Easy Visa Project

Ensemble Technique

25th July, 2023

Contents / Agenda

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

Executive Summary

- Summary of observations and conclusions
    - Decision tree, Bagging Classifier, Tuned Bagging Classifier and Random Forest are overfitting the training data set before tuning.
    - Most training and test performance has become comparable after hyperparameter tuning.
    - XG Boost can be seen no improvement after hyperparameter tuning.
    - All F1 score is improved after hyperparameter tuning.
    - If the model performance is selected based its simplicity and ease of interpretation, tuned Decision tree is the best compared with other model.
    - If the model performance is selected based on to improve/reduce the bias, tuned gradient boost is the best compared with other model.

Business Problem Overview and Solution Approach

- Business problem overview:
  - It is found the higher the education and job experience, the OFLC will approve for visa certification.
  - Majority requestor is from Asia, but the highest visa being certified is from Europe.
  - Less requestor being certified from Midwest and Island, may cause the region to less develop or take a long time for city development compared with other regions.

Business Problem Overview and Solution Approach

- Solution approach/business improvement/recommendation
  - Further investigations are needed on the root cause of why visa is being denied.
  - The hiring company should recommend the OFLC to certify Doctorate employee more as they are usually bring new expertise to the country.
  - OFLC should recommend the hiring company to increase the job training for the visa requestor to attract more talent.
  - More investigation should be perform, such as in critical industry needed in the regions vs qualification and prevailing wage; such as medical, finance, applied science, IT etc.

EDA Results – Univariate Analysis

- Checking the first 5 data
  - Data.shape – contains 25,480 rows and 12 columns

	case_id	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
0	EZYV01	Asia	High School	N	N	14513	2007	West	592.2029	Hour	Y	Denied
1	EZYV02	Asia	Master's	Y	N	2412	2002	Northeast	83425.6500	Year	Y	Certified
2	EZYV03	Asia	Bachelor's	N	Y	44444	2008	West	122996.8600	Year	Y	Denied
3	EZYV04	Asia	Bachelor's	N	N	98	1897	West	83434.0300	Year	Y	Denied
4	EZYV05	Africa	Master's	Y	N	1082	2005	1082	149907.3900	Year	Y	Certified

EDA Results – Univariate Analysis

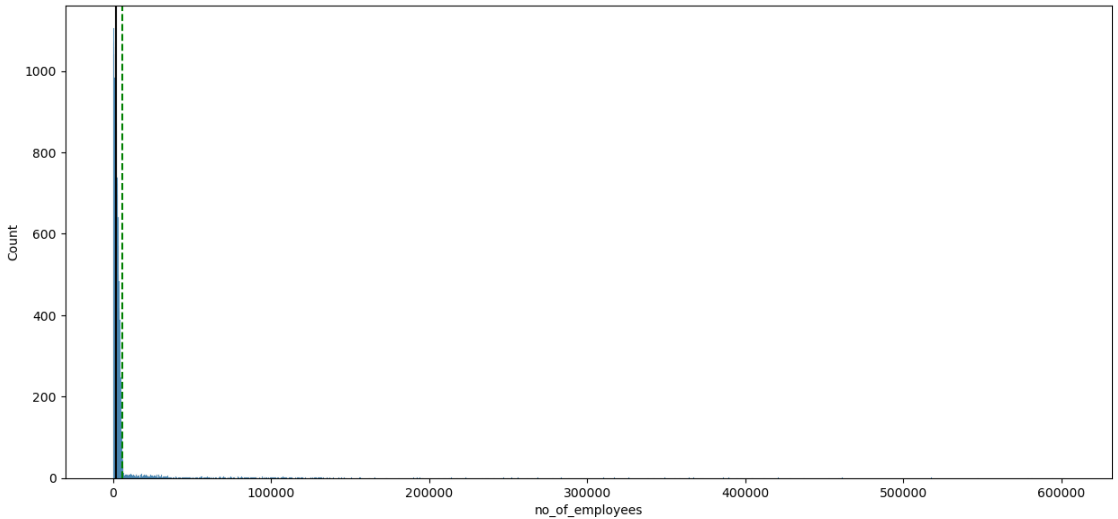
- Data.info as below, with no data duplication found.
- It is found the highest continent with education is Asia, with Bachelor's degree.
- The highest region of employment is at northeast.
- The average number of employees involved is about 5667, with average of prevailing wage is 74,455.81

```
<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
```

	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

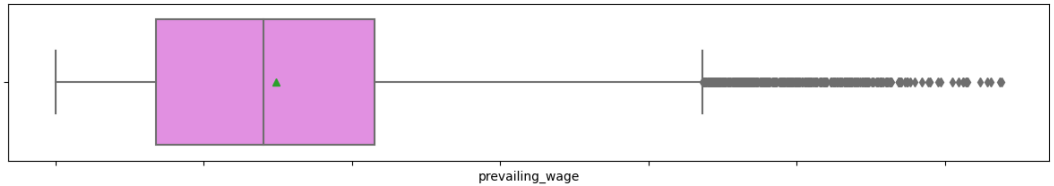
	count	mean	std	min	25%	50%	75%	max
no_of_employees	25480.0	5667.043210	22877.928848	-26.0000	1022.00	2109.00	3504.0000	602069.00
yr_of_estab	25480.0	1979.409929	42.366929	1800.0000	1976.00	1997.00	2005.0000	2016.00
prevailing_wage	25480.0	74455.814592	52815.942327	2.1367	34015.48	70308.21	107735.5125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

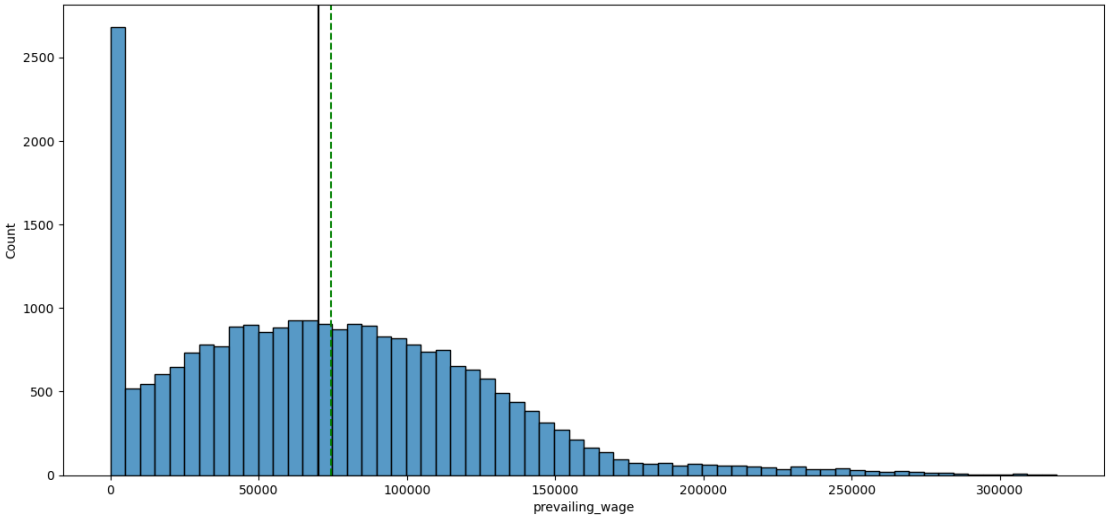
EDA Results – Univariate Analysis: 1. No of employees



- No_of_employees are observed being skewed to the right.
- However, we are not going to remove the outlier as this is real number.

EDA Results – Univariate Analysis: 2. Prevailing wage





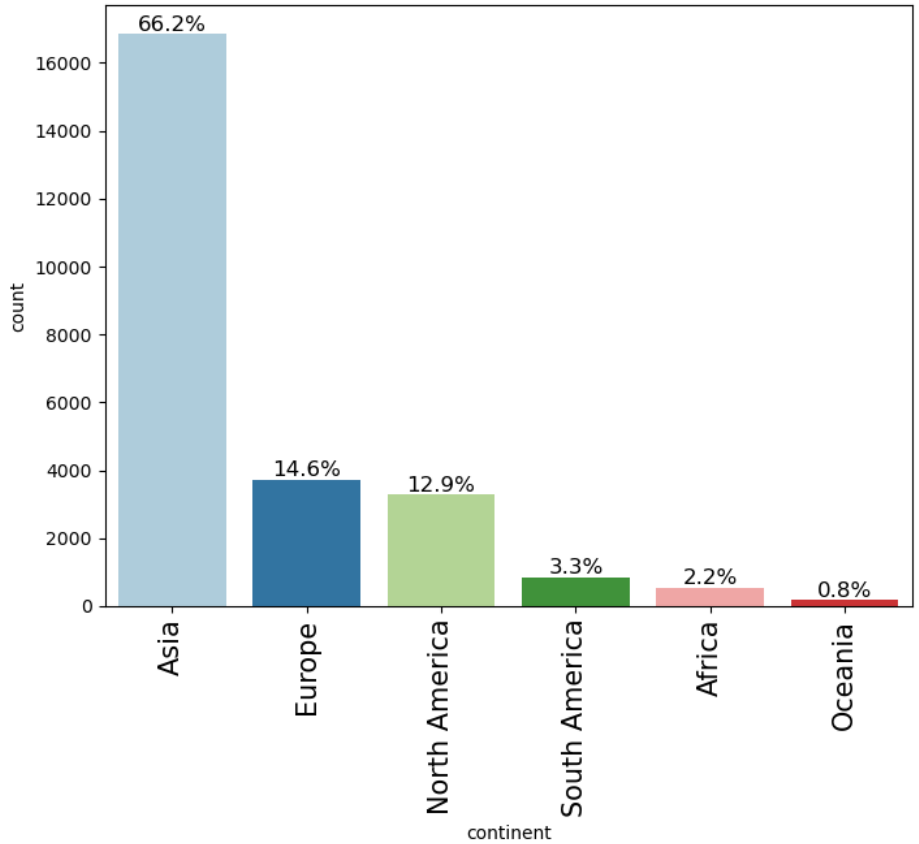
- Prevailing wage graph is observed having skewed to the right due to some outlier.
- It is observed the mean of prevailing wage is estimated about USD 74,000.

EDA Results – Univariate Analysis: 3. Prevailing wage < 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	Certified
634	Asia	Master's	N	N	834	1977	Northeast	3.3188	Hour	Y	Denied
839	Asia	High School	Y	N	4537	1999	West	61.1329	Hour	Y	Denied
876	South America	Bachelor's	Y	N	731	2004	Northeast	82.0029	Hour	Y	Denied
995	Asia	Master's	N	N	302	2000	South	47.4872	Hour	Y	Certified
...	...	...	...	...	...	...	...	...	...	...	...
25023	Asia	Bachelor's	N	Y	3200	1994	South	94.1546	Hour	Y	Denied
25258	Asia	Bachelor's	Y	N	3659	1997	South	79.1099	Hour	Y	Denied
25308	North America	Master's	N	N	82953	1977	Northeast	42.7705	Hour	Y	Denied
25329	Africa	Bachelor's	N	N	2172	1993	Northeast	32.9286	Hour	Y	Denied
25461	Asia	Master's	Y	N	2861	2004	West	54.9196	Hour	Y	Denied

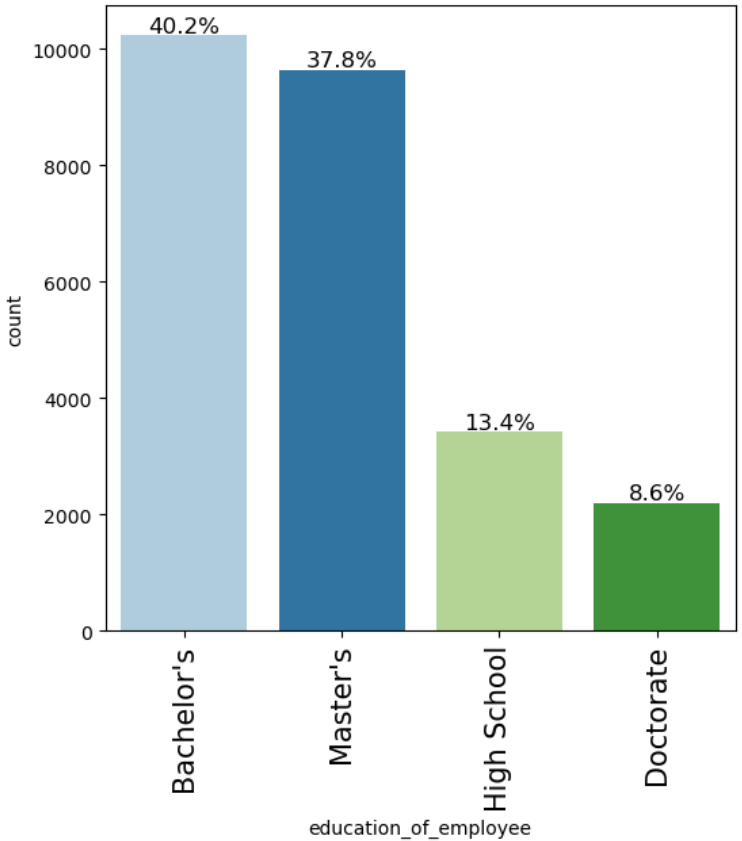
- It is found about 176 data is having prevailing wage less than 100. It is vary from region of employment and education of employees.

EDA Results – Univariate Analysis: 4. Observations on continent



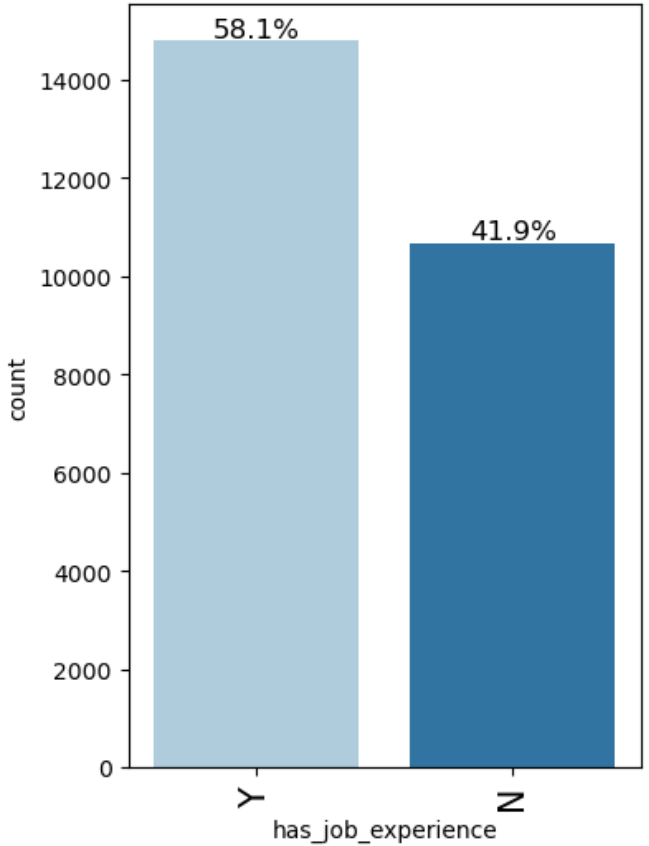
- It is observed employee from Asia is the highest and dominate about 66.2%, behid then is Europe 14.6% and North America 12.9%.
- South America, Africa and Oceania continent become the bottom 3 detractor with each contribute about 3.3%, 2.2% and 0.8%

EDA Results – Univariate Analysis: 5. Observations on education of employee



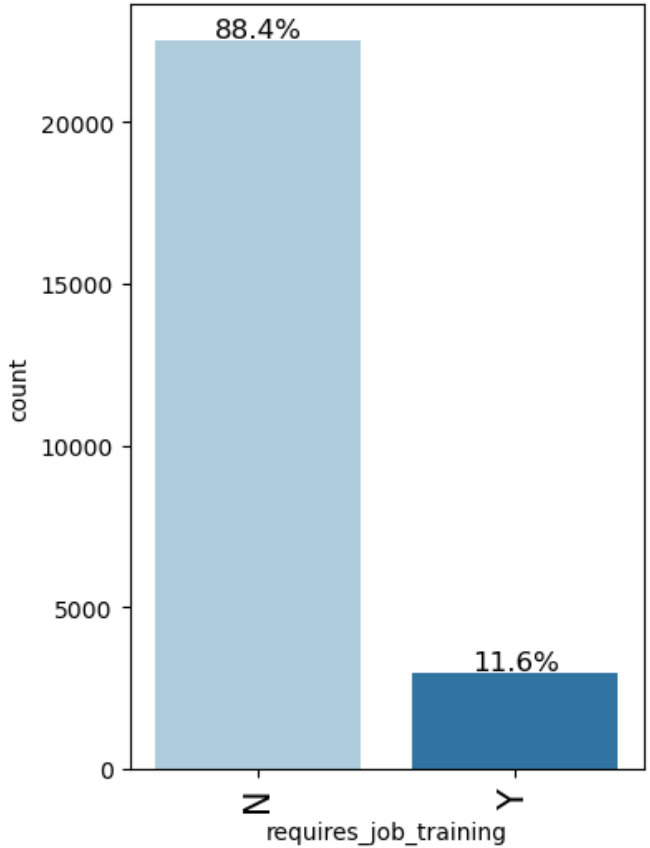
- Employee's having Bachelor's degree become the highest contributor (40.2%), followed by Master's degree (37.8%).

EDA Results – Univariate Analysis: 6. Observations on job experience



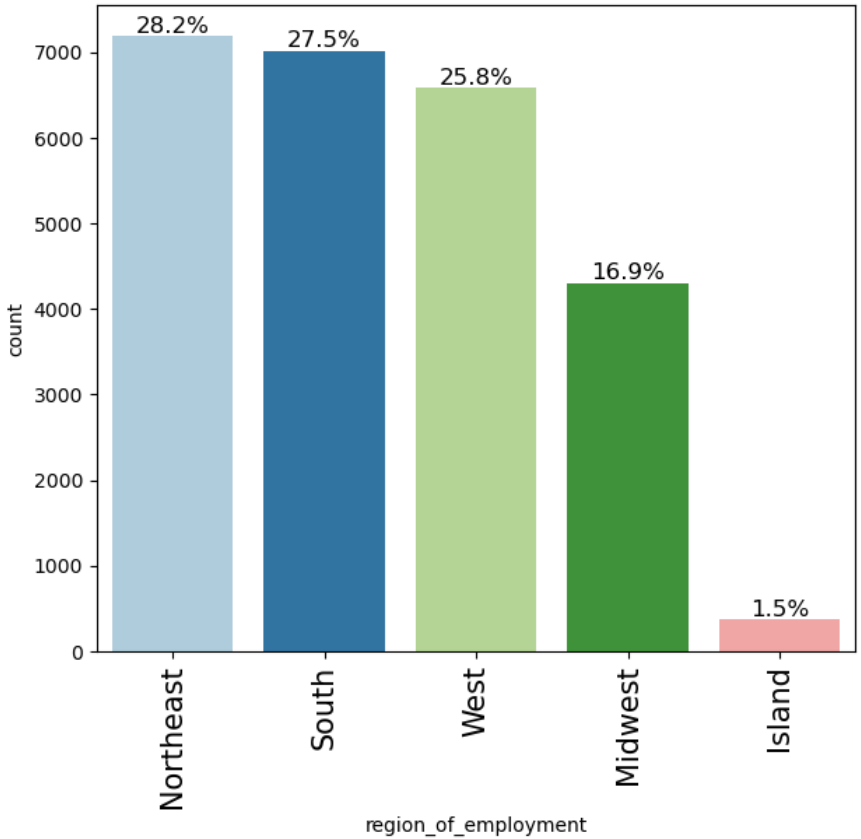
- Estimated around 58.1% of employees has a previous work experience and 41.9% has none.

EDA Results – Univariate Analysis: 7. Observations on job training



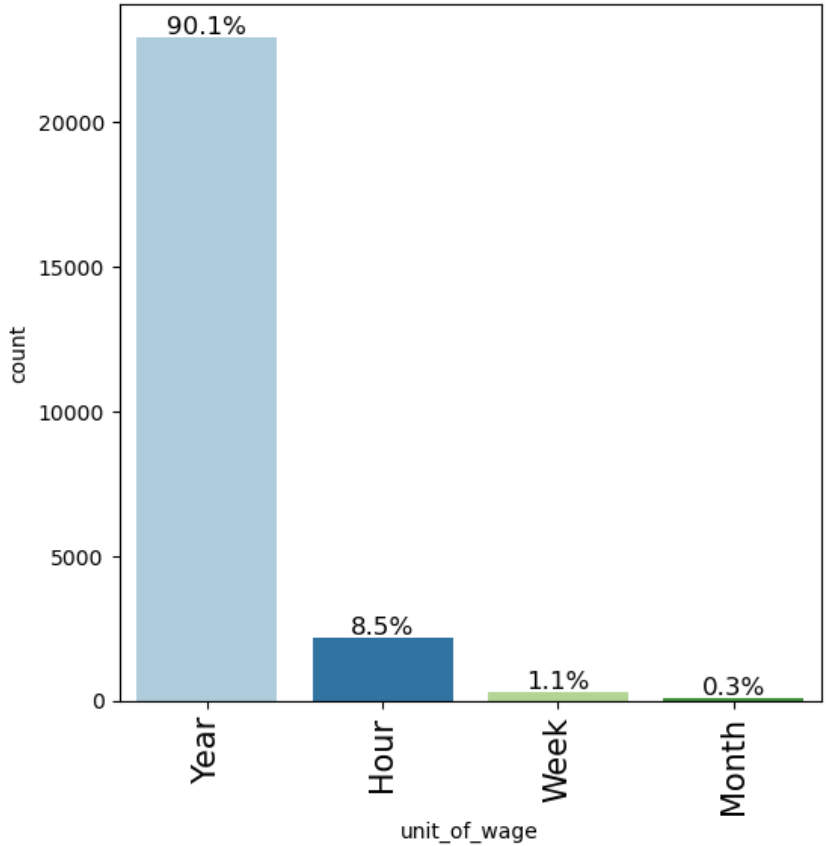
- 88.4% employees with no job training was identified, while 11.6% does requires job training.

EDA Results – Univariate Analysis: 8. Observations on region of employment



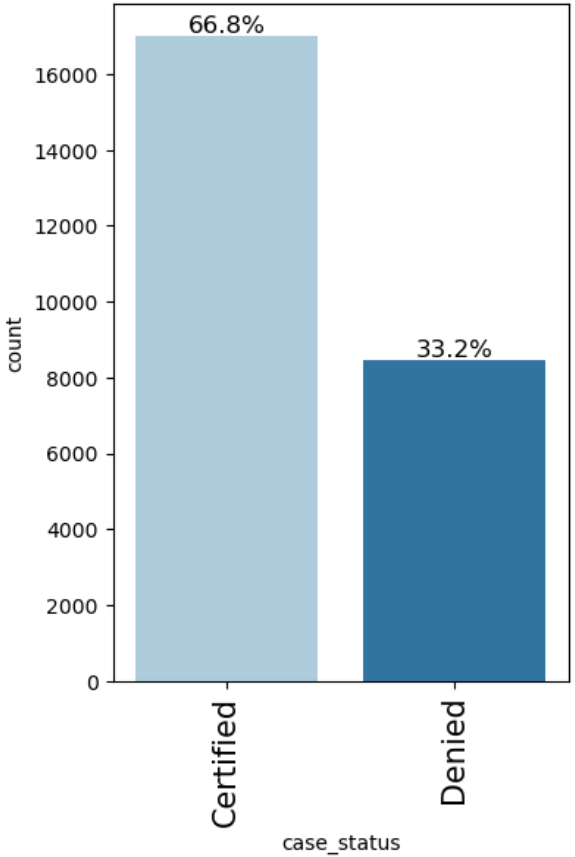
- 28.2% employess is going to Northeast for employment, followed by South (27.5%), West (25.8%) and Midwest (16.9%). Island becomes the bottom list on region of employment with only 1.5%.

EDA Results – Univariate Analysis: 9. Observations on unit of wage



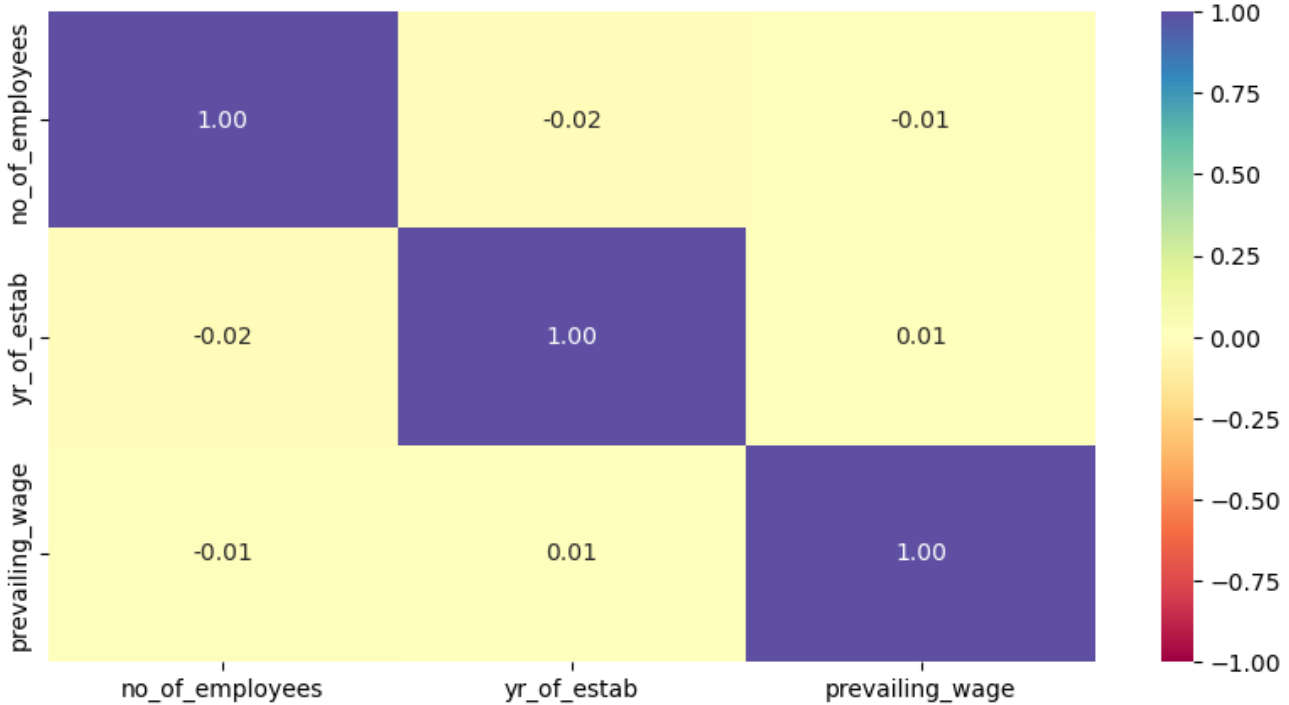
- Most preferred unit of wage is yearly, as it becomes the highest contributor, about 90.1%.

EDA Results – Univariate Analysis: 10. Observations on case status



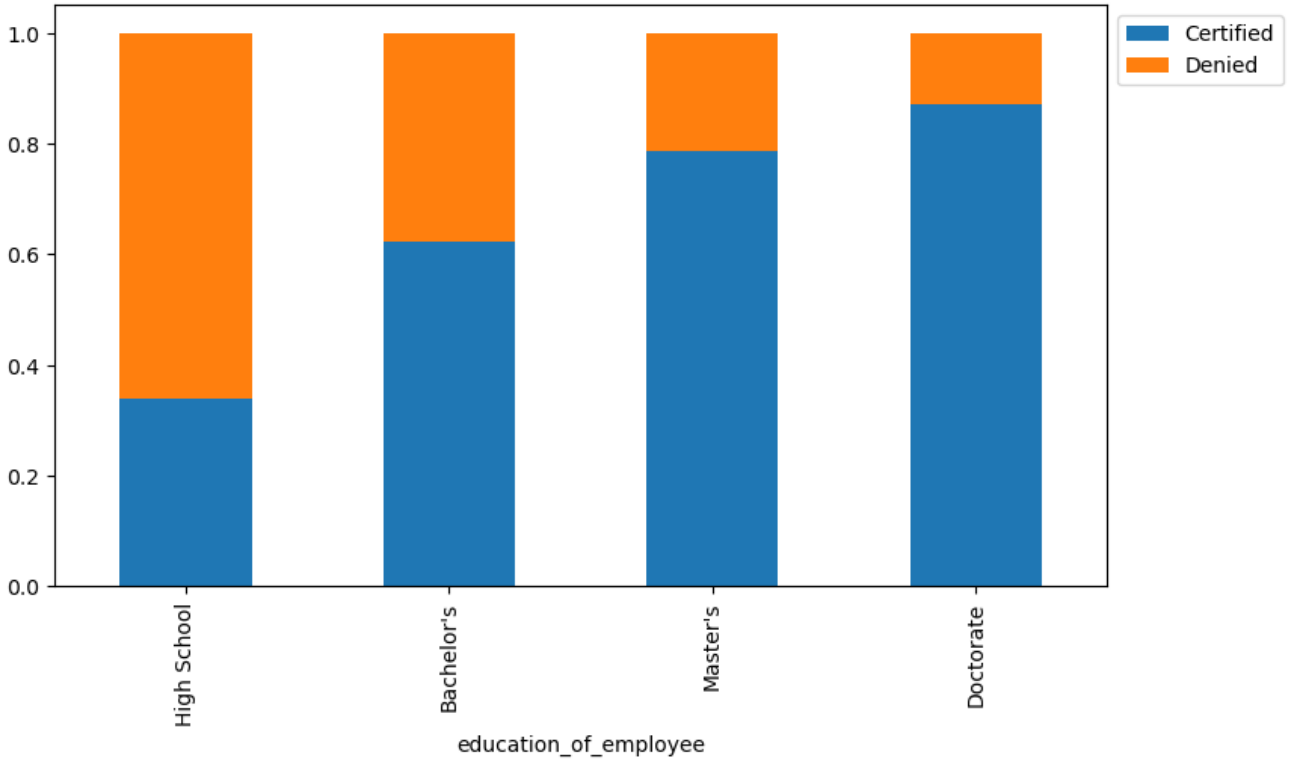
- It is observed about 66.8% was certified for a visa while 33.2% was denied.

EDA Results – Bivariate Analysis



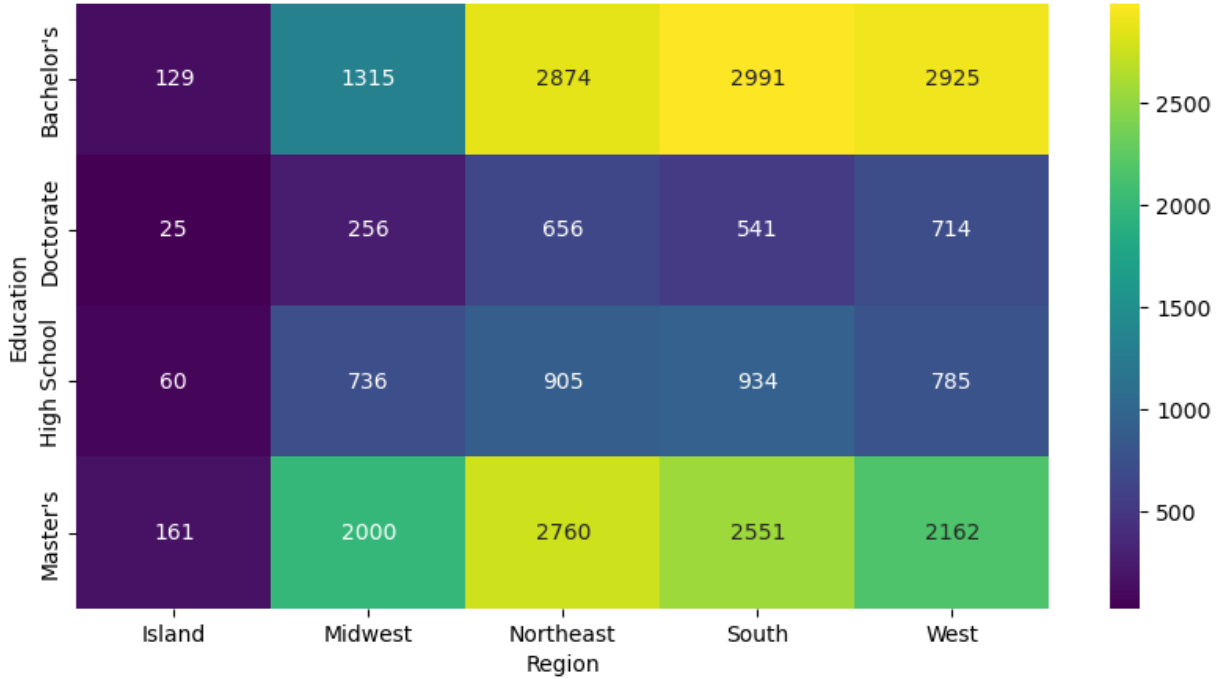
- All tested parameter is having low/weak relationship among each other.

EDA Results – Bivariate Analysis: 1. Does education has any impact on visa certification



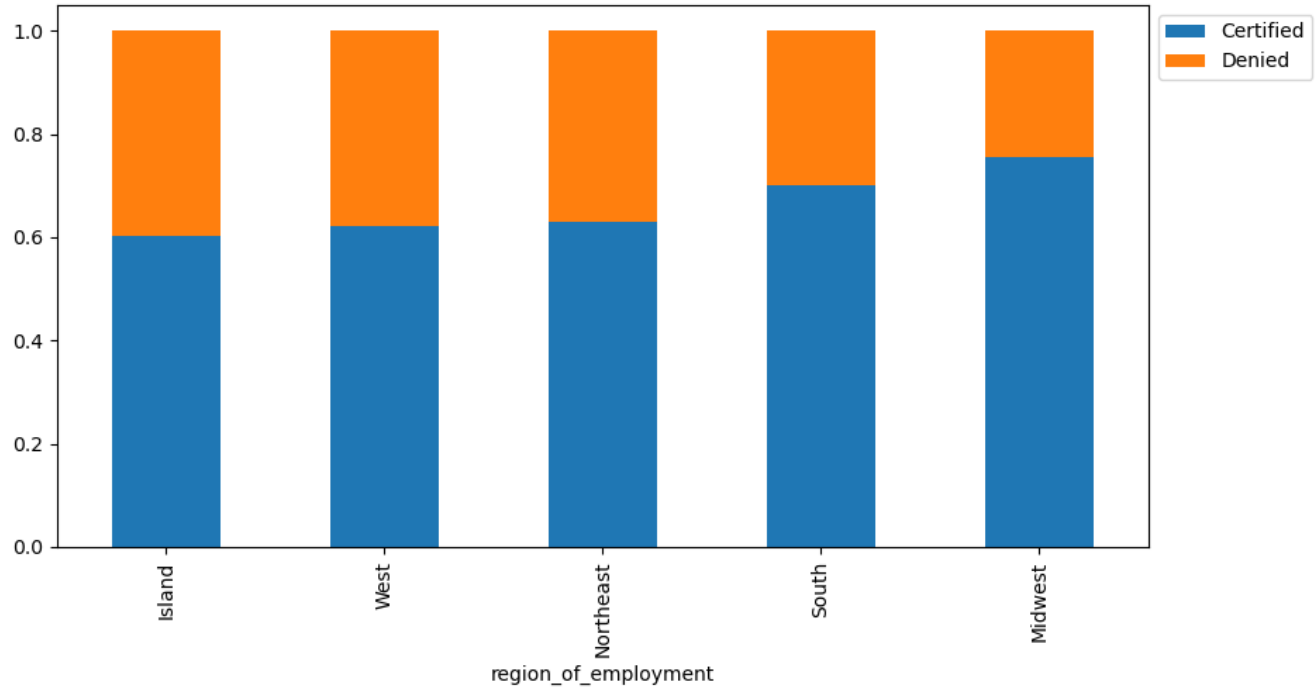
- The highest certified on visa was given to the Doctorate requestor, followed by Master's and Bachelor's degree requestor.

EDA Results – Bivariate Analysis: 2. Different regions have different requirements of talents



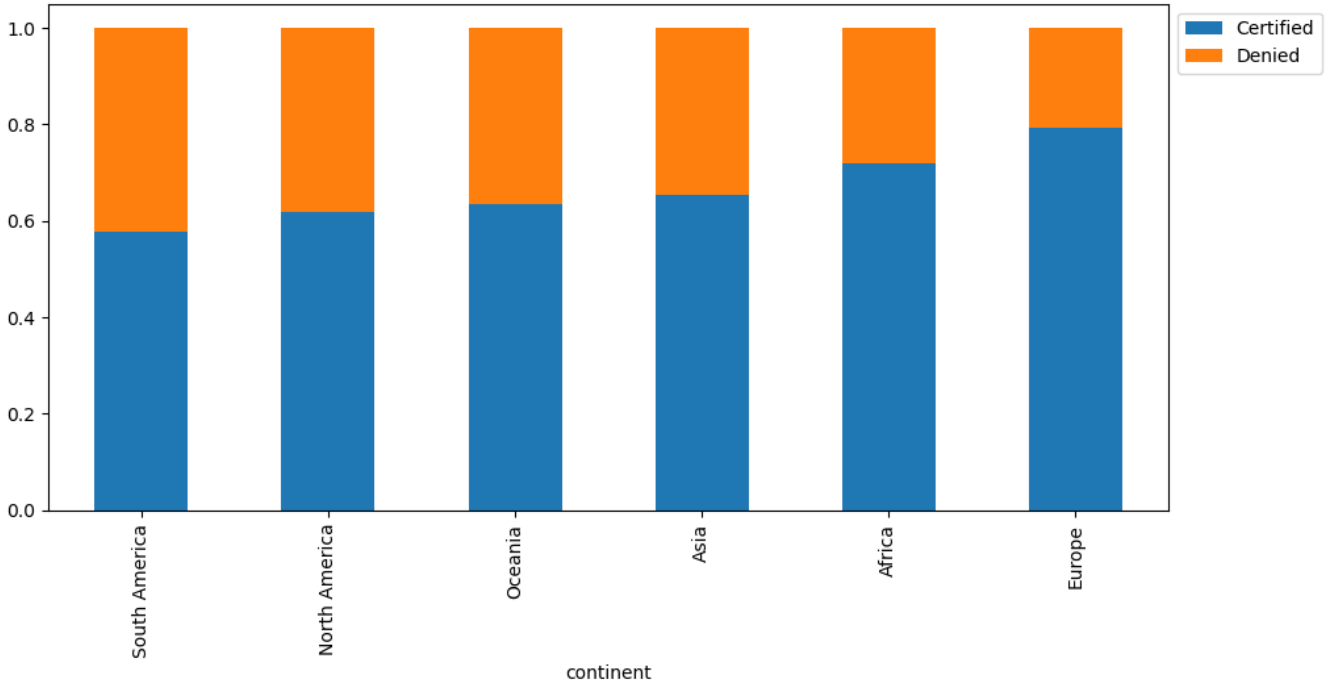
- Requirements of talents are varied based on regions.
- It is shown Northeast, South and West region is having Master's and Machelor's degree, followed by Midwest region.

EDA Results – Bivariate Analysis: 3. Visa certifications across each region



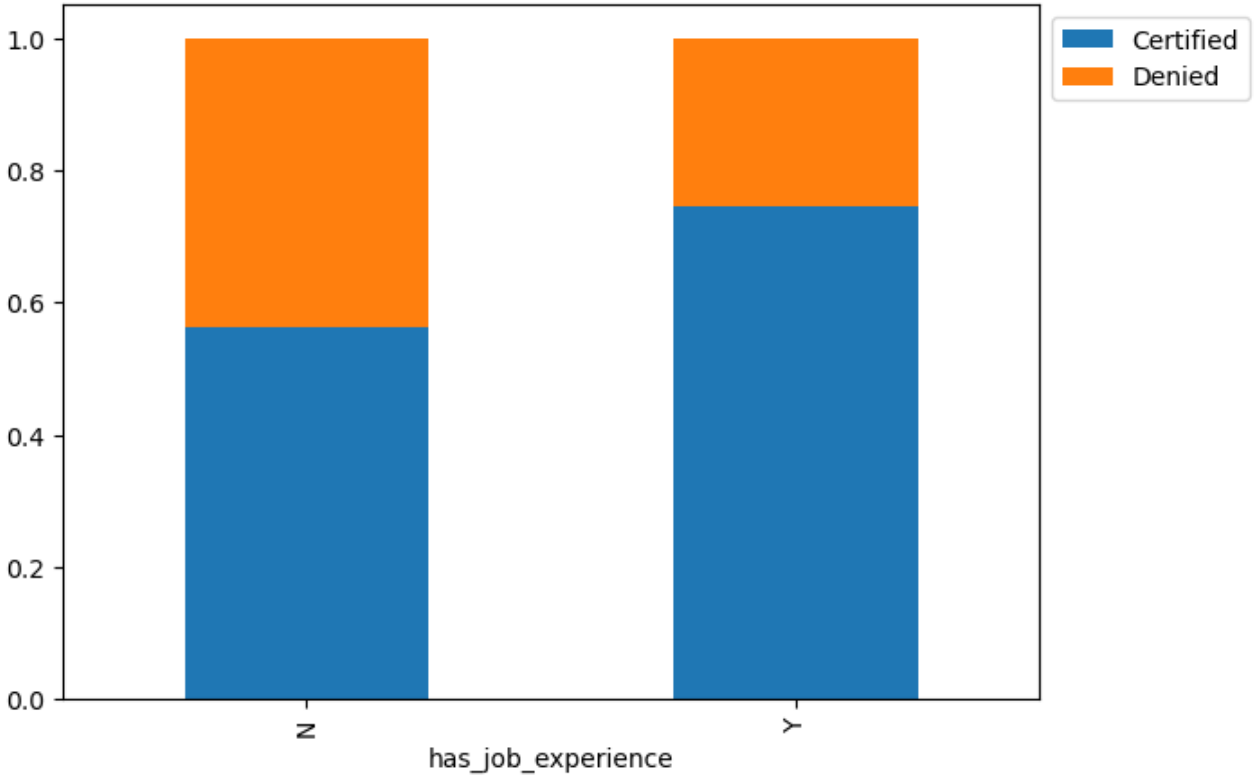
- Visa certifcations almost comparable for each regions.
- Nonetheless, the highest region is Midwest followed by South.

EDA Results – Bivariate Analysis: 4. Visa status vary across different continents



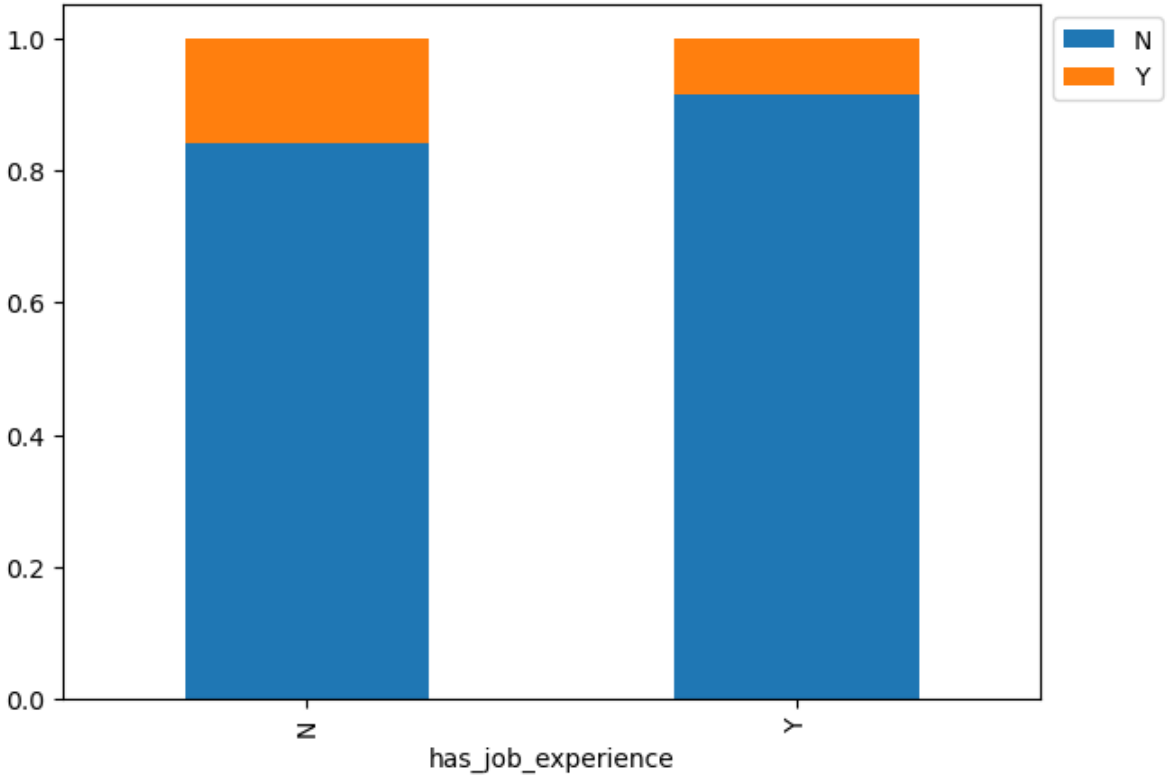
- Europe candidates become the highest contributor for certified visa, followed by Africa.
- Other continents; Asia, Oceania, North America and South America is comparable.

EDA Results – Bivariate Analysis: 5. Work experience has any influence over visa certification



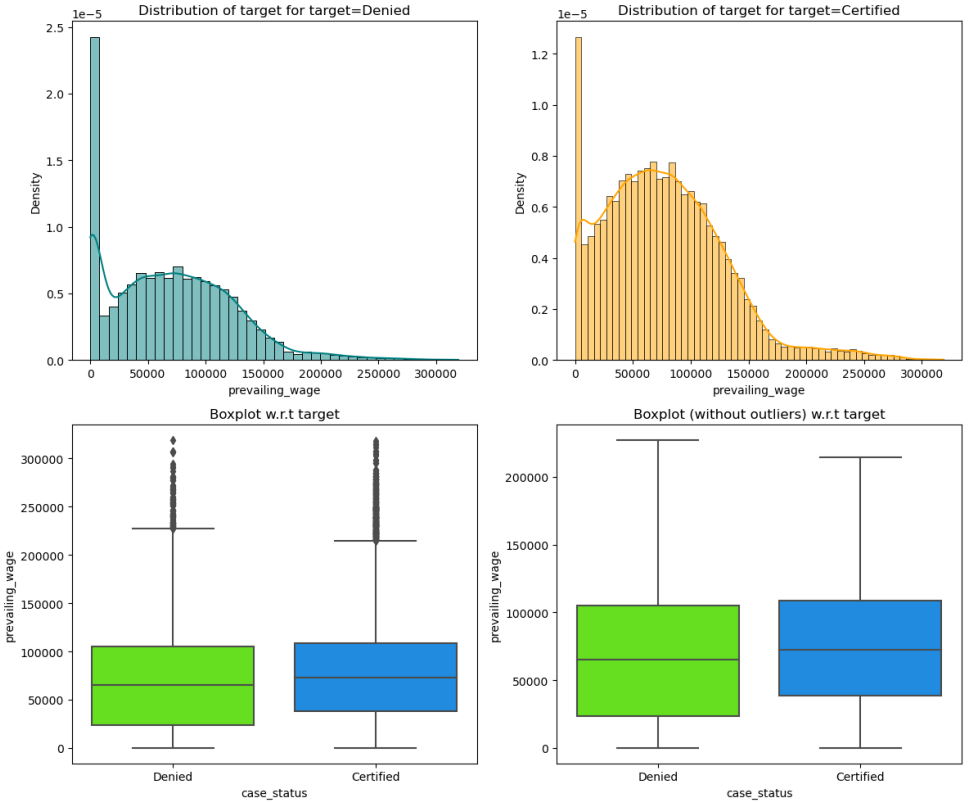
- Employee with job experience is having higher visa certification compared with no job experience.

EDA Results – Bivariate Analysis: 6. Prior work experience require any job training



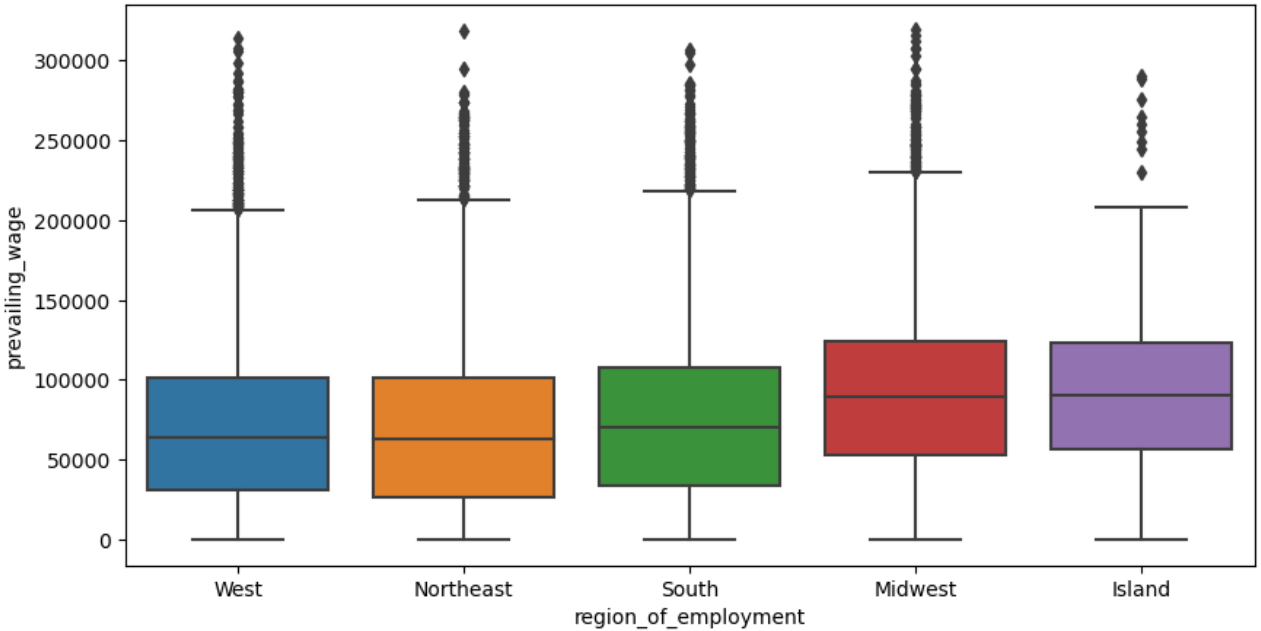
- It is also observed employee has previous job experience having higher percentage that require job traning compared with employee with no working experience.

EDA Results – Bivariate Analysis: 7. Visa status changes with the prevailing wage



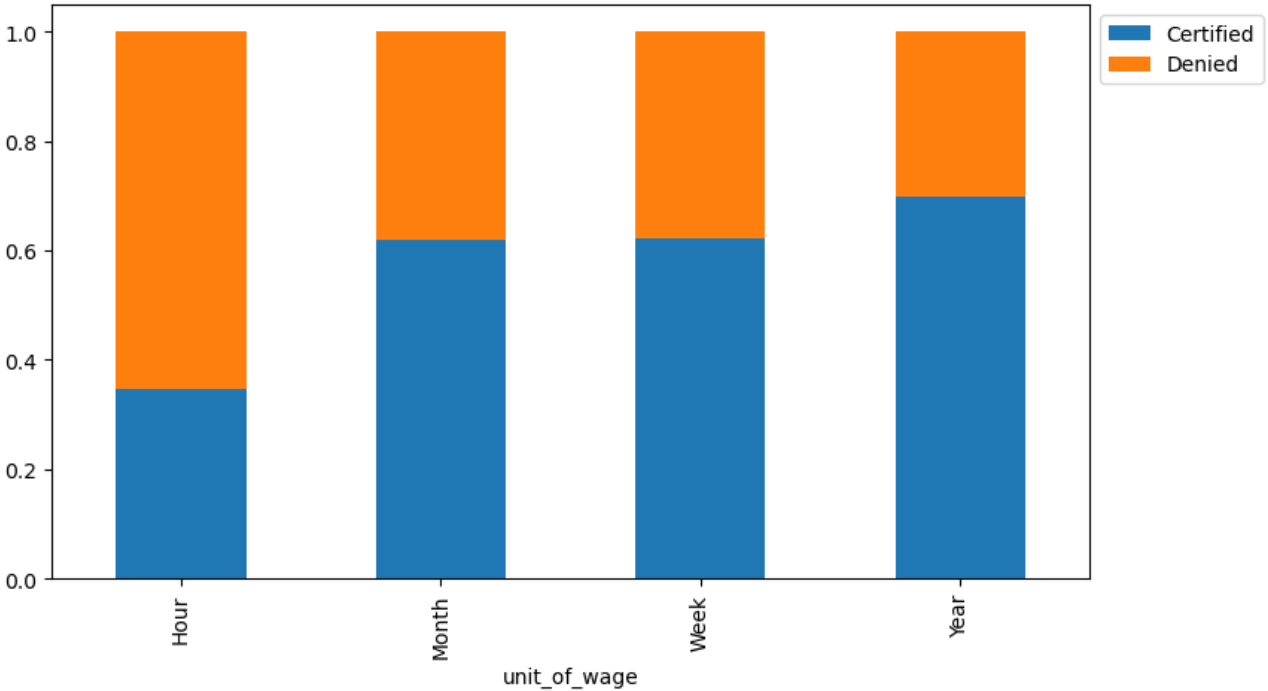
- No impact was observed on visa status, either certified or denied with the prevailing wage.

EDA Results – Bivariate Analysis: 8. is the prevailing wage similar across all regions of the US



- The prevailing wage is observed being simila/no significant difference across all region in the US.

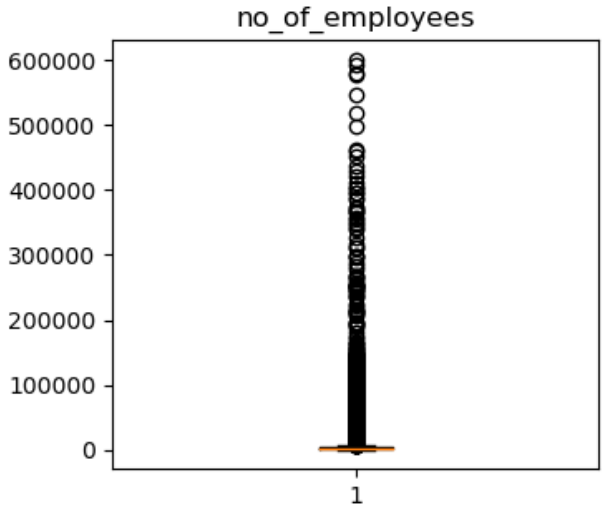
EDA Results – Bivariate Analysis: 9. Impact on visa applications getting certified

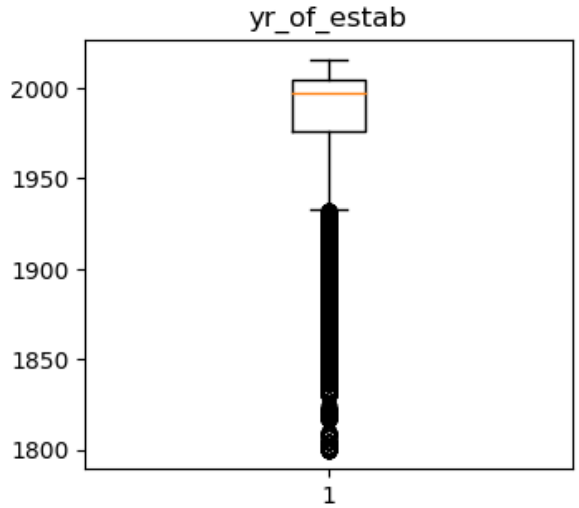


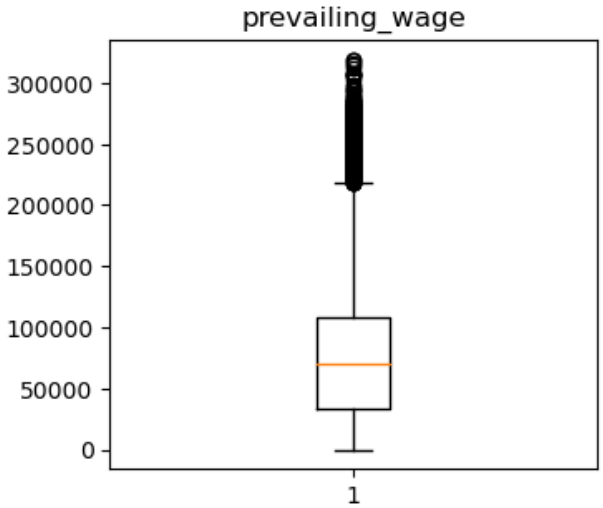
- No impact/significant change on yearly, weekly and monthly unit of wage with visa applications getting certified.

Data Preprocessing – Outlier Check

- No missing value was found in the data
- Found outlier on no_of_employees, yr_of_estab and prevailing_wage. However, data will be maintained as per original and will not be treated, as this is real data.







Data Preprocessing – Data preparation for modeling

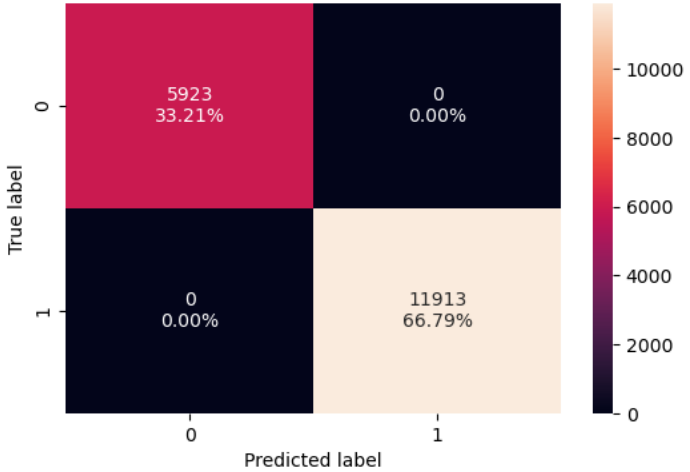
- Data preparation for modeling

	no_of_employees	yr_of_estab	prevailing_wage	continent_Asia	continent_Europe	continent_North America	continent_Oceania	continent_South America	education_of_employee_Doctorate	education_of_employee_High School	education_of_employee_Master's	has_job_experience_Y	requires_job_training_Y	region_of_employment_Midwest	region_of_employment_Northeast	region_of_employment_South	region_of_employment_West	unit_of_wage_Month	unit_of_wage_Week	unit_of_wage_Year	full_time_position_Y
0	14513	2007	592.2029	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1
1	2412	2002	83425.6500	1	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	1	1
2	44444	2008	122996.8600	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	1
3	98	1897	83434.0300	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1
4	1082	2005	149907.3900	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	1	1

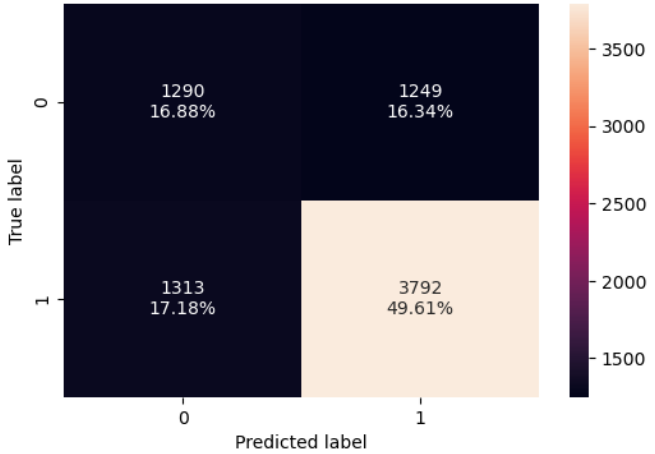
```
Number of rows in train data = (17836, 21)
Number of rows in test data = (7644, 21)
```

Model Building – Decision Tree

- Confusion matrix on training set:



- Confusion matrix on test set:



- Training performance:

	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

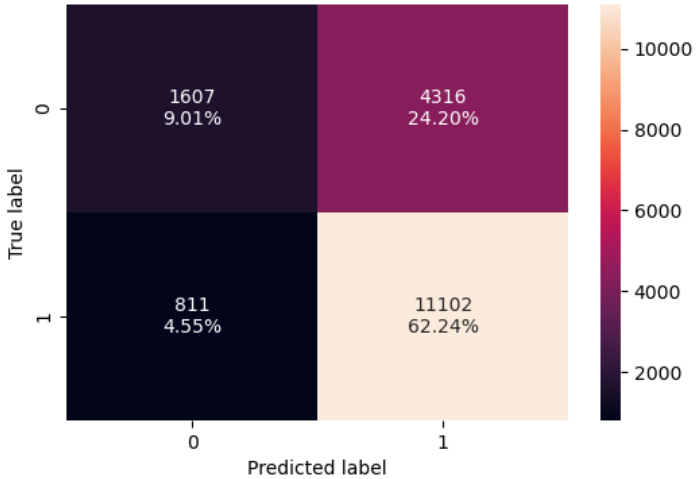
- Test performance:

	Accuracy	Recall	Precision	F1
0	0.664835	0.742801	0.752232	0.747487

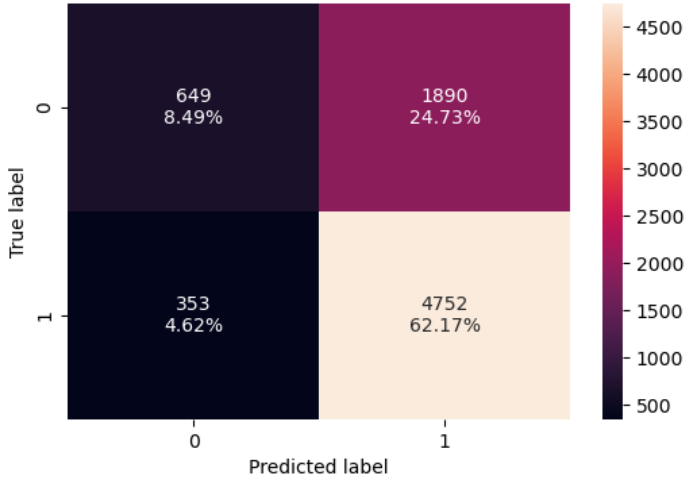
- Test performance starts to drop compared with training performance, a signs of overfitting. Therefore, we need to perform hyperparameter tuning.

Model Improvement – Decision Tree: Hyperparameter Tuning

- Confusion matrix on training set:



- Confusion matrix on test set:



- Training performance:

	Accuracy	Recall	Precision	F1
0	0.712548	0.931923	0.720067	0.812411

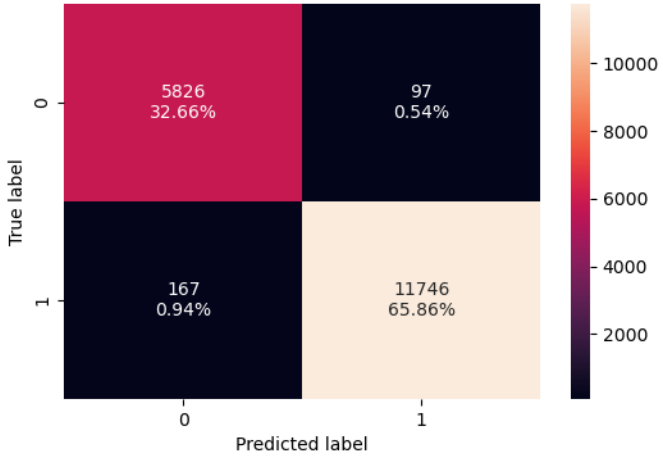
- Test performance:

	Accuracy	Recall	Precision	F1
0	0.706567	0.930852	0.715447	0.809058

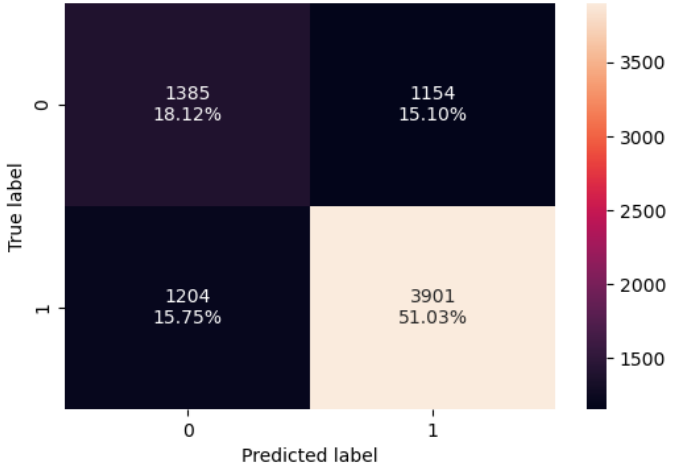
- Both result on training and test performance is comparable. Decision tree after hyperparameter tuning give a good and high percentage of F1 score.

Model Building – Bagging Classifier

- Confusion matrix on training set:



- Confusion matrix on test set:



- Training performance:

	Accuracy	Recall	Precision	F1
0	0.985198	0.985982	0.99181	0.988887

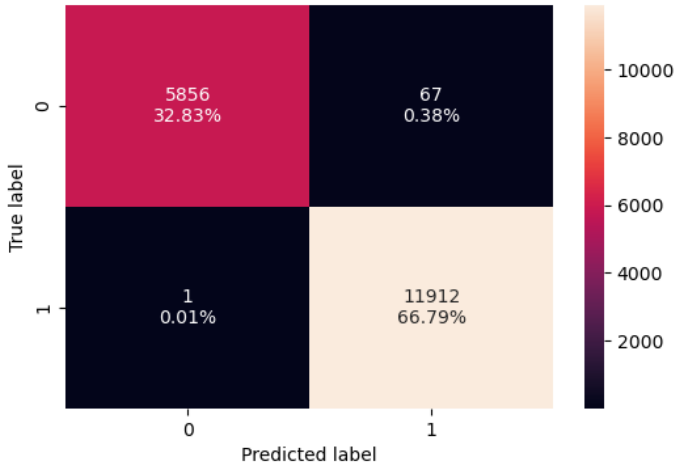
- Test performance:

	Accuracy	Recall	Precision	F1
0	0.691523	0.764153	0.771711	0.767913

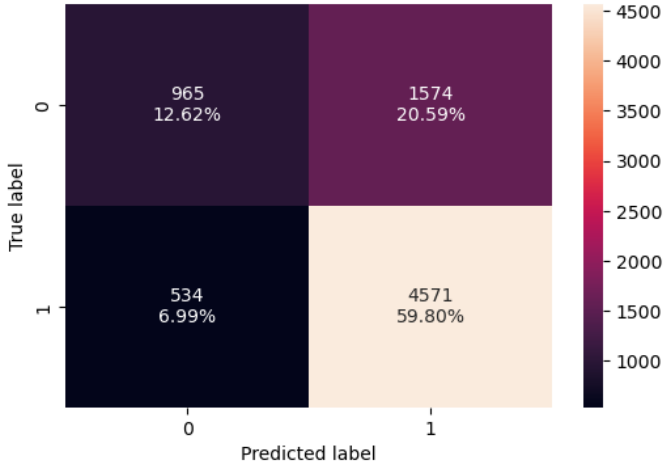
- Test performance starts to drop compared with training performance, a signs of overfitting. Therefore, we need to perform hyperparameter tuning.

Model Improvement – Bagging Classifier Hyperparameter Tuning

- Confusion matrix on training set:



- Confusion matrix on test set:



- Training performance:

	Accuracy	Recall	Precision	F1
0	0.996187	0.999916	0.994407	0.997154

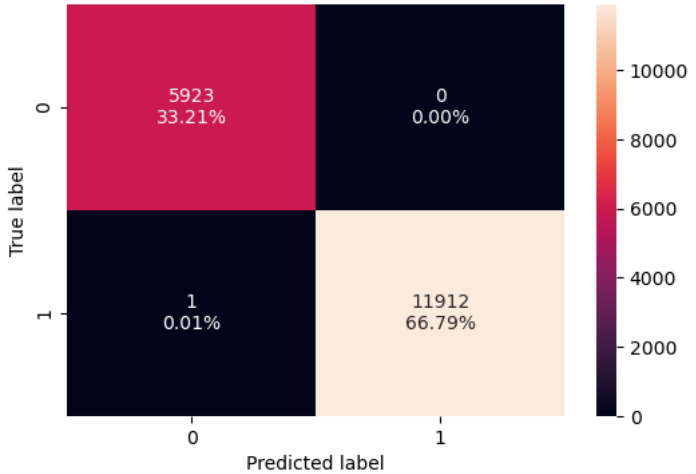
- Test performance:

	Accuracy	Recall	Precision	F1
0	0.724228	0.895397	0.743857	0.812622

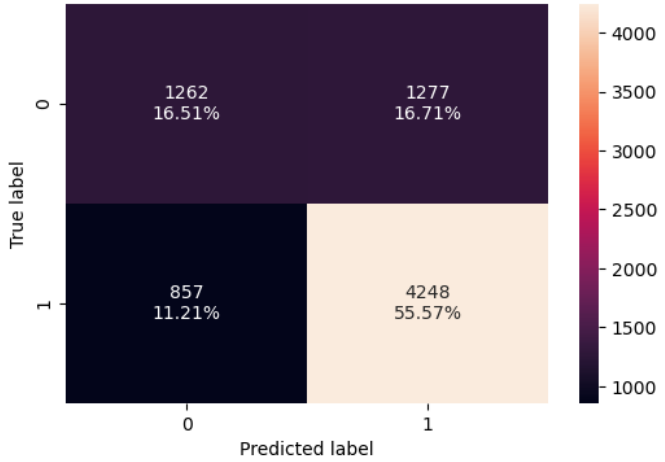
- Both performances improved after hyperparameter tuning. However, it still can be seen as overfitting due large gap between training and test performance.

Model Building – Random Forest

- Confusion matrix on training set:



- Confusion matrix on test set:



- Training performance:

	Accuracy	Recall	Precision	F1
0	0.999944	0.999916	1.0	0.999958

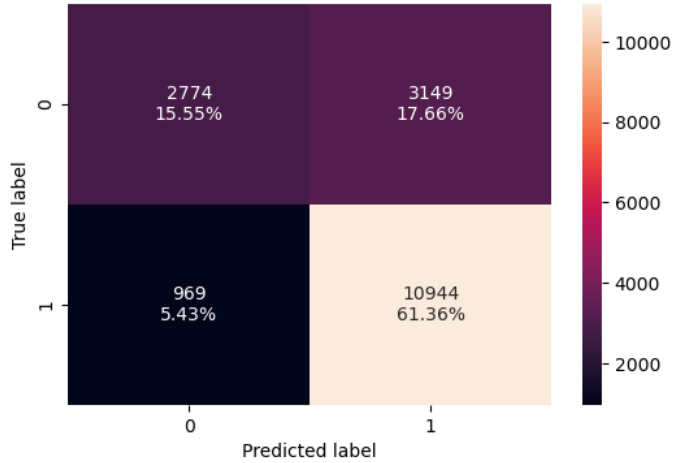
- Test performance:

	Accuracy	Recall	Precision	F1
0	0.720827	0.832125	0.768869	0.799247

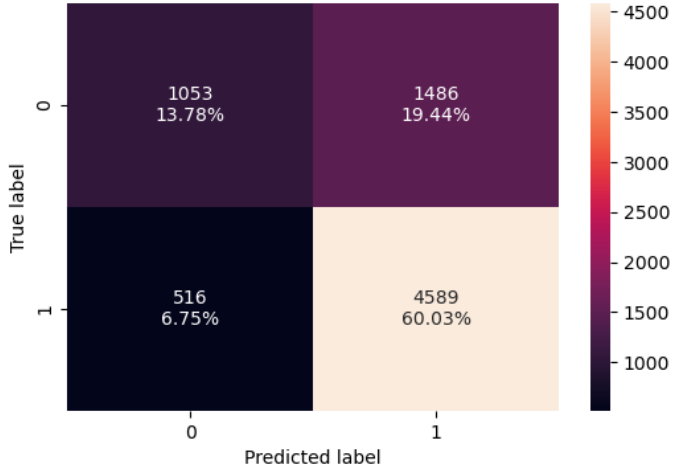
- Test performance starts to drop compared with training performance, a signs of overfitting. Therefore, we need to perform hyperparameter tuning.

Model Improvement – Random Forest Hyperparameter Tuning

- Confusion matrix on training set:



- Confusion matrix on test set:



- Training performance:

	Accuracy	Recall	Precision	F1
0	0.769119	0.91866	0.776556	0.841652

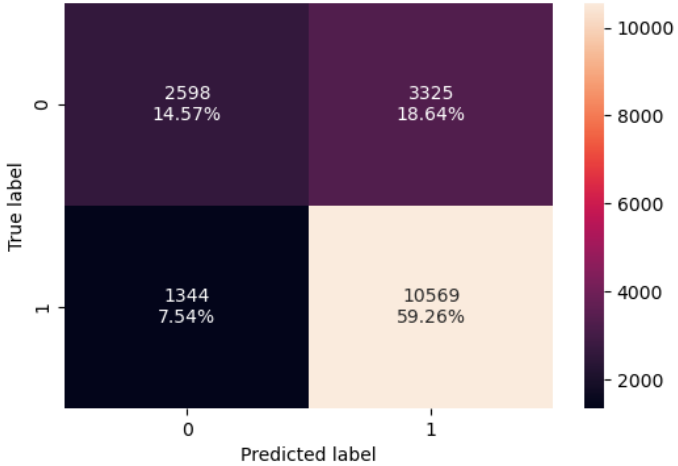
- Test performance:

	Accuracy	Recall	Precision	F1
0	0.738095	0.898923	0.755391	0.82093

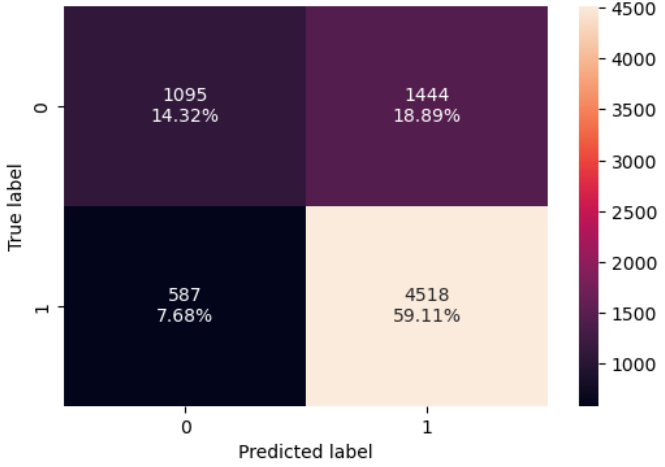
- Both performances improved after hyperparameter tuning. Precision and F1 score data after hyperparameter tuning has increased.

Model Building – AdaBoost Classifier

- Confusion matrix on training set:



- Confusion matrix on test set:



- Training performance:

	Accuracy	Recall	Precision	F1
0	0.738226	0.887182	0.760688	0.81908

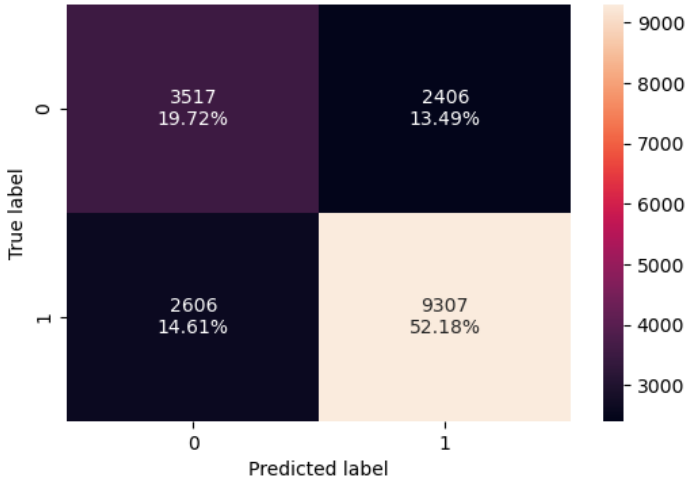
- Test performance:

	Accuracy	Recall	Precision	F1
0	0.734301	0.885015	0.757799	0.816481

- Both result are comparable, no signs of overfitting. However, we still might need to try hyperparameter tuning to improve the model.

Model Improvement – AdaBoost Classifier Hyperparameter Tuning

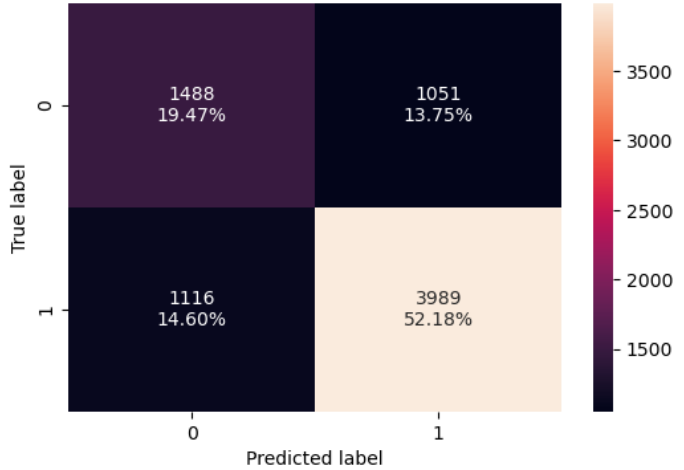
- Confusion matrix on training set:



- Training performance:

	Accuracy	Recall	Precision	F1
0	0.718995	0.781247	0.794587	0.787861

- Confusion matrix on test set:



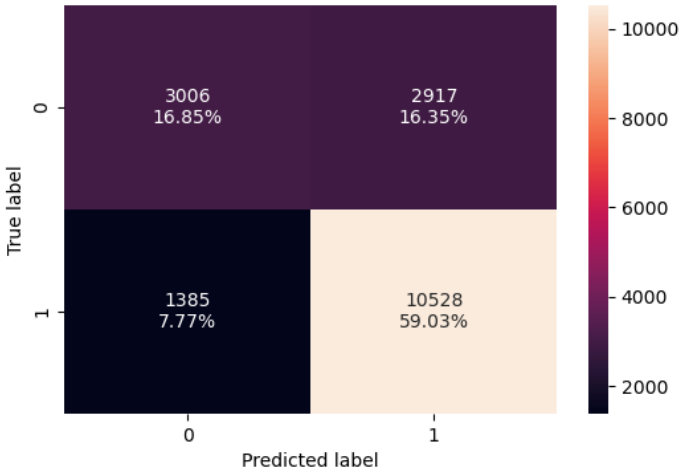
- Test performance:

	Accuracy	Recall	Precision	F1
0	0.71651	0.781391	0.791468	0.786397

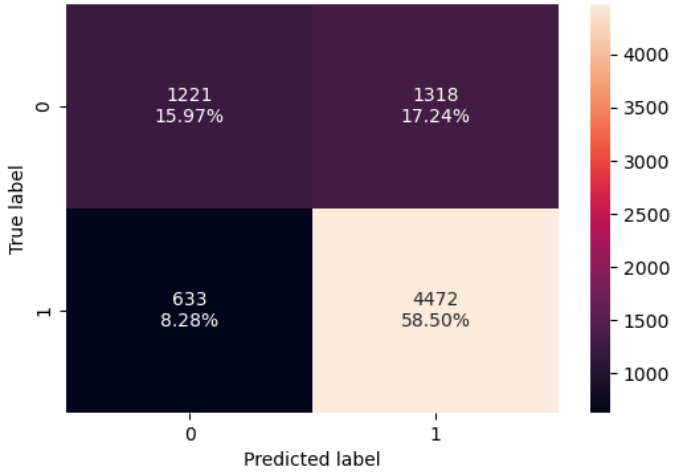
- No significant improvement can be observed after hyperparameter tuning.

Model Building – Gradient Boosting Classifier

- Confusion matrix on training set:



- Confusion matrix on test set:



- Training performance:

	Accuracy	Recall	Precision	F1
0	0.758802	0.88374	0.783042	0.830349

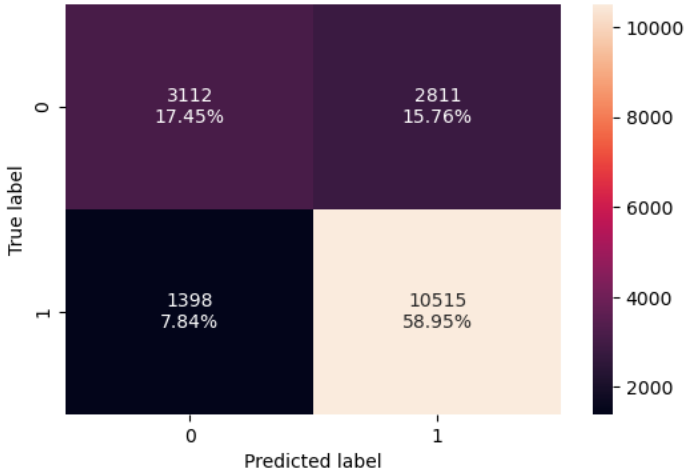
- Test performance:

	Accuracy	Recall	Precision	F1
0	0.744767	0.876004	0.772366	0.820927

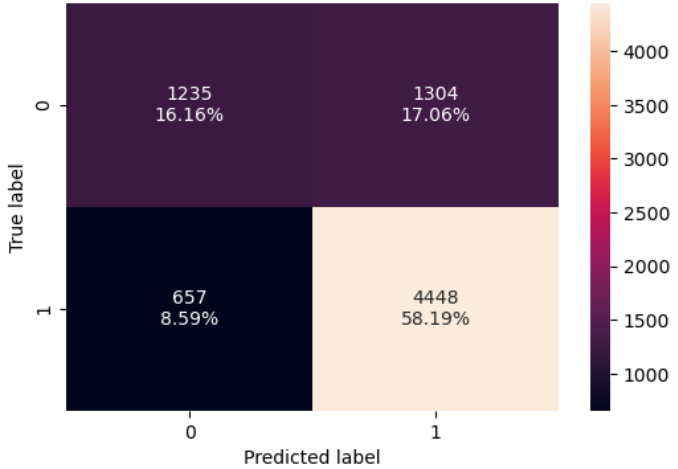
- Both result are comparable, no signs of overfitting. However, we still might need to try hyperparameter tuning to improve the model.

Model Improvement – Gradient Boosting Classifier Hyperparameter Tuning

- Confusion matrix on training set:



- Confusion matrix on test set:



- Training performance:

	Accuracy	Recall	Precision	F1
0	0.764017	0.882649	0.789059	0.833234

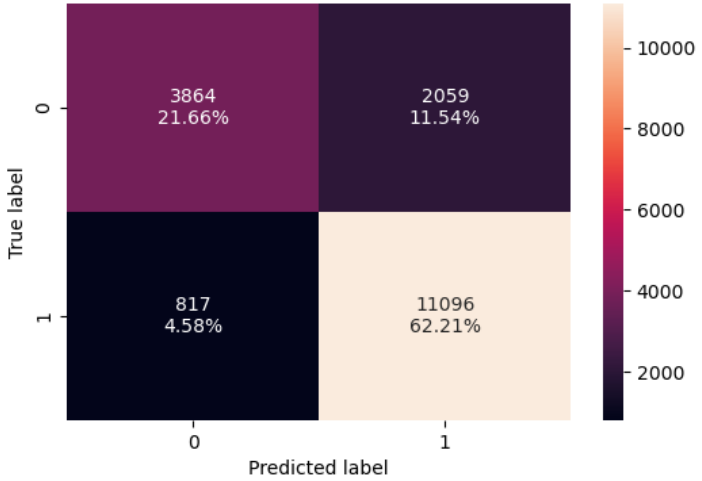
- Test performance:

	Accuracy	Recall	Precision	F1
0	0.743459	0.871303	0.773296	0.819379

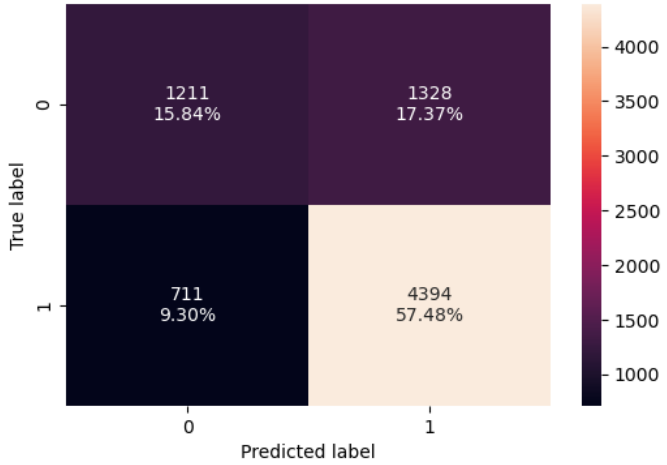
- No significant improvement can be observed after hyperparameter tuning.

Model Building – XG Boost Classifier

- Confusion matrix on training set:



- Confusion matrix on test set:



- Training performance:

	Accuracy	Recall	Precision	F1
0	0.838753	0.931419	0.843482	0.885272

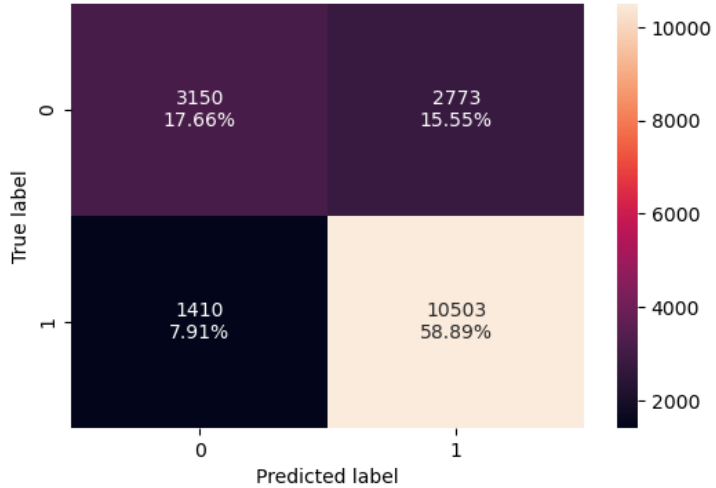
- Test performance:

	Accuracy	Recall	Precision	F1
0	0.733255	0.860725	0.767913	0.811675

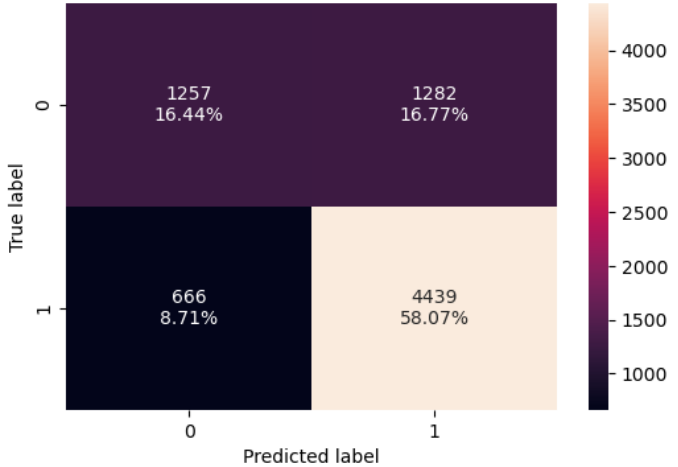
- Test performance starts to drop slightly compared with training performance, some signs of overfitting. Therefore, we need to perform hyperparameter tuning.

Model Improvement – XG Boost Classifier Hyperparameter Tuning

- Confusion matrix on training set:



- Confusion matrix on test set:



- Training performance:

	Accuracy	Recall	Precision	F1
0	0.765474	0.881642	0.791127	0.833935

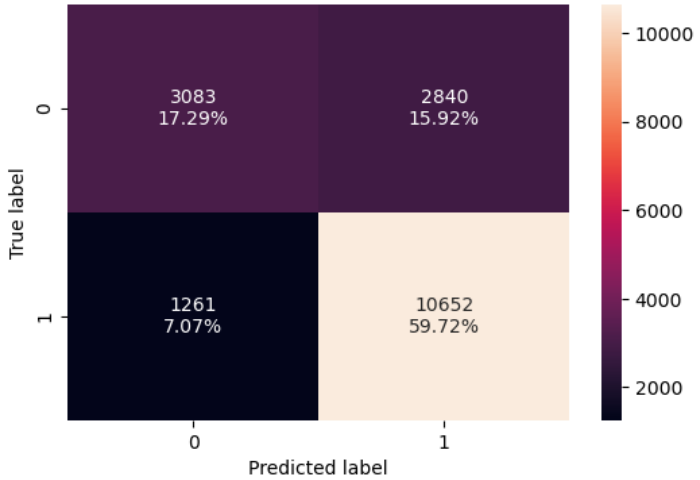
- Test performance:

	Accuracy	Recall	Precision	F1
0	0.74516	0.86954	0.775913	0.820063

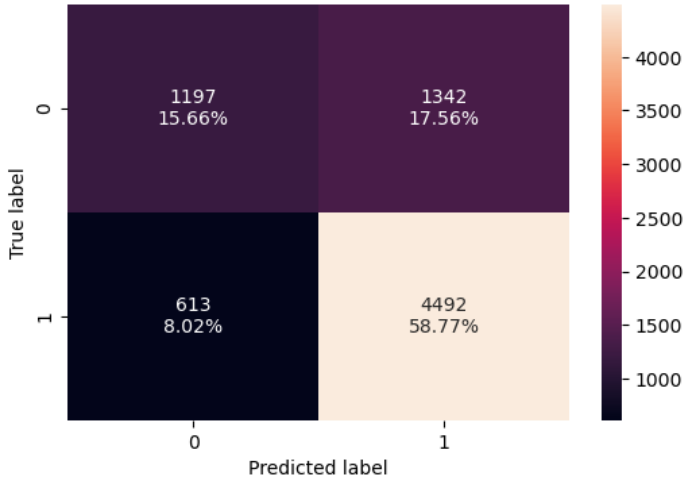
- Both performances improved after hyperparameter tuning. Precision and F1 score on test data after hyperparameter tuning is slightly increased.

Model Building – Stacking Classifier

- Confusion matrix on training set:



- Confusion matrix on test set:



- Training performance:

	Accuracy	Recall	Precision	F1
0	0.770072	0.894149	0.789505	0.838575

- Test performance:

	Accuracy	Recall	Precision	F1
0	0.744244	0.879922	0.769969	0.821282

- Stacking classifier result is comparable with XG Boost after hyperparameter tuning. Both data on training and test performance is comparable.

Model Performance Summary & Final Model Selection

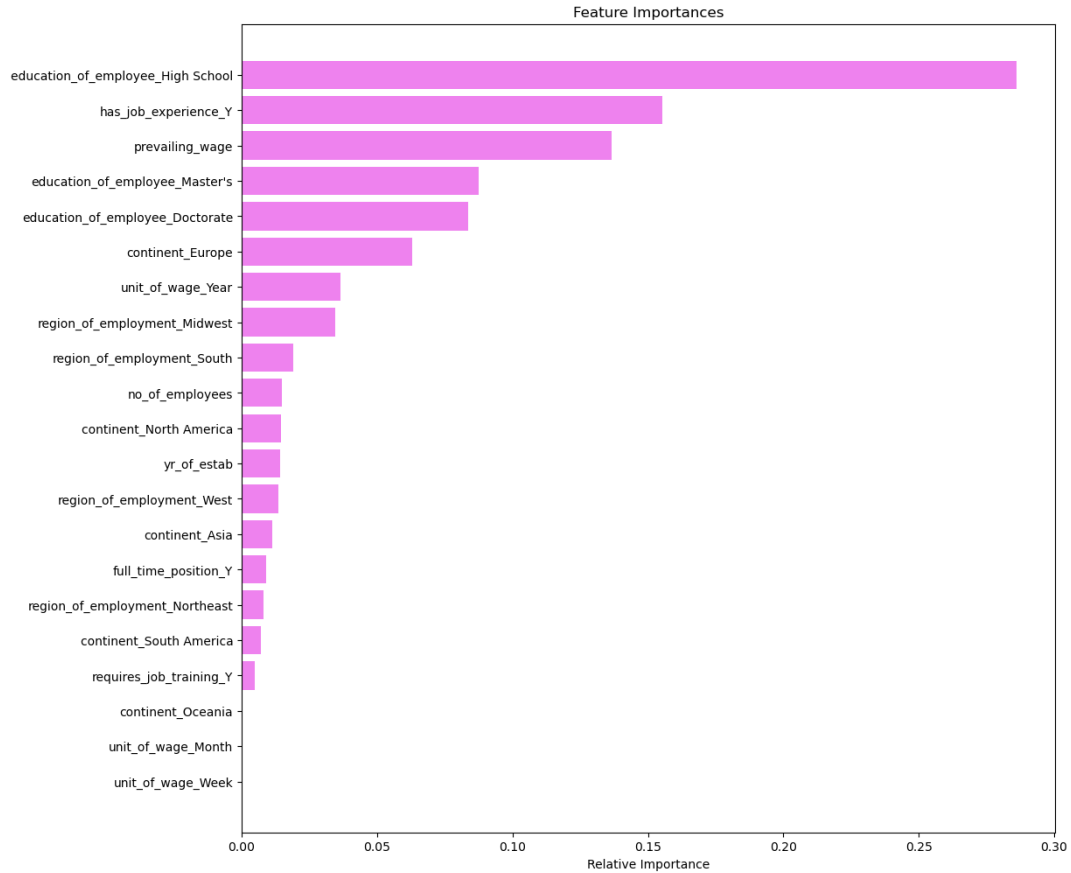
- Training performance:

	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	0.712548	0.985198	0.996187	0.999944	0.769119	0.738226	0.718995	0.758802	0.764017	0.838753	0.765474	0.770072
Recall	1.0	0.931923	0.985982	0.999916	0.999916	0.918660	0.887182	0.781247	0.883740	0.882649	0.931419	0.881642	0.894149
Precision	1.0	0.720067	0.991810	0.994407	1.000000	0.776556	0.760688	0.794587	0.783042	0.789059	0.843482	0.791127	0.789505
F1	1.0	0.812411	0.988887	0.997154	0.999958	0.841652	0.819080	0.787861	0.830349	0.833234	0.885272	0.833935	0.838575

- Testing performance:

	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	0.664835	0.706567	0.720827	0.738095	0.691523	0.724228	0.734301	0.716510	0.744767	0.743459	0.733255	0.745160	0.744244
Recall	0.742801	0.930852	0.832125	0.898923	0.764153	0.895397	0.885015	0.781391	0.876004	0.871303	0.860725	0.869540	0.879922
Precision	0.752232	0.715447	0.768869	0.755391	0.771711	0.743857	0.757799	0.791468	0.772366	0.773296	0.767913	0.775913	0.769969
F1	0.747487	0.809058	0.799247	0.820930	0.767913	0.812622	0.816481	0.786397	0.820927	0.819379	0.811675	0.820063	0.821282

Model Performance Summary & Final Model Selection



- Significant importance features on the top five; education_of_employee_High school, has_job_experience_Y, prevailing_wage, education_of_employee_Master's and education_of_employee_Doctorate.

Model Performance Summary & Final Model Selection

- Overview of final ML model and its parameters:
  - Decision tree, Bagging Classifier, Tuned Bagging Classifier and Random Forest are overfitting the training data set.
  - Most training and test performance has become comparable after hyperparameter tuning.
  - XG Boost can be seen no improvement after hyperparameter tuning.
  - All F1 score is improved after hyperparameter tuning.
  - If the model performance is selected based its simplicity and ease of interpretation, tuned Decision tree is the best compared with other model.
  - If the model performance is selected based on to improve/reduce the bias, tuned gradient boost is the best compared with other model.

APPENDIX

Data Background and Contents

- No data missing found

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

- Unique 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
-------------------------------------------------
```

Great Learning

Happy Learning !