Easy Visa Project: Ensemble methods

Contents / Agenda

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

Executive Summary

- the most important features are the following in order:

- ['education_of_employee_HighSchool', 'has_job_experience_Y', 'prevailing_wage',"education_of_employee_Master's",'education_of_employee_Doctorate','continent_Europe', 'unit_of_wage_Year','region_of_employment_Midwest','region_of_employment_South','no_of_employees','continent_North America','yr_of_estab','region_of_employment_West','continent_Asia','full_time_position_Y','region_of_employment_Northeast','continent_South America','requires_job_training_Y']
  - We can use the Gradient Boost predictive model to predict the certified visa cases. The ability to predict the certified visa per year can allow to oversee the Visa system to manage them more efficiently and fairly. In addition, reduce the labore work for big data.
  - to facilitate the process of visa approvals, a model of the important features will be built to generate prediction and make a list of the approved visa cases. E.g. Applicants with high job experience, from Europe, and high educations are on the top of this list. This list varies by regions as well. The data itself can be reduced by dropping the data with high likelihood to fail to be approved: high school education, and from south America, with no job experience and the job is not yearly.

Business Problem Overview and Solution Approach

- **Problems: The Office of Foreign Labor Certification (OFLC) processes job certification applications for employers seeking to bring foreign workers into the United States and grants certifications in those cases based on specific rolls. The process of reviewing every application is becoming a tedious task as the number of applicants is increasing every year. Machine Learning based solution will help in shortlisting the candidates having higher chances of VISA approval.**

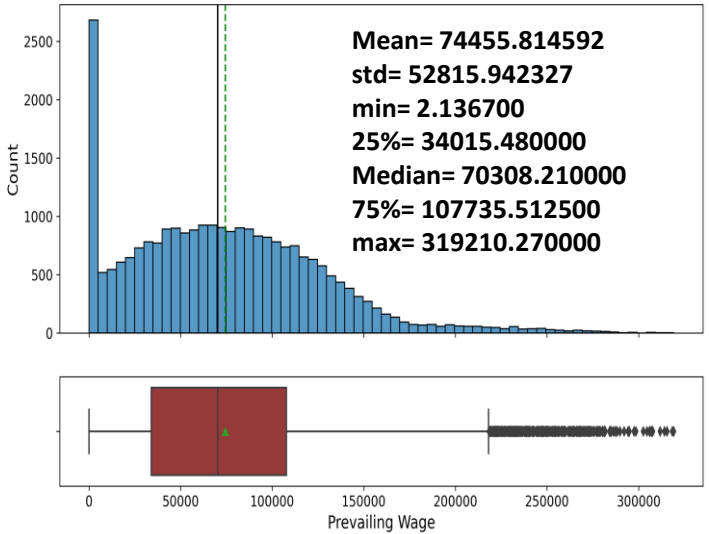
- **Solution approach / methodology**

  - **Classification supervised machine leaning models (logistic regression, Decision Tree, Random Forest, and GBM) applied and choose the best one to predict who's the applicants most likely to have Visa approved. And find what are the most important features affect the chosen model.**

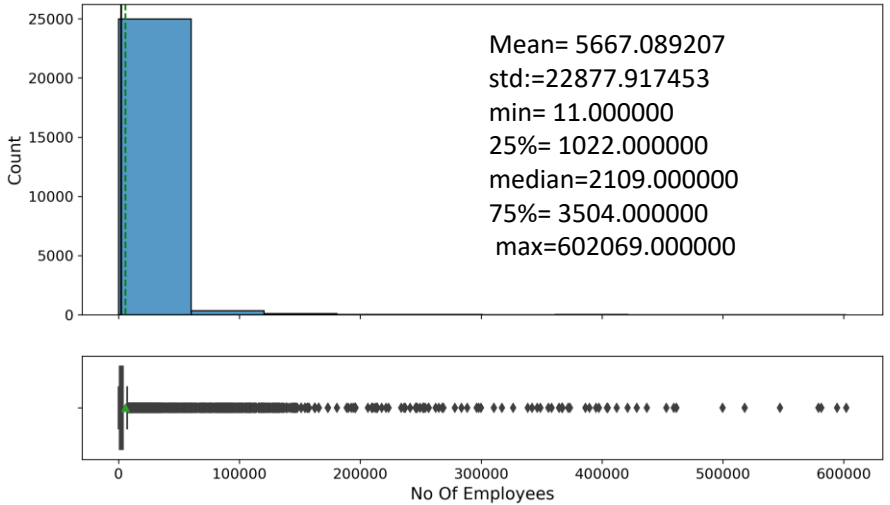
EDA Results

- Univariate Analysis

Observations on prevailing wage

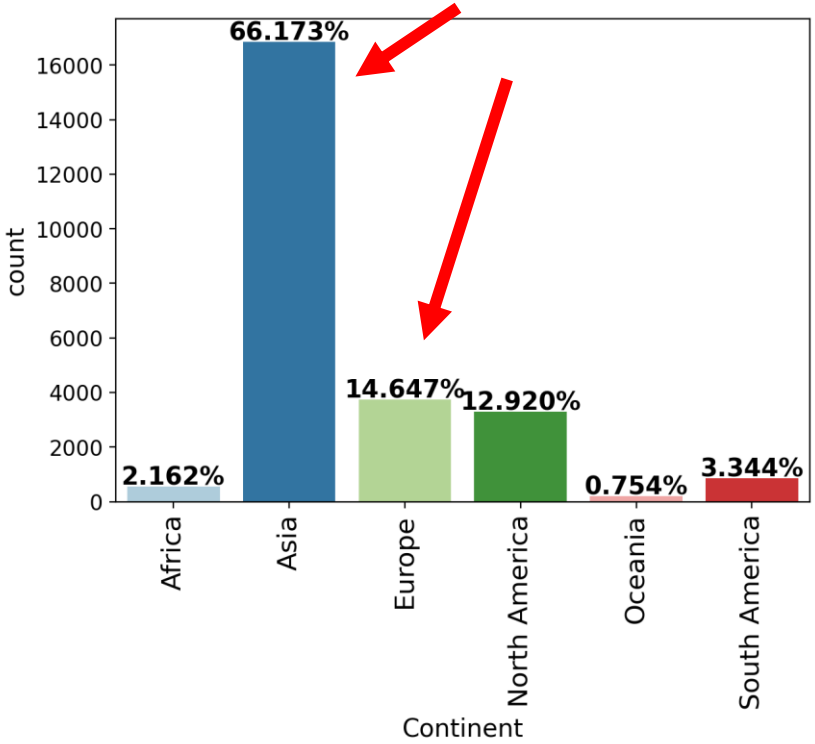


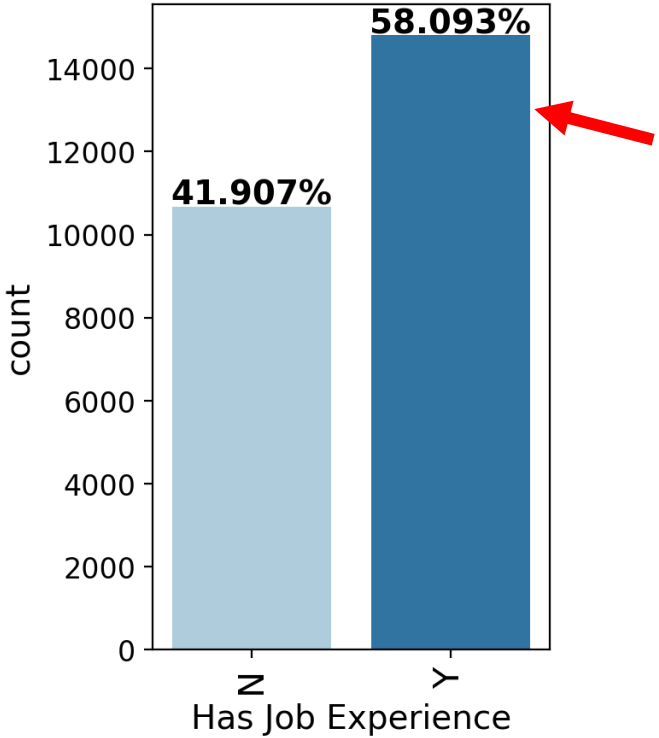
Observation of no of employees



EDA Results

- Univariate Analysis

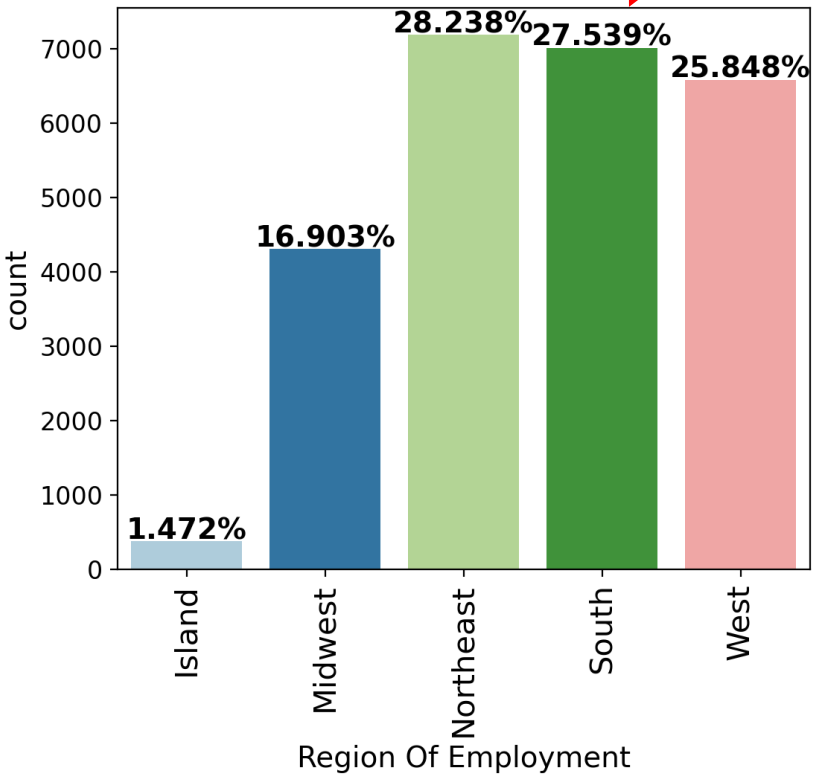


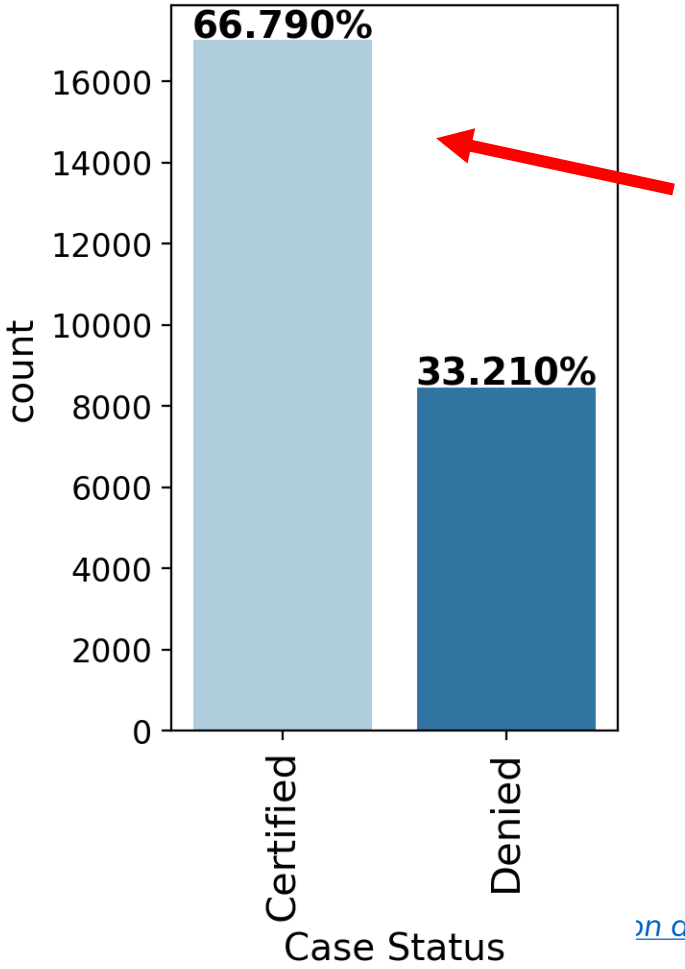


Link to Appendix slide on data background check

EDA Results

- Univariate Analysis

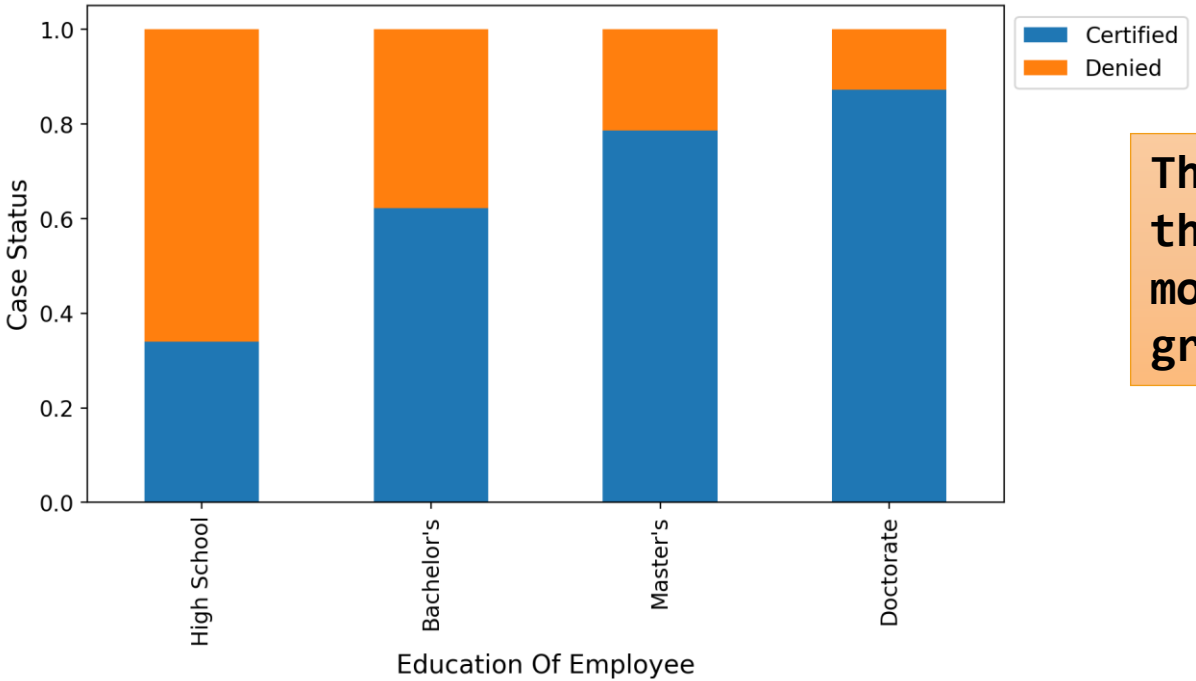




[ɔn data background check]

EDA Results

Does education play a role in Visa certification?

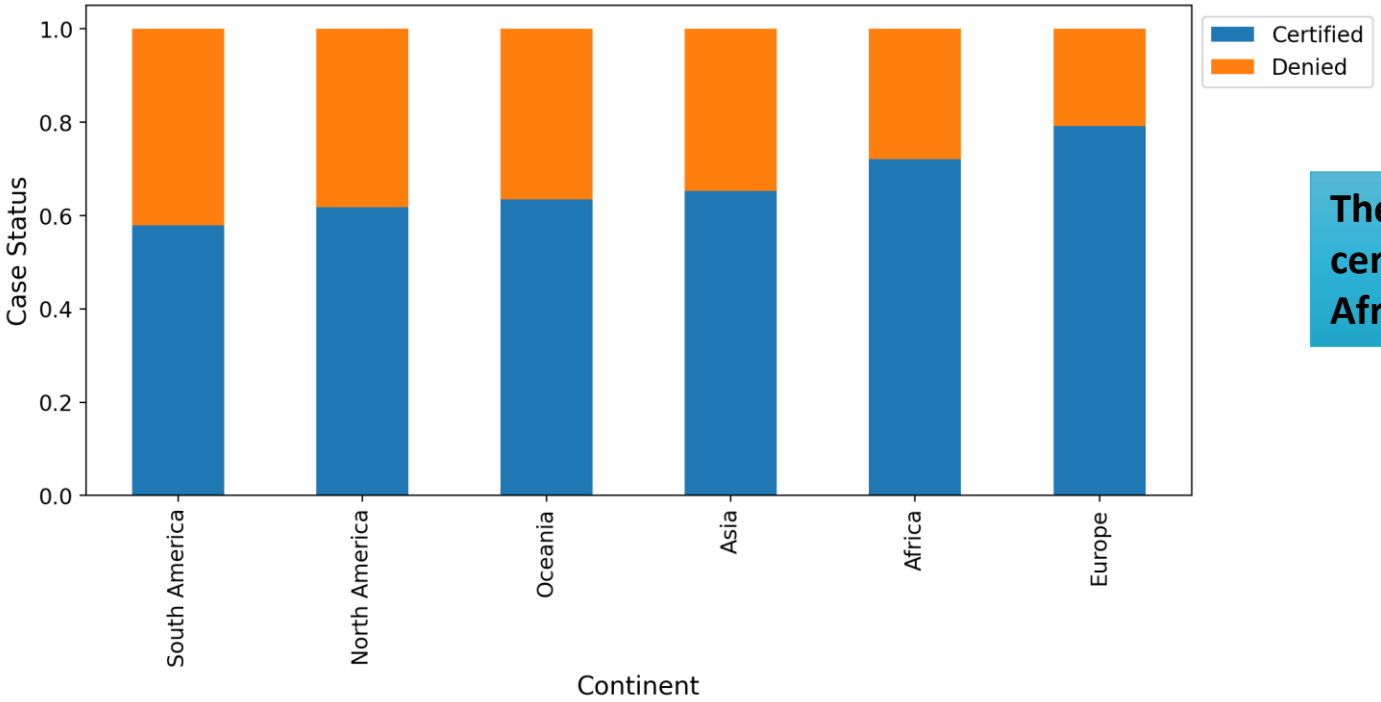


The higher education the applicants, the most likely to be granted the visa

Link to *Appendix slide on data background check*

EDA Results

How does the visa status vary across different continents?

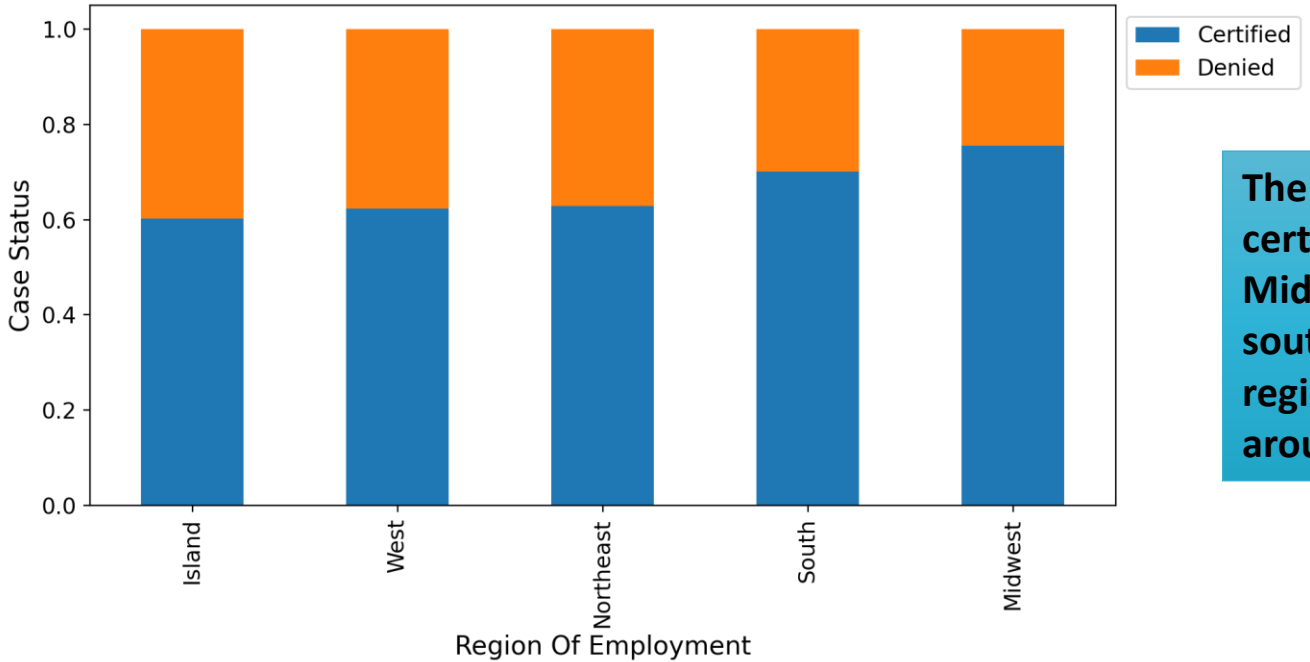


The percentage of visa certified is higher for African and European.

slide on data background check

EDA Results

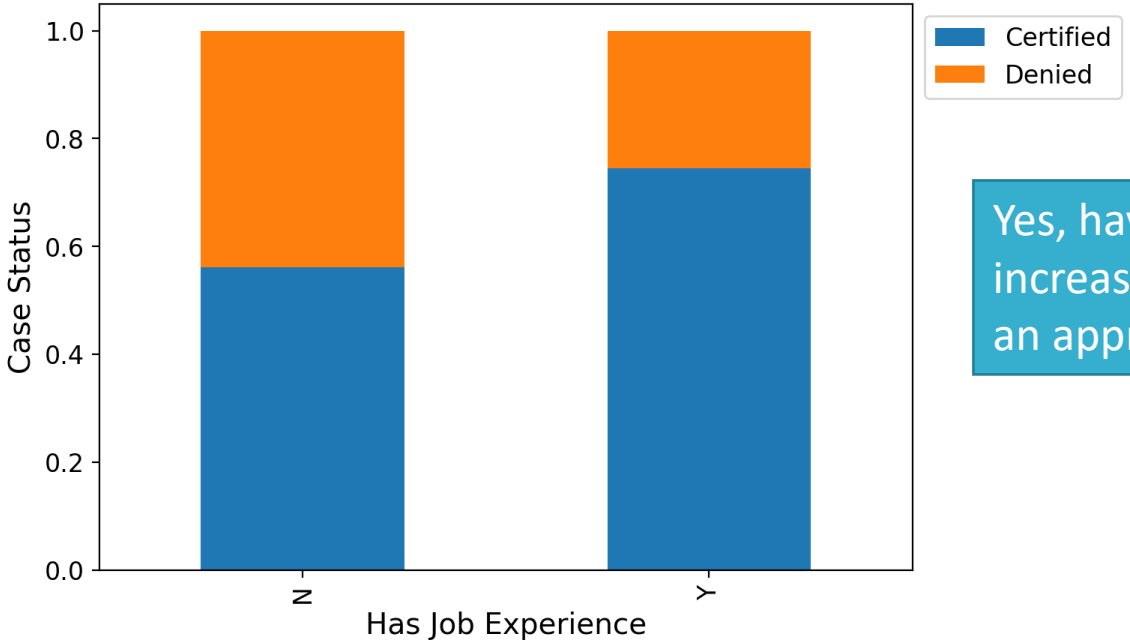
How does the visa status vary across USA Regions?



The percentage of visa certified is higher for Midwest (75%) and south(70%) while other regions the percentage around 60%.

EDA Results

Does work experience influence visa status?

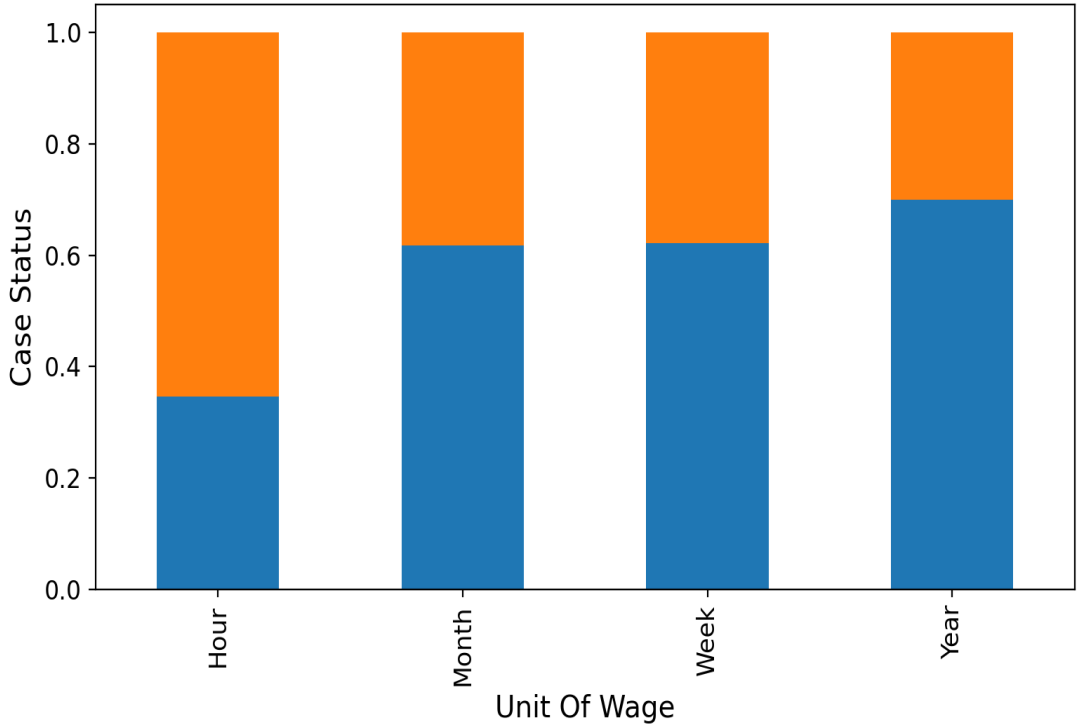


Yes, having work experience increases the probability to have an approved Visa.

Link to Appendix slide on data background check

EDA Results

Which pay unit is most likely to be certified for a visa?



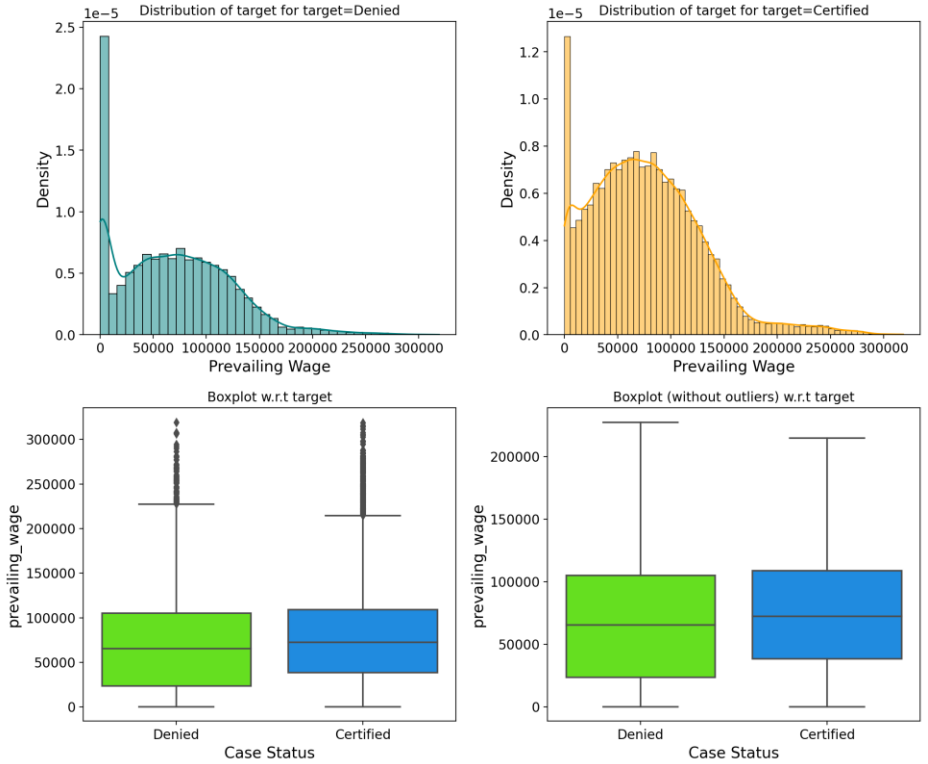
Yearly wage has highest Visa approval while the hourly waves has the lowest visa approval.

dix slide on data background check

EDA Results

How does the visa status change with the prevailing wage?

- There is no a clear relationship between the case status and the prevailing waves. However, the first quantile of prevailing wage is higher for certified visa.
- The outliers does not affect the results.
- Outliers look real values.



Data Preprocessing

- Duplicate value check
  - There is no duplicated Data
- Missing value treatment
  - No missing values
- Outlier check (treatment if needed)
  - Checked the oulier and we did not apply any treatment.
- Feature engineering
  - Drop Case_id feature.
- Data preparation for modeling
  - The defind the target as the Case status and the predictores the rest of the features .
  - Encode categorical features()
  - Split into training and test set with taking acount of class weight.

- ***Note**: You can use more than one slide if needed*

Models Performance Summary

Which case of false prediction is more important?

- False positive and false negative both the cases are important as if a visa is certified when it had to be denied a wrong employee will get the job position while US citizens will miss the opportunity to work on that position. If a visa is denied when it had to be certified the U.S. will lose a suitable human resource that can contribute to the economy.
- To reduce the losses, F1 Score can be used as a metric for evaluation of the model, greater the F1 score higher are the chances of minimizing False Negatives and False Positives.

Models Performance Summary

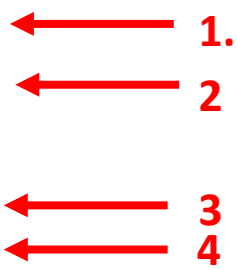
- Overview of final ML model and its parameters

Training Model	Accuracy	F1	Accuracy	F1
Decision Tree	1.000000	1.000000	0.664835	0.747487
Tuned Decision Tree	0.712548	0.812411	0.706567	0.809058
Bagging Classifier	0.985198	0.988887	0.691523	0.767913
Tuned Bagging Classifier	0.996692	0.997529	0.731293	0.811179
Random Forest	1.000000	1.000000	0.727368	0.805851
Tuned Random Forest	0.769119	0.841652	0.738095	0.820930
Adaboost Classifier	0.738226	0.819080	0.734301	0.816481
Tuned Adaboost Classifier	0.718995	0.787861	0.716510	0.786397
Gradient Boost Classifier	0.758802	0.830349	0.744767	0.820927
Tuned Gradient Boost Classifier	0.764802	0.833921	0.745029	0.820319
XGBoost Classifier	0.838753	0.885272	0.733255	0.811675
XGBoost Classifier Tuned	0.765474	0.833935	0.745160	0.820063
Stacking Classifier	0.769679	0.838889	0.742151	0.820312

Gradient Boosting Classifier is the best model . It has higher values and the least difference between training and testing values and fast to compute.

GB classifier parameters are

[loss="deviance", learning_rate=0.1, n_estimators=100, subsample=1, criterion="friedman_mse", min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0, max_depth=3, min_impurity_decrease=0, init=None, random_state=None, max_features=None, verbose=0, max_leaf_nodes=None, warm_start=False, validation_fraction=0.1, n_iter_no_change=None, tol=0.0001, ccp_alpha=0]



Gradient boosting, XGBoost, and stacking regressor are the top models. They are all giving a similar performance.

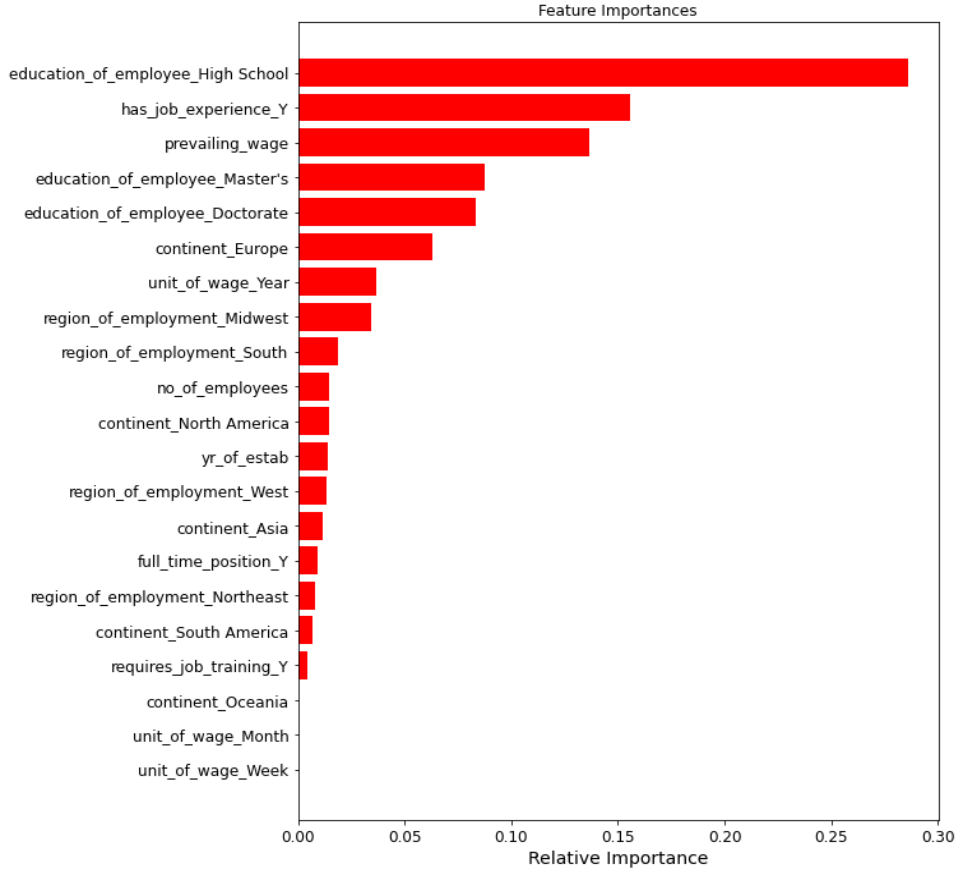
Model Performance Summary

- Summary of key performance metrics for training and test data in tabular format for comparison

	Accuracy	Recall	Precision	F1
Training	0.758802	0.88374	0.783042	0.830349
Test	0.744767	0.876004	0.772366	0.820927

Models Performance Summary

- **Summary of most important factors used by the Gradient Boosting model for prediction**



Most important features

Med important features

Not important features

APPENDIX

Data Background and Contents

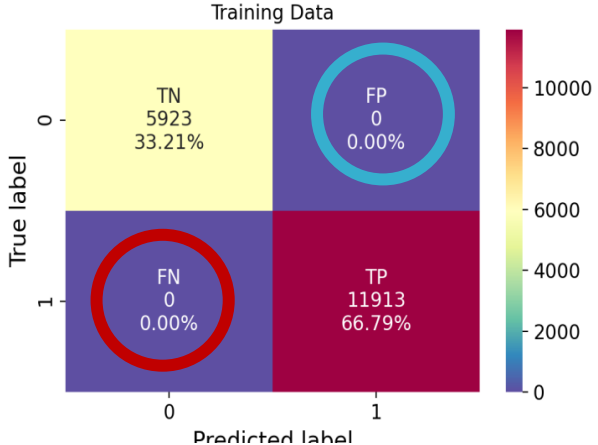
- The data contains the different attributes of employee and the employer. The detailed data dictionary is given below.

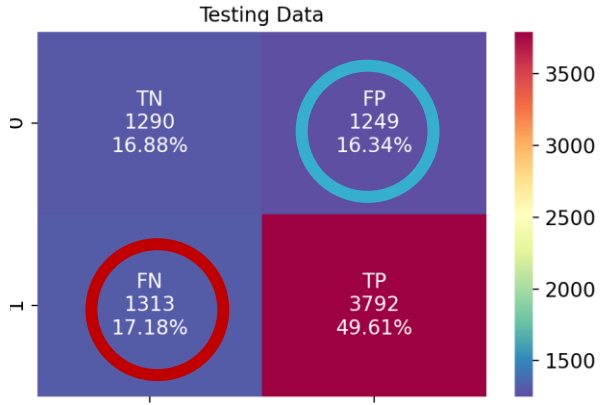
  - case_id: ID of each visa application
  - continent: Information of continent the employee
  - education_of_employee: Information of education of the employee
  - has_job_experience: Does the employee has any job experience? Y= Yes; N = No
  - requires_job_training: Does the employee require any job 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

Observations:

* 66% of the visa applicants are Asian, and 14% European
* 40% has Bachelor degree and 37% has Master degree,8% Doctrorat, 135 high school
* 58% has a job experiences
* 88% of applicant does't need job training
* Northeast,south,and west of region of employment has the highest percentage
* 66% of visa was approved.
* there is no correlation between features.
* The higher education the applicants, the most likely to have granted the visa
* applicant with Bachelor's mostly will be accepted in south and west. applicant with Doctorate will be accepted in west and northeast, High school most likely will be accepted in south, Master's will be accepted in Northeast and south.
* The percentage of visa certified is higher is midwest and south.
* The percentage of visa certified is higher in Africa and Europe.
* the visa certification is higher for applicant with jobs experience.
* the employees who have prior work experience require less job training.
* the higher the prevailing wage the most visa certified.
* island and Midwest has the higher prevailing wages

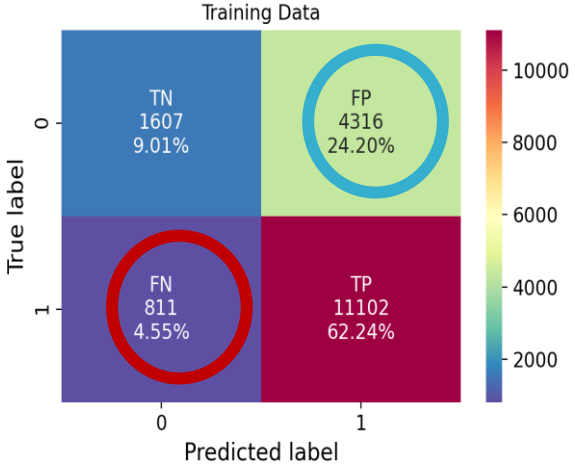
Model Building — Decision Tree

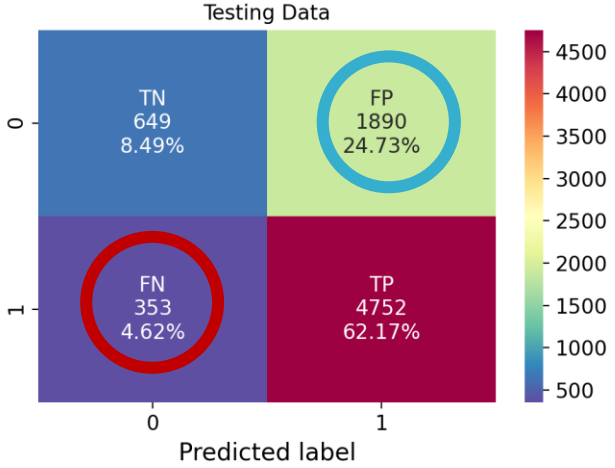




→ **Over Fitting**

F1=1.0::7474 87





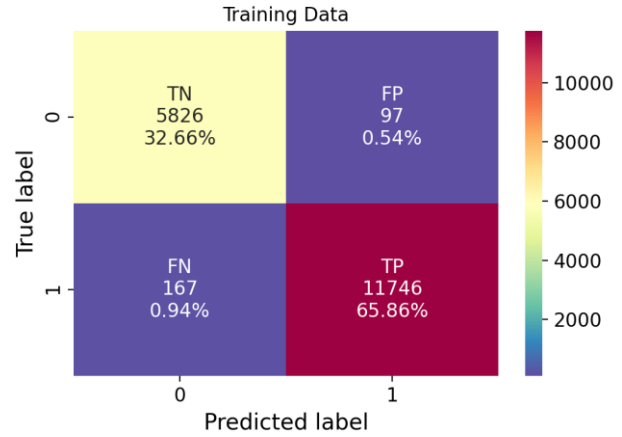
The goal of tunning hyperparameters is overcome the overfitting and reducing the FN and FP.
Parameters:
[class_weight='balanced', max_depth=5, max_leaf_nodes=2,min_impurity_decrease=0.0001, min_samples_leaf=3, random_state=1]

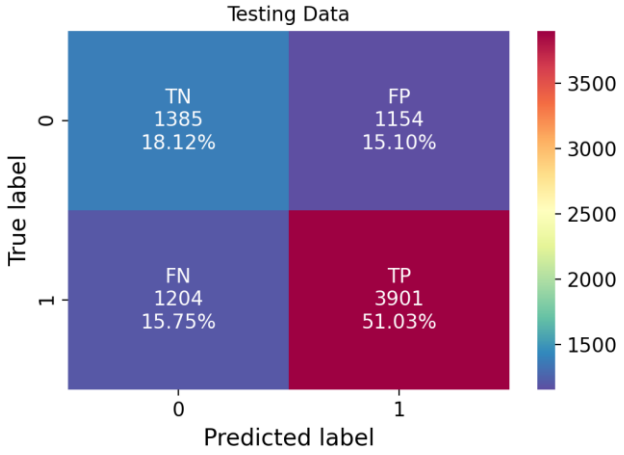
F1=0.812::0.80 9058

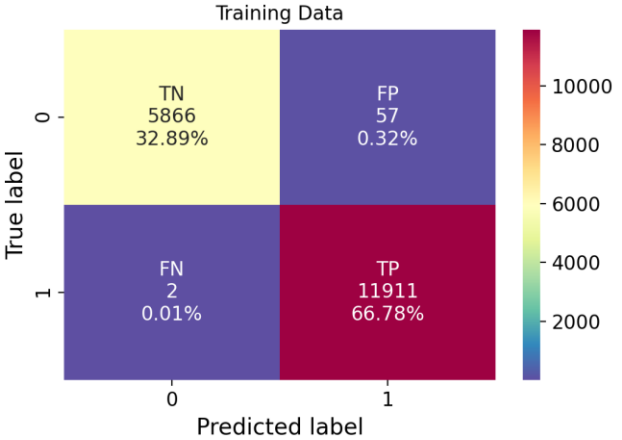
Improved model

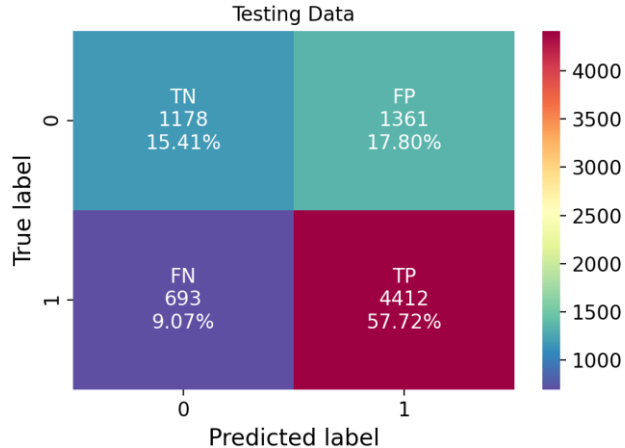
Model Building - Bagging

Bagging Classifier









Over Fitting

F1_test=0.988::0.76

The goal of tunning hyperparameters is reducing the FN and FP. Tunned Parameters: [max_features=0.9, max_samples=0.7, n_estimators=110, random_state=1]

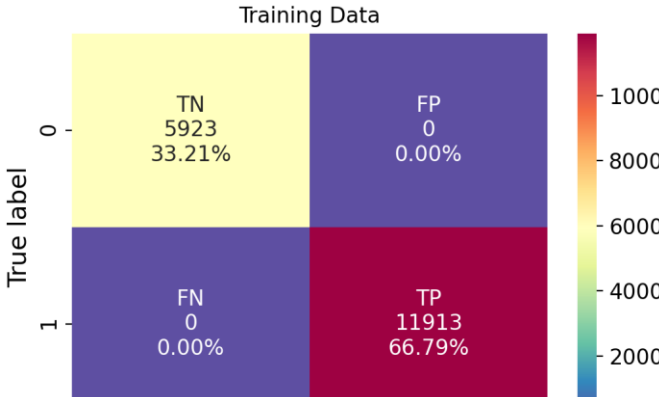
Overfit model

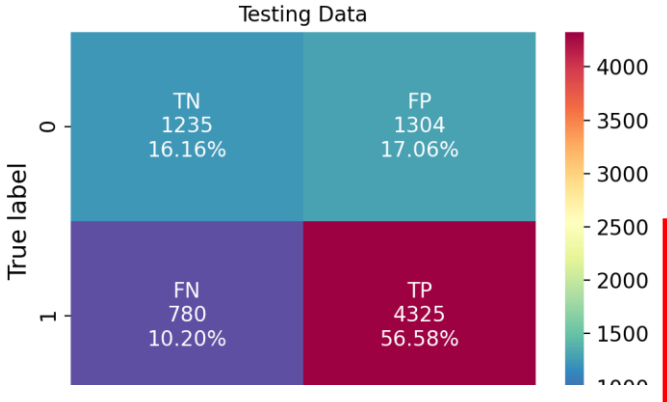
F1=0.997::0.81

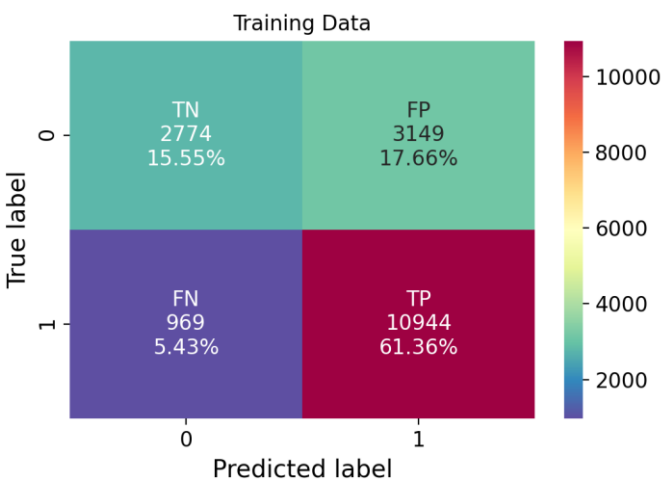
1179

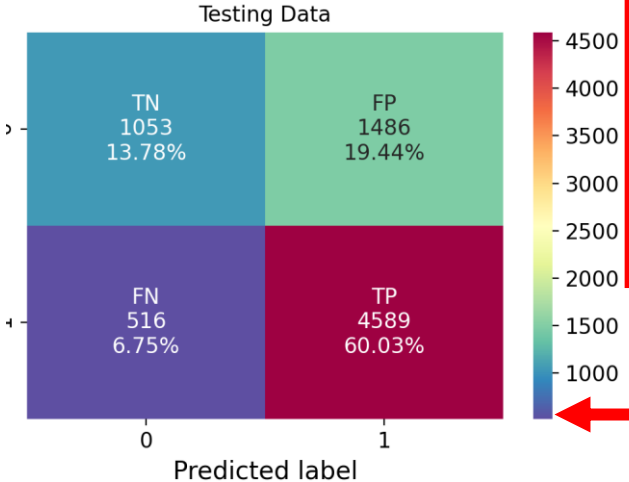
Model Building - Bagging

Random Forest









→ ***Over Fitting***

F1=1.0::0.805

851

Running hyperparameters to overcome over fitting and reducing the FN and FP.

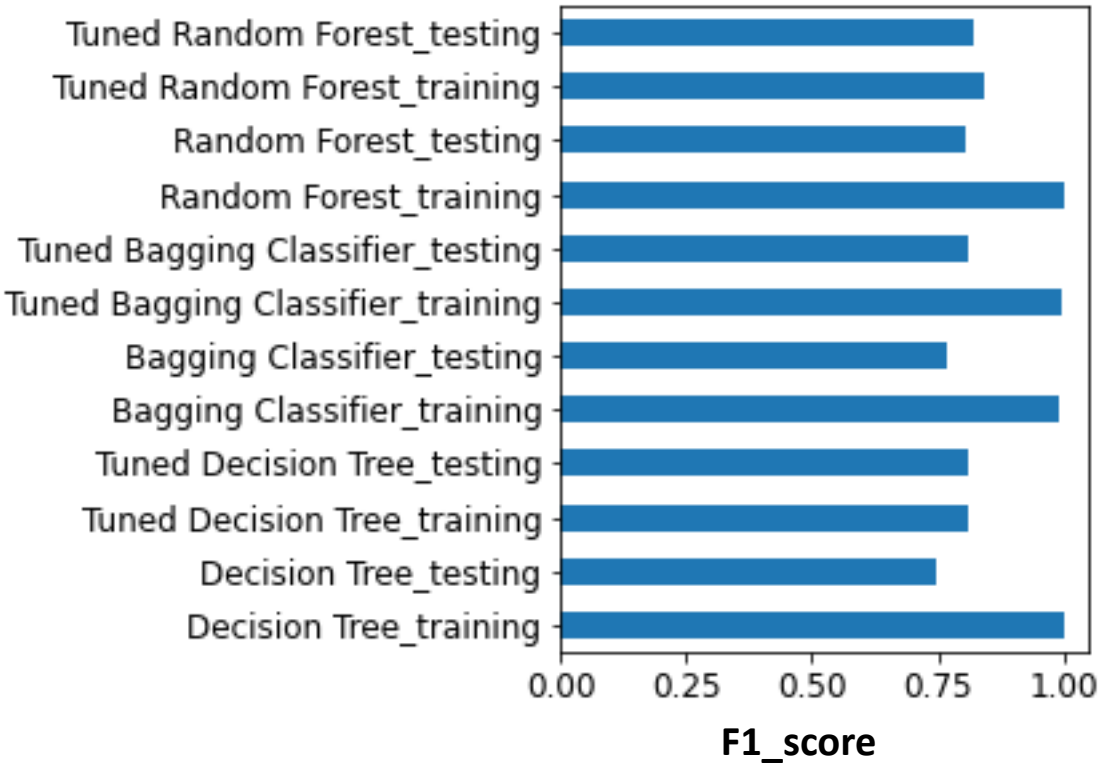
Parameters :

```
[max_depth=10,
max_features='sqrt',
min_samples_split=7,
n_estimators=20,
oob_score=True,
random_state=1]
```

F1=0.841::0.82

Improved

Model Improvement - Bagging

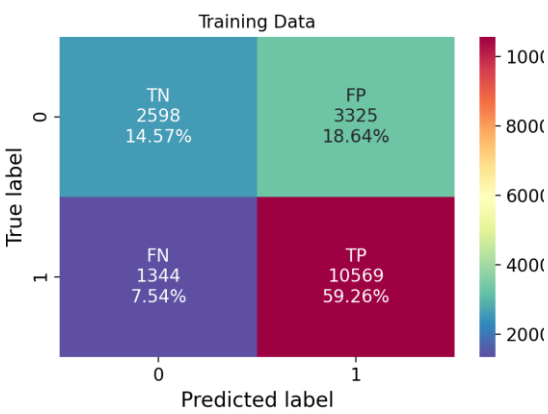


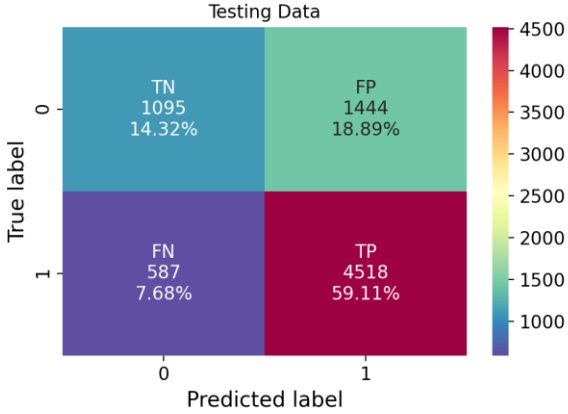
Good Model (Train F1=Test F1)

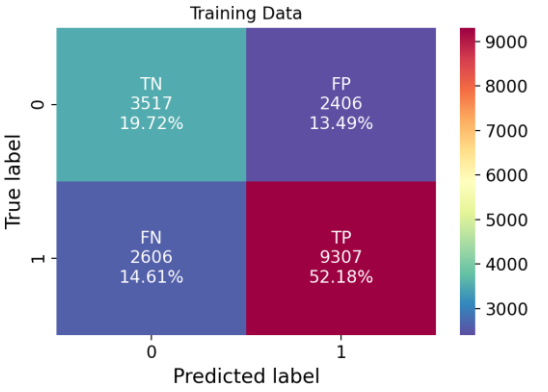
Good Model (Train F1=Test F1)

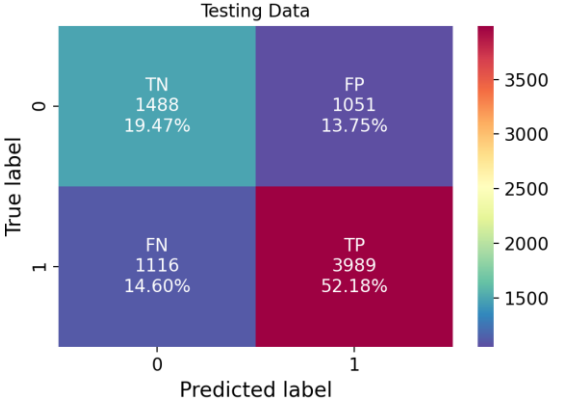
Boosting models

Adaboost models









Adaboost classifier
F1 of AdaBoost
Training:0.81908,
Testing: 0.816481

Good model
(No Over Fit)

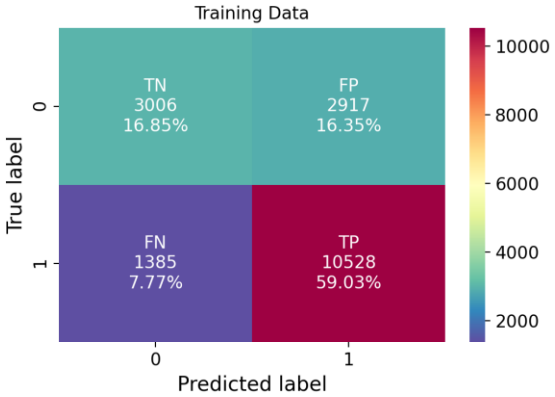
Tunning
Hyperparameter

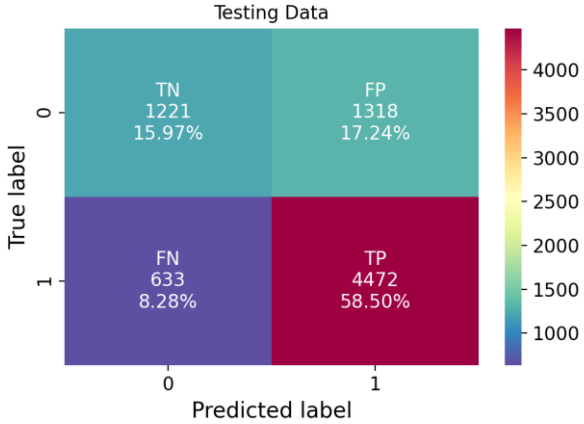
Tunned Adaboost classifier
F1 of Tunned AdaBoost
Training: 0.787861,
Testing:0.786397

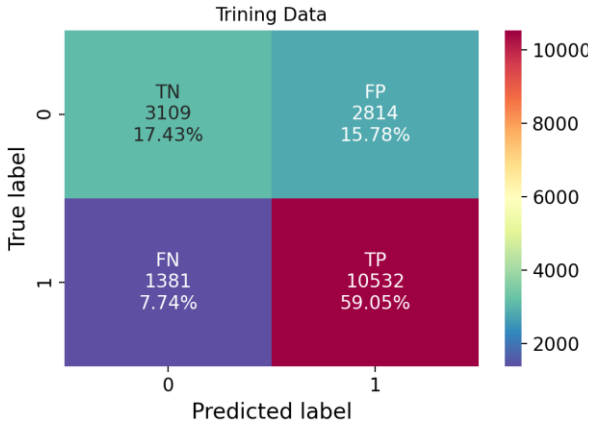
Low values

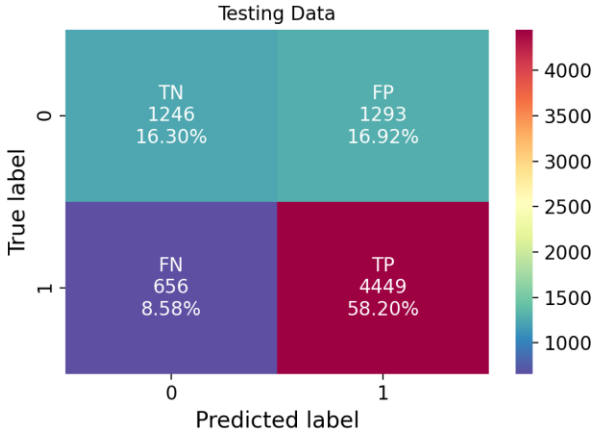
Boosting models

Gradient Boost Models









F1 of GBM
Training:0.830349,
Testing: 0.820927

No over fitting

Improve the F1 values by Tuning the hyperparameters

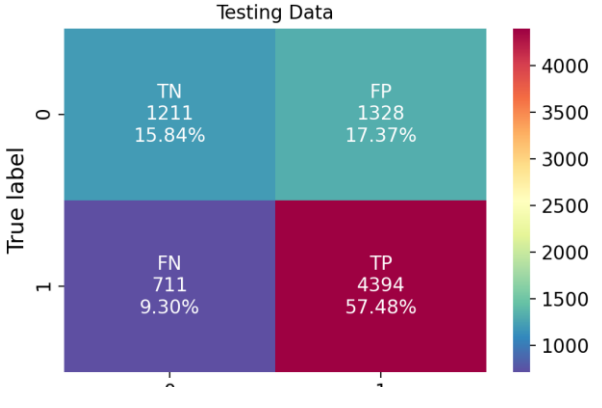
F1 of Tunned GBM
Training: 0.8339,
Testing:0.820319

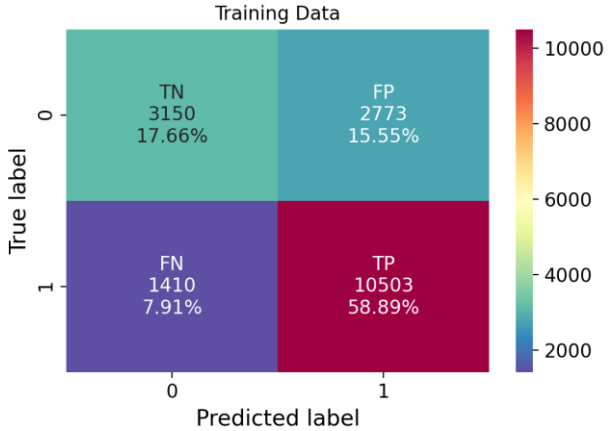
Good result

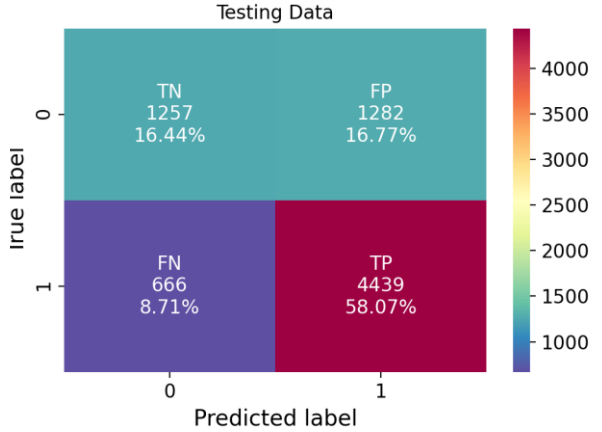
Not big improve in the results

Boosting models

XGradient Boost Models







F1 of XGM Training:0.885272, Testing: 0.811675

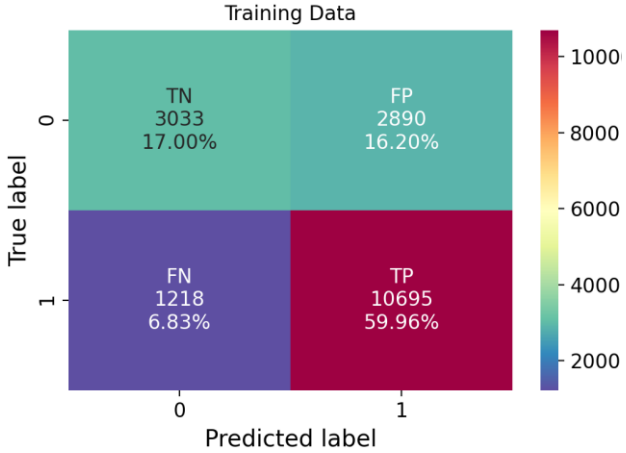
Over Fit Model

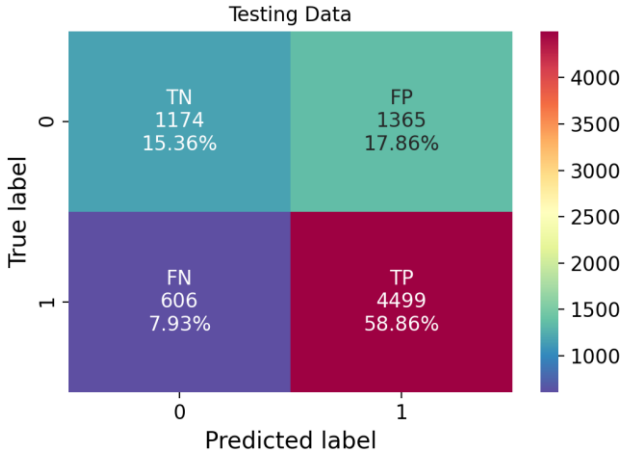
Tunning Hyperparameters

F1 of Tunned XGM Training:0.833935, Testing: 0.820063

Good result

Stacking Classifier consists of AdaBoost classifier, Gradient Boosting classifier, Random Forest classifier





F1 of Stacking Classifier Training:0.838889, Testing: 0.820312

No Over Fitting

Good result