BIRLA INSTITUTE OF TECHNOLOGY , MESRA

Jaipur Campus

EC400M MINOR PROJECT

BANK LOAN APPROVAL ON CREDIT HISTORY OF APPLICANTS USING MACHINE LEARNING ALGORITHM

Name: Bhavya Singh (BTECH/25045/20)
Anisha Agarwal (BTECH/25081/20)

Branch: ECE, 7th Sem

Project Guide: Mr. Gaurav Jain

Acknowledgement

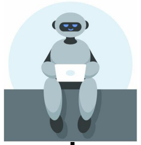
We would like to express my profound gratitude to my project guide, Mr. Gaurav Jain for his guidance and support during my thesis work. We benefited greatly by working under his guidance. His encouragement motivation and support has been invaluable throughout our studies at BIT, Mesra, Ranchi. We convey my sincere gratitude to Dept. of ECE, BIT, Mesra, Ranchi, for providing us various facilities needed to complete our project work. We would also like to thank all the faculty members of ECE department who have directly or indirectly helped during the course of the study.
Finally, We must express my very profound gratitude to our parents for providing us with unfailing support and continuous encouragement throughout the years of our study. This accomplishment would not have been possible without them. Our apologies and heartfelt gratitude to all who have assisted us yet have not been acknowledged by name.

Thank you.

Abstract

This project aims to dive into the crucial realm of financial losses that occur in credit card fraud transactions. Focusing on the need of real-world application, this project overview combines Exploratory Data Analysis (EDA) techniques with risk analysis for mitigating the financial risks associated with lending to customers. The project employs EDA to analyze data patterns, ensuring that loans are not being rejected for applicants who have the capability to repay. The company faces two risk scenario overall which are rejecting loans to creditworthy applicants which might results in business loss or while approving loans to potential defaulters and that leads to financial loss. By using comprehensive EDA techniques, this project overview aims to identify patterns that indicates applicant's likelihood to repay. The study extends its exploration by applying Machine Learning (ML) algorithms for Fraud Classification, specifically addressing missing credit history in the loan approval process.

Index



Introduction

Loan approval is a critical decision impacting both individuals and financial institutions.

- ➔ Applicant is capable of repaying the loan?
- ➔ Why EDA and ML?
  - ➢ Role
  - ➢ Algorithms' Advantage

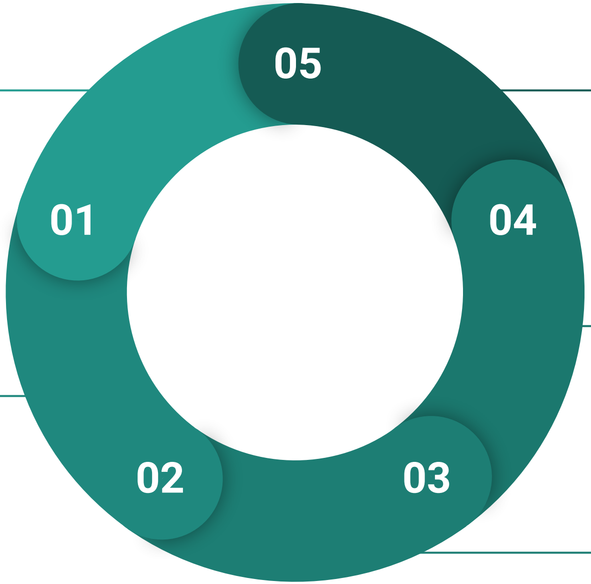
Libraries Used

Scikit-learn (Sklearn)

several modules such as RandomForestClassifier, LogisticRegression, SVC, train_test_split, GridSearchCV etc

Matplotlib

Various Plots including box plots



NumPy

for numerical operations and array manipulations.

Pandas

For data manipulation and analysis.

Seaborn

Creating count plots and heatmaps to visualise data distribution and correlation

Methodology

About Dataset

- Information about applicants.
- Have 122 attributes .

Target Attribute:

- 1: Defaulter cases
- 0: Non-defaulter cases

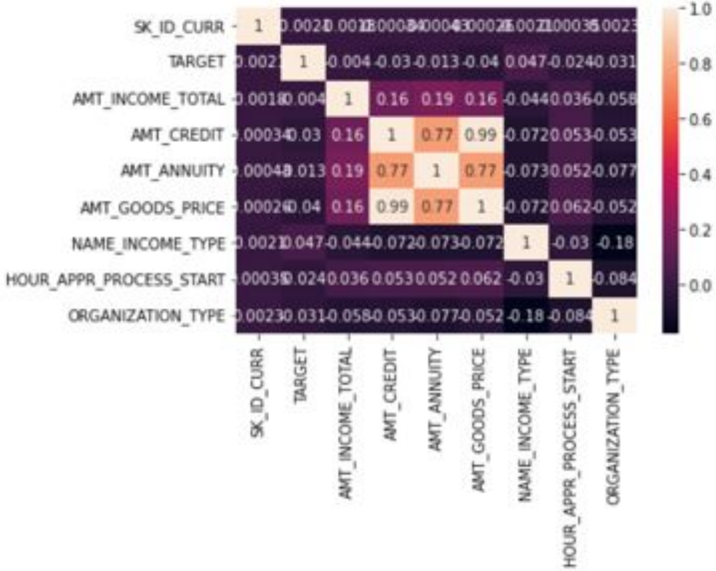
- Analyzing applicant information to predict default cases.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	TARGET	SK_ID_CURR	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_TYPE_SUITE	NAME_IN(
2	1	100002	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompanied	Working
3	0	100003	Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Family	State serva
4	0	100004	Revolving loans	M	Y	Y	0	67500	135000	6750	135000	Unaccompanied	Working
5	0	100006	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccompanied	Working
6	0	100007	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccompanied	Working
7	0	100008	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, partner	State serva
8	0	100009	Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Unaccompanied	Commerci
9	0	100010	Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Unaccompanied	State serva
10	0	100011	Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Children	Pensioner
11	0	100012	Revolving loans	M	N	Y	0	135000	405000	20250	405000	Unaccompanied	Working
12	0	100014	Cash loans	F	N	Y	1	112500	652500	21177	652500	Unaccompanied	Working
13	0	100015	Cash loans	F	N	Y	0	38419.155	148365	10678.5	135000	Children	Pensioner
14	0	100016	Cash loans	F	N	Y	0	67500	80865	5881.5	67500	Unaccompanied	Working
15	0	100017	Cash loans	M	Y	N	1	225000	918468	28966.5	697500	Unaccompanied	Working
16	0	100018	Cash loans	F	N	Y	0	189000	773680.5	32778	679500	Unaccompanied	Working
17	0	100019	Cash loans	M	Y	Y	0	157500	299772	20160	247500	Family	Working
18	0	100020	Cash loans	M	N	N	0	108000	509602.5	26149.5	387000	Unaccompanied	Working
19	0	100021	Revolving loans	F	N	Y	1	81000	270000	13500	270000	Unaccompanied	Working
20	0	100022	Revolving loans	F	N	Y	0	112500	157500	7875	157500	Other_A	Working
21	0	100023	Cash loans	F	N	Y	1	90000	544491	17563.5	454500	Unaccompanied	State serva
22	0	100024	Revolving loans	M	Y	Y	0	135000	427500	21375	427500	Unaccompanied	Working
23	0	100025	Cash loans	F	Y	Y	1	202500	1132573.5	37561.5	927000	Unaccompanied	Commerci
24	0	100026	Cash loans	F	N	N	1	450000	497520	32521.5	450000	Unaccompanied	Working
25	0	100027	Cash loans	F	N	Y	0	83250	239850	23850	225000	Unaccompanied	Pensioner
26	0	100029	Cash loans	M	Y	N	2	135000	247500	12703.5	247500	Unaccompanied	Working
27	0	100030	Cash loans	F	N	Y	0	90000	225000	11074.5	225000	Unaccompanied	Working
28	1	100031	Cash loans	F	N	Y	0	112500	979992	27076.5	702000	Unaccompanied	Working
29	0	100032	Cash loans	M	N	Y	1	112500	327024	23827.5	270000	Family	Working
30	0	100033	Cash loans	M	Y	Y	0	270000	790830	57676.5	675000	Unaccompanied	State serva
31	0	100034	Revolving loans	M	N	Y	0	90000	180000	9000	180000	Unaccompanied	Working
32	0	100035	Cash loans	F	N	Y	0	292500	665892	24592.5	477000	Unaccompanied	Commerci
33	0	100036	Cash loans	F	N	Y	0	112500	512064	25033.5	360000	Family	Working
34	0	100037	Cash loans	F	N	N	0	90000	199008	20893.5	180000	Unaccompanied	Working
35	0	100039	Cash loans	M	Y	N	1	360000	733315.5	39069	679500	Unaccompanied	Commerci

application_data

Data Preprocessing

- Dropped unwanted columns, resulting in 9 remaining columns.
- Handled missing and null values using mean, median, and mode.
- Split numerical and categorical values.
- Undertook undersampling and outlier detection.
- Examined correlations between variables.



Feature Selection

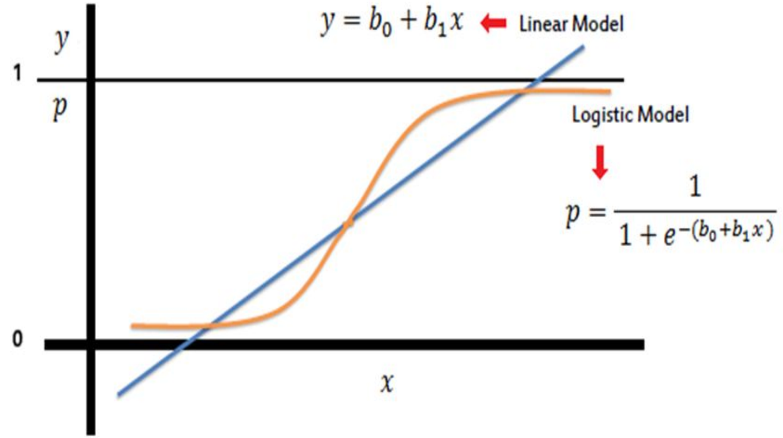
- Feature selection involves reducing the number of input variables in a predictive model.
  - X includes all attributes except the "Target" attribute.
  - Y includes only the "Target" attribute.

- The dataset is split into training and testing datasets:
  - X_train, X_test, y_train, y_test.

Algorithms and Results Obtained

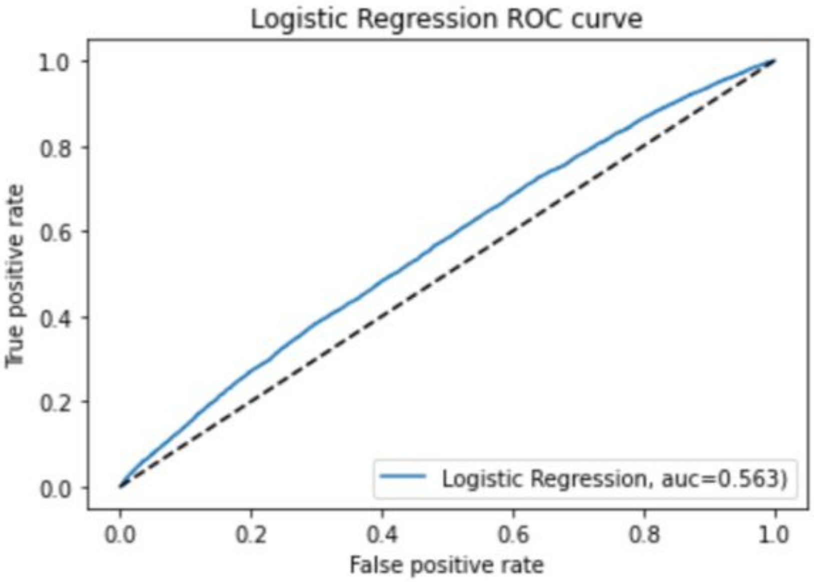
1. Logistic Regression

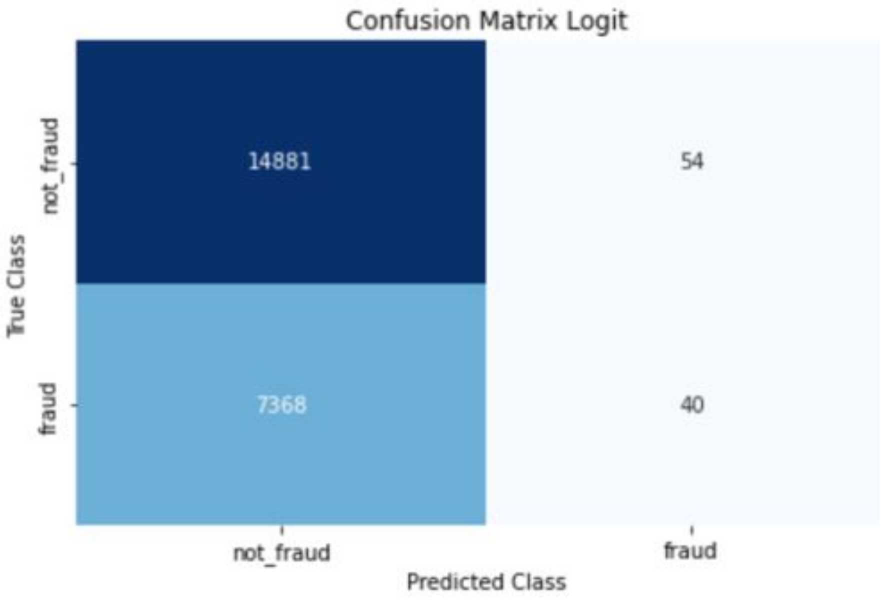
- It is a method for solving binary classification problem
- Uses sigmoid function
- If the sigmoid output is close to 1, it implies a high probability that the loan application will be approved.
- If the sigmoid output is close to 0, it indicates a high probability of rejection.



- Accuracy, Precision, Recall , F1 Score values are determined.
- Confusion Matrix and ROC curve are plotted.

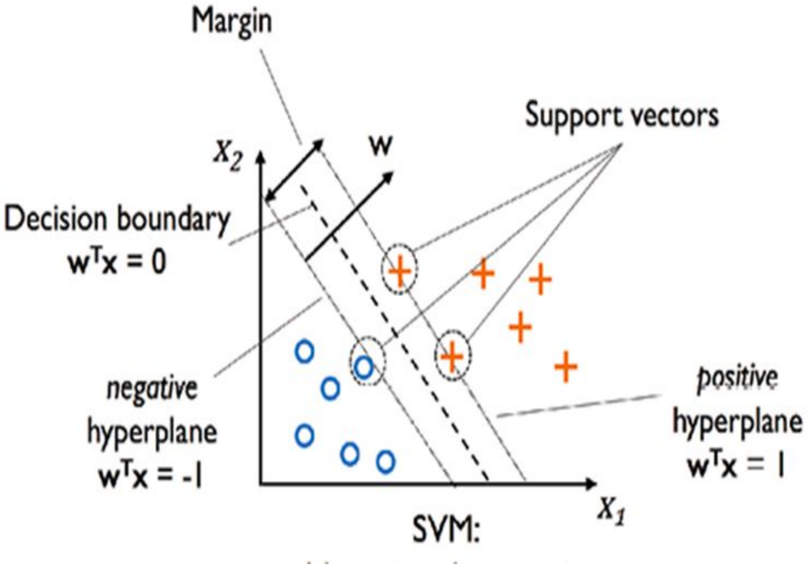
```
Accuracy Logit: 0.663026451237524
Precision Logit: 0.4444444444444444
Recall Logit: 0.005319856363878175
F1 Score Logit: 0.010513865159679328
```





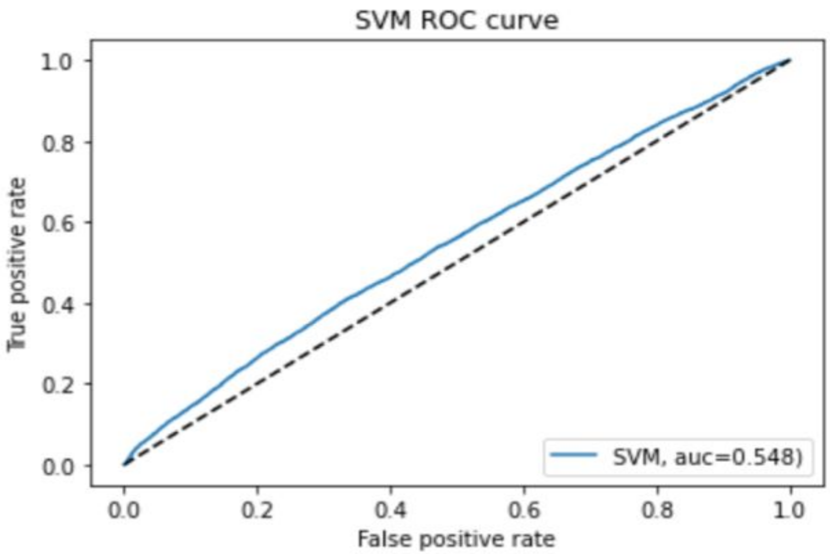
2. Support Vector Machine(SVM)

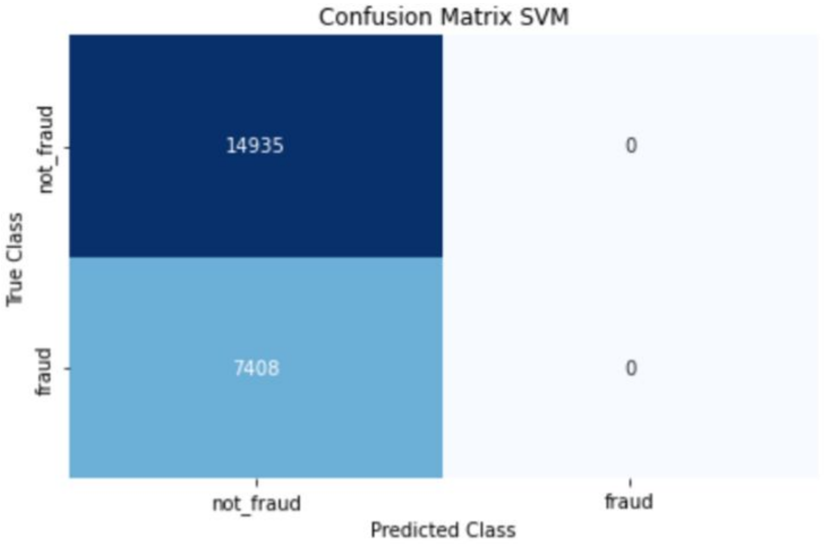
- Picture a plot with income on one side and credit score on the other.
- The hyperplane is the best line separating approved and denied applications.
- The margin is the space around this line, and approval or denial is decided based on which side an application falls.



- Accuracy, Precision, Recall , F1 Score values are determined.
- Confusion Matrix, AUC and ROC curve are plotted.

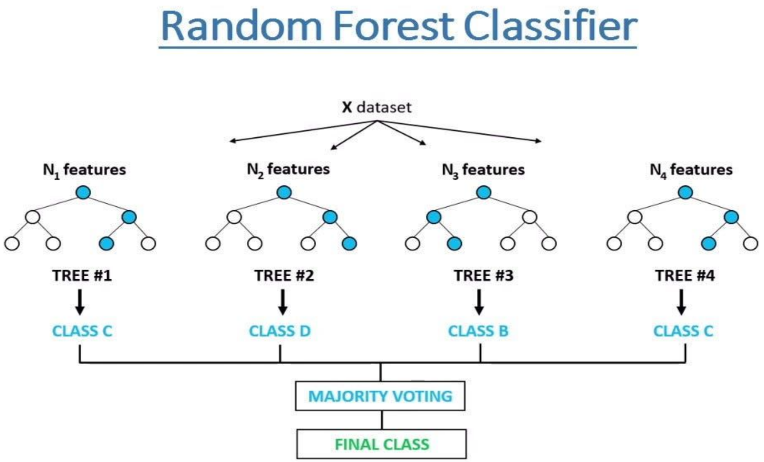
```
Accuracy SVM: 0.6634740187083202
Precision SVM: 1.0
Recall SVM: 0.0
F1 Score SVM: 0.0
```





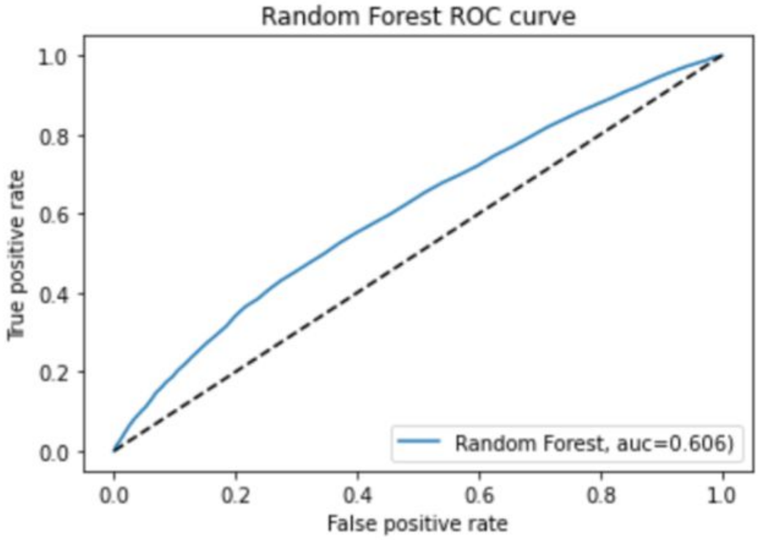
3. Random Forest

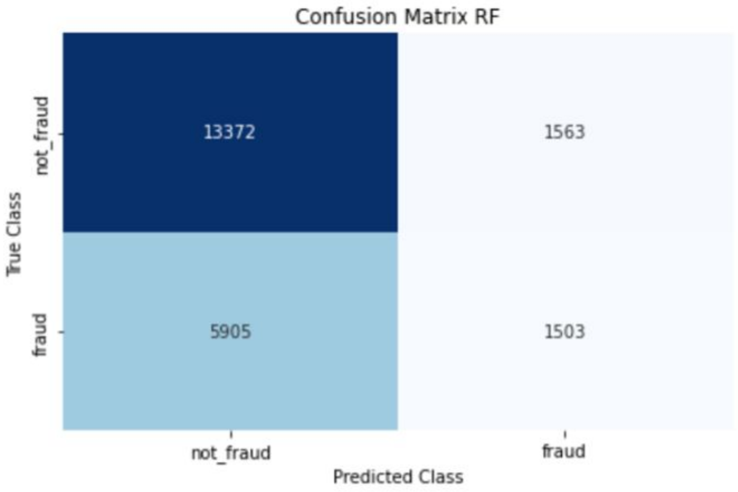
- Team of decision trees collaborates for loan approval decisions.
- Each tree specializes in different aspects like income, credit score, and employment history.
- Independent predictions are made by each tree, and the final decision is based on a majority vote, classifying the loan application



- Accuracy, Precision, Recall , F1 Score values are determined.
- Confusion Matrix, AUC and ROC curve are plotted

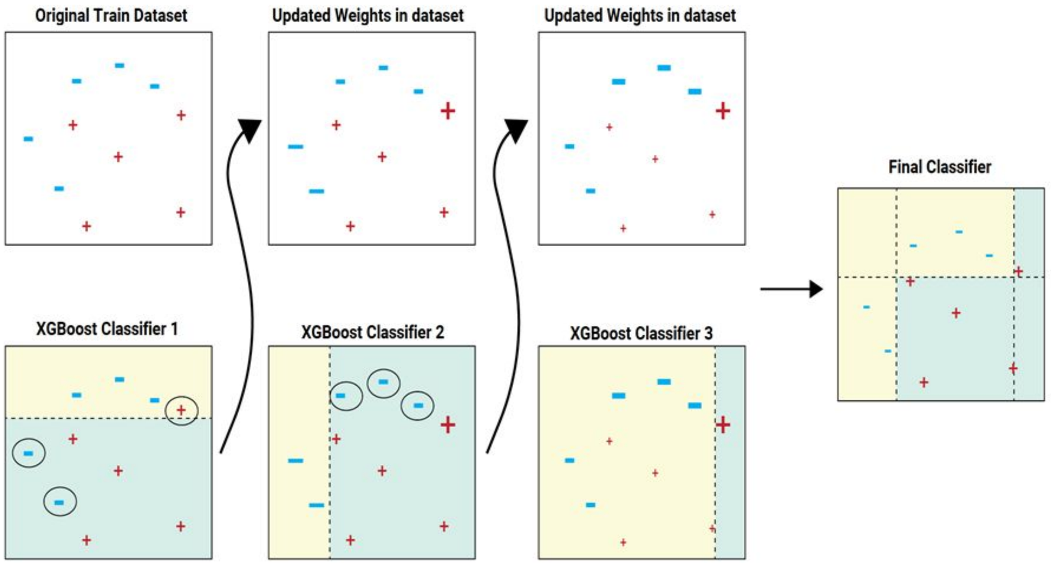
```
Accuracy RF: 0.6656223425681421
Precision RF: 0.5076677316293929
Recall RF: 0.21133129405506051
F1 Score RF: 0.29843177763170253
```





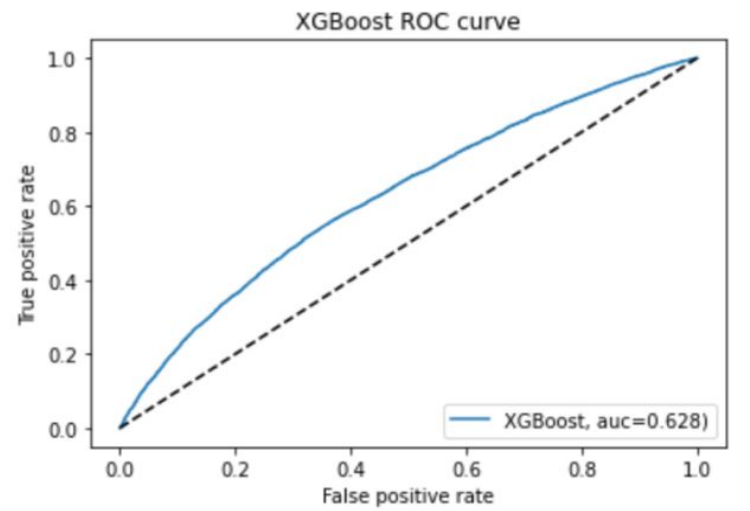
4. XGboost Classifier

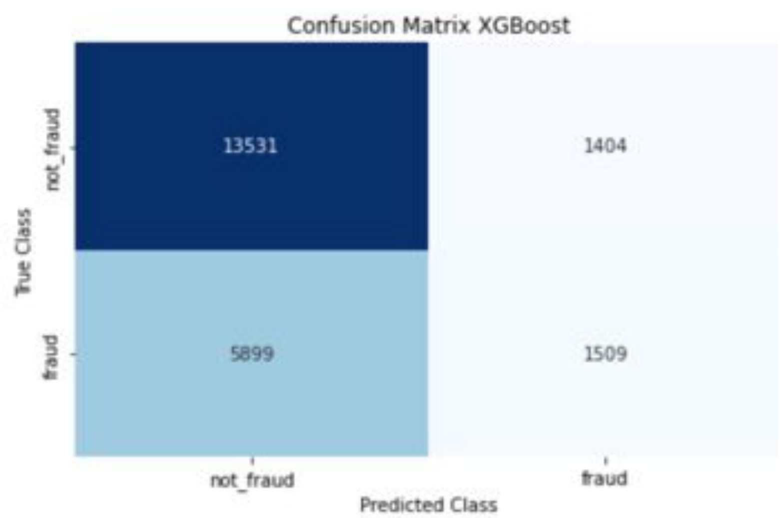
- XGBoost builds a series of decision-makers for accurate loan predictions.
- It's like creating a team of experts, where each corrects the mistakes of the previous one, improving accuracy with each iteration.



- Accuracy, Precision, Recall , F1 Score values are determined.
- Confusion Matrix, AUC and ROC curve are plotted
- Gave best accuracy among other models.

```
Accuracy XGB: 0.6739918542720315
Precision XGB: 0.540281107302022
Recall XGB: 0.2096023407368012
F1 Score XGB: 0.3020314296653895
```





Conclusion

The choice of the "better" model depends on the specific goals and trade-offs relevant to the loan approval scenario.

Minimizing False Positives

Metrics: Precision, Confusion Matrix, ROC Curve, AUC Score

Variables: model1, y_pred_logit

Logistic Regression

Capturing as Many Defaults as Possible

Metrics: Recall, Confusion Matrix, ROC Curve, AUC Score

Variables: model2, y_pred_svm

Support Vector Machine

Balancing Precision and Recall

Metrics: F1 Score, Confusion Matrix, ROC Curve, AUC Score

Variables: model3, y_pred_rf

Random Forest

Overall Performance

Metrics: Accuracy, Confusion Matrix, ROC Curve, AUC Score

Variables: model4, y_pred_xgb

XGBoost

References

[1] Gowthami, K., et al. Credit card fraud detection using logistic regression. Journal of Engineering Sciences, 11.

[2] Jayakumar et.al., A New Procedure of Clustering based on Multivariate Outlier Detection. Journal of Data Science 2013; 11: 69-84

[3] Sheikh, Mohammad Ahmad, Amit Kumar Goel, and Tapas Kumar. "An approach for prediction of loan approval using machine learning algorithm." 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE, 2020.

[4] X.Frencis Jensy, V.P.Sumathi, Janani Shiva Shri, "An exploratory Data Analysis for Loan Prediction based on nature of clients", International Journal of Recent Technology and Engineering (IJRTE),Volume-7 Issue-4S, November 2018.

[5] X. Zhang, Y. Han, W. Xu, and Q. Wang, "HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture," Inf. Sci., vol. 557, pp. 302–316, May 2021, doi: 10.1016/j.ins.2019.05.023.

[6] Source Code: https://drive.google.com/drive/folders/16RQztUqCfJOlbooHqYlJrp6Q7i L65uZB

[7] www.AnalyticsVidya.com


Thank
you