Business Presentation

Contents

- Business Problem Overview and Solution Approach
- Data Overview
- EDA
- Bagging and Boosting Model
- Model Performance Summary
- Model Performance Comparison and Conclusions
- Business Insights and Recommendations

Business Problem Overview and Solution Approach

Core business idea

- The increasing number of applications for Visa approval every year is becoming a tedious task. Idea is to shortlist the candidates having higher chances of VISA approval.

Problem to tackle

- Identify deficiencies in current target segmentation.
- Indentifying potential cancelations

Financial implications

- Reduce resource wastage and improve efficiency

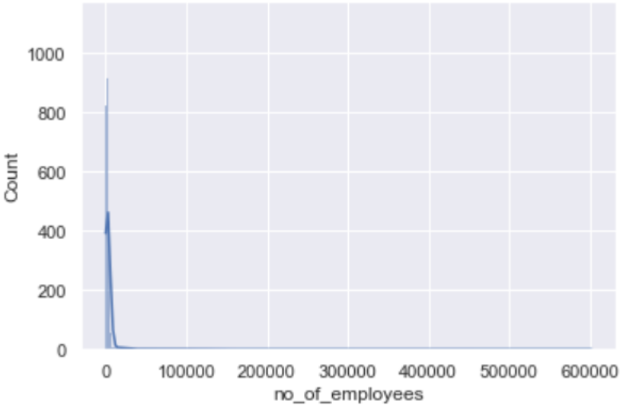
How to use ML model to solve the problem

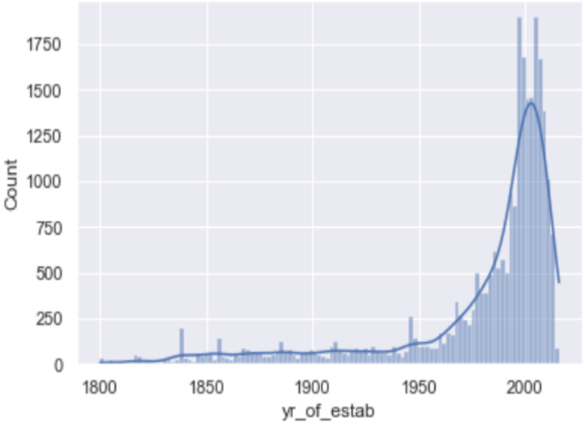
- Decision Trees, Bagging and Boosting Model techniques have been used to predict the model that can be implemented to predict the applications that are likely to be denied.

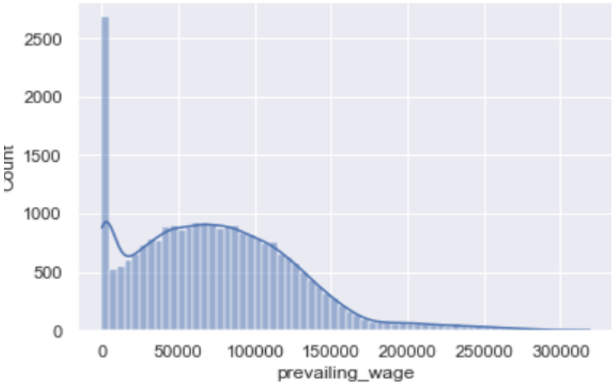
Data Overview

- The data contains information about 25480 Visa applicants and their details.
- The characteristics include has_job_experience, requires_job_training, full_time_position, case_status, region_of_employee and many more. There are 11 such characteristics.
- Object Variables 'continent', 'education_of_employee', 'region_of_employment', 'unit_of_wage’ have been converted to Categorical Variables.
- There are no duplicate rows.
- There are no missing values.
- Factors like full time position, unit of wage, region of employment and many more can affect the rejections.

EDA – no_of_employees, yr_of_estab,prevailing_wage

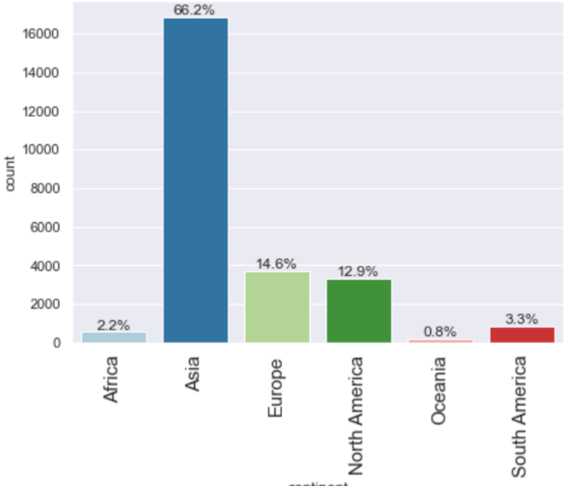


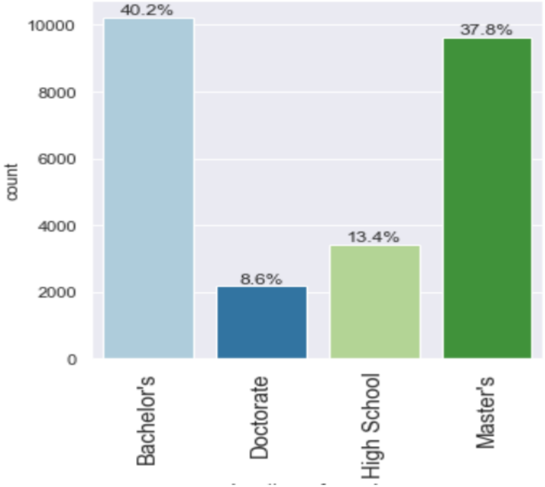


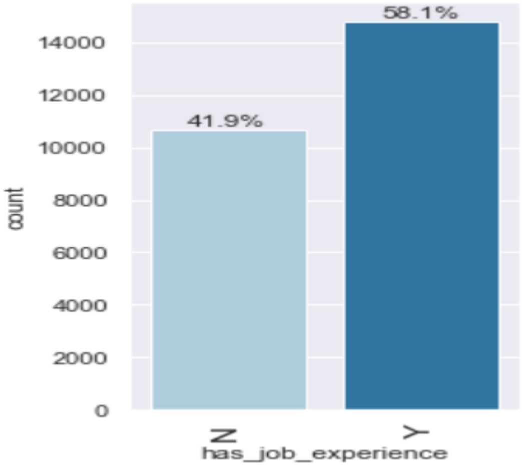


- Most of the company's employee count is less than 20,000 Although there are companies where the employee count is 600000.
- Most of the companies are established between 1980 and 2000.
- Most of the applicants have high prevailing wages.

EDA – Categorical Variable - Continent, education_of_employee, has_job_experience

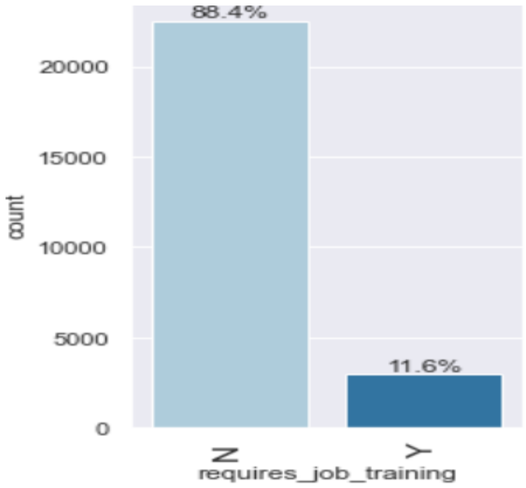


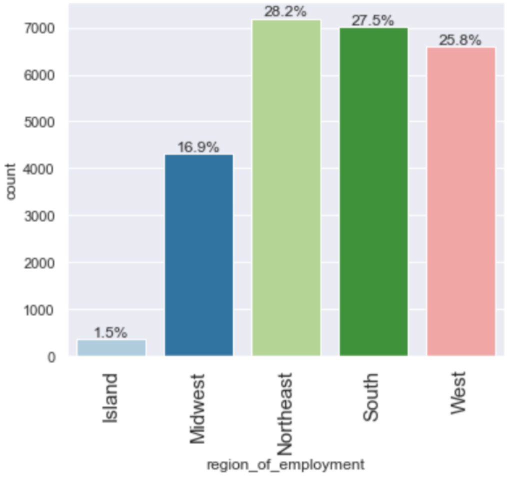


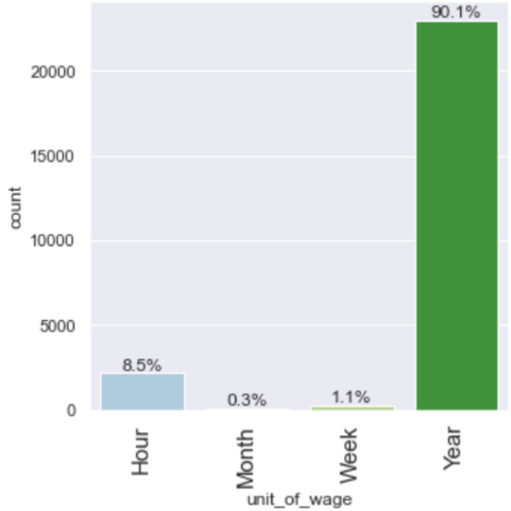


1. Most of the employees are from Asia. Very less number of employees are from Oceania.
2. 40.2% of the employees have Bachelors degree. 37.8% of the employees have Master's degree.
3. 58.1% of the applicants have job experience.

EDA – requires_job_training, region_of_employment, unit_of_wage

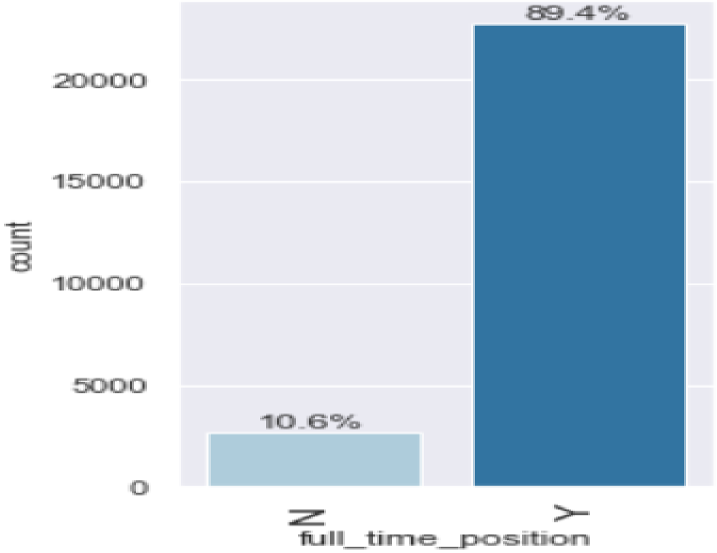


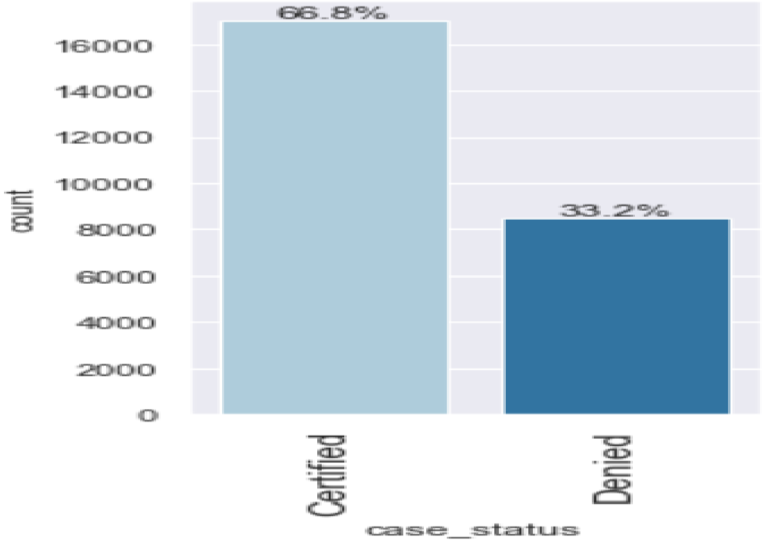




- 88.4% of the applicants do not require job training
- Very less(1.5%) of the applicants are from Island. Most of the employees are from Northeast, South and West regions.
- 90.1% of the employees receive year wages. 8.5% of the employees receive Hour wages.

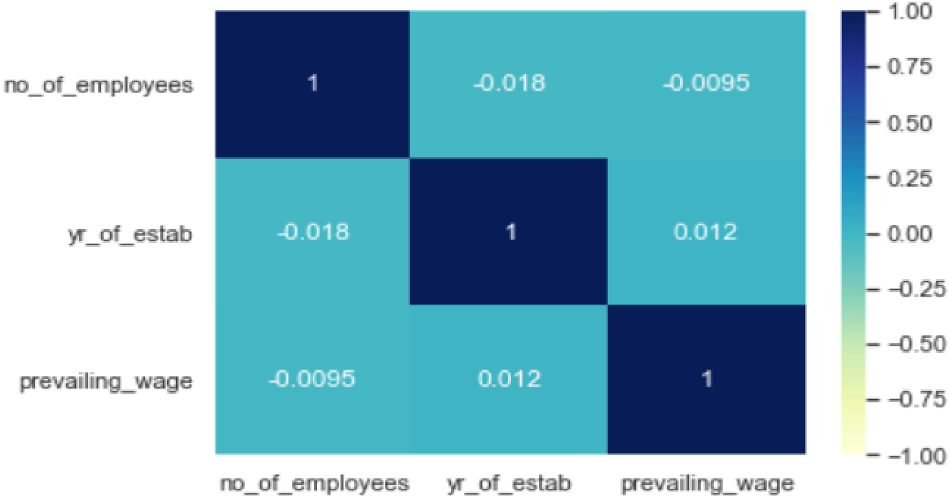
EDA – full_time_position, case_status





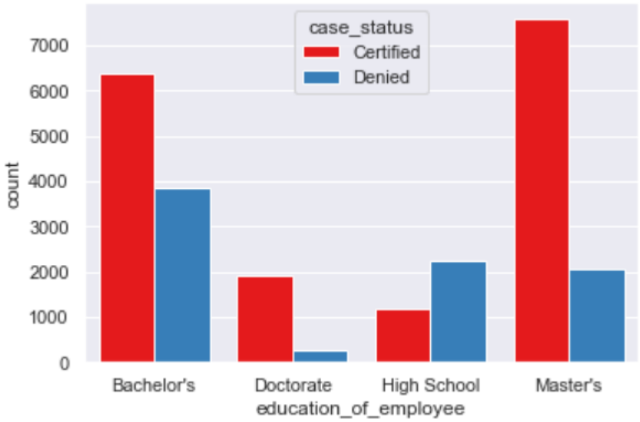
- 89.4% of the employees are in full time position. Rest are not.
- 66.8% of the cases are Certified.

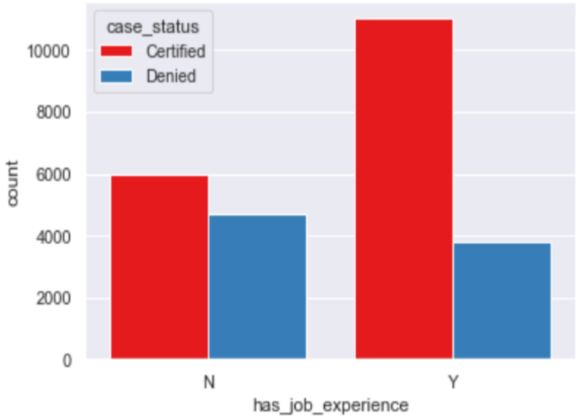
Bivariate Analysis – Heat Map

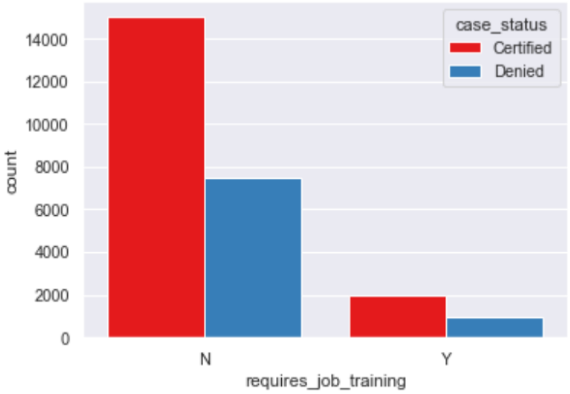


1. Highest correlation from the above heatmap is between year of establishment and number of employees (0.018) which is very less.
2. It is important to note that correlation does not imply causation.
3. Least correlation is between number of employees and prevailing wage

EDA – continent , education_of_employee, has_job_experience vs case_status

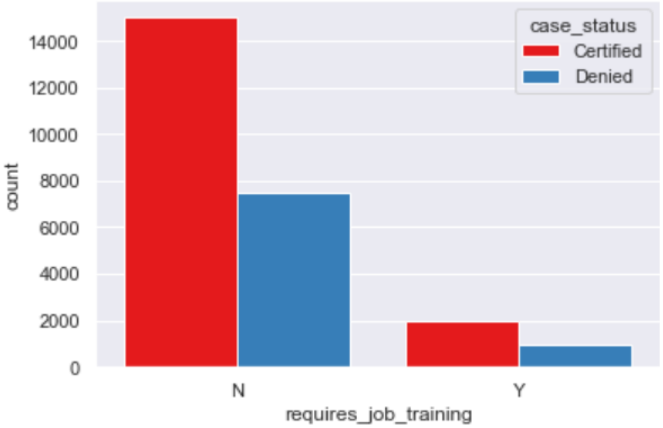


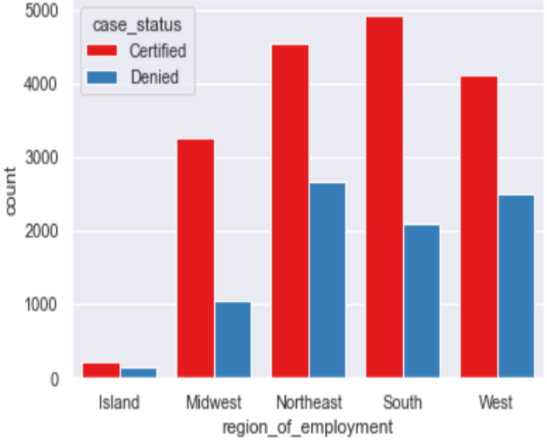


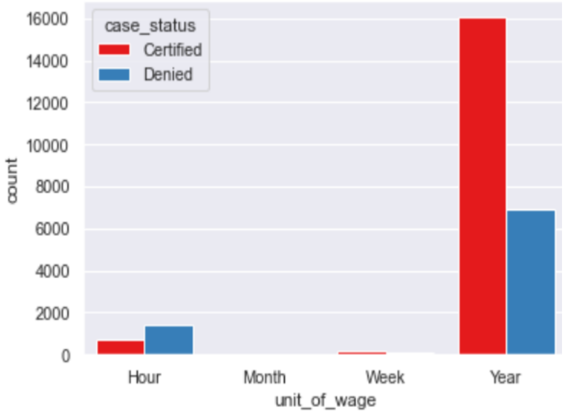


- Most of the applicants whose case status is certified have Master's followed by candidates with Bachelor's degree.
- Most of the applicants whose case status is certified have job experience.
- Most of the candidates whose case is certified are those who did not require job training.

EDA – **region_of_employment, unit_of_wage ,** full_time_position vs case_status

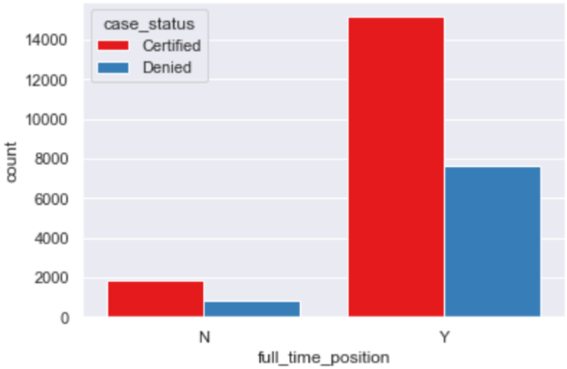


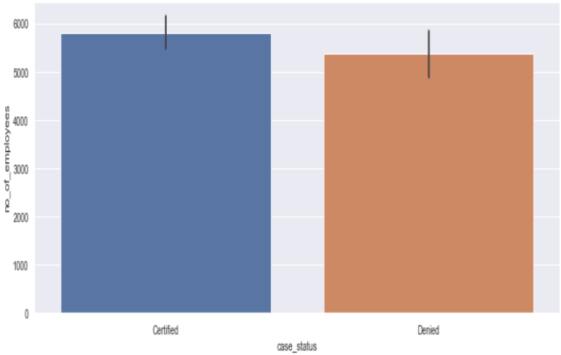


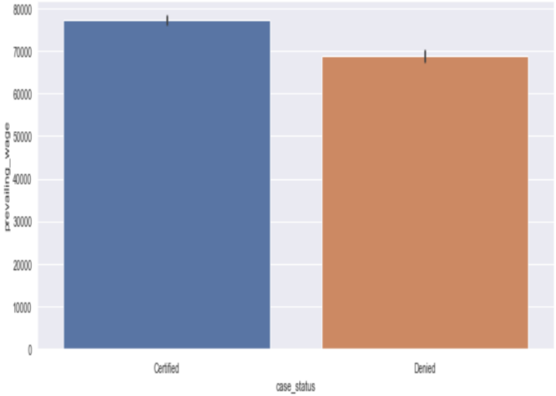


1. Most of the candidates whose case is certified are those who did not require job training.
2. Most of the certified applicantion are from south region followed by Northeast and West.
3. Most of the employees receive yearly wages. The maximum cases are Certified among Yearly waged applicants.

EDA – full_time_position, no_of_employees, prevailing_wages vs case_status



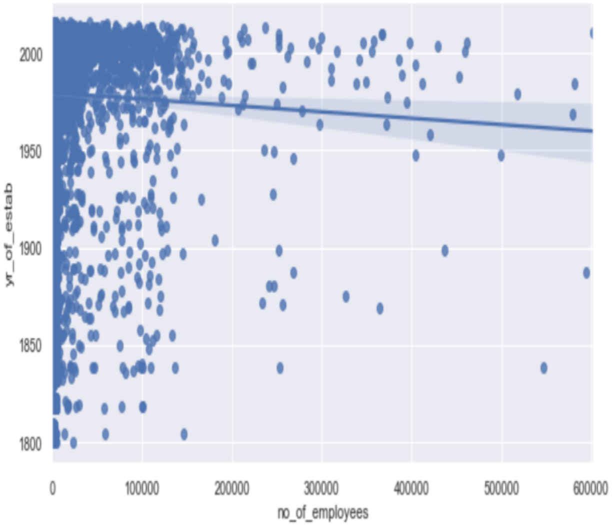


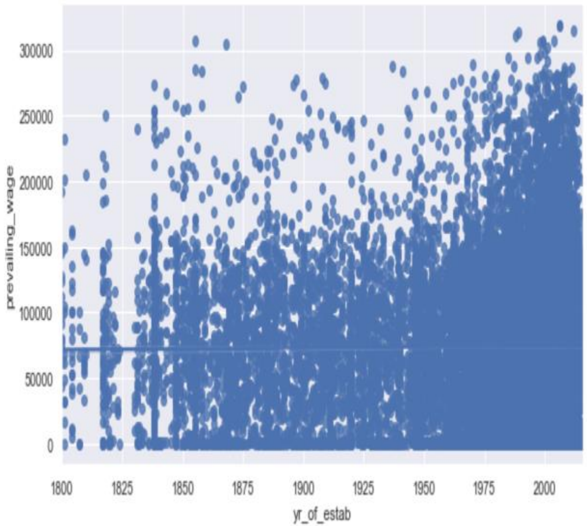


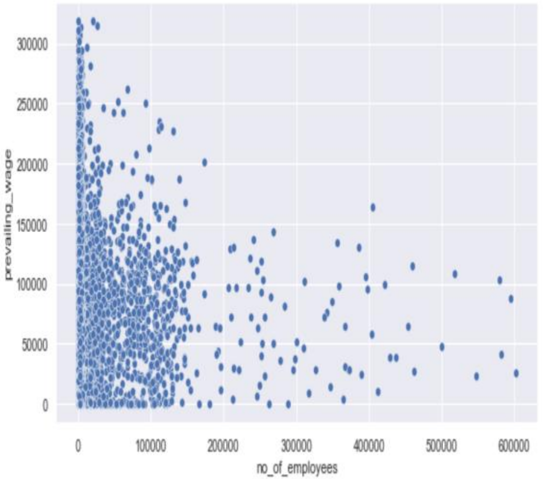
- Most of the certified applicants have full time position.
- There are many Certified applicants than rejected.
- Certified applicants have more prevailing wages than rejected applicants.

EDA(contd)

- The number of employees are more in the companies established between 1950 and 2020
- High Prevailing wages are among companies established between 1975 and 2020.
- High prevailing wages are given to the employees where the company has employee count less than 105000.

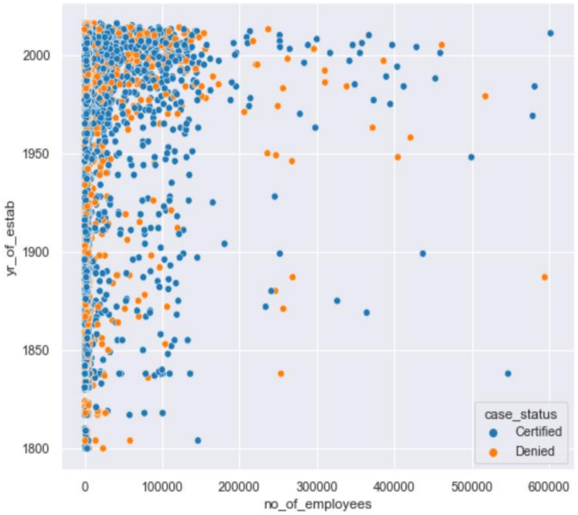


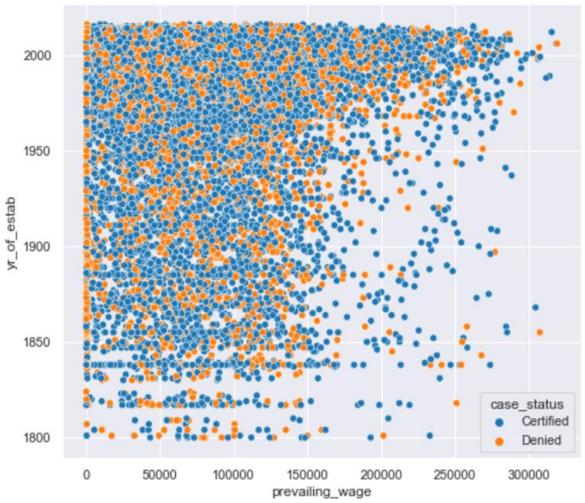


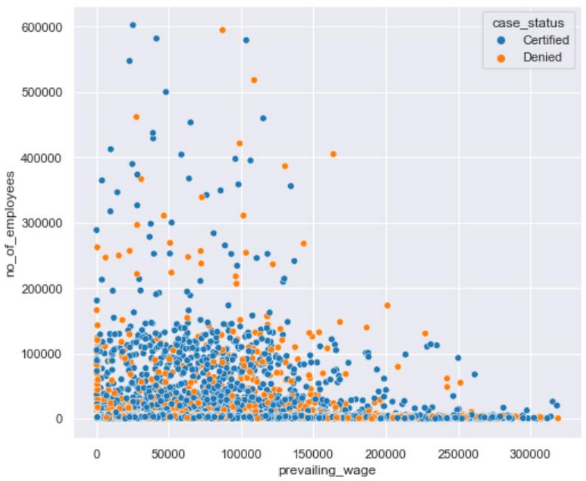


EDA(Correlation)

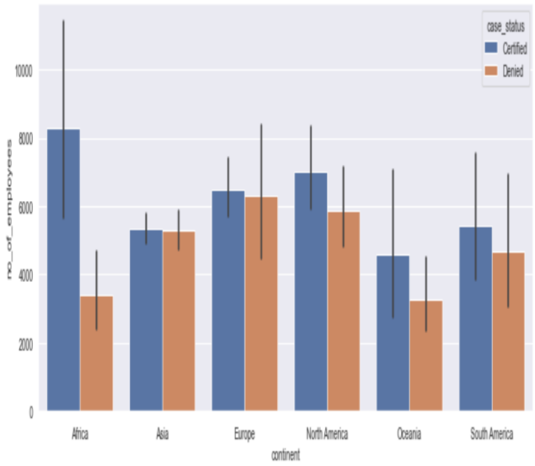
- Most of the cases are certified among the applicants who belong to the companies with employee count between 150000 and 200000 and established between 1950 and 2020.
- Most of the cases certified are employees who belong to the companies established between 1950 and 2000 and have prevailing wages less than 150000.
- Most of the Certified cases are among employees having prevailing wages less than 150000 and belong to the companies with employee count less than 150000.

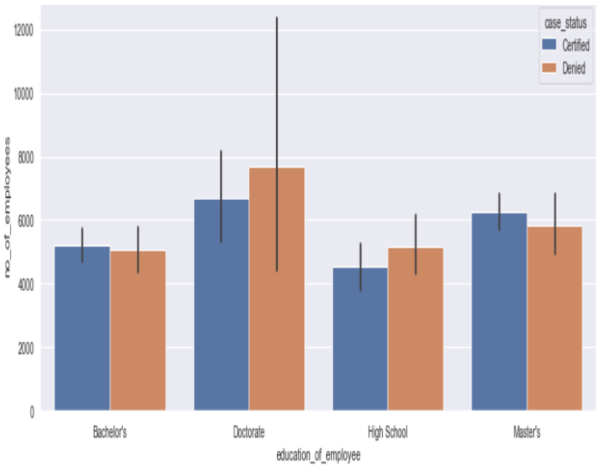


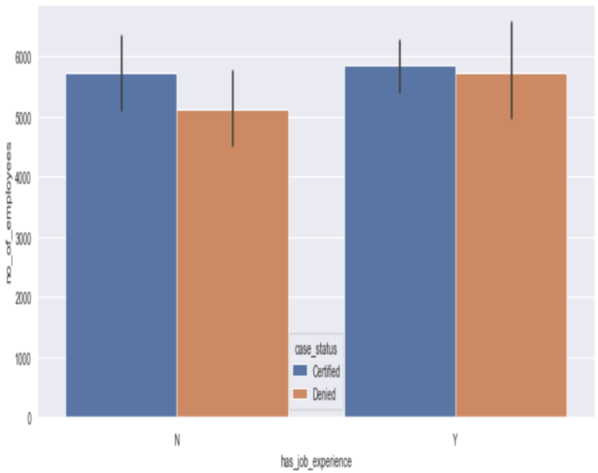




EDA(Correlation)

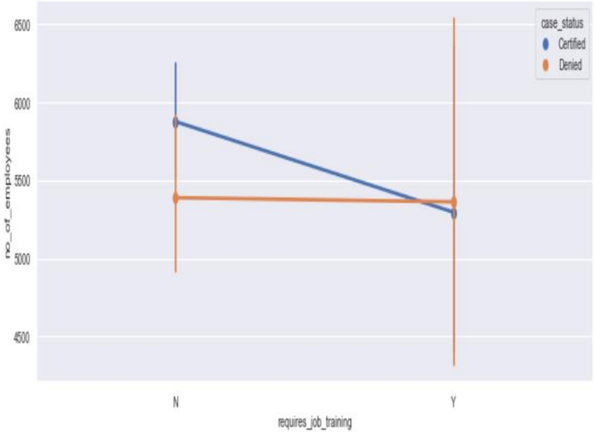


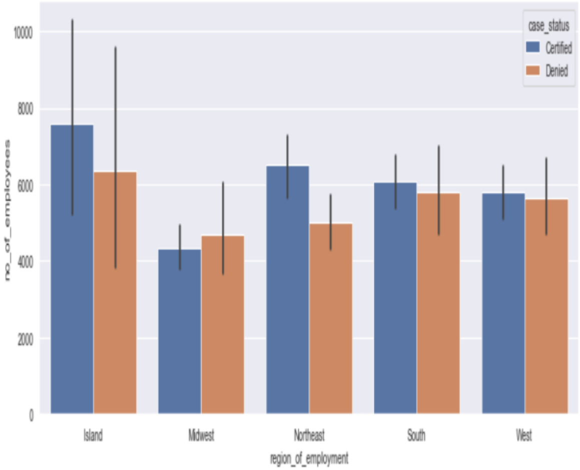


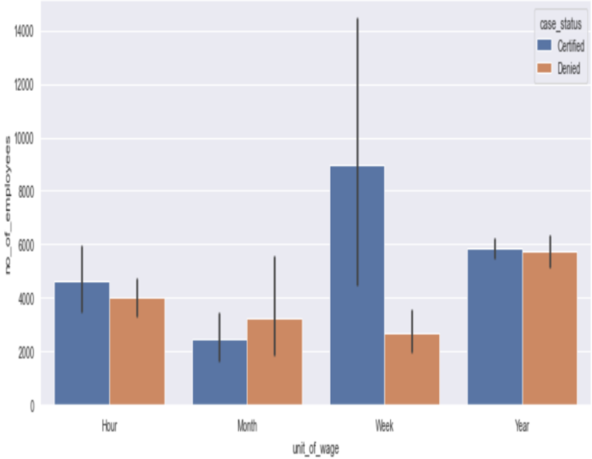


- Less Rejections are among the applicants who belong to the companies with employee count less than 4000 and are from Africa and Oceania.
- More cases are certified among the applicants who belong to the company with more employee count.
- Job experience criteria is less dependent on the case status.

EDA(Correlation)







- There are more cases certified among the candidates who donot require job training.

1. More cases are certified among the applicants from Island followed by Northeast and then south.

- The number of cases certified are more the applicants who belong to companies with more employee count and having weekly wages.

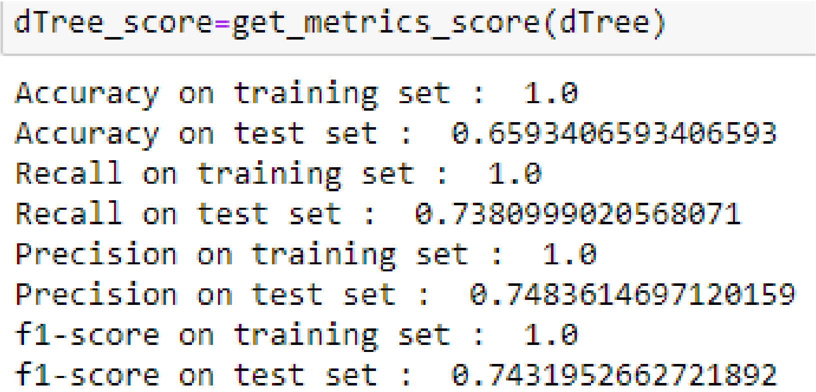
EDA - Summary

1. Most of the cases certified are among the applicants from Asia. Less rejections are from Oceania. There are more certified applicants from Africa, followed by North America and Europe.
2. 40.2% of the employees have Bachelor’s degree. 37.8% of the employees have Master's degree. A smaller number of employees(8.6%) have Doctor's degree.
3. 58.1% of the applicants have job experience. Equal number of cases are certified among the applicants with or without job experience. Job experience criteria is less dependent on the case status.
4. 88.4% of the applicants do not require job training. Most of the candidates whose case is certified are those who did not require job training.
5. 90.1% of the employees receive year wages. 8.5% of the employees receive Hour wages..
6. 89.4% of the employees are in full time position. Most of the certified applicants have full time position. Very less rejected candidates have full time position.
7. 66.8% of the cases are certified. Most of the cases certified are among the employees who belong to the companies established between 1950 and 2000.

Model Performance Summary

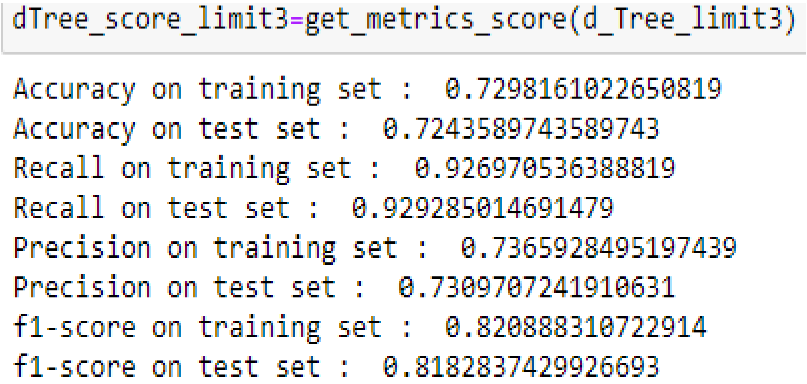
- Decision Tree, Bagging and Boosting Models are used to implement Machine Learning Models.
- Factors Accuracy, Recall, Precision and F1-score values are calculated and compared to find the best model.

Decision Tree



```
dTree_score=get_metrics_score(dTree)
```

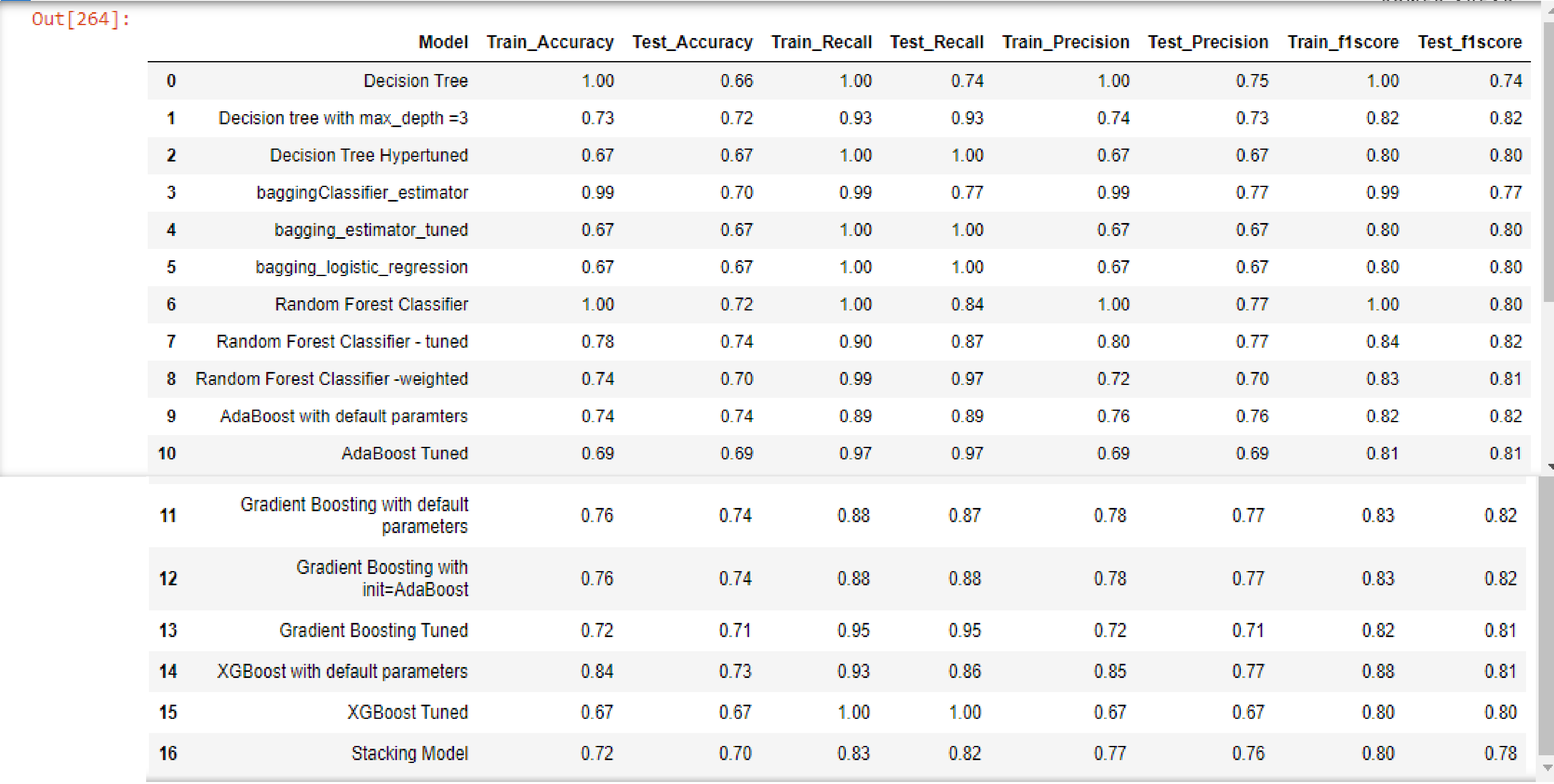
```
Accuracy on training set :  1.0
Accuracy on test set :  0.6593406593406593
Recall on training set :  1.0
Recall on test set :  0.7380999020568071
Precision on training set :  1.0
Precision on test set :  0.7483614697120159
f1-score on training set :  1.0
f1-score on test set :  0.7431952662721892
```



```
dTree_score_limit3=get_metrics_score(d_Tree_limit3)
```

```
Accuracy on training set :  0.7298161022650819
Accuracy on test set :  0.7243589743589743
Recall on training set :  0.926970536388819
Recall on test set :  0.929285014691479
Precision on training set :  0.7365928495197439
Precision on test set :  0.7309707241910631
f1-score on training set :  0.820888310722914
f1-score on test set :  0.8182837429926693
```

Model Performance Summary



Out[264]:

	Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision	Train_f1score	Test_f1score
0	Decision Tree	1.00	0.66	1.00	0.74	1.00	0.75	1.00	0.74
1	Decision tree with max_depth =3	0.73	0.72	0.93	0.93	0.74	0.73	0.82	0.82
2	Decision Tree Hypertuned	0.67	0.67	1.00	1.00	0.67	0.67	0.80	0.80
3	baggingClassifier_estimator	0.99	0.70	0.99	0.77	0.99	0.77	0.99	0.77
4	bagging_estimator_tuned	0.67	0.67	1.00	1.00	0.67	0.67	0.80	0.80
5	bagging_logistic_regression	0.67	0.67	1.00	1.00	0.67	0.67	0.80	0.80
6	Random Forest Classifier	1.00	0.72	1.00	0.84	1.00	0.77	1.00	0.80
7	Random Forest Classifier - tuned	0.78	0.74	0.90	0.87	0.80	0.77	0.84	0.82
8	Random Forest Classifier -weighted	0.74	0.70	0.99	0.97	0.72	0.70	0.83	0.81
9	AdaBoost with default paramters	0.74	0.74	0.89	0.89	0.76	0.76	0.82	0.82
10	AdaBoost Tuned	0.69	0.69	0.97	0.97	0.69	0.69	0.81	0.81
11	Gradient Boosting with default parameters	0.76	0.74	0.88	0.87	0.78	0.77	0.83	0.82
12	Gradient Boosting with init=AdaBoost	0.76	0.74	0.88	0.88	0.78	0.77	0.83	0.82
13	Gradient Boosting Tuned	0.72	0.71	0.95	0.95	0.72	0.71	0.82	0.81
14	XGBoost with default parameters	0.84	0.73	0.93	0.86	0.85	0.77	0.88	0.81
15	XGBoost Tuned	0.67	0.67	1.00	1.00	0.67	0.67	0.80	0.80
16	Stacking Model	0.72	0.70	0.83	0.82	0.77	0.76	0.80	0.78

Model Comparison Summary

1. Factors Accuracy, Recall, Precision and F1-score values are calculated and compared to find the best model.
2. According to the Recall definition, high recall means less false negatives. Lower chances of rejected cases as certified. Recall should be maximized, the greater the Recall higher the chances of identifying.
3. Here, the highest recall has been obtained for Decision Tree - Hypertuned, bagging logistic regression, bagging estimator - tuned and XGBoost - Tuned. But, the accuracy of these models are very low (0.67)
4. Coming to the second best model with better recall and accuracy are - Random Forest Classifier - weighted and Gradient Boosting Tuned
5. The difference between recall values of Random Forest Classifier -weighted and Gradient Boosting - Tuned models is 2%. The recall value of Random Forest Classifier -weighted is slightly more.
6. The accuracy of Gradient Boosting Tuned models is 1% more than Random Forest Classifier - weighted.
7. The differences between respective training vs testing recall values is less in Gradient Boosting Tuned model.
8. Hence, Gradient Boosting Tuned model is preferred.

Business Insights and Recommendations

From the data analysis, we can see definite patterns.

- 1. Most applications and certifications are from Asia
- 2. Close to 70% have Masters or Bachelors.
- 3. People from Europe have a low certification rate.
- 4. People from South America, North America and Asia have high certification rate.
- 5. Certification is highest for people with Doctorate or Masters degree.
- 6. Certification is lowest for people with High school degree.
- 7. Rejection rate is high in people with no job experience.
- 8. All weekly wage workers are certified.
- 9. Employees of Northeast has higher percentage of certification while employees of Midwest have the highest percentage of rejection.
- Hence Education of the candidates (High school / Masters / Doctorate), Job experience, Prevailing wage, Unit of wage and region of employment play a key role in the certification rate.

Business Insights and Recommendations

- OFLC can save a lot of resources and time by pre-filtering/ pre-sorting the candidates based on the model. They can start processing the applications where the candidates have high level of education, job experience, prevailing wage and being employed in regions like the North East.
- They could also raise the minimum requirements for prevailing wage, Education, job experience so that they get only high quality candidates who are more likely to be certified.
- Since the model selected has high recall and a decent accuracy, the model predictions on positive certifications will help reduce resource wastage while keeping the opportunity cost of losing good candidates to the minimum.

greatlearning

Power Ahead

Happy Learning !