Office of Foreign Labor Certification (OFLC)
& EasyVisa

Machine Learning Based Solution using Ensemble Techniques

Julie Kistler _ January 18, 2024

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix

Executive Summary

Businesses in the United States face a growing demand for human resources, leading to challenges in identifying and attracting the right talent. The Immigration and Nationality Act (INA) permits foreign workers to address this demand. The Office of Foreign Labor Certification (OFLC) administers immigration programs, processing applications for temporary and permanent labor certifications.

To streamline the visa approval process, EasyVisa, hired by OFLC, seeks a machine learning solution. As a data scientist, the goal is to analyze provided data and recommend a classification model that facilitates visa approvals based on key influencing factors.

The model highlights the significance of three key features: education level, job experience, and prevailing wage. These factors play a crucial role in determining visa approval outcomes.

Key Features for Model Importance:

- The top factors influencing visa approval are the education level of the employee, job experience, and prevailing wage. These features significantly contribute to the model's decision-making process.

Executive Summary, Cont.

Insights cont:

- Best profile for Visa Approval:
  - Education Level: Higher the education the better – most approvals require a Bachelors' degree or better – Doctorate and Master's degree are highly preferred
  - Job Experience: Job experience is essential
  - Prevailing Wage: Average prevailing wage is around $70k.
  - Additional Factors: Applicants from Europe, Africa, and Asia, with yearly unit of wage, and applying to the Mid-West region have higher chances of approval.

- Best profile for Visa Denied:
  - Education Level: High School Education and/or no degree
  - Job Experience: Lack of job experience
  - Prevailing Wage: Average prevailing wage is around $65k.
  - Additional Factors: Applicants with hourly unit of wage, from Oceania, North America, and South America are more likely to face denial.

Executive Summary, Cont.

Recommendations:

- Utilize the XGBoostClassifier Tuned model for its outstanding performance, achieving an F1 Score of 83% for the training set and 82% for the testing set. This model is recommended for accurate predictions in the visa approval process.

Recommended Further Analysis:

  - **Additional Data Collection:** Gather more information from both employers and employees to extract deeper insights
    - **Job Type and Prevailing Wage Data**: Analyze prevailing wage data based on job types such as IT, service, administration, etc.
    - **Required Education and Years of Experience**: Understand the correlation between visa approval and specific education levels and work experience.
    - **Industry Sector Analysis**: Explore visa approval trends within different industry sectors.
    - **Regional Analysis**: Segment data based on the type of company (industry sector) and size in different regions.
    - **Visa Length Analysis**: Examine the impact of visa length on approval rates.
    - **Applicant Segmentation by Company Size**: Classify applicants based on company size (small, median, large) to discern patterns in approved/denied applications.

Business Problem Overview and Solution Approach

The Office of Foreign Labor Certification (OFLC) continues to grapple with a substantial surge in visa applications, leading to a growing backlog. The current manual review process is proving to be laborious and inefficient, hindering the timely processing of applications. This is impacting the overall effectiveness of OFLC's visa approval procedures.

To address the challenges faced by OFLC, EasyVisa has been hired to develop a machine learning-based solution. The proposed solution aims to streamline operations by leveraging a classification model that can analyze and categorize visa applications. EasyVisa's objectives are:

- **Facilitate the Visa Approval Process:** EasyVisa intends to enhance the efficiency of the visa approval process by automating the initial screening of applications. By leveraging machine learning algorithms, the system will identify and shortlist applicants who are more likely to receive visa approval.
- **Recommend Suitable Profiles:** The solution will go beyond mere automation and provide OFLC with actionable insights. EasyVisa's classification model will consider various factors influencing the visa approval process and recommend suitable profiles for certification or denial. This approach ensures that the decision-making process is not only expedited but also guided by data-driven insights.

In summary, the collaboration between OFLC and EasyVisa seeks to transform the visa application process by, introducing efficiency through automation and informed decision-making based on machine learning analysis. This solution is poised to alleviate the strain caused by the increasing number of applications, ultimately leading to a more streamlined and effective visa approval workflow.

Data Overview

Data Dictionary	
case_id	ID of each visa application
continent	Information of continent the employee
education_of_employee	Information of education of the employee
has_job_experience	Does the employee has any job experience? Y= Yes; N = No
requires_job_training	Does the employee require any job training? Y = Yes; N = No
no_of_employees	Number of employees in the employer's company
yr_of_estab	Year in which the employer's company was established
region_of_employment	Information of foreign worker's intended region of employment in the US
prevailing_wage	Average wage paid to similarly employed workers in a specific occupation in the area of intended employment. The purpose of the prevailing wage is to ensure that the foreign worker is not underpaid compared to other workers offering the same or similar service in the same area of employment
unit_of_wage	Unit of prevailing wage. Values include Hourly, Weekly, Monthly, and Yearly
full_time_position	Is the position of work full-time? Y = Full Time Position; N = Part Time Position
case_status	Flag indicating if the Visa was certified or denied

Data Overview cont...

Column	Dtype
case_id	object
continent	object
education_of_employee	object
has_job_experience	object
requires_job_training	object
no_of_employees	int64
yr_of_estab	int64
region_of_employment	object
prevailing_wage	float64
unit_of_wage	object
full_time_position	object
case_status	object

Rows	Columns
25480	12

- There are no duplicate values
- There are no missing values

- **9 object data types** (case_id, continent, education_of_employee, has_job_experience, requires_job_training, region_of_employment, unit_of_wage, full_time_position, case_status)
- **2 integer data types** (no_of_employees, yr_of_estab)
- **1 float data type** (prevailing_wage)

EDA Results

- Statistical Summary
    - no_of_employees, yr_of_estab, prevailing_wage columns are all numerical features - the remaining columns are objects
    - The no_of_employees has a mean of 5667 with a median of 2109 indicating the distribution of data may be skewed
    - There are negative numbers in the no_of_employees – this could indicate an error
    - The yr_of_estab has a broad range from 1800 – 2016
    - The most prevalent continent is Asia
    - The most prevalent education level is a Bachelors Degree
    - Most applicants do not require job training
    - Most applicants do have job experience
    - The most prominent region is the Northeast
    - The average prevailing wage is ~ $75,456 (Min: $2.14 / Max: $319210) – data distribution is skewed
    - Annual salary if the most prevalent unit of wage
    - Most applicants are applying for a full time job
    - A majority of applicants are certified ~ 67%

Link to Appendix slide on statistical summary of data

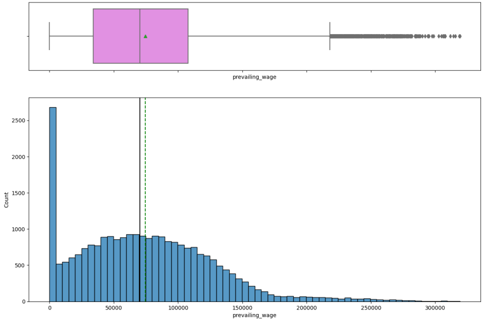
EDA Results

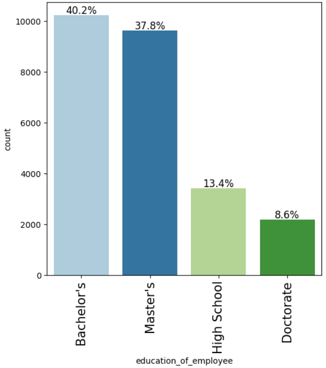
- Correct the negative values in the number of employees column
  - Assuming these negative numbers may be a result of a data entry error
  - There are 33 instances -- these were converted by using the absolute value of the numbers
- Number of observations for each unique category of the categorical variables
  - In the case_id category, there are 25,480 unique values
  - In the continent category, Asia has the most applicants at 16891 (66%)
  - In the education of employee category, Bachelor’s degree is the top with 10234 (40%)
  - In the has_job_experience category, over half have experience at 14802 (58%)
  - In the requires_job_training catagory, a vast majority do not need training at 22525 (88%)
  - In the region_of_employment catagory, the northeast the most requested at 7195 (28%) with the south close behind with 7017 (27%) – over 50% of the applicants are applying in these two regions.
  - In the unit_of_wage category, the most prevalent wage is year at 22962 (90%)
  - In the full_time_position category, the majority of applicants are full time at 22773 (89%)
  - In the case_status category, approx. 2/3 are certified at 17018 (67%)
- Removed “case_id” from data as it is not needed for analysis

Link to Appendix slide on # of observations for each unique category of data

EDA Results _ Univariate Analysis (Observations)

- Number of employees data distribution is heavily skewed right with lots of outliers
  - This may indicate there is a large variety of company sizes
- Prevailing wage data distribution is skewed right with lots of outliers
  - There are wages above the 200k mark
  - There is a large disparity between wages
  - There appears to be several wages at or close to the 0 mark – further analysis is recommended
    - There were 176 rows that have a prevailing hourly wage of less than 100
- Continent data indicated approximately 66% of all applications are originated out of Asia and 28% originated from Europe and North America
- The education of employee data reflected that approximately 87% of applicants had a higher education degree with the Bachelor's degree being the most common

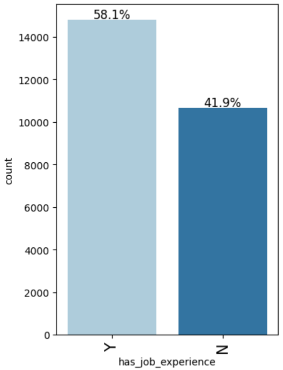


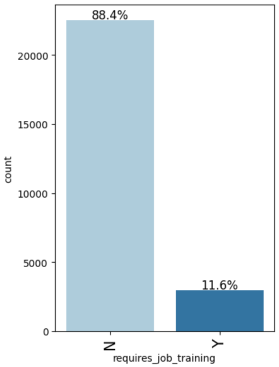


Link to Appendix slide to supporting EDA Univariant Analysis

EDA Results _ Univariate Analysis (Observations)

- Approximately 58% of applicants have job experience and approximately 88% of applicants do not require job training.



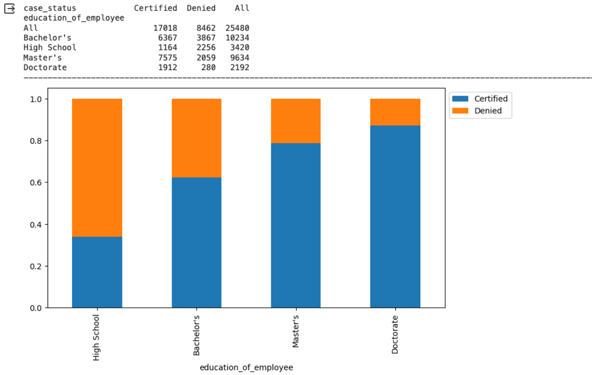


- The region of employment looks to be pretty evenly distributed across three regions (Northeast, South, and West). Approximately 82% of applications identify one of these three regions
- Approximately 90% of all applications have an hourly unit of wage
- Approximately 67% of of the Visas were certified

Link to Appendix slide to supporting EDA Univariant Analysis

EDA Results _ Bivariate Analysis (Observations)

- The level of education does appear to have an impact on visa certifications
  - Those with higher level of education appear to have a greater chance to be certified



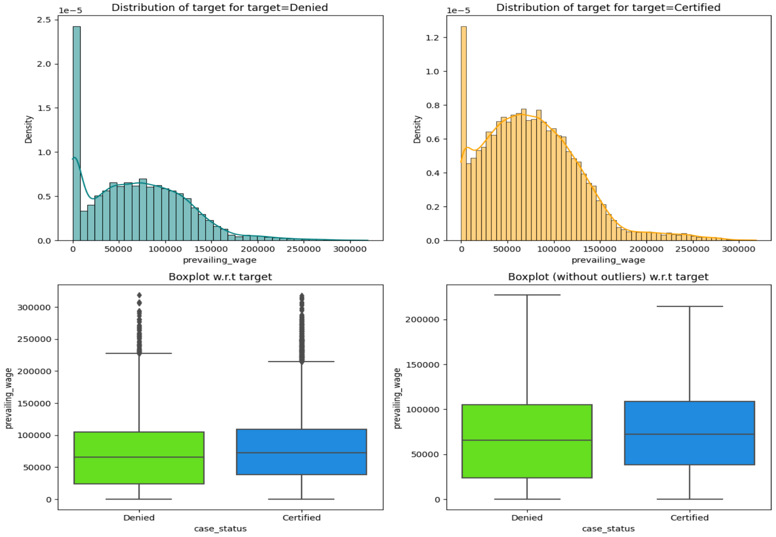
```
case_status            Certified  Denied    All
education_of_employee
All                        17018    8462  25480
Bachelor's                  6367    3867  10234
High School                 1164    2256   3420
Master's                    7575    2059   9634
Doctorate                   1912  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

- Educational Requirement and Regions
  - The west has the highest demand for applicants with a Doctoral education requirement
  - The northeast region has the highest demand for applicants with a Master's degree requirement
  - The south has the highest demand for applicants with a Bachelor's degree requirement
  - The south has the highest demand for applicants with a High School education requirement
- The Midwest Region currently has the highest certification rate – (~75%)
- Applicants from Europe appear to have the highest chance of certification while South America has the lowest
- Applicants with work experience are more likely get certified as they are less likely to need job training

Link to Appendix slide to supporting EDA Bivariate Analysis

EDA Results _ Bivariate Analysis (Wage Observations)

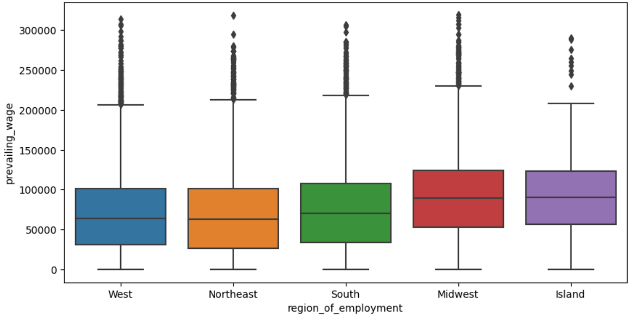
- It appears the median prevailing wage is slightly higher for applicants that obtain a visa certification



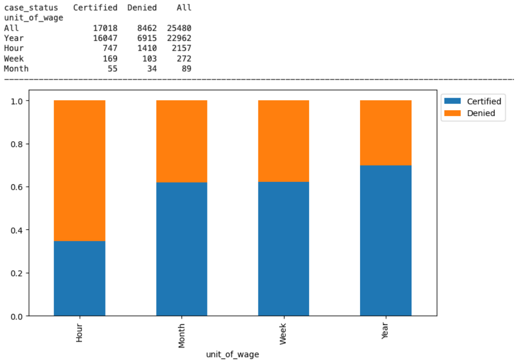
Link to Appendix slides to supporting EDA Bivariate Analysis

EDA Results _ Bivariate Analysis (Wage Observations)

- It appears the median prevailing wage is slightly higher in the Midwest and Island regions



- It appears applicants with “Year” unit of wage have a higher chance to obtain visa certification vs. the applicants that have “hour” 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
------------------------------------------------------------------------------------------------------------------------
```

Link to Appendix slide to supporting EDA Bivariate Analysis

Data Preprocessing

- There were no duplicates found. - no treatment necessary
- There were no missing values founds – no treatment necessary
- Outlier check completed – all data appears to be in the dataset and appears to be valuable information no treatment necessary
- Feature engineering
  - The “case_status column was encoded a 1 for certified and 0 for all other values
- Data preparation for modeling
  - The "case_status" column was dropped from the feature set X.
  - Dummy variables were created for categorical features in X using pd.get_dummies().
  - The data was split into training and testing sets using with a 70:30 ratio, and the stratify parameter was set to ensure that the class distribution is maintained in the splits.

Shape of Training and Testing Data Sets

Shape of Training Set	(17836, 21)
Shape of Testing Set	(7644, 21)

Percentage of Classes

Percentage of Classes	Training Set	Test Set
0	0.667919	0.667884
1	0.332081	0.332156

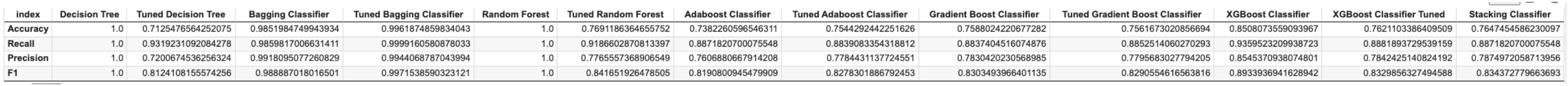
- The model_performance_classification_sklearn function was used to check the model performance of models
- The confusion_matrix_sklearn function was used to plot the confusion matrix
- The F1 Score was used as an evaluation metric because it considers both false positives and false negatives
- Balanced class weights were used so the model focuses equal on both classes

Link to Appendix slide to supporting Data Preprocessing

Model Performance Summary

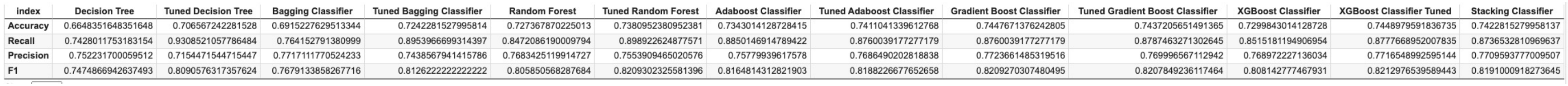
- It is recommended to use the XGBoost Classifier Tuned model as it appears to be performing the best reflecting the best F1 scores for the Training performance and the testing performance
  - XGBoost Classifier Tuned (Training) F1 Score: .832
  - XGBoost Classifier Tuned (Test) F1 Score: .821
- The top feature of importance for the model are:
  1) Education of Employee _ High School
  2) Has Job Experience _y
  3) Prevailing wage
  4) Education of Employee _ Masters
  5) Education of Employee _ Doctorate

Training Performance Comparison



index	Decision Tree	Tuned Decision Tree	Bagging Classifier	Tuned Bagging Classifier	Random Forest	Tuned Random Forest	Adaboost Classifier	Tuned Adaboost Classifier	Gradient Boost Classifier	Tuned Gradient Boost Classifier	XGBoost Classifier	XGBoost Classifier Tuned	Stacking Classifier
Accuracy	1.0	0.7125476564252075	0.9851984749943934	0.9961874859834043	1.0	0.7691186364655752	0.7382260596546311	0.7544292442251626	0.7588024220677282	0.7561673020856694	0.8508073559093967	0.7621103386409509	0.7647454586230097
Recall	1.0	0.9319231092084278	0.9859817006631411	0.9999160580878033	1.0	0.9186602870813397	0.8871820700075548	0.8839083354318812	0.8837404516074876	0.8852514060270293	0.9359523209938723	0.8881893729539159	0.8871820700075548
Precision	1.0	0.7200674536256324	0.9918095077260829	0.9944068787043994	1.0	0.7765557368906549	0.7606880667914208	0.7784431137724551	0.7830420230568985	0.7795683027794205	0.8545370938074801	0.7842425140824192	0.7874972058713956
F1	1.0	0.8124108155574256	0.988887018016501	0.9971538590323121	1.0	0.841651926478505	0.8190800945479909	0.8278301886792453	0.830349396640135	0.8290554616563816	0.8933936941628942	0.8329856327494588	0.834372779663693

Testing Performance Comparison



index	Decision Tree	Tuned Decision Tree	Bagging Classifier	Tuned Bagging Classifier	Random Forest	Tuned Random Forest	Adaboost Classifier	Tuned Adaboost Classifier	Gradient Boost Classifier	Tuned Gradient Boost Classifier	XGBoost Classifier	XGBoost Classifier Tuned	Stacking Classifier
Accuracy	0.6648351648351648	0.706567242281528	0.6915227629513344	0.7242281527995814	0.727367870225013	0.7380952380952381	0.7343014128728415	0.7411041339612768	0.7447671376242805	0.7437205651491365	0.7299843014128728	0.7448979591836735	0.7422815279958137
Recall	0.7428011753183154	0.9308521057786484	0.764152791380999	0.8953966699314397	0.8472086190009794	0.89892262487757 1	0.8850146914789422	0.8760039177277179	0.8760039177277179	0.8787463271302645	0.8515181194906954	0.8777668952007835	0.873653281096963 7
Precision	0.752231700059512	0.7154471544715447	0.7717111770524233	0.7438567941415786	0.7683425119914727	0.7553909465020576	0.75779939617578	0.7686490202818838	0.7723661485319516	0.769996567112942	0.768972227136034	0.7716548992595144	0.7709593777009507
F1	0.7474866942637493	0.8090576317357624	0.7679133858267716	0.8126222222222222	0.805850568287684	0.8209302325581396	0.8164814312821903	0.8188226677652658	0.8209270307480495	0.8207849236117464	0.808142777467931	0.8212976539589443	0.8191000918273645

Link to Appendix slide to supporting Model Building w/analysis and feature importance

Insights and Recommendations

Insights:

- Key Features for Model Importance:
  - The top factors influencing visa approval are the education level of the employee, job experience, and prevailing wage. These features significantly contribute to the model's decision-making process.

- Best profile for Visa Approval:
  - Education Level: Higher the education the better – most approvals require a Bachelors' degree or better – Doctorate and Master's degree are highly preferred
  - Job Experience: Job experience is essential
  - Prevailing Wage: Average prevailing wage is around $70k.
  - Additional Factors: Applicants from Europe, Africa, and Asia, with yearly unit of wage, and applying to the Mid-West region have higher chances of approval.

- Best profile for Visa Denied:
  - Education Level: High School Education and/or no degree
  - Job Experience: Lack of job experience
  - Prevailing Wage: Average prevailing wage is around $65k.
  - Additional Factors: Applicants with hourly unit of wage, from Oceania, North America, and South America are more likely to face denial.

Insights and Recommendations

Recommendations:

- Utilize the XGBoostClassifier Tuned model for its outstanding performance, achieving an F1 Score of 83% for the training set and 82% for the testing set. This model is recommended for accurate predictions in the visa approval process.

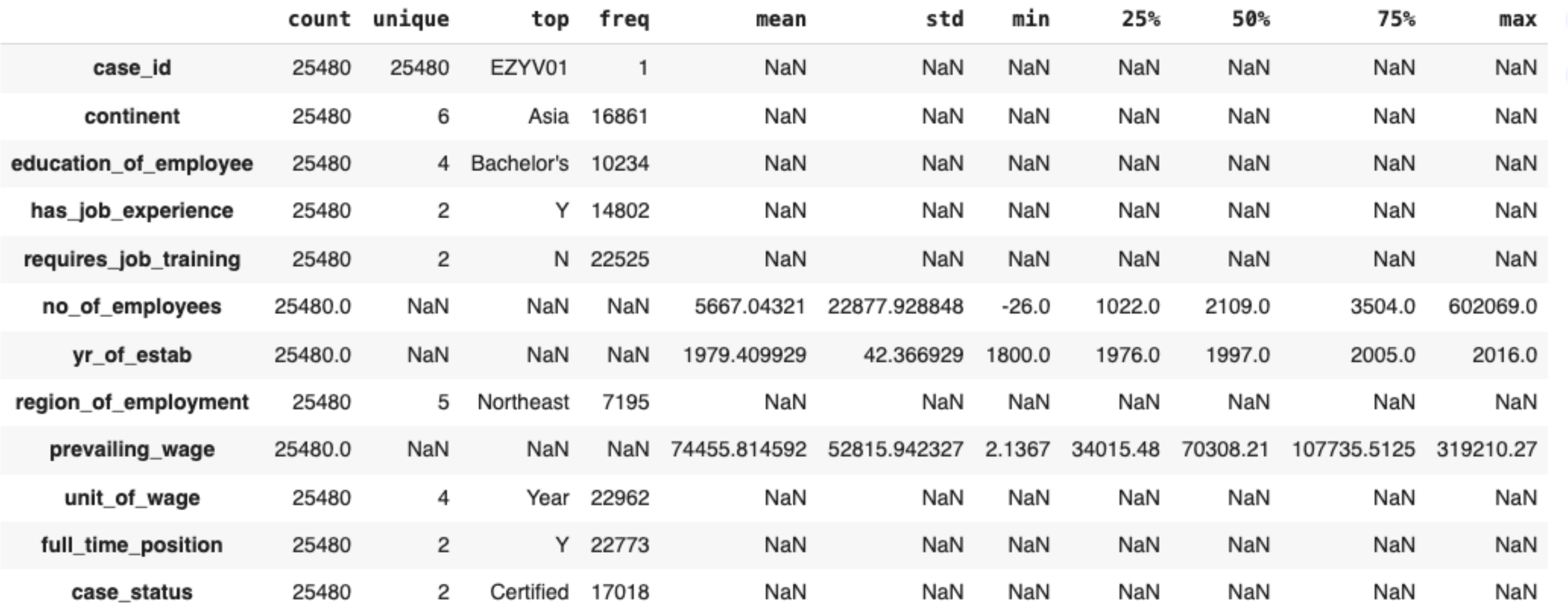
Recommended Further Analysis:

- Additional Data Collection: Gather more information from both employers and employees to extract deeper insights
  - **Job Type and Prevailing Wage Data**: Analyze prevailing wage data based on job types such as IT, service, administration, etc.
  - **Required Education and Years of Experience**: Understand the correlation between visa approval and specific education levels and work experience.
  - **Industry Sector Analysis**: Explore visa approval trends within different industry sectors.
  - **Regional Analysis**: Segment data based on the type of company (industry sector) and size in different regions.
  - **Visa Length Analysis**: Examine the impact of visa length on approval rates.
  - **Applicant Segmentation by Company Size**: Classify applicants based on company size (small, median, large) to discern patterns in approved/denied applications.

APPENDIX

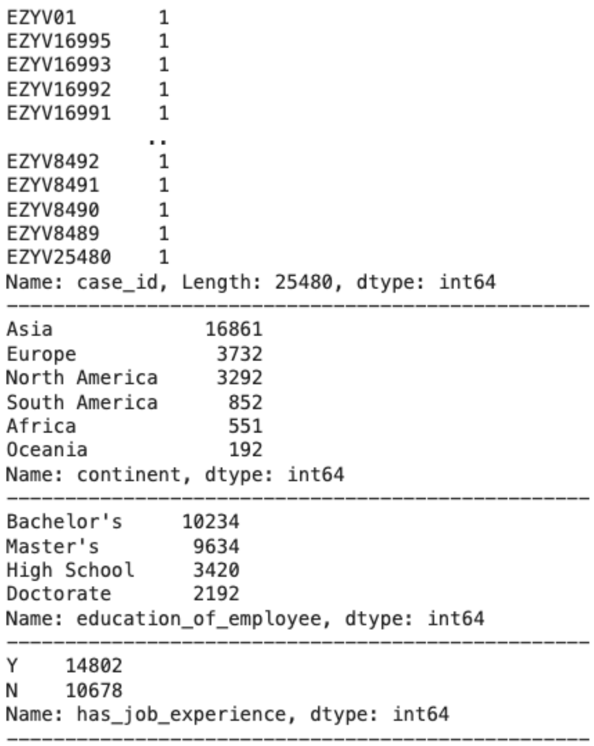
EDA

Statistical Summary of the Data

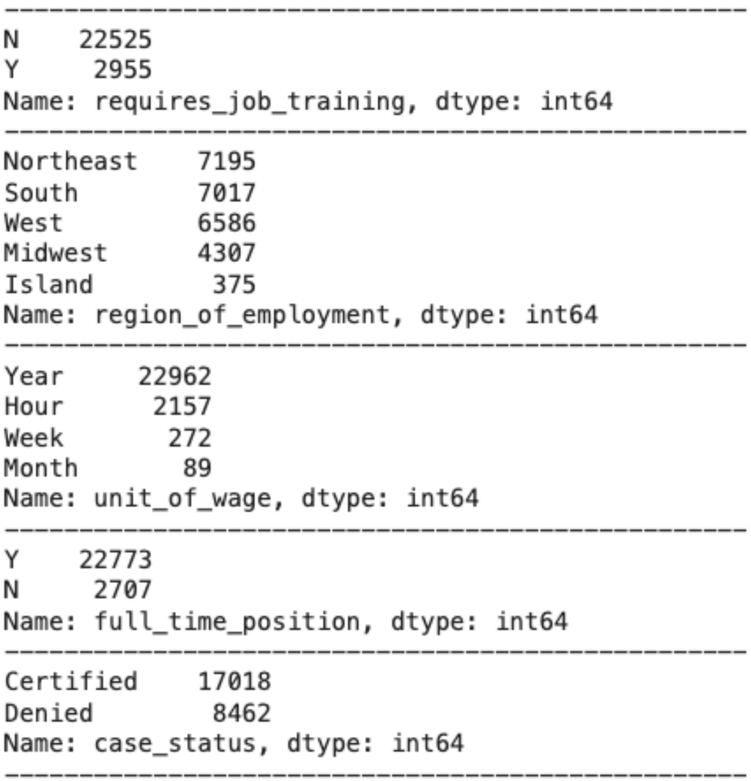


	count	unique	top	freq	mean	std	min	25%	50%	75%	max
case_id	25480	25480	EZYV01	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
continent	25480	6	Asia	16861	NaN	NaN	NaN	NaN	NaN	NaN	NaN
education_of_employee	25480	4	Bachelor's	10234	NaN	NaN	NaN	NaN	NaN	NaN	NaN
has_job_experience	25480	2	Y	14802	NaN	NaN	NaN	NaN	NaN	NaN	NaN
requires_job_training	25480	2	N	22525	NaN	NaN	NaN	NaN	NaN	NaN	NaN
no_of_employees	25480.0	NaN	NaN	NaN	5667.04321	22877.928848	-26.0	1022.0	2109.0	3504.0	602069.0
yr_of_estab	25480.0	NaN	NaN	NaN	1979.409929	42.366929	1800.0	1976.0	1997.0	2005.0	2016.0
region_of_employment	25480	5	Northeast	7195	NaN	NaN	NaN	NaN	NaN	NaN	NaN
prevailing_wage	25480.0	NaN	NaN	NaN	74455.814592	52815.942327	2.1367	34015.48	70308.21	107735.5125	319210.27
unit_of_wage	25480	4	Year	22962	NaN	NaN	NaN	NaN	NaN	NaN	NaN
full_time_position	25480	2	Y	22773	NaN	NaN	NaN	NaN	NaN	NaN	NaN
case_status	25480	2	Certified	17018	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Counts for each unique category (Categorical Variables)



```
EZYV01       1
EZYV16995    1
EZYV16993    1
EZYV16992    1
EZYV16991    1
            ..
EZYV8492     1
EZYV8491     1
EZYV8490     1
EZYV8489     1
EZYV25480    1
Name: case_id, Length: 25480, dtype: int64
--------------------------------------------------
Asia             16861
Europe            3732
North America     3292
South America      852
Africa             551
Oceania            192
Name: continent, dtype: int64
--------------------------------------------------
Bachelor's     10234
Master's        9634
High School     3420
Doctorate       2192
Name: education_of_employee, dtype: int64
--------------------------------------------------
Y    14802
N    10678
Name: has_job_experience, dtype: int64
--------------------------------------------------
```

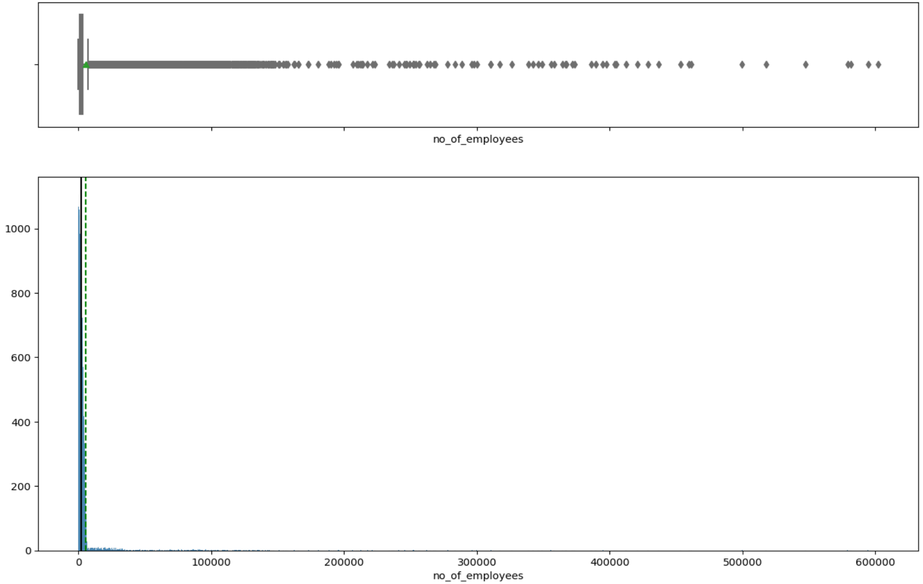


```
--------------------------------------------------
N    22525
Y     2955
Name: requires_job_training, dtype: int64
--------------------------------------------------
Northeast    7195
South        7017
West         6586
Midwest      4307
Island        375
Name: region_of_employment, dtype: int64
--------------------------------------------------
Year     22962
Hour      2157
Week       272
Month       89
Name: unit_of_wage, dtype: int64
--------------------------------------------------
Y    22773
N     2707
Name: full_time_position, dtype: int64
--------------------------------------------------
Certified    17018
Denied        8462
Name: case_status, dtype: int64
--------------------------------------------------
```

Univariate Analysis

EDA _Univariate Analysis

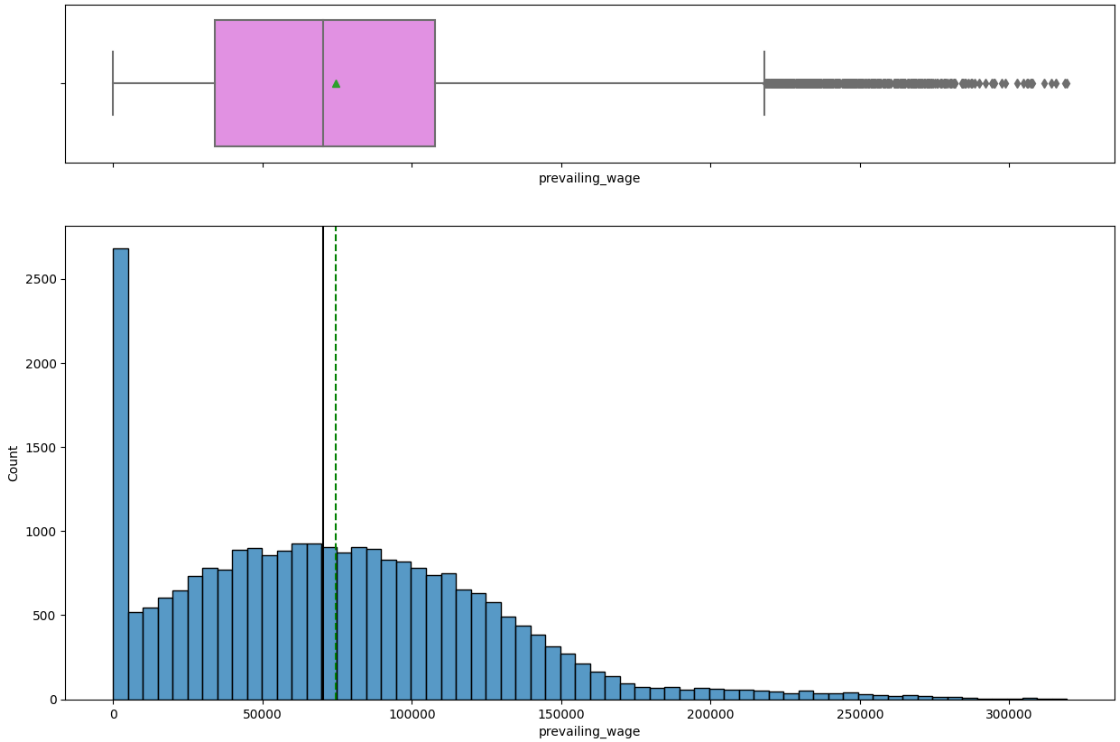
Number of Employees



- Data appears heavily skewed right with lots of outliers
- Data indicates a large variety of company size

EDA _Univariate Analysis cont…

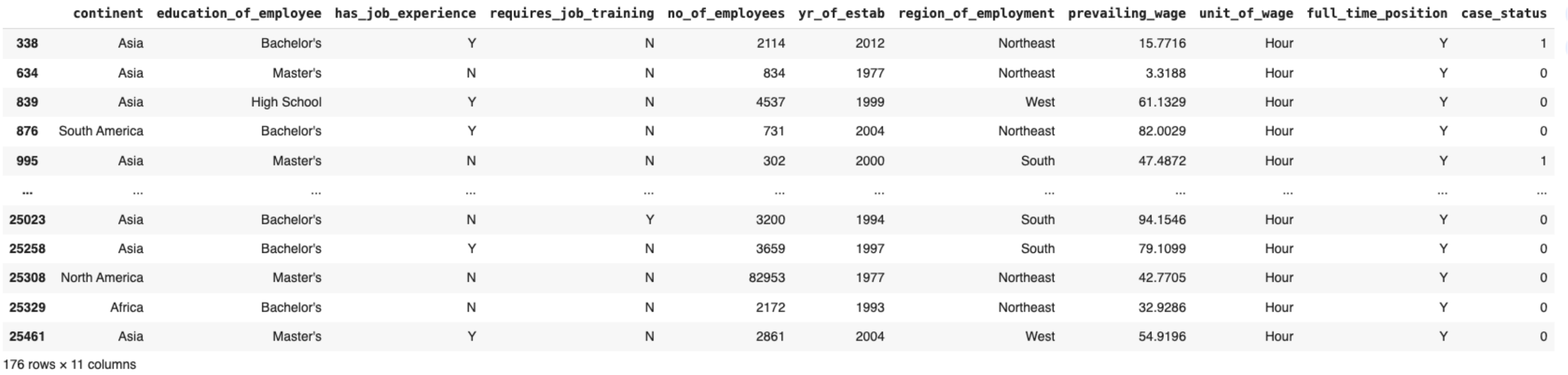
Prevailing Wage



- Data appears skewed right with lots of outliers
- There is a large disparity within the prevailing wage data
- There is a large amount of wages around the 0 mark – this may require further analysis
- There may be a data entry error in the data using hourly data
- There are wages above the 200k mark

EDA _Univariate Analysis cont…

Prevailing Wage _ Analysis less than 100



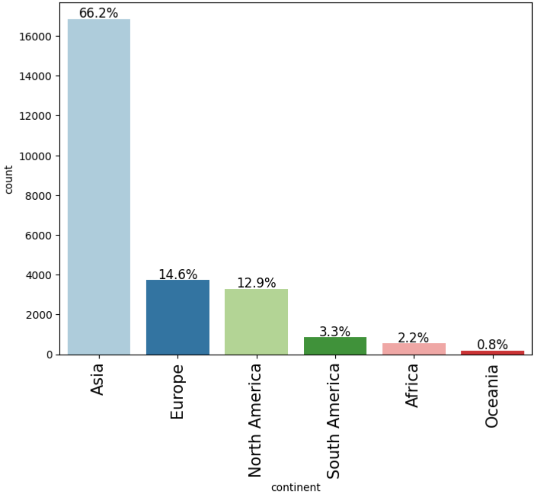
	continent	education_of_employee	has_job_experience	requires_job_training	no_of_employees	yr_of_estab	region_of_employment	prevailing_wage	unit_of_wage	full_time_position	case_status
338	Asia	Bachelor's	Y	N	2114	2012	Northeast	15.7716	Hour	Y	1
634	Asia	Master's	N	N	834	1977	Northeast	3.3188	Hour	Y	0
839	Asia	High School	Y	N	4537	1999	West	61.1329	Hour	Y	0
876	South America	Bachelor's	Y	N	731	2004	Northeast	82.0029	Hour	Y	0
995	Asia	Master's	N	N	302	2000	South	47.4872	Hour	Y	1
...	...	...	...	...	...	...	...	...	...	...	...
25023	Asia	Bachelor's	N	Y	3200	1994	South	94.1546	Hour	Y	0
25258	Asia	Bachelor's	Y	N	3659	1997	South	79.1099	Hour	Y	0
25308	North America	Master's	N	N	82953	1977	Northeast	42.7705	Hour	Y	0
25329	Africa	Bachelor's	N	N	2172	1993	Northeast	32.9286	Hour	Y	0
25461	Asia	Master's	Y	N	2861	2004	West	54.9196	Hour	Y	0

176 rows × 11 columns

- 176 rows had less than 100 in hourly prevailing wage

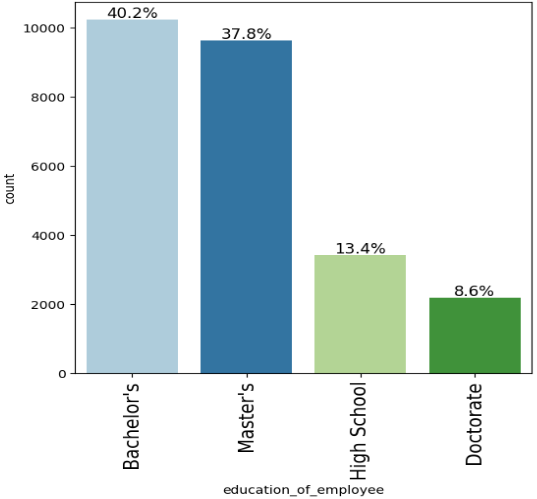
EDA _Univariate Analysis cont…

Continent Observations



- ~66% Asia
- ~15% Europe
- ~13% North America
- ~3% South America
- ~2% from Africa
- ~1% from Oceania

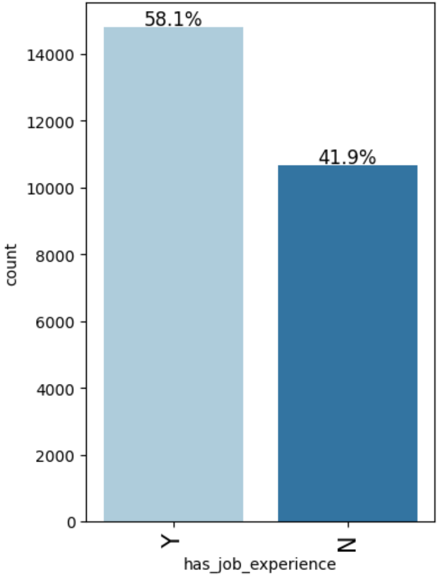
Education of Employee Observations



- Bachelor’s degree is the top education (~40%)
- Master’s degree is second highest (~38%)
- Approximately 87% have a degree in higher education

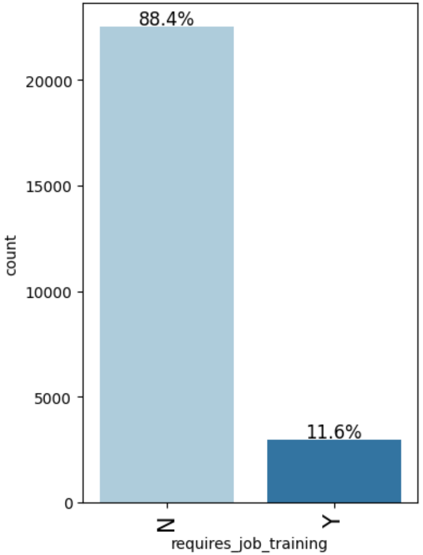
EDA _Univariate Analysis cont…

Job Experience Observations



- ~58% of applicants have job experience
- ~42% do not have job experience

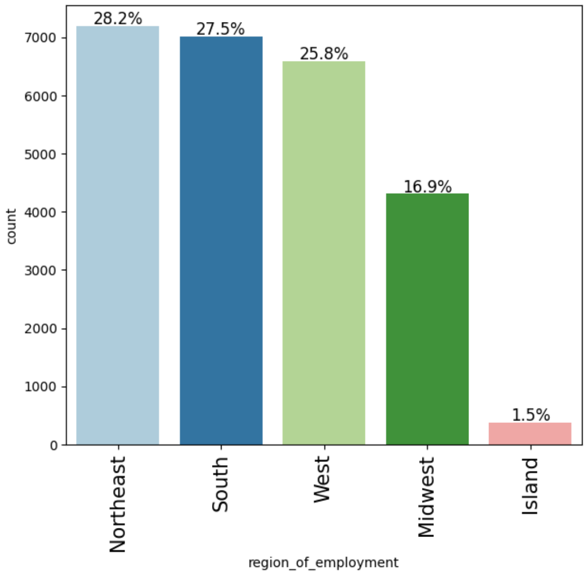
Requires Job Training Observations



- ~88% of applicants do not need job training
- ~12% of applicants will need job training

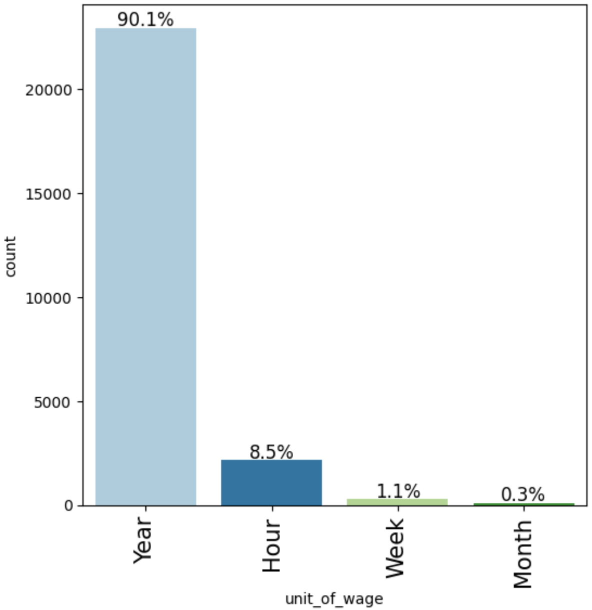
EDA _Univariate Analysis cont...

Region of Employment Observations



- Data looks pretty evenly distributed with the top three regions
- ~82% of applications designate the Northeast, South and West regions.
- ~17% of applicants designate the Midwest
- The island region is the lest designated at 1.5%

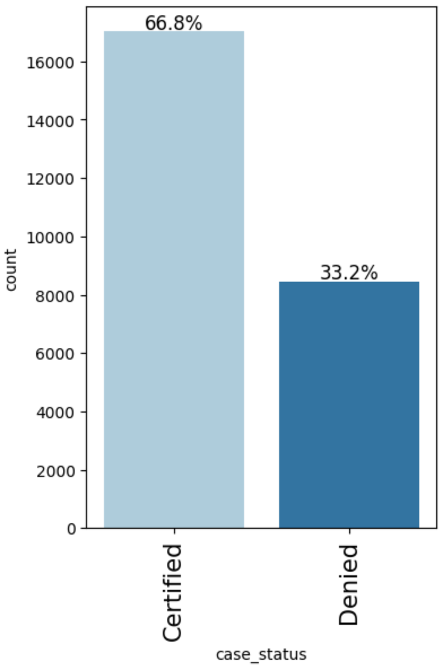
Unit of Wage Observations



- ~90% of applicants have a yearly unit of wage

EDA _Univariate Analysis cont…

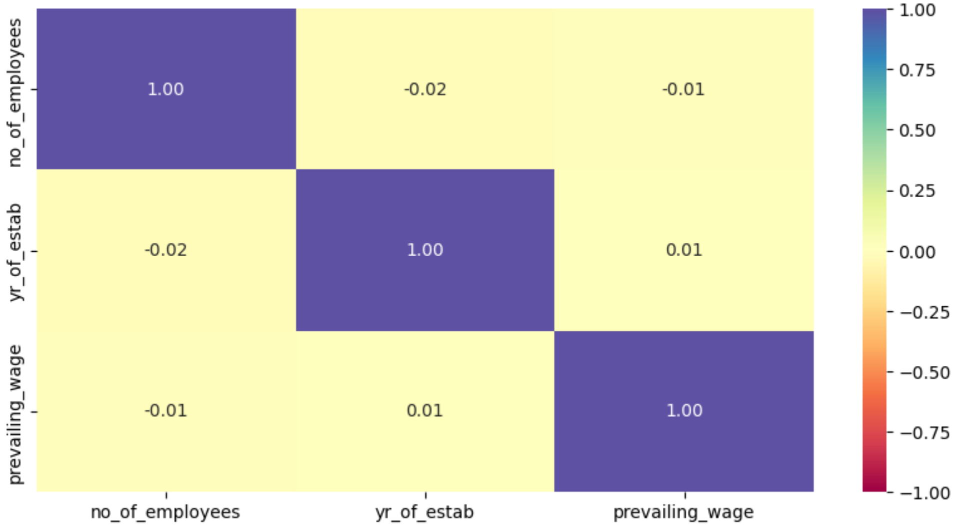
Case Status Observations



- ~67% of all visas are certified
- ~33% of all visas are denied

Bivariate Analysis

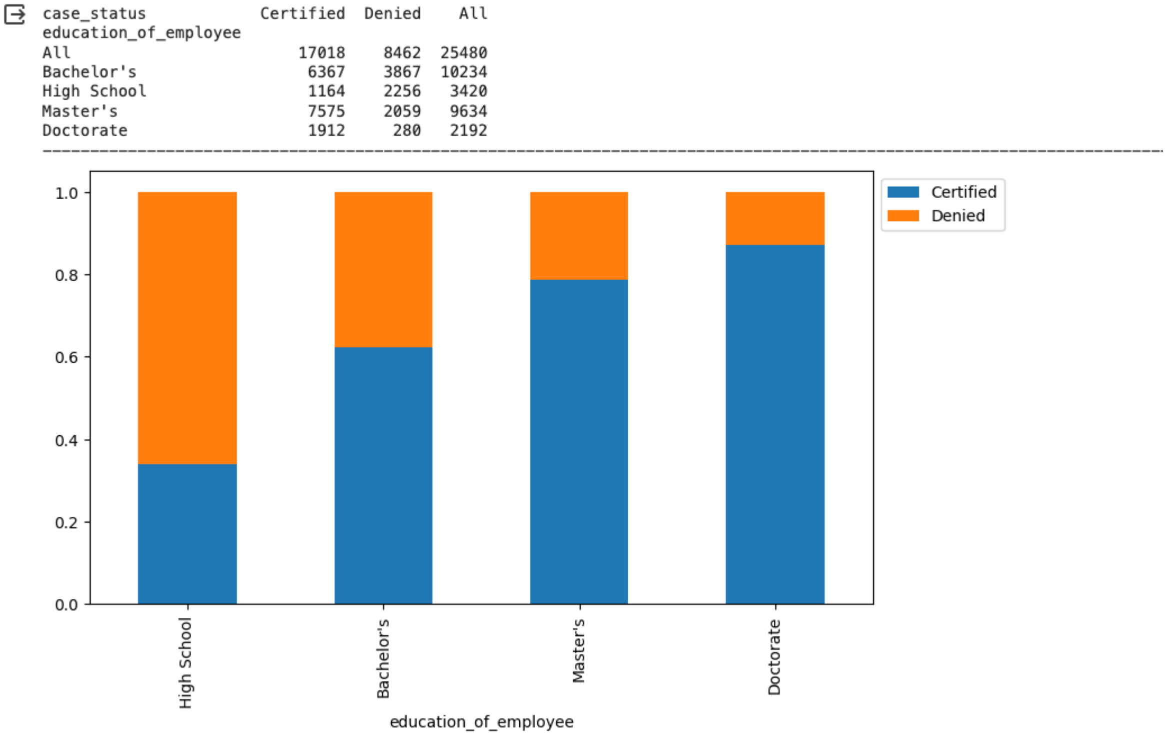
EDA _Bivariate Analysis



- There does not appear to be any correlation among the numerical variables

EDA _Bivariate Analysis, cont...

Does education level impact visa certification?

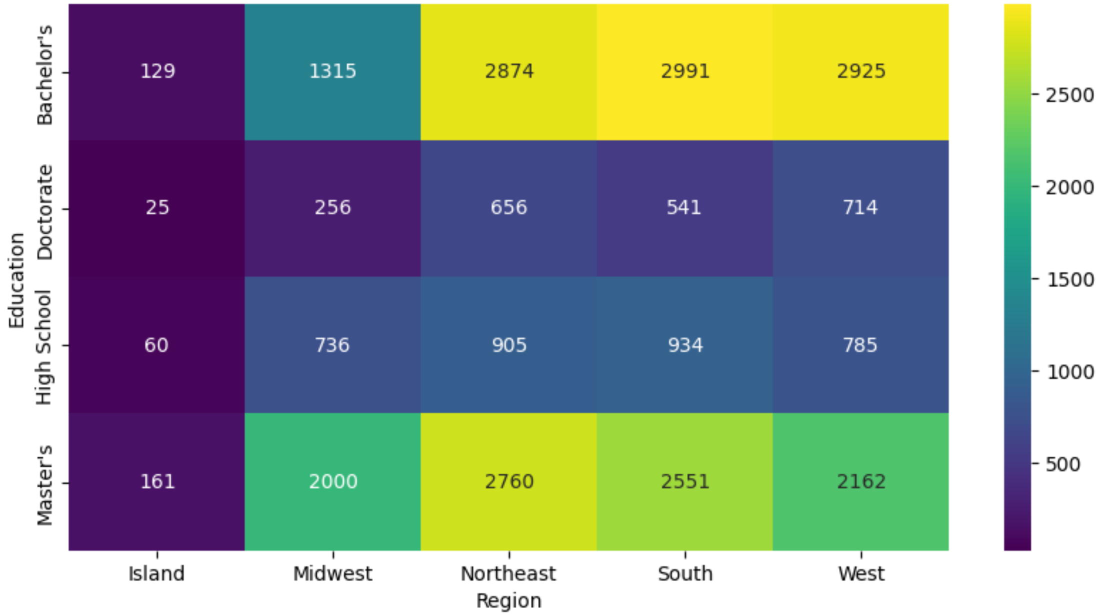


```
case_status          Certified  Denied    All
education_of_employee
All                      17018    8462  25480
Bachelor's                6367    3867  10234
High School               1164    2256   3420
Master's                  7575    2059   9634
Doctorate                 1912     280   2192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

- Yes, it appears the higher the level of education to better chance to have the visa certified

EDA _Bivariate Analysis, cont…

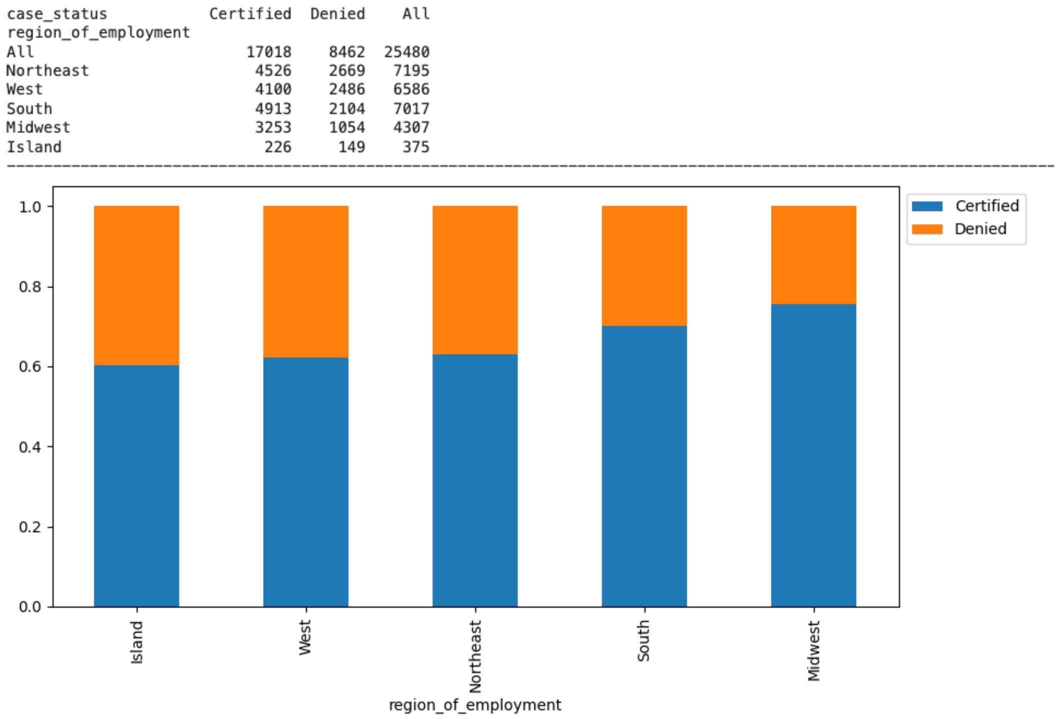
Regions vs. diverse talent and educational backgrounds



- High school education requirement if highest in the South Region followed by the Northeast Region
- Bachelor’s Degree education requirement if highest in South Region followed by the West Region
- Master’s Degree education requirement if highest in the Northeast Region followed by the South Region
- Doctorate Degree education requirement is highest in the West Region followed by the Northeast Region

EDA _Bivariate Analysis, cont...

Number of visa certification across each region

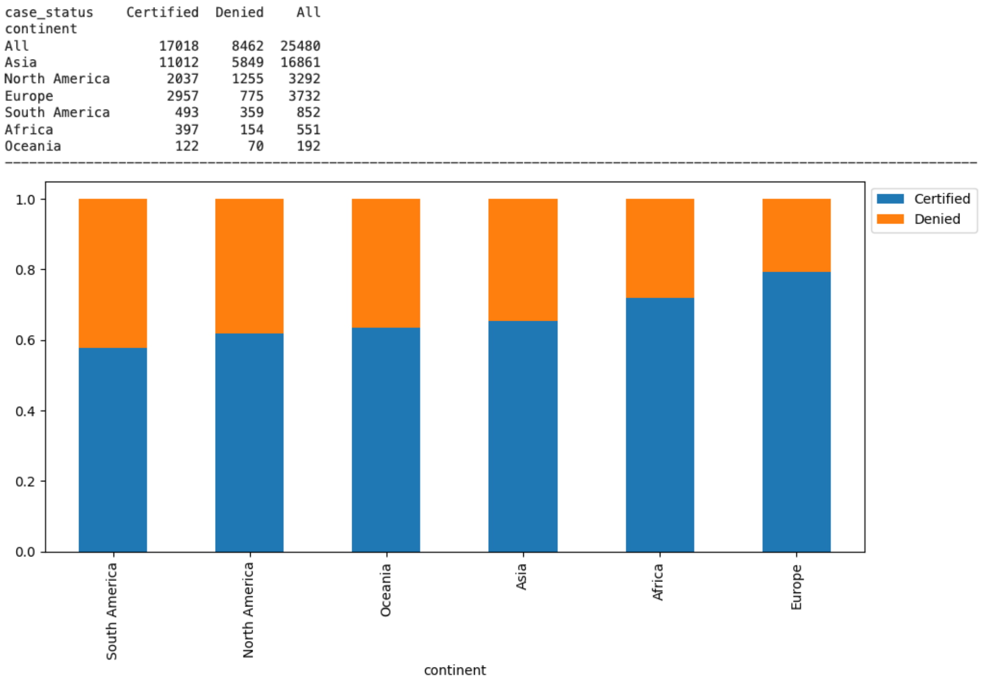


```
case_status          Certified  Denied    All
region_of_employment
All                      17018    8462  25480
Northeast                 4526    2669   7195
West                      4100    2486   6586
South                     4913    2104   7017
Midwest                   3253    1054   4307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

- The Midwest region has the highest certification rate
- The Island region has the lowest certification rate

EDA _Bivariate Analysis, cont…

Number of visa certification across continents

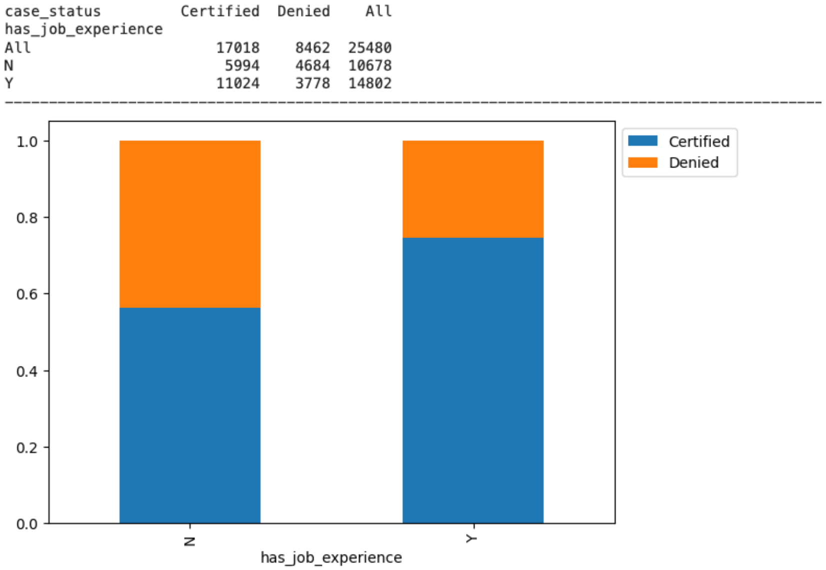


```
case_status    Certified  Denied    All
continent
All                17018    8462  25480
Asia               11012    5849  16861
North America       2037    1255   3292
Europe              2957     775   3732
South America        493     359    852
Africa               397     154    551
Oceania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

- Europe has the highest certification rate
- South America has the lowest certification rate

EDA _Bivariate Analysis, cont...

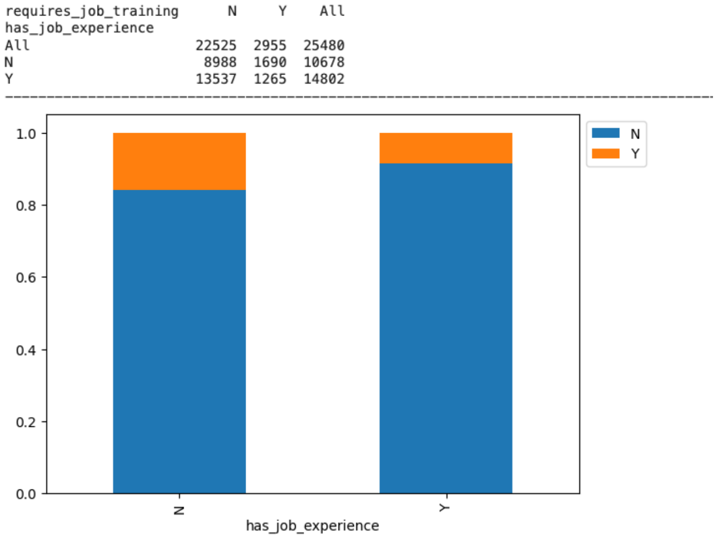
Does work experience have an influence on getting certified?



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

- Having work experience increases the applicants chances to get certified

Do the employees who have prior work experience require any job training?

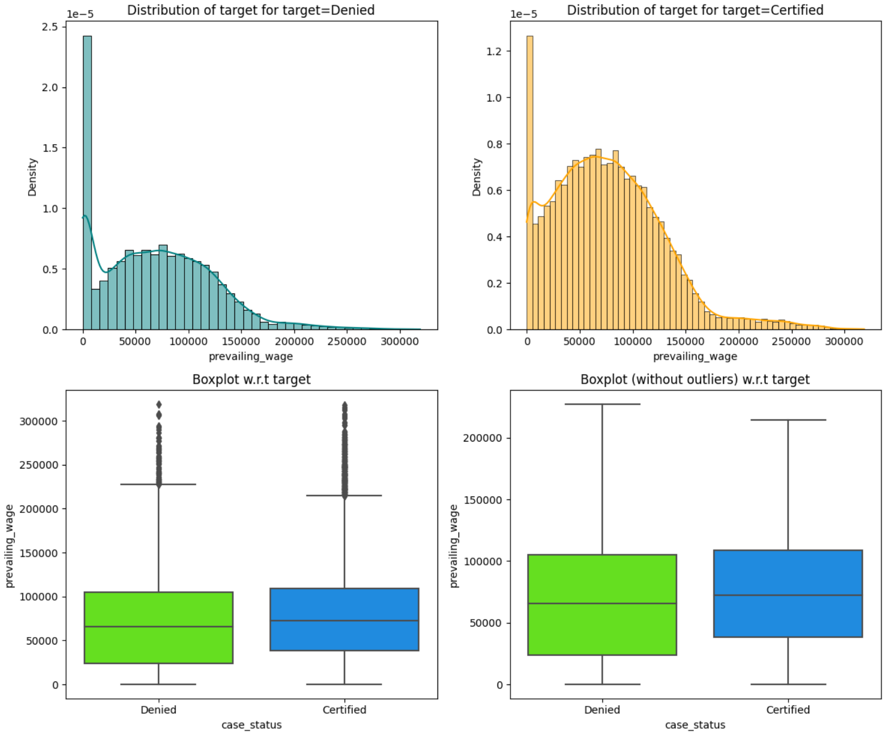


```
requires_job_training      N     Y    All
has_job_experience
All                    22525  2955  25480
N                       8988  1690  10678
Y                      13537  1265  14802
------------------------------------------------------------------------------------------------------------------------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

- Applicants that do have job experience and less likely going to need job training

EDA _Bivariate Analysis, cont...

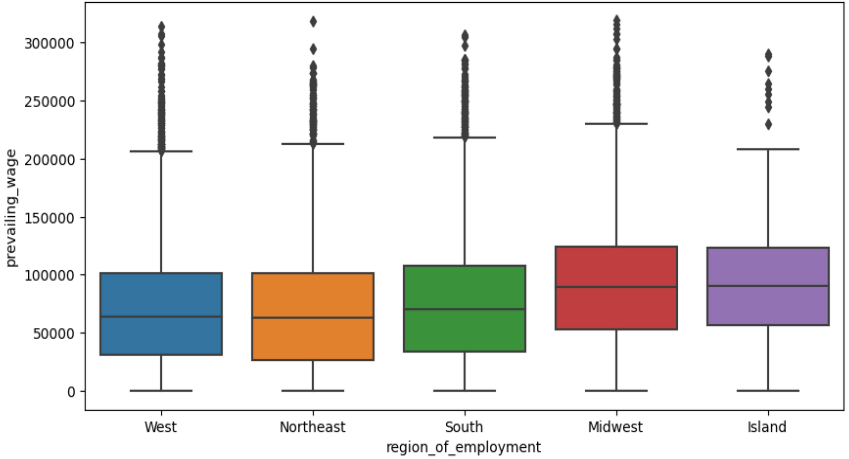
- Does Visa status change with prevailing wage?



- It appears the median prevailing wage of applicants that have obtained a visa certification is slightly higher than those applicates that were denied.

EDA _Bivariate Analysis, cont…

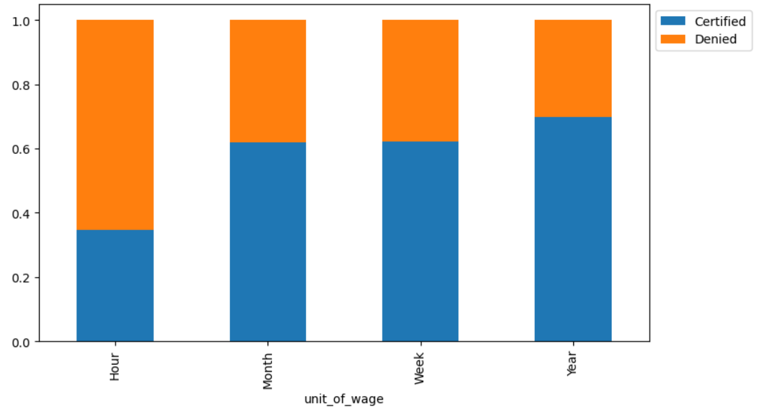
- Is the prevailing wage is similar across all the regions of the US?



- The prevailing wages appear to higher in the Midwest and Island regions

- Does the type of prevailing wage unit type have any impact on certification?

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

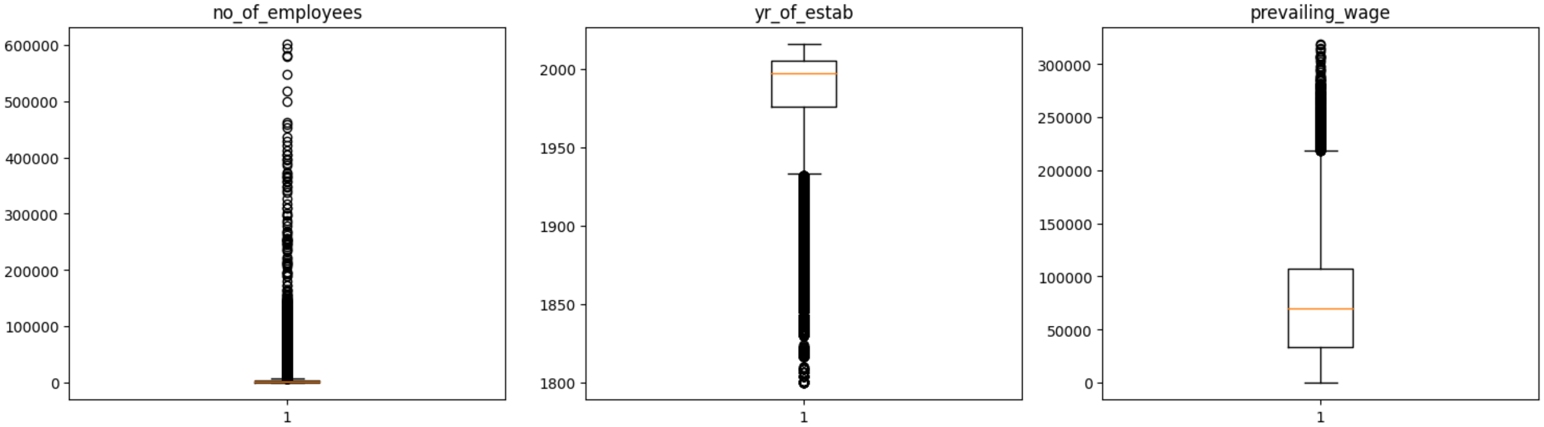


- It appears applicants with “year” units of wage have a higher can for certification where applicants with a “hour” unit of wage have the lowest chance for certification

Data Preprocessing

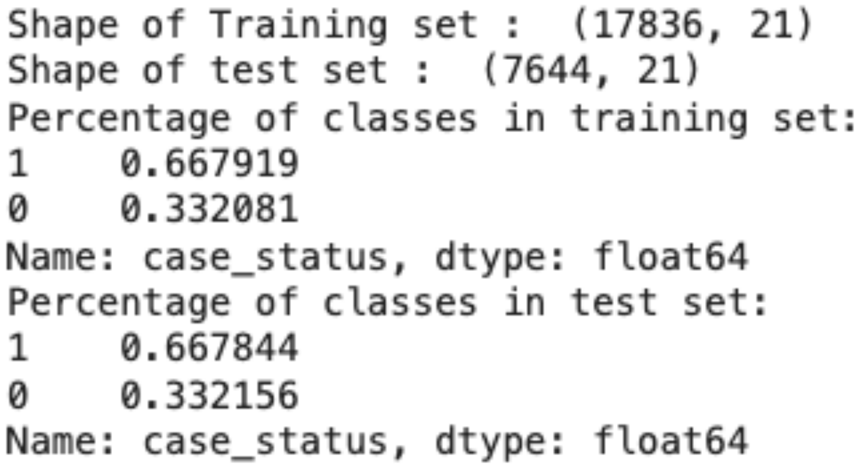
Data Preprocessing _ Outlier Check

- There are quite a few outliers in the data- these all seem to appear in the dataset
- However, we will not treat them as they provide valuable information



Data Preprocessing _ Train and Test Set

- The data was split into training and testing sets using with a 70:30 ratio
- The stratify parameter was set to ensure that the class distribution is maintained in the splits



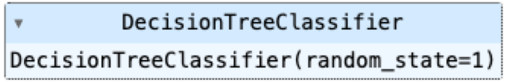
```
Shape of Training set :  (17836, 21)
Shape of test set :  (7644, 21)
Percentage of classes in training set:
1    0.667919
0    0.332081
Name: case_status, dtype: float64
Percentage of classes in test set:
1    0.667844
0    0.332156
Name: case_status, dtype: float64
```

Model Building

Model Building – Bagging (Decision Tree)

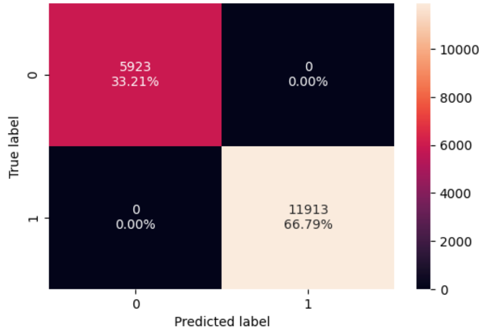
Decision Tree Model building steps

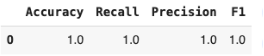
- Used the DecisionTreeClassifier function with random state = 1



```
▾ DecisionTreeClassifier
DecisionTreeClassifier(random_state=1)
```

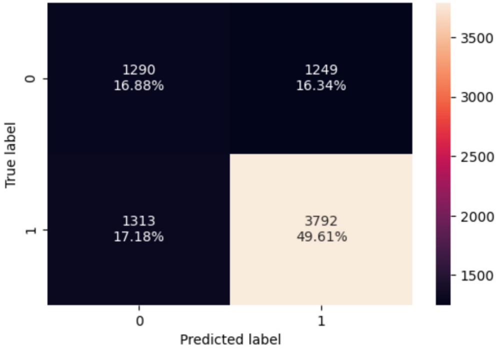
Train Performance

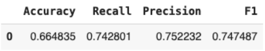




	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Test Performance





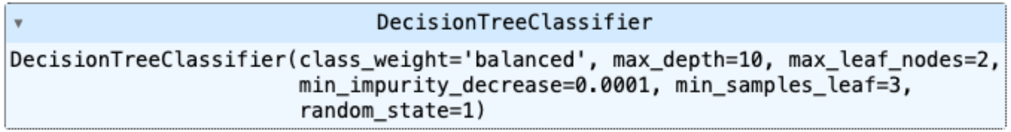
	Accuracy	Recall	Precision	F1
0	0.664835	0.742801	0.752232	0.747487

- The model is overfitting in the train data (F1 is 1.0)
- The test data is not performing as well (F1 is .747)

Model Building – Bagging (Decision Tree) cont...

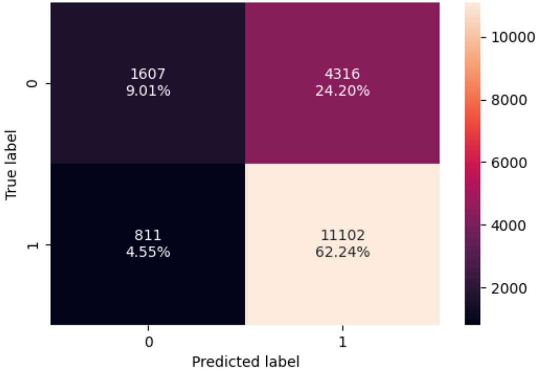
Hyperparameter Tuning - Decision Tree

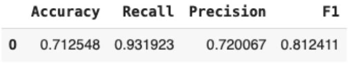
- Used the DecisionTreeClassifier function (class_weight=‘balanced’, random state = 1)



```
▾                          DecisionTreeClassifier
DecisionTreeClassifier(class_weight='balanced', max_depth=10, max_leaf_nodes=2,
                       min_impurity_decrease=0.0001, min_samples_leaf=3,
                       random_state=1)
```

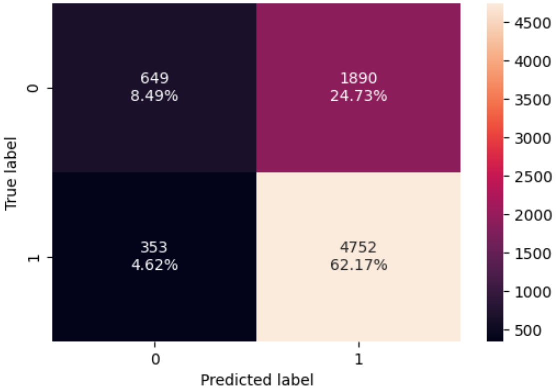
Train Performance

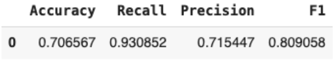




	Accuracy	Recall	Precision	F1
0	0.712548	0.931923	0.720067	0.812411

Test Performance





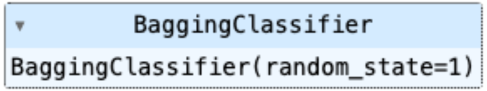
	Accuracy	Recall	Precision	F1
0	0.706567	0.930852	0.715447	0.809058

The model is demonstrating a good fit
and is not overfitting

Model Building – Bagging (Bagging Classifier)

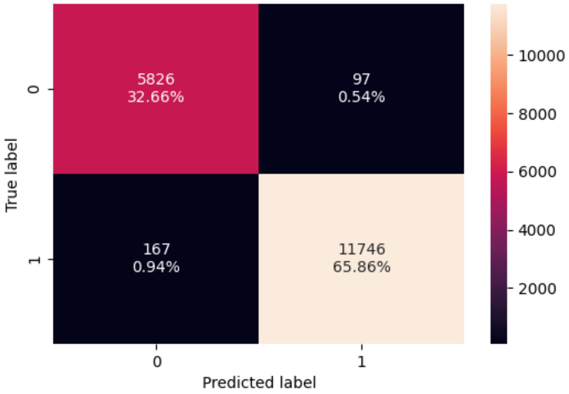
Bagging Classifier Model building steps

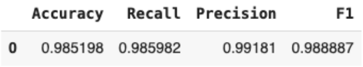
- Used the BaggingClassifier function with random state = 1



```
▾ BaggingClassifier
BaggingClassifier(random_state=1)
```

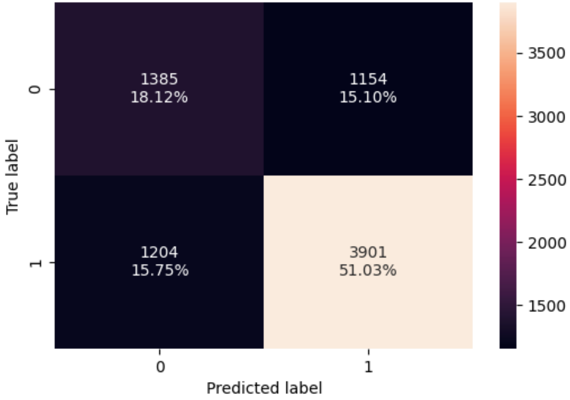
Train Performance

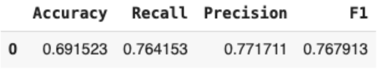




	Accuracy	Recall	Precision	F1
0	0.985198	0.985982	0.99181	0.988887

Test Performance





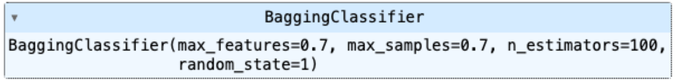
	Accuracy	Recall	Precision	F1
0	0.691523	0.764153	0.771711	0.767913

The model appears to be overfitting

Model Building – Bagging (Bagging Classifier) cont...

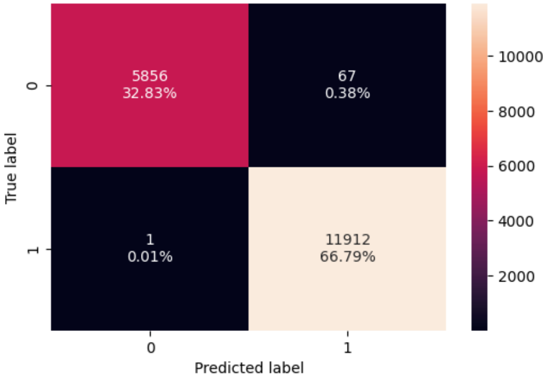
Hyperparameter Tuning – Bagging Classifier

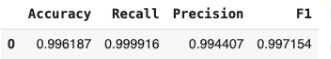
- Used the BaggingClassifier function with random state = 1



```
▾                          BaggingClassifier
BaggingClassifier(max_features=0.7, max_samples=0.7, n_estimators=100,
                  random_state=1)
```

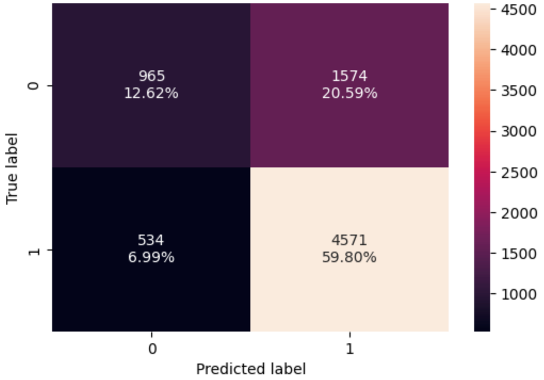
Train Performance

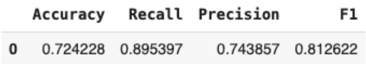




	Accuracy	Recall	Precision	F1
0	0.996187	0.999916	0.994407	0.997154

Test Performance





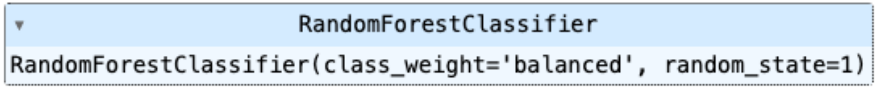
	Accuracy	Recall	Precision	F1
0	0.724228	0.895397	0.743857	0.812622

The model appears to be overfitting even after tuning

Model Building – Bagging (Random Forest)

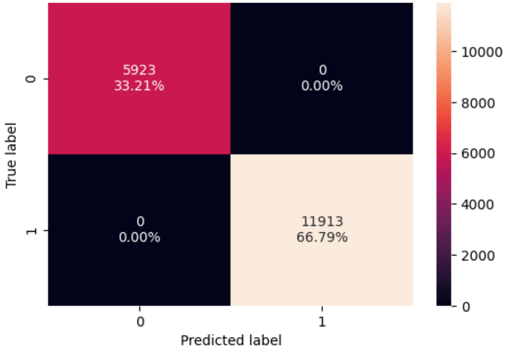
Random Forest Model building steps

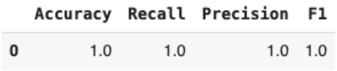
- Used the RandomForestClassifier (random state = 1, class_weight=‘balanced’,)



```
▾                     RandomForestClassifier
RandomForestClassifier(class_weight='balanced', random_state=1)
```

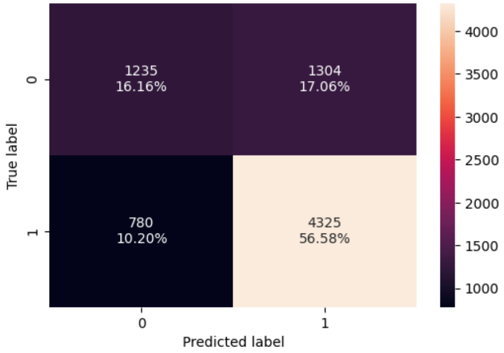
Train Performance

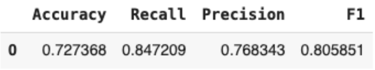




	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Test Performance





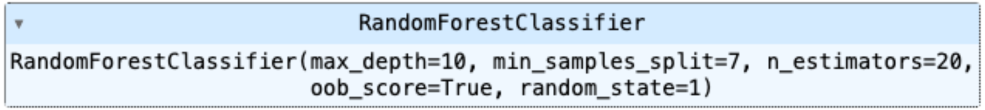
	Accuracy	Recall	Precision	F1
0	0.727368	0.847209	0.768343	0.805851

- The random forest classifier appears to be overfitting in the train data (F1 is 1.0)
- Test data is not performing as well (F1 is .805)

Model Building – Bagging (Random Forest), cont...

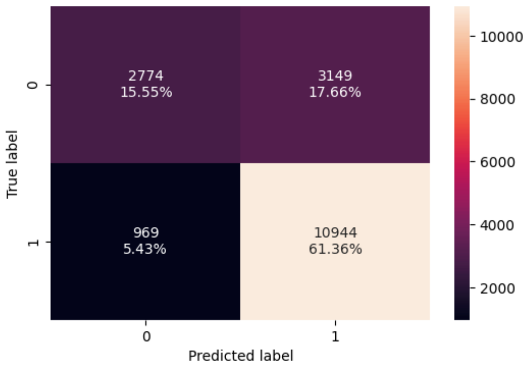
Hyperparameter Tuning - Random Forest

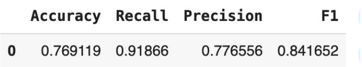
- Used the RandomForestClassifier (random state = 1, obb_score=True, bootstrap=True)



```
▾                        RandomForestClassifier
RandomForestClassifier(max_depth=10, min_samples_split=7, n_estimators=20,
                       oob_score=True, random_state=1)
```

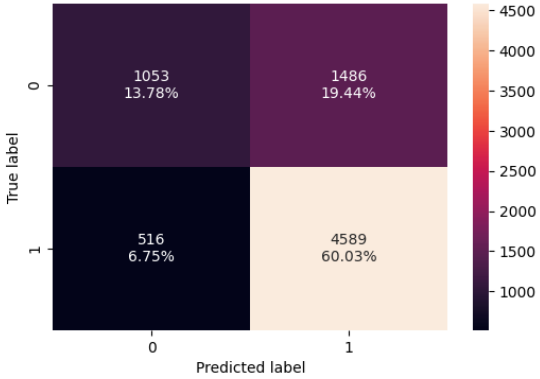
Train Performance

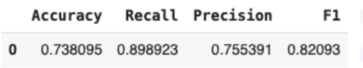




	Accuracy	Recall	Precision	F1
0	0.769119	0.91866	0.776556	0.841652

Test Performance





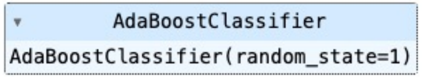
	Accuracy	Recall	Precision	F1
0	0.738095	0.898923	0.755391	0.82093

The model is demonstrating a good fit
and is not overfitting

Model Building – Boosting (AdaBoost)

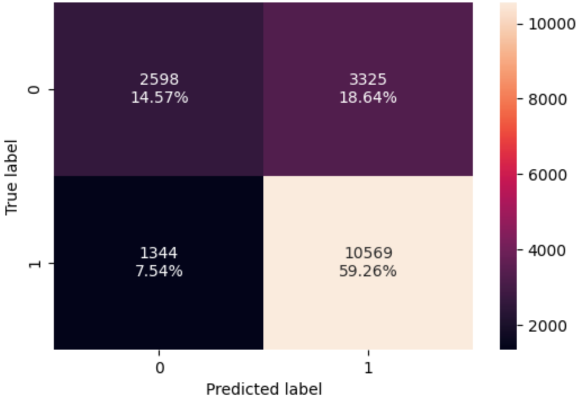
Boosting Classifier Model building steps

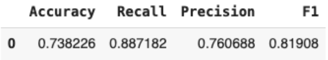
- Used the AdaBoostClassifier (random state = 1)



```
▾ AdaBoostClassifier
AdaBoostClassifier(random_state=1)
```

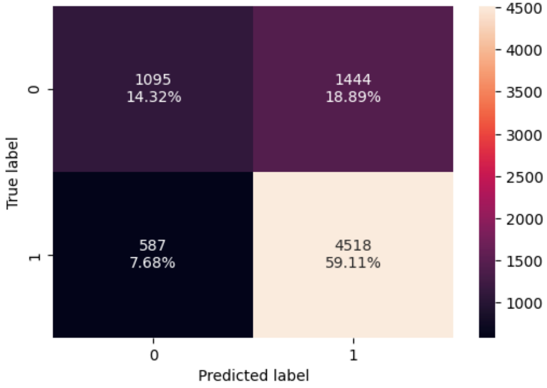
Train Performance

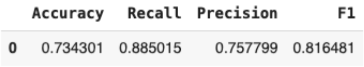




	Accuracy	Recall	Precision	F1
0	0.738226	0.887182	0.760688	0.81908

Test Performance





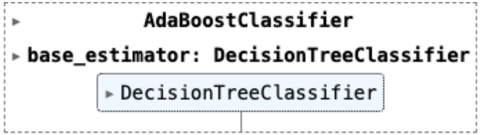
	Accuracy	Recall	Precision	F1
0	0.734301	0.885015	0.757799	0.816481

The model is demonstrating a good fit
and is not overfitting

Model Building – Boosting (AdaBoost)

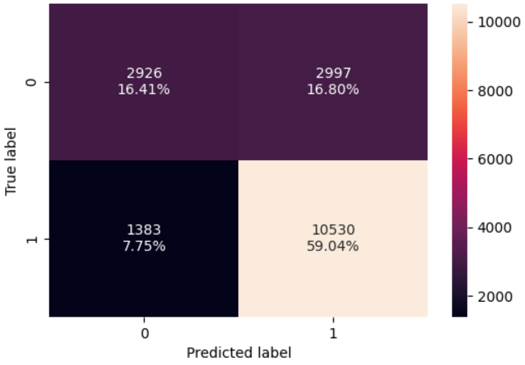
Hyperparameter Tuning – AdaBoost Classifier

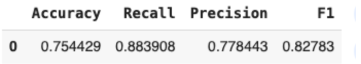
- Used the AdaBoostClassifier (random state = 1)



```
▸ AdaBoostClassifier
▸ base_estimator: DecisionTreeClassifier
  ▸ DecisionTreeClassifier
```

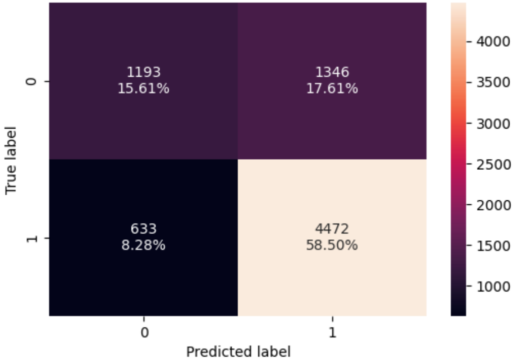
Train Performance

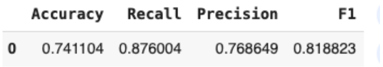




	Accuracy	Recall	Precision	F1
0	0.754429	0.883908	0.778443	0.82783

Test Performance





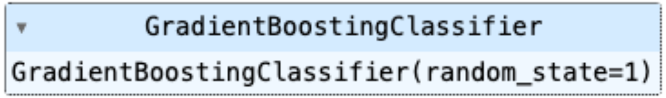
	Accuracy	Recall	Precision	F1
0	0.741104	0.876004	0.768649	0.818823

The model is demonstrating a good fit and is not overfitting – The F1 score has increased in both the train and test data

Model Building – Boosting (GradientBoosting)

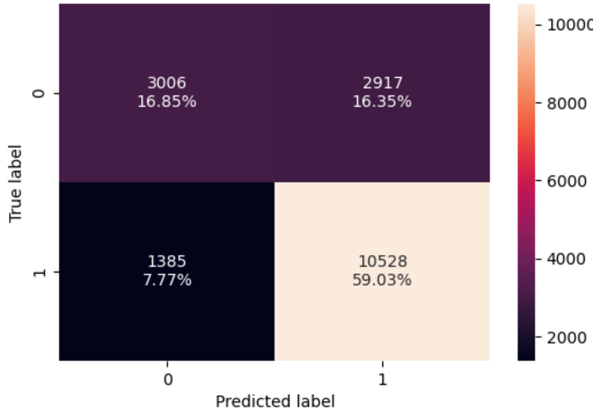
Boosting Classifier Model building steps

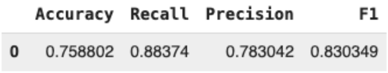
- Used the GradientBoostingClassifier (random state = 1)



```
▾ GradientBoostingClassifier
GradientBoostingClassifier(random_state=1)
```

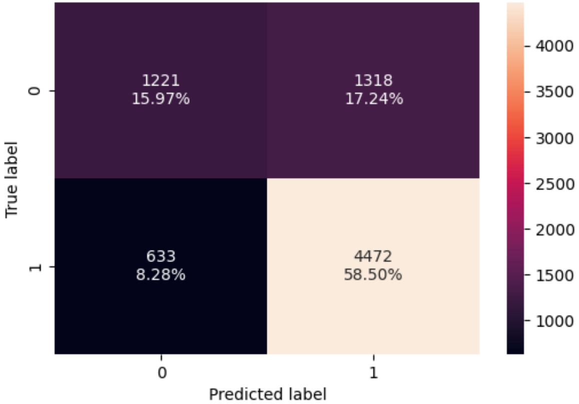
Train Performance

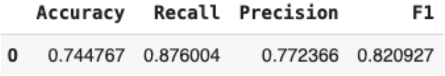




	Accuracy	Recall	Precision	F1
0	0.758802	0.88374	0.783042	0.830349

Test Performance





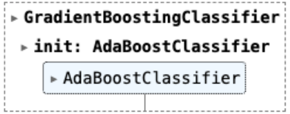
	Accuracy	Recall	Precision	F1
0	0.744767	0.876004	0.772366	0.820927

The model is demonstrating a
good fit and is not overfitting

Model Building – Boosting (GradientBoosting)

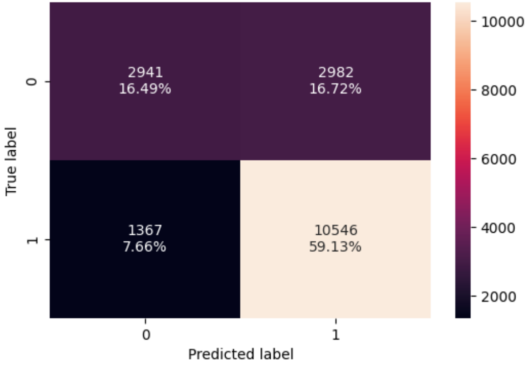
Hyperparameter Tuning – Gradient Boosting Classifier

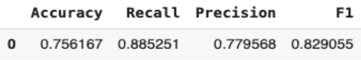
- Used the GradientBoostingClassifier(init=AdaBoostClassifier(Random_State=1), Random_State=1



```
▸ GradientBoostingClassifier
 ▸ init: AdaBoostClassifier
   ▸ AdaBoostClassifier
```

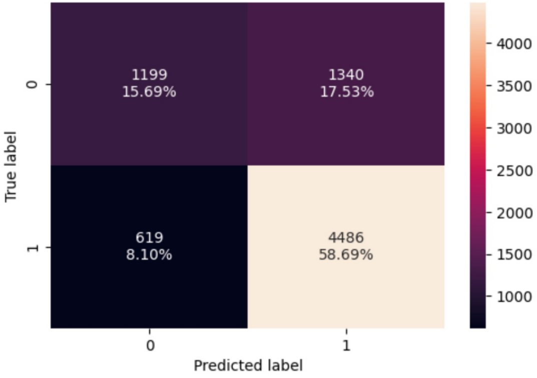
Train Performance

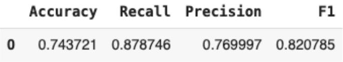




	Accuracy	Recall	Precision	F1
0	0.756167	0.885251	0.779568	0.829055

Test Performance





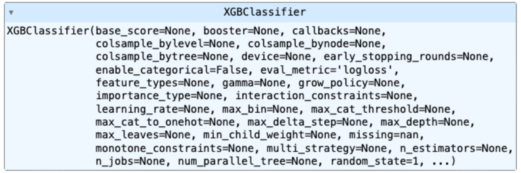
	Accuracy	Recall	Precision	F1
0	0.743721	0.878746	0.769997	0.820785

The model is demonstrating a good fit and is not overfitting – The F1 score has decreased in both the train and test data

Model Building – Boosting (XGBoost)

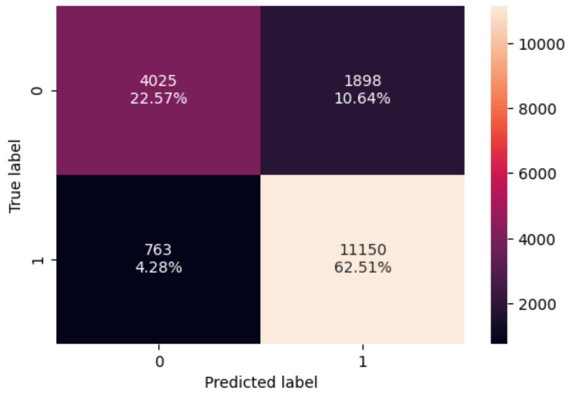
XGBoost Classifier Model building steps

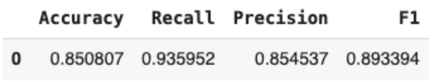
- Used the XGBClassifier(random_state=1, eval_metrics=‘logloss’)



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

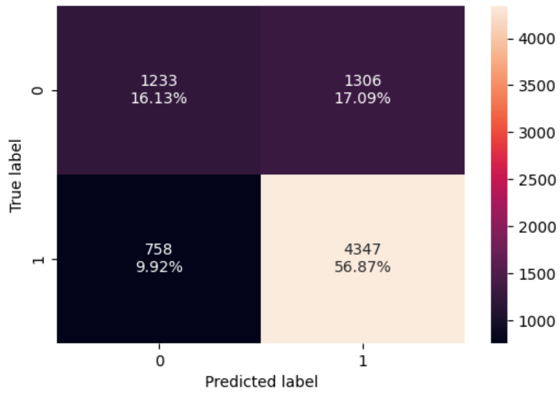
Train Performance

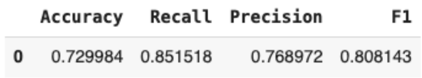




	Accuracy	Recall	Precision	F1
0	0.850807	0.935952	0.854537	0.893394

Test Performance





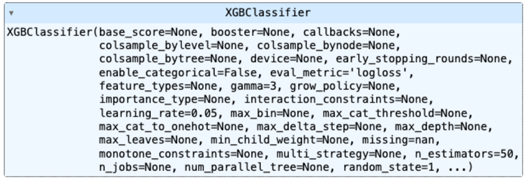
	Accuracy	Recall	Precision	F1
0	0.729984	0.851518	0.768972	0.808143

It appears the XGBoost may be overfitting

Model Building – Boosting (XGBoost)

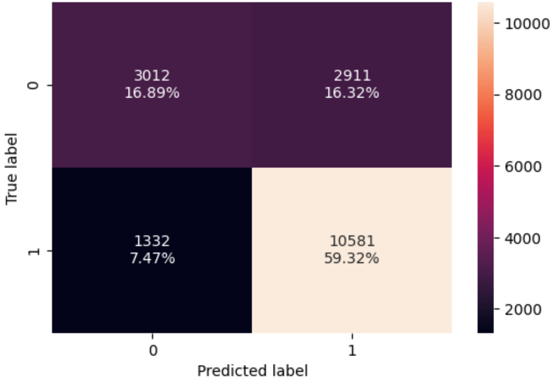
Hyperparameter Tuning – XGBoost Classifier

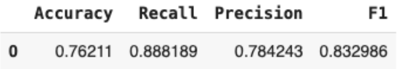
- Used the XGBClassifier(random_state=1, eval_metric='logloss'



```
▾                                   XGBClassifier
XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=None, device=None, early_stopping_rounds=None,
              enable_categorical=False, eval_metric='logloss',
              feature_types=None, gamma=3, grow_policy=None,
              importance_type=None, interaction_constraints=None,
              learning_rate=0.05, max_bin=None, max_cat_threshold=None,
              max_cat_to_onehot=None, max_delta_step=None, max_depth=None,
              max_leaves=None, min_child_weight=None, missing=nan,
              monotone_constraints=None, multi_strategy=None, n_estimators=50,
              n_jobs=None, num_parallel_tree=None, random_state=1, ...)
```

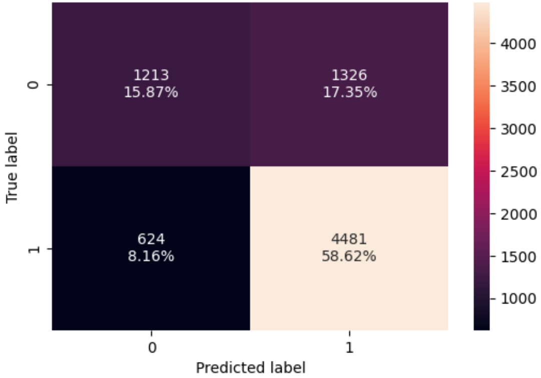
Train Performance

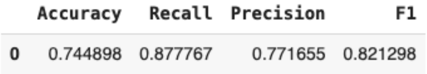




	Accuracy	Recall	Precision	F1
0	0.76211	0.888189	0.784243	0.832986

Test Performance

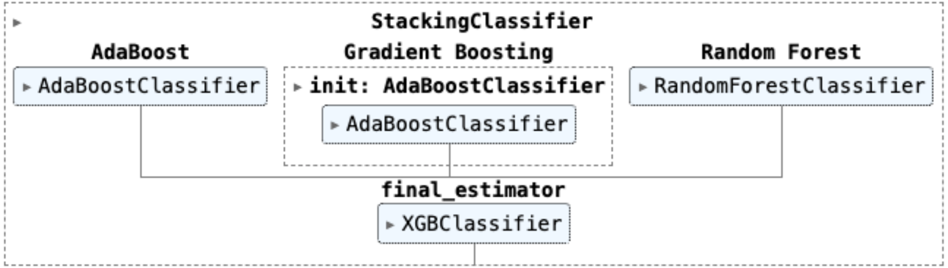




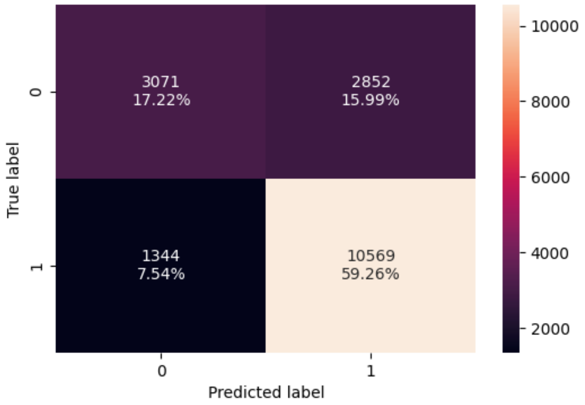
	Accuracy	Recall	Precision	F1
0	0.744898	0.877767	0.771655	0.821298

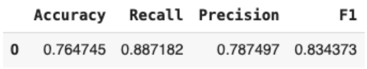
The model is demonstrating a
good fit and is not overfitting

Stacking Classifier



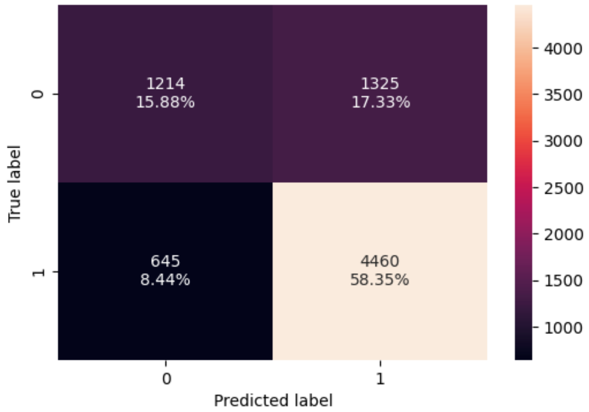
Train Performance

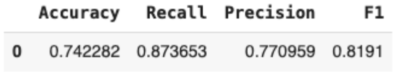




	Accuracy	Recall	Precision	F1
0	0.764745	0.887182	0.787497	0.834373

Test Performance

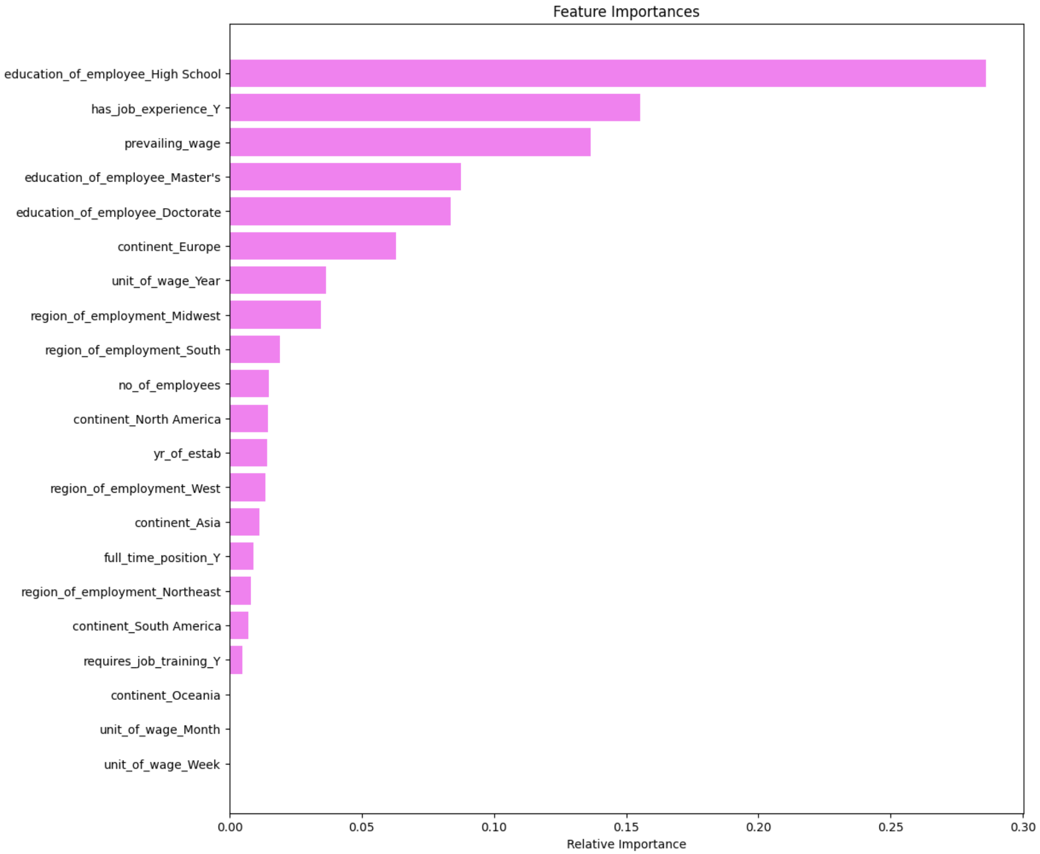




	Accuracy	Recall	Precision	F1
0	0.742282	0.873653	0.770959	0.8191

The model is demonstrating a good fit and is not overfitting. Comparable to XGBClassifier results. F1 Scores are Train (.83) and Test (.81)

Feature Importance


Feature Importances
education_of_employee_High School
has_job_experience_Y
prevailing_wage
education_of_employee_Master's
education_of_employee_Doctorate
continent_Europe
unit_of_wage_Year
region_of_employment_Midwest
region_of_employment_South
no_of_employees
continent_North America
yr_of_estab
region_of_employment_West
continent_Asia
full_time_position_Y
region_of_employment_Northeast
continent_South America
requires_job_training_Y
continent_Oceania
unit_of_wage_Month
unit_of_wage_Week
Relative Importance