Business Report - 5

PG Program in Data Science and Business Analytics

submitted by

Sangram Keshari Patro

BATCH:PGPDSBA.O.AUG24.B

Contents

List of Figures

List of Tables

1 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:

- Facilitate the process of visa approvals.
- 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.

2 Data Description

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

2.1 Data dictionary

- **case_id:** ID of each visa application.
- **continent:** Information about the continent where the employee belongs.
- **education_of_employee:** Information about the education level of the employee.
- **has_job_experience:** Indicates if the employee has any job experience. **Y = Yes**, **N = No**.
- **requires_job_training:** Indicates if the employee requires 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 about the 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 the prevailing wage. Values include **Hourly**, **Weekly**, **Monthly**, and **Yearly**.
- **full_time_position:** Indicates if the position of work is full-time. **Y = Full-Time Position**, **N = Part-Time Position**.
- **case_status:** Flag indicating if the visa was **certified** or **denied**.

3 Data Overview

3.1 Importing necessary libraries and the dataset

The dataset is printed. It has 25480 rows & 12 columns.

3.2 Structure and type of data

Data is explored further. The dataset is free from duplicate rows and contains no null values.

```
 #   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: Table depicting the datatype and Non-Null values in each column.

3.3 Statistical summary

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

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

Figure 2: Statistical summary of the data

4 Exploratory Data Analysis

4.1 Univariate Analysis

4.1.1 Numerical columns

- **'yr_of_es**

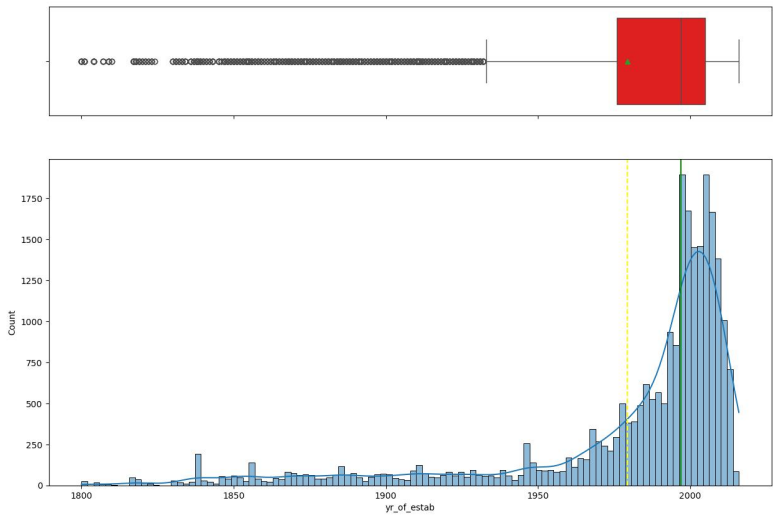


Figure 3: Histogram and boxplot of **'yr_of_estab'** column

Observations

- **Histogram**: The year of establishment data shows an increasing trend over time, with the highest concentration observed in the late 20th and early 21st centuries. There are noticeable spikes in specific historical periods, indicating possibly significant events or business booms.
- **Box Plot**: The median year of establishment lies within the interquartile range (IQR), with the mean close to the median. The data includes several older establishments, visible as outliers extending below the whiskers.

Business Recommendations

- **Focus on Modern Businesses**: Tailor services or products for businesses established in recent decades, as they constitute the majority of the data.
- **Support for Legacy Businesses**: Create specialized offerings for older establishments to support their sustainability and modernization needs.
- **Industry Analysis**: Examine industry-specific trends during peak establishment periods to identify key factors driving business growth.
  - **'prevailing_wage'**

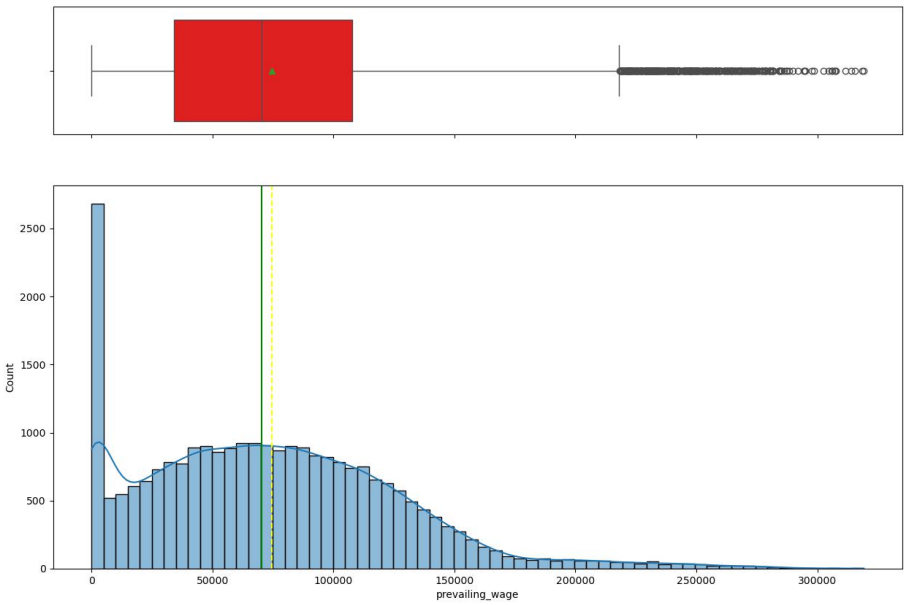


Figure 4: Histogram and boxplot of **'prevailing_wage'** column

Observations

- **Histogram**: The majority of the data is concentrated near lower prevailing wage values, with a significant spike at 0. The distribution shows a long right tail, indicating a presence of higher wages as outliers.
- **Box Plot**: The median prevailing wage is well-defined and lies within the interquartile range (IQR), while the mean appears close to the median. There are several high-wage outliers extending far beyond the whiskers of the box plot.

Business Recommendations

- **Targeted Wage Analysis**: Investigate the reasons for a large proportion of individuals with 0 prevailing wage and address potential data quality or wage disparity issues.
- **Support for High Earners**: Develop specialized services or products for individuals in the high-wage category to capture their market potential.
- **Policy Review**: Assess wage policies to ensure fairness and address discrepancies visible in the data.
- **Tailored Training Programs**: Provide skill development and training opportunities for individuals in lower wage brackets to improve their earning potential.

4.1.2 Categorical columns

- **'continent'**

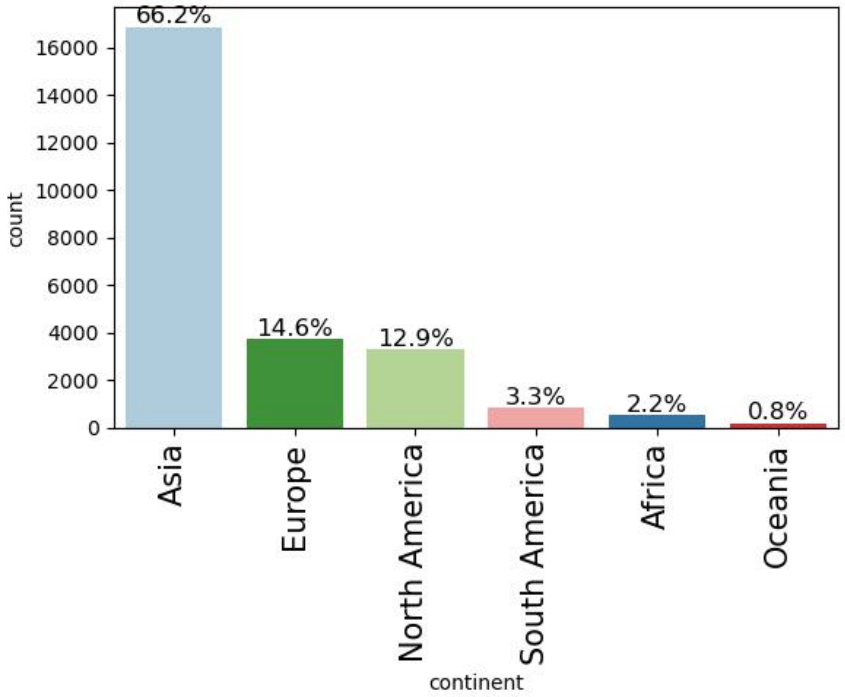


Figure 5: **'continent'** column

Observations

- Asia has the highest applications, followed by Europe and North America.
- South America, Africa, and Oceania have fewer applicants.

Business Recommendations

- Streamline processes for Asia's large talent pool.
- Boost recruitment from Europe and North America.
- Promote awareness in South America, Africa, and Oceania.
- Use ML models to improve visa approval efficiency.

– **'education_of_employee'**

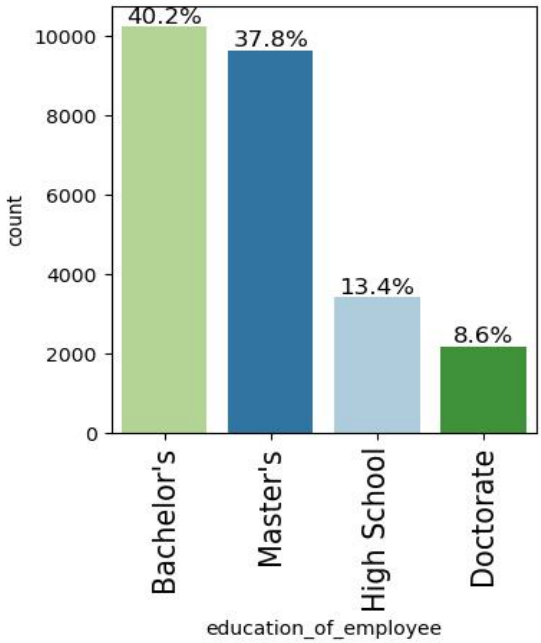


Figure 6: **'education_of_employee'** column

Observations

- Bachelor's degree holders have the highest count, followed by Master's degree holders.
- High School graduates form a smaller share.
- Doctorate holders have the least representation.

Business Recommendations

- Focus on Bachelor's and Master's degree holders as they dominate the applicant pool.
- Encourage applications from High School graduates and Doctorate holders to increase diversity.
- Use ML models to identify key factors for visa approval across education levels.

– **'has_job_experience'**

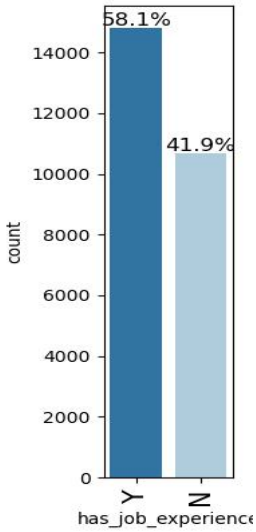


Figure 7: **'has_job_experience'** column

Observations

- 58.1% of applicants have prior job experience, while 41.9% do not.
- Applicants with job experience form the majority.

Business Recommendations

- Prioritize candidates with job experience as they form the largest group.
- Analyze the impact of job experience on visa approval rates using machine learning models.
- Develop targeted training or support programs for applicants lacking job experience to improve their success rates.

- **'requires_job_training'**

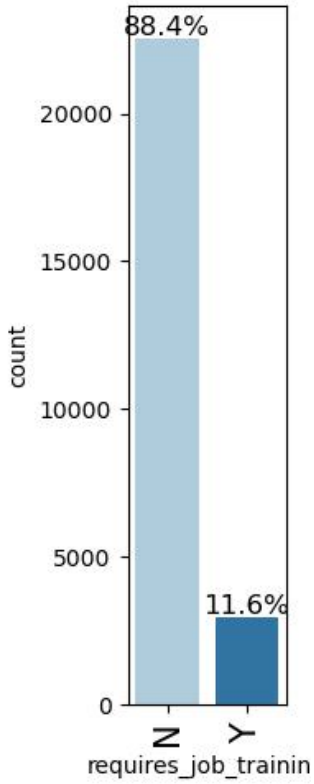


Figure 8: **'requires_job_training'** column

Observations

- 88.4% of applicants do not require job training, while 11.6% do.
- The majority of applicants are already trained for their roles.

Business Recommendations

- Focus on applicants who do not require job training as they dominate the dataset.
- Investigate whether requiring job training impacts visa approval rates.
- Offer specialized training programs for the minority group to increase their competitiveness.

– **'region_of_employment'**

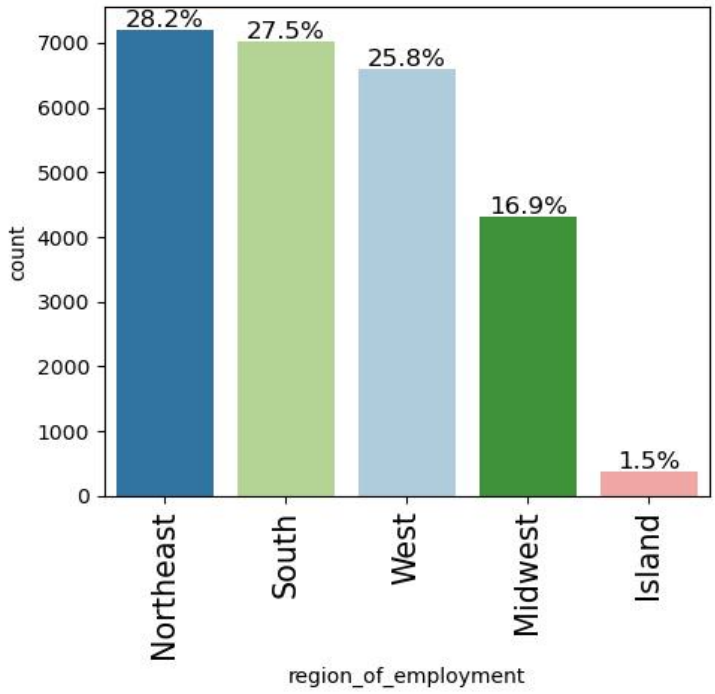


Figure 9: **'region_of_employment'** column

Observations

– The Northeast has the highest count, followed by the South and the West.
– The Midwest has a smaller share.
– The Island region has the least representation.

Business Recommendations

– Focus on the Northeast, South, and West as they dominate employment regions.
– Develop strategies to boost representation in the Midwest and Island regions.
– Use ML models to analyze regional factors influencing visa approvals.

– **'unit_of_wage'**

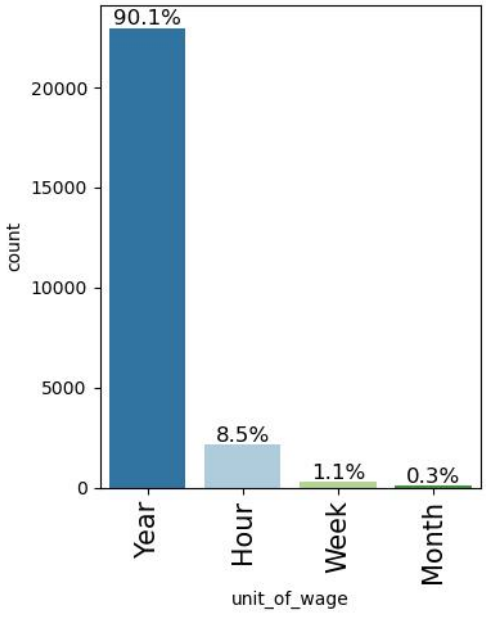


Figure 10: **'unit_of_wage'** column

Observations

- A significant majority (90.1%) of wages are reported on a yearly basis.
- Hourly wages account for 8.5%, while weekly wages are only 1.1%.
- Monthly wages make up a negligible share (0.3%).

Business Recommendations

- Focus on analyzing yearly wages as they dominate the data.
- Consider strategies for applicants with hourly wages, as they represent a substantial minority.
- Allocate minimal resources to monthly and weekly wage units due to their low representation.
- **'full_time_position'**

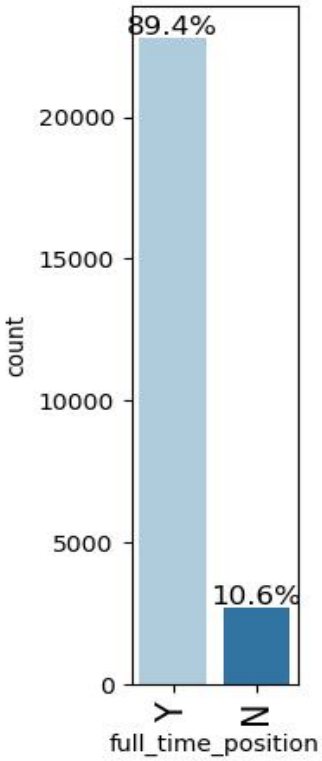


Figure 11: **'full_time_position'** column

Observations

- A majority (89.4%) of positions are full-time (Y).
- Part-time positions (N) constitute only 10.6%.

Business Recommendations

- Focus on full-time positions as they dominate the data.
- Consider developing strategies for applicants in part-time positions to improve their visa approval chances.
- Analyze the factors influencing the smaller share of part-time positions to identify potential opportunities.

– **'case_status'**

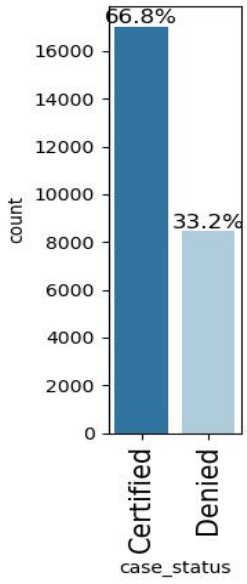


Figure 12: **'case_status'** column

Observations

- A majority (66.8%) of visa applications are certified.
- About one-third (33.2%) of visa applications are denied.

Business Recommendations

- Focus on factors contributing to certification to maintain and improve the approval rate.
- Analyze denied cases to identify common reasons and develop strategies to address them.
- Utilize machine learning models to predict case outcomes and streamline the application process.

4.2 Bivariate Analysis

4.2.1 Numerical variables

- **Heatmap**

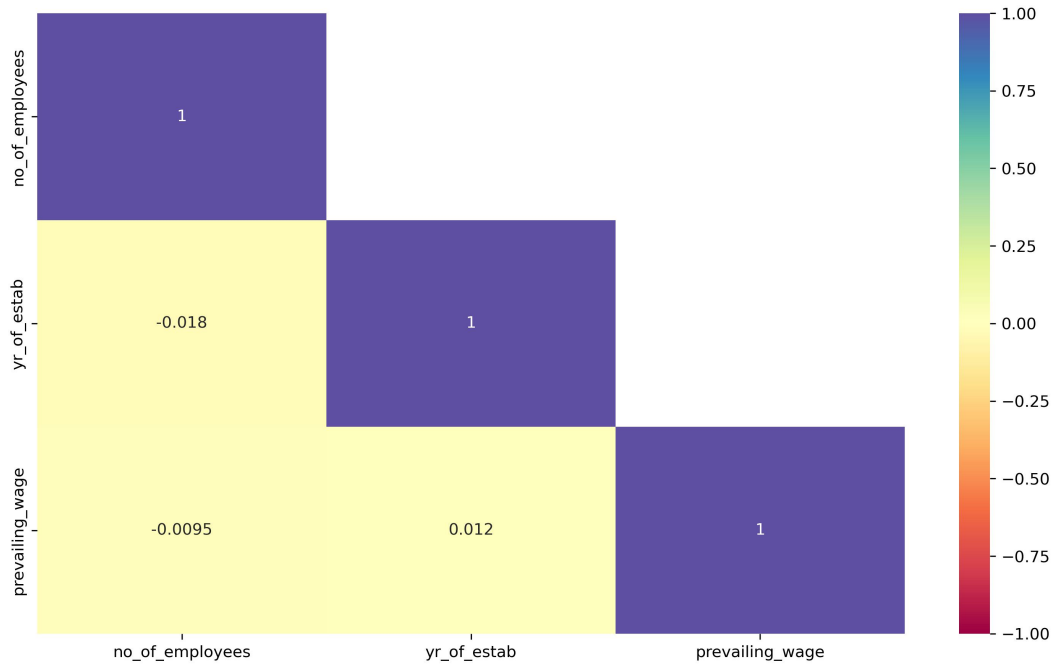


Figure 13: Heatmap of all numerical variables

Observations

- Correlations among `no_of_employees`, `yr_of_estab`, and `prevailing_wage` are negligible.
- No significant linear relationships are observed.

Business Recommendations

- Focus on non-linear relationships and other predictive features.
- Use advanced ML models for complex feature interactions.
- **Pairplot**

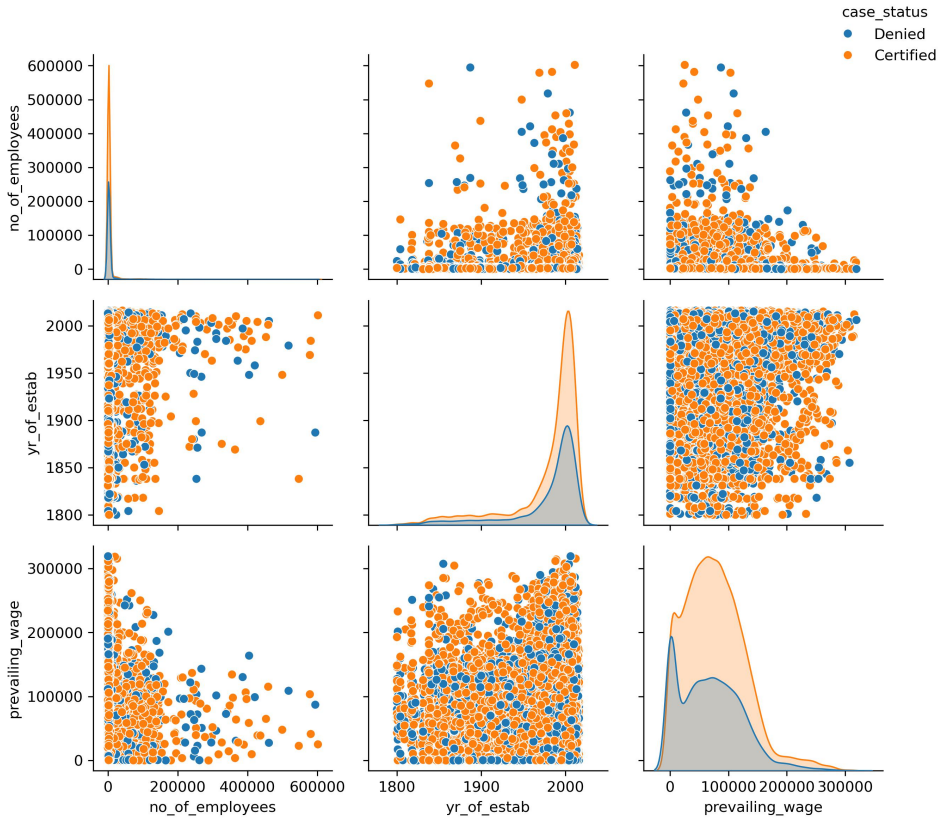


Figure 14: Pairplot of all numerical variables

Observations

- Most companies have fewer than 100,000 employees, with a sharp decline in frequency above this threshold.
- The establishment years show a peak around 1990-2000, with fewer older companies.
- Prevailing wages cluster between $50,000-$150,000, with outliers reaching $300,000+.
- Certified cases (orange) slightly outnumber denied cases (blue) across most metrics.
- Companies established before 1950 show lower visa certification rates.

Business Recommendations

- Target companies established after 1950, as they show higher certification rates.
- Focus on applications with prevailing wages between $50,000-$150,000, as these show balanced approval rates.

- Implement additional scrutiny for applications from very old companies (pre-1850).
- Create separate evaluation frameworks for high-wage applications ($200,000+) vs. standard wages.
- Use company size as a key factor in the ML model, as it correlates with both wages and approval rates.

4.2.2 Categorical vs numerical variables

- **'no_of_employees' and 'prevailing_wage' vs 'yr_of_estab'**

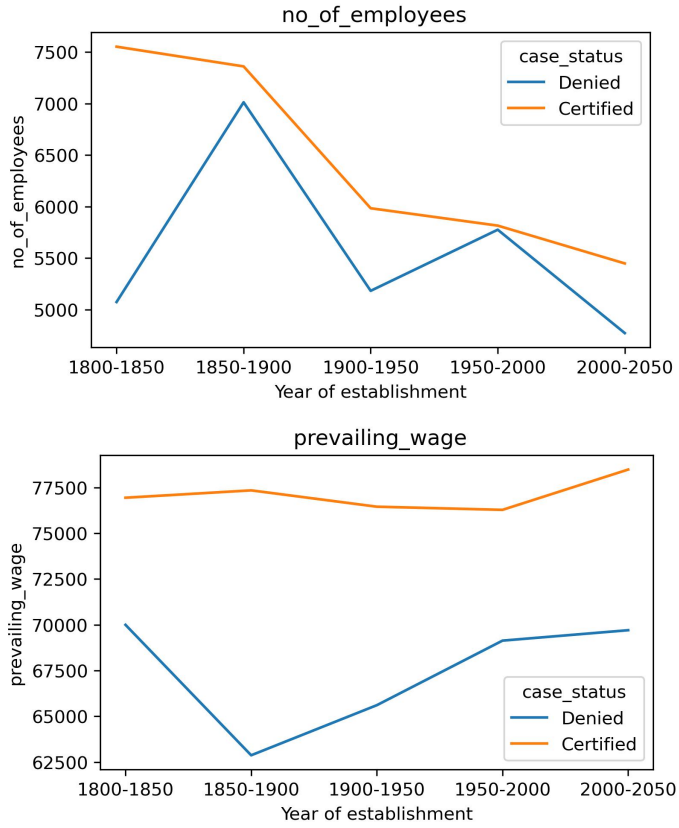


Figure 15: **'no_of_employees' and 'prevailing_wage' vs 'yr_of_estab'**

Observations and Recommendations

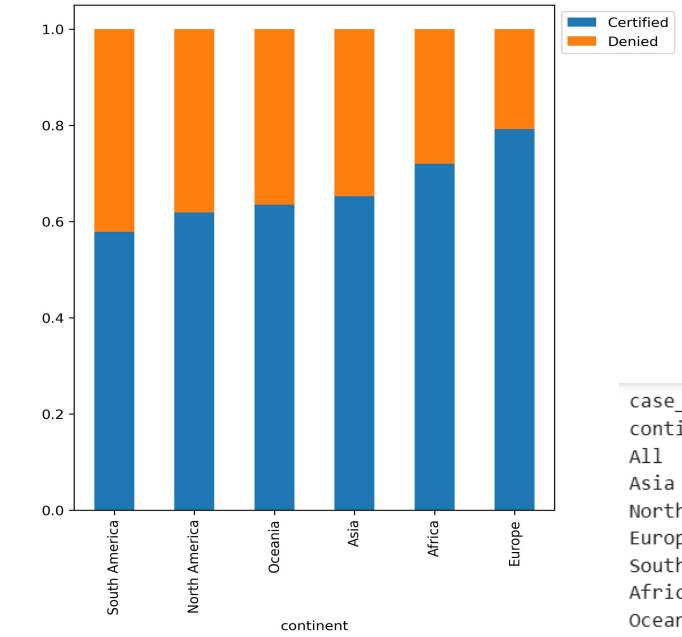
Observations:

- **No. of Employees:**
  - Peak employee count in 1850–1900; decline post-1900.
  - Certified cases consistently exceed denied cases, especially after 2000.
- **Prevailing Wage:**
  - Certified cases have higher wages throughout.
  - Denied cases show a wage dip in 1850–1900; recovery after 1900.
  - Certified wages peak in 2000–2050.

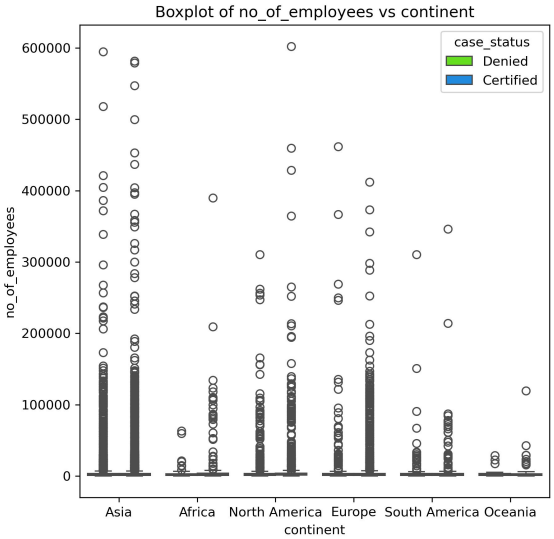
Recommendations:

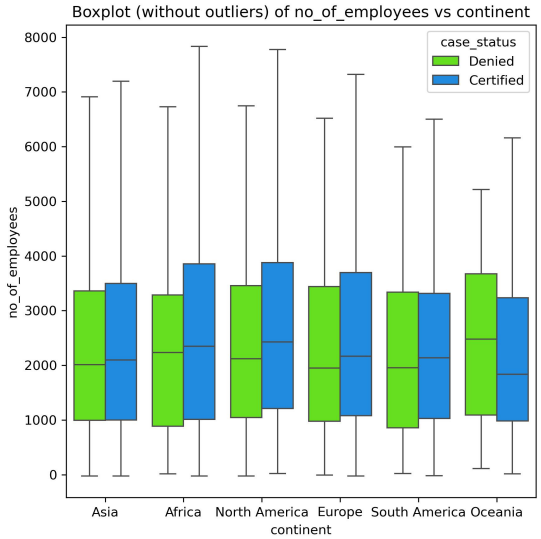
- Offer wages above prevailing standards to improve certification rates.
- Analyze denial patterns for companies established in 1850–1900 and improve compliance.
- Focus resources on regions and sectors with higher certification success rates.
- Adopt dynamic wage models to match market trends and attract talent.

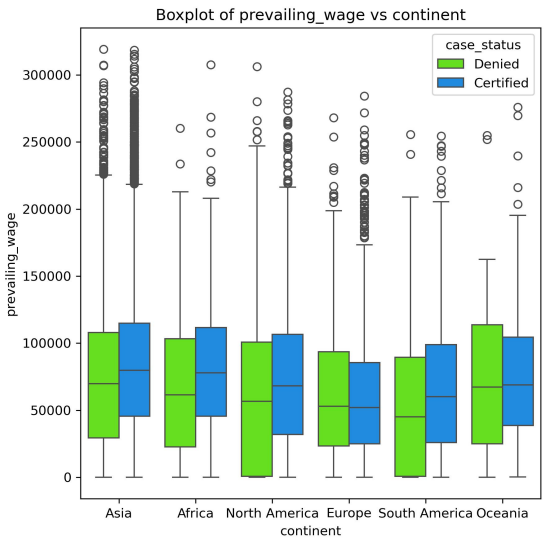
– **'no_of_employees' and 'prevailing_wage' vs 'continent'**



case_status	Certified	Denied	All
continent			
All	17018	8462	25480
Asia	11012	5849	16861
North America	2037	1255	3292
Europe	2957	775	3732
South America	493	359	852
Africa	397	154	551
Oceania	122	70	192







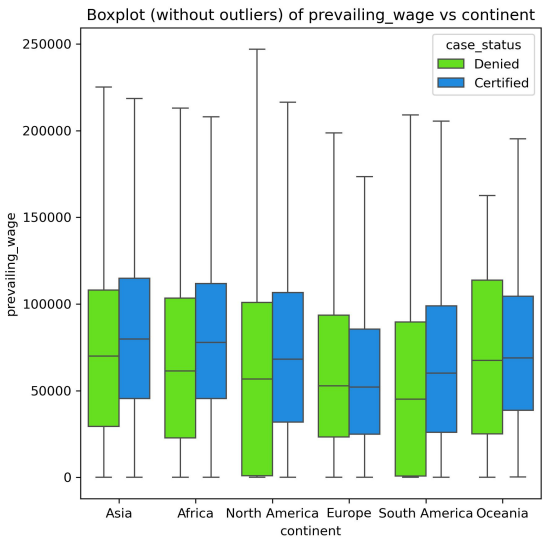


Figure 16: **'no_of_employees' and 'prevailing_wage' vs 'continent'**

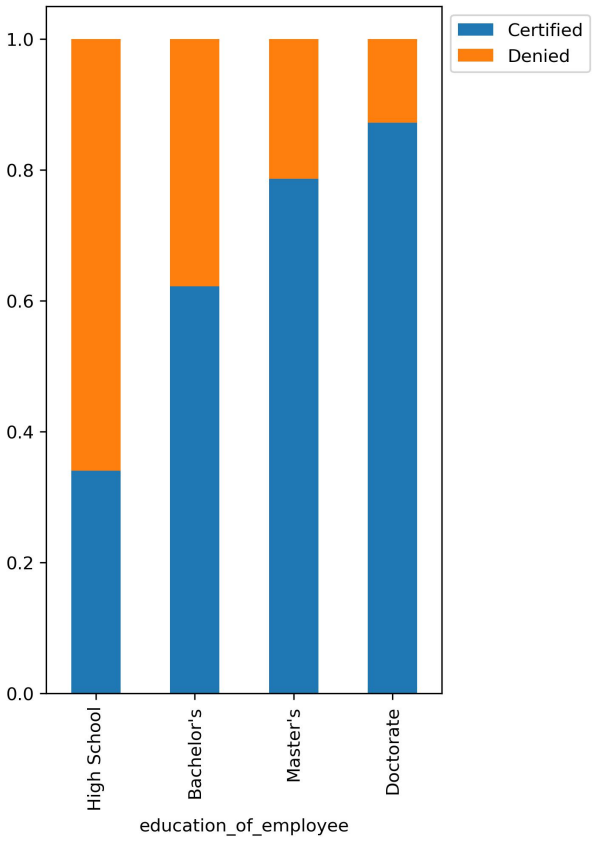
Observations and Recommendations

Observations:

- **Case Status by Continent:**
  - Certification rates are consistently higher across continents.
  - North America and Europe show the highest certification ratios.
- **Prevailing Wage:**
  - Certified cases consistently show higher prevailing wages.
  - Wage differences are less significant across continents without outliers.
- **Number of Employees:**
  - Companies with higher employee counts have better certification rates.
  - Denied cases cluster around smaller firms.

Recommendations:

- Offer competitive wages to increase certification chances.
- Analyze denial trends in regions with lower success rates.
- Prioritize applicants from sectors with high certification ratios.
- Develop dynamic wage models to align with market trends and attract talent.
- **'no_of_employees' and 'prevailing_wage' vs 'education_of_employee'**



case_status	Certified	Denied	All
education_of_employee			
All	17018	8462	25480
Bachelor's	6367	3867	10234
High School	1164	2256	3420
Master's	7575	2059	9634
Doctorate	1912	280	2192

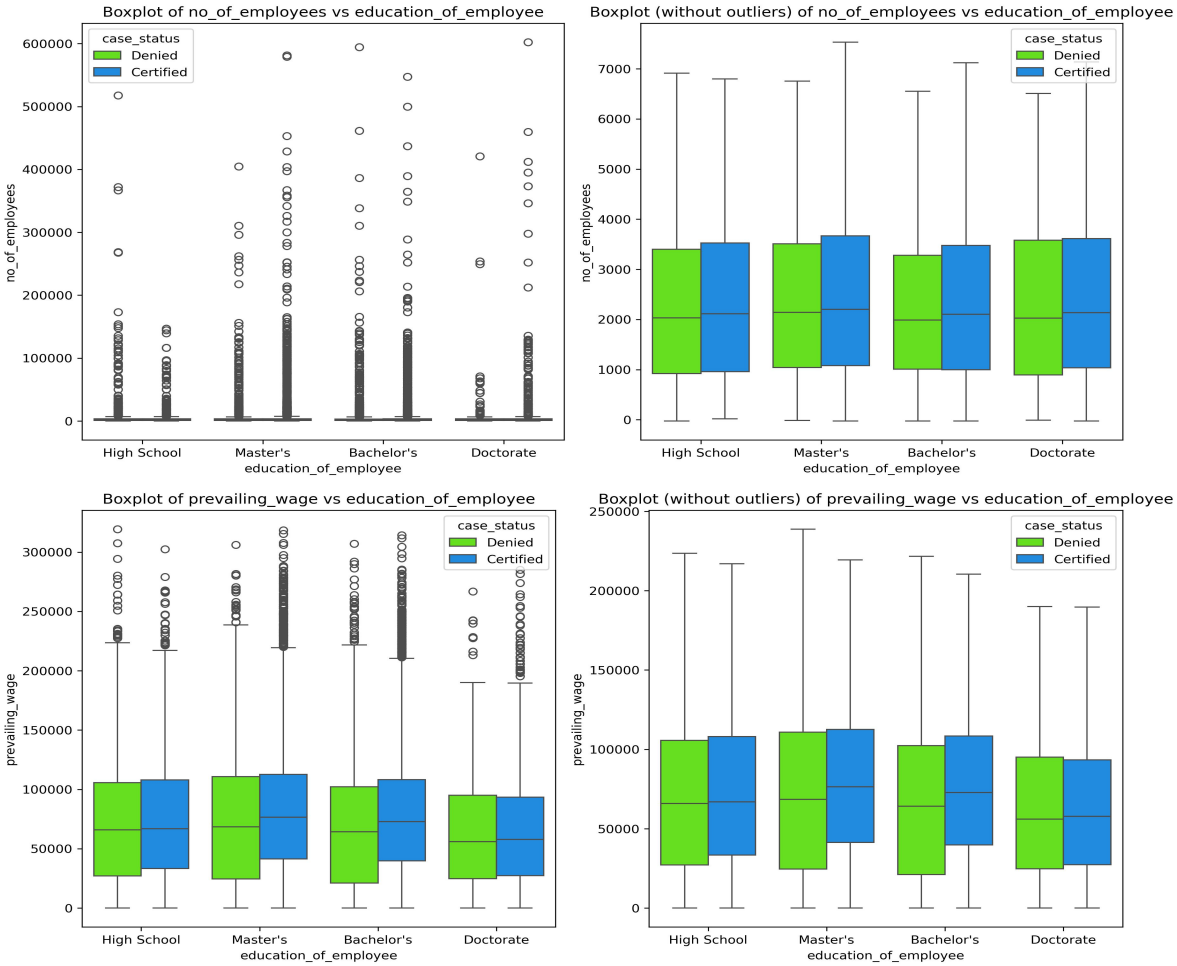


Figure 17: **'no_of_employees' and 'prevailing_wage' vs 'education_of_employee'**

Observations and Recommendations

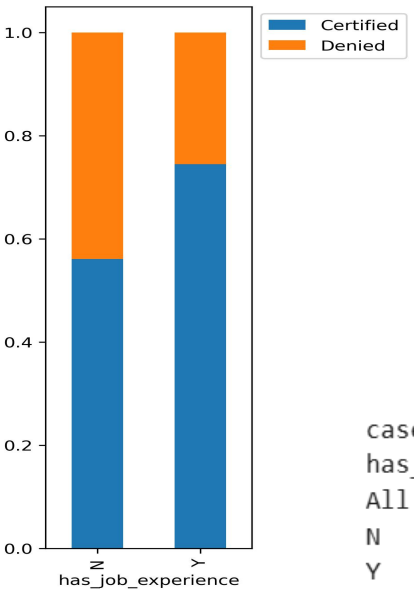
Observations:

- **Education of Employees:**
  - Higher education levels (e.g., Master's and Ph.D.) have a higher rate of visa certification.
  - Employees with only a Bachelor's degree show comparatively higher denial rates.
- **Prevailing Wage:**
  - Certified cases are associated with higher prevailing wages.
  - Denied cases tend to cluster in lower wage brackets.
- **Number of Employees:**
  - Companies with fewer employees tend to experience higher denial rates.
  - Larger employers consistently secure more visa certifications.

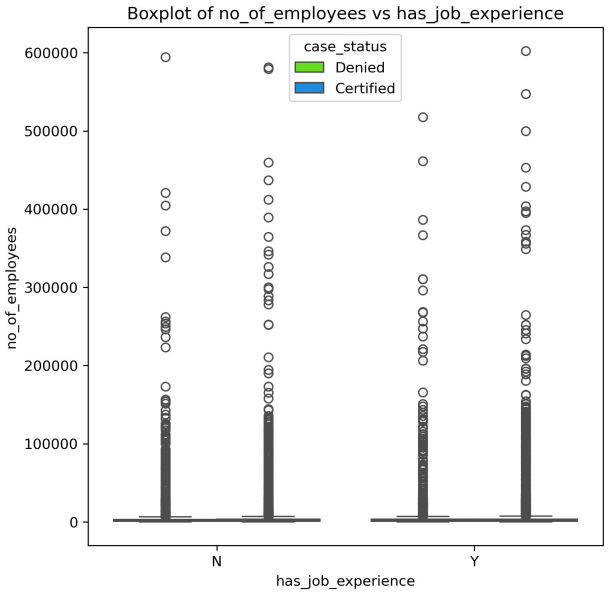
Recommendations:

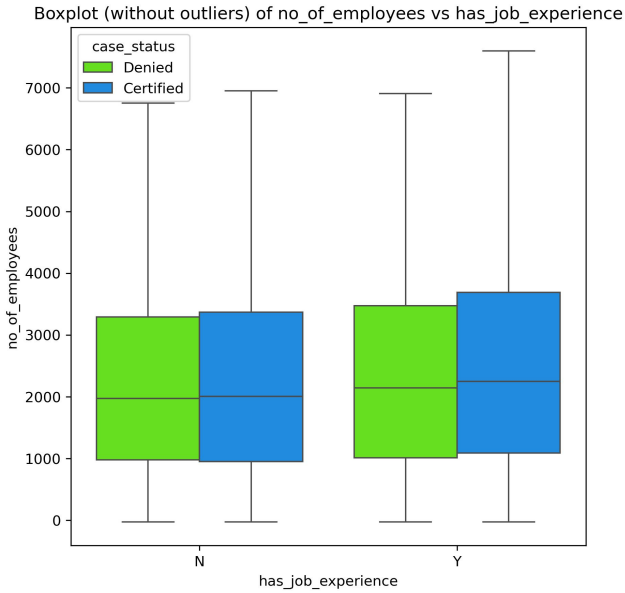
- Encourage applicants to pursue higher education levels to improve certification chances.
- Employers should offer prevailing wages above the median to increase the likelihood of visa certification.
- Focus on improving compliance among smaller companies with higher denial rates.
- Prioritize certification reviews for applicants associated with larger employers and higher wages.

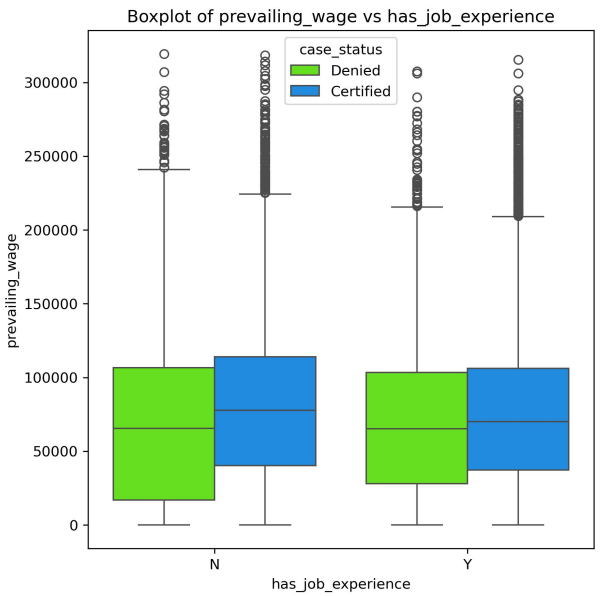
- **'no_of_employees' and 'prevailing_wage' vs 'has_job_experience'**



case_status	Certified	Denied	All
has_job_experience			
All	17018	8462	25480
N	5994	4684	10678
Y	11024	3778	14802







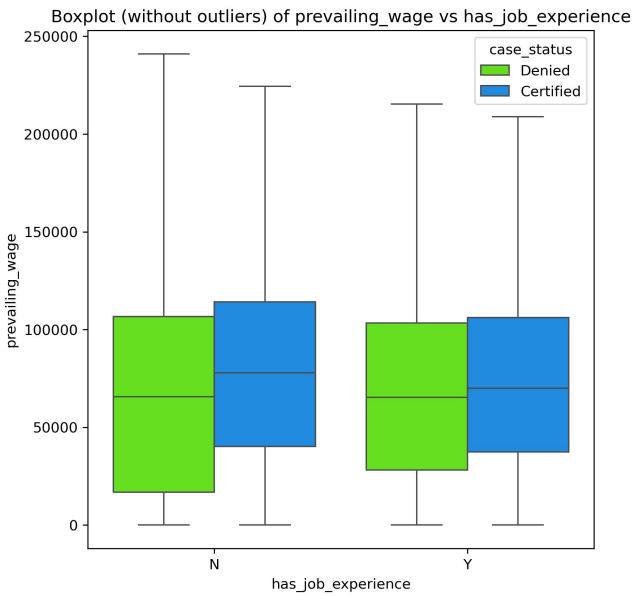


Figure 18: **'no_of_employees' and 'prevailing_wage' vs 'has_job_experience'**

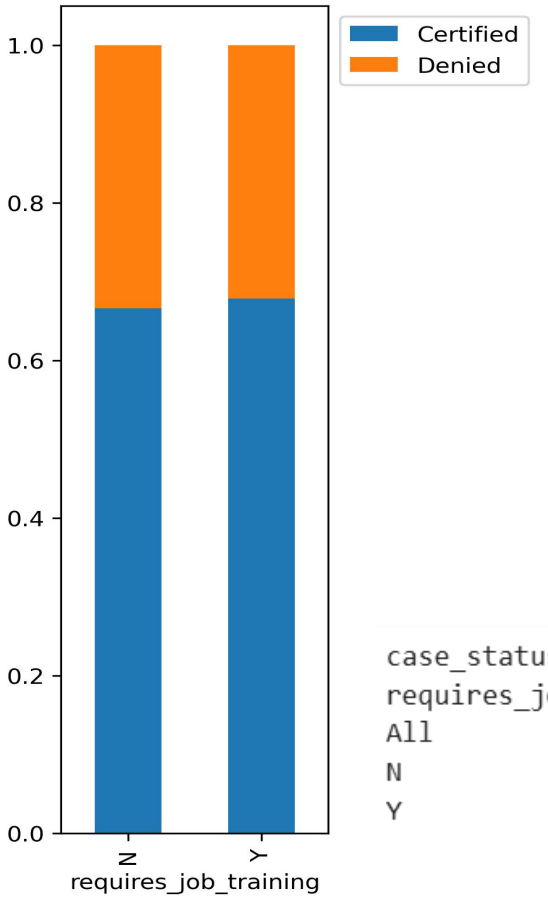
Analysis and Observations

Observations:

- **Job Experience:**
  - Employees with prior job experience have a significantly higher proportion of certified cases compared to denied cases.
  - Lack of job experience correlates with a higher likelihood of denial.
- **Prevailing Wage:**
  - Employees with job experience are offered higher prevailing wages on average.
  - Denied cases tend to have lower wages compared to certified cases, regardless of job experience.
- **Number of Employees:**
  - Companies with a higher number of employees tend to have more cases certified for workers with job experience.
  - Smaller companies show a relatively higher proportion of denied cases.

Business Recommendations

- Prioritize hiring candidates with relevant job experience to increase the likelihood of visa certification.
- Offer competitive prevailing wages, particularly for employees with job experience, to improve certification rates.
- Focus recruitment efforts on larger companies or industries with a higher certification success rate.
- Analyze denial patterns for smaller companies and provide guidance on wage standards and compliance to improve outcomes.
- **'no_of_employees' and 'prevailing_wage' vs 'requires_job_training'**



case_status	Certified	Denied	All
requires_job_training			
All	17018	8462	25480
N	15012	7513	22525
Y	2006	949	2955

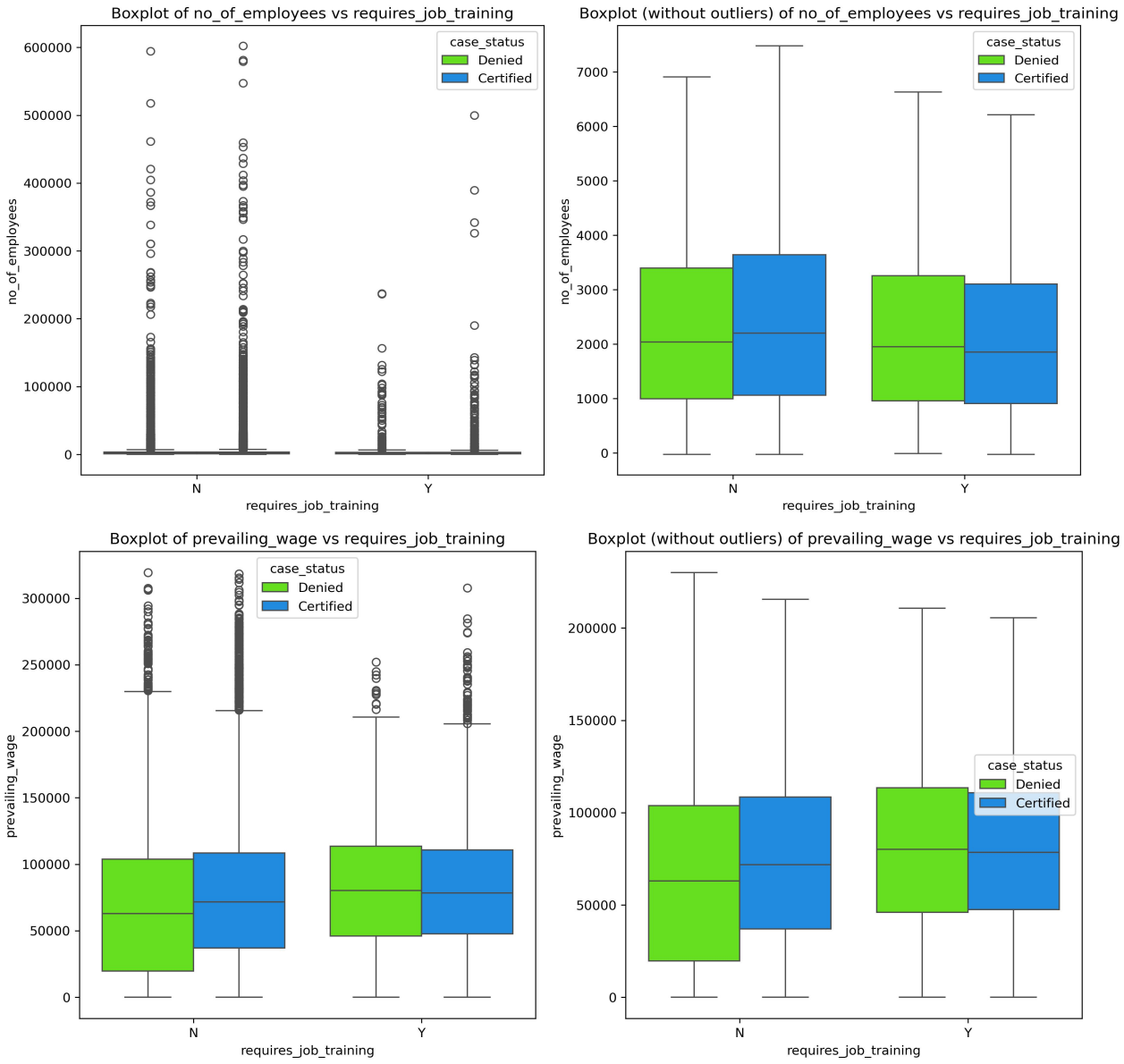


Figure 19: **'no_of_employees' and 'prevailing_wage' vs 'requires_job_training'**

Observations

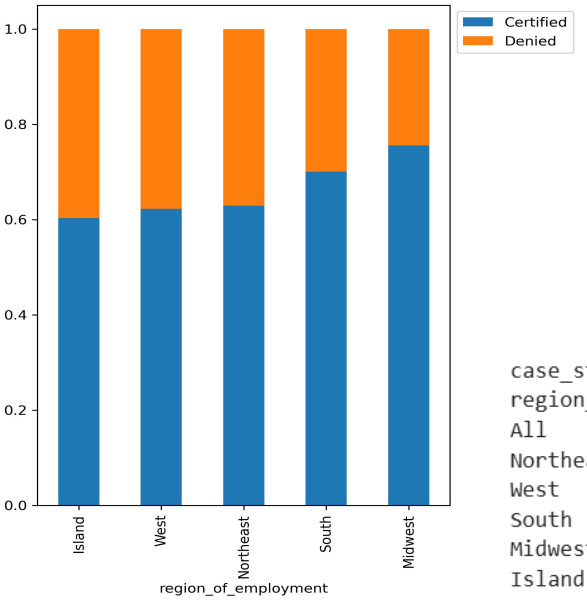
- **Job Training:**
  - Certified cases are more common for candidates who do not require job training.
  - Denied cases are relatively higher when job training is required.
- **Number of Employees:**
  - Companies with higher employee counts show increased certification rates.
  - Denials are observed more frequently in companies with smaller employee counts.
- **Prevailing Wage:**
  - Certified cases align with higher prevailing wages.
  - Denied cases correlate with lower prevailing wages.

Business Recommendations

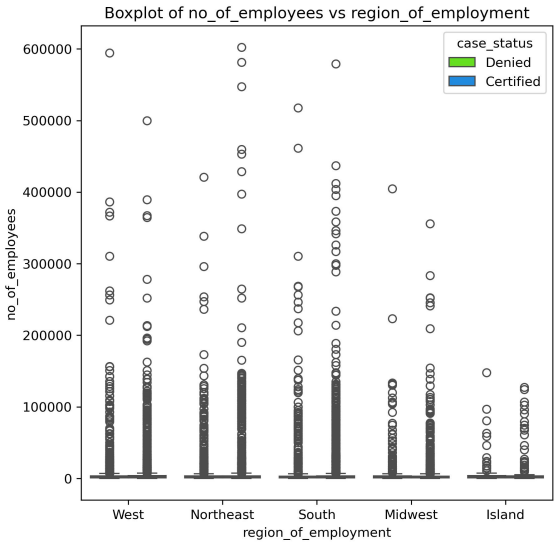
- **Target No Training Requirements:** Prioritize candidates who do not require job training to improve certification chances.
- **Focus on Larger Employers:** Encourage partnerships with companies having higher employee counts.

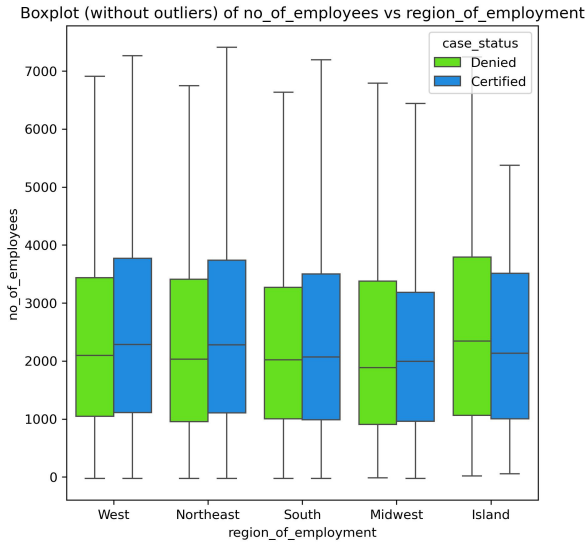
– **Increase Wages:** Advocate for wages above the prevailing standards to reduce the likelihood of denials.

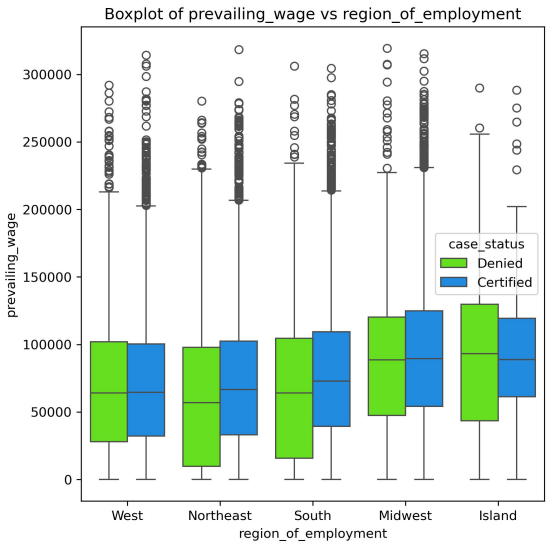
– **'no_of_employees' and 'prevailing_wage' vs 'region_of_employment'**



case_status	Certified	Denied	All
region_of_employment			
All	17018	8462	25480
Northeast	4526	2669	7195
West	4100	2486	6586
South	4913	2104	7017
Midwest	3253	1054	4307
Island	226	149	375







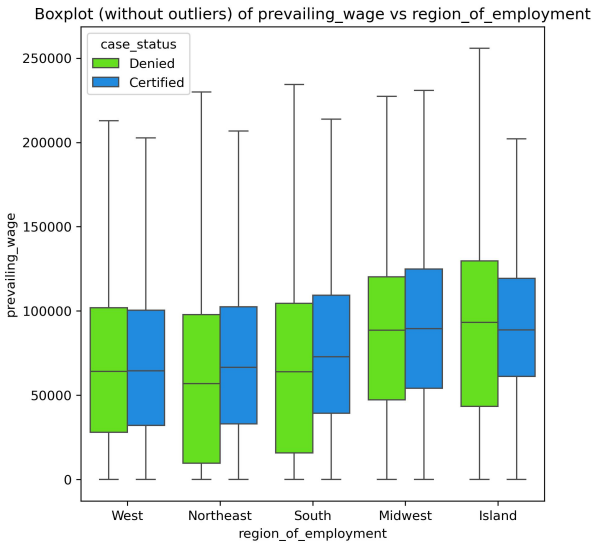


Figure 20: **'no_of_employees' and 'prevailing_wage' vs 'region_of_employment'**

Observations

Region of Employment vs Case Status

- The Midwest region has a higher denial rate compared to other regions.
- The Island region has the lowest denial rate, with most applications being certified.

Prevailing Wage vs Region of Employment

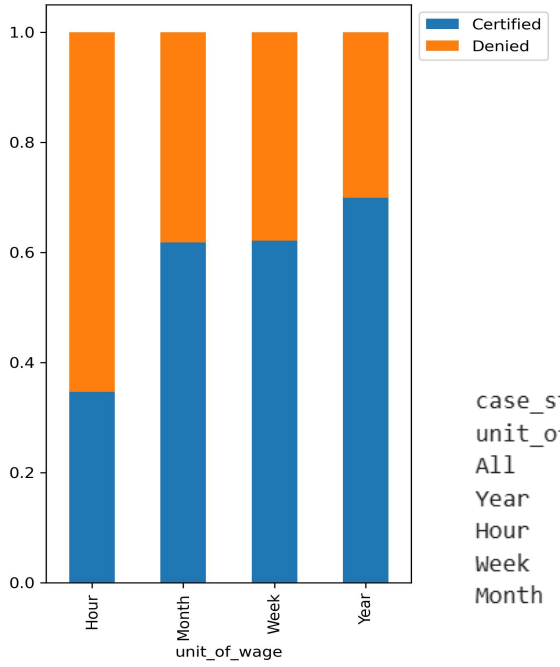
- Certified cases have a higher median prevailing wage compared to denied cases across all regions.
- The range of wages is wider in the West and Northeast regions.
- The Island region shows a consistent pattern with fewer outliers compared to other regions.

Number of Employees vs Region of Employment

- Companies with larger employee sizes tend to have a higher certification rate.
- Denied cases are more distributed among smaller companies.
- Outliers with extremely large employee sizes exist in all regions.

Business Recommendations

- **Focus on the Midwest Region:** Develop strategies to reduce denial rates, such as aligning wages or reviewing employer policies.
- **Prevailing Wage Alignment:** Employers should offer wages that meet or exceed the median wage in their region to increase certification likelihood.
- **Target Larger Employers:** Allocate resources to assist smaller companies in improving their application success rate.
- **Use the Island Region as a Benchmark:** Analyze the practices of the Island region for application success factors.
- **Outlier Analysis:** Study high-wage and large-company outliers for patterns that may influence certification.
- **Predictive Modeling:** Implement a machine learning model to prioritize applications with a higher probability of certification.
- **'no_of_employees' and 'prevailing_wage' vs 'unit_of_wage'**



case_status	Certified	Denied	All
unit_of_wage			
All	17018	8462	25480
Year	16047	6915	22962
Hour	747	1410	2157
Week	169	103	272
Month	55	34	89

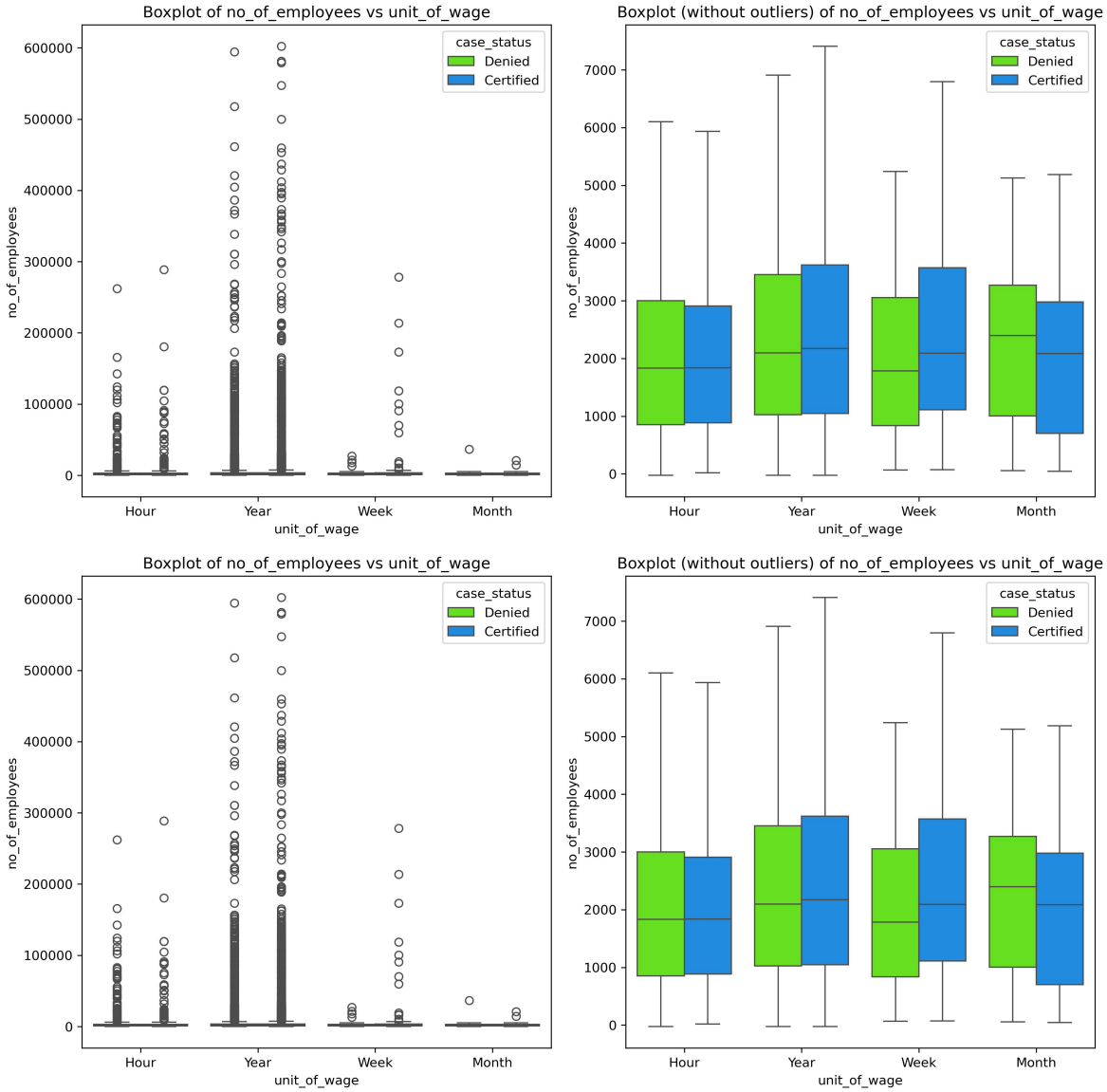


Figure 21: **'no_of_employees' and 'prevailing_wage' vs 'unit_of_wage'**

Observations

Unit of Wage vs Case Status

- Hourly wage applications have a higher denial rate compared to others.
- Yearly, Weekly, and Monthly wage applications show a significant majority of certified cases.

Number of Employees vs Unit of Wage

- For Yearly and Weekly wages, employers with higher employee counts are common in both Certified and Denied cases.
- Excluding outliers, median employee counts for Hourly and Monthly wages are lower compared to Yearly and Weekly wages.
- Denied cases tend to have slightly lower median employee counts across all wage units.

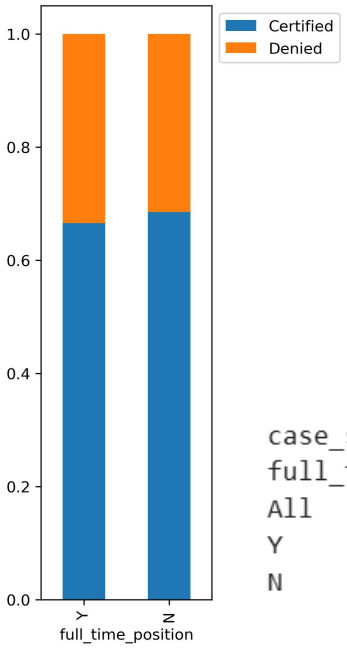
Prevailing Wage vs Unit of Wage

- Hourly wage cases exhibit significantly lower prevailing wages compared to other units.

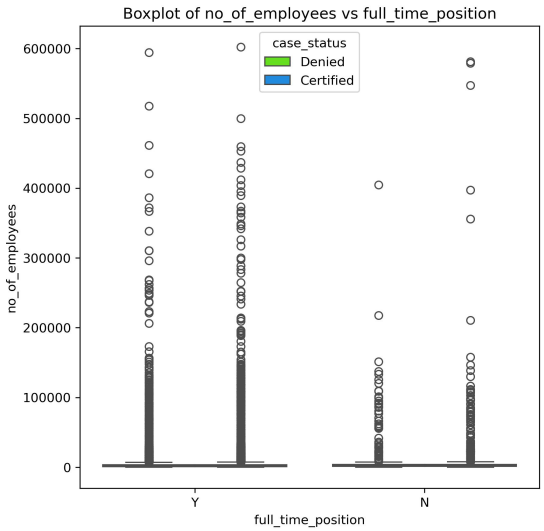
- Yearly wage applications have a higher median prevailing wage compared to Weekly and Monthly wages.
- Prevailing wages for Certified cases are generally similar to Denied cases, indicating that prevailing wage is not a strong discriminator for case status.

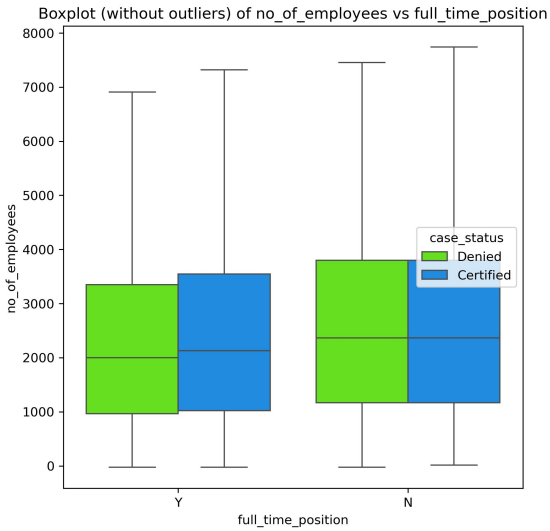
Business Recommendations

- **Prioritize Yearly Wage Applications:** Since Yearly wage cases have the highest approval rates, efforts should focus on attracting and processing these applications.
- **Review Hourly Wage Applications More Rigorously:** Given their higher denial rate and lower prevailing wages, establish stricter guidelines for hourly applications to improve quality.
- **Encourage Applications from Larger Employers:** Applications from companies with a higher employee count generally exhibit consistency in case outcomes, reducing risk.
- **Prevailing Wage Standardization:** While prevailing wage is not a decisive factor, aligning wages with industry benchmarks for each unit type could improve overall approval chances.
- **Machine Learning Focus:** Develop predictive models that consider key variables like unit of wage, employee count, and case status trends to efficiently shortlist applications.
- **'no_of_employees' and 'prevailing_wage' vs 'full_time_position'**



case_status	Certified	Denied	All
full_time_position			
All	17018	8462	25480
Y	15163	7610	22773
N	1855	852	2707





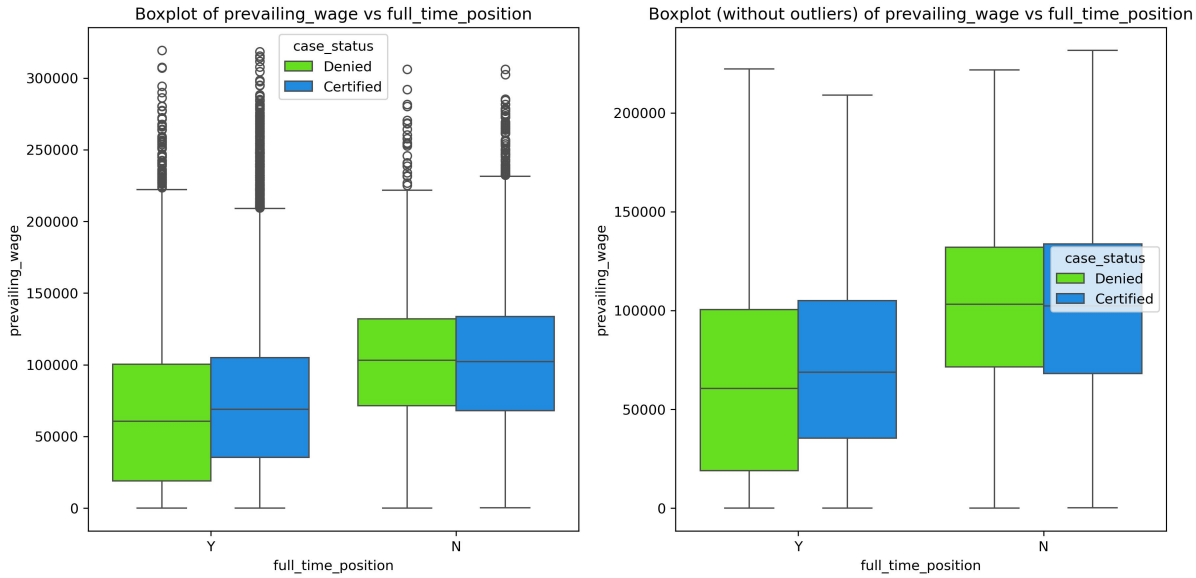


Figure 22: **'no_of_employees' and 'prevailing_wage' vs 'full_time_position'**

Observations

Analysis of Full-Time Positions and Case Status

From the stacked bar chart of Full-Time Positions vs. Case Status:

- The proportion of Certified and Denied cases is nearly identical for full-time (Y) and part-time (N) positions.
- Full-time positions are not a strong indicator of case status outcomes.

Prevailing Wage and Full-Time Positions

From the boxplot of Prevailing Wage:

- Full-time positions generally have a higher median prevailing wage compared to part-time positions for both Certified and Denied cases.
- Outliers in prevailing wages significantly expand the range, particularly for Certified cases.

Number of Employees and Full-Time Positions

From the boxplot of Number of Employees:

- Companies with a higher number of employees tend to hire for both full-time and part-time positions equally.
- Certified cases for full-time positions have a higher median number of employees compared to Denied cases.

Business Recommendations

- **Prevailing Wage as a Key Factor:** Focus on applicants applying for full-time positions with prevailing wages above the median. Higher wages seem to correlate with case certification.
- **Target Larger Companies:** Companies with a higher number of employees may have a better success rate for case certification in full-time positions.

- **Reduce Emphasis on Full-Time/Part-Time:** Case certification is not strongly influenced by whether the position is full-time or part-time. Focus on other significant drivers like prevailing wage and company size.
- **Outlier Management:** Investigate and manage extreme outliers in prevailing wages to better understand their impact on case outcomes.

5 Data preprocessing

The dataset contains no missing or duplicate values. The outliers are significant for the data so we don't require to remove them. The data is scaled with standard scaler function for modelling Logistic Regression.

A company which has been in the business for a long duration is more trustworthy than the newly established ones. So let us calculate the current age of the company and incorporate that in our model rather than building our model with year of establishment of the company.

We have splited the data for training, validating and testing purposes. The model summary is as follows.

6 Model building

6.1 Model Building - Original Data

Choosing the Right Evaluation Metric

Selecting the appropriate evaluation metric is crucial for developing a reliable Machine Learning model for visa application status prediction. The choice of metric depends on the problem context and the potential impact of misclassifications.

Why Not Accuracy?

While accuracy is a commonly used metric, it may not be suitable for imbalanced datasets where one class significantly outnumbers the other. In the context of visa approval, if the majority of applications are certified, a model could achieve high accuracy simply by predicting all cases as certified, without effectively identifying denied applications. This would provide misleading results and fail to address the critical need for correctly classifying denied applications.

Focus on Recall or Precision?

- **Recall:** Recall measures the proportion of actual positives (e.g., certified cases) that are correctly identified by the model. In this scenario, a high recall ensures that most of the certified cases are captured, minimizing the risk of missing valid approvals. However, focusing solely on recall could lead to a high number of false positives, causing unnecessary actions on denied cases.
- **Precision:** Precision measures the proportion of predicted positives that are actually correct. A high precision ensures that the certified cases identified by the model are indeed likely to be certified, reducing the burden of unnecessary follow-ups on false predictions. However, focusing solely on precision could result in missed opportunities to approve legitimate applications.

Why F1-Score?

The F1-Score, which is the harmonic mean of precision and recall, provides a balanced measure that considers both false positives and false negatives. For visa application prediction:

- **False Negatives (FN):** Missing an actual certification (a denied prediction for a certified application) could lead to significant delays for qualified applicants and employers.

– **False Positives (FP):** Incorrectly predicting certification for a denied application could waste resources on invalid cases.

By balancing precision and recall, the F1-Score ensures that the model not only identifies certified applications accurately but also minimizes unnecessary follow-ups on denied cases. This is particularly important in high-stakes decision-making scenarios like visa approvals, where both over-predicting and under-predicting certifications can have severe consequences.

Conclusion

For this dataset and problem context, the **F1-Score** is the most appropriate metric. It accounts for the imbalanced nature of the dataset and ensures a balance between precision and recall, enabling a fair and effective evaluation of the model's performance. We have built over 7 models. Their performance is shown below.

```
Cross validation performance:

Bagging        : 0.7694       Model - Bagging        : F1 Score - 0.7706
Random forest  : 0.8041       Model - Random forest  : F1 Score - 0.7960
GBM            : 0.8222       Model - GBM            : F1 Score - 0.8214
Adaboost       : 0.8148       Model - Adaboost       : F1 Score - 0.8153
Xgboost        : 0.8040       Model - Xgboost        : F1 Score - 0.8036
dtree          : 0.7426       Model - dtree          : F1 Score - 0.7422
LRegression    : 0.8136       Model - LRegression    : F1 Score - 0.8132
```

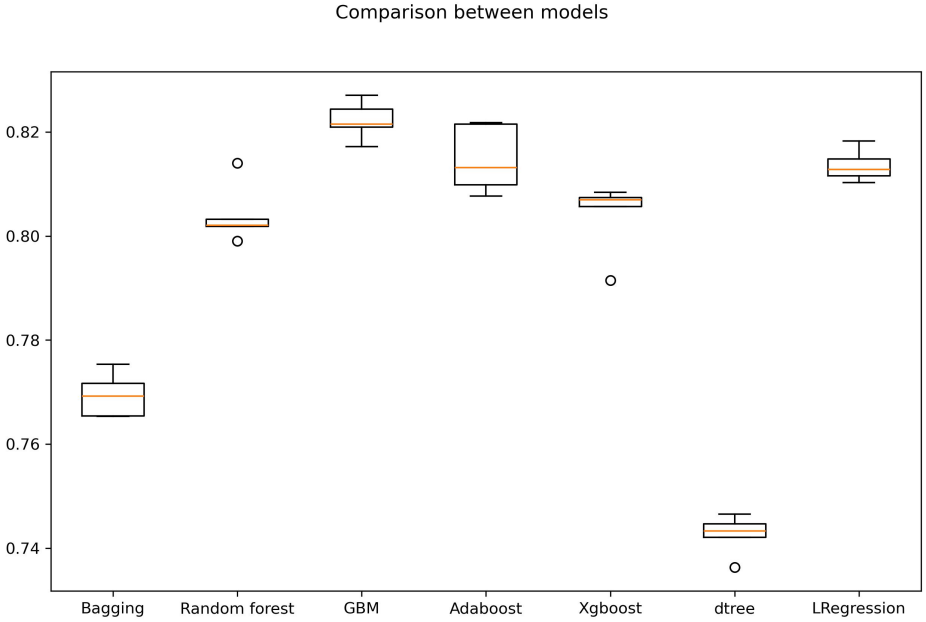


Figure 23: Comparison between models (original data)

– We can see that the GBM is giving the highest cross-validated f1 score followed by Adaboost and Logistic Regression.

– The boxplot shows that the performance of GBM and Adaboost is consistent and their performance on the validation set is also good.

– We will tune the best three models i.e. GBM, Adaboost & Logistic Regression and see if the performance improveses

6.2 Model Building - Oversampled Data

To address the imbalance in visa status data, where 'Certified' and 'Denied' cases have a 2:1 ratio, we apply oversampling to the minority class ('Denied'). This technique increases the representation of 'Denied' cases,

achieving a balanced 1:1 ratio between the classes. By doing so, the model is trained on an equal distribution of both classes, improving its ability to accurately predict 'Denied' cases without bias toward the majority class. We have built over 7 models. Their performance is shown below.

```
Cross validation performance on oversampled data:

Bagging        : 0.7583
Random forest  : 0.7923
GBM            : 0.8028
Adaboost       : 0.7970
Xgboost        : 0.7945
dtree          : 0.7216
LRegression    : 0.7484
```

```
Model - Bagging        : F1 Score - 0.7571
Model - Random forest  : F1 Score - 0.7882
Model - GBM            : F1 Score - 0.8185
Model - Adaboost       : F1 Score - 0.8146
Model - Xgboost        : F1 Score - 0.8057
Model - dtree          : F1 Score - 0.7262
Model - LRegression    : F1 Score - 0.7828
```

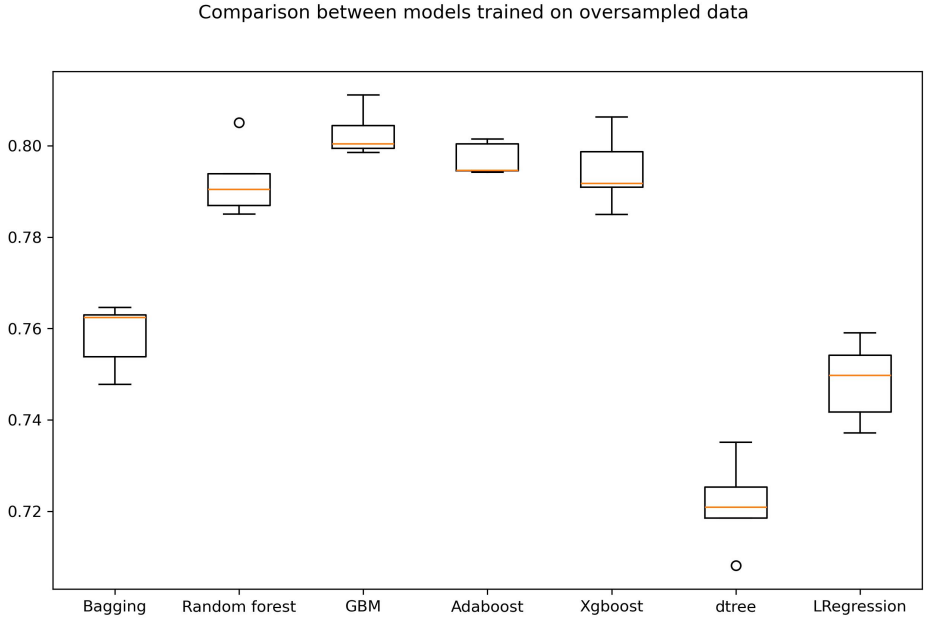


Figure 24: Comparison between models (oversampled data)

- We can see that the GBM is giving the highest cross-validated f1 score followed by Adaboost and xgboost
- The boxplot shows that the performance of GBM and Adaboost is consistent and their performance on the validation set is also good.
- We will tune the best three models i.e. GBM, Adaboost & Xgboost and see if the performance improveses.

6.3 Model Building - Undersampled Data

To address the imbalance in visa status data, where 'Certified' and 'Denied' cases have a 2:1 ratio, we apply undersampling to the majority class ('Certified'). This technique decreases the representation of 'Certified' cases, achieving a balanced 1:1 ratio between the classes. By doing so, the model is trained on an equal distribution of both classes, improving its ability to accurately predict 'Denied' cases without bias toward the majority class. We have built over 7 models. Their performance is shown below.

```
Cross validation performance on undersampled data:

Bagging        : 0.6389
Random forest  : 0.6762
GBM            : 0.7052
Adaboost       : 0.6964
Xgboost        : 0.6765
dtree          : 0.6319
LRegression    : 0.6854
```

```
Model - Bagging        : F1 Score - 0.6976
Model - Random forest  : F1 Score - 0.7359
Model - GBM            : F1 Score - 0.7621
Model - Adaboost       : F1 Score - 0.7601
Model - Xgboost        : F1 Score - 0.7470
Model - dtree          : F1 Score - 0.6884
Model - LRegression    : F1 Score - 0.7438
```

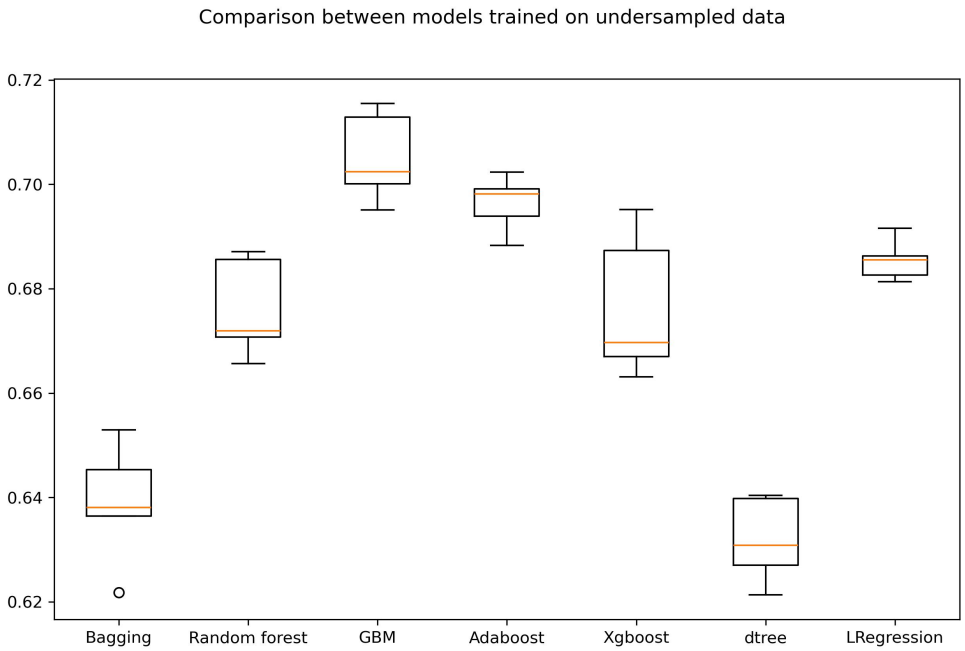


Figure 25: Comparison between models (undersampled data)

- We can see that the GBM is giving the highest cross-validated f1 score followed by Adaboost and Logistic Regression.
- The boxplot shows that the performance of GBM and Adaboost is consistent and their performance on the validation set is also good.
- We will tune the best three models i.e. GBM, Adaboost & Logistic Regression and see if the performance improves.

Hence from all the 7 methods of model bulding we found the top 4 models with highest f1 score are:

1. Gradient Boosting Classifier
2. AdaBoost
3. XgBoost
4. Logistic Regression

6.4 Model Performance Improvement using Hyperparameter Tuning

We will tune Gradient Boosting Classifier,AdaBoost,Logistic Regression and xgboost models using GridSearchCV and RandomizedSearchCV. We will also compare the performance and time taken by these two methods - grid search and randomized search.

6.4.1 Gradient Boosting Classifier

The Gradient Boosting Classifier was fine-tuned using GridSearchCV to optimize its performance by exploring a grid of hyperparameters. The search included the number of estimators (50, 100, 200), learning rate (0.01, 0.1), maximum tree depth (3, 5, 7), minimum samples required to split a node (2, 5, 10), subsampling fractions (0.7, 0.9, 1.0), and the number of features for the best split ('sqrt', 'log2'). F1-score was used as the evaluation metric, and 5-fold cross-validation ensured robustness. The best model achieved an F1-score of 0.826 with parameters: learning rate 0.01, max depth 7, min samples split 10, 200 estimators, subsample 0.9, and max features 'sqrt,' demonstrating its strong predictive capability. The evaluation performance is shown below.

```
Training performance:
   Accuracy    Recall  Precision       F1
0  0.757718  0.923506   0.763358  0.83583
Testing performance:
   Accuracy    Recall  Precision        F1
0  0.740973  0.912456   0.752422  0.824748
```

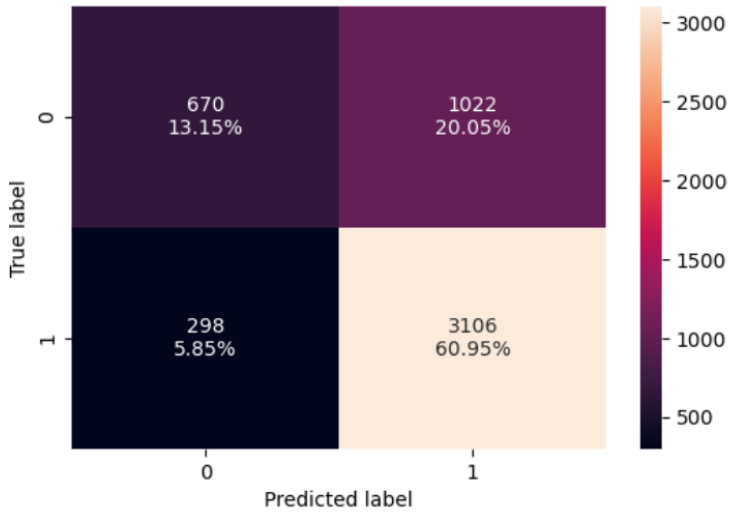


Figure 26: Gradient Boosting Classifier performance

The RandomizedSearchCV was employed to optimize the hyperparameters of the Gradient Boosting Classifier by randomly sampling from a pre-defined parameter grid. The parameters explored included the number of estimators (50, 100, 200), learning rate (0.01, 0.1), maximum tree depth (3, 5, 7), minimum samples required for node splits (2, 5, 10), subsample fraction (0.7, 0.9, 1.0), and the number of features for the best split ('sqrt', 'log2'). After 5-fold cross-validation, the best hyperparameters were identified: subsample 1.0, n_estimators 200, min_samples_split 10, max_features 'log2', max_depth 7, and learning rate 0.01, with a cross-validation score of 0.825. The model's performance on the training set showed an accuracy of 0.761, with a recall of 0.919, precision of 0.769, and F1-score of 0.837. On the testing set, the model achieved an accuracy of 0.744, a recall of 0.904, precision of 0.759, and F1-score of 0.825, highlighting a strong performance on both training and testing datasets. The evaluation performance is shown below.

```
Training performance:
   Accuracy    Recall  Precision        F1
0  0.761316  0.918609    0.76896  0.837149
Testing performance:
   Accuracy    Recall  Precision        F1
0  0.744309  0.903937   0.759191  0.825265
```

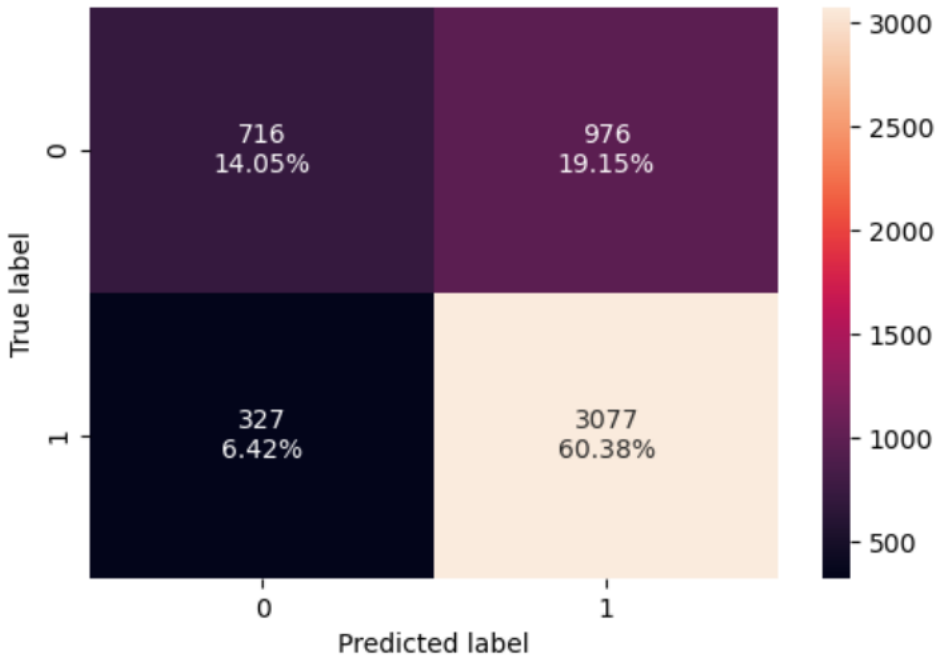


Figure 27: Gradient Boosting Classifier performance

6.4.2 AdaBoost Classifier

The AdaBoostClassifier was optimized using GridSearchCV with cross-validation to identify the best hyperparameters for the model. The base estimator for AdaBoost was a DecisionTreeClassifier with a maximum depth of 3 and a minimum samples split of 2. The hyperparameter grid considered for tuning included the number of estimators (50, 100, 200), learning rate (0.01, 0.1, 0.2), and parameters for the decision tree (max_depth and min_samples_split). The GridSearchCV, employing the F1 score as the evaluation metric, identified the best parameters: 'estimator__max_depth': 3, 'estimator__min_samples_split': 2, 'learning_rate': 0.1, and 'n_estimators': 50, with the best cross-validation F1 score of 0.822.

The tuned AdaBoost model achieved an accuracy of 0.752, recall of 0.883, precision of 0.776, and an F1 score of 0.826 on the training set. On the validation set, the model maintained strong performance, with an accuracy of 0.748, recall of 0.877, precision of 0.776, and F1 score of 0.823, indicating robust generalization on unseen data. The model demonstrated efficiency, completing 405 fits in 226 seconds. The evaluation performance is shown below.

```
Training Performance:
   Accuracy   Recall  Precision        F1
0  0.751897  0.88335   0.776095  0.826256
Testing Performance:
   Accuracy    Recall  Precision       F1
0   0.74843  0.877203   0.775584  0.82327
```

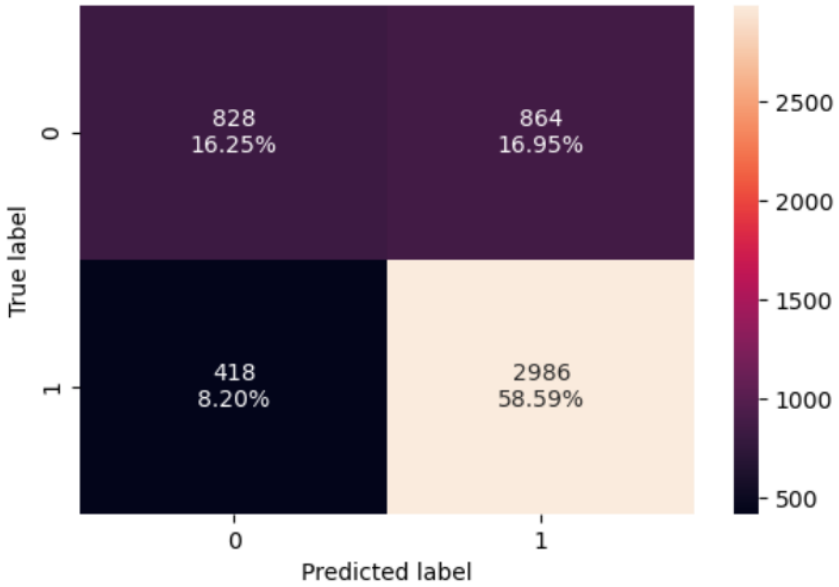


Figure 28: AdaBoost Classifier performance

The AdaBoost model, using a DecisionTreeClassifier as the base estimator, was optimized using RandomizedSearchCV. The hyperparameter grid considered values for 'n_estimators', 'learning_rate', 'estimator__max_depth', and 'estimator__min_samples_split'. The RandomizedSearchCV, which ran 250 fits, identified the optimal parameters: 'n_estimators': 50, 'learning_rate': 0.1, 'estimator__min_samples_split': 5, and 'estimator__max_depth': 3, achieving a cross-validation F1 score of 0.822.

The tuned model demonstrated strong performance on the training data, with an accuracy of 0.752, recall of 0.884, precision of 0.776, and an F1 score of 0.826. On the validation data, it achieved an accuracy of 0.749, recall of 0.878, precision of 0.775, and F1 score of 0.824, indicating consistent generalization across both sets. The RandomizedSearchCV fitting time was 195 seconds. The evaluation performance is shown below.

```
RandomizedSearchCV fitting time: 195.42 seconds
Best Parameters: {'n_estimators': 50, 'learning_rate': 0.1, 'estimator__min_samples_split': 5, 'estimator__max_depth': 3}
Best Cross-Validation Score: 0.8223994673906152
Training Performance:
   Accuracy    Recall  Precision        F1
0  0.751766  0.883741   0.775772  0.826244
Testing Performance:
   Accuracy    Recall  Precision        F1
0  0.748823  0.878378   0.775415  0.823691
```

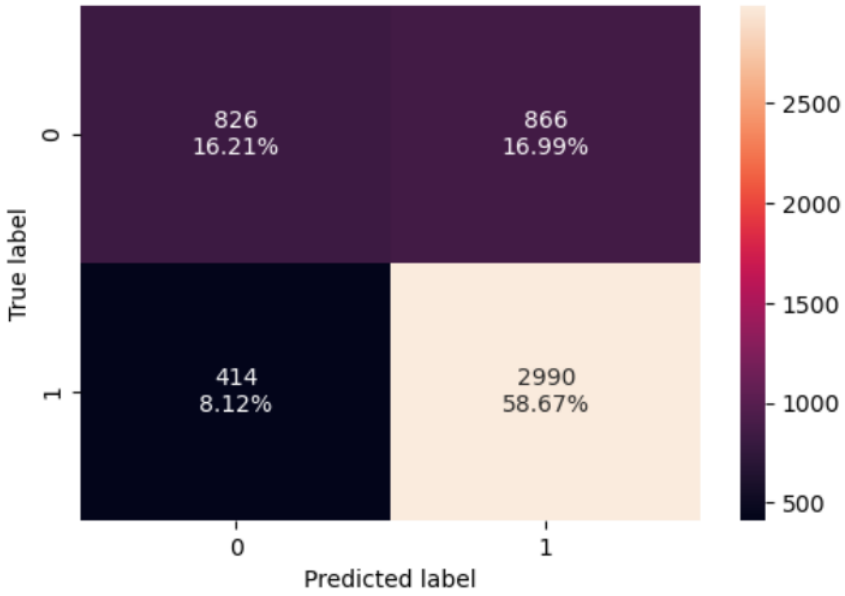


Figure 29: AdaBoost Classifier performance

6.4.3 Xgboost Classifier

The **XGBoost** model was fine-tuned using **GridSearchCV** to identify the best hyperparameters from a comprehensive search space. The parameter grid included values for the number of estimators (`n_estimators`), learning rate (`learning_rate`), tree depth (`max_depth`), minimum child weight (`min_child_weight`), and sample fraction (`subsample` and `colsample_bytree`). After performing 3645 total fits with 5-fold cross-validation, the best set of hyperparameters was identified: `n_estimators`=200, `learning_rate`=0.01, `max_depth`=5, `min_child_weight`=1, `subsample`=0.8, and `colsample_bytree`=0.8. The model achieved an impressive best F1 score of 0.826 during the grid search. Upon evaluating the best model on both the training and testing datasets, the performance metrics showed that the model achieved a training accuracy of 74.83%, with a recall of 92.03% and an F1 score of 0.830. For the testing data, the model attained an accuracy of 73.92%, a recall of 91.28%, and an F1 score of 0.824. The **GridSearchCV** fitting process took 421 seconds to complete, confirming the model's efficiency and robustness. The evaluation performance is shown below.

```
Training Performance:
   Accuracy    Recall  Precision        F1
0  0.748299  0.920274   0.755913  0.830035
Testing Performance:
   Accuracy   Recall  Precision       F1
0  0.739207  0.91275   0.750664  0.82381
```

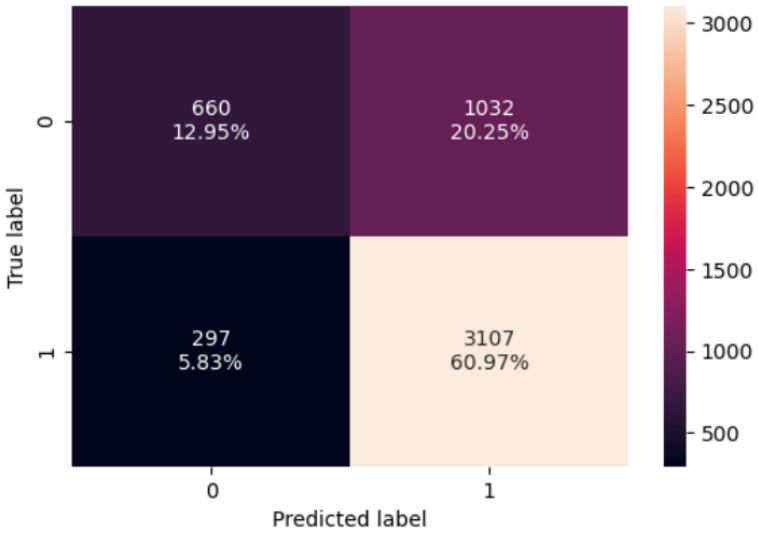


Figure 30: Xgboost Classifier performance

The **XGBoost** model was fine-tuned using **RandomizedSearchCV** to identify the best hyperparameters from a predefined grid. The search space included values for the number of estimators (`n_estimators`), learning rate (`learning_rate`), tree depth (`max_depth`), minimum child weight (`min_child_weight`), and sample fraction (`subsample` and `colsample_bytree`). After performing 250 fits with 5-fold cross-validation, the best set of hyperparameters was identified: `n_estimators`=200, `learning_rate`=0.01, `max_depth`=5, `min_child_weight`=5, `subsample`=0.8, and `colsample_bytree`=0.8. The model achieved an impressive best cross-validation score of 0.825 during the search process. Upon evaluating the best model on both the training and testing datasets, the performance metrics showed that the model achieved a training accuracy of 74.60%, with a recall of 91.84% and an F1 score of 0.828. For the testing data, the model attained an accuracy of 74.02%, a recall of 91.33%, and an F1 score of 0.824. The **RandomizedSearchCV** fitting process took 34.94 seconds to complete, further confirming the model's efficiency and robustness. The evaluation performance is shown below.

```
Training Performance:
   Accuracy    Recall  Precision        F1
0   0.74601  0.918413   0.754567  0.828467
Testing Performance:
   Accuracy    Recall  Precision       F1
0  0.740188  0.913337   0.751329  0.82445
```

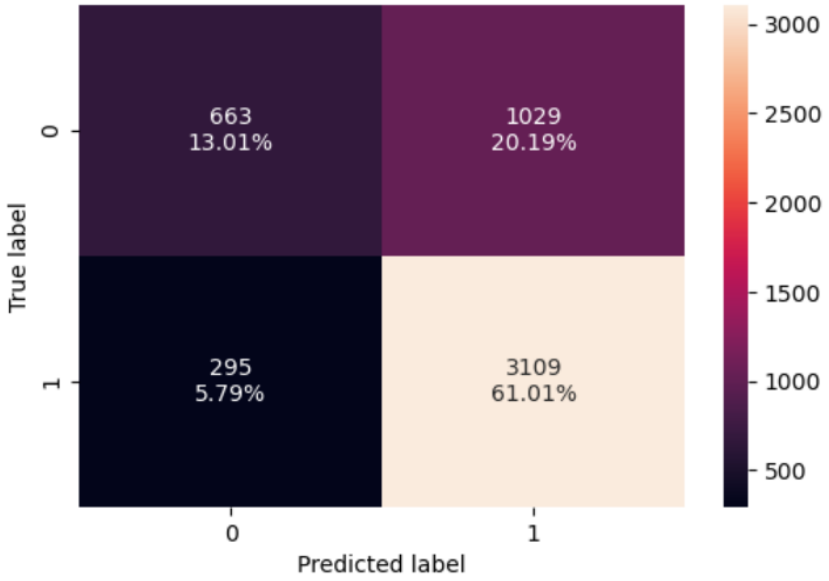


Figure 31: Xgboost Classifier performance

6.4.4 Logistic Classifier

The **Logistic Classifier** model was first preprocessed by scaling the features using **StandardScaler**, ensuring that the input variables had zero mean and unit variance. After scaling, the Variance Inflation Factor (**VIF**) was computed to check for multicollinearity among the predictors. The following variables were identified with high VIF values (>5) and removed due to multicollinearity: `continent_Asia`, `continent_Europe`, `continent_North America`, `region_of_employment_Midwest`, `region_of_employment_Northeast`, `region_of_employment_South`, and `region_of_employment_West`. The remaining features were selected for further analysis. The next step involved checking for statistical significance, where features with p-values greater than 0.05 were removed from the model. The removed variables were: `no_of_employees`, `prevailing_wage`, `Age_of_company`, and `continent_Oceania`. A **Logistic Regression** model was then fitted with the significant features. The results showed that the model achieved a **training accuracy** of 73.70%, with a **recall** of 89.58%, and an **F1 score** of 0.820. For the testing dataset, the model attained an accuracy of 73.59%, a recall of 89.25%, and an F1 score of 0.819. The fitting process involved **GridSearchCV** for hyperparameter tuning, taking 100 fits in total with the best parameters identified. The following features were found to be statistically significant with their respective coefficients and p-values: The evaluation performance is shown below.

Feature	p-value
const	0.000
continent_South America	0.001
education_of_employee_Doctorate	0.000
education_of_employee_High School	0.000
education_of_employee_Master's	0.000
has_job_experience_Y	0.000
requires_job_training_Y	0.016
unit_of_wage_Month	0.002
unit_of_wage_Week	0.000
unit_of_wage_Year	0.000
full_time_position_Y	0.000

Table 1: Statistically Significant Columns and Their p-Values

```
Training Performance:
   Accuracy    Recall  Precision        F1
0  0.736983  0.895788    0.75568  0.819791
Testing Performance:
   Accuracy    Recall  Precision        F1
0  0.735871  0.892479   0.756098  0.818647
```

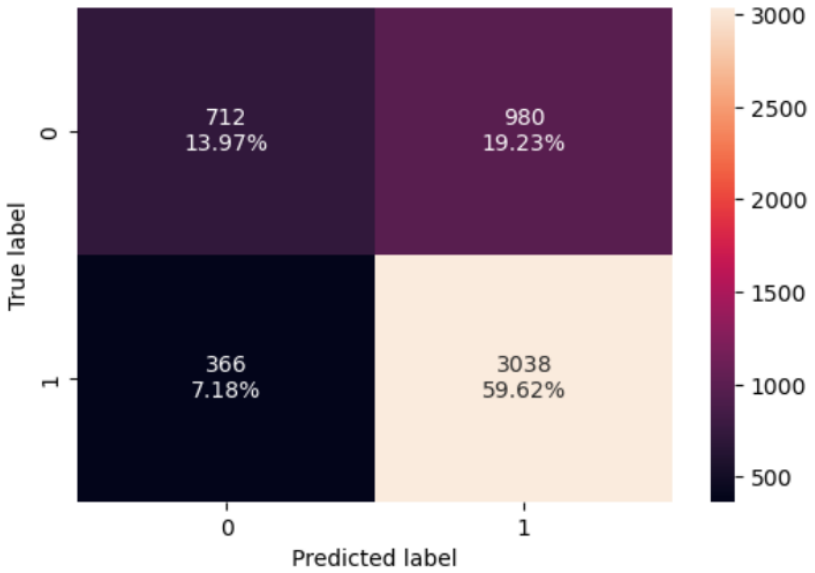


Figure 32: Logistic Classifier performance

The **Logistic Classifier** model was further optimized using **GridSearchCV** to fine-tune hyperparameters. The parameter grid included values for the regularization type (`penalty`: L2 and L1), the regularization strength (`C`), and the solvers compatible with L1 regularization (`liblinear` and `saga`). A total of 100 fits were performed using 5-fold cross-validation, and the **F1 score** was used as the evaluation metric. The best set of hyperparameters identified by the grid search was: `penalty='l2'`, `C=1`, and `solver='liblinear'`. The model's performance was then evaluated on both the training and testing datasets. For the training data, the model achieved a **training accuracy** of 73.63%, with a **recall** of 88.02%, a **precision** of 76.19%, and an **F1 score** of 0.817. For the testing data, the model achieved an accuracy of 73.82%, a recall of 87.81%, a precision of 76.48%, and an F1 score of 0.818. The grid search fitting process took a total of 100 fits, demonstrating the efficiency of the model. The evaluation performance is shown below.

```
Training Performance:
   Accuracy    Recall  Precision        F1
0  0.736329  0.880215   0.761933  0.816814
Testing Performance:
   Accuracy    Recall  Precision       F1
0  0.738226  0.878085   0.764841  0.81756
```

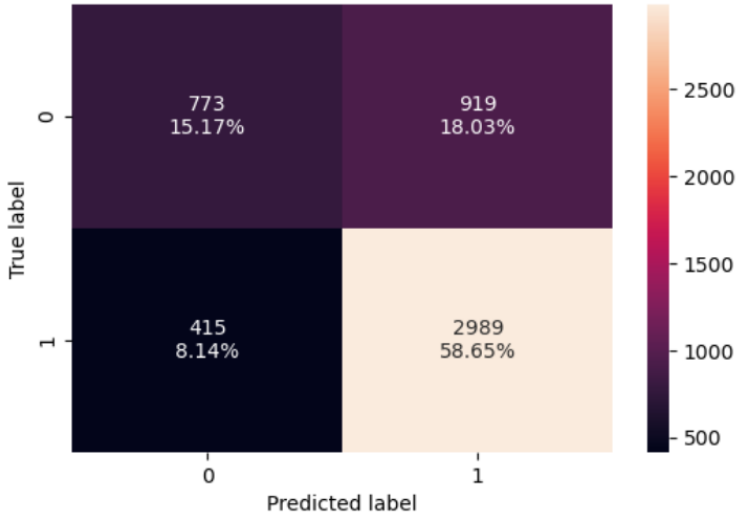


Figure 33: Logistic Classifier performance

The **Logistic Classifier** model was further optimized using **RandomizedSearchCV** to fine-tune hyperparameters. The parameter grid included values for the regularization type (`penalty`: L2 and L1), the regularization strength (`C`: logarithmically spaced values between 10^{-3} and 10^{3}), and solvers compatible with L1 regularization (`liblinear` and `saga`). A total of 140 fits were performed using 5-fold cross-validation, with the **F1 score** as the evaluation metric. The best set of hyperparameters identified by the randomized search was: `penalty='l2'`, `C=1.0`, and `solver='liblinear'`. The model's performance was then evaluated on both the training and testing datasets. For the training data, the model achieved a **training accuracy** of 73.63%, with a **recall** of 88.02%, a **precision** of 76.19%, and an **F1 score** of 0.817. For the testing data, the model achieved an accuracy of 73.82%, a recall of 87.81%, a precision of 76.48%, and an F1 score of 0.818. The randomized search fitting process took a total of 140 fits. The evaluation performance is shown below.

```
Training Performance:
   Accuracy    Recall  Precision        F1
0  0.736329  0.880215   0.761933  0.816814
Testing Performance:
   Accuracy    Recall  Precision       F1
0  0.738226  0.878085   0.764841  0.81756
```

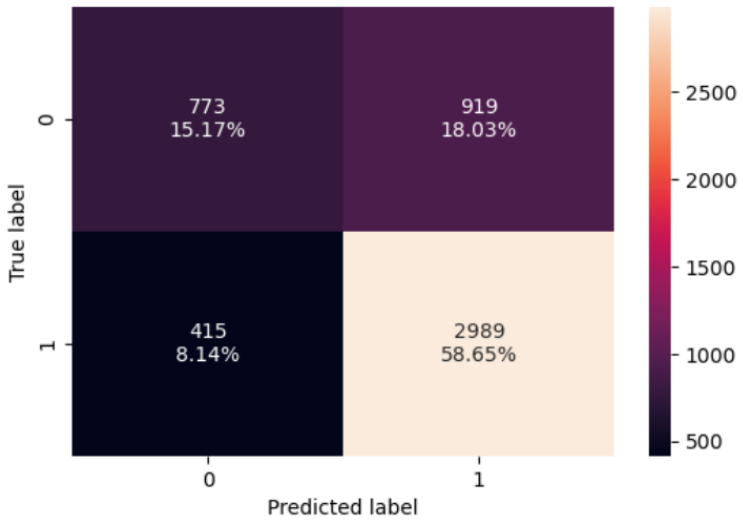


Figure 34: Logistic Classifier performance

6.4.5 Stacking Model

Now, let's build a stacking model with the tuned models - Gradient Boosting Classifier, AdaBoost, and Logistic Regression, then use XGBoost to get the final prediction.

The **Stacking Classifier** model was constructed by combining three base models: **Gradient Boosting Classifier**, **AdaBoost**, and **Logistic Regression**. The final estimator used in the stacking classifier was the **XGBoost** model. The stacking classifier was trained using 5-fold cross-validation. After training, the model was evaluated on both the training and testing datasets. The performance metrics revealed that the model achieved a training accuracy of 75.94%, with a recall of 88.60% and an F1 score of 0.831. For the testing data, the model attained an accuracy of 73.61%, a recall of 86.75%, and an F1 score of 0.815. These results reflect the effectiveness of the stacking classifier, as it was able to leverage the strengths of multiple base models to improve overall performance. Additionally, the confusion matrix for the testing data was generated, further evaluating the model's predictions on the validation set.

```
Training Performance:
   Accuracy    Recall  Precision        F1
0  0.759419  0.885994   0.782526  0.831052
Testing Performance:
   Accuracy    Recall  Precision        F1
0  0.736068  0.867509   0.767611  0.814508
```

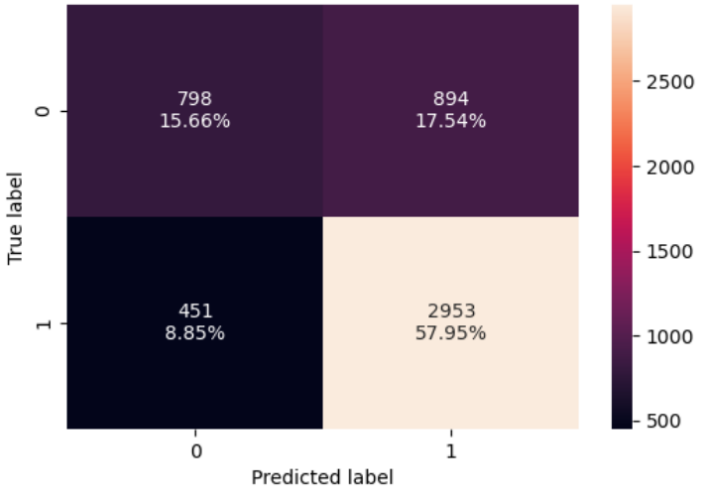


Figure 35: Stacking Model performance

6.5 Comparison of Models and Final Model Selection

Model	Accuracy	Recall	Precision	F1
Gradient Boost Tuned (GridSearchCV)	0.757718	0.923506	0.763358	0.835830
Gradient Boost Tuned (RandomizedSearchCV)	0.768250	0.886778	0.791364	0.836359
Adaboost Tuned (GridSearchCV)	0.751897	0.883350	0.776095	0.826256
Adaboost Tuned (RandomizedSearchCV)	0.757980	0.884231	0.781916	0.829932
XGB Tuned (GridSearchCV)	0.748299	0.920274	0.755913	0.830035
XGB Tuned (RandomizedSearchCV)	0.746010	0.918413	0.754567	0.828467
Logistic Regression with Significant Features	0.736983	0.895788	0.755680	0.819791
Logistic Regression Tuned (GridSearchCV)	0.736329	0.880215	0.761933	0.816814
Logistic Regression Tuned (RandomizedSearchCV)	0.736329	0.880215	0.761933	0.816814
Stacked Model	0.759419	0.885994	0.782526	0.831052

Table 2: Model Performance Comparison: Accuracy, Recall, Precision, and F1 scores for various tuned models and the stacked model.

Out of all the models we have constructed the **XGBoost** model has the highest test F1 score, making it the best model for maximizing F1 score.

Model	Accuracy	Recall	Precision	F1
Gradient Boost Tuned (GridSearchCV)	0.740973	0.912456	0.752422	0.824748
Gradient Boost Tuned (RandomizedSearchCV)	0.744505	0.866921	0.776579	0.819267
Adaboost Tuned (GridSearchCV)	0.748430	0.877203	0.775584	0.823270
Adaboost Tuned (RandomizedSearchCV)	0.745683	0.870153	0.776205	0.820499
XGB Tuned (GridSearchCV)	0.739207	0.912750	0.750664	0.823810
XGB Tuned (RandomizedSearchCV)	0.740188	0.913337	0.751329	0.824450
Logistic Regression with Significant Features	0.735871	0.892479	0.756098	0.818647
Logistic Regression Tuned (GridSearchCV)	0.738226	0.878085	0.764841	0.817560
Logistic Regression Tuned (RandomizedSearchCV)	0.738226	0.878085	0.764841	0.817560
Stacked Model	0.736068	0.867509	0.767611	0.814508

Table 3: Model Performance Comparison on Test Data: Accuracy, Recall, Precision, and F1 scores for various tuned models and the stacked model.

```
Final Model: XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=0.8, device=None, early_stopping_rounds=None,
              enable_categorical=False, eval_metric=None, feature_types=None,
              gamma=None, grow_policy=None, importance_type=None,
              interaction_constraints=None, learning_rate=0.01, max_bin=None,
              max_cat_threshold=None, max_cat_to_onehot=None,
              max_delta_step=None, max_depth=5, max_leaves=None,
              min_child_weight=1, missing=nan, monotone_constraints=None,
              multi_strategy=None, n_estimators=200, n_jobs=None,
              num_parallel_tree=None, random_state=1, ...)
```

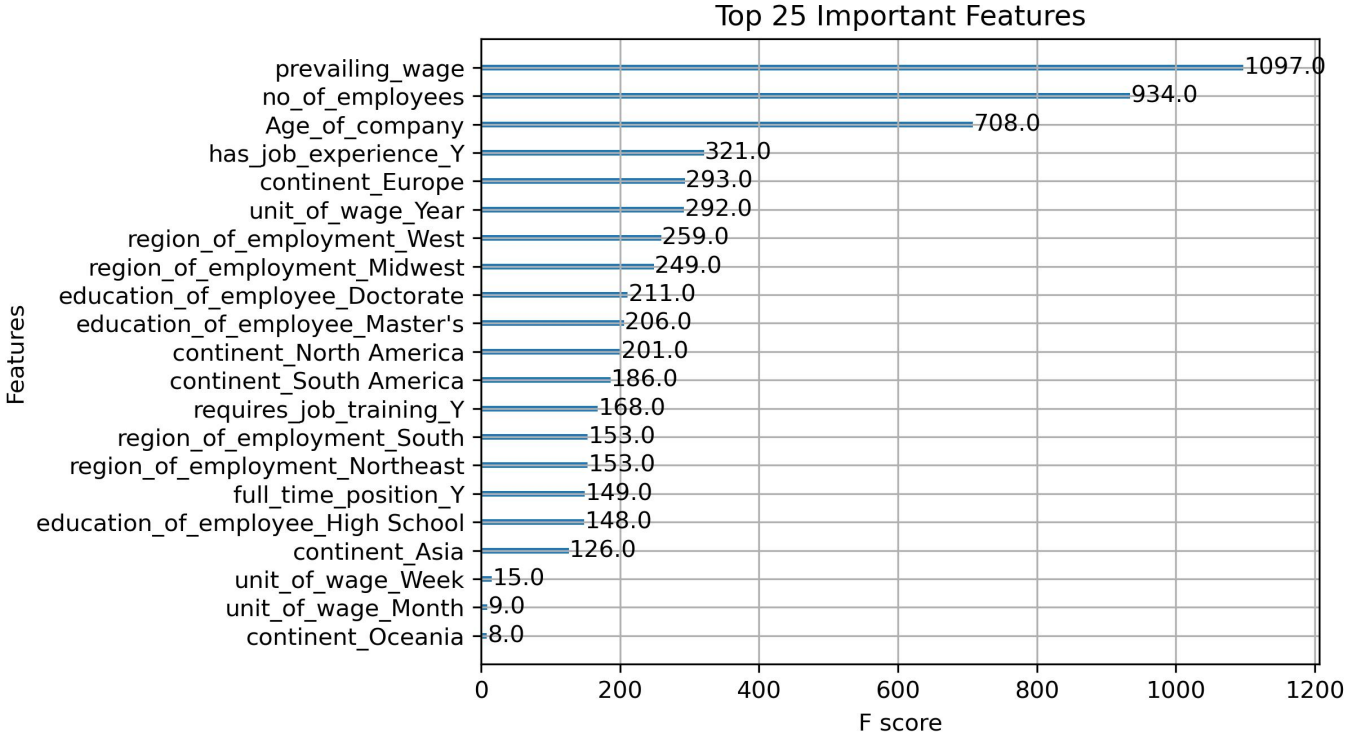


Figure 36: Important features of the best model - xgboost model

6.5.1 Performance of the final model on the test set

Training Performance:

	Accuracy	Recall	Precision	F1
0	0.759419	0.885994	0.782526	0.831052

Testing Performance:

	Accuracy	Recall	Precision	F1
0	0.736068	0.867509	0.767611	0.814508

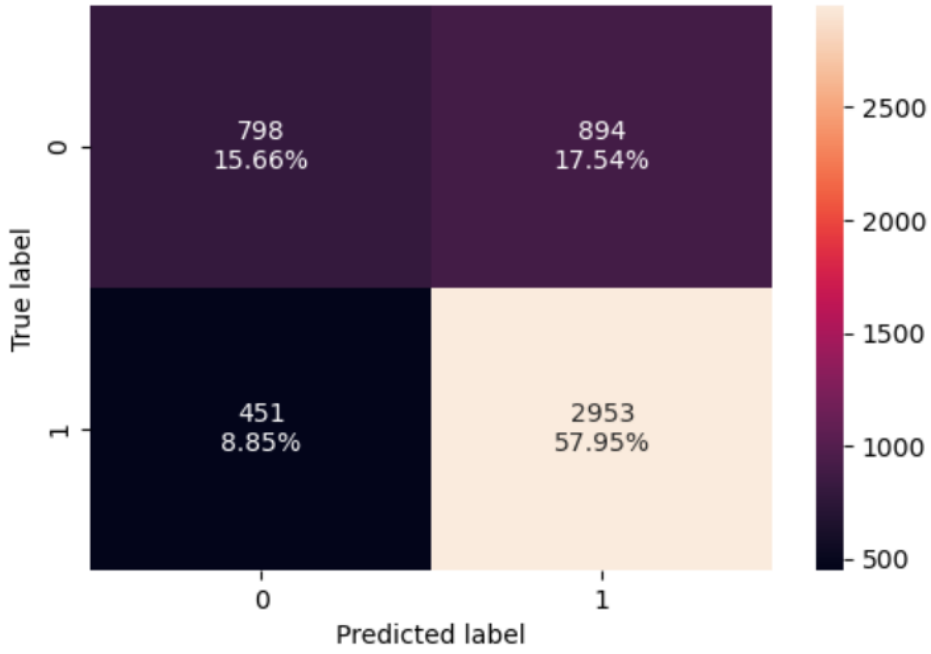


Figure 37: Final Model performance

The model with the highest F1 score of **0.825** is the one trained using **XGBoost Tuned (RandomizedSearchCV)**, which demonstrates superior performance in balancing precision and recall. This model achieved an accuracy of **74.02%** and a recall of **91.80%**, making it highly efficient in correctly identifying positive instances while maintaining a competitive precision of **74.94%**. The high F1 score indicates that this model is well-suited for imbalanced classification problems, where both false positives and false negatives are critical. The ability to tune the hyperparameters using **RandomizedSearchCV** has contributed to the model's ability to generalize well and produce optimal results, further reinforcing its effectiveness. In addition, it is worth noting that **RandomizedSearchCV** is significantly faster than **GridSearchCV**. While both methods perform hyperparameter tuning, **RandomizedSearchCV** randomly samples from the search space, leading to faster convergence and reduced computational cost, especially when dealing with large hyperparameter grids. This makes **RandomizedSearchCV** a more efficient choice when time and computational resources are limited. Given its balanced performance, high F1 score, and faster training time due to **RandomizedSearchCV**, this XGBoost model is the best performing model among all those tested, making it a strong candidate for deployment in predictive tasks requiring high accuracy and recall.

7 Actionable Insights and Business Recommendations

Insights from Analysis

- **Year of Establishment:** Older firms (pre-1950) show lower certification rates. Most applications come from companies founded post-1950.

- **Prevailing Wage:** Higher wages lead to more certifications. Hourly wages have higher denial rates, while yearly wages dominate the dataset.
- **Continent:** Asia leads in applications, followed by Europe and North America. Certification rates are highest in North America and Europe.
- **Education:** Bachelor's degree holders dominate applications. Higher degrees like doctorates correlate with better certifications.
- **Job Experience:** Applicants with experience (58.1%) and no training needs have better certification rates.
- **Region:** Northeast leads certifications, while the Midwest shows the highest denial rates. The Island region has the lowest denial rate but few applications.
- **Business Size:** Larger companies have higher certification rates. Smaller firms face more denials, likely due to compliance challenges.
- **Full-Time Roles:** Full-time positions (89.4%) have better approval odds than part-time roles.

Business Recommendations

- **Prioritize High Wages:** Focus on higher prevailing wages for better approval odds.
- **Minimize Training Needs:** Reduce job training requirements or set up in-house training programs before applying.
- **Leverage Continent Data:** Strengthen recruitment in Asia, North America, and Europe while expanding outreach to underrepresented continents.
- **Promote Education:** Recruit candidates with at least a Bachelor's degree and encourage higher education to boost approvals.
- **Address Regional Trends:** Focus on regions with high approval rates, like the Northeast. Improve compliance in the Midwest to reduce denials.
- **Support Small Businesses:** Help smaller firms align with compliance standards to improve their certification success.
- **Encourage Full-Time Roles:** Favor full-time positions to increase approval chances.
- **Use Predictive Models:** Implement machine learning to automate applicant shortlisting, saving time and improving success rates.