EasyVisa Project

Ensemble Techniques

August 23, 2022

Sunny Amirize, MBA

Contents / Agenda

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

Executive Summary

To analyze the data provided to build a Machine Learning based solution that can help in shortlisting the candidates having higher chances of VISA approval - with the help of a classification model, the follow steps were followed;

- Sanity checks on the dataset/Data overview.
- Exploratory data analysis/EDA was done.
- Data preprocessing was done.
- Model building and performance checks.
- Model assumptions check.
- Calculate performance metrics and create confusion matrix for different models.
- Compare all models.
- Feature importance of XGBoost Hyperparameter Tuned Model.

After building the final model to get the decision tree results, the insights and recommendation are as follows;
For the model, we have;

- Education of employee - employee with a doctorate degree have 65% chance of getting visa certified, and employee with high school certification has over 65% of getting visa denied.
- Unit of wage - employee with non-hourly pay has 70% chance of getting visa certified, and employee with hourly pay has 65% of getting visa denied.
- Continent - employee with work experience and from Europe has 75% and 80% chance of getting visa certified compared to employee with no work experience have 50% chance of getting visa denied.
- Region of employment - employees from Midwest and South have 70% chances of getting visa certified.
- Attributes such as; full time or part time position, require job training, prevailing wage and year of establishment do not have much impact for visas to get certified or denied.
- We built a model that can capture over 80% of the information while making predictions, the findings can help build a suitable profile of candidates to facilitate the process of visa.

Business Problem Overview and Solution Approach

Problem Statement

Business communities in the United States are facing high demand for human resources, but one of the constant challenges is identifying and attracting the right talent, which is perhaps the most important element in remaining competitive. Companies in the United States look for hard-working, talented, and qualified individuals both locally as well as abroad.

The Immigration and Nationality Act (INA) of the US permits foreign workers to come to the United States to work on either a temporary or permanent basis. The act also protects US workers against adverse impacts on their wages or working conditions by ensuring US employers' compliance with statutory requirements when they hire foreign workers to fill workforce shortages. The immigration programs are administered by the Office of Foreign Labor Certification (OFLC).

OFLC processes job certification applications for employers seeking to bring foreign workers into the United States and grants certifications in those cases where employers can demonstrate that there are not sufficient US workers available to perform the work at wages that meet or exceed the wage paid for the occupation in the area of intended employment.

Solution approach/methodology

In FY 2016, the OFLC processed 775,979 employer applications for 1,699,957 positions for temporary and permanent labor certifications. This was a nine percent increase in the overall number of processed applications from the previous year. The process of reviewing every case is becoming a tedious task as the number of applicants is increasing every year.

- The increasing number of applicants every year calls for a Machine Learning based solution that can help in shortlisting the candidates having higher chances of VISA approval. The objective is to analyze the data provided and, with the help of a classification model:
- Facilitate the process of visa approvals.
- Recommend a suitable profile for the applicants for whom the visa should be certified or denied based on the drivers that significantly influence the case status.

EDA Results

- **Key results from EDA**
- Majority of the employees come from the the continent Asia.
- Majority of the employee have Bachelor's degree.
- About 58% have job experience and 42% have no job experience.
- More foreign workers intend applying for jobs in the Northeast region (28%) followed by South (28%) and West (26%) regions.
- 90% Unit prevailing wage falls under Yearly, 9% falls under Hourly.
- Employee with High School are most denied and employee with Doctorate degrees are most certified or offered visa.

- **Please mention answers to the insight-based questions provided**
- Education of employee - employee with a doctorate degree have 65% chance of getting visa certified, and employee with high school certification has over 65% of getting visa denied.
- Unit of wage - employee with non-hourly pay has 70% chance of getting visa certified, and employee with hourly pay has 65% of getting visa denied.
- Continent - employee with work experience and from Europe has 75% and 80% chance of getting visa certified compared to employee with no work experience have 50% chance of getting visa denied.
- Region of employment - employees from Midwest and South have 70% chances of getting visa certified.

We built a model that can capture over 80% of the information while making predictions, the findings can help build a suitable profile of candidates to facilitate the process of visa.

Link to Appendix slide on data background check

EDA Results

View the top 5 rows of the dataset

	0	1	2	3	4
case_id	EZYV01	EZYV02	EZYV03	EZYV04	EZYV05
continent	Asia	Asia	Asia	Asia	Africa
education_of_employee	High School	Master's	Bachelor's	Bachelor's	Master's
has_job_experience	N	Y	N	N	Y
requires_job_training	N	N	Y	N	N
no_of_employees	14513	2412	44444	98	1082
yr_of_estab	2007	2002	2008	1897	2005
region_of_employment	West	Northeast	West	West	South
prevailing_wage	592.2029	83425.65	122996.86	83434.03	149907.39
unit_of_wage	Hour	Year	Year	Year	Year
full_time_position	Y	Y	Y	Y	Y
case_status	Denied	Certified	Denied	Denied	Certified

View the last 5 rows of the dataset

	25475	25476	25477	25478	25479
case_id	EZYV25476	EZYV25477	EZYV25478	EZYV25479	EZYV25480
continent	Asia	Asia	Asia	Asia	Asia
education_of_employee	Bachelor's	High School	Master's	Master's	Bachelor's
has_job_experience	Y	Y	Y	Y	Y
requires_job_training	Y	N	N	Y	N
no_of_employees	2601	3274	1121	1918	3195
yr_of_estab	2008	2006	1910	1887	1960
region_of_employment	South	Northeast	South	West	Midwest
prevailing_wage	77092.57	279174.79	146298.85	86154.77	70876.91
unit_of_wage	Year	Year	Year	Year	Year
full_time_position	Y	Y	N	Y	Y
case_status	Certified	Certified	Certified	Certified	Certified

Link to Appendix slide on data background check

EDA Results

Checking the shape/dimension of the dataset.

The dataset has 25480 rows and 12 columns.

```
(25480, 12)
```

Link to Appendix slide on data background check

EDA Results

Checking the data types of the columns for the dataset.

There are 5 columns of the dtype object,
1 column of the dtype float64,
and 2 columns of the dtype int64.

```
<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
```

Link to Appendix slide on data background check

EDA Results

Checking for duplicate values.

```
False    25480
dtype: int64
```

Link to Appendix slide on data background check

EDA Results

Statistical summary of the data.

	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

Link to Appendix slide on data background check

EDA Results

Checking for negative values in the employee column.

```
(0,)
```

[Link to Appendix slide on data background check]

EDA Results

Checking the count of each unique category in each of the the 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
--------------------------------------------------
```

Link to Appendix slide on data background check

EDA Results

Checking unique values in the mentioned column.

```
array(['EZYV01', 'EZYV02', 'EZYV03', ..., 'EZYV25478', 'EZYV25479',
       'EZYV25480'], dtype=object)
```

Link to Appendix slide on data background check

EDA Results

Checking unique values in the mentioned column.

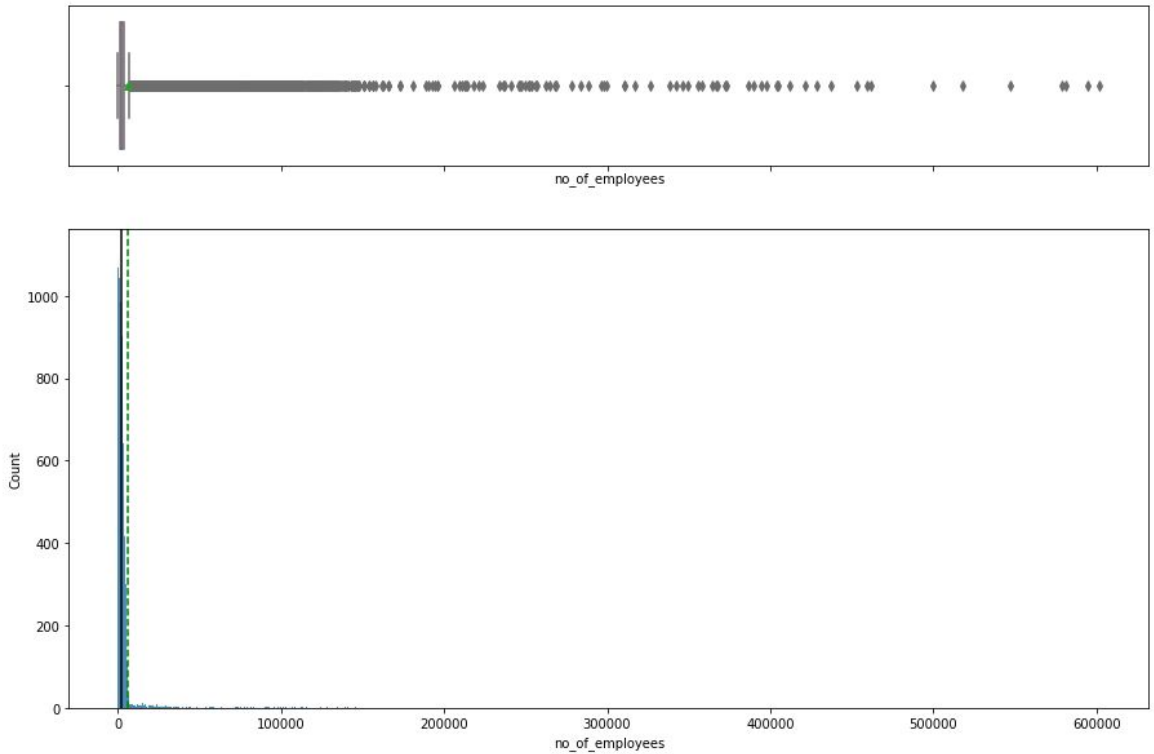
```
array(['EZYV01', 'EZYV02', 'EZYV03', ..., 'EZYV25478', 'EZYV25479',
       'EZYV25480'], dtype=object)
```

Link to Appendix slide on data background check

EDA Results
Univariate Analysis

Observations on number of employees

- *The distribution of the number of employees is right-skewed.*
- *The boxplot shows that there are lots of outliers to the right.*

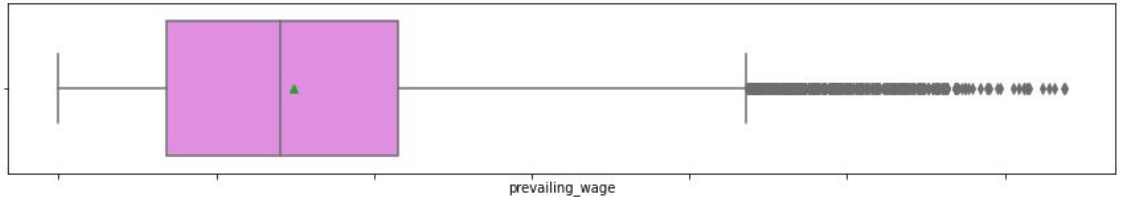


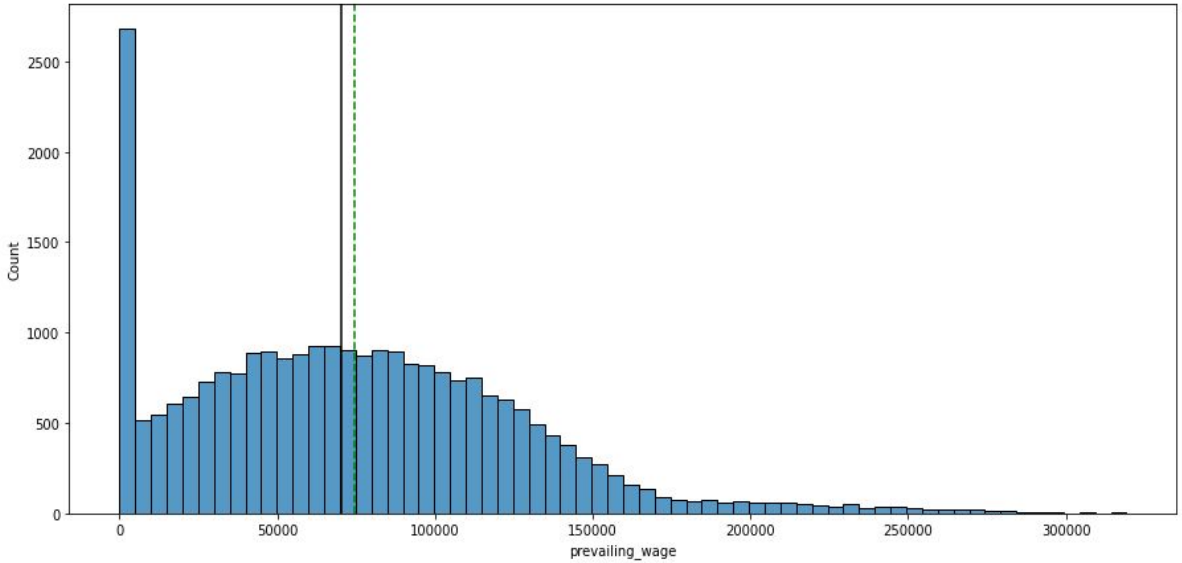
Link to Appendix slide on data background check

EDA Results
Univariate Analysis

Observations on prevailing wage

Visual analysis of both distributions shows
- right-skewed.
- the mean is around USD 70,000.
- outliers in the income bracket between USD 200,000 to USD 300,000





Link to Appendix slide on data background check

EDA Results

Rows with less than 100 prevailing wage

	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

176 rows × 11 columns

Link to Appendix slide on data background check

EDA Results

Count of the values in the mentioned column

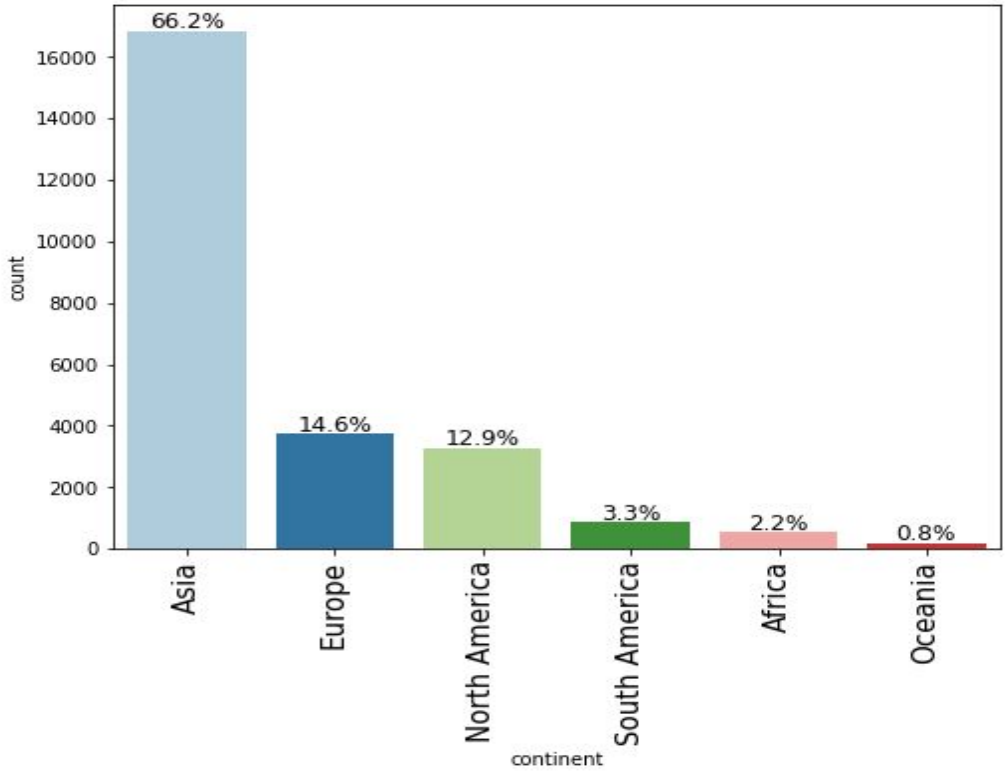
```
Hour    176
Name: unit_of_wage, dtype: int64
```

Link to Appendix slide on data background check

EDA Results
Univariate Analysis

Observations on continent

Visual analysis of bar plot shows
- Majority of the employees come from the the continent Asia.

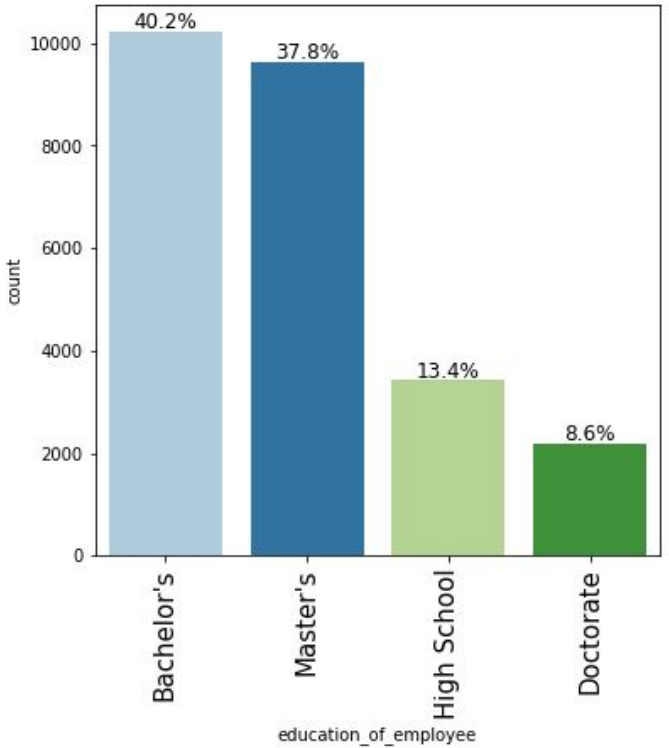


Link to Appendix slide on data background check

EDA Results
Univariate Analysis

Observations on education of employee

Visual analysis of bar plot shows
- Majority of the employee have Bachelor's degree.

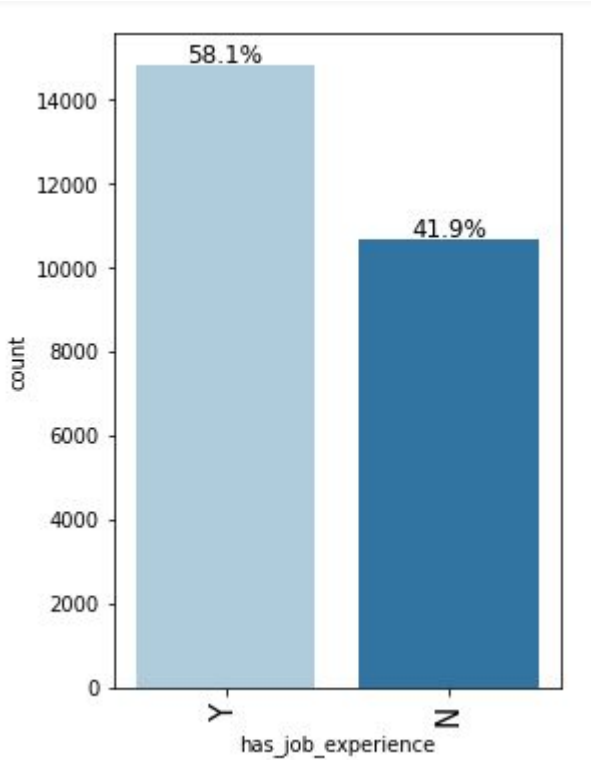


Link to Appendix slide on data background check

EDA Results
Univariate Analysis

Observations on job experience

Visual analysis of the bar plot shows
- About 58% have job experience and 42% have No job experience.

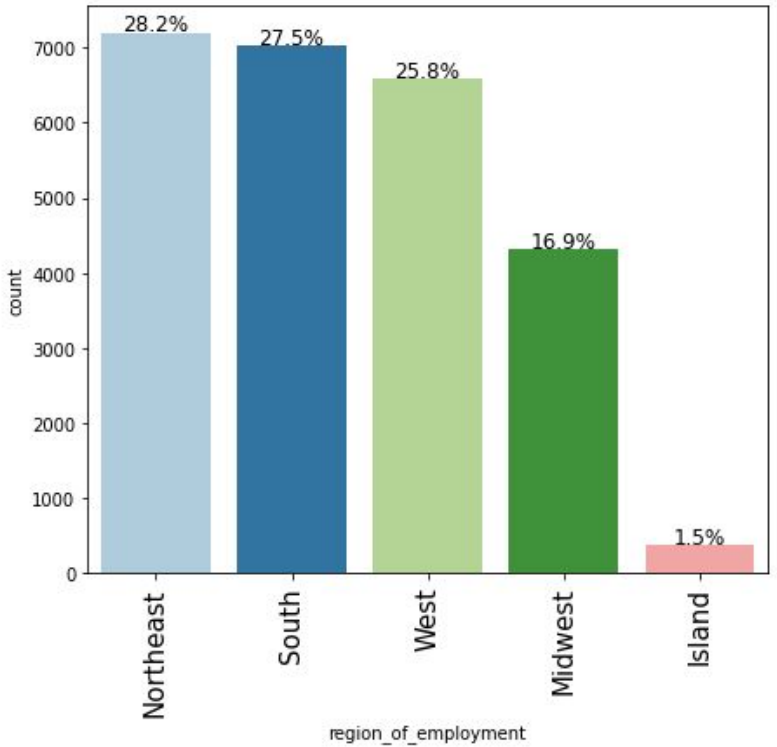


Link to Appendix slide on data background check

EDA Results
Univariate Analysis

Observations on region of employment

Visual analysis of the bar plot shows
- More foreign workers intend applying For jobs in the Northeast region (28%) followed by South (28%) and West (26%) regions.

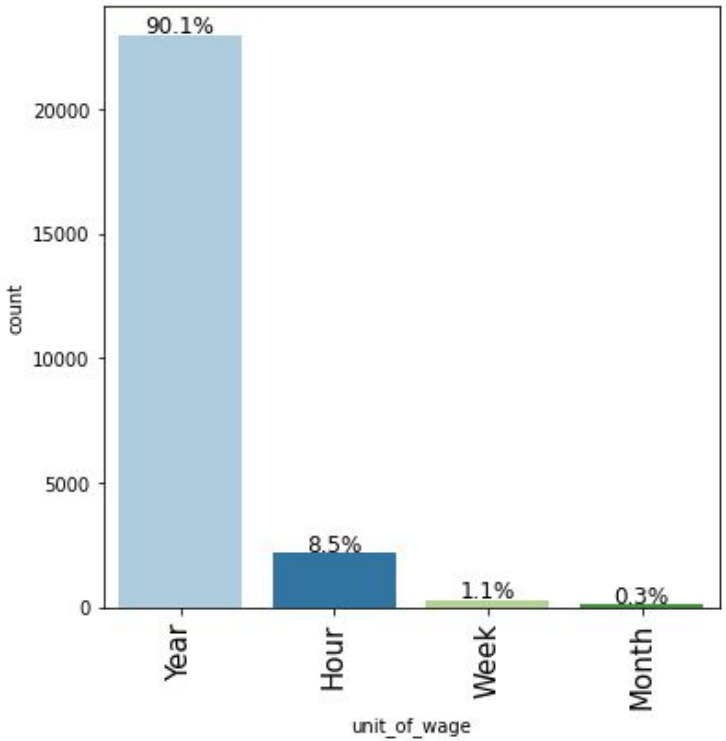


Link to Appendix slide on data background check

EDA Results
Univariate Analysis

Observations on unit wage

Visual analysis of the bar plot shows
- 90% Unit prevailing wage falls under Yearly,
9% falls under Hourly.

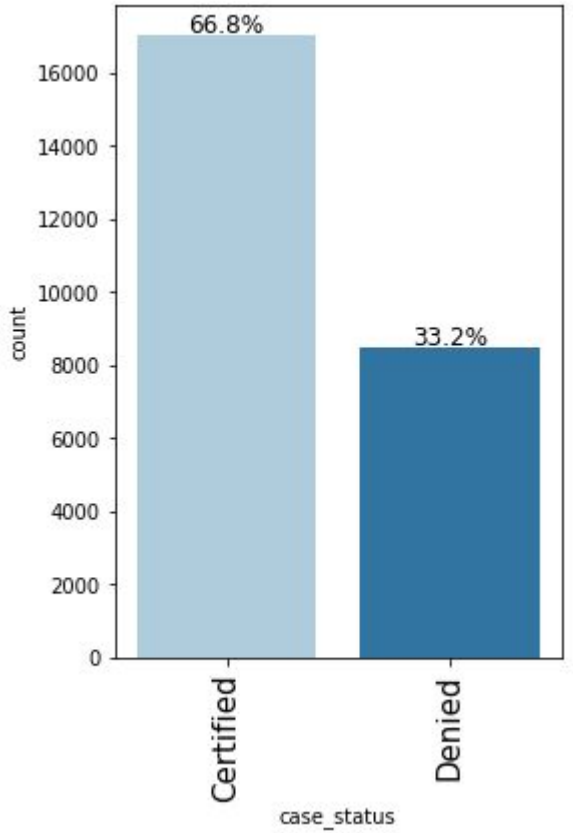


Link to Appendix slide on data background check

EDA Results
Univariate Analysis

Observations on case status

Visual analysis of the bar plot shows
- About 67% of cases are certified or approved and 33% are denied.

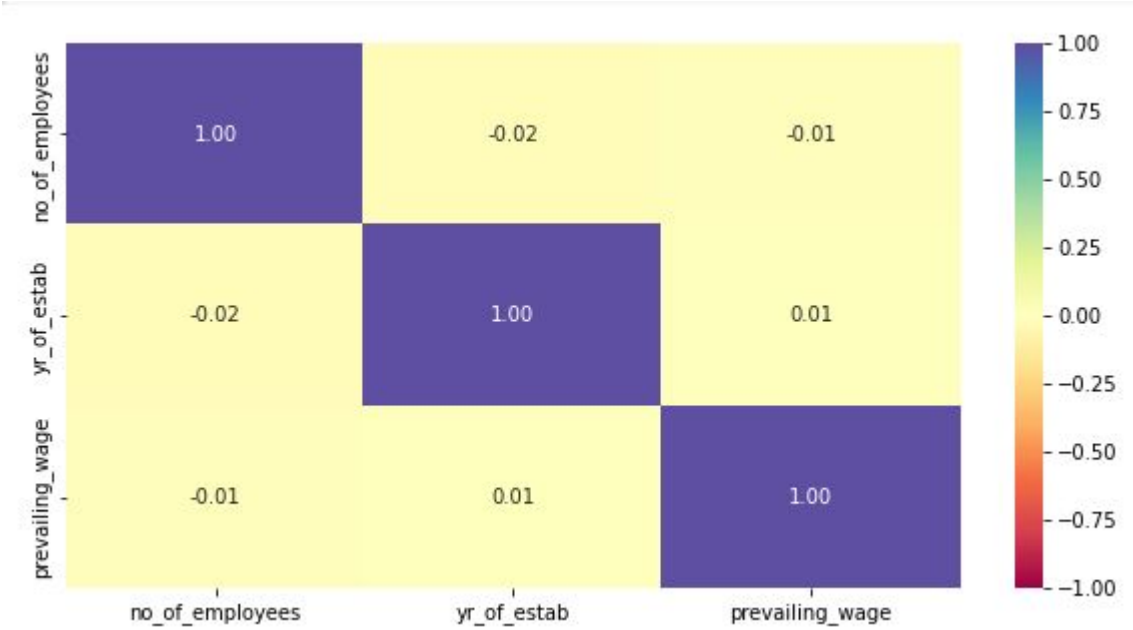


Link to Appendix slide on data background check

EDA Results
Bivariate Analysis

Observations on correlation heat map

Visual analysis of the heatmap shows
- There is positive correlation between no_of_employees, yr_of_estab and Prevailing_wage.

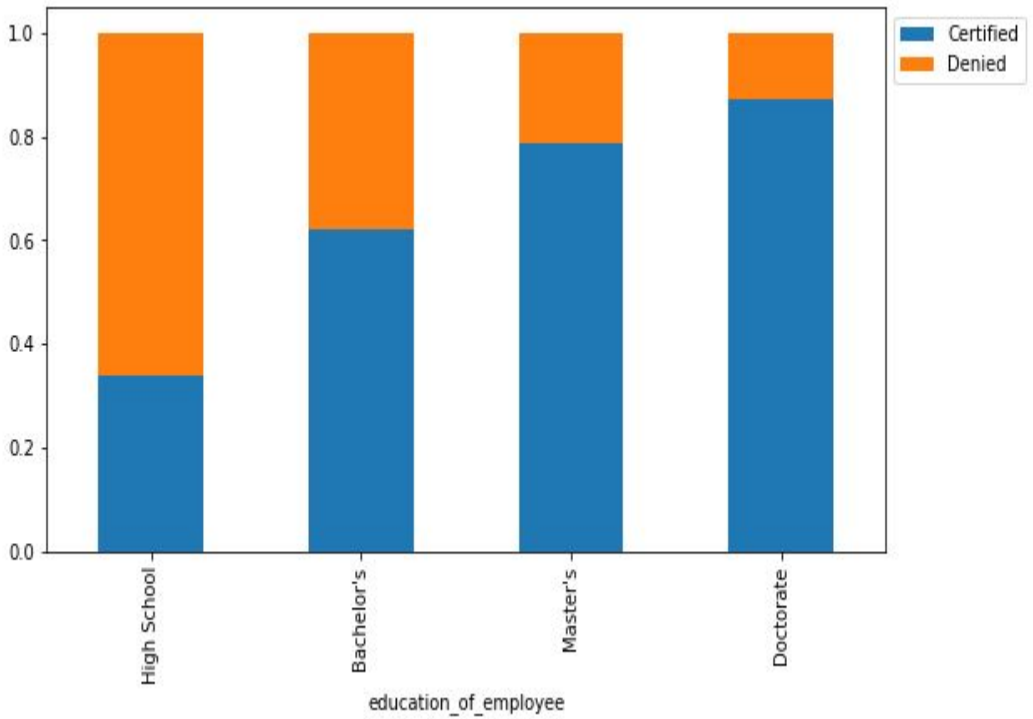


Link to Appendix slide on data background check

EDA Results
Bivariate Analysis

Observations on education impact on visa certification

Visual analysis of the barplot shows
- employee with High School are most denied and employee with Doctorate degrees are most certified or offered visa.



Link to Appendix slide on data background check

EDA Results
Bivariate Analysis

Observations on educational background requirement for different regions

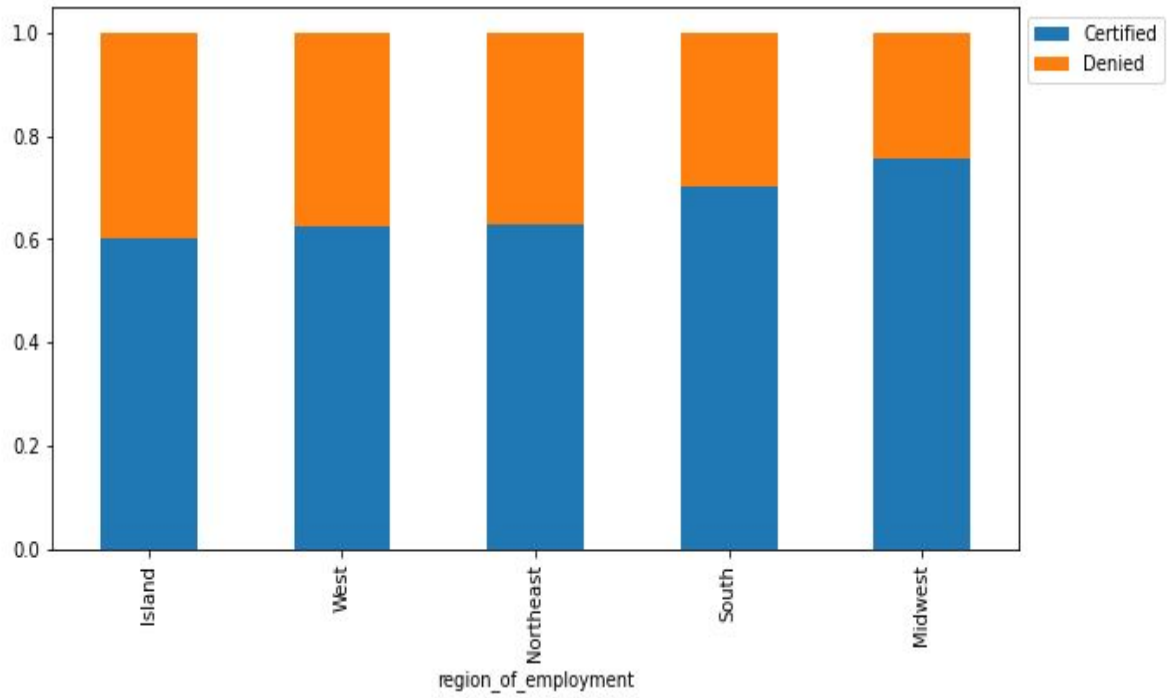
Visual analysis of the heatmap shows
- *The requirement for Bachelors in more in the South, requirement for Doctorate is more in the West, requirement for High School is more in the South and Requirement for Master's is more in the Northeast region.*

[Link to Appendix slide on data background check]

EDA Results
Bivariate Analysis

Observations on percentage of visa certifications across each region

Visual analysis of the stacked barplot shows
- Midwest region have the highest visa certifications, Island and Northeast have the lowest.

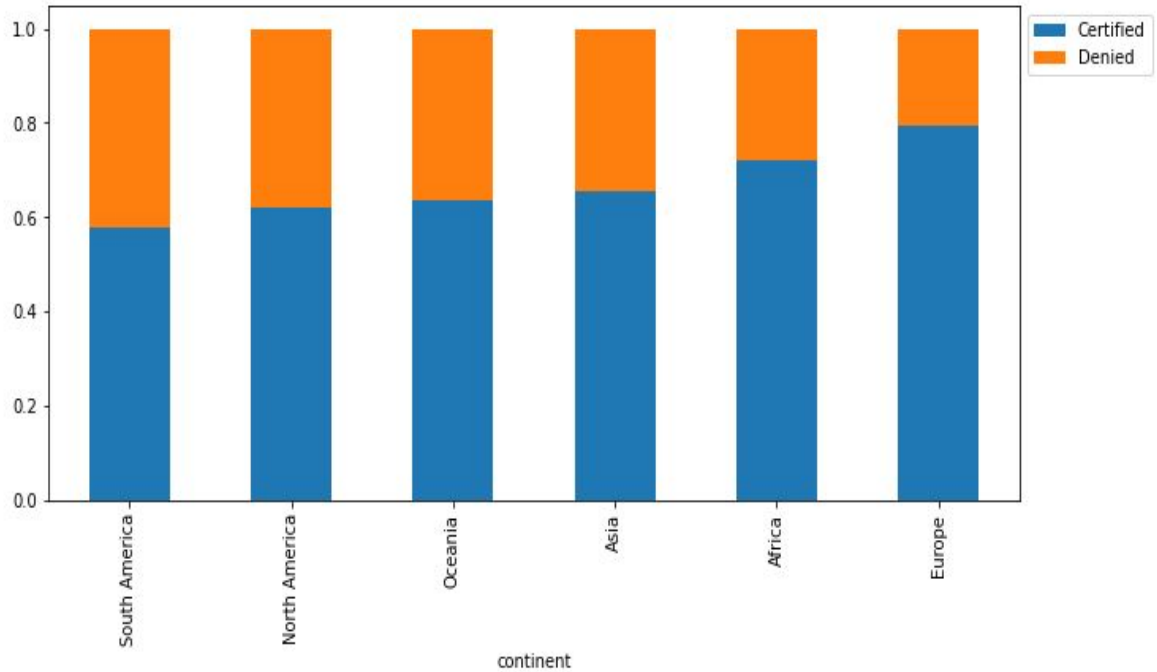


Link to Appendix slide on data background check

EDA Results
Bivariate Analysis

Observations on how visa status vary across different continents.

Visual analysis of the barplot shows
- Europe and Africa leads getting their visa certified compared to other continents.

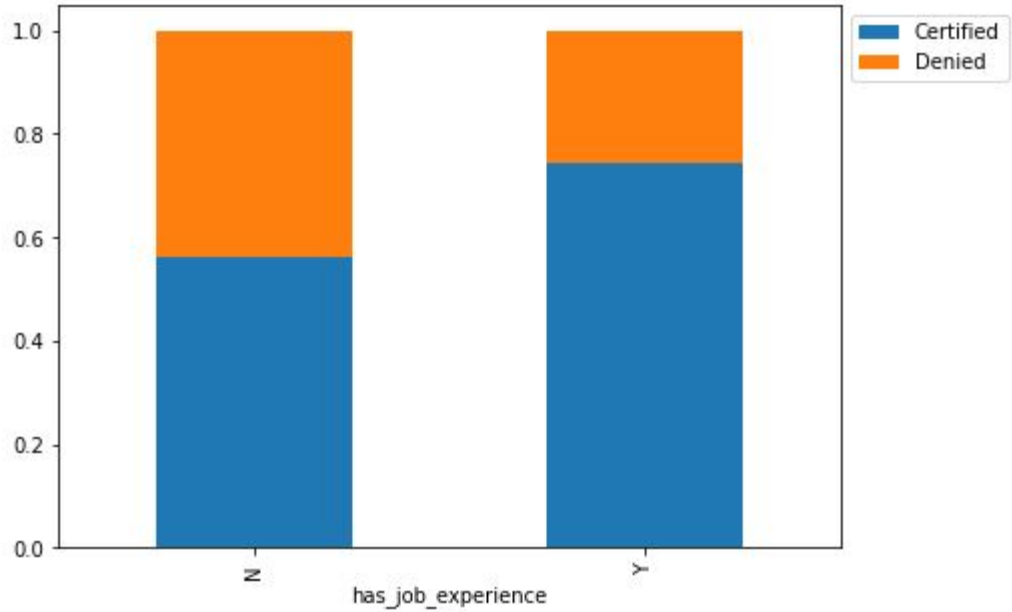


Link to Appendix slide on data background check

EDA Results
Bivariate Analysis

Observations on job experience and case status

Visual analysis of the barplot shows
- applicants who have job experience are more likely to get visa certified. About 80% of those who have job experience got certified.

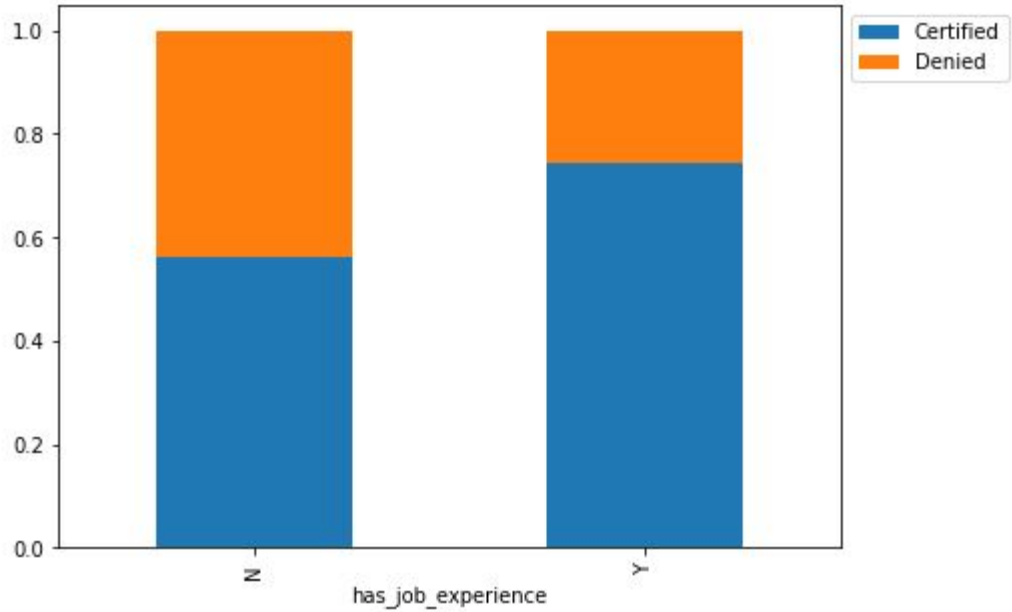


Link to Appendix slide on data background check

EDA Results
Bivariate Analysis

Observations on job experience and require job training

Visual analysis of the barplot shows
- applicants who have job experience requires less job training.

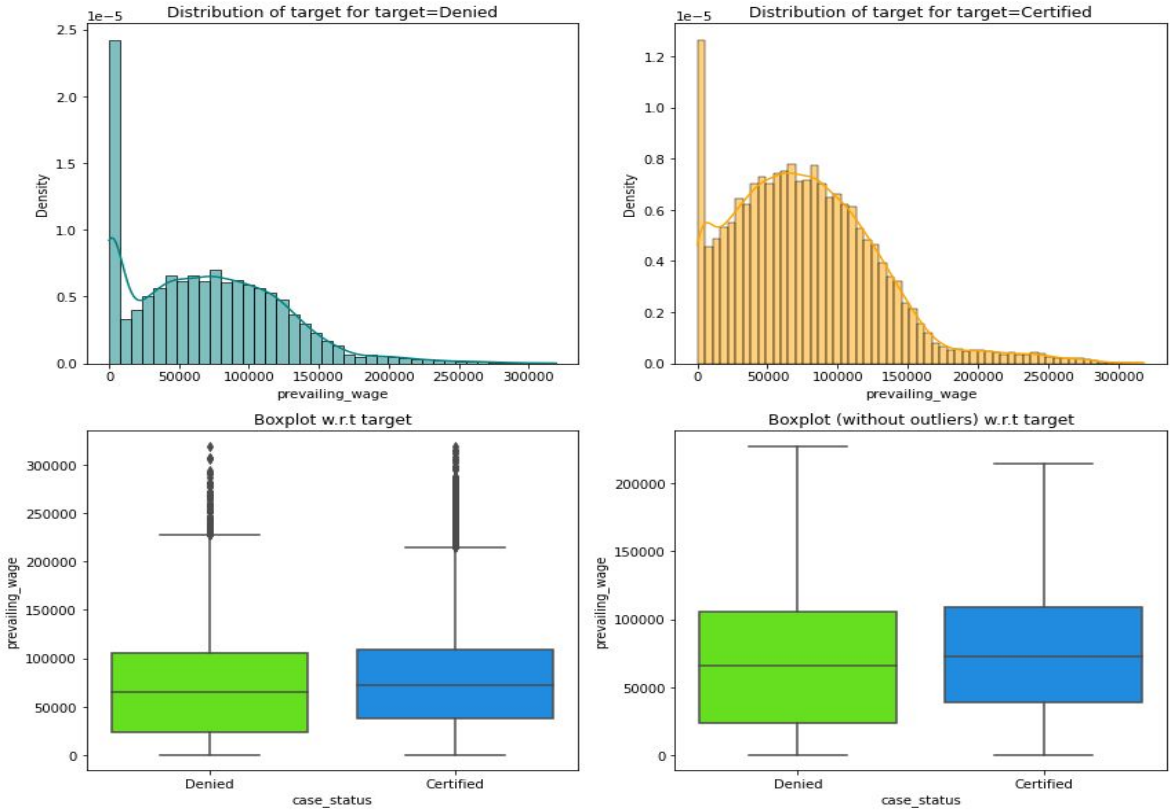


Link to Appendix slide on data background check

EDA Results
Bivariate Analysis

Observations on distribution of prevailing wage and case status

Visual analysis of the distribution plot shows
- the median prevailing wage for certified applications is slightly higher compared to the denied applications.

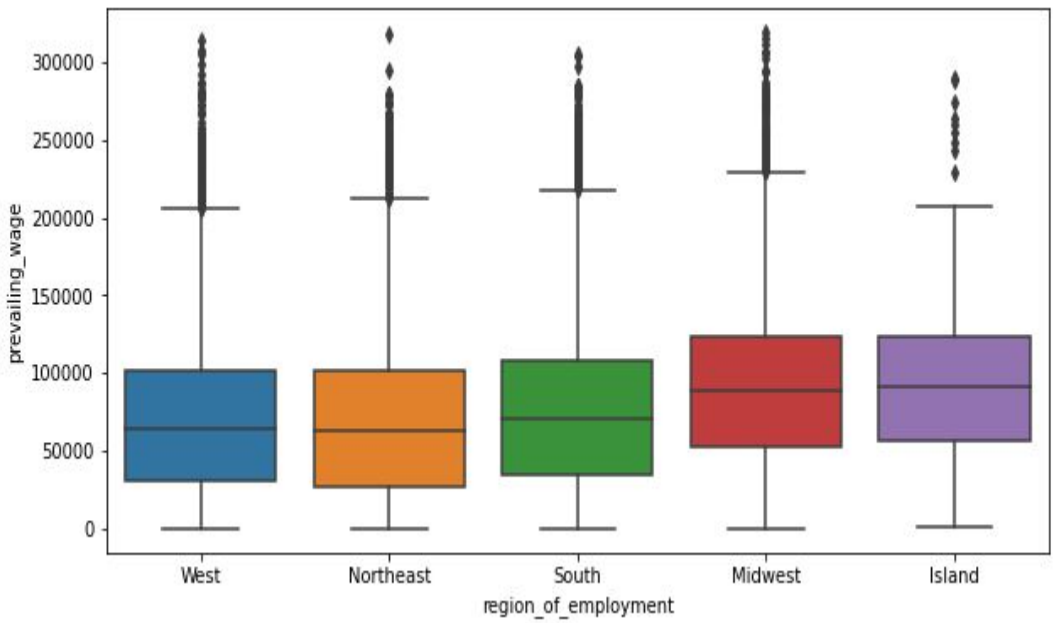


Link to Appendix slide on data background check

EDA Results
Bivariate Analysis

Observations on region of employment and prevailing wage

Visual analysis of the distribution plot shows
- Midwest and Island have slightly higher prevailing wages compared to rest.

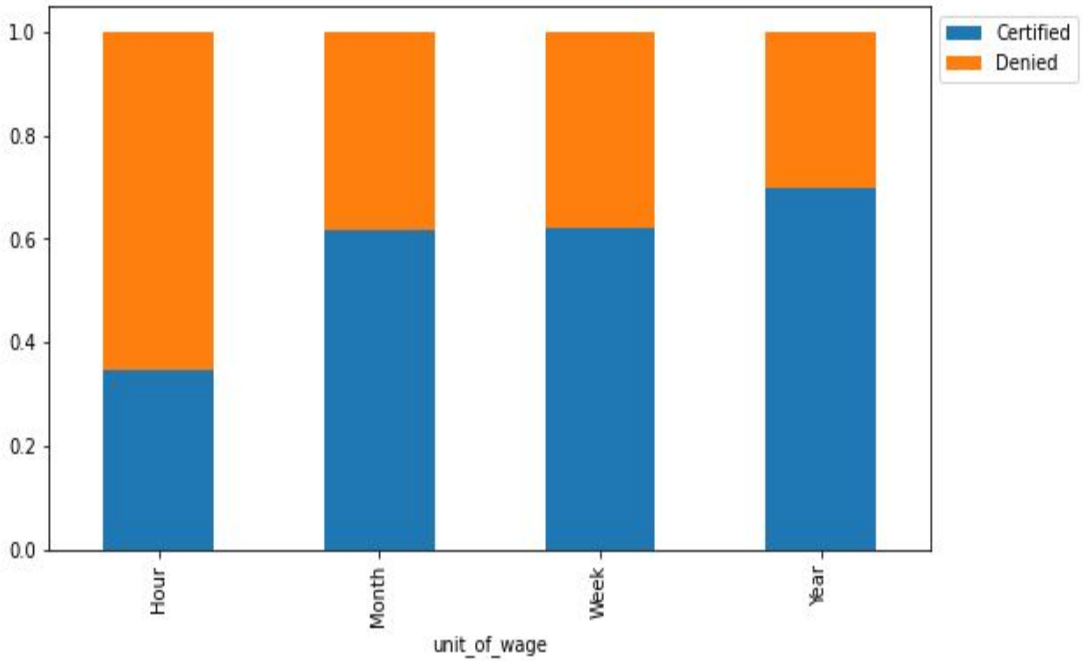


Link to Appendix slide on data background check

EDA Results
Bivariate Analysis

Observations on unit wage and case status

Visual analysis of the barplot shows
- applicant with yearly unit wage have high chance of getting certified followed by applicants with weekly and monthly unit wage. Hourly unit wage applicants have the lowest chance of getting certified.

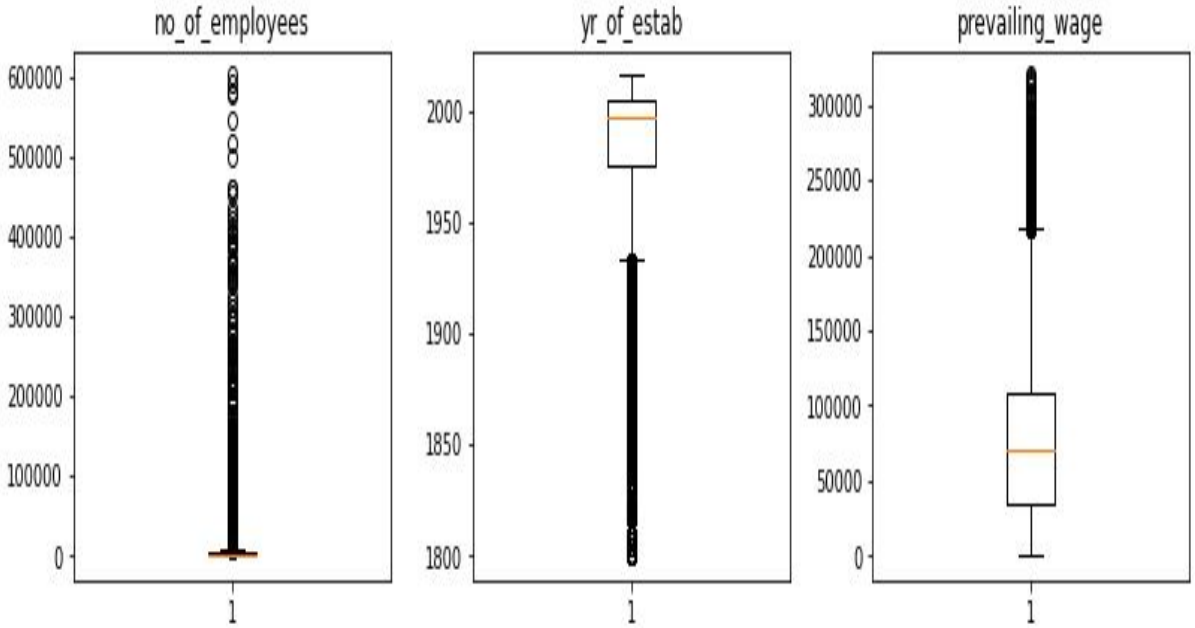


Link to Appendix slide on data background check

Data Preprocessing

Check for outliers in the data

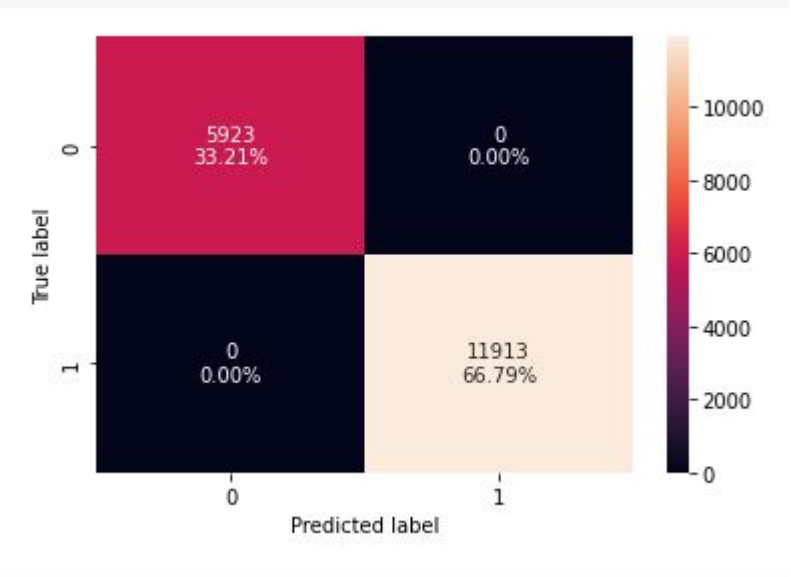
Visual analysis of the boxplot shows
- few outliers are present in the data.



Model Performance Summary

Checking model performance on training set

Visual analysis of the confusion matrix for train data shows zero errors 0% on training set and overfitting.



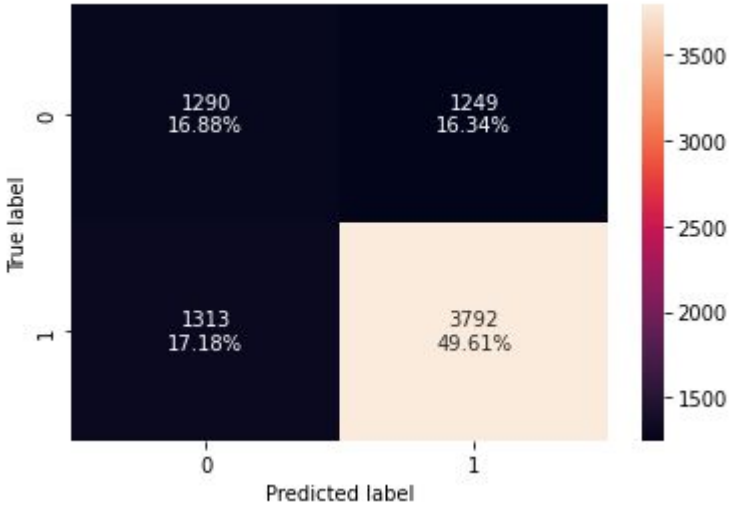
	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Link to Appendix slide on model assumptions

Model Performance Summary

Checking model performance on test set

Visual analysis of the confusion matrix for test data shows overfitting. Needs improvement.



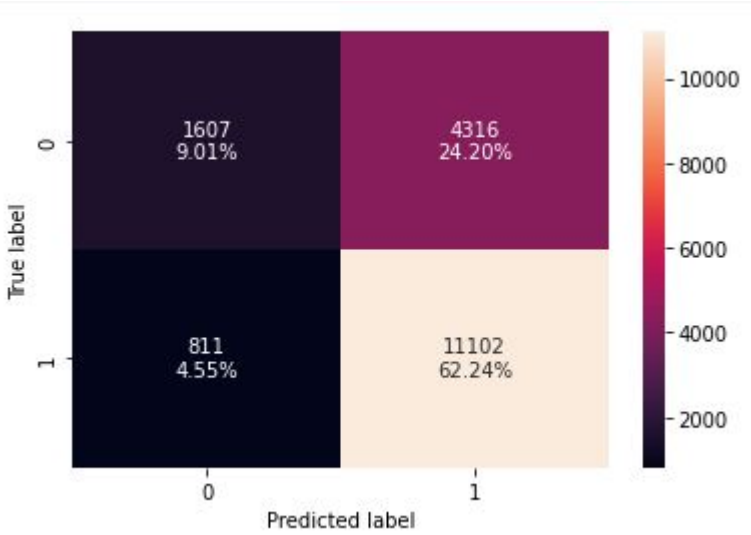
	Accuracy	Recall	Precision	F1
0	0.664835	0.742801	0.752232	0.747487

Link to Appendix slide on model assumptions

Model Performance Summary

Hyperparameter Tuning - Decision Tree

Visual analysis of the confusion matrix for training data on tuned estimator.



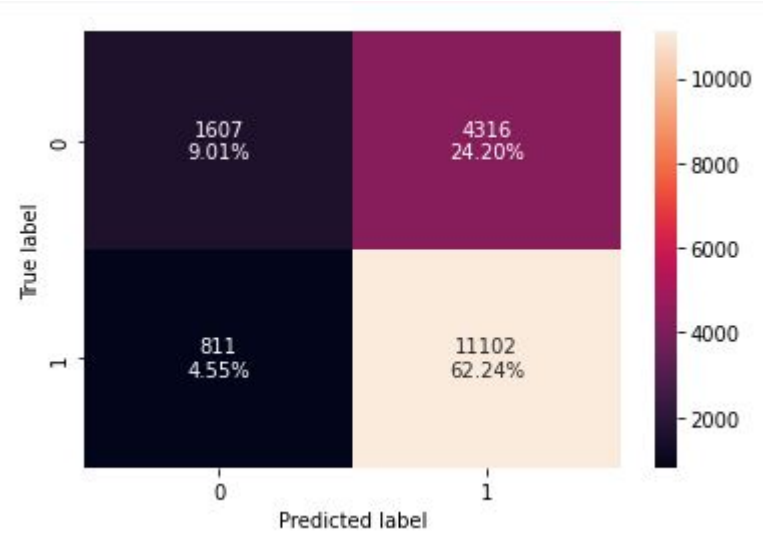
	Accuracy	Recall	Precision	F1
0	0.706567	0.930852	0.715447	0.809058

Link to Appendix slide on model assumptions

Model Performance Summary

Checking model performance on test set

Visual analysis of the confusion matrix for test data on tuned estimator shows reduction in overfitting, F1 score for both training and test set is 0.80.



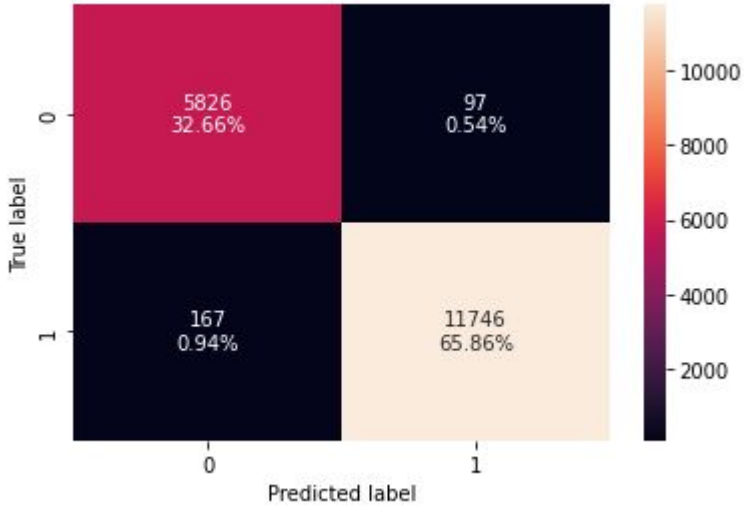
	Accuracy	Recall	Precision	F1
0	0.706567	0.930852	0.715447	0.809058

Link to Appendix slide on model assumptions

Model Performance Summary

Bagging - Model Building and Hyperparameter

Visual analysis of the confusion matrix for training data shows overfitting.



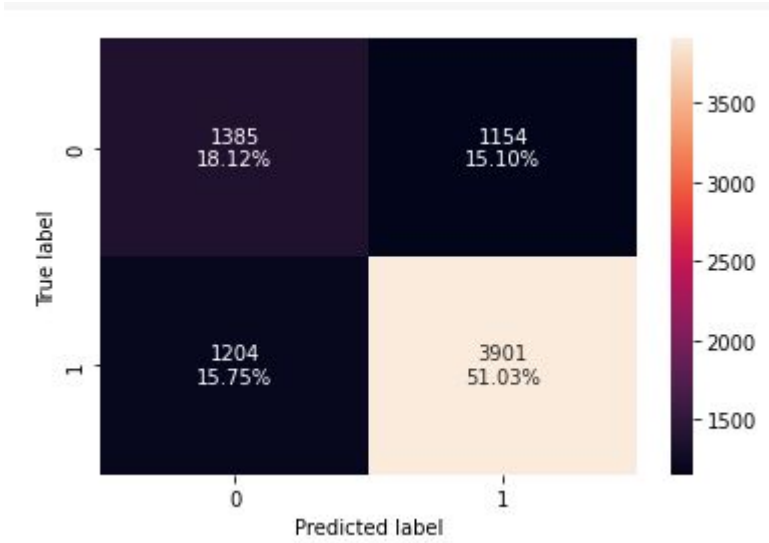
	Accuracy	Recall	Precision	F1
0	0.985198	0.985982	0.99181	0.988887

Link to Appendix slide on model assumptions

Model Performance Summary

Checking model performance on test set

Visual analysis of the confusion matrix for test data on bagging classifier shows reduction in overfitting, F1 score for both training and test set is 0.80



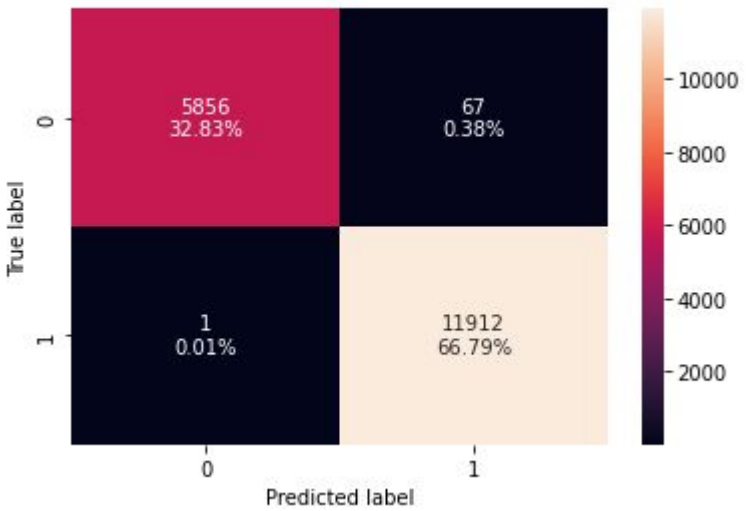
	Accuracy	Recall	Precision	F1
0	0.691523	0.764153	0.771711	0.767913

Link to Appendix slide on model assumptions

Model Performance Summary

Hyperparameter Tuning - Bagging Classifier

Visual analysis of the confusion matrix for training data shows overfitting.



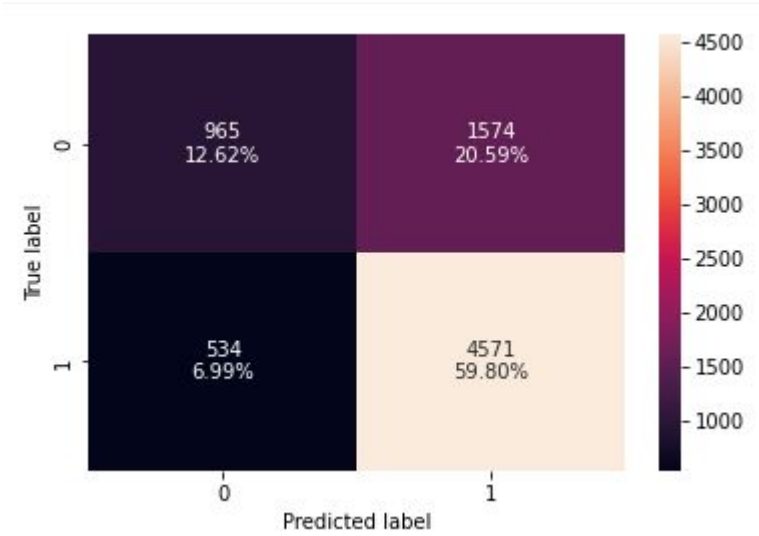
	Accuracy	Recall	Precision	F1
0	0.996187	0.999916	0.994407	0.997154

Link to Appendix slide on model assumptions

Model Performance Summary

Checking model performance on test set

Visual analysis of the confusion matrix for test data on bagging classifier shows big difference between training and test data.



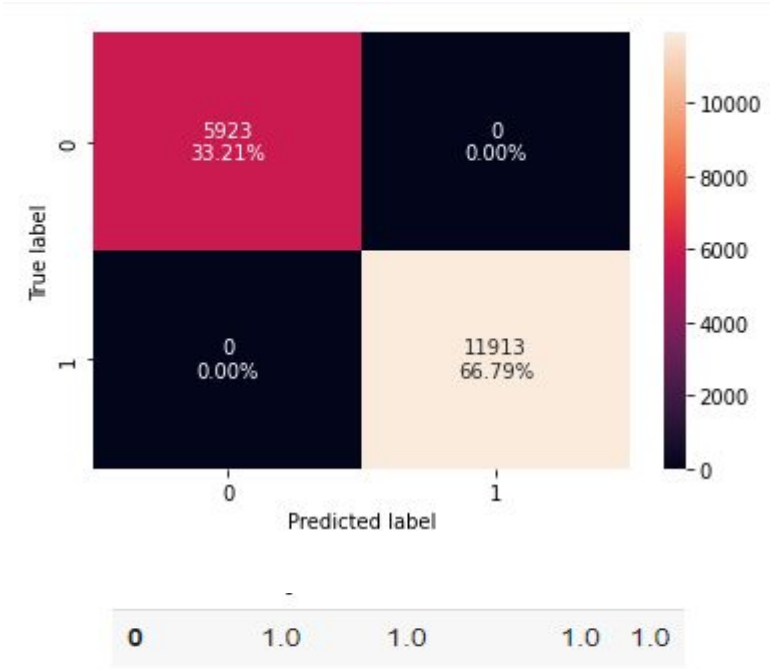
	Accuracy	Recall	Precision	F1
0	0.724228	0.895397	0.743857	0.812622

Link to Appendix slide on model assumptions

Model Performance Summary

Random Forest

Visual analysis of the confusion matrix for training data shows overfitting.



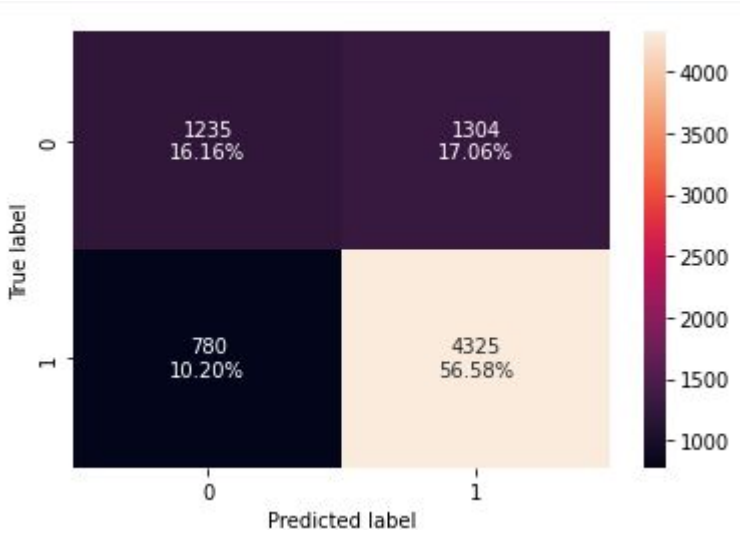
0	1.0	1.0	1.0	1.0

Link to Appendix slide on model assumptions

Model Performance Summary

Random Forest

Visual analysis of the confusion matrix for test data on bagging classifier showing F1 as 0.80



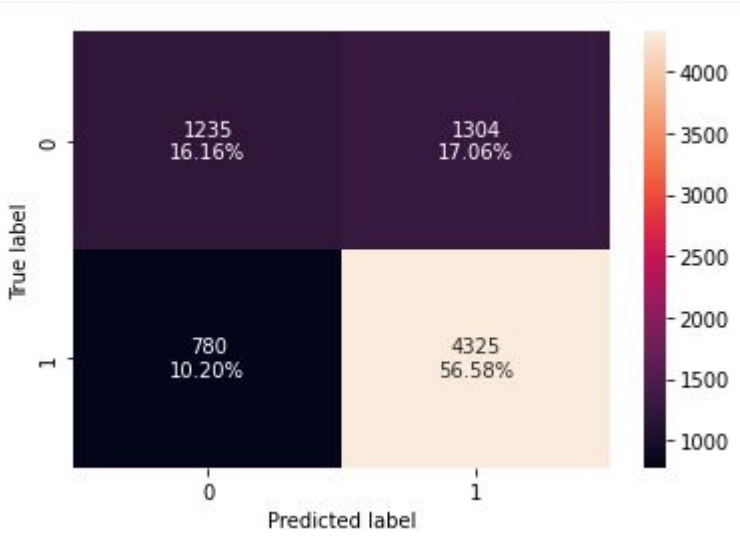
	Accuracy	Recall	Precision	F1
0	0.727368	0.847209	0.768343	0.805851

Link to Appendix slide on model assumptions

Model Performance Summary

Random Forest

Visual analysis of the confusion matrix for test data on bagging classifier showing F1 as 0.80



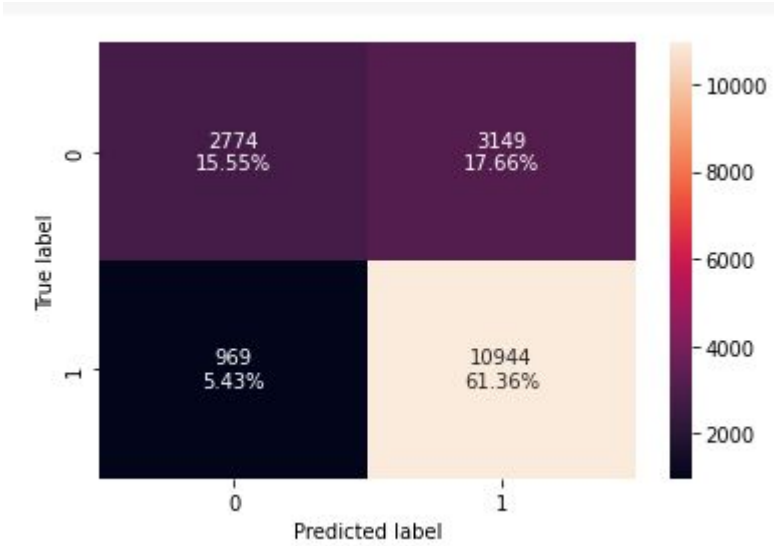
	Accuracy	Recall	Precision	F1
0	0.727368	0.847209	0.768343	0.805851

Link to Appendix slide on model assumptions

Model Performance Summary

Hyperparameter Tuning - Random Forest

Visual analysis of the confusion matrix for training data shows it has generalized.



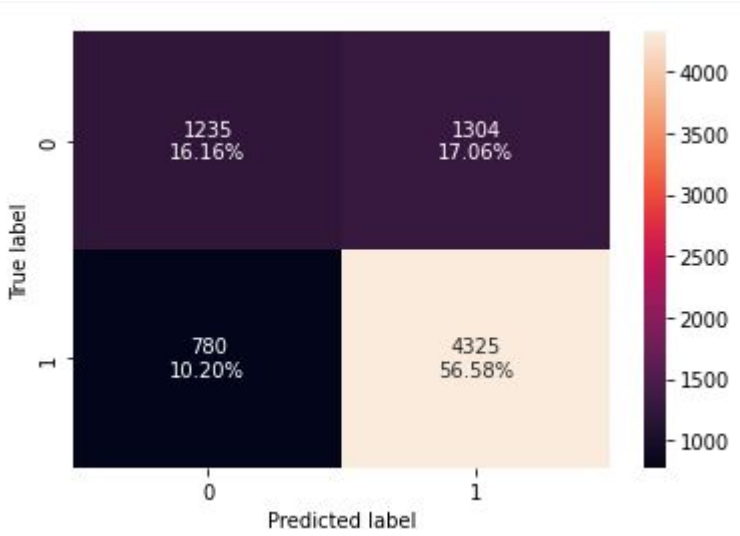
	Accuracy	Recall	Precision	F1
)	0.769119	0.91866	0.776556	0.841652

Link to Appendix slide on model assumptions

Model Performance Summary

Hyperparameter Tuning - Random Forest

Visual analysis of the confusion matrix for test data shows good precision. F1 scores are 0.84 and 0.82 on training and test data. No overfitting.



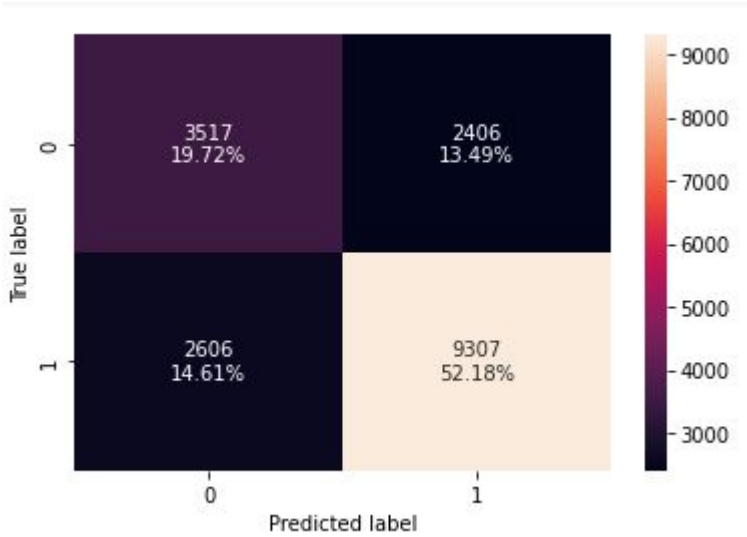
	Accuracy	Recall	Precision	F1
0	0.727368	0.847209	0.768343	0.805851

Link to Appendix slide on model assumptions

Model Performance Summary

Hyperparameter Tuning - AdaBoost Classifier

Visual analysis of the confusion matrix for training data shows it has generalized.



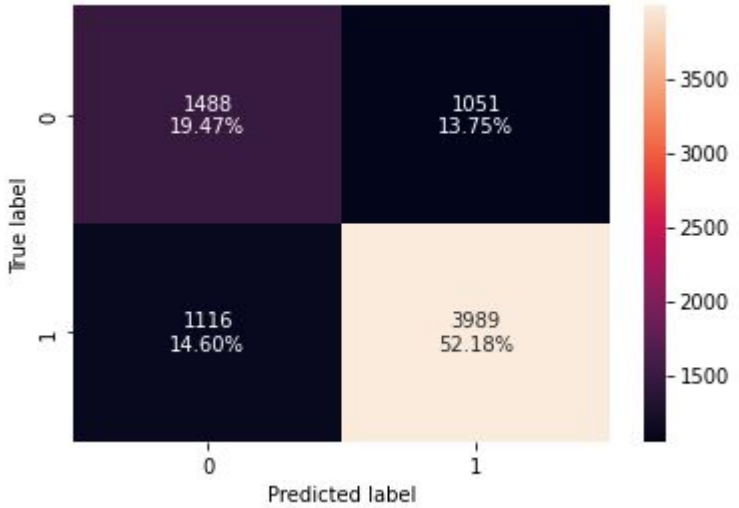
	Accuracy	Recall	Precision	F1
0	0.718995	0.781247	0.794587	0.787861

Link to Appendix slide on model assumptions

Model Performance Summary

Hyperparameter Tuning - AdaBoost Classifier

Visual analysis of the confusion matrix for test data shows good precision. F1 scores are 0.78 and 0.78 on training and test data. Model has good precision and high recall.



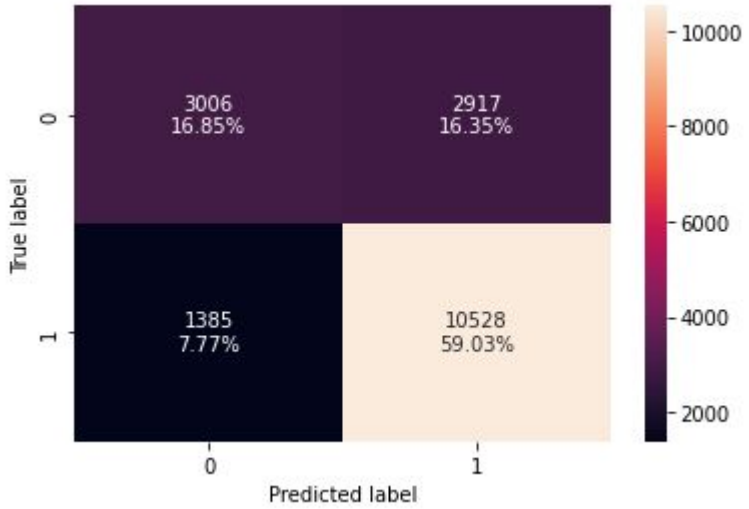
	Accuracy	Recall	Precision	F1
0	0.71651	0.781391	0.791468	0.786397

Link to Appendix slide on model assumptions

Model Performance Summary

Gradient Boosting Classifier

Visual analysis of the confusion matrix for training data shows it has generalized.



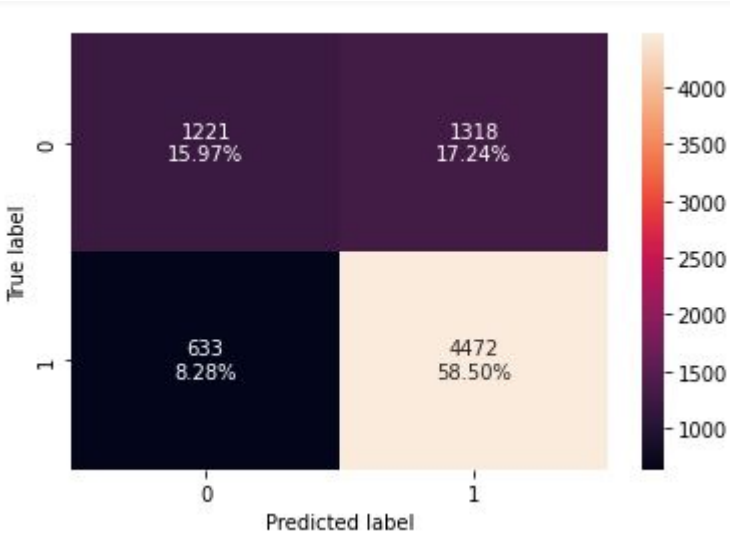
	Accuracy	Recall	Precision	F1
0	0.758802	0.88374	0.783042	0.830349

Link to Appendix slide on model assumptions

Model Performance Summary

Gradient Boosting Classifier

Visual analysis of the confusion matrix for test data shows model has generalized performance, with F1 score 0.83 and 0.82 for training and test set.



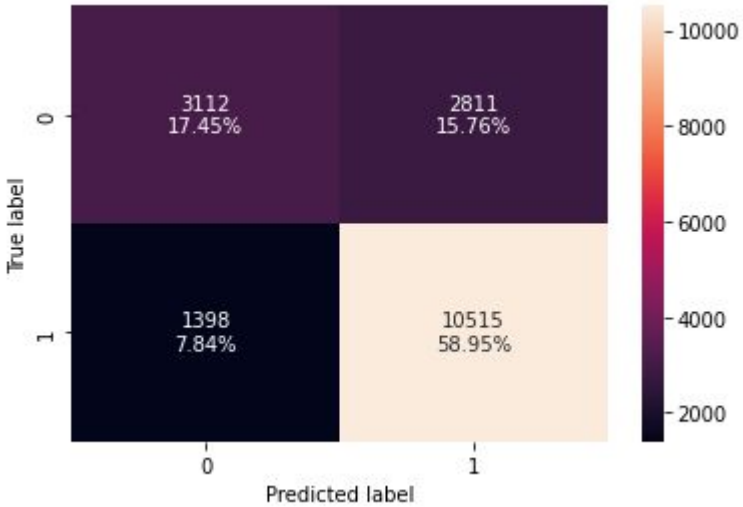
	Accuracy	Recall	Precision	F1
0	0.744767	0.876004	0.772366	0.820927

Link to Appendix slide on model assumptions

Model Performance Summary

Hyperparameter Tuning - Gradient Boosting Classifier

Visual analysis of the confusion matrix for train data



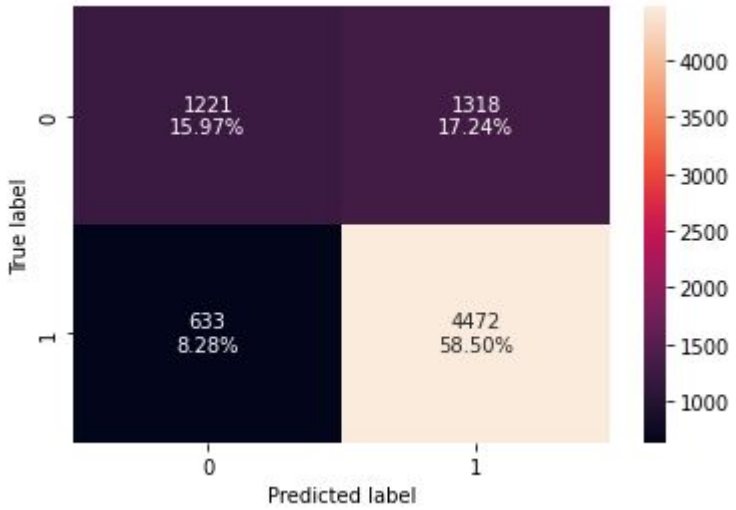
	Accuracy	Recall	Precision	F1
0	0.764017	0.882649	0.789059	0.833234

Link to Appendix slide on model assumptions

Model Performance Summary

Hyperparameter Tuning - Gradient Boosting Classifier

Visual analysis of the confusion matrix for test data shows no much difference after hyperparameter tuning.



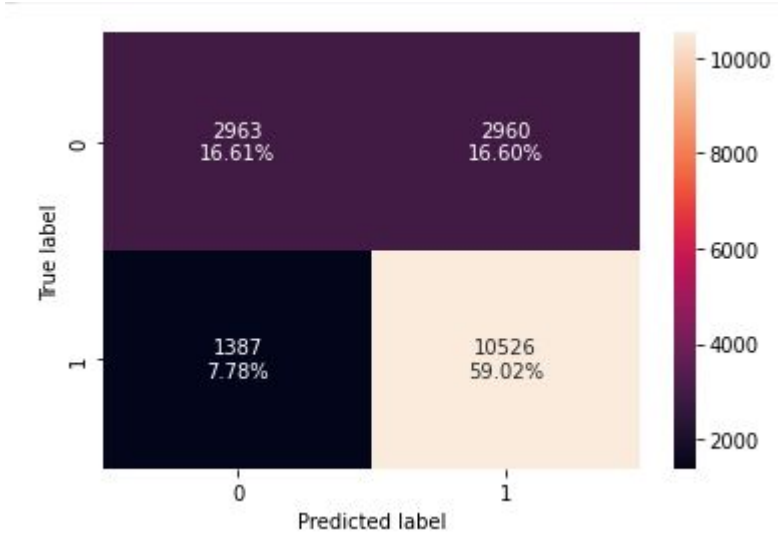
	Accuracy	Recall	Precision	F1
0	0.744767	0.876004	0.772366	0.820927

Link to Appendix slide on model assumptions

Model Performance Summary

XGBoost Classifier

Visual analysis of the confusion matrix for train data



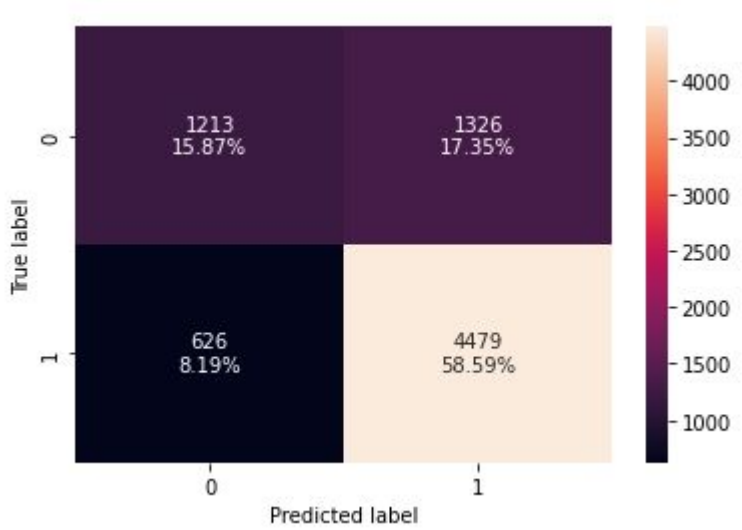
	Accuracy	Recall	Precision	F1
0	0.756279	0.883573	0.780513	0.828852

Link to Appendix slide on model assumptions

Model Performance Summary

XGBoost Classifier

Visual analysis of the confusion matrix for test data shows generalized performance, but will further tune the hyperparameters to see if any further improvement.



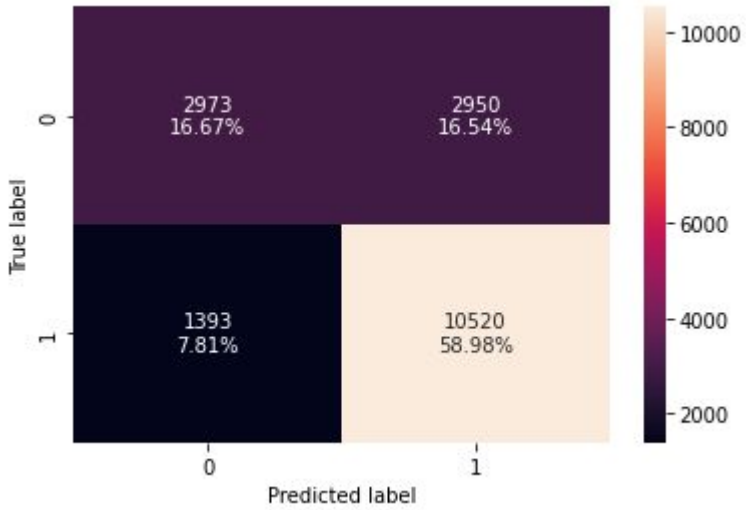
	Accuracy	Recall	Precision	F1
0	0.744636	0.877375	0.771576	0.821082

Link to Appendix slide on model assumptions

Model Performance Summary

Hyperparameter Tuning - XGBoost Classifier

Visual analysis of the confusion matrix for train data



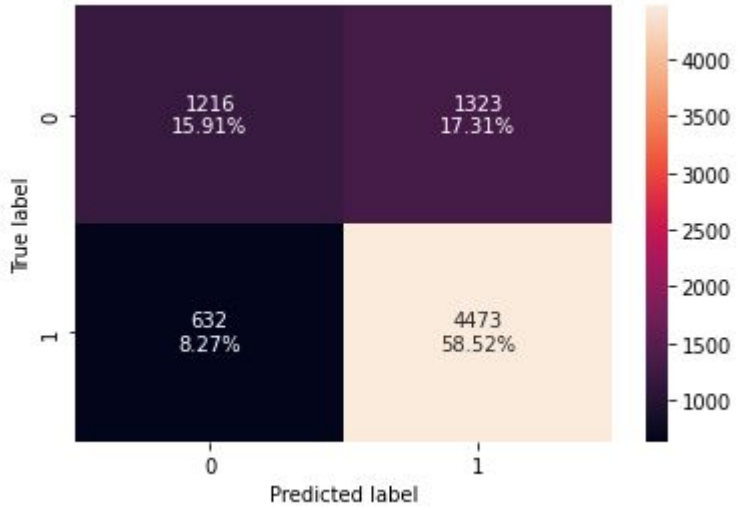
	Accuracy	Recall	Precision	F1
0	0.756504	0.883069	0.780995	0.828901

Link to Appendix slide on model assumptions

Model Performance Summary

Hyperparameter Tuning - XGBoost Classifier

Visual analysis of the confusion matrix for test data shows generalized performance, with F1 score of 0.82 for both training and test set.



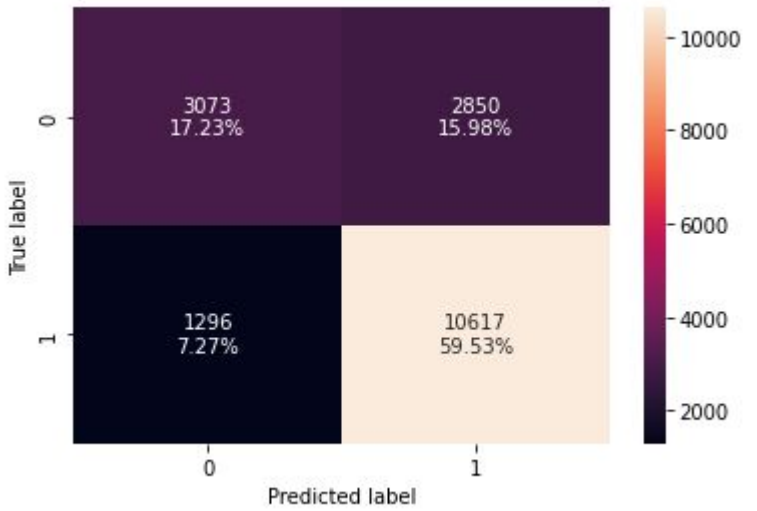
	Accuracy	Recall	Precision	F1
0	0.744244	0.8762	0.771739	0.820659

Link to Appendix slide on model assumptions

Model Performance Summary

Stacking Classifier

Visual analysis of the confusion matrix for train data



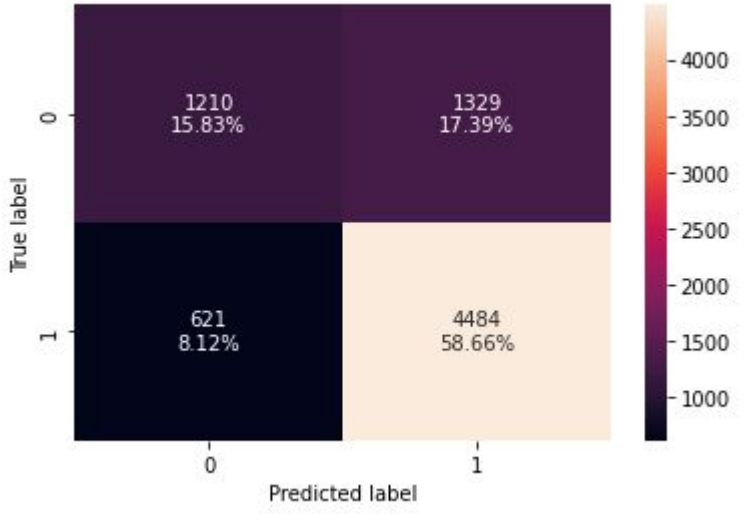
	Accuracy	Recall	Precision	F1
0	0.744244	0.8762	0.771739	0.820659

Link to Appendix slide on model assumptions

Model Performance Summary

Hyperparameter Tuning - XGBoost Classifier

Visual analysis of the confusion matrix for test data shows generalized performance, with F1 score of 0.82 for both training and test set, as we have in XGBoost model (no difference)



	Accuracy	Recall	Precision	F1
0	0.744898	0.878355	0.771375	0.821396

Link to Appendix slide on model assumptions

Model Performance Summary

Model Performance Comparison and Final Model Selection

Training performance comparison

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	0.706567	0.985198	0.996187	1.0	0.769119	0.738226	0.718995	0.758802	0.764017	0.756279	0.756504	0.744244
Recall	1.0	0.930852	0.985982	0.999916	1.0	0.918660	0.887182	0.781247	0.883740	0.882649	0.883573	0.883069	0.876200
Precision	1.0	0.715447	0.991810	0.994407	1.0	0.776556	0.760688	0.794587	0.783042	0.789059	0.780513	0.780995	0.771739
F1	1.0	0.809058	0.988887	0.997154	1.0	0.841652	0.819080	0.787861	0.830349	0.833234	0.828852	0.828901	0.820659

Link to Appendix slide on model assumptions

Model Performance Summary

Model Performance Comparison and Final Model Selection

Test performance comparison

- *Findings shows Decision Tree, Bagging Classifier (Default and Tuned), Random Forest (Default and Tuned) were found to overfit the training set.*
- *Decision Tree (Tuned), Random Forest (Tuned), AdaBoost Default), Gradient Boost (Default and Tuned), XGBoost (Default and Tuned) and Stacking (Default) gave a generalized performance on both training and testing datasets.*
- *Stacking Classifier has the highest F1 score.*

Testing performance comparison:

	0	0	0	0	0	0	0	0	0	0	0	0	0
Accuracy	0.706567	0.706567	0.691523	0.724228	0.727368	0.738095	0.734301	0.716510	0.744767	0.744767	0.744636	0.744244	0.744898
Recall	0.930852	0.930852	0.764153	0.895397	0.847209	0.898923	0.885015	0.781391	0.876004	0.876004	0.877375	0.876200	0.878355
Precision	0.715447	0.715447	0.771711	0.743857	0.768343	0.755391	0.757799	0.791468	0.772366	0.772366	0.771576	0.771739	0.771375
F1	0.809058	0.809058	0.767913	0.812622	0.805851	0.820930	0.816481	0.786397	0.820927	0.820927	0.821082	0.820659	0.821396

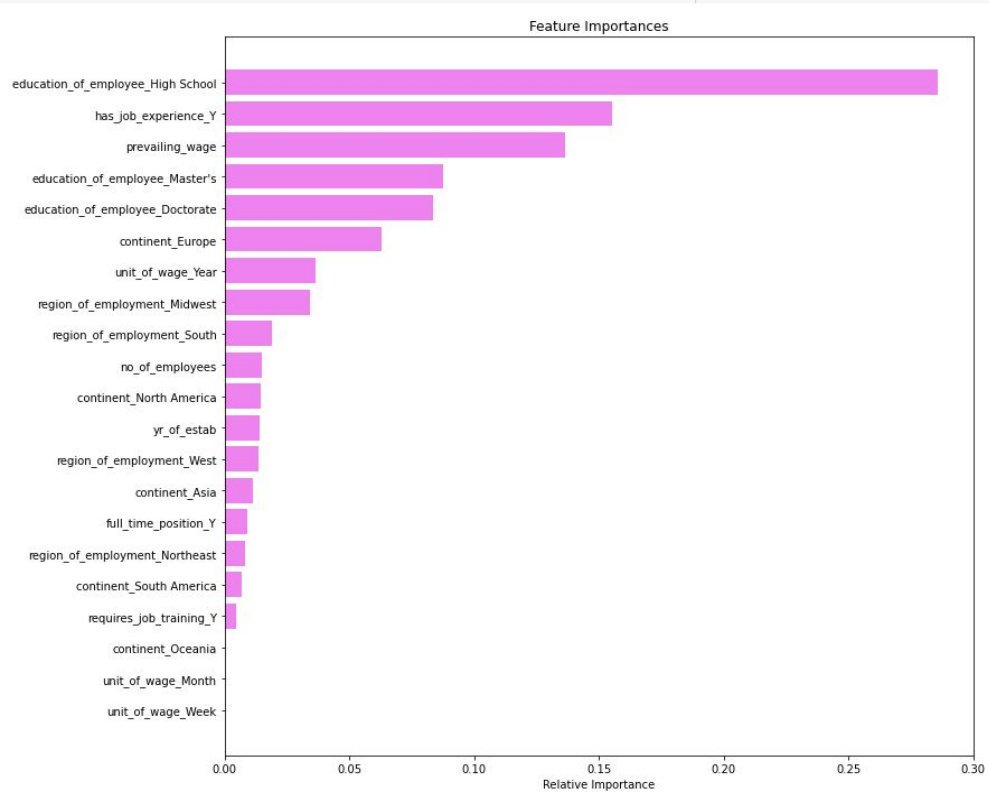
Link to Appendix slide on model assumptions

Model Performance Summary

Important features of the final model

The findings shows education of employee is the most important attribute having an influence on visa certifications.

Other important attributes are; employee having a prior job experience, prevailing wage and continent of the employee.



Link to Appendix slide on model assumptions

Actionable Insights and Recommendations

Based on the EDA and Stacking Classifier model, the following were observed as important factor for visas to get certified or get denied;

- Education of employee - employee with a doctorate degree have 65% chance of getting visa certified, and employee with high school certification has over 65% of getting visa denied.
- Unit of wage - employee with non-hourly pay has 70% chance of getting visa certified, and employee with hourly pay has 65% of getting visa denied.
- Continent - employee with work experience and from Europe has 75% and 80% chance of getting visa certified compared to employee with no work experience have 50% chance of getting visa denied.
- Region of employment - employees from Midwest and South have 70% chances of getting visa certified.

Attributes such as; full time or part time position, require job training, prevailing wage and year of establishment do not have much impact for visas to get certified or denied.

We built a model that can capture over 80% of the information while making predictions, the findings can help build a suitable profile of candidates to facilitate the process of visa.

Great Learning

Happy Learning !