EASYVISA Project

By: Nihal Kala

Problem Statement

- Companies are trying to find a way to hire the most talented and competent candidates for the job.
- Companies are looking to fill up shortages in companies by allowing foreign workers to come with a visa on a permanent or temporary basis.
- The number of processed applications increased by 9% in 2016.
- The objective is to find a machine learning solution that will be able to filter out the most qualified candidates for having a better chance of Visa approval.
- In that way, the facilitation of Visa approvals and a profile will be characterized based on factors that would significantly impact approval and denial status will be implemented.

Head and Tail of Data

	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	South	149907.3900	Year	Y	Certified

	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
25475	EZYV25476	Asia	Bachelor's	Y	Y	2601	2008	South	77092.57	Year	Y	Certified
25476	EZYV25477	Asia	High School	Y	N	3274	2006	Northeast	279174.79	Year	Y	Certified
25477	EZYV25478	Asia	Master's	Y	N	1121	1910	South	146298.85	Year	N	Certified
25478	EZYV25479	Asia	Master's	Y	Y	1918	1887	West	86154.77	Year	Y	Certified
25479	EZYV25480	Asia	Bachelor's	Y	N	3195	1960	Midwest	70876.91	Year	Y	Certified

Shape of Data: 25480 rows, 12 columns
No duplicated values

Data Info

```
<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
```

Data Summary

	no_of_employees	yr_of_estab	prevailing_wage
count	25480.000000	25480.000000	25480.000000
mean	5667.043210	1979.409929	74455.814592
std	22877.928848	42.366929	52815.942327
min	-26.000000	1800.000000	2.136700
25%	1022.000000	1976.000000	34015.480000
50%	2109.000000	1997.000000	70308.210000
75%	3504.000000	2005.000000	107735.512500
max	602069.000000	2016.000000	319210.270000

Column Counts

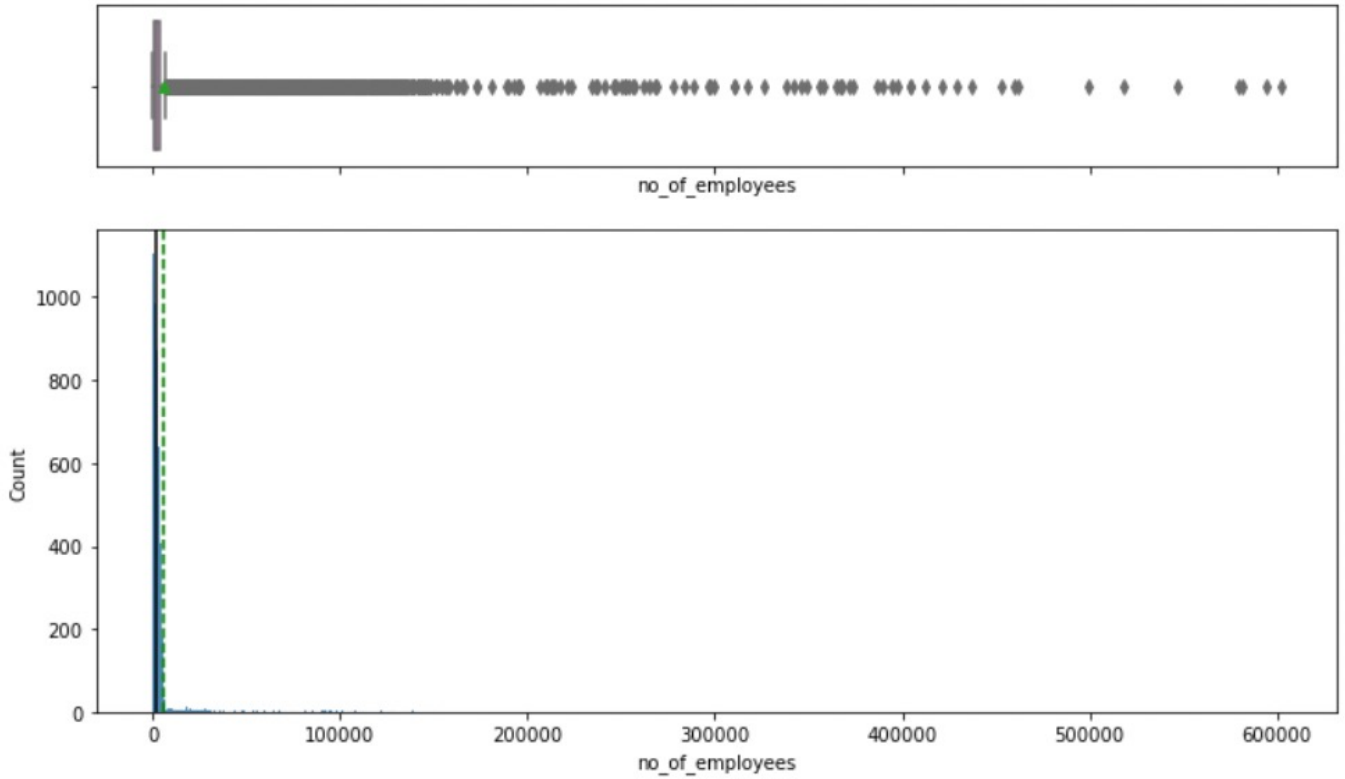
- Negative numbers in employee columns: 33
- Counting each unique category in categorical variables seen in right photo.

Number of Unique values in bottom photo.

```
<bound method Series.unique of 0          EZYV01
1           EZYV02
2           EZYV03
3           EZYV04
4           EZYV05
           ...
25475    EZYV25476
25476    EZYV25477
25477    EZYV25478
25478    EZYV25479
25479    EZYV25480
Name: case_id, Length: 25480, dtype: object>
```

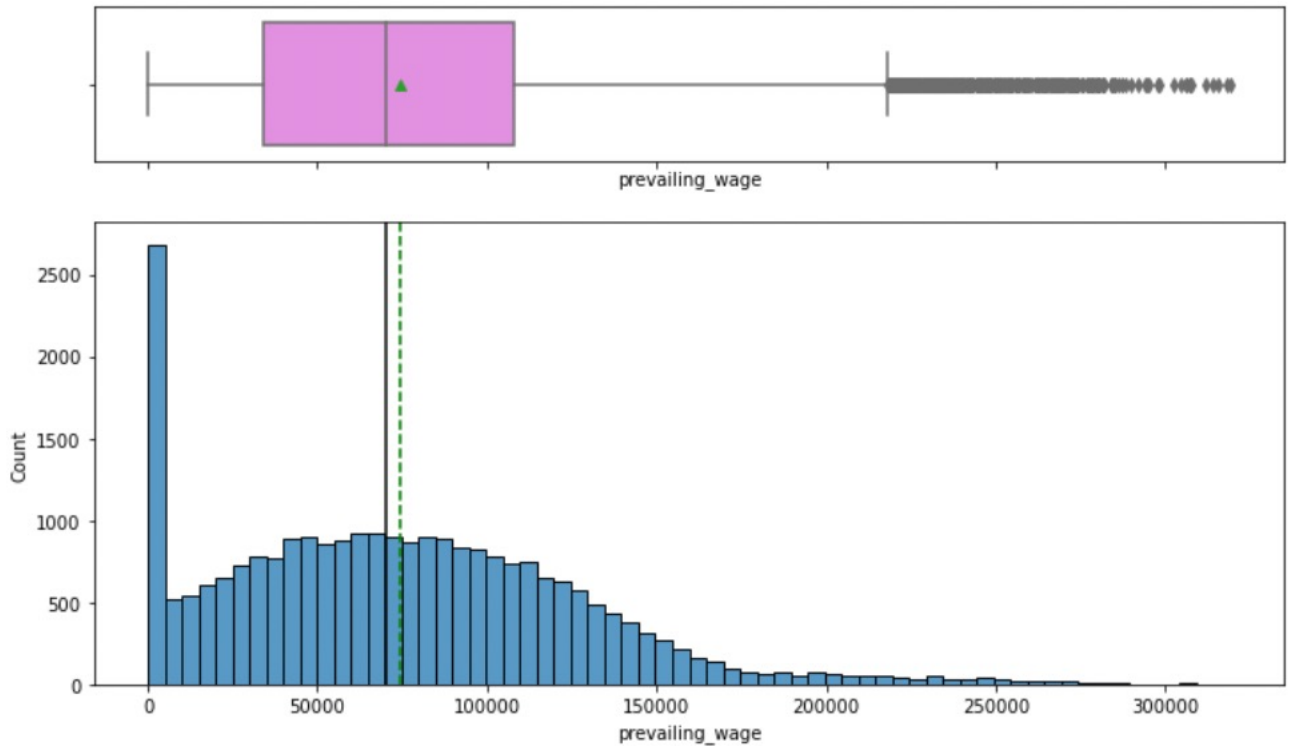
```
EZYV9205     1
EZYV23550    1
EZYV6077     1
EZYV3366     1
EZYV20968    1
            ..
EZYV3249     1
EZYV2615     1
EZYV25320    1
EZYV16584    1
EZYV23624    1
Name: case_id, Length: 25480, dtype: int64
-------------------------------------------------
Asia             16861
Europe            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
-------------------------------------------------
```

Histogram-Box Plot (No. of Employees)



-Distribution is right-skewed.
-Many outliers to the right in variable.
-Majority of companies have less than 100,000 employees.
-With drop of negative values in number of employees, this impacts the mean and median values.

Histogram-Box Plot (Prevailing Wage)



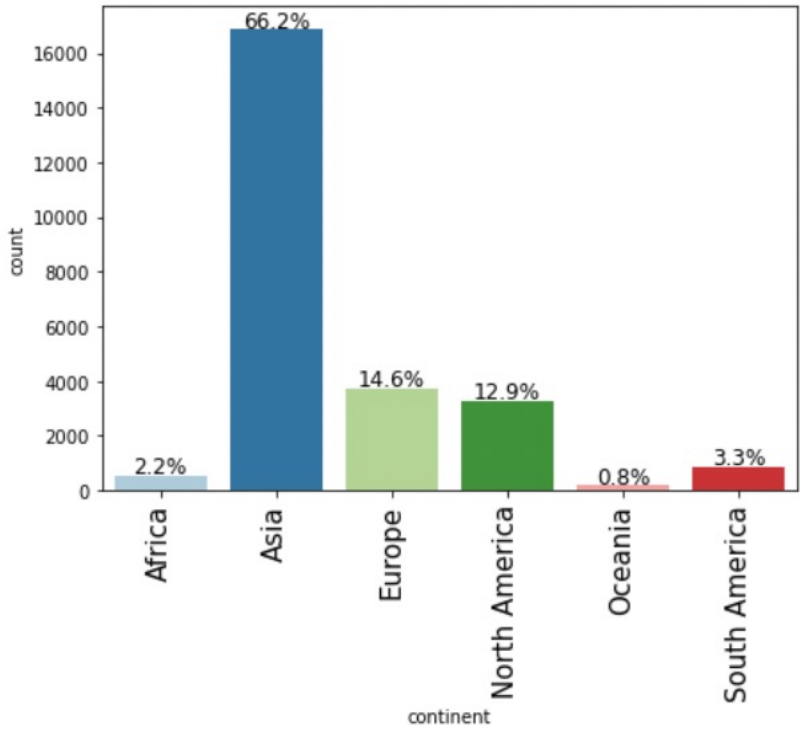
-Distribution is right-skewed with many outliers to right.
Range from 2.13 to 391210
Mean of 74455.8
Median of 70308.21
IQR from 34015.48 to 107735.51

Observations with less than 100 prevailing wage

```
338      15.7716
634       3.3188
839      61.1329
876      82.0029
995      47.4872
          ...
25023    94.1546
25258    79.1099
25308    42.7705
25329    32.9286
25461    54.9196
Name: prevailing_wage, Length: 176, dtype: float64
```

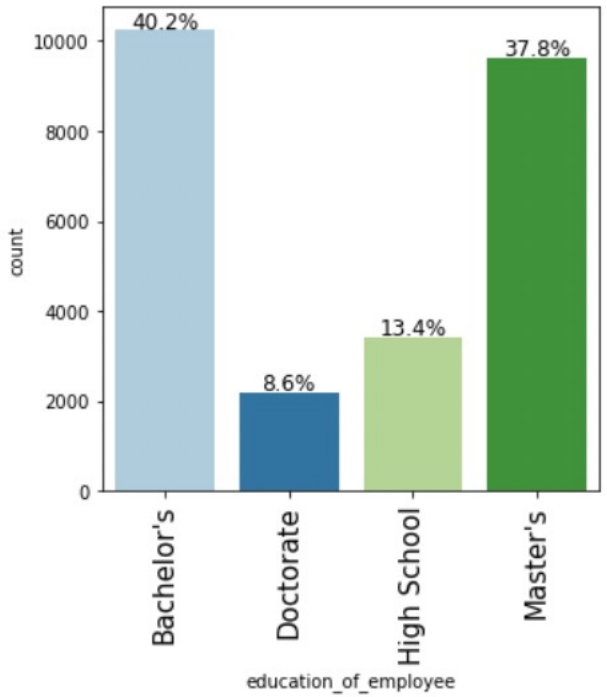
176 observations with less than 100 prevailing wage

Barplot (Continent)



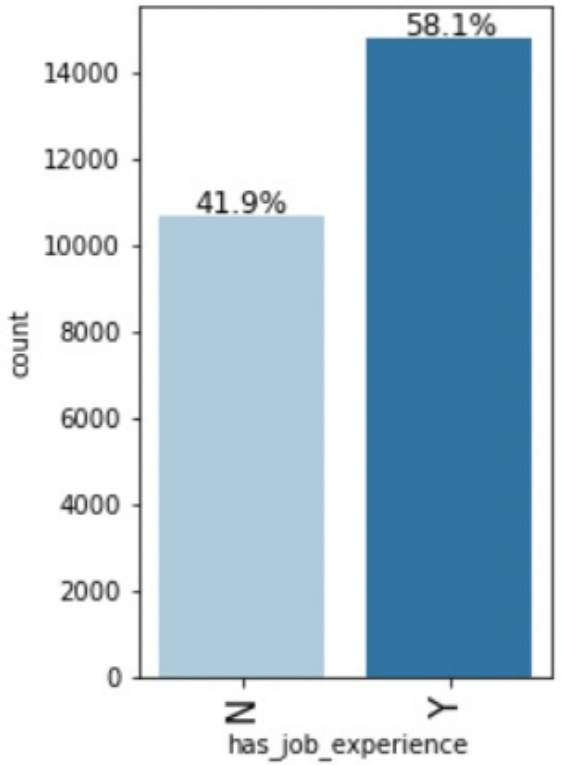
Most employees are from the continent of Asia with the least from Oceania.

Barplot (Education of Employee)



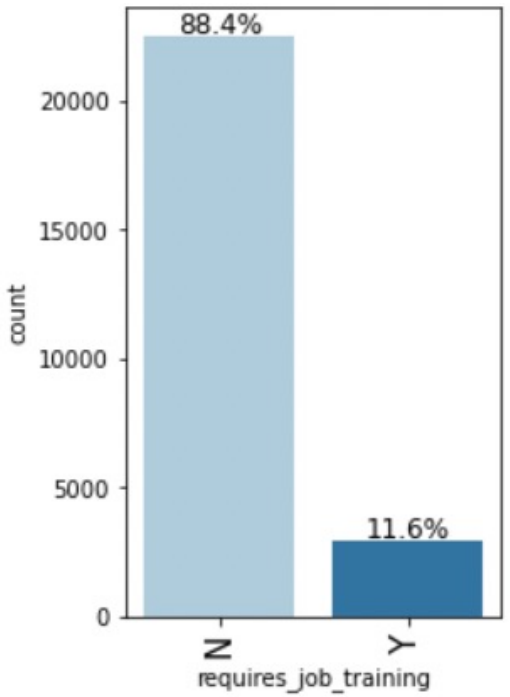
Most employees have a bachelor's degree, while the least employees have a doctorate degree.

Barplot (Job Experience)



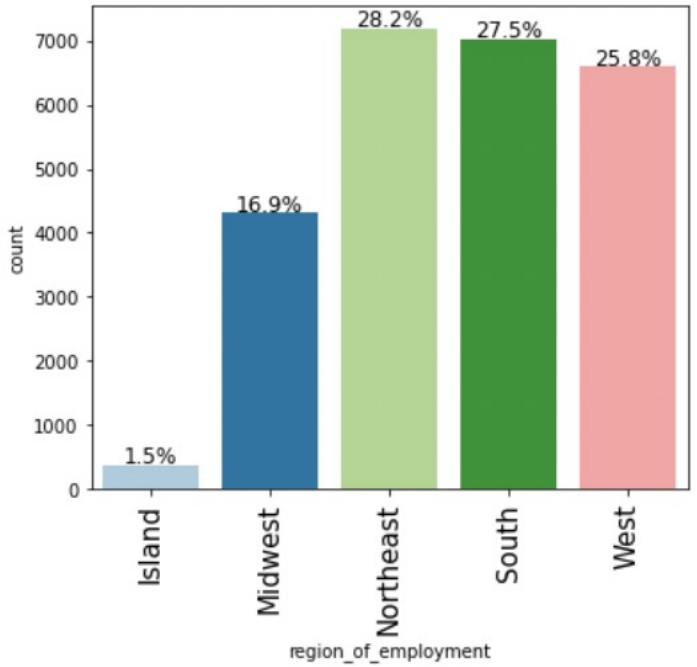
More employees have job experience than employees without job experience.

Barplot (Job Training)



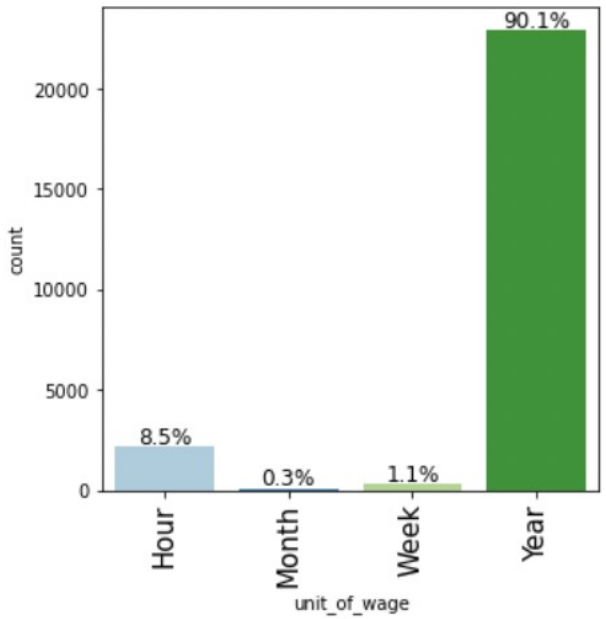
More employees don’t require job training than employees that do require job training.

Barplot (Region of Employment)



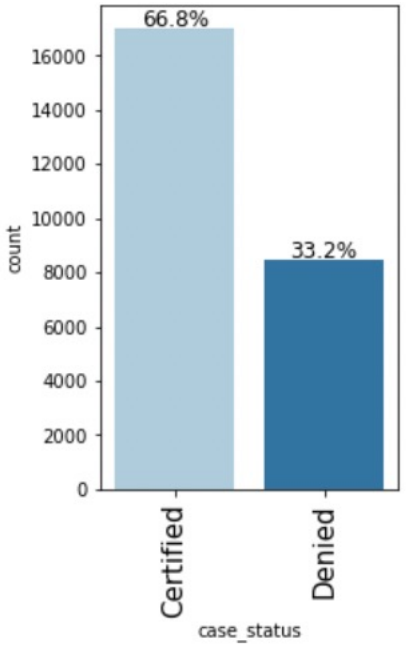
Most foreign employees intended region of employment in the US is in the Northeast and the least intended region of employment is an island.

Barplot (Unit of Wage)



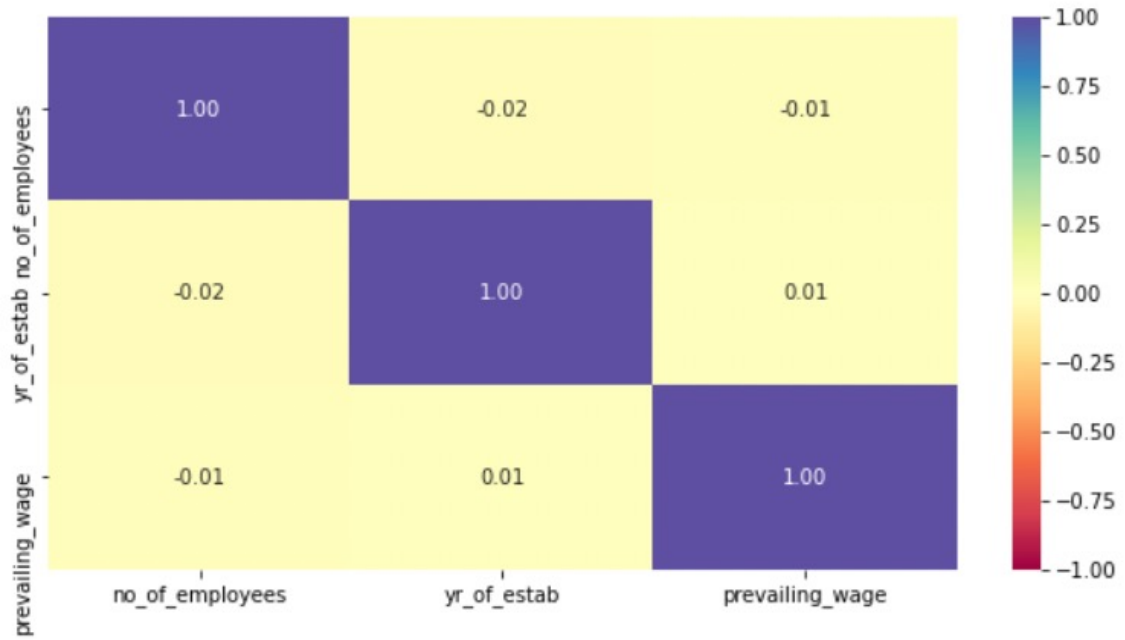
The majority of employees' unit of prevailing wage is on a yearly basis and the least on a monthly basis.

Barplot (Case Status)



More employees have
a certified visa than a denied visa.

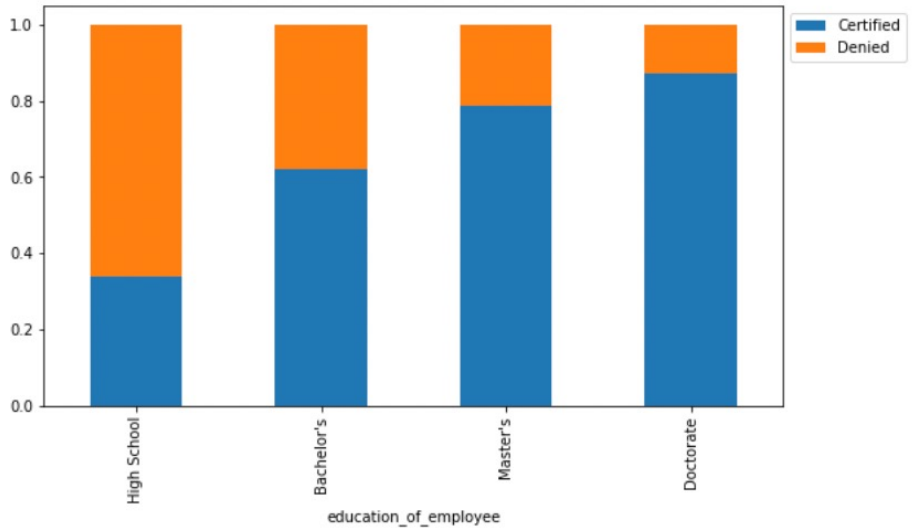
Correlation Heatmap



There are significantly weak correlations between all the variables: prevailing wage, year of establishment, and number of employees.

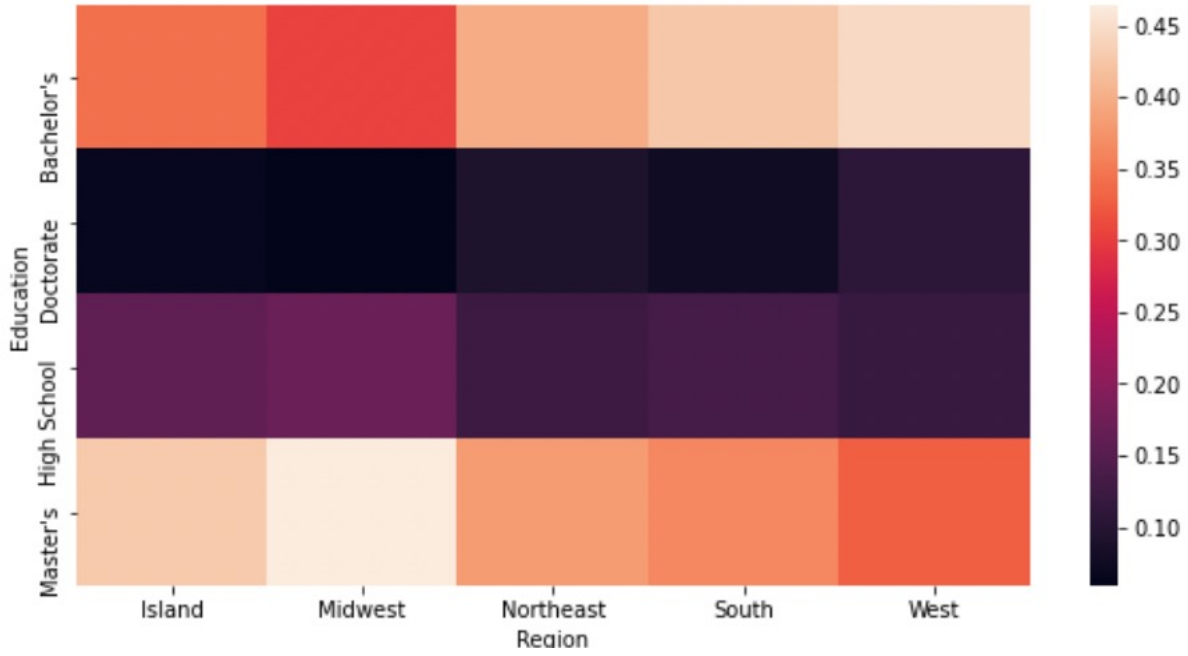
Stacked Barplot (Education of Employee vs. Case Status)

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



The higher the education one receives, the likelier the visa status will be certified.

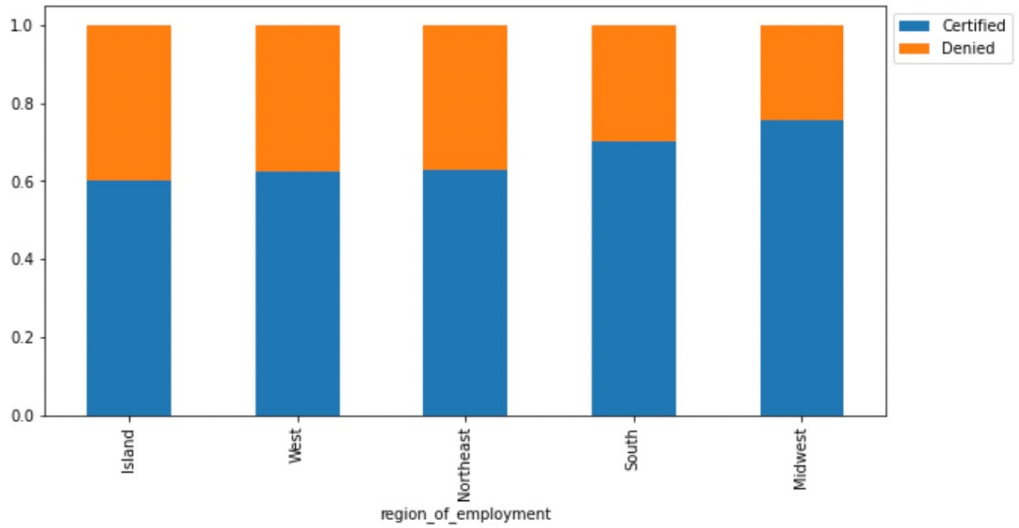
Heatmap (Education of employee vs. Region of Employment)



Having a bachelor's or master's degree has a stronger influence of living in a region of employment compared to doctorate and high school degrees.
A bachelor's degree prefers to be employed in the West Coast, while a master's degree prefers to be employed in the Midwest.

Stacked Barplot (Region of Employment vs. Case Status)

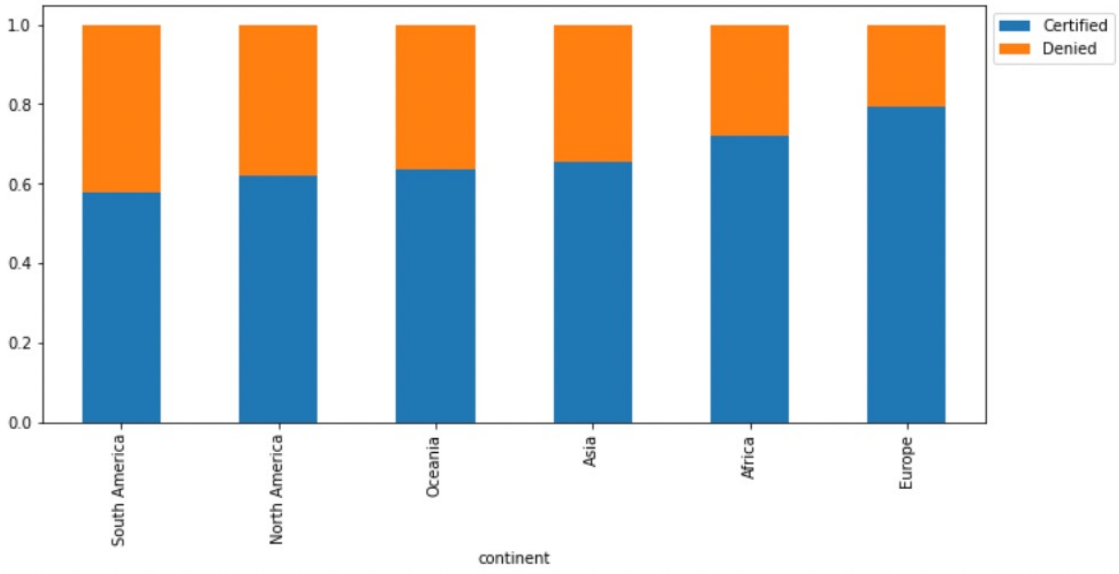
case_status	Certified	Denied	All
region_of_employment			
All	17018	8462	25480
Northeast	4526	2669	7195
West	4100	2486	6586
South	4913	2104	7017
Midwest	3253	1054	4307
Island	226	149	375



Most employees whose intended region of employment is in the Northeast have their visa certified and denied, with the least amount certified and denied on an island. The proportion of visa denial is greatest for the island and least for the Midwest. The proportion of visa certification is greatest for the Midwest and least for the island.

Stacked Barplot (Continent vs. Case Status)

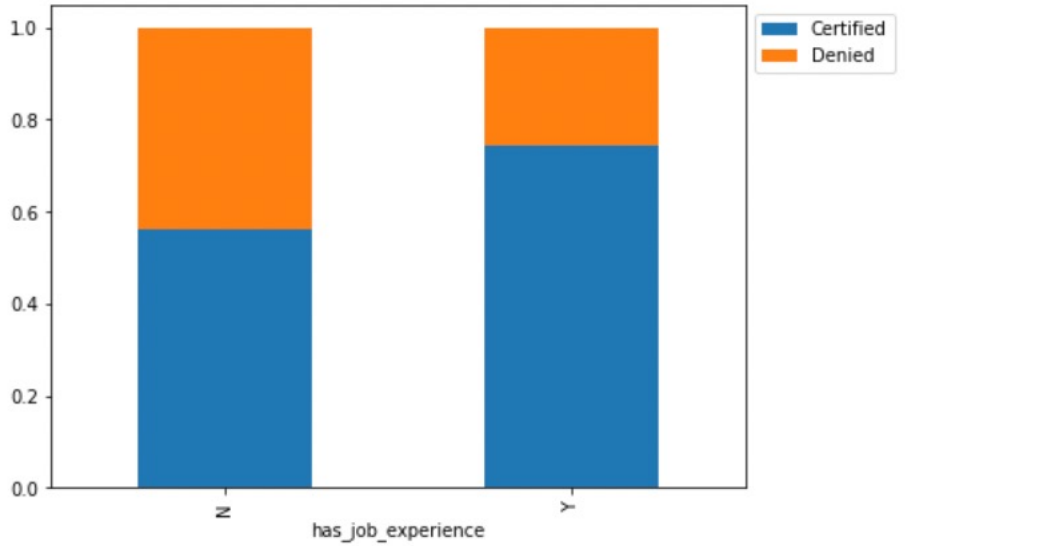
case_status	Certified	Denied	All
continent			
All	17018	8462	25480
Asia	11012	5849	16861
North America	2037	1255	3292
Europe	2957	775	3732
South America	493	359	852
Africa	397	154	551
Oceania	122	70	192



Most employees from the continent of Asia had their visas certified and denied, while the least employees from the continent of Oceania had their visas certified and denied. Proportion of denial is greatest in South America and least in Europe. Proportion of certified is greatest in Europe and least in South America.

Stacked Barplot (Job Experience vs. Case Status)

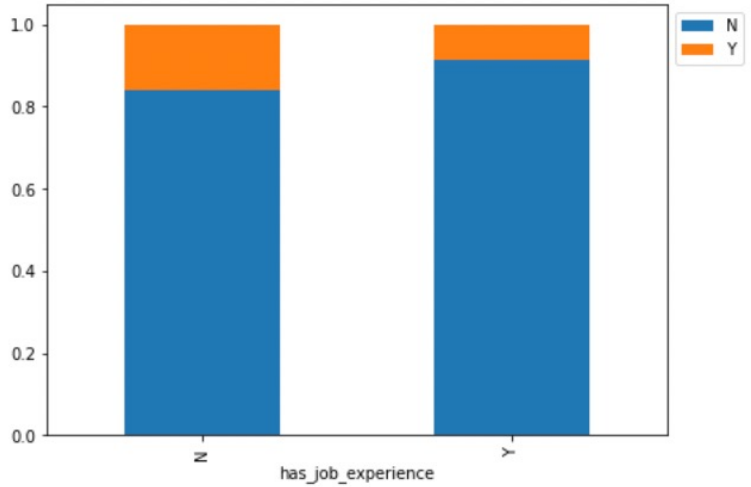
case_status	Certified	Denied	All
has_job_experience			
All	17018	8462	25480
N	5994	4684	10678
Y	11024	3778	14802



More people with job experience have a certified visa compared to people without job experience. More people without job experience have a denied visa than people with job experience.

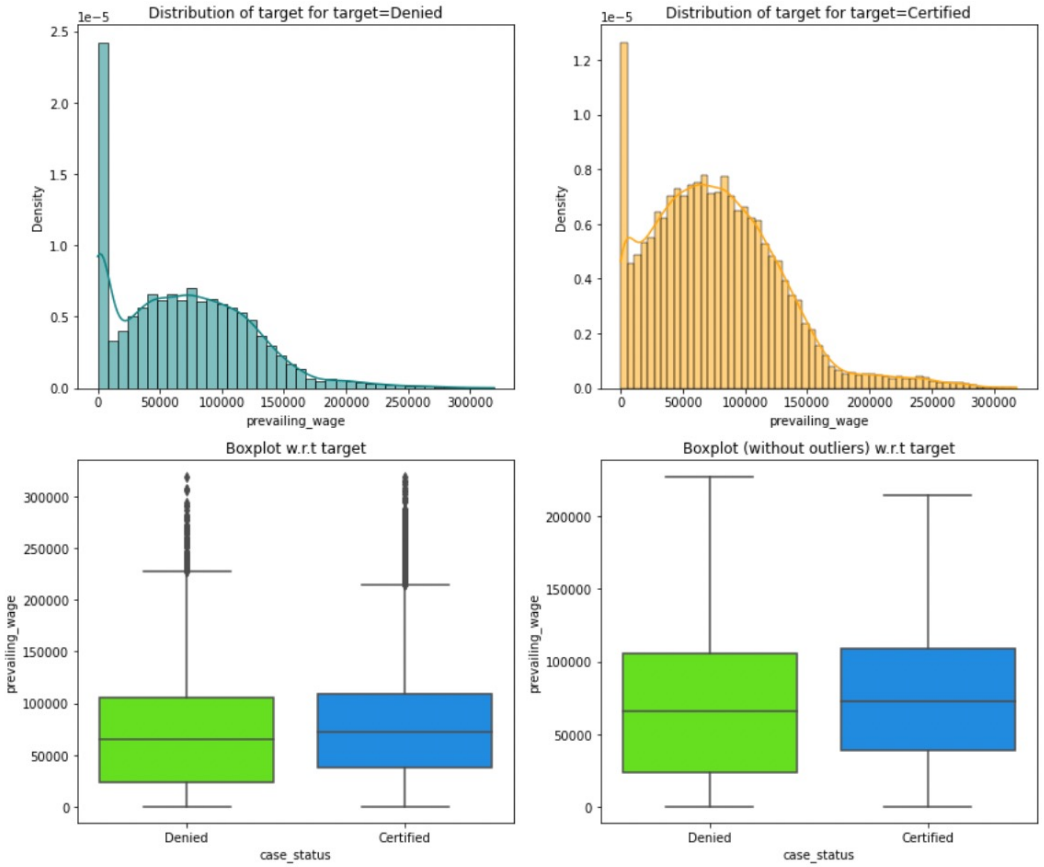
Stacked Barplot (Job Experience vs. Job Training)

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



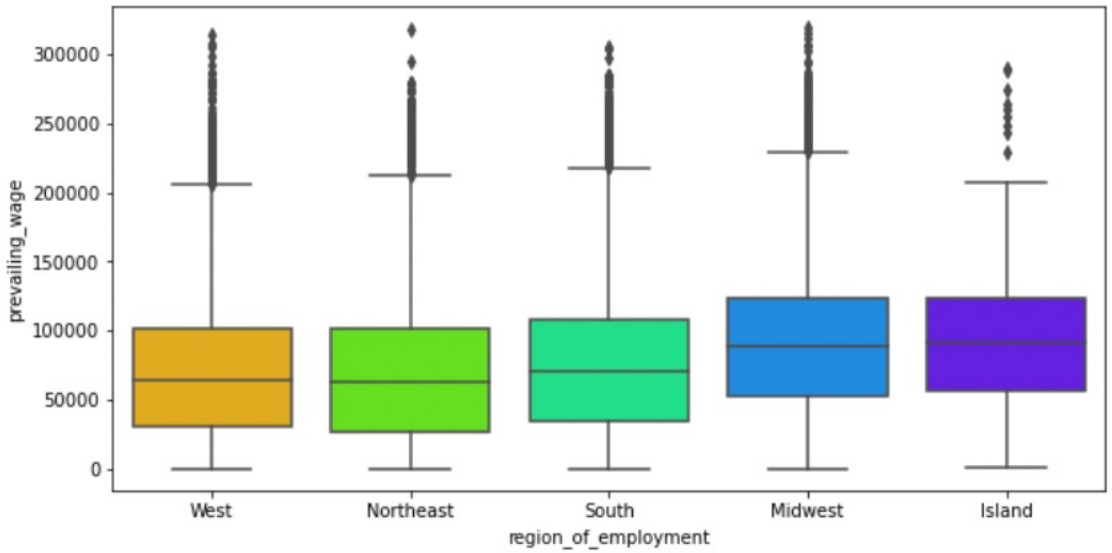
More people who have job experience don't need job training than people without job experience.

Distribution Plot (Prevailing Wage vs. Case Status)



Right-skewed distribution with and without outliers.
Average prevailing wage is greater for certified visa than denied visa.
IQR Range is greater for denied visa status than certified status.

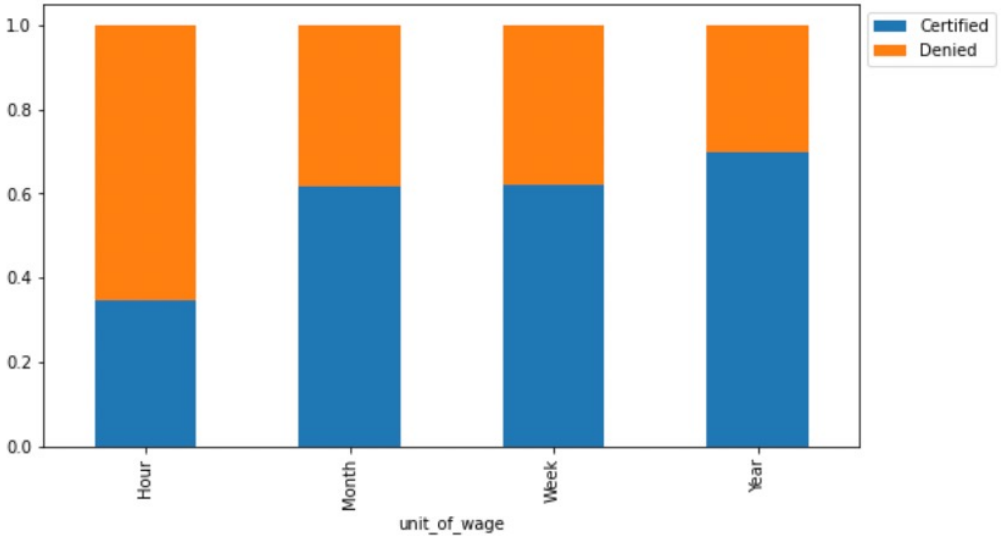
Boxplot (Region of Employment vs. Prevailing Wage)



Highest average prevailing wage is for employees' intended region in the Island and least in the Northeast. There are many outliers in the West, Northeast, South, and Midwest compared to the Island in region of employment.

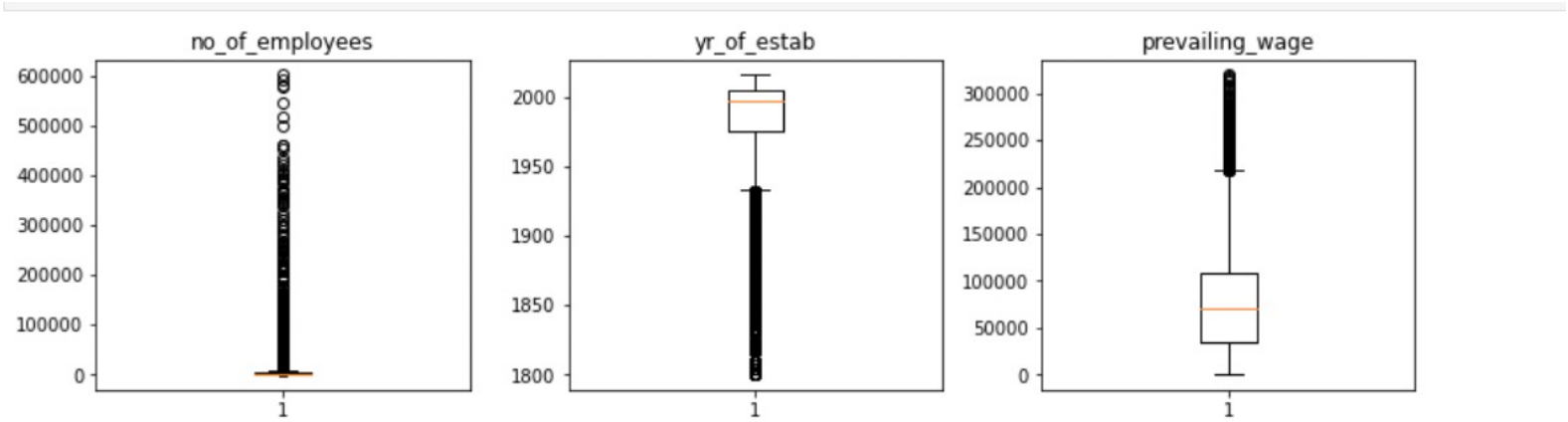
Stacked Barplot (Unit of Wage vs. Case Status

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



Most employees whose unit of wage is yearly have their visa status certified and least for hourly. Most employees whose unit of wage if hourly have their visa denied and least for visa denial is yearly. Greatest proportion of denial is hourly and least is yearly. Greatest proportion of certified is yearly and least is hourly.

Outlier Check

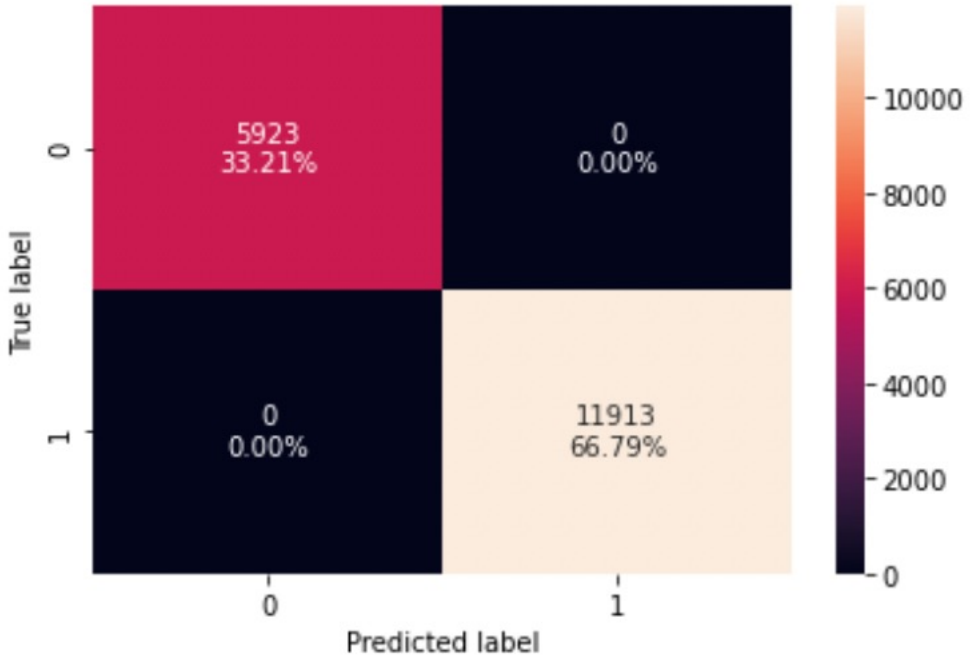


Many outliers to the right of number of employees and prevailing wage and outliers to the left of year of establishment.

Data Preparation for Modeling

```
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
```

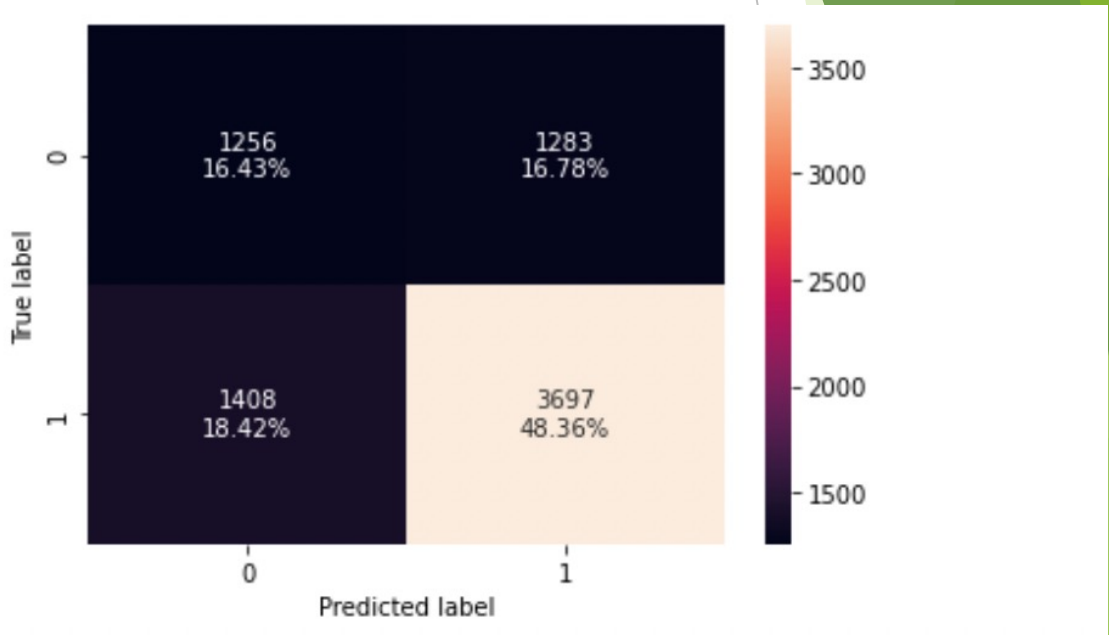
Decision Tree Model Evaluation



	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Training

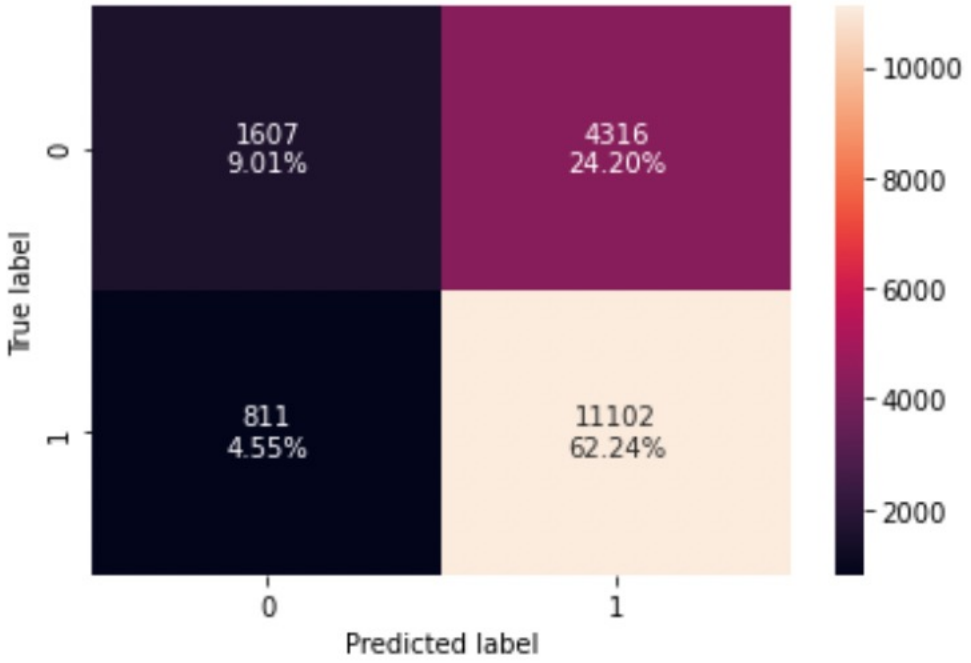
Decision tree is working well and overfitting on training data, but not able to generalize well concerning test data due to recall.



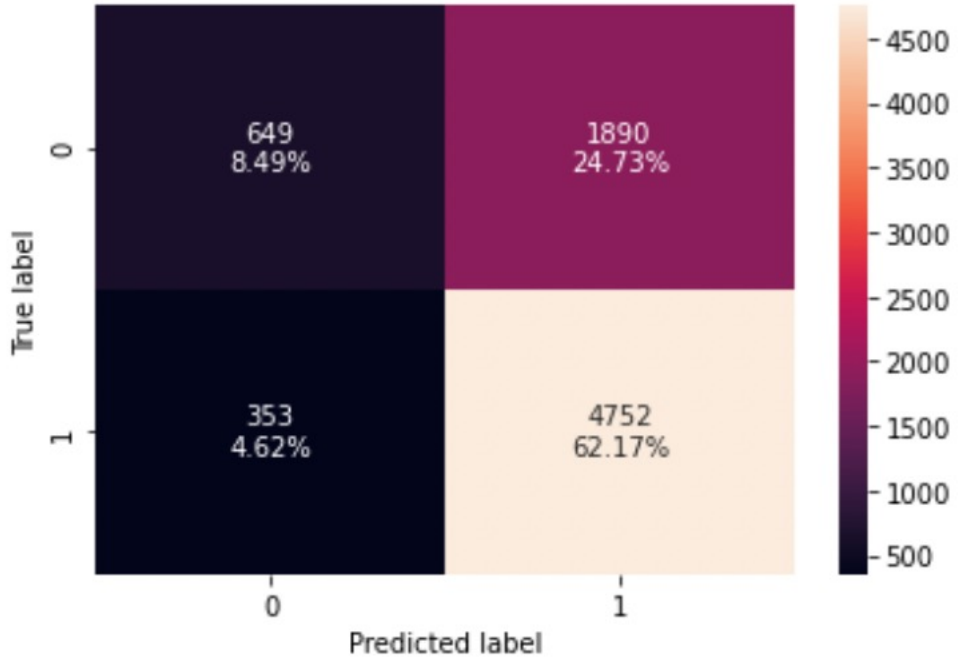
	Accuracy	Recall	Precision	F1
0	0.647959	0.724192	0.742369	0.733168

Test

Decision Tree Model Hyperparameter Tuning



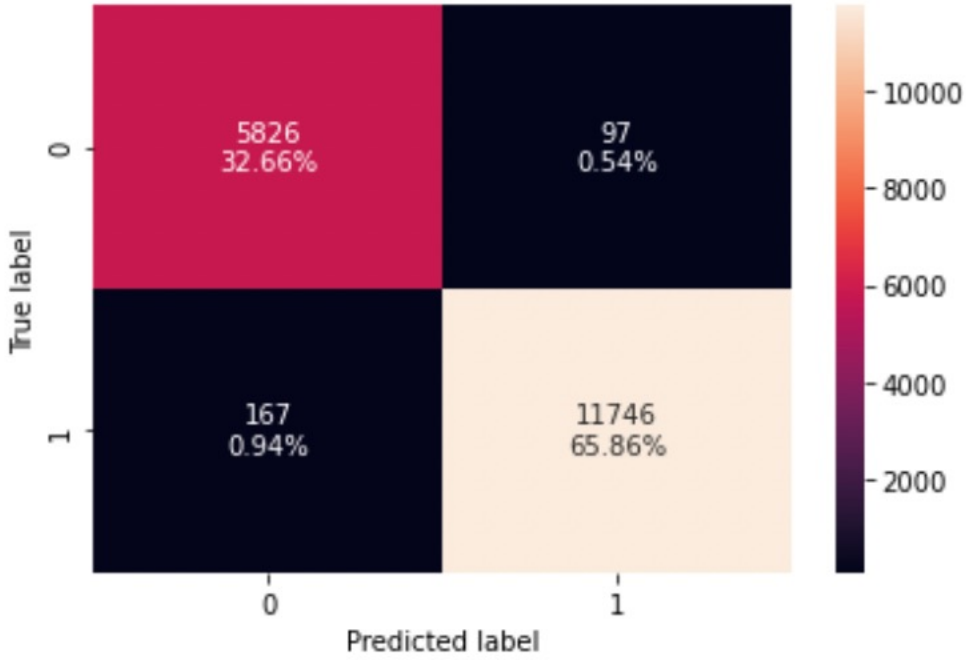
	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0



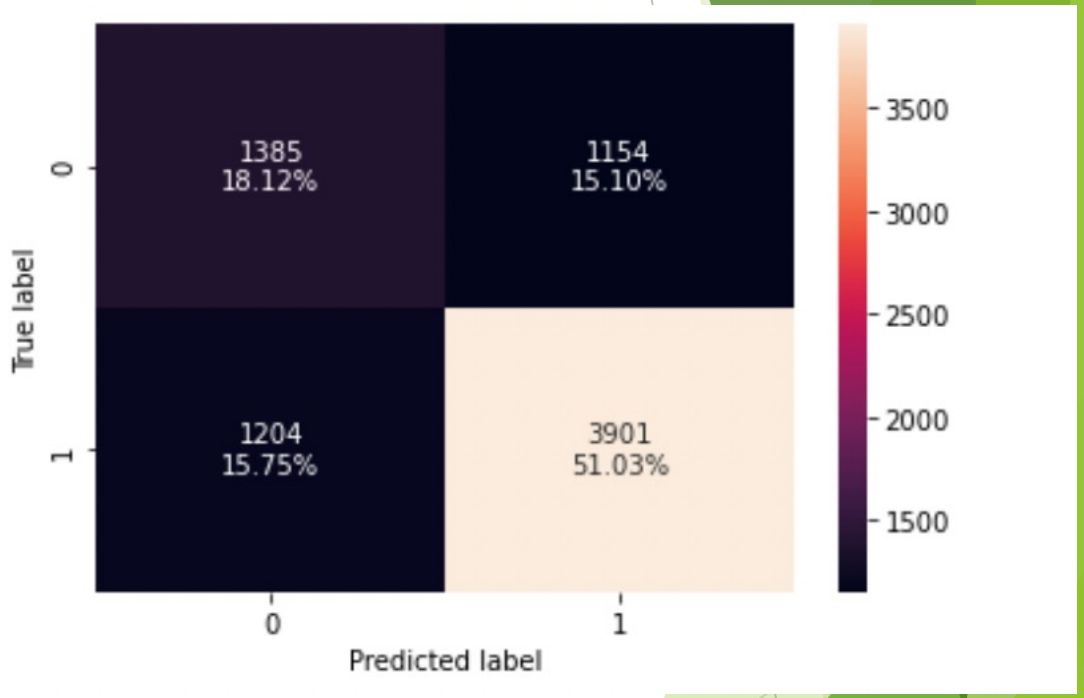
	Accuracy	Recall	Precision	F1
0	0.647959	0.724192	0.742369	0.733168

Results are the same with hyperparameter tuning.

Bagging Classifier Evaluation



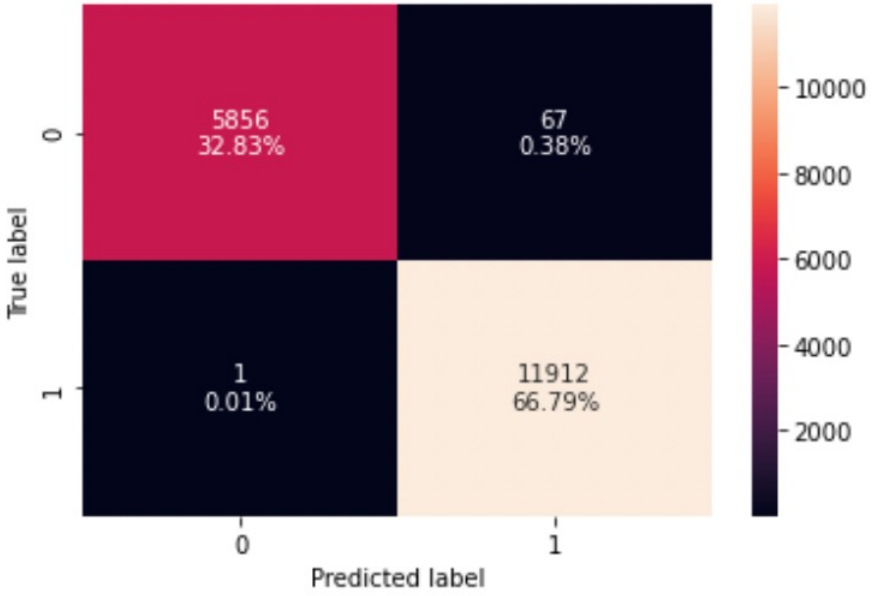
	Accuracy	Recall	Precision	F1
0	0.985198	0.985982	0.99181	0.988887



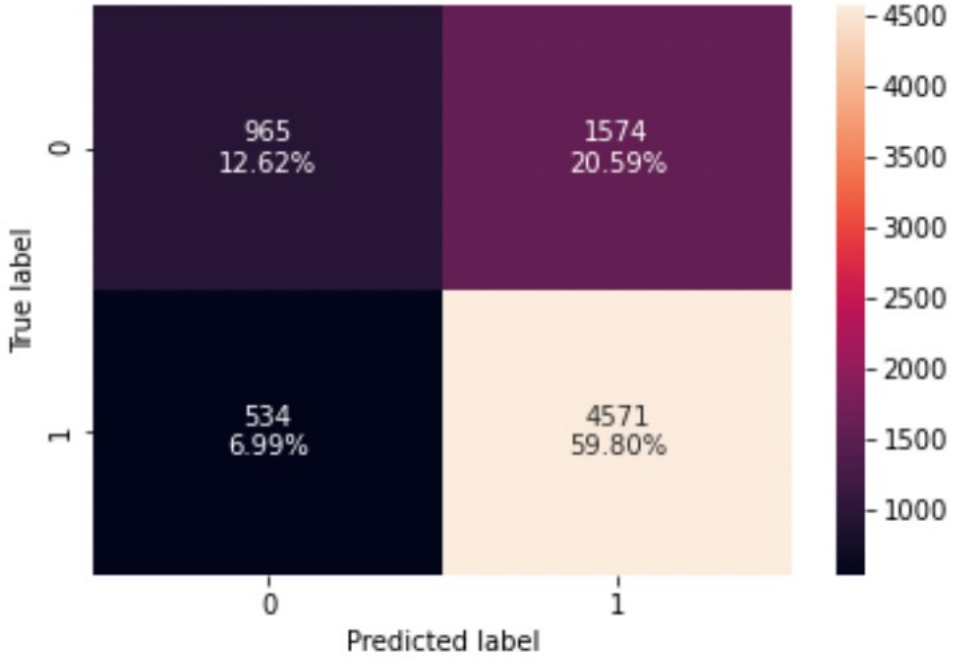
	Accuracy	Recall	Precision	F1
0	0.691523	0.764153	0.771711	0.767913

Bagging classifier is overfitting on training set and has a slightly better recall than decision trees, yet disappointing performance on test set.

Bagging Classifier Hyperparameter Tuning



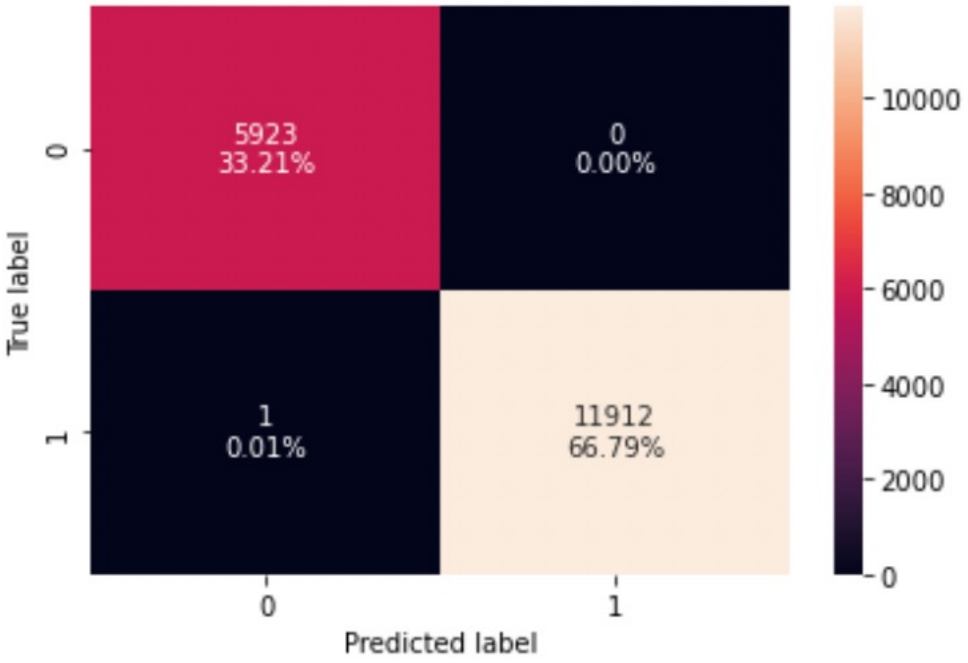
	Accuracy	Recall	Precision	F1
0	0.996187	0.999916	0.994407	0.997154



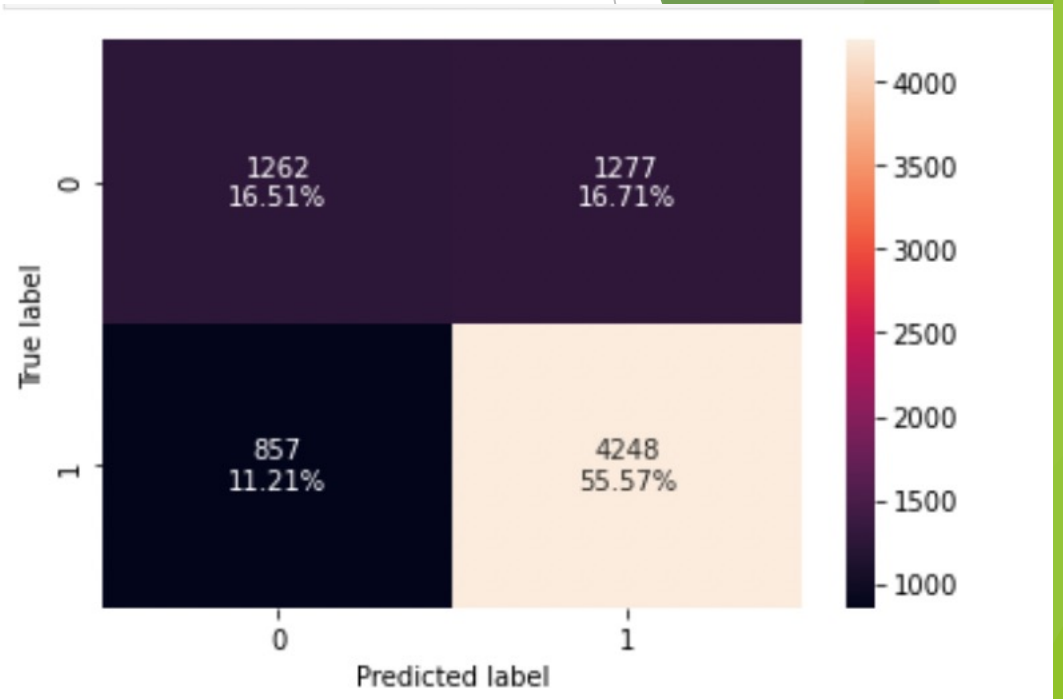
	Accuracy	Recall	Precision	F1
0	0.724228	0.895397	0.743857	0.812622

With hyperparameter tuning, accuracy has slightly improved with recall showing a significant improvement in the test set with bagging classifier giving generalized performance.

Random Forest Evaluation



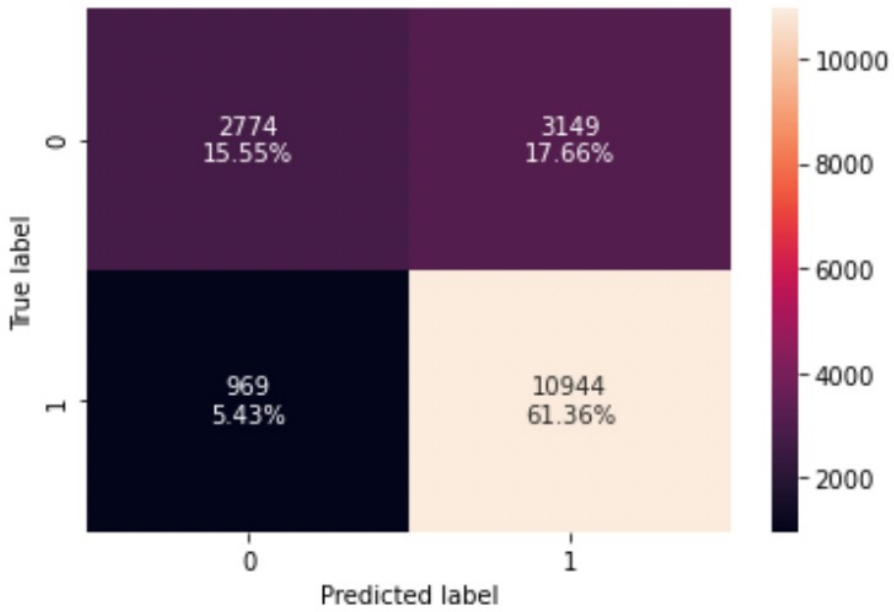
	Accuracy	Recall	Precision	F1
0	0.999944	0.999916	1.0	0.999958



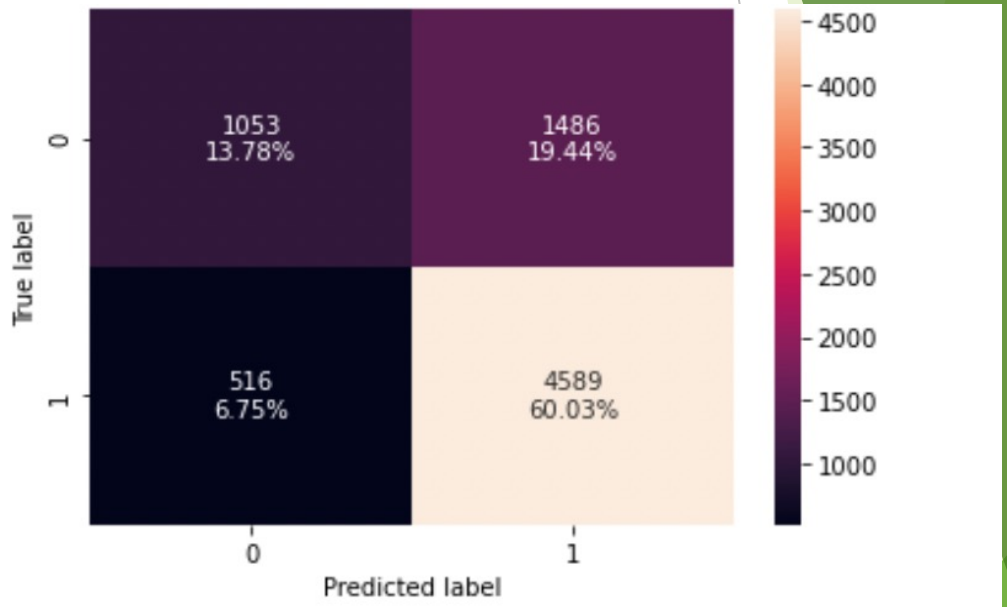
	Accuracy	Recall	Precision	F1
0	0.720827	0.832125	0.768869	0.799247

Random forest doesn’t have the best accuracy and precision in test set compared to training set, recall needs to be improved to generalize well on test set data.

Random Forest Hyperparameter Tuning



	Accuracy	Recall	Precision	F1
0	0.769119	0.91866	0.776556	0.841652

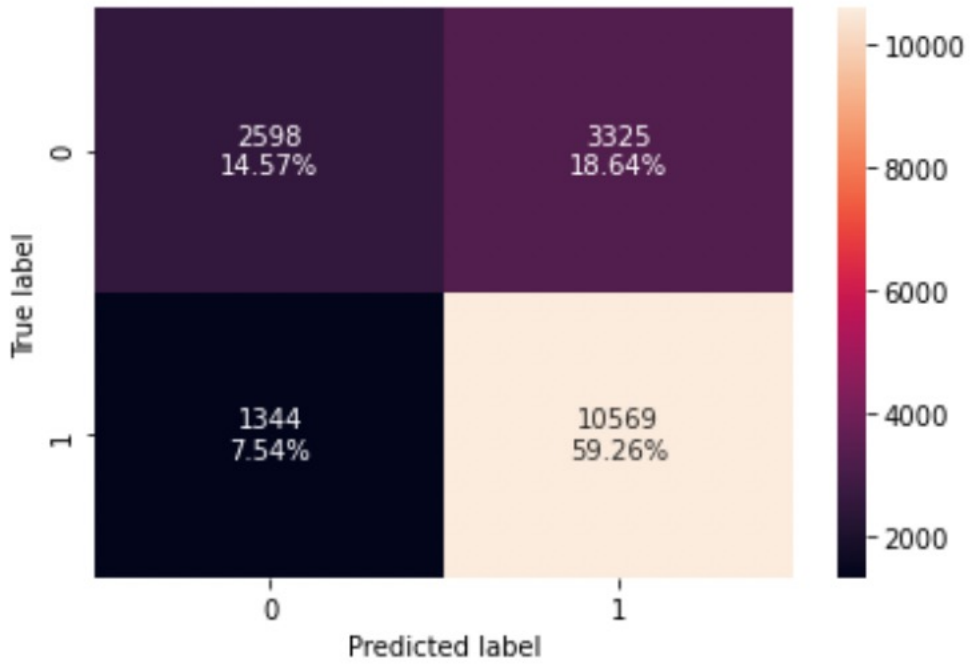


	Accuracy	Recall	Precision	F1
0	0.738095	0.898923	0.755391	0.82093

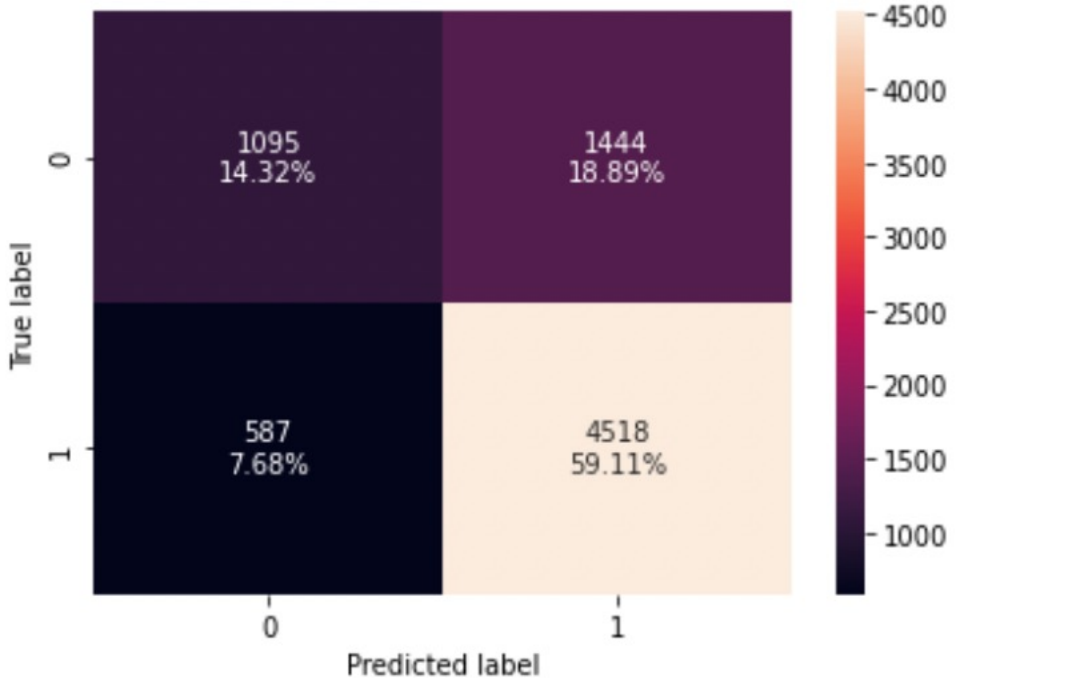
The test data recall has increased to generalize test data with accuracy and precision with slight changes. Training data accuracy and precision have dropped dramatically, suggesting errors in model.

AdaBoost Classifier Evaluation

Accuracy, recall, and precision for training and test data are nearly the same, however they can be improved.

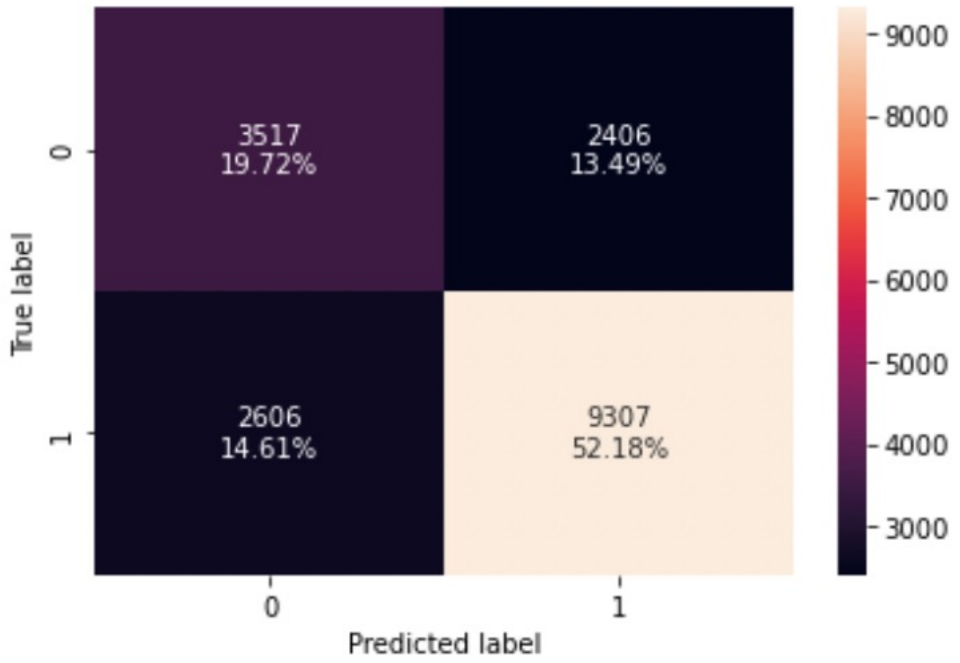


	Accuracy	Recall	Precision	F1
0	0.738226	0.887182	0.760688	0.81908

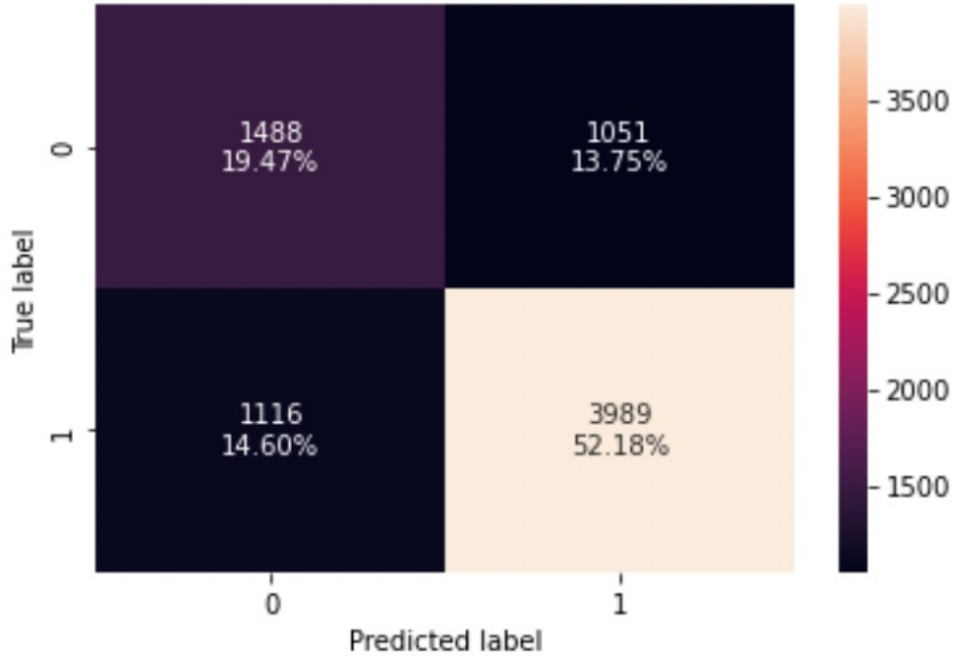


	Accuracy	Recall	Precision	F1
0	0.734301	0.885015	0.757799	0.816481

AdaBoost Classifier Hyperparameter Tuning



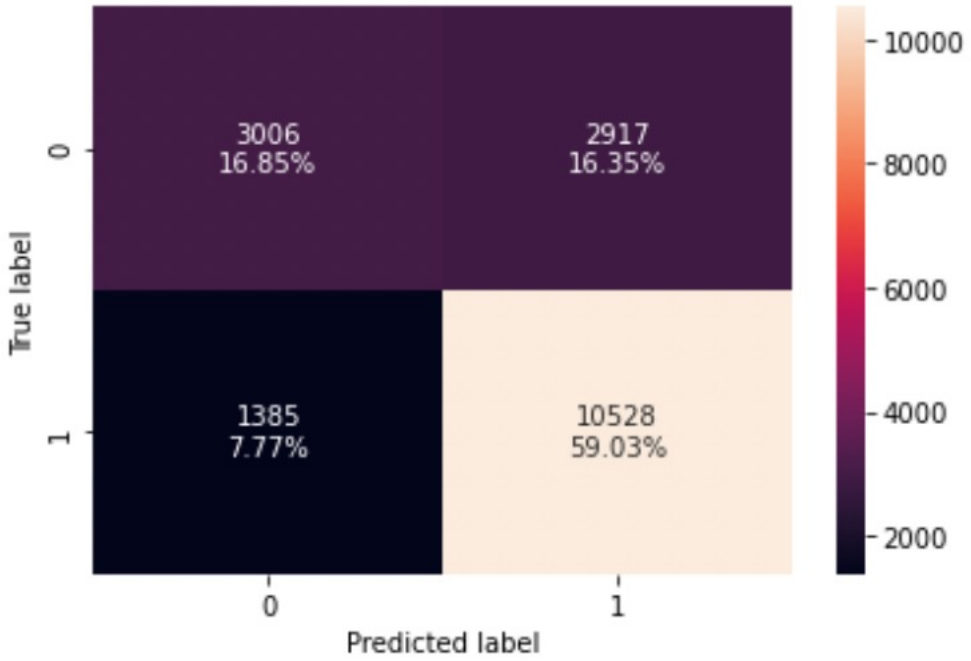
	Accuracy	Recall	Precision	F1
0	0.718995	0.781247	0.794587	0.787861

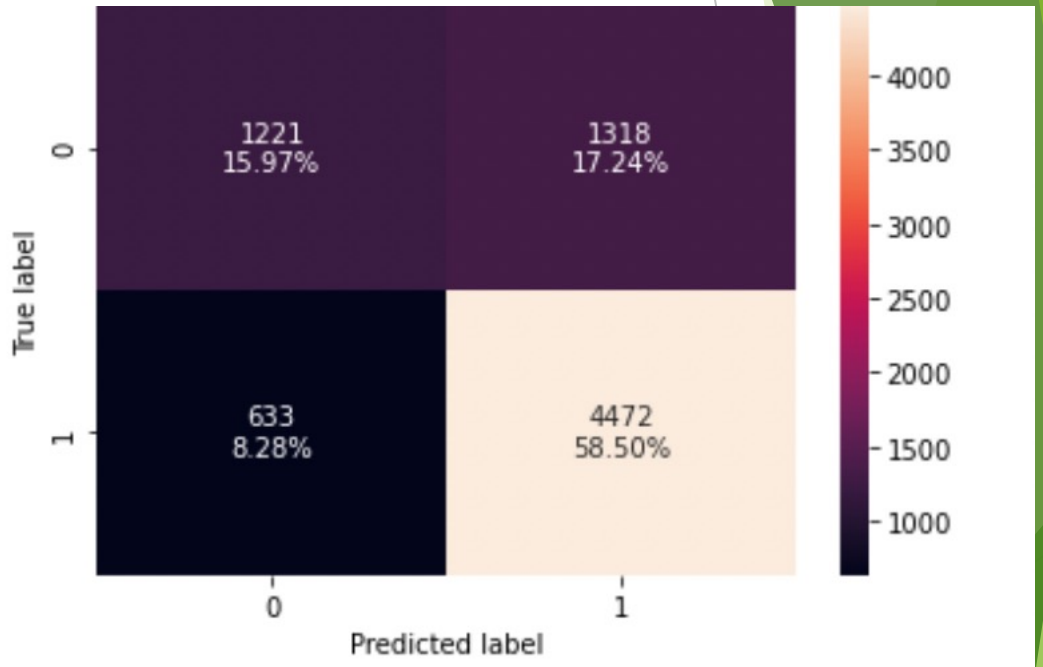


	Accuracy	Recall	Precision	F1
0	0.71651	0.781391	0.791468	0.786397

Training/test data accuracy and recall features have dropped and precision increased, but with lower test recall with tuning, the model won't be good at identifying defaulters.

Gradient Boosting Classifier Evaluation



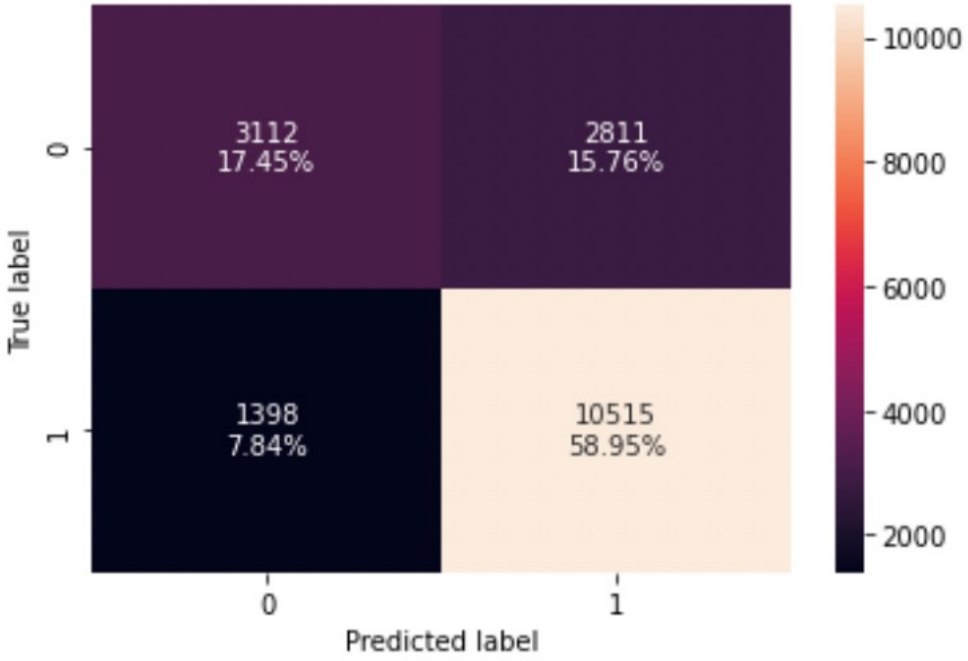


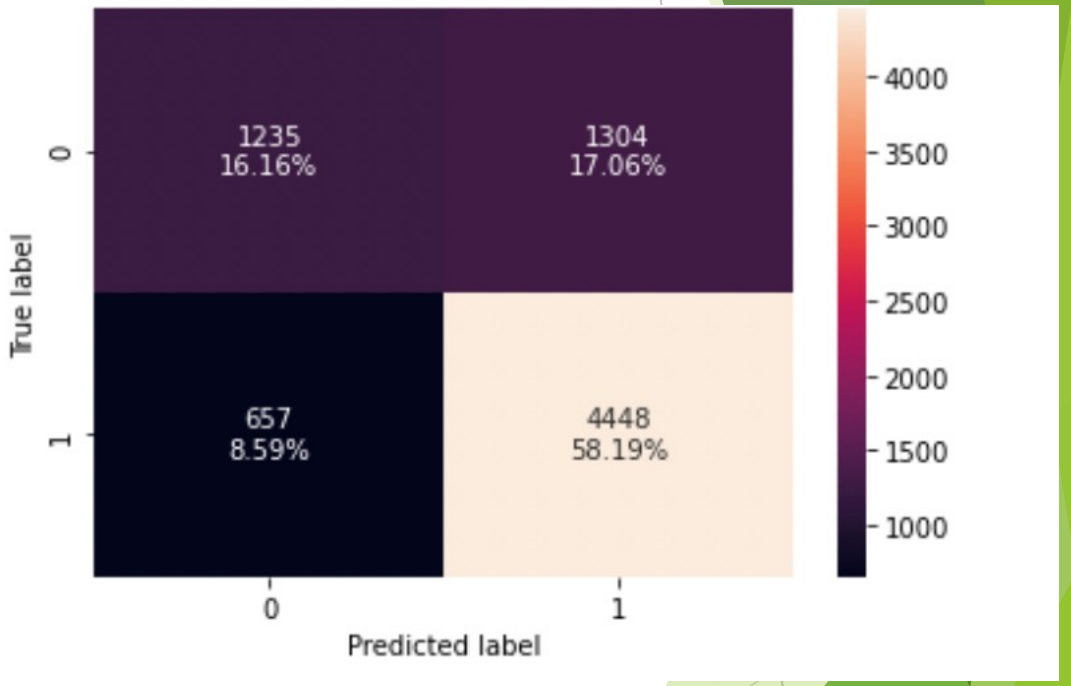
	Accuracy	Recall	Precision	F1
0	0.758802	0.88374	0.783042	0.830349

	Accuracy	Recall	Precision	F1
0	0.744767	0.876004	0.772366	0.820927

The training and test data have a pretty good recall to identify defaulters and non-defaulters, but they have a relatively lower accuracy and precision.

Gradient Boosting Classifier Hyperparameter Tuning



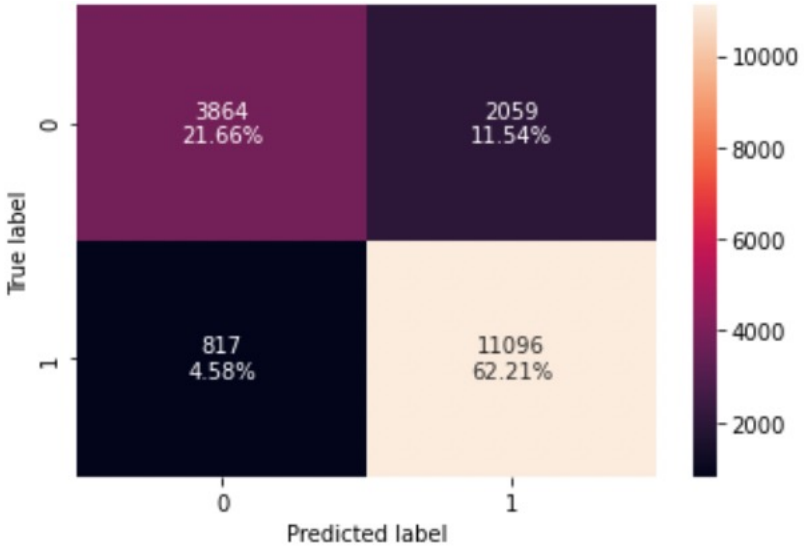


	Accuracy	Recall	Precision	F1
0	0.764017	0.882649	0.789059	0.833234

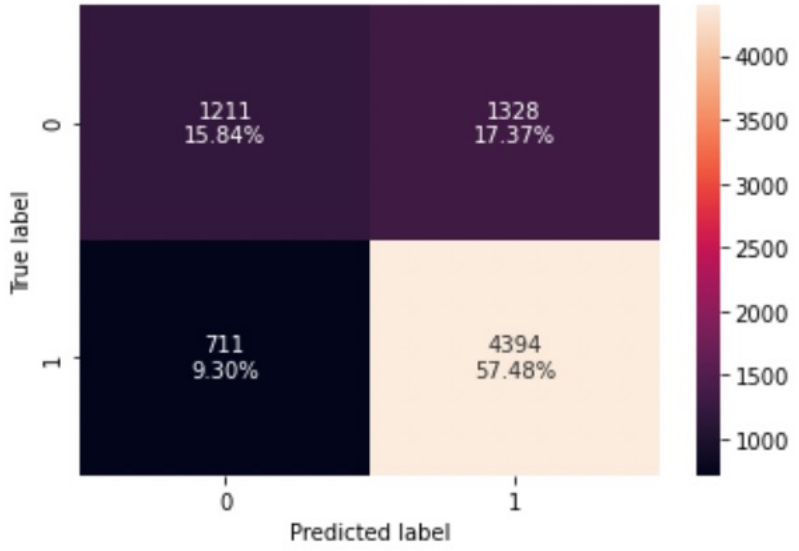
	Accuracy	Recall	Precision	F1
0	0.743459	0.871303	0.773296	0.819379

There wasn’t much significant change with tuning in test and training data.

XGBoost Classifier Evaulation



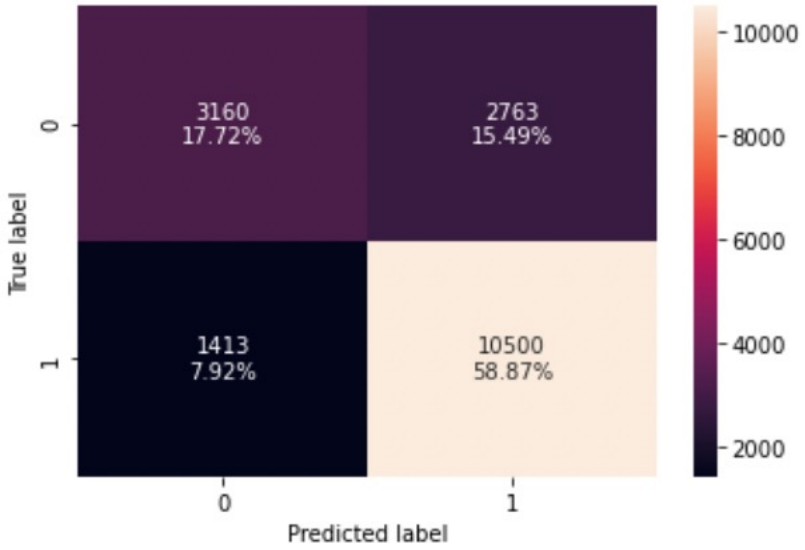
	Accuracy	Recall	Precision	F1
0	0.838753	0.931419	0.843482	0.885272



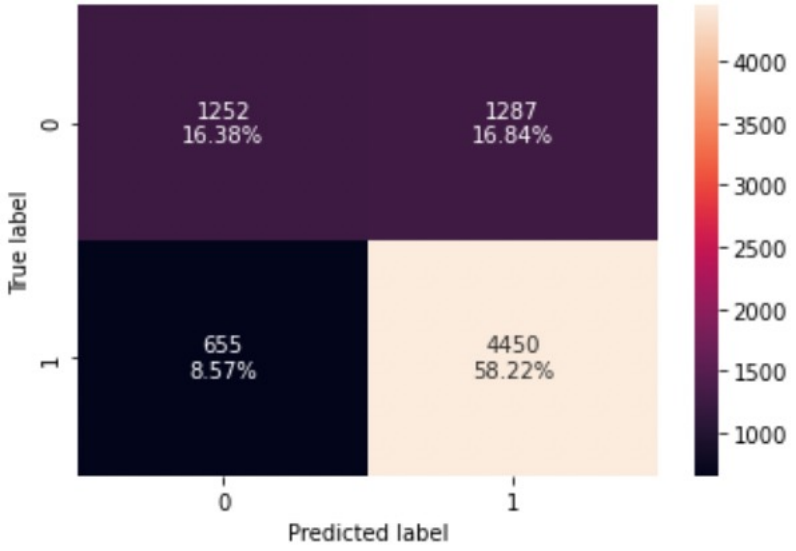
	Accuracy	Recall	Precision	F1
0	0.733255	0.860725	0.767913	0.811675

Accuracy and precision are considerably lower in test data than training data, but recall has higher recall in training data than test data due to overfitting, but one would preferably want higher recall in test data to identify defaulters.

XGBoost Classifier Hyperparameter Tuning



	Accuracy	Recall	Precision	F1
0	0.765867	0.88139	0.791676	0.834128



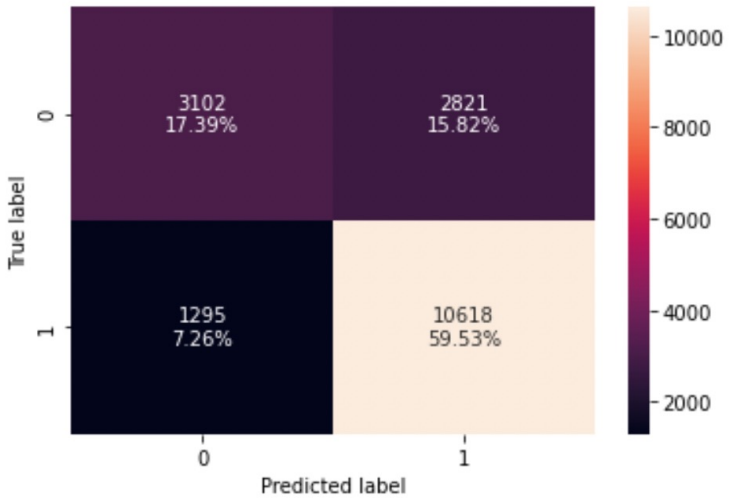
	Accuracy	Recall	Precision	F1
0	0.745945	0.871694	0.775667	0.820882

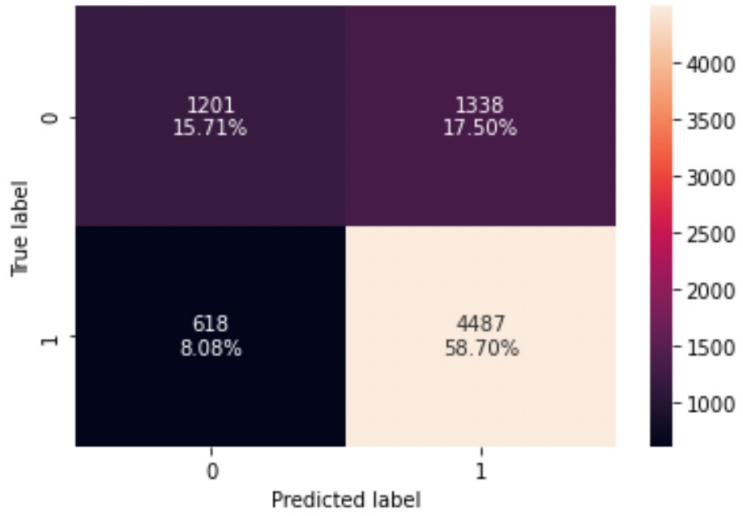
While accuracy, recall, and precision have decreased in training data, they have slightly increased in test data, giving a slightly better generalized performance.

Estimators for Stacking Classifier

```
StackingClassifier(estimators=[('AdaBoost', AdaBoostClassifier(random_state=1)),
                               ('Gradient Boosting',
                                GradientBoostingClassifier(init=AdaBoostClassifier(random_state=1),
                                                           max_features=0.8,
                                                           n_estimators=200,
                                                           random_state=1,
                                                           subsample=1)),
                               ('Random Forest',
                                RandomForestClassifier(max_depth=10,
                                                       max_features='sqrt',
                                                       min_samples_split=7,
                                                       n_estimators=20,
                                                       oob_score=Tru...
                                               eval_metric='logloss', gamma=5,
                                               gpu_id=-1,
                                               grow_policy='depthwise',
                                               importance_type=None,
                                               interaction_constraints='',
                                               learning_rate=0.1, max_bin=256,
                                               max_cat_to_onehot=4,
                                               max_delta_step=0, max_depth=6,
                                               max_leaves=0,
                                               min_child_weight=1,
                                               missing=nan,
                                               monotone_constraints='()',
                                               n_estimators=150, n_jobs=0,
                                               num_parallel_tree=1,
                                               predictor='auto',
                                               random_state=1, reg_alpha=0,
                                               reg_lambda=1, ...))
```

Stacking Classifier Evaluation





	Accuracy	Recall	Precision	F1
0	0.769231	0.891295	0.790089	0.837646

	Accuracy	Recall	Precision	F1
0	0.744113	0.878942	0.7703	0.821043

Training data has better accuracy, recall, and precision than test data, but there isn't much significant difference between them.

Training vs. Test Model Comparisons

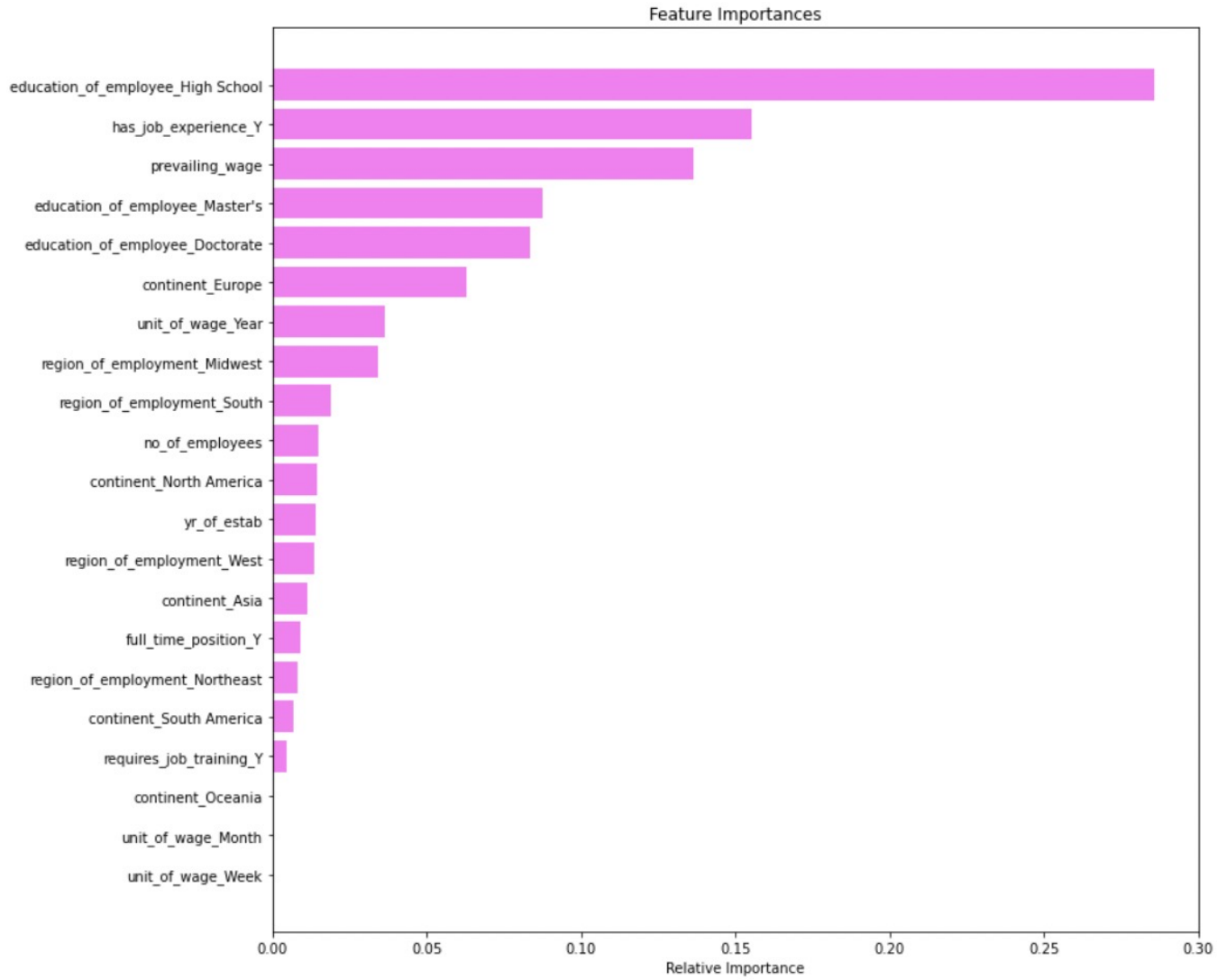
Training performance comparison:

	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	0.647959	0.691523	0.996187	0.999944	0.769119	0.738226	0.718995	0.758802	0.764017	0.838753	0.765867	0.769231
Recall	1.0	0.724192	0.764153	0.999916	0.999916	0.918660	0.887182	0.781247	0.883740	0.882649	0.931419	0.881390	0.891295
Precision	1.0	0.742369	0.771711	0.994407	1.000000	0.776556	0.760688	0.794587	0.783042	0.789059	0.843482	0.791676	0.790089
F1	1.0	0.733168	0.767913	0.997154	0.999958	0.841652	0.819080	0.787861	0.830349	0.833234	0.885272	0.834128	0.837646

Testing performance comparison:

	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	0.647959	0.647959	0.691523	0.724228	0.720827	0.738095	0.734301	0.716510	0.744767	0.743459	0.733255	0.745945	0.744113
Recall	0.724192	0.724192	0.764153	0.895397	0.832125	0.898923	0.885015	0.781391	0.876004	0.871303	0.860725	0.871694	0.878942
Precision	0.742369	0.742369	0.771711	0.743857	0.768869	0.755391	0.757799	0.791468	0.772366	0.773296	0.767913	0.775667	0.770300
F1	0.733168	0.733168	0.767913	0.812622	0.799247	0.820930	0.816481	0.786397	0.820927	0.819379	0.811675	0.820882	0.821043

Feature Importance



Education of employee: high school is most important feature, followed by having job experience and prevailing wage.

Conclusions

- Factors such as a higher education, intended region of employment in the Midwest, arriving from the continent of Europe, and yearly unit of wage would likely result in a certified visa status.
- High school education, having job experience, and prevailing wage are the most crucial factors in determining visa status.
- In hyperparameter tuning, random forests have the best recall, accuracy, and precision to give generalized performance of test data and identify defaulters, followed closely by bagged classifiers, which don't have much of a significant difference in these characteristics.
- A fair amount of training data models have overfitting due to high recall over 0.9.
- When analyzing boosters with tuning, a stacking classifier would provide the best recall in generalizing test performance data followed very closely by XGBoost Classifier and Gradient Boost Classifier respectively.

Recommendations

- Have job qualification tests/interviews for employees to see how much they can describe about their skillset and how their education and past job experience would be useful in determining visa status.
- Filter out applicants by other factors such as earning potential, field of education in university/college, and experience in job to get best field of qualified applicants for visa status approval.
- Assess the potential differences in wages employees made in native country/continent and how much they would make in the United States.