Easy Visa Project

DSBA

09/22/2023

Contents / Agenda

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

Executive Summary

- **Recommendations:-**

Considering how easy it is to use and how easy it can be to understand, OFLC recommends the Tuned decision tree model as the ultimate classifier.

If you want to reduce the bias and use an ensemble model, the Tuned gradient boosting model is a good option. OFLC will take into account the applicant's education level, job experience, and prevailing wage unit when estimating their visa certification chances.

People with higher education, experience in the US, and the wage unit of their US job are more likely to get a work visa eventually. Being from Europe can also increase the chances of getting a visa certification in some cases.

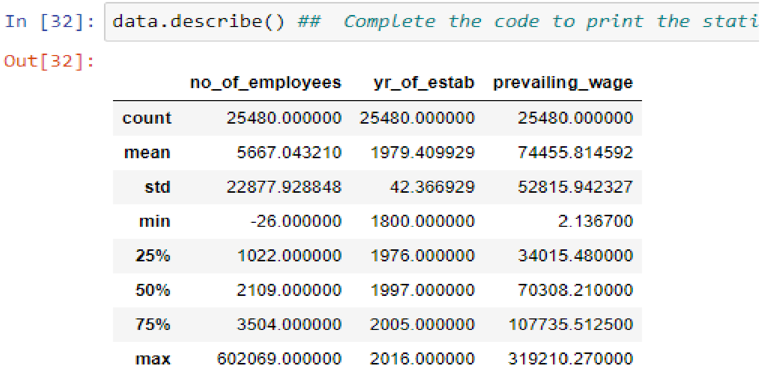
To make sure the US doesn't have a shortage of skilled workers, especially in industries that rely on foreign employees, OFLC will prioritize visa applications that have a higher chance of being certified based on the classification models.

Business Problem Overview and Solution Approach

- 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.
- 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:-

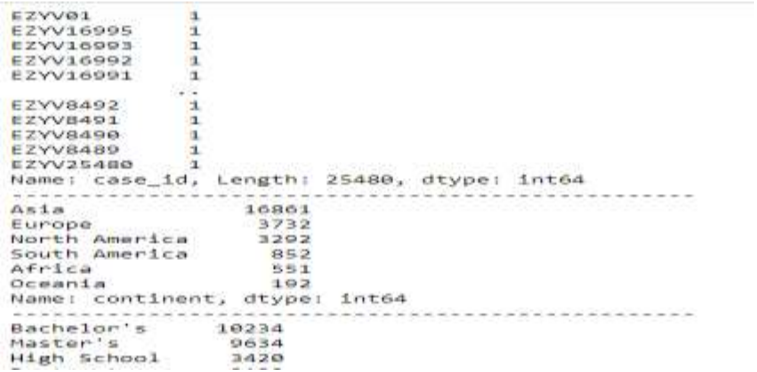
Facilitate the process of visa approvals.
Recommend a suitable profile for the applicants for whom the visa should be certified or denied based on the drivers that significantly influence the case status.

EDA Results



```
In [32]: data.describe() ##  Complete the code to print the stati
Out[32]:
```

	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

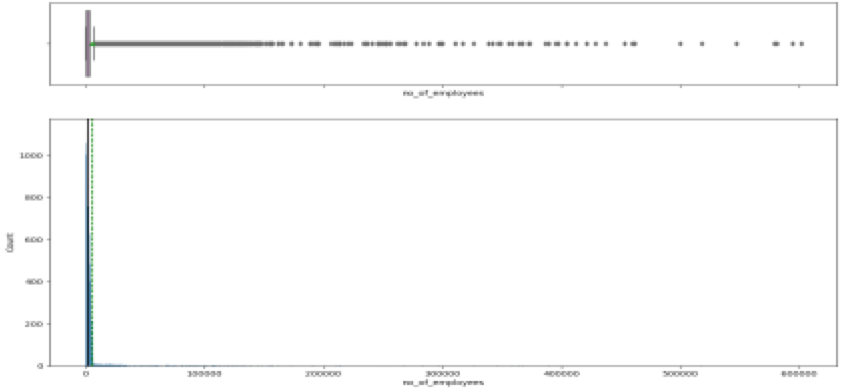


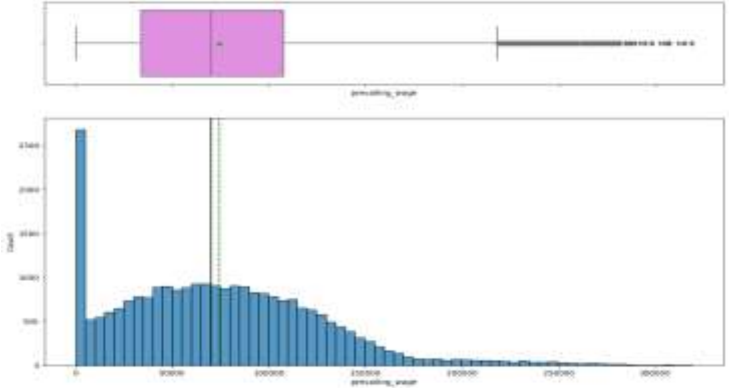
```
EZYV01          1
EZYV16995       1
EZYV16993       1
EZYV16992       1
EZYV16991       1
               ..
EZYV8492        1
EZYV8491        1
EZYV8490        1
EZYV8489        1
EZYV25480       1
Name: case_id, Length: 25480, dtype: int64
--------------------------------------------------
Asia                 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
```

Observation:-

- The data shows the statistical summary of data which includes mean value of employees, yr. of establishment , and prevailing wage rate.
- Majority of employees are form Asia and also have bachelors degree and also have job experience.
- The majority of the applications are full time positions.
- Nearly about 2/3 rd. of applications are certified.
- Majority of jobs do not require job training.

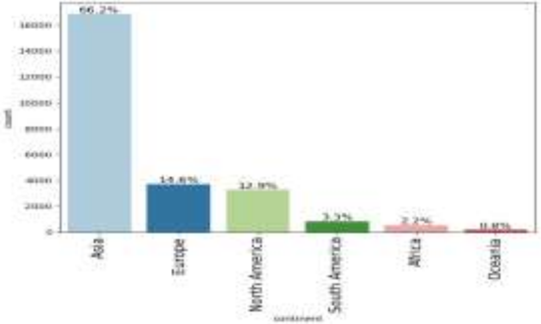
Univariate Analysis

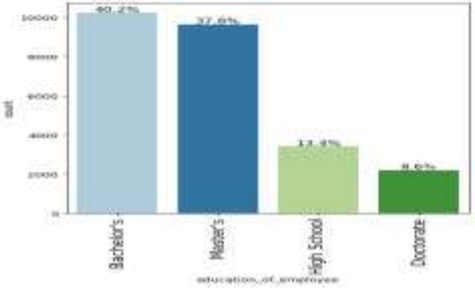


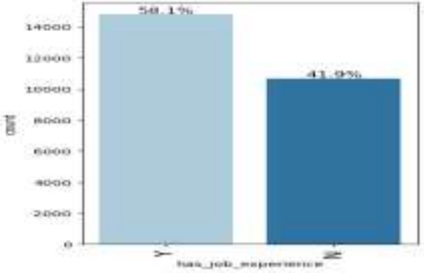


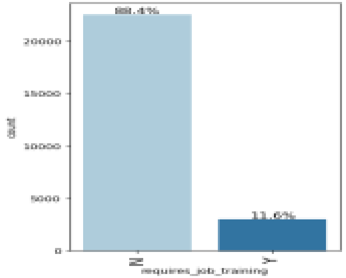
- The above graphs shows are of prevailing wage rate and no of employees.
- The majority of the applications are for the jobs whose prevailing wages are computed per year.
- There is a large variation in number of employees.
- The distribution is highly rightly skewed.
- Because there were firms with hundreds of thousands of employees in the United States in 2016, not all outliers that were discovered using the 1.5-IQR criteria had to be considered as outliers. According to the distribution displayed in this case, a cut-off value of 450000 is taken into account for the workforce size.

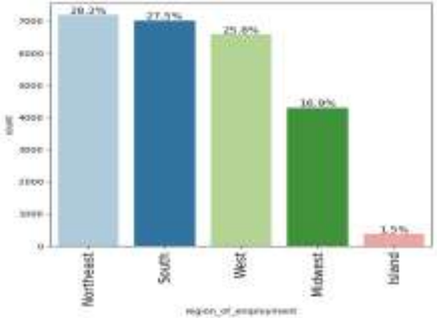
Univariate Analysis

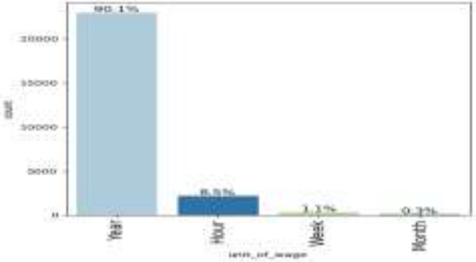


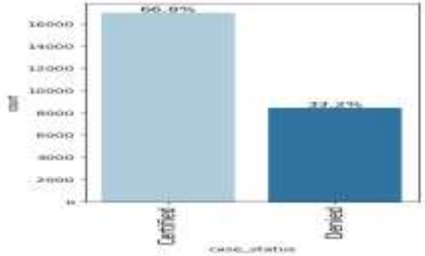








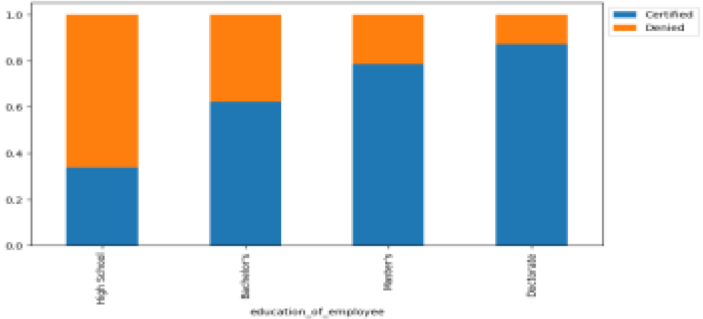




- This all graphs as per every data point it is represented in graph which shows how much we require and what is good and bad for every data point.

Bivariate Analysis

- **Those with higher education may want to travel abroad for a well-paid job. Let's find out if education has any impact on visa certification. Case Study v/s Education.**

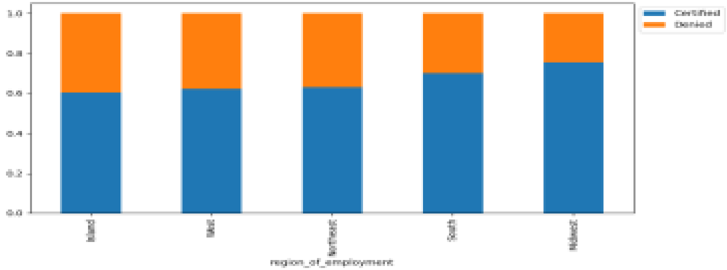


Observation:-

- It is obvious that an applicant's chances of receiving a visa are increased by their level of education.
- More particular, whereas the likelihood of visa approval for candidates with a PhD degree is 87%, it is only 34% for applicants with only a high school diploma.

Bivariate Analysis

- **Let's have a look at the percentage of visa certifications across each region.Case study v/s emplyoement.**

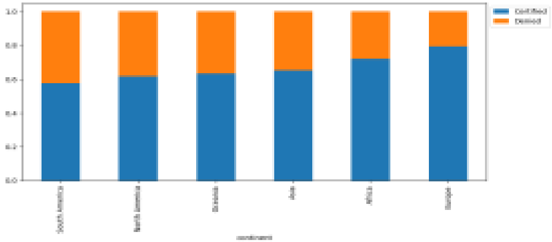


Observation:-

- The Northeast, South, and West of the United States receive the majority of the applicants for jobs. Given that the majority of tech businesses are located in those locations and that those regions have higher populations than other regions of the United States, this might be predicted.
- The least amount of applicants (1.5%) are from the Island region.

Bivariate Analysis

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

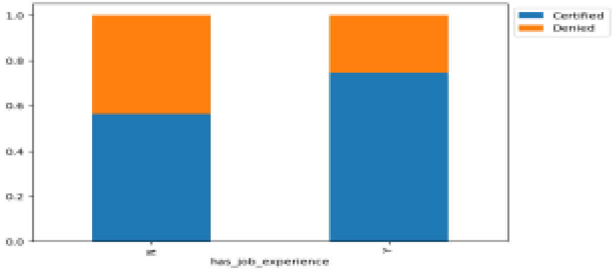


Observation:-

- Given that Asia has a large population, it makes reasonable that the majority (66%) of visa applicants are from this region.
- Oceania accounts for the smallest percentage of applicants (1%), which is also reasonable given its extremely tiny population.
- There are nearly as many candidates from North America and Europe (12.9% and 14.6%).

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.

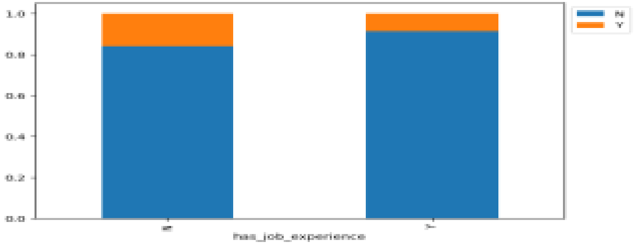


Observation:-

- More than half of the applicants have job experience.

Bivariate Analysis

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

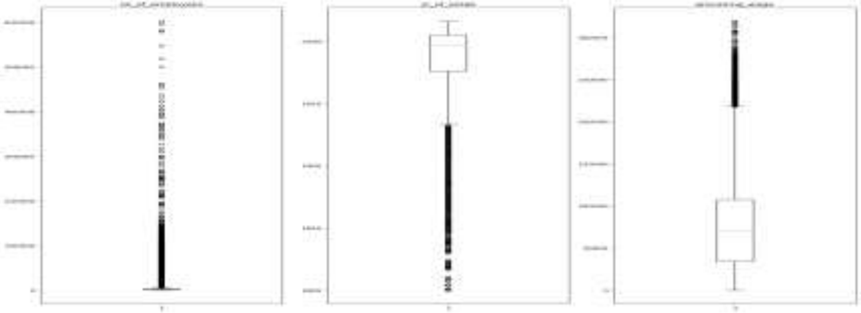


Overall observations:-

- Apart from the above analysis of the ratio of applications approved to denied based on education, prior job experience, region, prevailing wage and wage unit, whether or not the job requires training does not impact this ratio. If the job is full-time or not, the ratio does not change. If the job is in a Midwestern or Southern region, applicants tend to have a higher approval rate than applicants in the Northeastern, Western and especially Island regions. The jobs that candidates with doctorate degrees are applying to have the lowest median prevailing wage of all education. The temporary jobs that candidates are applying to have much higher prevailing wages than the full-time jobs.

Data Preprocessing

- **Outlier check (treatment if needed):-**

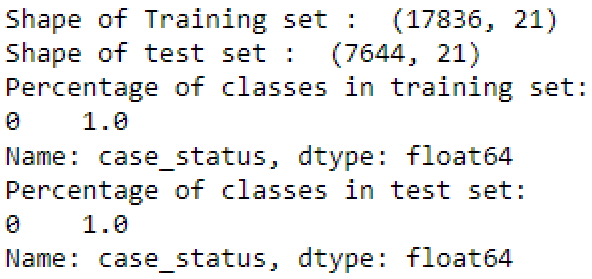


- Observation:-

It's important to note that not all of the "outliers" found in the first EDA section are actually real outliers. Just for the sake of completeness, here are the maximum cut-offs for the three variables.
No_of_Employees = 450000, SNC_Estab = 200, Hourly wage = 4000. If there aren't any outliers, the maximum values in the relevant columns will be used to treat them.

Data Preprocessing

- **Data Preparation for modeling:-**



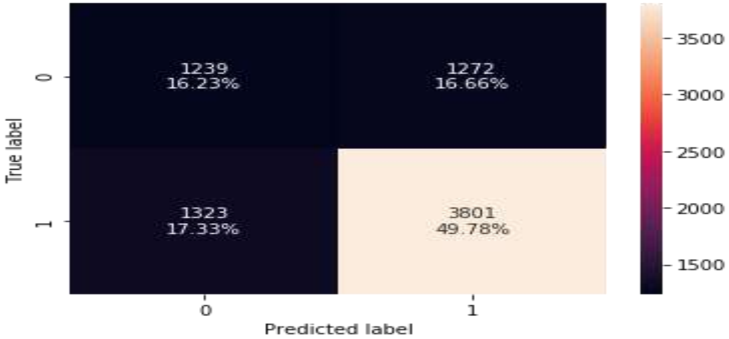
```
Shape of Training set :  (17836, 21)
Shape of test set :  (7644, 21)
Percentage of classes in training set:
0    1.0
Name: case_status, dtype: float64
Percentage of classes in test set:
0    1.0
Name: case_status, dtype: float64
```

- **Observation:-**

Here we can see that before making a decision tree it defines a shape of training and test set.

Decision Tree - Model Building and Hyperparameter Tuning

Confusion matrix on testing of data includes various points:-



- Decision tree is working well on training data but not able to get it on test data so it shows that it is over fitted.

  - Applicant was approved and the model predicted approval : True Positive
  - Applicant was denied and the model predicted approval : False Positive
  - Applicant was denied and the model predicted denial : True Negative
  - Applicant was approved and the model predicted denial : False Negative

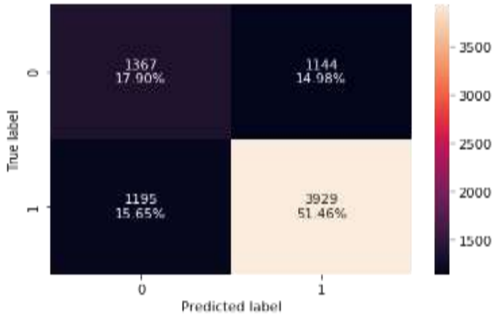
Bagging - Model Building and Hyperparameter Tuning

Confusion matrix on testing data:-

- Training performance Accuracy Recall Precision F1

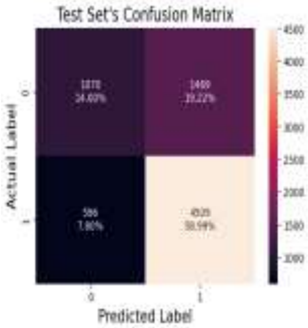
  0 0.98 0.98 0.99 0.99
- Testing performance Accuracy Recall Precision F1

  0 0.69 0.77 0.77 0.77



Observations:

- If we compare with initial model, this model is slightly better performance on test data.
- On the training data it shows a high performance so from this it shows that model is over fitting and needs hyper parameter tuning.
- After the hyperparameter tuning of data it shows that the tuned model performance is slightly better than bagging model-the F1 score is also increased form 0.77 to 0.81.

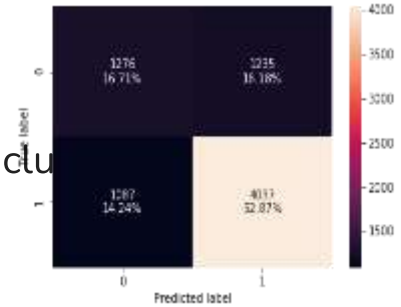


Bagging - Model Building and Hyperparameter Tuning

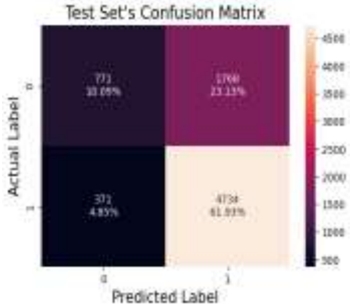
Random Forest and hyperparameter tuning of random forest:-

- This model is slightly better on performance on the test data.
- However, it shows the metrics equal to 1 for training of data set which conclu

that it is overfitting the data. So this also needs a hyperparameter tuning.



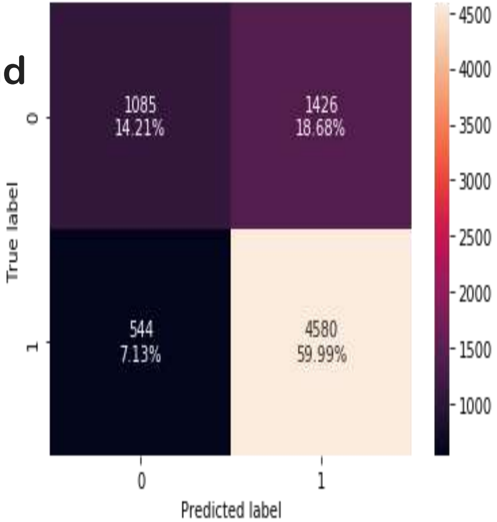
- The performance matrix are very close for training and test data so after hyperparameter tuning the model is not overfitting.
- If we try to compare with random forest precision has decreased and F-1 score have increased.



Bagging - Model Building and Hyperparameter Tuning

AdaBoost Classifier and hyperparameter tuning of it:-

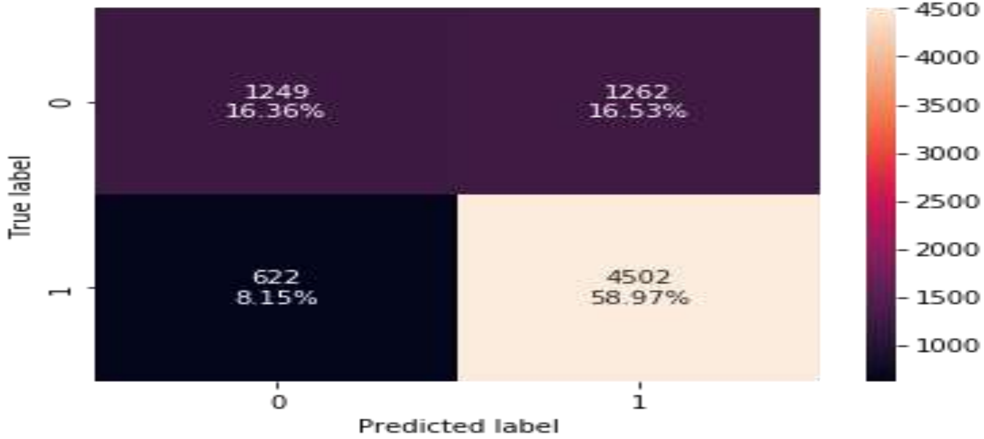
- **It shows that model have the performance metrics for training and test data are very close.**
- **But even need a hyperparameter tuning that helps in overall model performance.**
- **After hyperparameter also there is no change.**



Bagging - Model Building and Hyperparameter Tuning

Gradient Boosting Classifier:-

- This Gradient Boosting classifier performs equally well on the training and testing datasets, indicating that the model is not overfit.
- Additionally, this model performs even better than the AdaBoost model above.



Great Learning

Happy Learning !