EasyVisa Project

By: Esra Mercan

Contents

Context

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.

Objective

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. OFLC has hired the firm EasyVisa for data-driven solutions. You as a data scientist at EasyVisa have 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.

Data Overview

The data contains the different attributes of employee and the employer. The detailed data dictionary is given below.

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

Exploratory Data Analysis (EDA)

This is how our data looks in the beginning:

	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

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
```

Let's check the 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

1. There are 25480 rows and 12 columns in the data.
2. There is no missing values.
3. Average No_of_employees is 5667. It ranges from -26 to 602069. It has outliers.
4. No_of_employees has negative minimum value. Negative values has fixed to an absolute value to avoid errors.
5. Yr_of_estab indicates that some companies are established starting from 1800 to 2016. Median year is 1997. There are outliers.
6. Prevailing_wage average is about 74456. It ranges from 2.1 to 319210.

```
# checking for duplicate
data.duplicated().sum()
```

```
0
```

There is no duplicate values.

Exploratory Data Analysis (EDA)

Let's check the count of each unique category in each of the categorical variables.

```
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
```

```
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
```

```
N    22525
Y     2955
Name: requires_job_training, dtype: int64
```

```
Certified    17018
Denied        8462
Name: case_status, dtype: int64
```

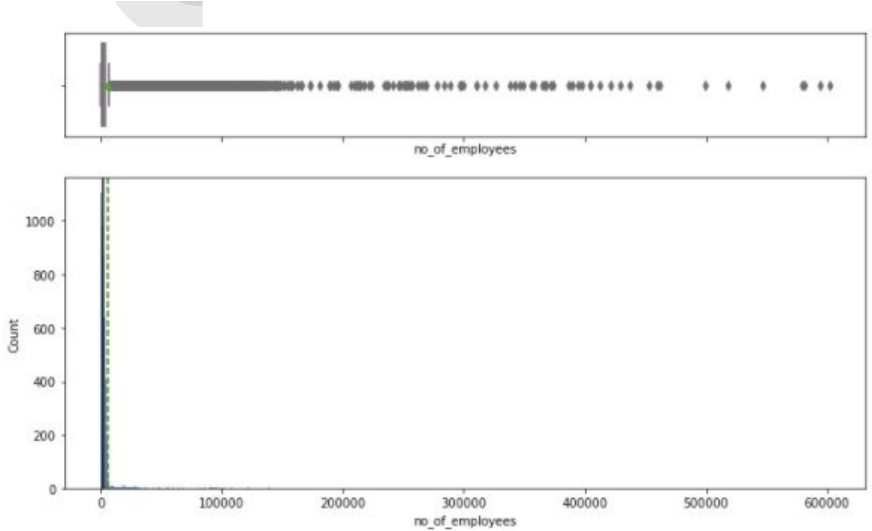
```
Y    22773
N     2707
Name: full_time_position, dtype: int64
```

```
Y    14802
N    10678
Name: has_job_experience, dtype: int64
```

Case_id column is dropped due to having unique category for each case. This would not contribute any benefit to the model.

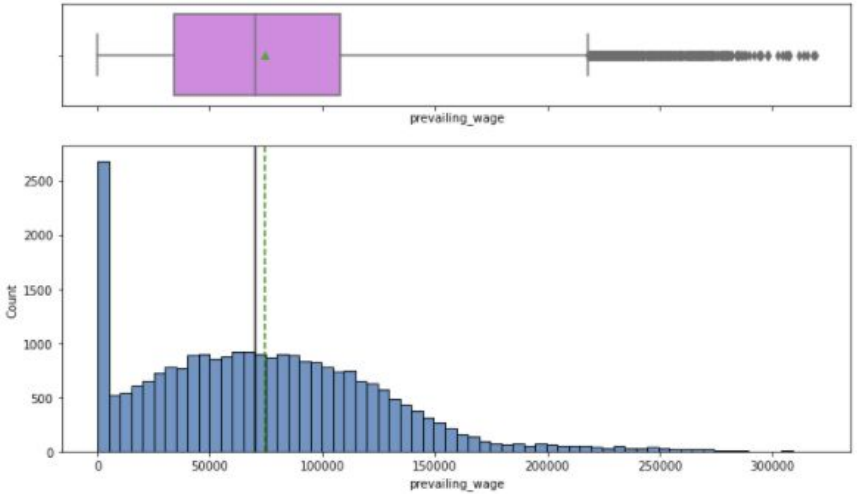
Exploratory Data Analysis (EDA)-Univariate Analysis

Observations on number of employees



The data is right skewed and there is outliers.

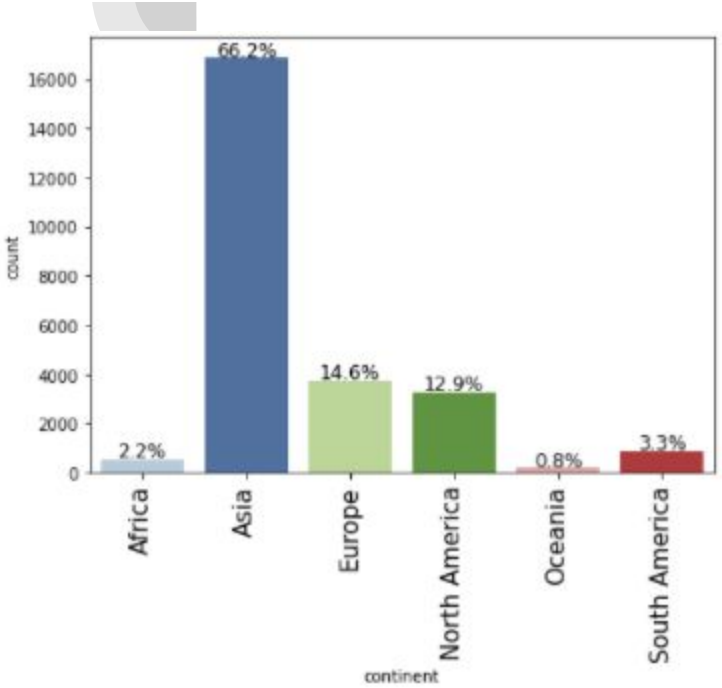
Observations on prevailing wage



The data is right skewed and there are outliers. Mean and median values are very close.
There are 176 case which has prevailing_wage less than $100 hourly.

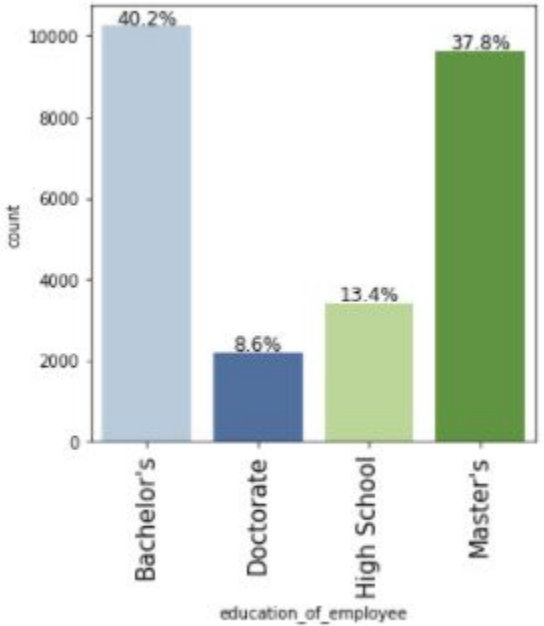
Exploratory Data Analysis (EDA)-Univariate Analysis

Observations on continent



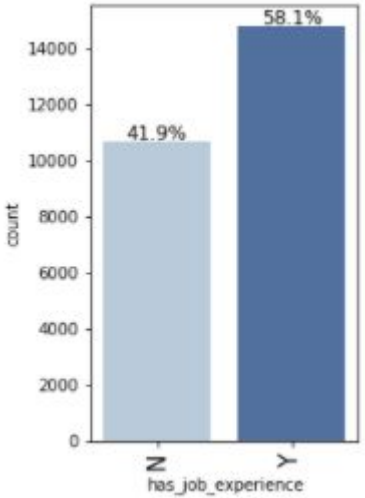
Asia is leading by 66.2% on number of cases.

Observations on education of employee



Most applicants has higher education.
Only 13.4% has high school diploma

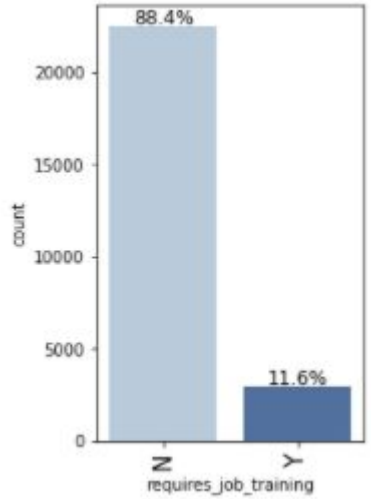
Observations on job experience



58% of the applicants has job experience.

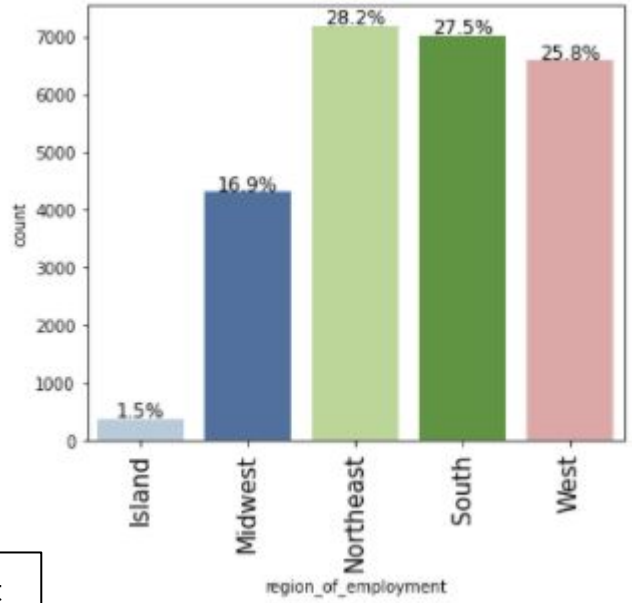
Exploratory Data Analysis (EDA)-Univariate Analysis

Observations on job training



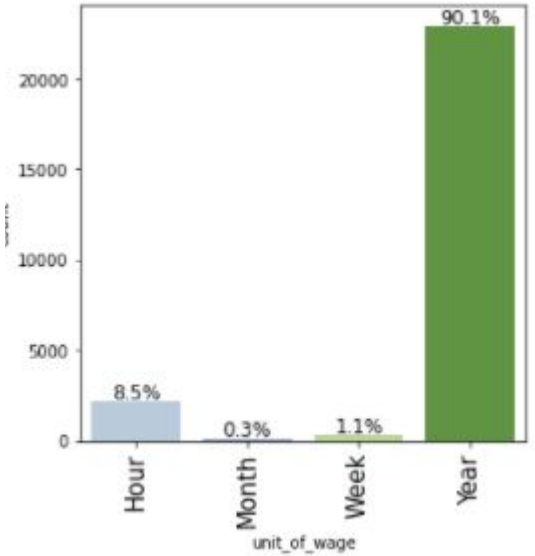
88.4% of the jobs does not require a job training

Observations on region of employment



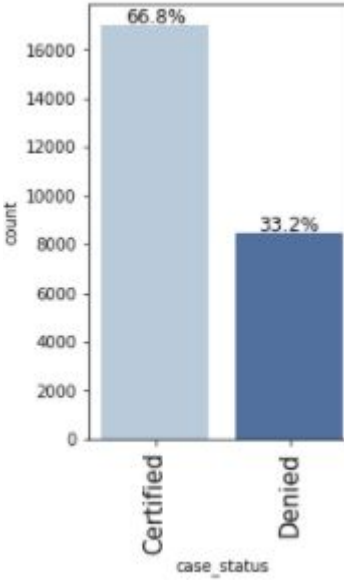
Northeast, South and west employments are very close to each other. Island has the least region of employment

Observations on unit of wage



90.1% of the wages are given yearly

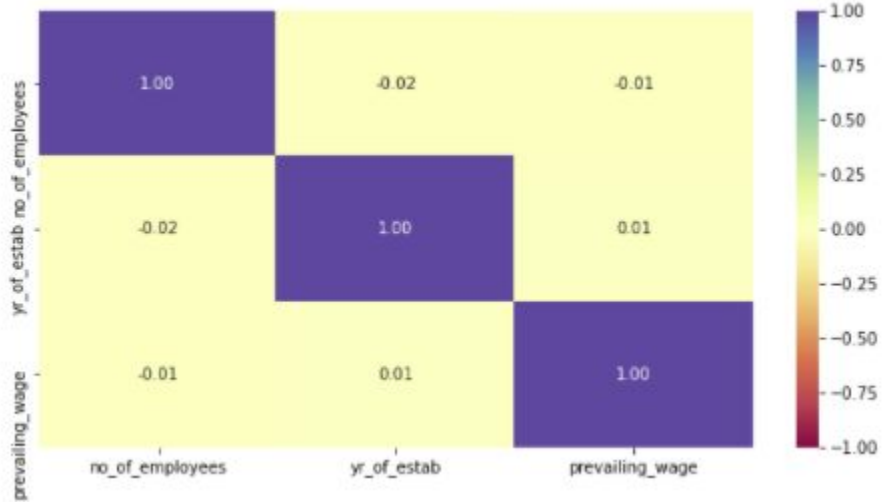
Observations on case status



66.8% of the case status are certified.

Exploratory Data Analysis (EDA)-Bivariate Analysis

Here is the correlation between the numerical variables



There is no correlation between the numerical variables

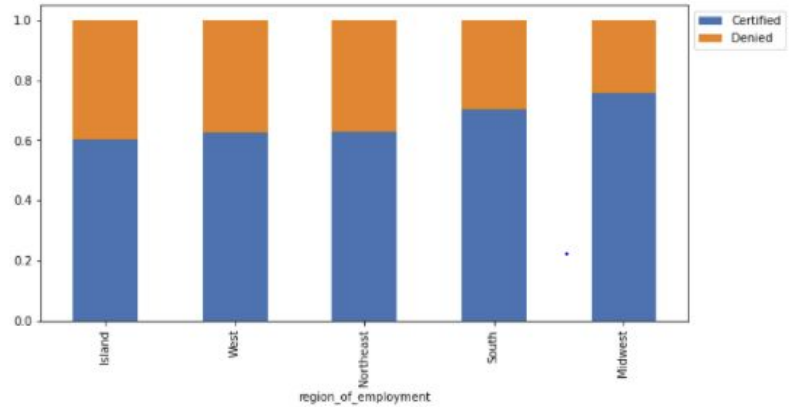
Different regions have different requirements of talent having diverse educational backgrounds

Doctorate degree and West has high correlation (0.33)
Bachelor's degree and West has high correlation (~0.29)
Northeast and South has employers from almost equal educational backgrounds.
Midwest has correlation with High School(~0.22) and Master's (~0.21)

Exploratory Data Analysis (EDA)-Bivariate Analysis

Let's have a look at the percentage of visa certifications across each region

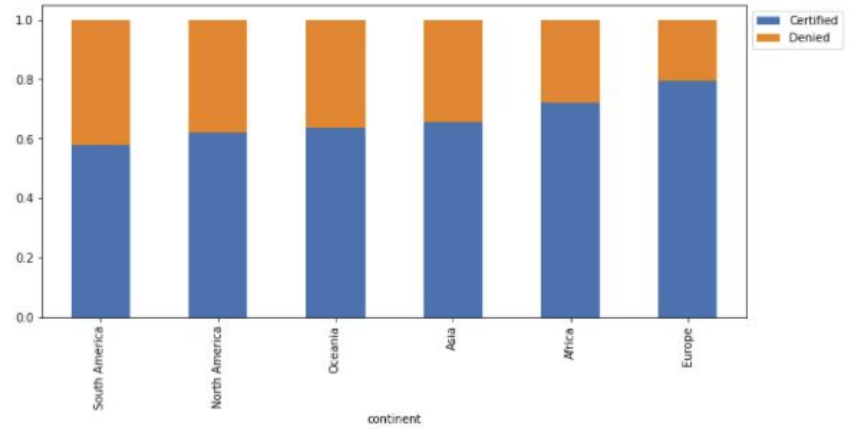
case_status	Certified	Denied	All
region_of_employment			
All	17018	8462	25480
Northeast	4526	2669	7195
West	4100	2486	6586
South	4913	2104	7017
Midwest	3253	1054	4307
Island	226	149	375



Midwest welcomed more employees than other regions as a **percentage**.
However, South has certified more visa applications **in numbers**.

Lets' similarly check for the continents and find out how the visa status vary across different continents

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

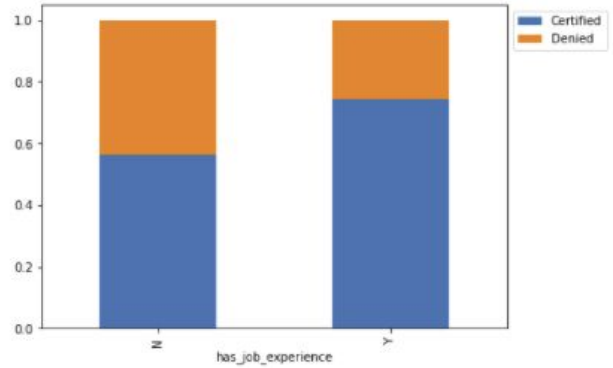


Europe has highest **percentage** on visa certified Africa is following next.
On the other hand, Asia has the largest visa applicants therefore **as a number** it has the largest visa certified status.

Exploratory Data Analysis (EDA)-Bivariate Analysis

Experienced professionals might look abroad for opportunities to improve their lifestyles and career development. Let's see if having work experience has any influence over visa certification

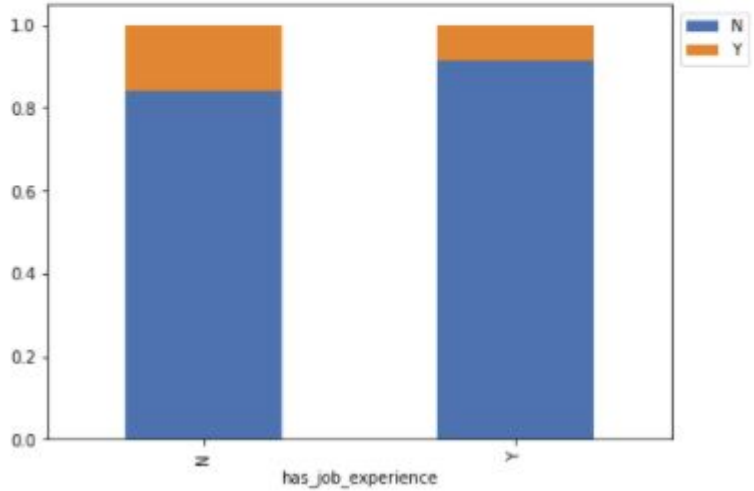
case_status	Certified	Denied	All
has_job_experience			
All	17018	8462	25480
N	5994	4684	10678
Y	11024	3778	14802



As we see on the graph having job experience has a significant effect on the visa certification. Whoever has job experience has much more chance to get visa.

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

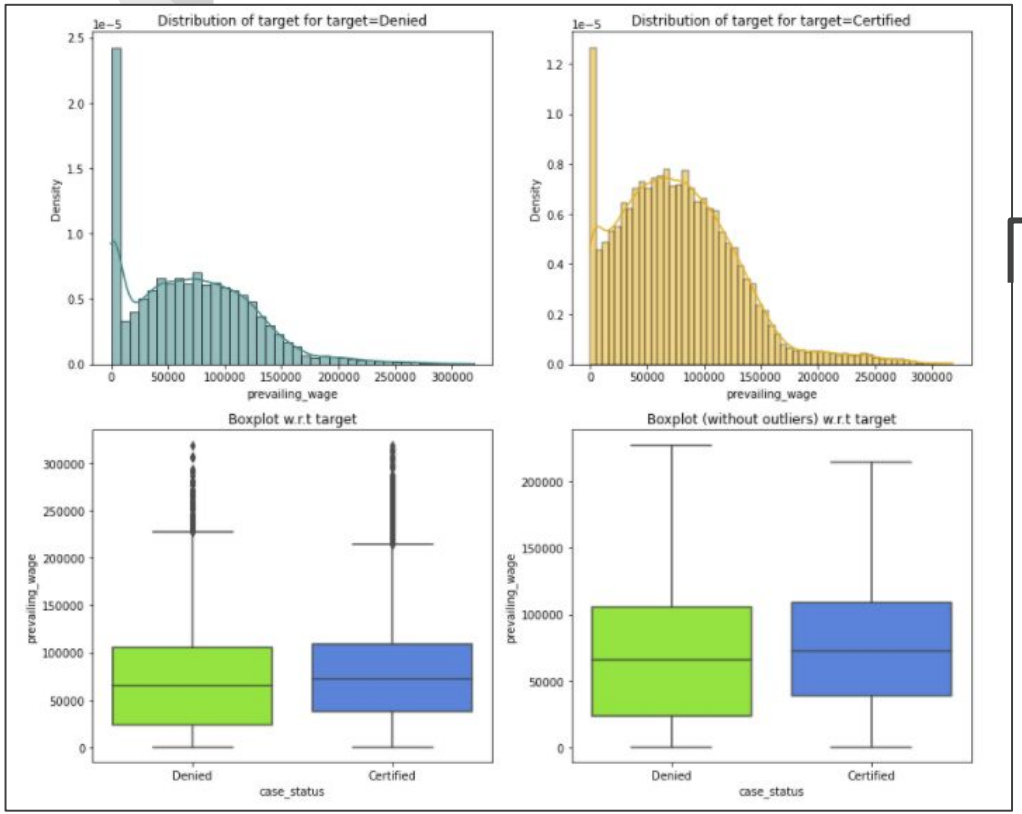
requires_job_training	N	Y	All
has_job_experience			
All	22525	2955	25480
N	8988	1690	10678
Y	13537	1265	14802



Employees who has job experience are not required to have a training(~90%)
On the other hand, in general, job training is not required in most workplaces

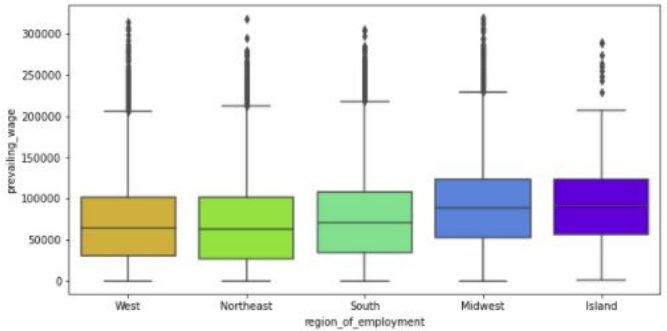
Exploratory Data Analysis (EDA)-Bivariate Analysis

The US government has established a prevailing wage to protect local talent and foreign workers. Let's analyze the data and see if the visa status changes with the prevailing wage



When prevailing wage increase certified status also increase.
Case status and prevailing wage has a normal distribution.
Certified status is slightly higher than denied status.

Checking if the prevailing wage is similar across all the regions of the US

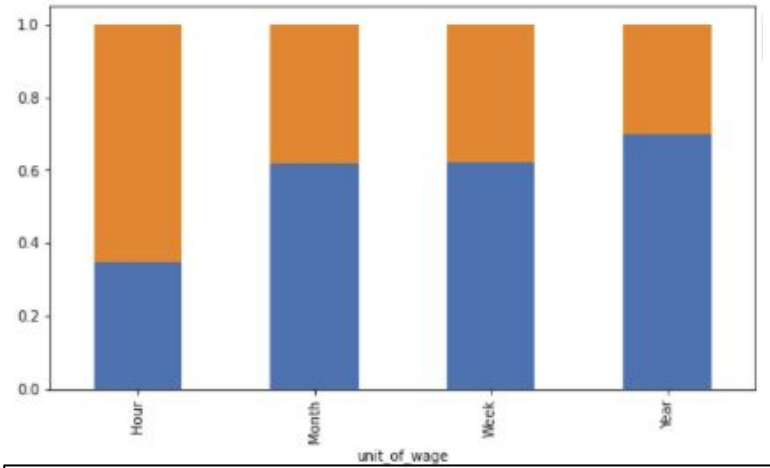


Island and Midwest averages are the highest in the prevailing wage.

Exploratory Data Analysis (EDA)-Bivariate Analysis

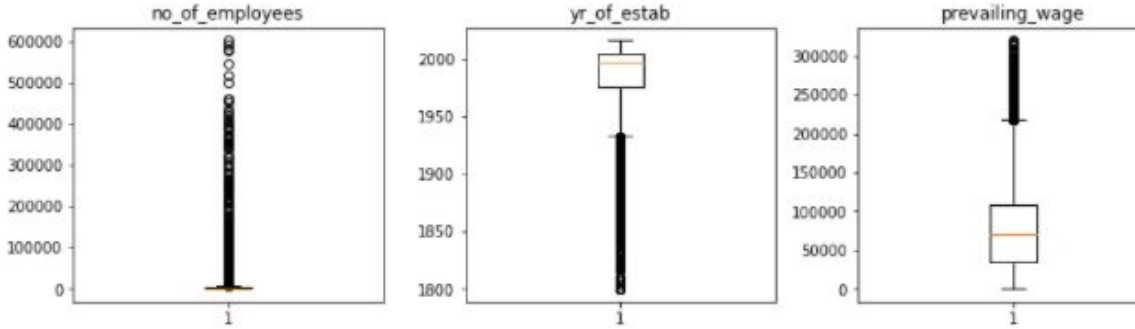
The prevailing wage has different units (Hourly, Weekly, etc). Let's find out if it has any impact on visa applications getting certified.

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



Yearly paychecks are getting certified mostly, it has a significant effect.

Outlier Check

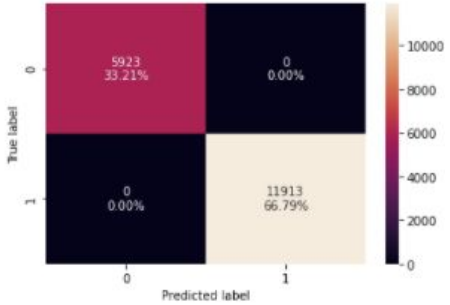


There are outliers ,but we are not treating the them since all are genuine values

Model Performance Summary

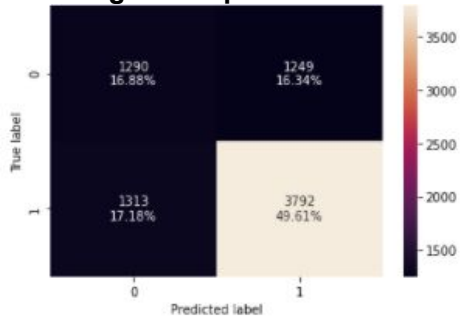
Decision Tree Model

Checking model performance on training set



	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

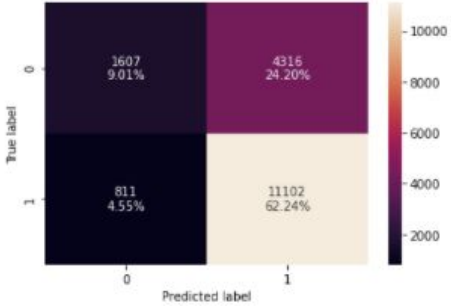
Checking model performance on test



	Accuracy	Recall	Precision	F1
0	0.664835	0.742801	0.752232	0.747487

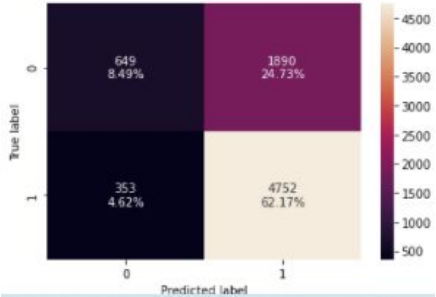
Hyperparameter Tuning - Decision Tree

Checking model performance on training set



	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Checking model performance on test set



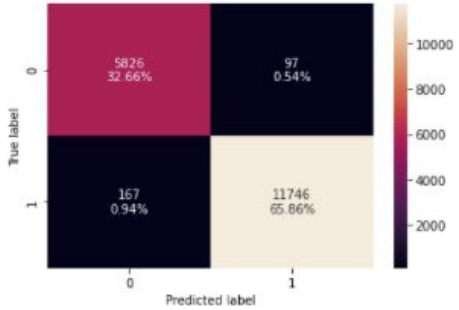
	Accuracy	Recall	Precision	F1
0	0.664835	0.742801	0.752232	0.747487

It is obvious that train data is over fitting.
Model performance did not change after hyperparameter tuning.

Model Performance Summary

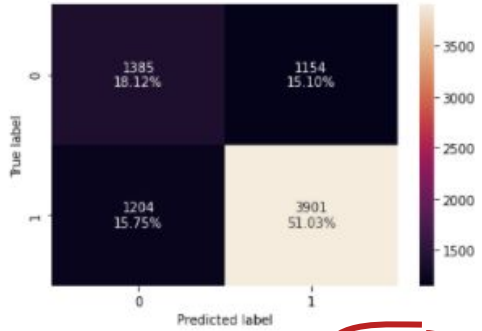
Bagging Classifier

Checking model performance on training set



	Accuracy	Recall	Precision	F1
0	0.985198	0.985982	0.99181	0.988887

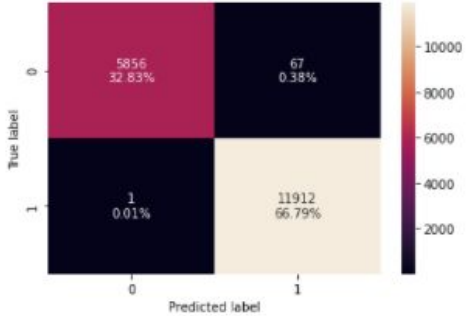
Checking model performance on test set



	Accuracy	Recall	Precision	F1
0	0.691523	0.764153	0.771711	0.767913

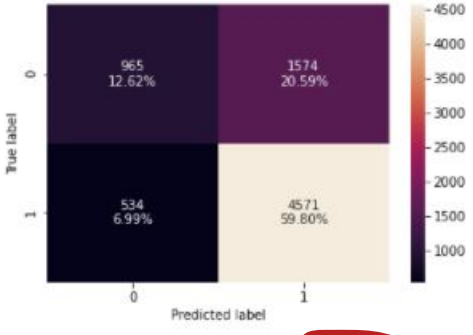
Hyperparameter Tuning - Bagging Classifier

Checking model performance on training set



	Accuracy	Recall	Precision	F1
0	0.996187	0.999916	0.994407	0.997154

Checking model performance on test set



	Accuracy	Recall	Precision	F1
0	0.724228	0.895397	0.743857	0.812622

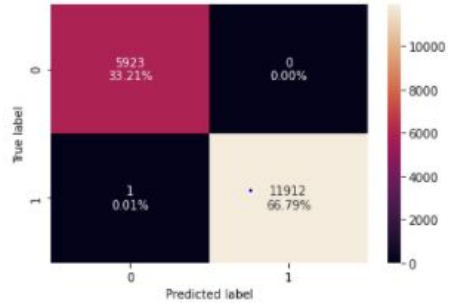
Training performance is overfitting on bagging classifier before and after hyperparameter tuning.

Test performance got better after hyperparameter tuning

Model Performance Summary

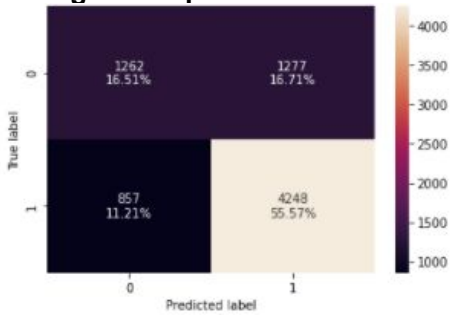
Random Forest

Checking model performance on training set



	Accuracy	Recall	Precision	F1
0	0.999944	0.999916	1.0	0.999958

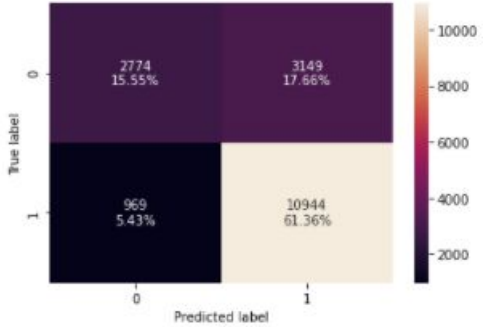
Checking model performance on test set



	Accuracy	Recall	Precision	F1
0	0.720827	0.832125	0.768869	0.799247

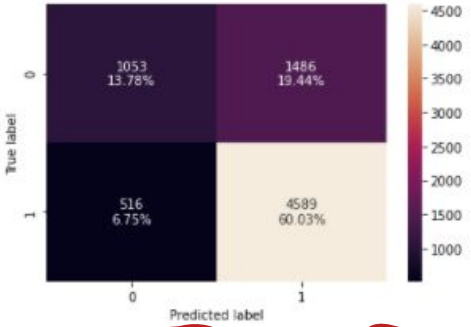
Hyperparameter Tuning - Random Forest

Checking model performance on training set



	Accuracy	Recall	Precision	F1
0	0.769119	0.91866	0.776556	0.841652

Checking model performance on test set



	Accuracy	Recall	Precision	F1
0	0.738095	0.898923	0.755391	0.82093

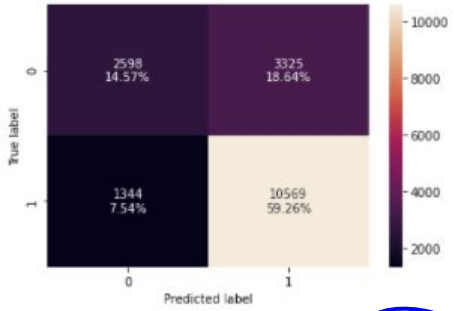
Random Forest Training performance was overfitting, it got better after hyperparameter tuning.

F1 score and Recall improved after hyperparameter tuning.

Model Performance Summary

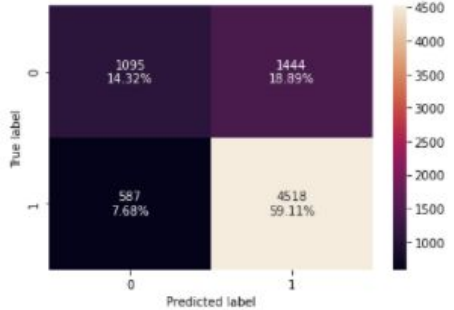
AdaBoost Classifier

Checking model performance on training set



	Accuracy	Recall	Precision	F1
0	0.738226	0.887182	0.760688	0.81908

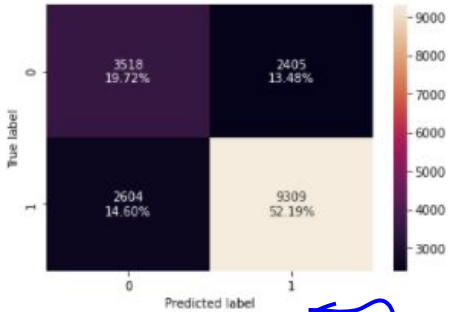
Checking model performance on test set



	Accuracy	Recall	Precision	F1
0	0.734301	0.885015	0.757796	0.816481

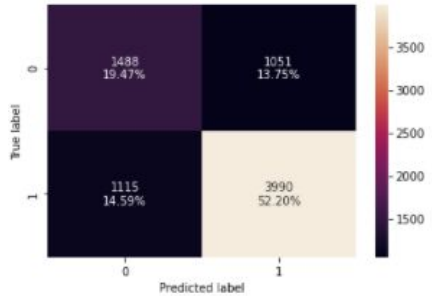
Hyperparameter Tuning - AdaBoost Classifier

Checking model performance on training set



	Accuracy	Recall	Precision	F1
0	0.719163	0.781415	0.79469	0.787997

Checking model performance on test set



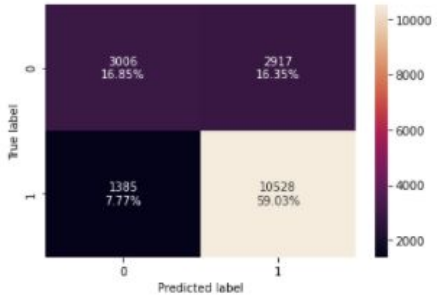
	Accuracy	Recall	Precision	F1
0	0.716641	0.781587	0.79151	0.786517

Training and testing data F1 scores are very close to each other in Ada Boost Clasifier, it means the model is getting better.
However, Hyperparameter tuning lowered the F1 scores on training and testing.

Model Performance Summary

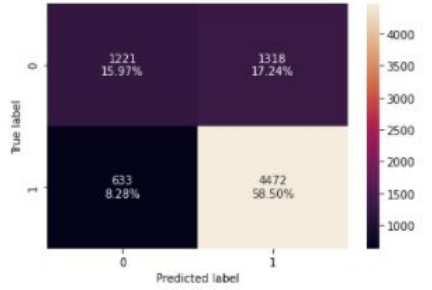
Gradient Boosting Classifier

Checking model performance on training set



	Accuracy	Recall	Precision	F1
0	0.758802	0.88374	0.783042	0.830349

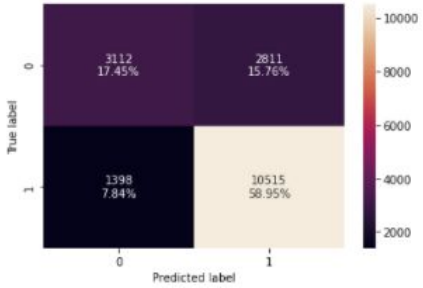
Checking model performance on test set



	Accuracy	Recall	Precision	F1
0	0.744767	0.876004	0.772366	0.820927

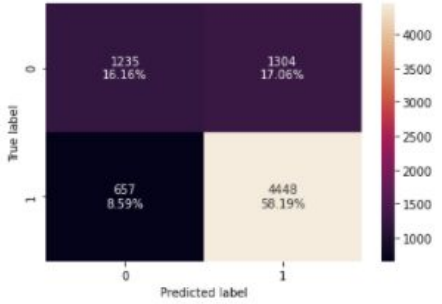
Hyperparameter Tuning-Gradient Boosting Classifier

Checking model performance on training set



	Accuracy	Recall	Precision	F1
0	0.764017	0.882649	0.789059	0.833234

Checking model performance on test set



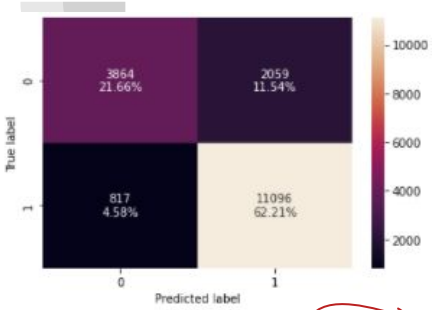
	Accuracy	Recall	Precision	F1
0	0.743459	0.871303	0.773296	0.819379

Like AdaBoost, Gradient Boosting training and testing data F1 scores are very close to each other.
Hyperparameter tuning did not affect much on the F1 score in training data, but caused a drop on F1 score in the testing data.

Model Performance Summary

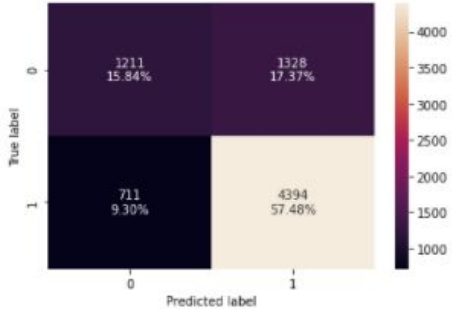
XGBoost Classifier

Checking model performance on training set



	Accuracy	Recall	Precision	F1
0	0.838753	0.931419	0.843482	0.885272

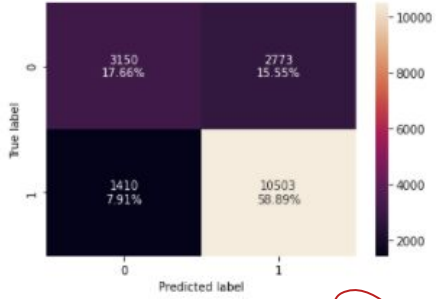
Checking model performance on test set



	Accuracy	Recall	Precision	F1
0	0.733255	0.860725	0.767913	0.811675

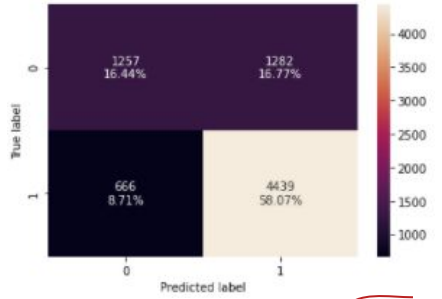
Hyperparameter Tuning - XGBoost Classifier

Checking model performance on training set



	Accuracy	Recall	Precision	F1
0	0.765474	0.881642	0.791127	0.833935

Checking model performance on test set



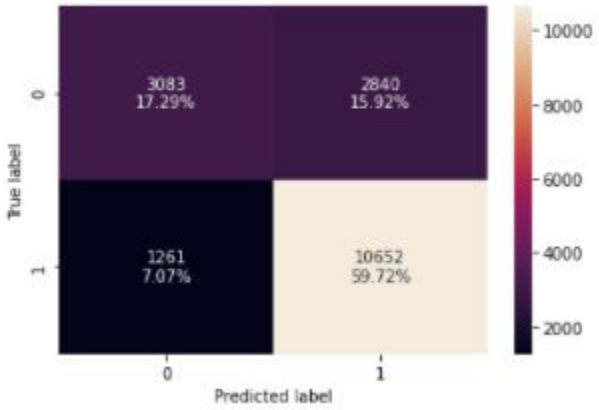
	Accuracy	Recall	Precision	F1
0	0.74516	0.86954	0.775913	0.820063

XGBoost Classifier training and testing values are not close to each other. However after hyperparameter tuning we got very close scores. The model will look good with this effect.

Model Performance Summary

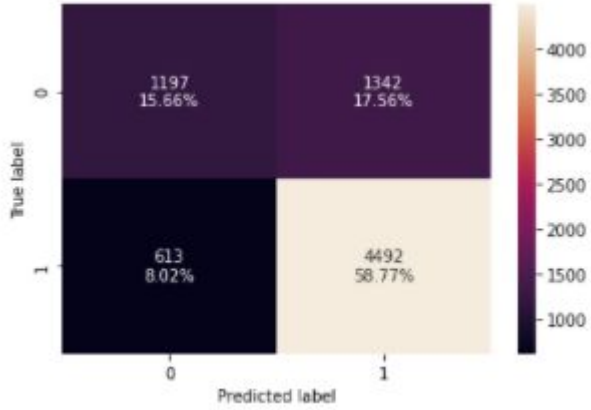
Stacking Classifier

Checking model performance on training set



	Accuracy	Recall	Precision	F1
0	0.770072	0.894149	0.789505	0.838575

Checking model performance on test set



	Accuracy	Recall	Precision	F1
0	0.744244	0.879922	0.769969	0.821282

F1 scores and all the other scores are very close to each other.

Conclusion

Comparing all models

Training performance comparison:

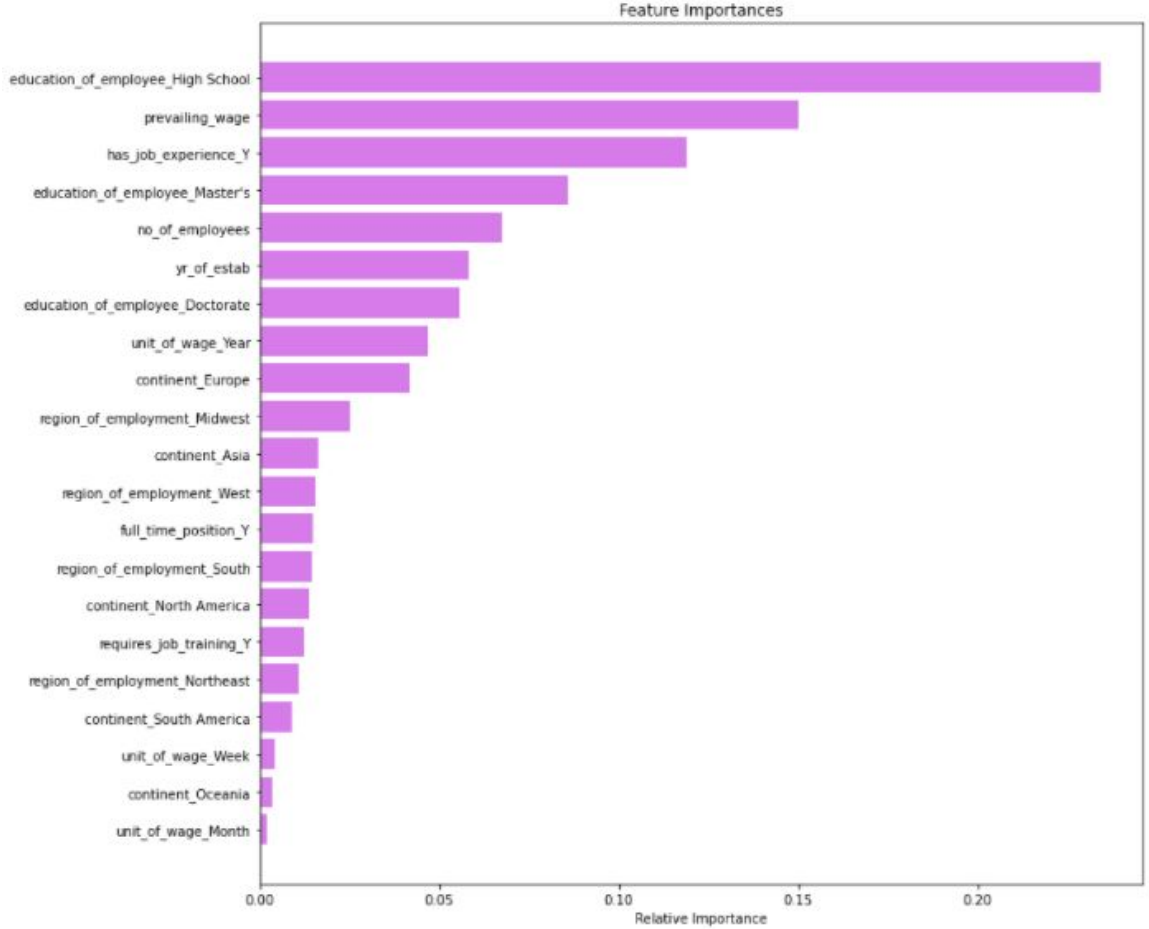
	Decision Tree	Tuned Decision Tree	Bagging Classifier	Tuned Bagging Classifier	Random Forest	Tuned Random Forest	Adaboost Classifier	Tuned Adaboost Classifier	Gradient Boost Classifier	Tuned Gradient Boost Classifier	XGBoost Classifier	XGBoost Classifier Tuned	Stacking Classifier
Accuracy	1.0	1.0	0.985198	0.996187	0.999944	0.769119	0.738226	0.719163	0.758802	0.764017	0.838753	0.765474	0.770072
Recall	1.0	1.0	0.985982	0.999916	0.999916	0.918660	0.887182	0.781415	0.883740	0.882649	0.931419	0.881642	0.894149
Precision	1.0	1.0	0.991810	0.994407	1.000000	0.776556	0.760688	0.794690	0.783042	0.789059	0.843482	0.791127	0.789505
F1	1.0	1.0	0.988887	0.997154	0.999958	0.841652	0.819080	0.787997	0.830349	0.833234	0.885272	0.833935	0.838575

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	0.664835	0.664835	0.691523	0.724228	0.720827	0.738095	0.734301	0.716641	0.744767	0.743459	0.733255	0.745160	0.744244
Recall	0.742801	0.742801	0.764153	0.895397	0.832125	0.898923	0.885015	0.781587	0.876004	0.871303	0.860725	0.869540	0.879922
Precision	0.752232	0.752232	0.771711	0.743857	0.768869	0.755391	0.757799	0.791510	0.772366	0.773296	0.767913	0.775913	0.769969
F1	0.747487	0.747487	0.767913	0.812622	0.799247	0.820930	0.816481	0.786517	0.820927	0.819379	0.811675	0.820063	0.821282

All the models after Random Tuned Forest Classifier has very close F1 scores in both train and testing performance. I will use Stacking Classifier for my model.

Important features of the final model



As we see on the left, most important feature of my model is education_of_employee_High School , prevailing_wage, has_job_expeience, and eduacation_of_employee_Master's respectively.

Recommendations

- Education type High School visa candidate should apply for weekly paid jobs to get visa certification especially in Midwest. Asia and South American's who has only high school education will have high chance on getting visa. Midwest also getting equal amount Master Degree candidates.
- Education type Masters and Doctoral degree will get more certified visa in Northeast and West. South region follows them next. Candidates who are applying from Europe, Asia and North America (respectively) have higher chance to get certified visa. They are also looking for yearly wage on application.
- Bachelor's degree candidates in West has high chance.
- Northeast and South has employers from almost equal educational backgrounds.
- Midwest welcomed more employees than other regions as a percentage. However, South has certified more visa applications in numbers.
- Europe has highest percentage on visa certified Africa is following next. On the other hand, Asia has the largest visa applicants therefore as a number it has the largest visa certified status.
- Prevailing wage is the most important factor for each education type to get certified visa. Applicants wants to ensure that the foreign worker is not underpaid compared to other workers offering the same or similar service in the same area of employment.
- Having a job experience is a big plus. Whoever has job experience has much more chance to get visa.
- Employees who has job experience are not required to have a training(~90%).Some companies may require job training, therefore applicants should consider that before they apply.
- Companies who has large number of employees will likely to get more qualified candidates.
- Full time positions are most preferable by the applicants.