapping for Encoded Variables

Feature	Categories	Train Set	Validation Set	Test Set
continent	Asia	10885	3395	3381
	Europe	2285	711	734
	North America	1944	655	693
	South America	528	173	151
	Africa	333	121	97
	Oceania	113	39	40
education_of_employee	Bachelor's	6141	2030	2060
	Master's	5792	1886	1956
	High School	2845	698	681
	Doctorate	1118	438	399
has_job_experience	Y	8045	2963	2994
	N	6443	2133	2102
requires_job_training	N	13477	4951	4547
	Y	1811	595	549
region_of_employment	Northeast	4312	1438	1453
	South	4248	1389	1188
	West	3928	1352	1314
	Midwest	2576	855	876
	Island	232	70	73
unit_of_wage	Year	13786	4576	4608
	Hour	1286	452	419
	Week	156	57	59
	Month	68	11	18

Feature	Categories	Train Set	Validation Set	Test Set
full_time_position	Y	13768	4552	4543
	N	1618	544	553
total_no_of_employees	2	7622	2579	2535
	1	3838	1282	1295
	3	2201	1235	1266
	4	885	1243	1277
years_of_established	2	7393	2464	2479
	3	4009	1666	1277
	4	3889	1266	1398

Table 1: Checking inverse mapped values/categories.

Observation:

Inverse mapping returned original labels.

4.3.4 Creating Dummy Variables

```
(15288, 24) (5096, 24) (5096, 24)
```

Figure 28: Creating Dummy Variables

Observation:

- After encoding there are 24 columns.

5. Model Building

Model evaluation criterion

Model can make wrong predictions as:

1. Predicting an applicant visa should be certified but in reality, the applicant visa is denied.
2. Predicting an applicant visa should be denied but in reality, the applicant visa is certified.

Which case is more important?

- **Both are important:** If an applicant is approved when they would have been denied, an unqualified employee will get a job that should have been filled by a US citizen. If an applicant is denied when they should have been approved, U.S. companies will not be able to fill critical positions and the overall economy will not be as productive.

How to reduce this losses?

- As the process of reviewing each application is time and resource-intensive, this model should identify those candidates predicted to be approved, so agents can prioritize these applications.
- F1 Score can be used a the metric for evaluation of the model, as the greater the F1 score, the higher the chances of minimizing False Negatives and False Positives.
- We will use balanced class weights, where applicable, so that model focuses equally on both classes.

Let's define a function to output different metrics (including recall) on the train and test set and a function to show confusion matrix so that we do not have to use the same code repetitively while evaluating models.

5.1 Model Building - Original Data

5.1.1 Training performance and validation performance

Training Performance				
Model	**Recall**	**Precision**	**F1 Score**	**Accuracy**
Bagging	0.986	0.9913	0.9887	0.9849
Random Forest	1	1	1	1
GBM	0.8784	0.7837	0.8283	0.7569
AdaBoost	0.8883	0.7601	0.8192	0.7382
Decision Tree	1	1	1	1

Validation Performance				
Model	**Recall**	**Precision**	**F1 Score**	**Accuracy**
Bagging	0.7709	0.7709	0.7709	0.6939
Random Forest	0.8161	0.7621	0.7882	0.707
GBM	0.8699	0.7817	0.8234	0.7508
AdaBoost	0.8772	0.7598	0.8143	0.7327
Decision Tree	0.738	0.7494	0.7436	0.6601

Table 2: Training performance and validation performance

Observation

- **Overfitting Issues:** The Random Forest and Decision Tree models show signs of overfitting, as evidenced by their perfect training metrics contrasted with much lower validation metrics. This suggests they have learned the training data too well, including noise and outliers.
- **Generalization:** Both GBM and AdaBoost exhibit more consistent performance between training and validation metrics, indicating better generalization to unseen data.
- **Best Model:** While Random Forest had the highest recall on the validation set, GBM and AdaBoost may be more reliable for practical applications due to their balanced performance and lower risk of overfitting.

5.1.2 Training and Validation Performance Difference:

Model	Metric	Training	Validation	Difference
Bagging	Recall	0.986	0.7709	0.2151
	Precision	0.9913	0.7709	0.2205
	F1 Score	0.9887	0.7709	0.2178
	Accuracy	0.9849	0.6939	0.291
Random Forest	Recall	1	0.8161	0.1839
	Precision	1	0.7621	0.2379
	F1 Score	1	0.7882	0.2118
	Accuracy	1	0.707	0.293
GBM	Recall	0.8784	0.8699	0.0085
	Precision	0.7837	0.7817	0.002
	F1 Score	0.8283	0.8234	0.0049
	Accuracy	0.7569	0.7508	0.0061
AdaBoost	Recall	0.8883	0.8772	0.0111
	Precision	0.7601	0.7598	0.0003
	F1 Score	0.8192	0.8143	0.0049
	Accuracy	0.7382	0.7327	0.0054
Decision Tree	Recall	1	0.738	0.262
	Precision	1	0.7494	0.2506
	F1 Score	1	0.7436	0.2564
	Accuracy	1	0.6601	0.3399

Table 3: Training and Validation Performance Difference

Observation:

Generalization Capability: GBM and AdaBoost show strong generalization capabilities with minimal differences in performance metrics between training and validation sets. This suggests they may be better suited for practical applications where the model needs to perform well on new data.

Overfitting Risks: Bagging, Random Forest, and Decision Tree demonstrate considerable overfitting, with substantial performance declines on validation datasets. This highlights the need for strategies to reduce overfitting, such as hyperparameter tuning, cross-validation, or employing regularization techniques.

Model Selection: When choosing a model for deployment, GBM and AdaBoost appear to be the most reliable options due to their consistent performance across training and validation datasets. In contrast, Bagging, Random Forest, and Decision Tree may require further adjustments to improve their generalization abilities before practical use.

5.2 Model Building - Oversampled Data

5.2.1 Training and Validation Performance

Model	Metric	Training	Validation
Bagging	Recall	0.9805	0.7335
	Precision	0.9925	0.7755
	F1 Score	0.9865	0.7539
	Accuracy	0.9865	0.6801
Random Forest	Recall	0.9999	0.7844
	Precision	1	0.7697
	F1 Score	1	0.777
	Accuracy	1	0.6992
GBM	Recall	0.8405	0.8352
	Precision	0.7583	0.7935
	F1 Score	0.7973	0.8138
	Accuracy	0.7863	0.7447
AdaBoost	Recall	0.8391	0.8381
	Precision	0.7406	0.7827
	F1 Score	0.7868	0.8095
	Accuracy	0.7726	0.7365
Decision Tree	Recall	1	0.7083
	Precision	1	0.7407
	F1 Score	1	0.7241
	Accuracy	1	0.6395

Table 4: Training and Validation Performance

Observation:

- **Overfitting Concerns:** Bagging, Random Forest, and Decision Tree models exhibit substantial overfitting, as seen by their high training scores compared to significantly lower validation scores.
- **Generalization Strength:** GBM and AdaBoost appear more robust, with minimal performance differences, suggesting that they can effectively generalize to unseen data.
- **Model Selection:** GBM and AdaBoost should be preferred for practical applications where generalization is critical. In contrast, models like Bagging, Random Forest, and Decision Tree may require further adjustments (e.g., tuning hyperparameters) to improve generalization capabilities.

5.2.2 Training and Validation Performance Difference

Model	Metric	Training	Validation	Difference
Bagging	Recall	0.9805	0.7335	0.247
	Precision	0.9925	0.7755	0.217
	F1 Score	0.9865	0.7539	0.2325
	Accuracy	0.9865	0.6801	0.3064
Random Forest	Recall	0.9999	0.7844	0.2155
	Precision	1	0.7697	0.2303
	F1 Score	1	0.777	0.223
	Accuracy	1	0.6992	0.3008
GBM	Recall	0.8405	0.8352	0.0054
	Precision	0.7583	0.7935	-0.0352
	F1 Score	0.7973	0.8138	-0.0165
	Accuracy	0.7863	0.7447	0.0416
Adaboost	Recall	0.8391	0.8381	0.0009
	Precision	0.7406	0.7827	-0.0421
	F1 Score	0.7868	0.8095	-0.0227
	Accuracy	0.7726	0.7365	0.0361
Decision Tree	Recall	1	0.7083	0.2917
	Precision	1	0.7407	0.2593
	F1 Score	1	0.7241	0.2759
	Accuracy	1	0.6395	0.3605

Table 5: Training and Validation Performance Difference

Observation:

- **Severe Overfitting:** Bagging, Random Forest, and Decision Tree models demonstrate significant overfitting, as indicated by their substantial performance declines in validation metrics.
- **Robust Generalization:** GBM and AdaBoost show strong performance consistency across training and validation datasets, indicating they can generalize better to unseen data.
- **Model Selection:** For practical applications, GBM and AdaBoost should be prioritized due to their balanced performance, while Bagging, Random Forest, and Decision Tree may need tuning to enhance generalization.

5.3 Model Building - Undersampled Data

5.3.1 Training and Validation Performance

Training Performance				
Model	**Recall**	**Precision**	**F1 Score**	**Accuracy**
Bagging	0.9677	0.9933	0.9803	0.9806
Random Forest	1	1	1	1
GBM	0.7475	0.7093	0.7279	0.7206
Adaboost	0.7127	0.6921	0.7022	0.6978
Decision Tree	1	1	1	1

Validation Performance				
Model	**Recall**	**Precision**	**F1 Score**	**Accuracy**
Bagging	0.6063	0.8174	0.6962	0.6466
Random Forest	0.6545	0.809	0.7236	0.666
GBM	0.7312	0.8291	0.7771	0.7198
Adaboost	0.7086	0.8238	0.7618	0.7041
Decision Tree	0.6316	0.772	0.6948	0.6293

Table 6: Training and Validation Performance

Observation

- **Overfitting:** Random Forest and Decision Tree are likely overfitting, given their perfect training metrics. Although they perform reasonably well on validation data, there is a notable difference in their recall and precision.
- **Balanced Generalization:** GBM and Adaboost demonstrate relatively good performance on both training and validation datasets, suggesting that they can generalize better while maintaining performance across metrics.
- **Model Selection:** For practical applications, selecting GBM or Adaboost may be beneficial due to their balanced performance, whereas Random Forest and Decision Tree might require further tuning to enhance generalization.

5.3.2 Training and Validation Performance Difference

Model	Metric	Training	Validation	Difference
Bagging	Recall	0.9677	0.6063	0.3614
	Precision	0.9933	0.8174	0.1759
	F1 Score	0.9803	0.6962	0.2841
	Accuracy	0.9806	0.6466	0.334
Random Forest	Recall	1	0.6545	0.3455
	Precision	1	0.809	0.191
	F1 Score	1	0.7236	0.2764
	Accuracy	1	0.666	0.334
GBM	Recall	0.7475	0.7312	0.0163
	Precision	0.7093	0.8291	-0.1198
	F1 Score	0.7279	0.7771	-0.0492
	Accuracy	0.7206	0.7198	0.0008
Adaboost	Recall	0.7127	0.7086	0.0041
	Precision	0.6921	0.8238	-0.1317
	F1 Score	0.7022	0.7618	-0.0596
	Accuracy	0.6978	0.7041	-0.0063
Decision Tree	Recall	1	0.6316	0.3684
	Precision	1	0.772	0.228
	F1 Score	1	0.6948	0.3052
	Accuracy	1	0.6293	0.3707

Table 7:Training and Validation Performance Difference

Observation:

- **Overfitting:** Both Bagging, Random Forest, and Decision Tree models exhibit signs of overfitting, as evidenced by their high training scores and significantly lower validation metrics. These models may require tuning to reduce complexity and improve generalization.
- **Good Generalization:** GBM and Adaboost demonstrate better balance in performance across training and validation datasets, suggesting that they are more robust and suitable for practical applications.
- **Model Selection Strategy:** Given the findings, prioritizing models like GBM or Adaboost for deployment may yield better outcomes, while further refining Random Forest and Decision Tree is advisable to enhance their generalization capabilities.
    - GBM has the best performance followed by AdaBoost model as per the validation performance
    - After building 15 models, it was observed that both the GBM and Adaboost models, trained on an undersampled dataset, as well as the 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 Hyperparameter Tuning

6.1 Tuning AdaBoostClassifier model with Undersampled data

```
Best parameters are {'n_estimators': 20, 'learning_rate': 0.2, 'base_estimator': DecisionTreeClassifier(max_depth=3, random_state=1)} with CV
f1 score=0.7154
Best recall score: 0.7434
Best precision score: 0.6897
Best f1 score: 0.7154
Best accuracy score: 0.7043
CPU times: total: 1.72 s
Wall time: 14.4 s
```

Figure 29: Tuning AdaBoostClassifier model with Undersampled data

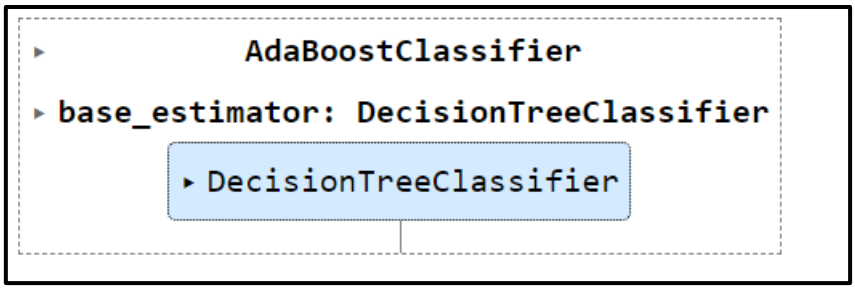


Figure 30: AdaBoostClassifier

Checking model's performance on training set

	Accuracy	Recall	Precision	F1
0	0.702	0.717	0.697	0.707

Figure 31: Checking model's performance on training set

Checking model's performance on validation set

	Accuracy	Recall	Precision	F1
0	0.709	0.714	0.828	0.766

Figure 32: Checking model's performance on validation set

6.2 Tuning Gradient Boosting model with Undersampled Data

```
Best parameters are {'subsample': 0.8, 'n_estimators': 125, 'max_features': 0.7, 'learning_rate': 0.05, 'init': AdaBoostClassifier(random_stat
e=1)} with CV f1 score=0.7140
Best recall score: 0.7359
Best precision score: 0.6935
Best f1 score: 0.7140
Best accuracy score: 0.7053
CPU times: total: 2.41 s
```

Figure 33: Tuning Gradient Boosting model with Undersampled Data

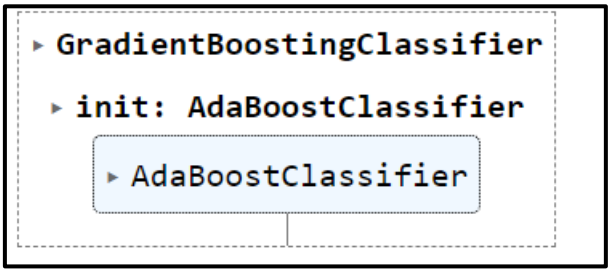


Figure 34: GradientBoostingClassifier

Checking model's performance on training set

	Accuracy	Recall	Precision	F1
0	0.703	0.713	0.699	0.706

Figure 35: Checking model's performance on training set

Checking model's performance on validation set

	Accuracy	Recall	Precision	F1
0	0.707	0.707	0.829	0.764

Figure 36: Checking model's performance on validation set

6.3 Tuning Gradient Boosting model with Oversampled data

```
Best parameters are {'subsample': 0.7, 'n_estimators': 125, 'max_features': 0.7, 'learning_rate': 0.2, 'init': AdaBoostClassifier(random_state
=1)} with CV f1 score=0.7935
Best recall score: 0.8437
Best precision score: 0.7623
Best f1 score: 0.7935
Best accuracy score: 0.7722
CPU times: total: 3.22 s
Wall time: 1min 2s
```

Figure 37: Tuning Gradient Boosting model with Oversampled data

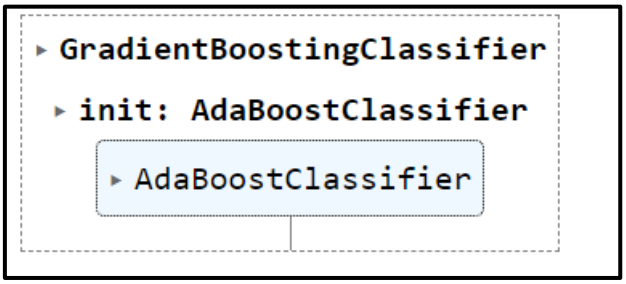


Figure 38: GradientBoostingClassifier

Checking model's performance on training set

	Accuracy	Recall	Precision	F1
0	0.689	0.714	0.680	0.697

Figure 39: Checking model's performance on training set

Checking model's performance on validation set

	Accuracy	Recall	Precision	F1
0	0.707	0.707	0.829	0.764

Figure 40: Checking model's performance on validation set

7. Model Performance Improvement using Hyperparameter Tuning

7.1 Training performance comparison

Training performance comparison:

	Gradient boosting trained with Undersampled data	Gradient boosting trained with Oversampled data	AdaBoost trained with Undersampled data
Accuracy	0.703	0.689	0.702
Recall	0.713	0.714	0.717
Precision	0.699	0.680	0.697
F1	0.706	0.697	0.707

Figure 41: Training performance comparison

7.2 Validation performance comparison

Validation performance comparison:

	Gradient boosting trained with Undersampled data	Gradient boosting trained with Oversampled data	AdaBoost trained with Undersampled data
Accuracy	0.707	0.707	0.709
Recall	0.707	0.707	0.714
Precision	0.829	0.829	0.828
F1	0.764	0.764	0.766

Figure 42: Validation performance comparison

Observation:

- Overall, AdaBoost trained with undersampled data shows slightly superior recall and F1 score during validation, making it a strong candidate for this classification problem.

7.3 Checking the performance on test set

	Accuracy	Recall	Precision	F1
0	0.706	0.726	0.813	0.767

Figure 43: Checking the performance on test set

Observation

- The Adaboost model trained on undersampled data has achieved an F1 score of approximately 77% on the test set.
- This performance is consistent with what we observed during training and validation, indicating that the model is not overfitting and generalizes well.
- Thus, we can conclude that this is a robust and generalized model suitable for predicting visa outcomes.

7.4 Feature Importance

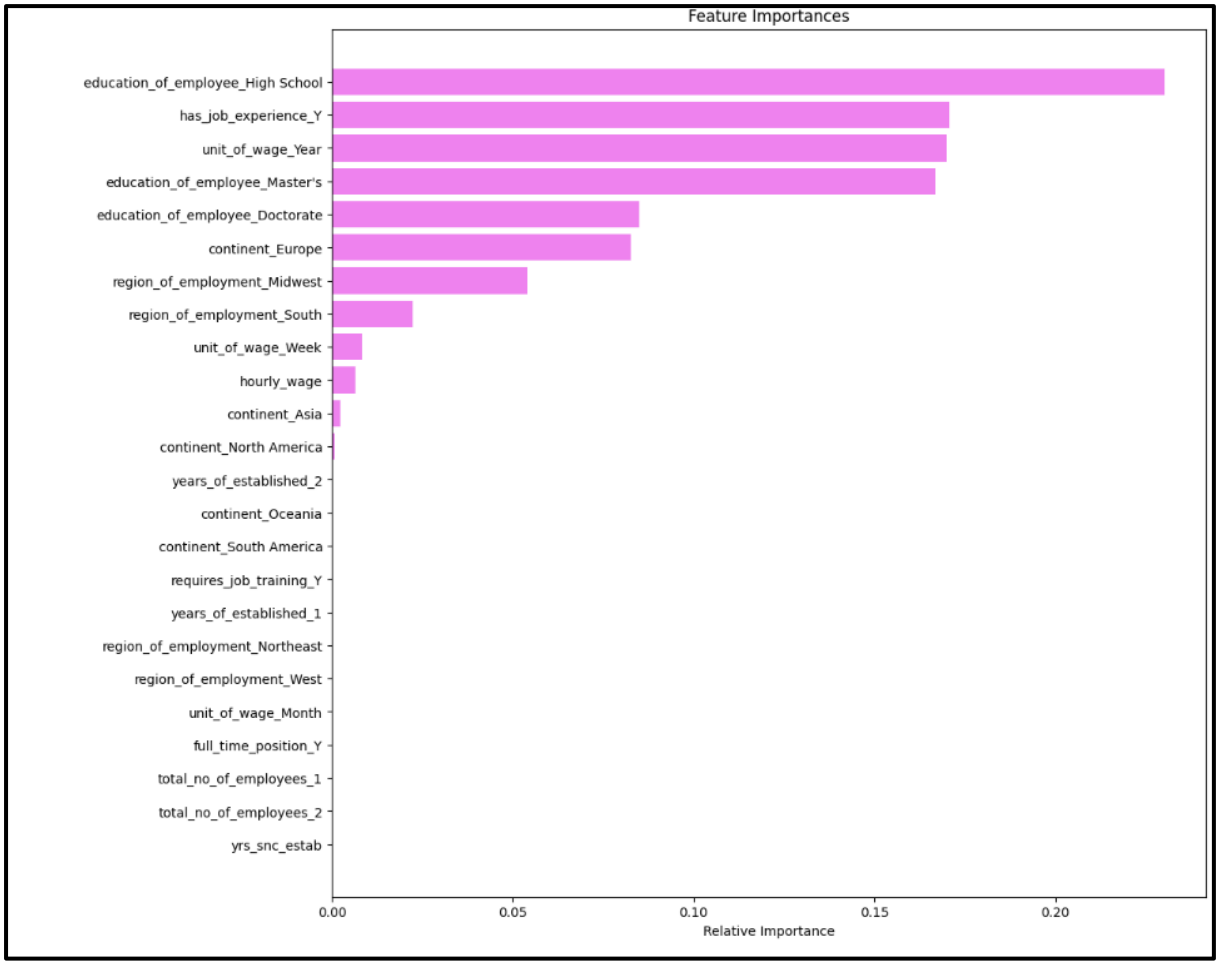


Figure 44: Feature Importance

Observation

- We can see that education of employee high school, has job experience yes, unit of wage year, education of employee master are the most important features for making predictions

8. Actionable Insights & Recommendations

8.1 Actionable Insights

- **Applicant Demographics:** A significant proportion of work visa applications (66%) originate from Asia, highlighting the need for tailored recruitment and outreach strategies to this demographic to attract qualified candidates.
- **Education and Skills:** The data shows that 78% of applicants possess at least a bachelor's or master's degree, with higher educational attainment correlating positively with visa certification rates. This indicates that targeting highly educated individuals can improve overall approval rates and enhance the skill level of the workforce.
- **Experience Matters:** With 58% of applicants having prior job experience, it's evident that experienced candidates are more likely to meet job requirements and have better chances of certification. Companies should prioritize applicants with relevant work history to ensure they meet the job criteria effectively.
- **Job Nature and Certification Rates:** The majority of offered positions (88%) do not require additional training, suggesting that employers prefer candidates who can quickly integrate into their roles. This preference indicates that candidates with a robust skill set and practical experience will likely be favored in the application process.
- **Regional Trends:** Over 81% of offered jobs are concentrated in the Northeast, South, and West regions of the US, with the Midwest showing higher certification rates. Companies should focus their recruitment efforts in these regions to align with job availability and certification likelihood.
- **Wage Structure Implications:** Positions with annual wage units have a higher chance of certification than those with hourly wages. This suggests that companies offering annual salaries may be more appealing to skilled foreign workers, thereby improving their chances of obtaining work visas.
- **Certification Disparities by Region:** The analysis indicates that applicants from different regions have varying certification rates. European applicants exhibit the highest likelihood of certification, while South American applicants face the most challenges. This disparity suggests that regional factors should be considered in recruitment and application strategies.

- **Impact of Employer Characteristics:** The age and size of the sponsoring employer appear to have negligible effects on visa certification likelihood. This insight suggests that other factors, such as job relevance and applicant qualifications, are more critical in the certification process.
- **Wage Trends:** Certified positions typically offer lower average equivalent hourly wages compared to denied applications. This trend could imply that competitive wage offerings might influence certification outcomes positively.
- **Long-Term Workforce Planning:** Given the importance of experience and education, companies can strategically invest in developing partnerships with educational institutions and training programs to build a pipeline of qualified candidates, ensuring that they meet future labor demands effectively.

8.2 Recommendations

- **Model Selection:** Given its simplicity and interpretability, the Tuned Decision Tree model is recommended to the Office of Foreign Labor Certification (OFLC) as the final classifier. If an ensemble approach is preferred, the Tuned Gradient Boosting model is recommended to reduce bias.
- **Focus on Key Factors:** The OFLC should prioritize applicants based on their education level, job experience, and wage unit when estimating visa certification probabilities. Applicants with higher education, relevant job experience, and annual wage units are more likely to receive certifications. Additionally, applicants from Europe have increased chances of certification in certain contexts.
- **Prioritize High-Demand Applications:** To avoid workforce shortages in the US, especially in industries reliant on foreign employees, it is recommended that the OFLC expedite processing for visa applications that demonstrate a higher likelihood of certification based on the developed classification models.
- **Streamline Denial Processes:** To minimize resource waste, the OFLC could rapidly deny applications predicted to have a very high chance of denial. Such applications could be redirected to a different section for appeal processes initiated by applicants or employers.
- **Consider Additional Variables:** It is recommended to explore other potentially important variables in future classification models, including the applicant's industry of employment (e.g., medical, engineering, finance, agriculture), years of experience, alignment of qualifications with job requirements, and the socioeconomic benefits of the employer to the US.
- **Experiment with Advanced Models:** Exploring more sophisticated machine learning-based classification models is encouraged to enhance predictive accuracy and insights into visa certification processes.

8.3 Conclusion

By implementing these insights and recommendations, the OFLC can improve its decision-making processes, ensuring a more efficient and equitable visa certification system while addressing labor needs within the United States.