EasyVisa Project

Business Presentation

Contents

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

Executive Summary

The profile of the applicants for whom the visa status can be approved:

Primary information to look at:

- Education level - At least has a Bachelor's degree - Master's and doctorate are preferred.
- Job Experience - Should have some job experience.
- Prevailing wage - The median prevailing wage of the employees for whom the visa got certified is around 72k.

Secondary information to look at:

- Unit of Wage - Applicants having a yearly unit of wage.
- Continent - Ideally the nationality and ethnicity of an applicant shouldn't matter to work in a country but previously it has been observed that applicants from Europe, Africa, and Asia have higher chances of visa certification.
- Region of employment - Analysis shows that Midwest region have more chances of visa approval.

Executive Summary

The profile of the applicants for whom the visa status can be denied:

Primary information to look at:

- Education level - Doesn't have any degree and has completed high school.
- Job Experience - Doesn't have any job experience.
- Prevailing wage - The median prevailing wage of the employees for whom the visa got certified is around 65k.

Secondary information to look at:

- Unit of Wage - Applicants having an hourly unit of wage.
- Continent - Ideally the nationality and ethnicity of an applicant shouldn't matter to work in a country but previously it has been observed that applicants from South America, North America, and Oceania have higher chances of visa applications getting denied.

Executive Summary

- Additional information of employers and employees can be collected to gain better insights. Information such as:
  - Employers: Information about the wage they are offering to the applicant, Sector in which company operates in, etc
  - Employee's: Specialization in their educational degree, Number of years of experience, etc

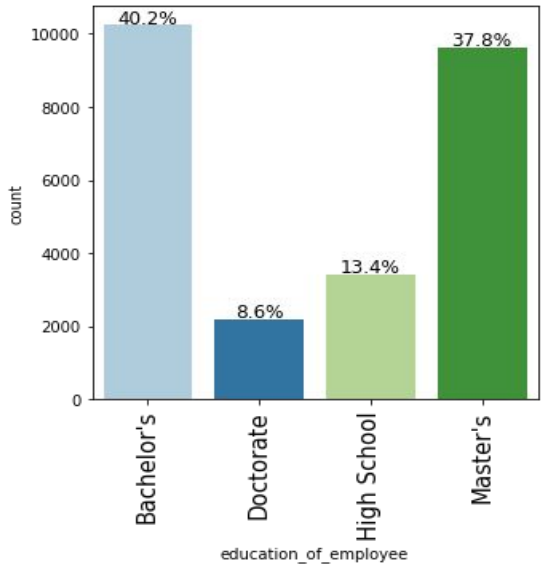
Business Problem Overview and Solution Approach

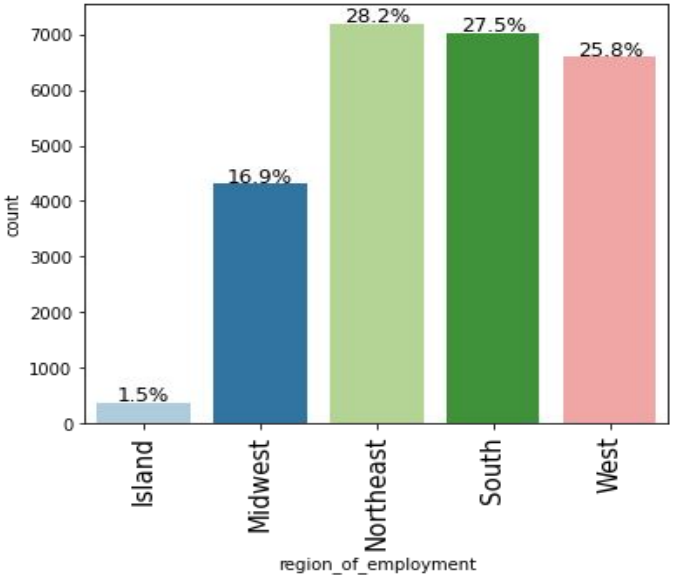
- OFLC processes job certification applications for employers seeking to bring foreign workers into the United States and grants certifications in those cases where employers can demonstrate that there are not sufficient US workers available to perform the work at wages that meet or exceed the wage paid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

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

- The task at hand to analyze the data provided to facilitate the process of visa approvals by building a machine learning model and to 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.

Exploratory Data Analysis Results

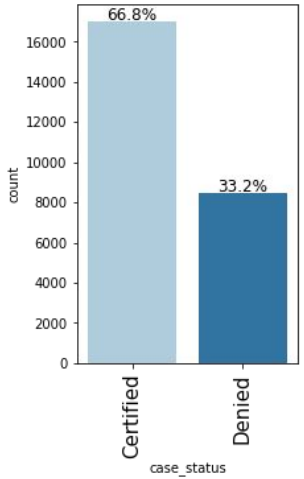


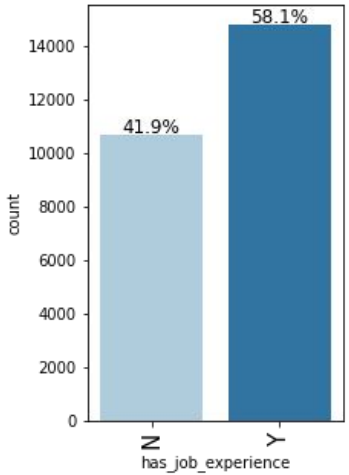


- 40.2% of the applicants have a bachelor's degree, followed by 37.8% having a master's degree.
- 8.6% of the applicants have a doctorate degree.

- Northeast, South, and West have almost equal percentages of applicants.
- The Island regions have only 1.5% of the applicants.

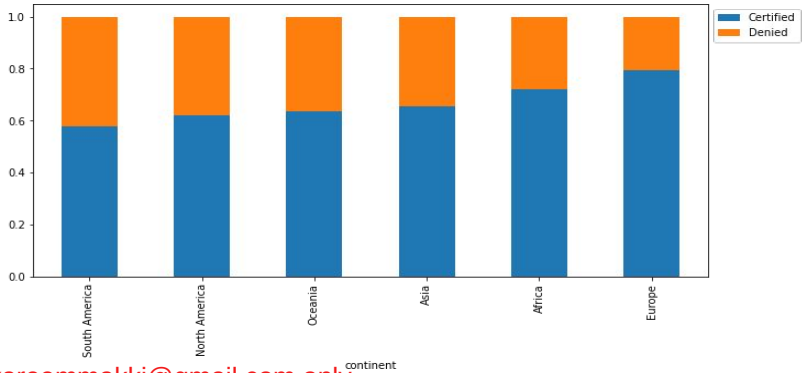
Exploratory Data Analysis



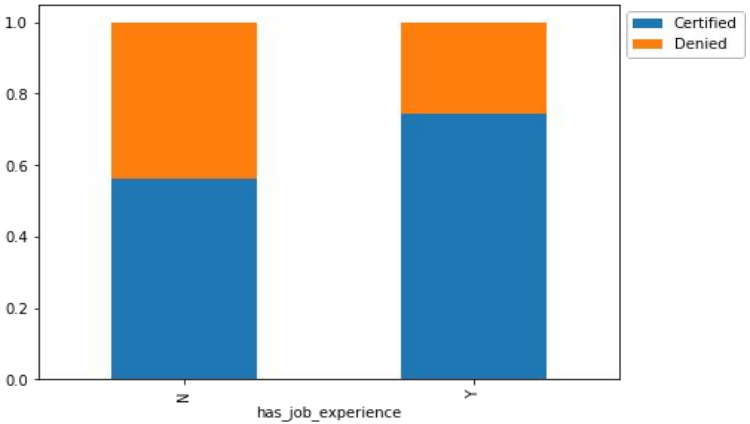


- 66.8% of the visas were certified
- 58.1% of the applicants have job experience.

- Applications from **Europe** and **Africa** have a higher chance of getting certified.
- Around 80% of the applications from **Europe** are certified.
- **Asia** has the third-highest percentage (Around 60%) of visa certification and has the highest number of applications.

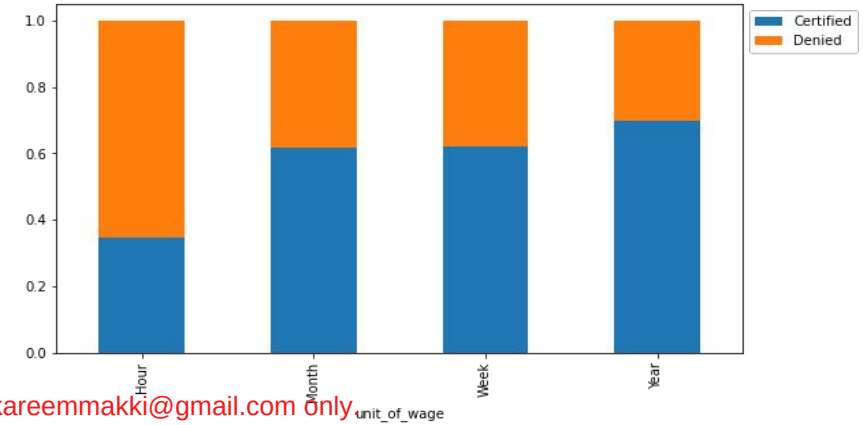


Exploratory Data Analysis

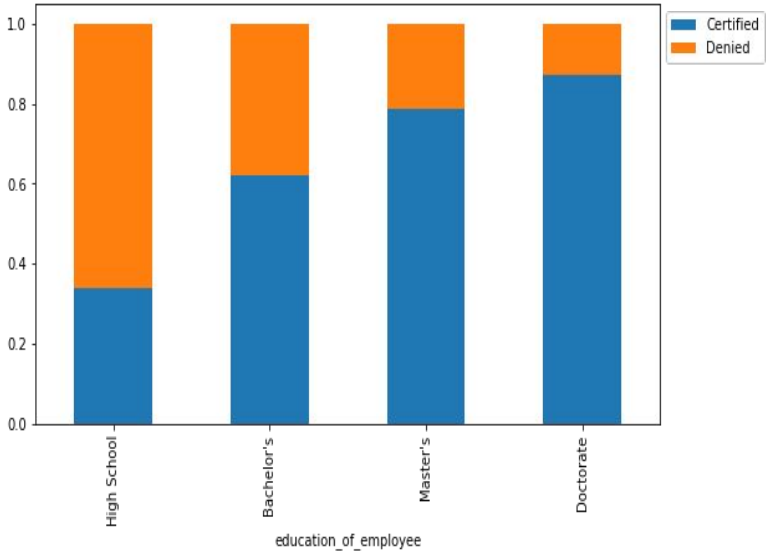


- **Having job experience** seems to be a key differentiator between visa applications getting certified or denied.
- Around 80% of the applications were certified for the applicants who have some job experience as compared to the applicants who do not have any job experience.
- Applicants without job experiences saw only 60% of the visa applications getting certified.

- **Unit of prevailing wage** is an important factor for differentiating between a certified and a denied visa application.
- If the unit of prevailing wage is Yearly, there's a high chance of the application getting certified.
- Around 75% of the applications were certified for the applicants who have a yearly unit of wage. While only 35% of the applications were certified for applicants who have an hourly unit of wage.

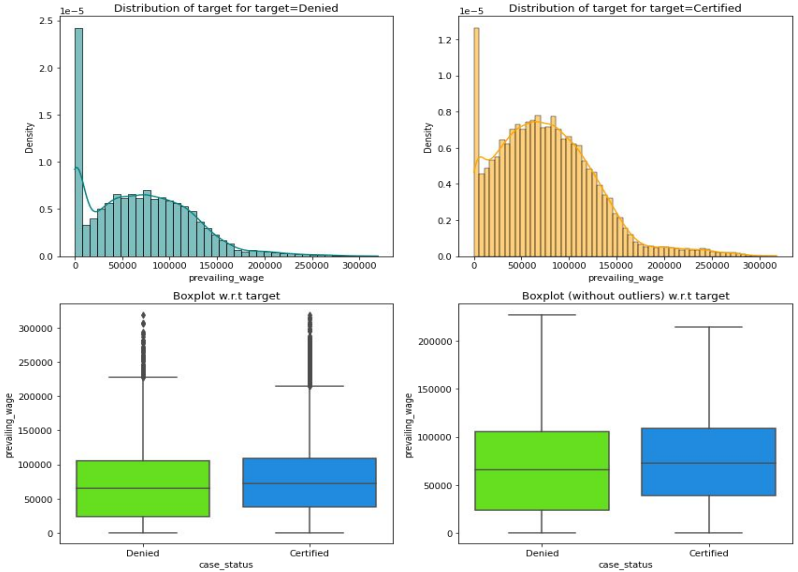


Exploratory Data Analysis

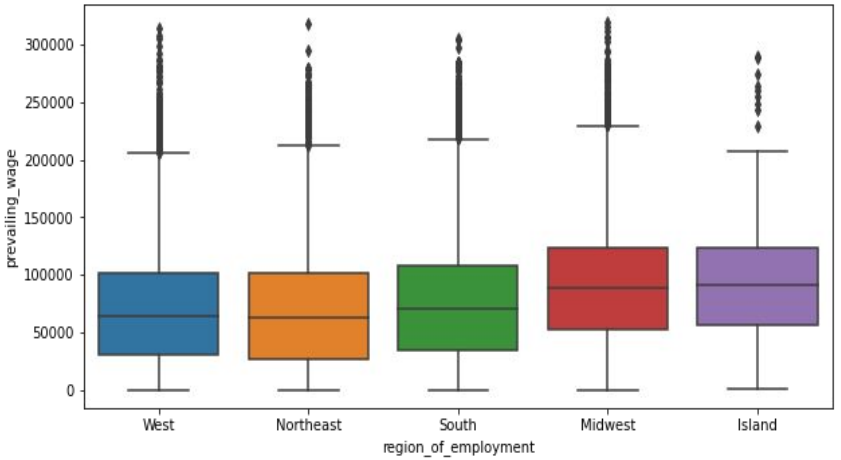


- **Education** seems to have a positive relationship with the certification of visa that is higher the education higher are the chances of visa getting certified.
- Around 85% of the visa applications got certified for the applicants with Doctorate degree. While 80% of the visa applications got certified for the applicants with Master's degree.
- Around 60% of the visa applications got certified for applicants with Bachelor's degrees.
- Applicants who do not have a degree and have graduated from high school are more likely to have their applications denied.

Exploratory Data Analysis



- The median prevailing wage for the **certified** applications is slightly higher as compared to **denied** applications.



- **Midwest** and **Island** regions have slightly higher prevailing wages as compared to other regions.
- The distribution of prevailing wage is similar across **West**, **Northeast**, and **South regions**.

Data Preprocessing

- There were no null and duplicate values in the dataset.
- The case_id column was dropped from the dataset since all the values in that column were unique.
- There were some negative values in the number of employees column which were replaced by their absolute values.
- There were quite a few outliers in the data, however, they were not treated since they were genuine values.

Model Performance Summary

- We want to predict whether the visa application will get certified or not using the information provided.
- We will use F1 Score as the metric for evaluation of the model because
    - 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 of the country.
    - F1 score will help us to minimize both false positives and false negatives.
- We will use **balanced class** weights so that model focuses equally on both classes.
- We will build different models - DecisionTreeClassifier, RandomForestClassifier, BaggingClassifier, AdaBoostClassifier, GradientBoostingClassifier, XGBClassifier, and StackingClassifier
- We will also perform **hyperparameter tuning** for these models and evaluate their performance using different metrics and confusion matrix.

Model Performance Summary

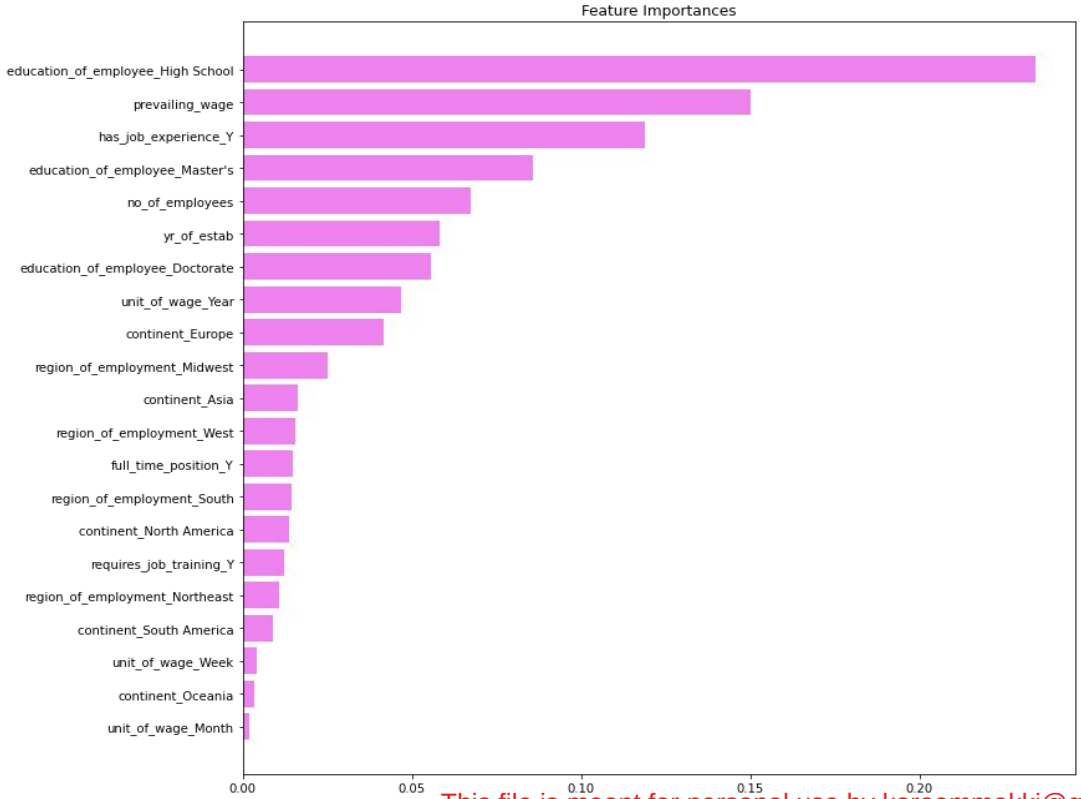
Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision	Train F1	Test F1
Decision Tree	0.712548	0.706567	0.931923	0.930852	0.720067	0.715447	0.812411	0.809058
Tuned Decision Tree	0.712548	0.706567	0.931923	0.930852	0.720067	0.715447	0.812411	0.809058
Bagging Classifier	0.985198	0.691523	0.985982	0.764153	0.99181	0.771711	0.988887	0.767913
Tuned Bagging Classifier	0.996187	0.724228	0.999916	0.895397	0.994407	0.743857	0.997154	0.812622
Random Forest	0.999944	0.721088	0.999916	0.840744	1	0.764926	0.999958	0.801045
Tuned Random Forest	0.769119	0.738095	0.91866	0.898923	0.776556	0.755391	0.841652	0.82093

Best Performing Model

Model Performance Summary

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision	Train F1	Test F1
Adaboost Classifier	0.738226	0.734301	0.887182	0.885015	0.760688	0.757799	0.81908	0.816481
Tuned Adaboost Classifier	0.719163	0.716641	0.781415	0.781587	0.79469	0.79151	0.787997	0.786517
Gradient Boost Classifier	0.758802	0.744767	0.88374	0.876004	0.783042	0.772366	0.830349	0.820927
Tuned Gradient Boost Classifier	0.764017	0.743459	0.882649	0.871303	0.789059	0.773296	0.833234	0.819379
XGBoost Classifier	0.838753	0.733255	0.931419	0.860725	0.843482	0.767913	0.885272	0.811675
XGBoost Classifier Tuned	0.765474	0.74516	0.881642	0.86954	0.791127	0.775913	0.833935	0.820063
Stacking Classifier	0.770296	0.74529	0.892554	0.879138	0.790558	0.771399	0.838465	0.821752

Feature Importance



The top **3 important features**

to look for while certifying a visa are -

- **Education of the employee**
- **Job experience**
- **Prevailing Wage.**

APPENDIX

Data Overview

- The data contains information of 25480 employees and their employers.
- The data includes information about employee's education, job experience, whether the employee requires training or not, continent, etc.
- The data also includes information about the number of employees in the company, year of establishment of the company, prevailing wage, etc.

Decision Tree

Accuracy	Recall	Precision	F1
1.0	1.0	1.0	1.0

Training Performance

- 0 errors on the training set, each sample has been classified correctly.
- Model has performed very well on the training set.
- As we know, a decision tree will continue to grow and classify each data point correctly if no restrictions are applied as the trees will learn all the patterns in the training set.

Accuracy	Recall	Precision	F1
0.664835	0.742801	0.752232	0.747487

Testing Performance

- The decision tree model is overfitting the data as expected and not able to generalize well on the test set.
- We will have to prune the decision tree.

Hyperparameter Tuning - Decision Tree

Accuracy	Recall	Precision	F1
0.712548	0.931923	0.720067	0.812411

Training Performance

Accuracy	Recall	Precision	F1
0.706567	0.930852	0.715447	0.809058

Testing Performance

- The decision tree model has a very high recall but, the precision is quite less.
- The performance of the model after hyperparameter tuning has become generalized.
- We are getting an F1 score of 0.81 and 0.80 on the training and test set, respectively.
- Let's try building some ensemble models and see if the metrics improve.

Bagging Classifier

Accuracy	Recall	Precision	F1
0.985198	0.985982	0.99181	0.988887

Training Performance

Accuracy	Recall	Precision	F1
0.691523	0.764153	0.771711	0.767913

Testing Performance

- The bagging classifier is overfitting on the training set like the decision tree model.
- We'll try to reduce overfitting and improve the performance by hyperparameter tuning.

Hyperparameter Tuning - Bagging Classifier

Accuracy	Recall	Precision	F1
0.996187	0.999916	0.994407	0.997154

Training Performance

Accuracy	Recall	Precision	F1
0.724228	0.895397	0.743857	0.812622

Testing Performance

- After tuning the hyperparameters the bagging classifier is still overfitting.
- There's a big difference in the training and the test recall.

Random Forest

Accuracy	Recall	Precision	F1
1.0	1.0	1.0	1.0

Training Performance

Accuracy	Recall	Precision	F1
0.727368	0.847209	0.768343	0.805851

Testing Performance

- With default parameters, random forest is overfitting the training data.
- We'll try to reduce overfitting and improve recall by hyperparameter tuning.

Hyperparameter Tuning - Random Forest

Accuracy	Recall	Precision	F1
0.769119	0.91866	0.776556	0.841652

Training Performance

Accuracy	Recall	Precision	F1
0.738095	0.898923	0.755391	0.82093

Testing Performance

- After hyperparameter tuning the model performance has generalized.
- We have an F1 score of 0.84 and 0.82 on the training and test data, respectively.
- The model has a high recall and a good precision.

AdaBoost Classifier

Accuracy	Recall	Precision	F1
0.738226	0.887182	0.760688	0.81908

Training Performance

Accuracy	Recall	Precision	F1
0.734301	0.885015	0.757799	0.816481

Testing Performance

- The model is giving a generalized performance.
- We have received a good F1 score of 0.81 on both the training and test set.

Hyperparameter Tuning - AdaBoost Classifier

Accuracy	Recall	Precision	F1
0.719163	0.781415	0.79469	0.787997

Training Performance

Accuracy	Recall	Precision	F1
0.716641	0.781587	0.79151	0.786517

Testing Performance

- After tuning the F1 score has reduced.
- The recall of the model has reduced but the precision has improved.

Gradient Boosting Classifier

Accuracy	Recall	Precision	F1
0.758802	0.88374	0.783042	0.830349

Training Performance

Accuracy	Recall	Precision	F1
0.744767	0.876004	0.772366	0.820927

Testing Performance

- The model is giving a good and generalized performance.
- We are getting the F1 score of 0.83 and 0.82 on the training and test set, respectively.
- Let's see if the performance can be improved further by hyperparameter tuning.

Hyperparameter Tuning - Gradient Boosting Classifier

Accuracy	Recall	Precision	F1
0.764017	0.882649	0.789059	0.833234

Training Performance

Accuracy	Recall	Precision	F1
0.743459	0.871303	0.773296	0.819379

Testing Performance

- After tuning there is not much change in the model performance as compared to the model with default values of hyperparameters.

XGBoost Classifier

Accuracy	Recall	Precision	F1
0.838753	0.931419	0.843482	0.885272

Training Performance

Accuracy	Recall	Precision	F1
0.733255	0.860725	0.767913	0.811675

Testing Performance

- The XGBoost model on the training set has performed very well but it is not able to generalize on the test set.
- Let's try and tune the hyperparameters and see if the performance can be generalized.

Hyperparameter Tuning - XGBoost Classifier

Accuracy	Recall	Precision	F1
0.765474	0.881642	0.791127	0.833935

Training Performance

Accuracy	Recall	Precision	F1
0.74516	0.86954	0.775913	0.820063

Testing Performance

- XGBoost model after tuning is giving a good and generalized performance.
- We have received the F1 score of 0.83 and 0.82 on the training and the test set, respectively.

Stacking Classifier

Accuracy	Recall	Precision	F1
0.770296	0.892554	0.790558	0.838465

Training Performance

Accuracy	Recall	Precision	F1
0.74529	0.879138	0.771399	0.821752

Testing Performance

- For the Stacking Classifier, the adaboost classifier, the tuned gradient boosting classifier and the tuned random forest models were used as the initial estimators while the tuned xgboost classifier was used as the final estimator.
- Stacking model has also given a good and generalized performance.
- The performance is comparable to the XGBoost model.
- We have received F1 scores of 0.83 and 0.81 on the training and test set, respectively.

greatlearning
Power Ahead

Happy Learning !