EasyVisa - Problem Statement

PGP - Data Science & Business Analytics
December 6, 2024

Leslieane Beltran

Contents/Agenda

Executive Summary

- Businesses face growing challenges in attracting qualified talent domestically and internationally, leading to the need of effective solutions to meet workforce demands while complying with labor regulations
- The Immigration and Nationality Act (INA) allows foreign workers to fill workforce shortages while monitoring US labor market conditions through oversight by the Office of Foreign Labor Certification (OFLC)
- EasyVisa has been hired to streamline visa approval processes by identifying key factors influencing case outcomes
- EasyVisa will be using a machine learning-based classification model
- The model will allow for a more efficient process by predicting visa approval likelihood, aiding decision-making, and recommending profiles for certification or denial to support OFLC’s mission effectively

Business Problem Overview

- The OFLC faces a surge in labor certification applications, processing nearly 776,000 in FY 2016, with demand increasing annually.
- The manual review of applications has become time consuming, impacting efficiency and decision-making speed.
- Ensuring that the case follow all the rules while still getting through the workload is difficult
- Identifying key factors influencing visa approvals is critical to make the decision-making process more efficient and allocate resources effectively.
- A data-driven machine learning model is required to assist in automating the applicant shortlisting process and improve outcomes.

Solution Approach

The following describe the solution approach:

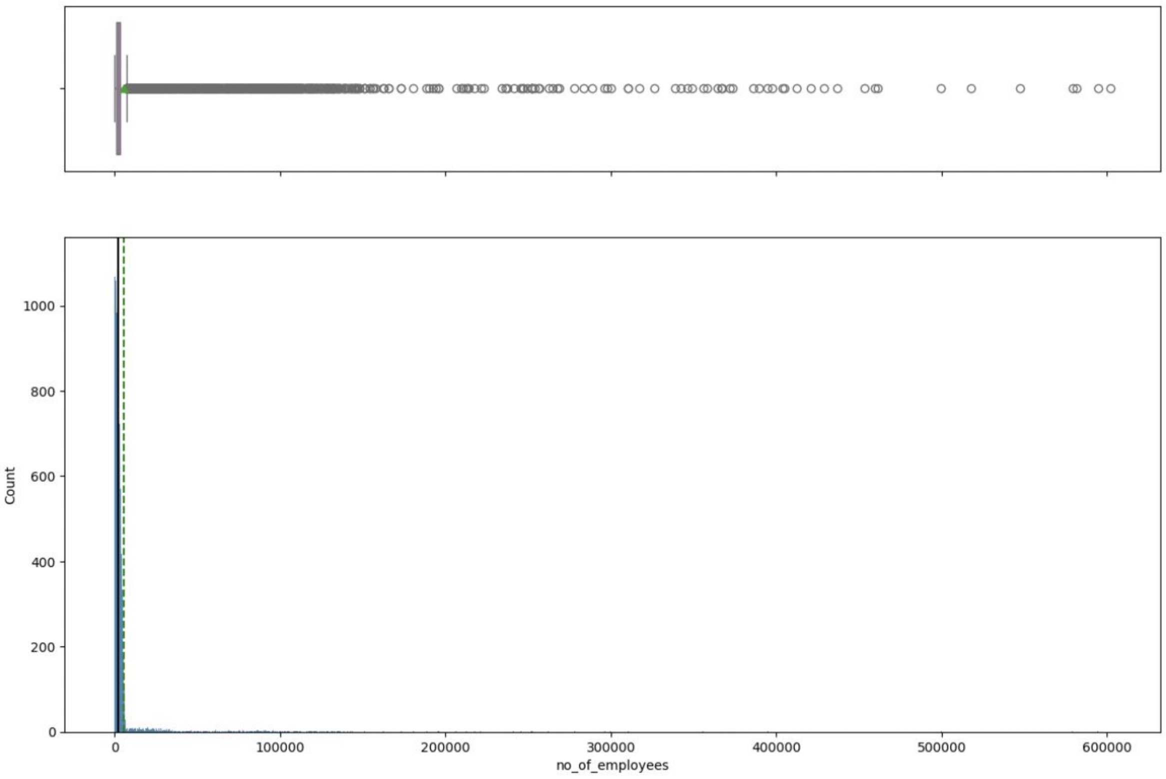
- Analyze historical data from past applications to discover patterns and identify the key factors that influence visa approval outcomes.
- Build a machine learning classification model to accurately predict the likelihood of visa approval for each applicant based on relevant factors.
- Implement an automated system to prioritize applications with higher chances of approval, lowering manual effort and speeding up the process.
- Provide the OFLC with detailed insights and recommendations to help them efficiently certify or deny applications while maintaining compliance.
- Design a flexible framework that can adapt to increasing application volumes and evolving needs over time.

```
<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
```

-There are 9 columns of the dtype object, 1 column of the dtype float64, and 2 columns of the dtype int64.
- There are no missing values nor any duplicates in the data.

EDA Results - Univariate Analysis

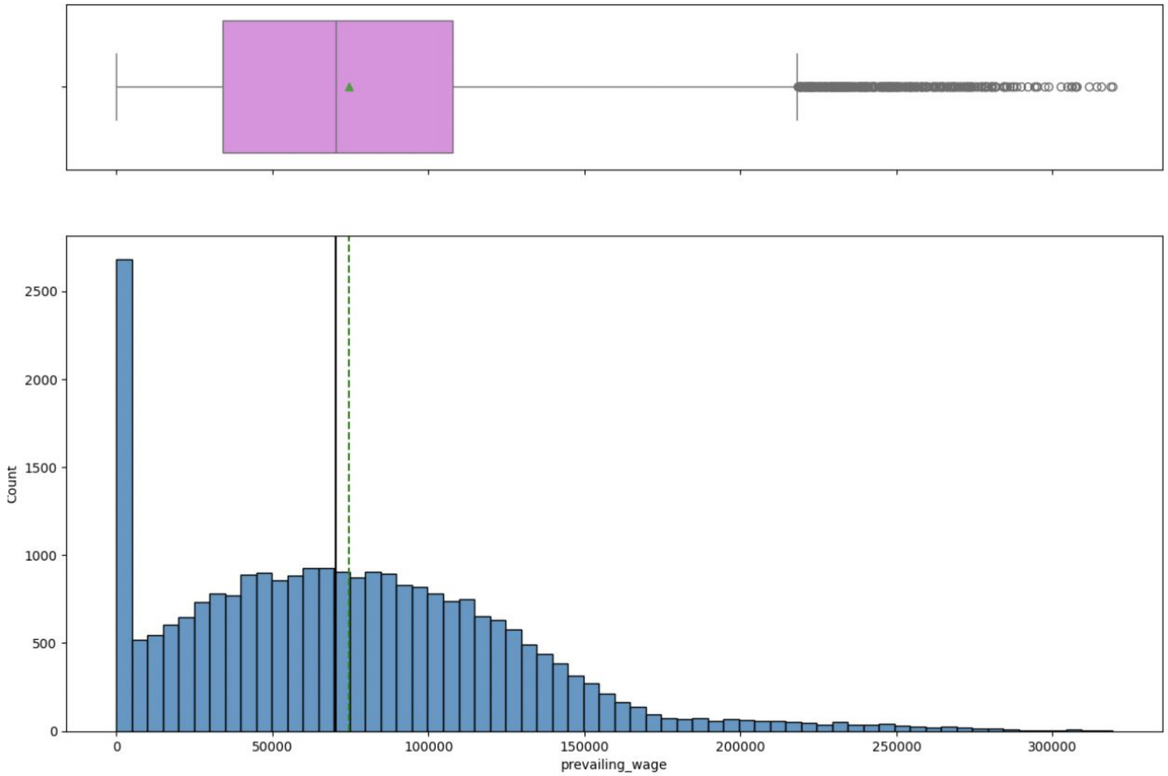
Observations on number of employees



The data and graphs above, indicate the distribution of companies by number of employee is heavily right skewed.

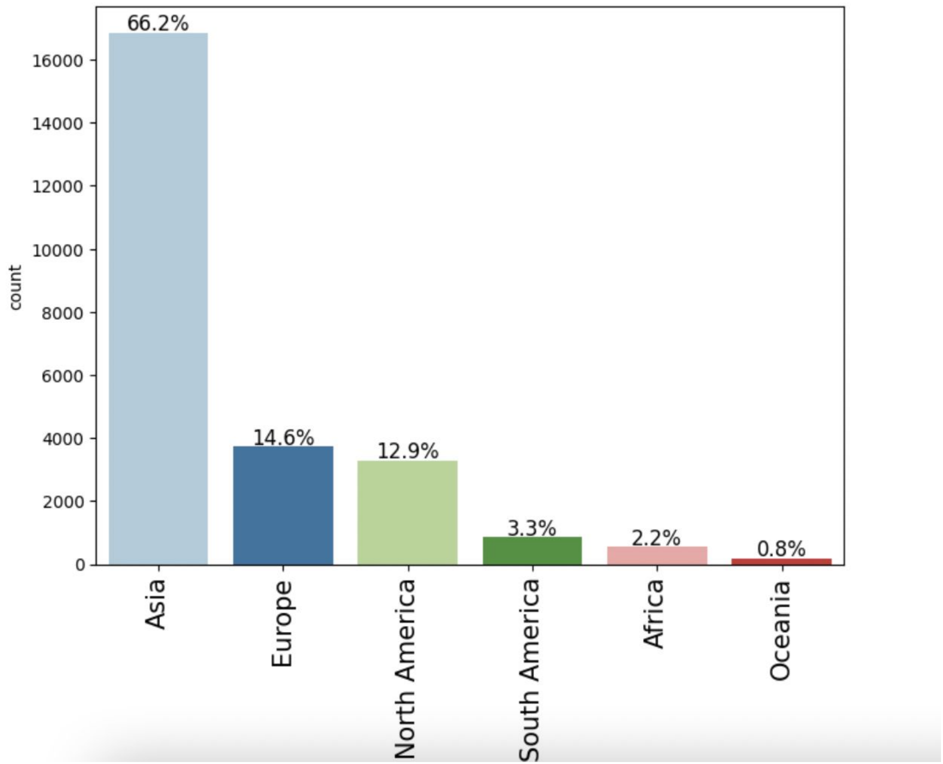
Observations on prevailing wage

- Outliers
- median is the 50th percentile
-



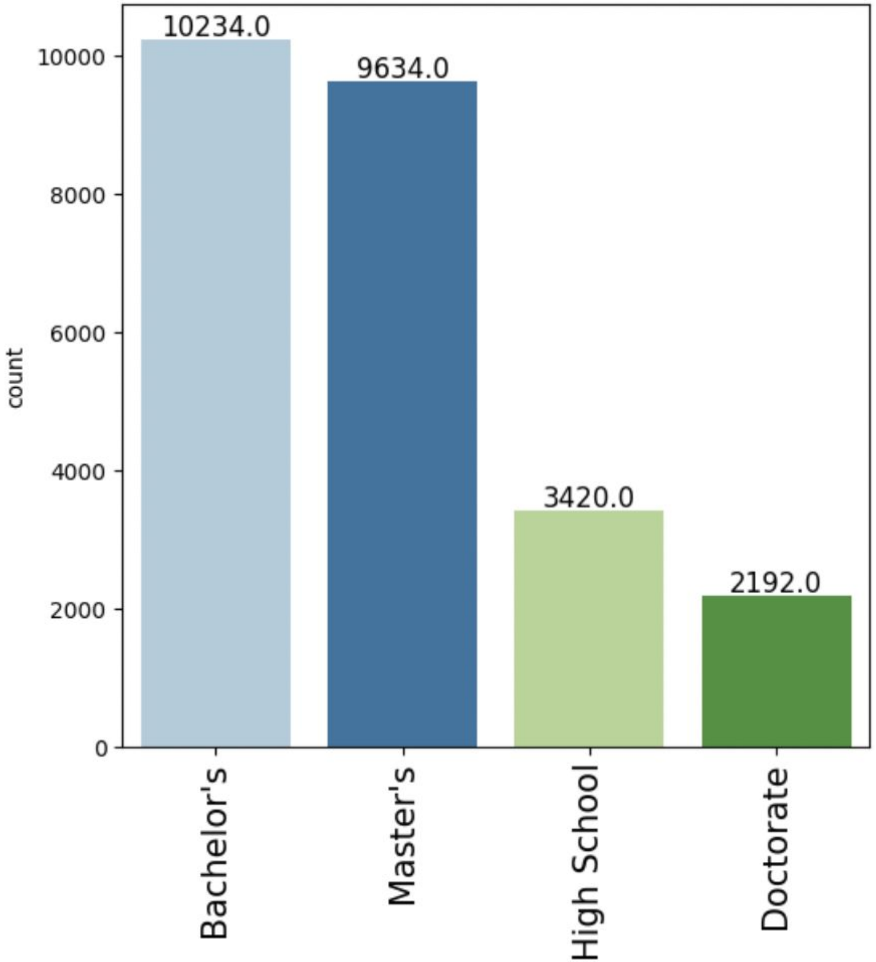
Observations on continent

- most of the applicants are from asia (more than half)

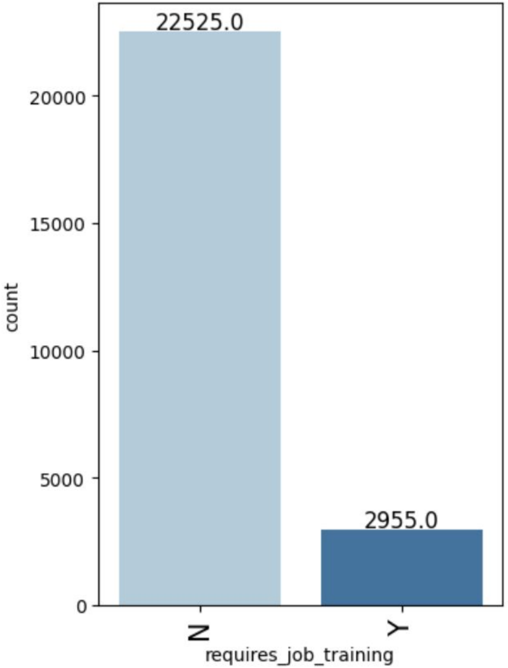


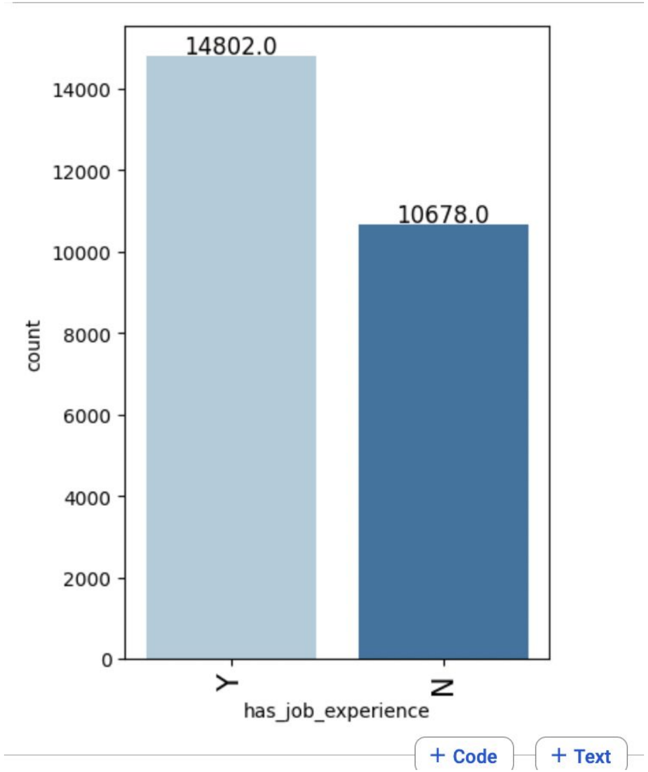
Observations on education of employee

- Most visa applicants in the dataset have a Bachelor's degree. A substantial amount have Master’s.



Observations on job experience and job training

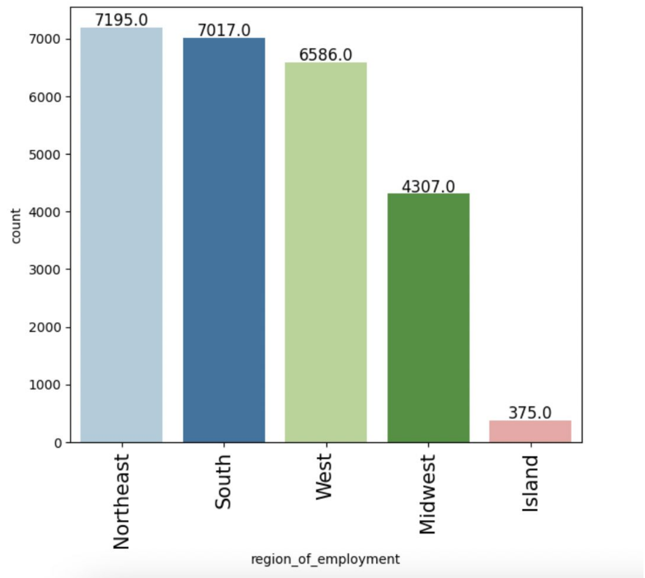


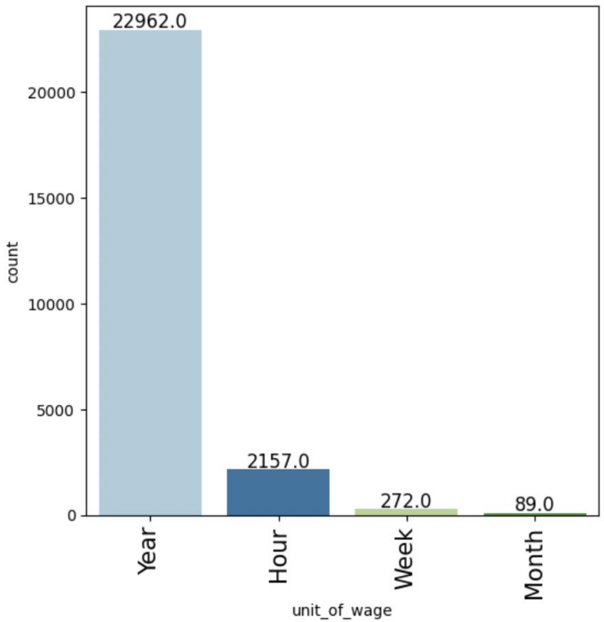


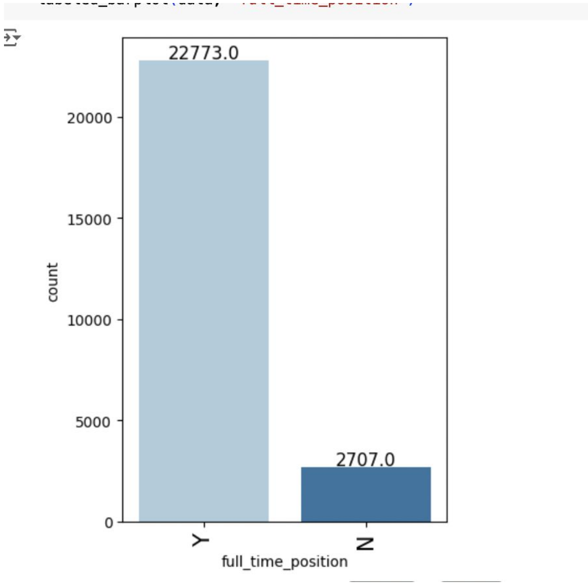
Majority of the applicants do not require job training, meaning they are well equipped for a job.

More Visa applicants do have job experience, but a good amount still do not have experience.

Observations on region of employment, unit of wage & full time position







Most are paid in yearly wage

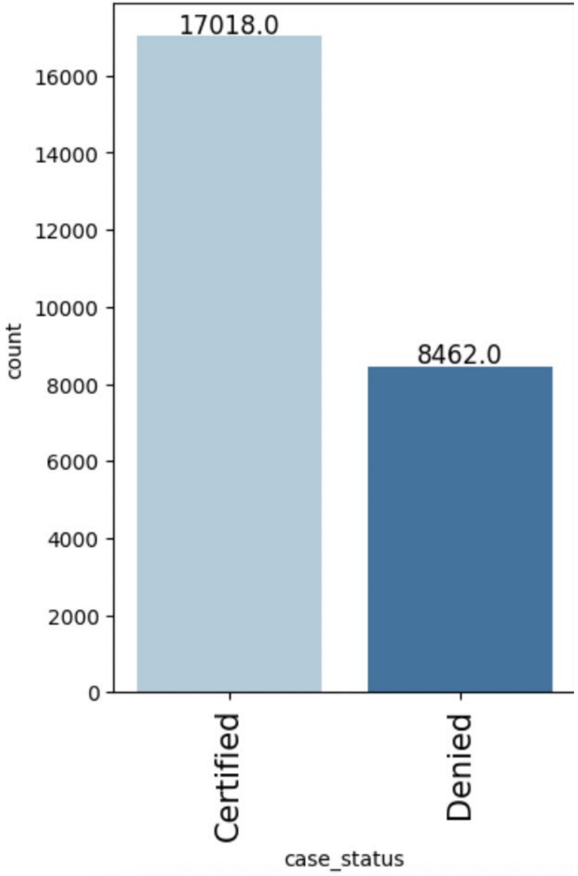
Most are full time positions

Northeast, South, and West have almost equal percentages of applicants. The Island region

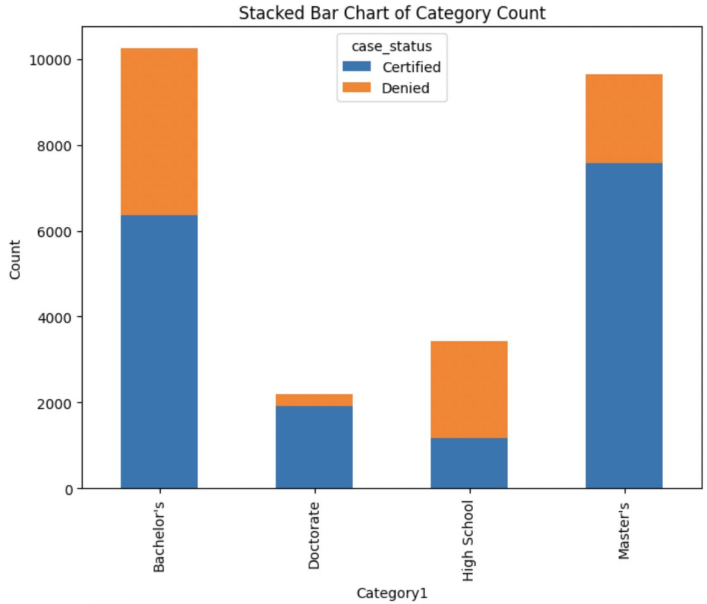
Might be an outlier

Observations on case status

It does not seem that guests who require a parking space have a substantial effect on cancellations.



case_status	Certified	Denied
education_of_employee		
Bachelor's	6367	3867
Doctorate	1912	280
High School	1164	2256
Master's	7575	2059

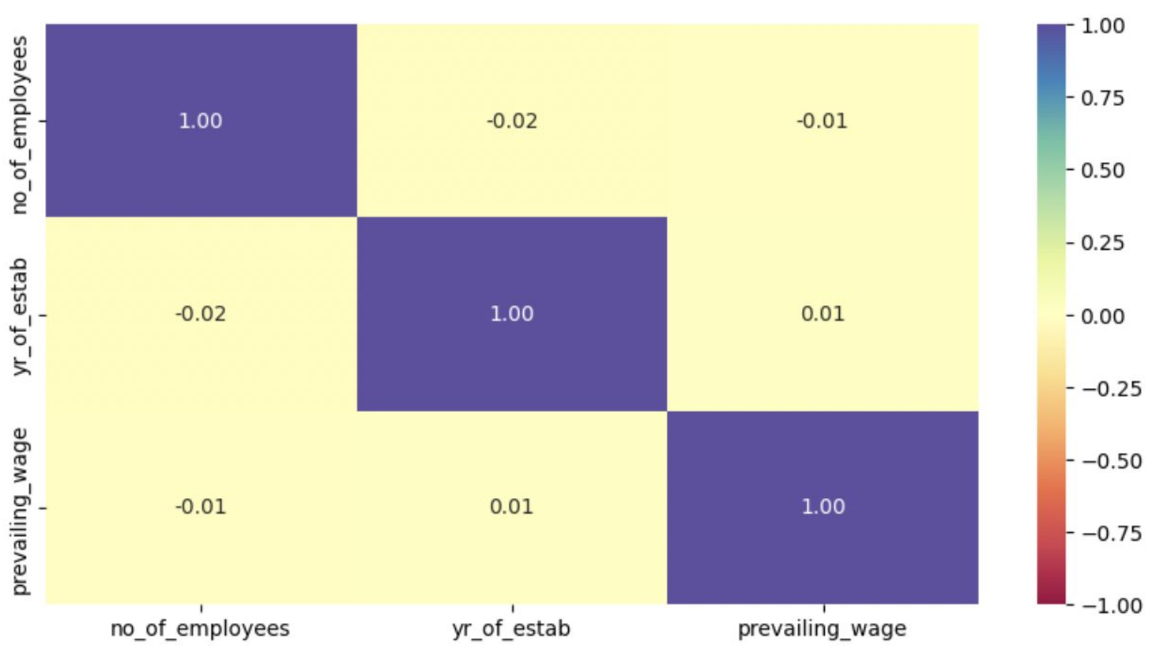


EDA Results - Bivariate Analysis

Heat Map

The correlation coefficient between these variables is very close to zero

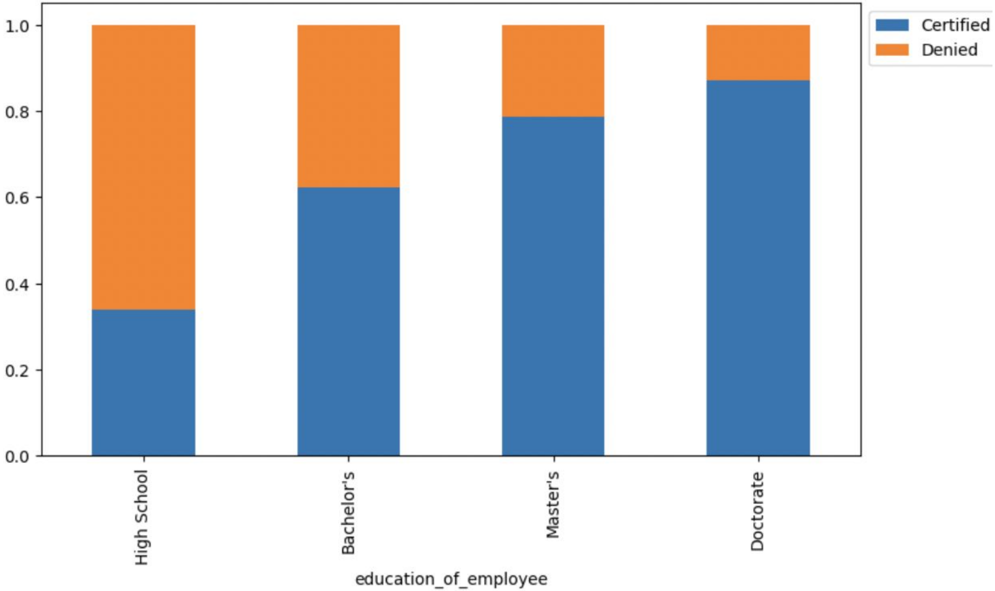
This indicates a weak correlation or almost no linear relationship



Higher education may want to travel abroad for a well-paid job.

The higher the degree one has, the more likely your visa will be accepted

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



Regions and special requirements

Requirement for applicants who have passed high school is most in the South region, followed by Northeast region.

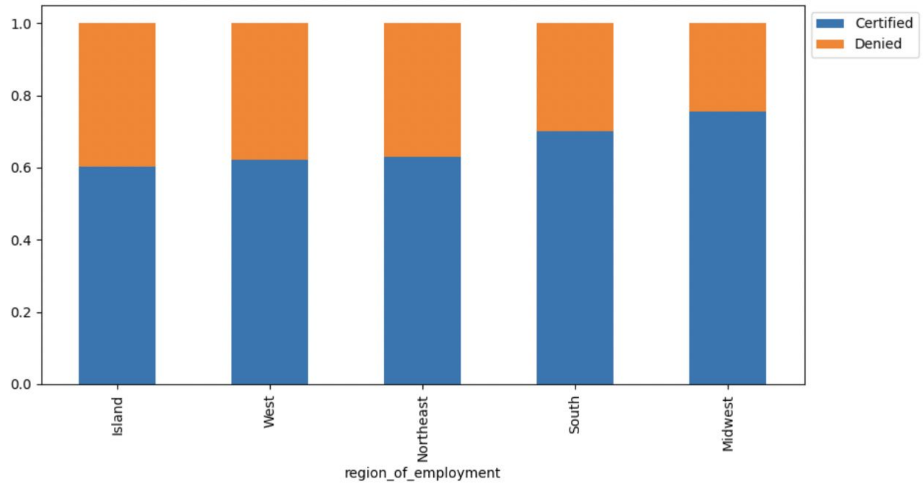
Requirement for Bachelor's is mostly in South region, followed by West region.

For Master's is most in Northeast region, followed by South region. The requirement for Doctorate's is mostly in West region, followed by Northeast region.

percentage of visa certifications across each region

Midwest has highest number of visa certifications but its not the most picked from any level of education

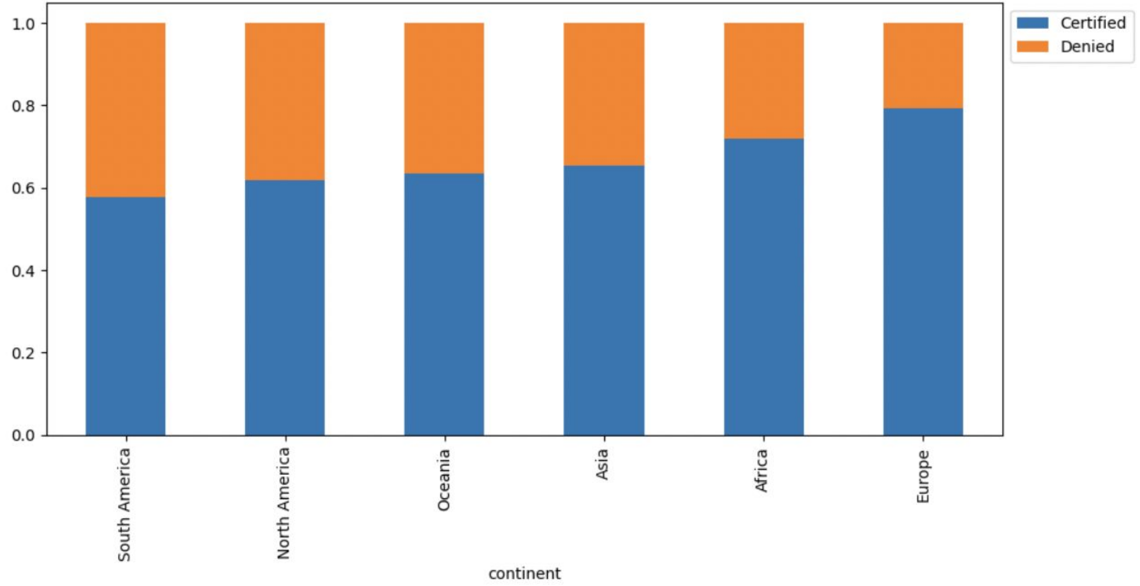
case_status region_of_employment	Certified	Denied	All
All	17018	8462	25480
Northeast	4526	2669	7195
West	4100	2486	6586
South	4913	2104	7017
Midwest	3253	1054	4307
Island	226	149	375



how the visa status vary across different continents.

```
stacked_barplot(data, 'continent', 'case_status') ## Complete the code to plot stacked barplc
```

```
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
------------------------------------------------------------------------------------------------
```



ɔpe has the highest visa acceptance followed by africa

Europe has the highest visa acceptance followed by africa

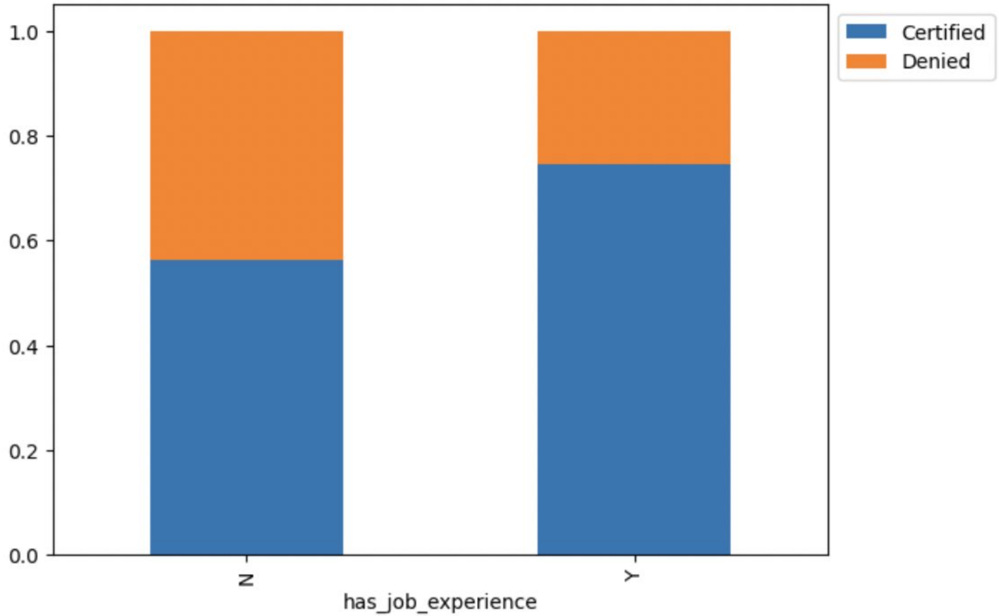
asia has the 3rd highest visa certification but has the highest no of application

Work experience and visa certification

People with job experience have a higher chance of their visa being certified.

but a good amount of people without have a job experience also got their visa certified

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

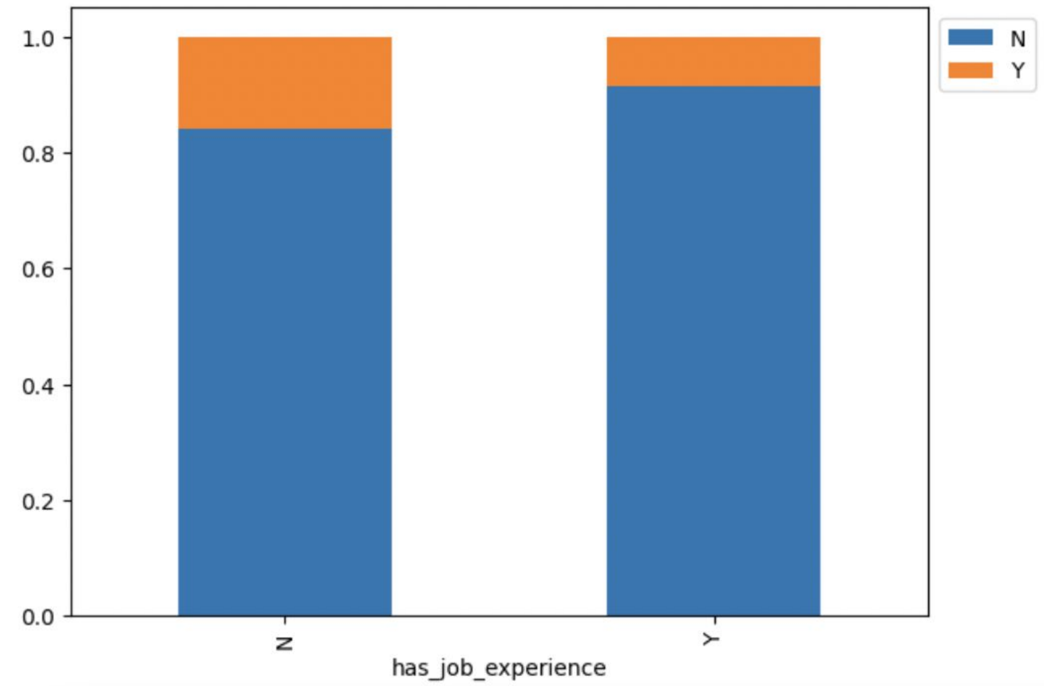


Prior work experience and job training?

minimum percentage of applicants dont require job training

but less if they have job experience

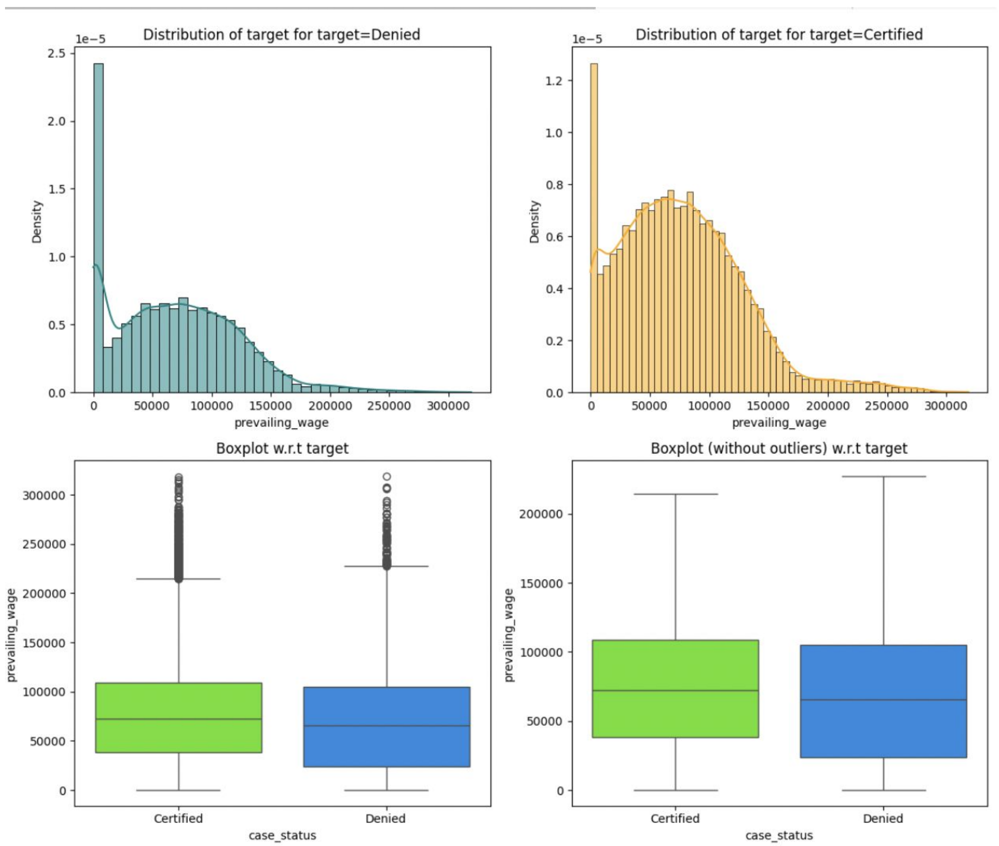
```
requires_job_training      N     Y    All
has_job_experience
All                    22525  2955  25480
N                       8988  1690  10678
Y                      13537  1265  14802
------------------------------------------------------------------------------------------------------------------------
```



visa status and the prevailing wage

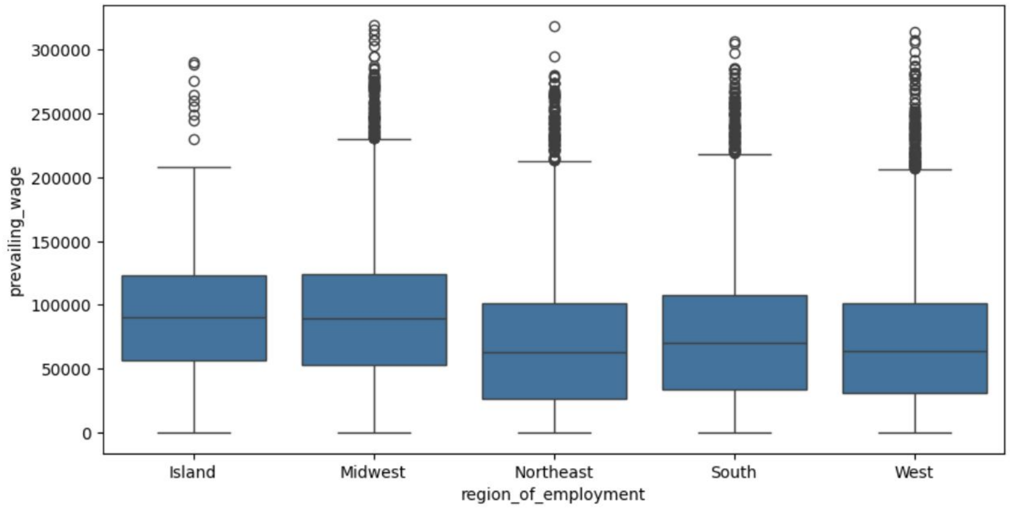
the median wage for the certified applications is slightly higher than the denied application

but is the prevailing wage similar across all region



Prevailing wage across regions of the US

midwest and inland have slightly higher median wages compared to other regions



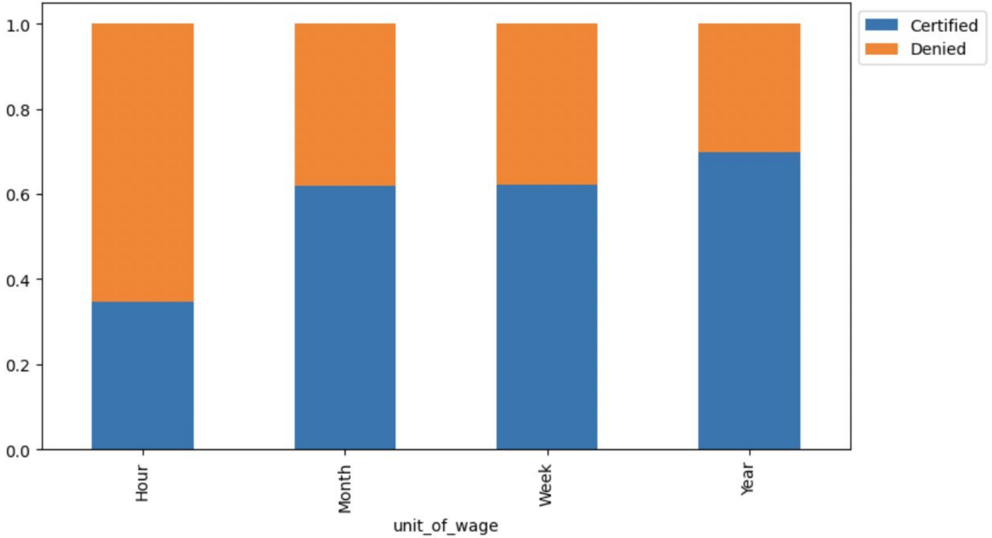
est and inland have slightly higher median wages compared to other regions

Unit of wage impact on visa applications getting certified

yearly is most likely to be certified.

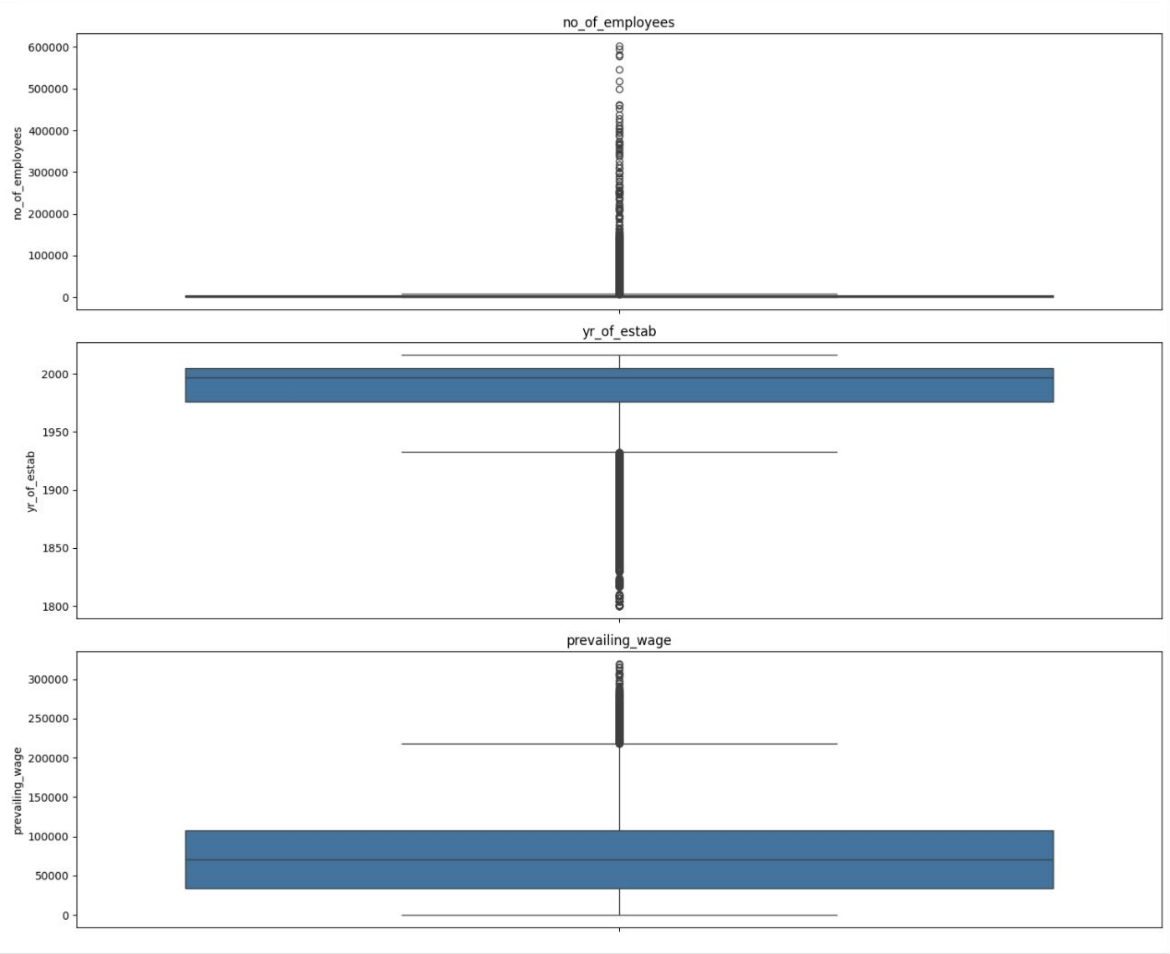
Week and month's percentage of employees certified is almost the same

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



Data Preprocessing

Outlier Detection



Model Building

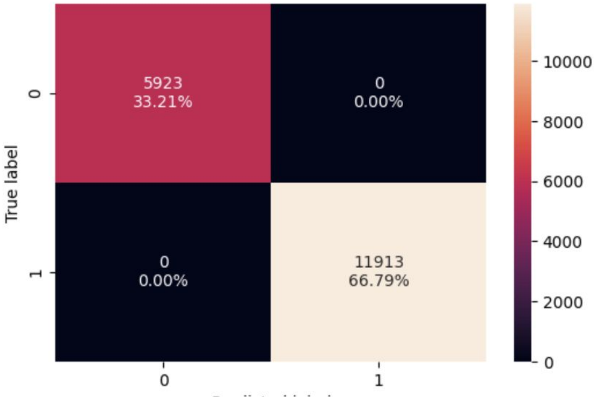
Decision Tree - Model Building and Hyperparameter Tuning

Relatively higher accuracy and F1-score for class 1,

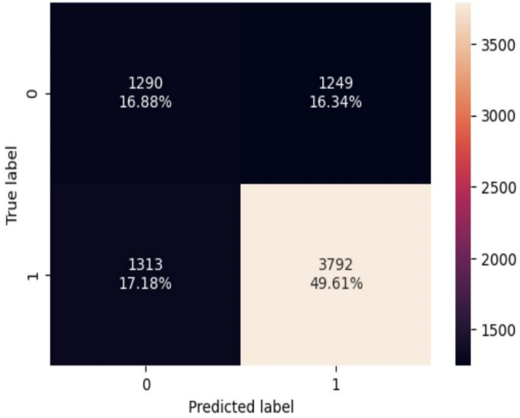
Performance on class 0 is lower, as evident from the lower recall and F1-score for class 0.

Decision Tree Model

	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

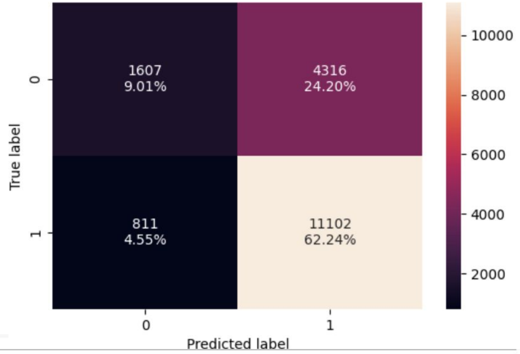


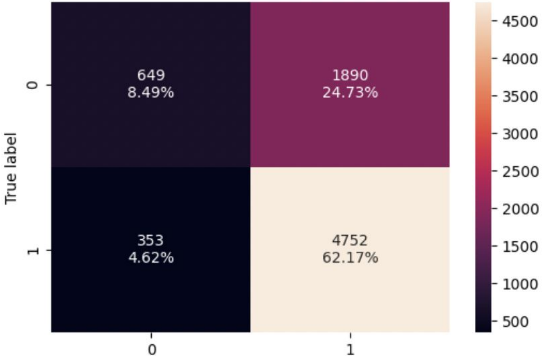
	Accuracy	Recall	Precision	F1
0	0.664835	0.742801	0.752232	0.747487



Hyperparameter Tuning

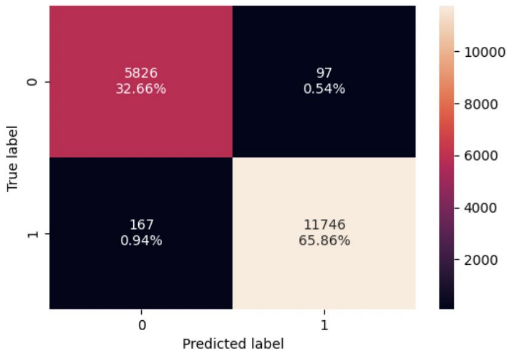
	Accuracy	Recall	Precision	F1
0	0.712548	0.931923	0.720067	0.812411



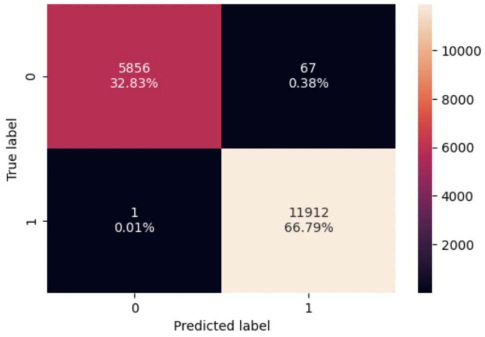


	Accuracy	Recall	Precision	F1
0	0.706567	0.930852	0.715447	0.809058

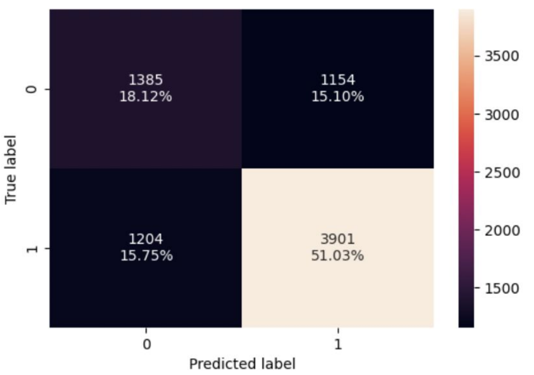
Bagging - Model Building and Hyperparameter Tuning



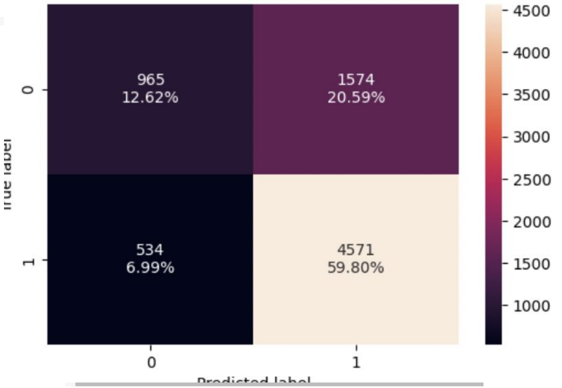
	Accuracy	Recall	Precision	F1
0	0.985367	0.986317	0.991729	0.989016



	Accuracy	Recall	Precision	F1
0	0.996187	0.999916	0.994407	0.997154



	Accuracy	Recall	Precision	F1
0	0.691523	0.764153	0.771711	0.767913



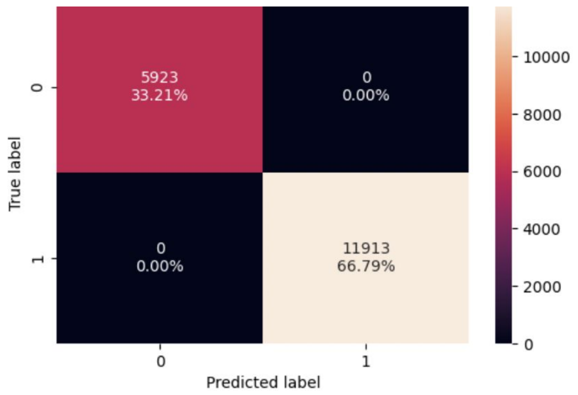
	Accuracy	Recall	Precision	F1
0	0.724228	0.895397	0.743857	0.812622

Bagging Classifier

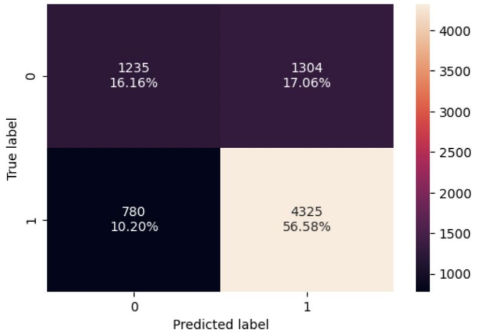
The tuned bagging classifier model shows good performance with the test set showing an F1 score of .81. The first class is still good.

Tuned

Rf

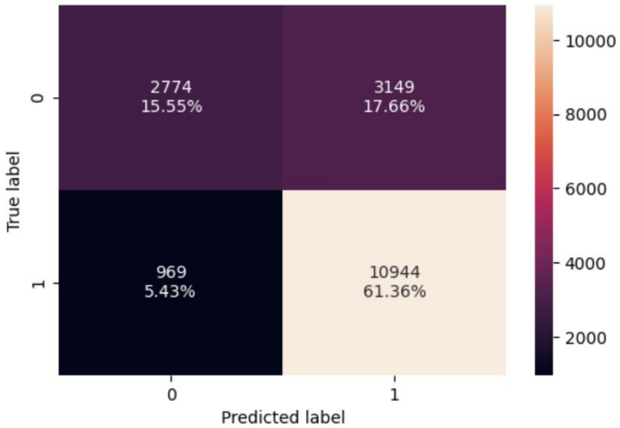


	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

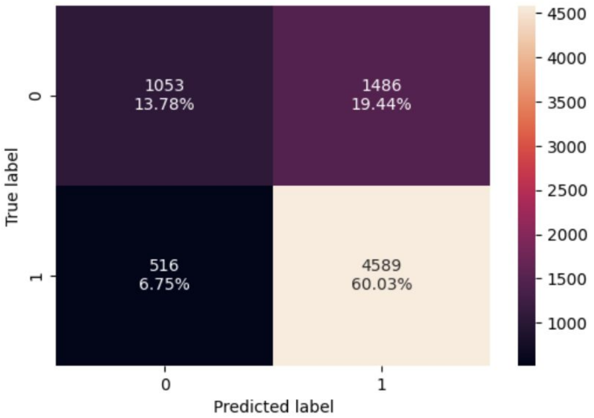


	Accuracy	Recall	Precision	F1
0	0.727368	0.847209	0.768343	0.805851

Rf tuned



	Accuracy	Recall	Precision	F1
0	0.769119	0.91866	0.776556	0.841652



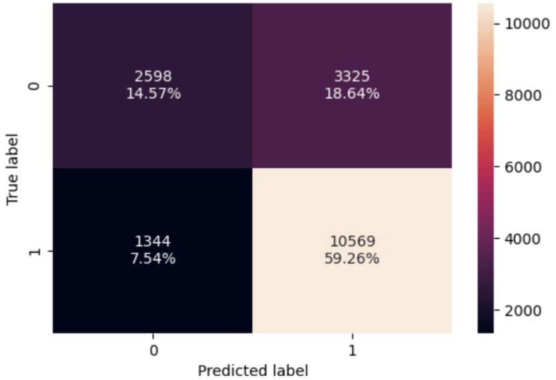
	Accuracy	Recall	Precision	F1
0	0.738095	0.898923	0.755391	0.82093

The rf tuned F1 score sits at .82 vs the rf F1 score of .80. Based on these F1 scores, the rt tuned model will be more accurate.

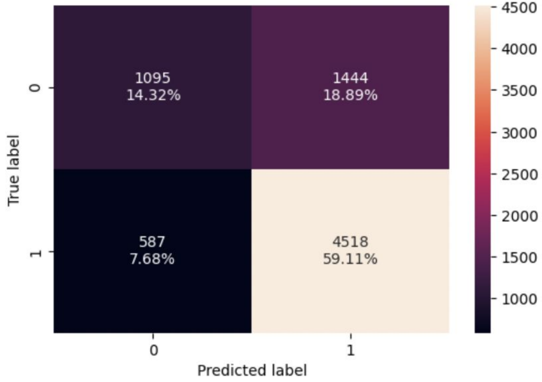
Boosting - Model Building and Hyperparameter Tuning

An F1 score of 0.81 for the AdaBoost Classifier means the model is relatively effective vs the F1 score of 0.78 for the tuned version.

ab

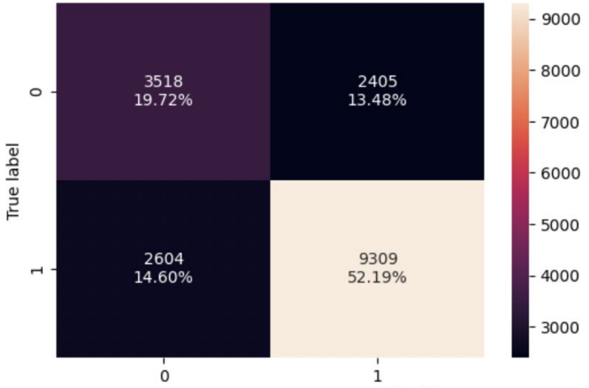


	Accuracy	Recall	Precision	F1
0	0.738226	0.887182	0.760688	0.81908

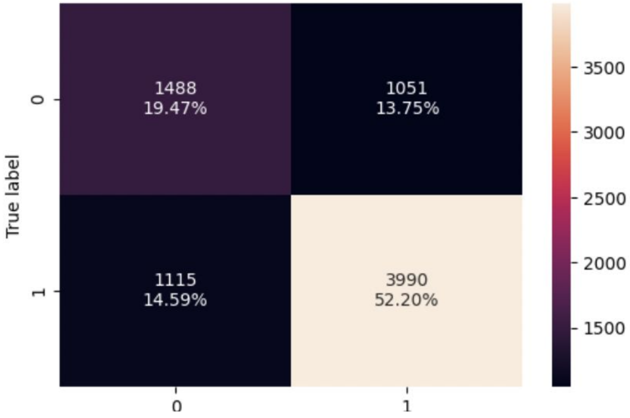


	Accuracy	Recall	Precision	F1
0	0.734301	0.885015	0.757799	0.816481

tuned



	Accuracy	Recall	Precision	F1
0	0.719163	0.781415	0.79469	0.787997

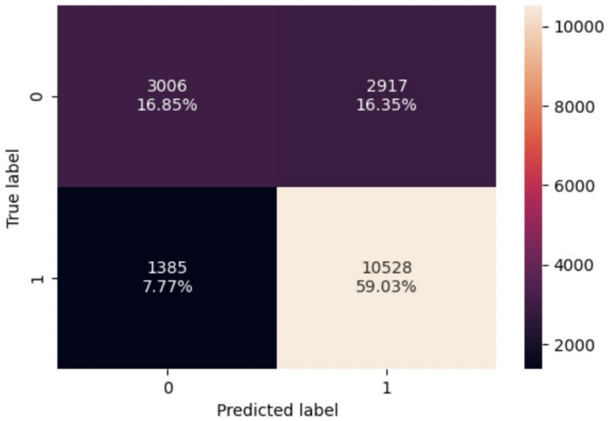


	Accuracy	Recall	Precision	F1
0	0.716641	0.781587	0.79151	0.786517

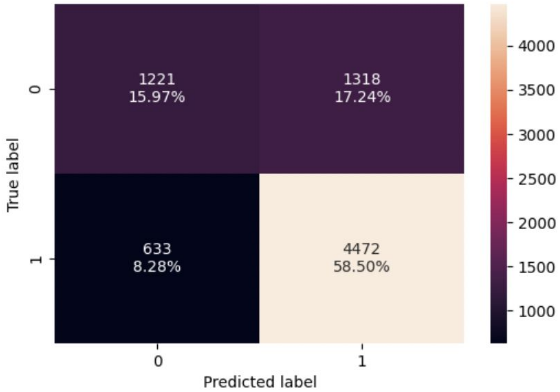
Gradient Boosting Classifier

Gb model F1 score of .82 vs the tuned version F1 score of .81 suggests that the original model performed slightly better. Shows that tuning doesn't always guarantee performance improvements.

gb

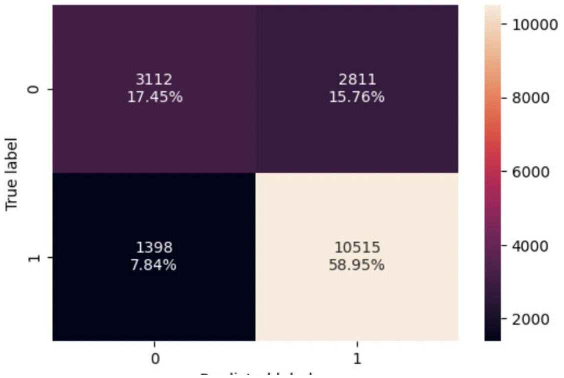


	Accuracy	Recall	Precision	F1
0	0.758802	0.88374	0.783042	0.830349

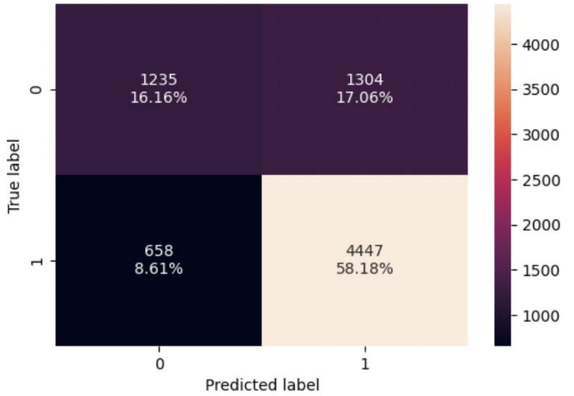


	Accuracy	Recall	Precision	F1
0	0.744767	0.876004	0.772366	0.820927

Tuned



	Accuracy	Recall	Precision	F1
0	0.764017	0.882649	0.789059	0.833234



	Accuracy	Recall	Precision	F1
0	0.743328	0.871107	0.773257	0.81927

Model Performance Comparison and Final Model Selection

Model performance Summary

Training performance comparison:

	Decision Tree	Random Forest	Tuned Random Forest	Adaboost Classifier	Tuned Adaboost Classifier	Gradient Boost Classifier	Tuned Gradient Boost Classifier
Accuracy	1.0	1.0	0.769119	0.738226	0.719163	0.758802	0.764017
Recall	1.0	1.0	0.918660	0.887182	0.781415	0.883740	0.882649
Precision	1.0	1.0	0.776556	0.760688	0.794690	0.783042	0.789059
F1	1.0	1.0	0.841652	0.819080	0.787997	0.830349	0.833234

Based on the training performance summary, it would be best to go with the decision tree or random forest model. These models have perfect scoresvfor precision, recall, and accuracy.

Insights & Recommendations

Recommendations

- The analysis shows that education level, job experience, and prevailing wage are key factors in predicting application approvals.
- OFLC should use these insights to streamline its pre-screening process. The simplicity of the Decision-Tree model also hints at potential biases against less skilled or entry-level applicants, which should be addressed to ensure fairness and transparency.

To efficiently allocate resources for screening applications likely to be approved, the OFLC should:

- Prioritize applications by education level, reviewing those with higher qualifications first.
- Sort by job experience, reviewing applicants with relevant experience first.
- Separate applications by wage type (hourly vs. annual), rank each group by prevailing wage, and prioritize salaried jobs from highest to lowest wage.