Predicting Credit Risk

TEAM NAME: FRAUD DETECTORS

Name	Student ID
Dontireddy Jyotsna	16370175
Emani Nandhitha	16358354
Gopu Sakshitha	12612128
Munnangi Geetha Reddy	16358418

Objective

- To design and evaluate machine learning models that can forecast the likelihood of a loan applicant defaulting, using financial, demographic, and credit-related attributes.

Introduction

- By leveraging machine learning models, we can better identify high-risk applicants, whose loan will be rejected. Since we had unbalanced data we have used another dataset similar to this and performed the development of models to compare the results for verification.
- Models Used: Random Forest, Decision Tree Classifier, Xtreme Gradient Boosting algorithm, Light Gradient Boosting, AdaBoost, Logistic regression.

Dataset Description

- We are having 7 features and 1 target variable for the first dataset and 11 features and 1 target variable for second dataset.
- Dataset

```
import numpy as np
import pandas as pd
data = pd.read_csv("/content/lending_data.csv")
data.head()
```

	loan_size	interest_rate	borrower_income	debt_to_income	num_of_accounts	derogatory_marks	total_debt	loan_status
0	10700.0	7.672	52800	0.431818	5	1	22800	0
1	8400.0	6.692	43600	0.311927	3	0	13600	0
2	9000.0	6.963	46100	0.349241	3	0	16100	0
3	10700.0	7.664	52700	0.430740	5	1	22700	0
4	10800.0	7.698	53000	0.433962	5	1	23000	0

Data Preprocessing

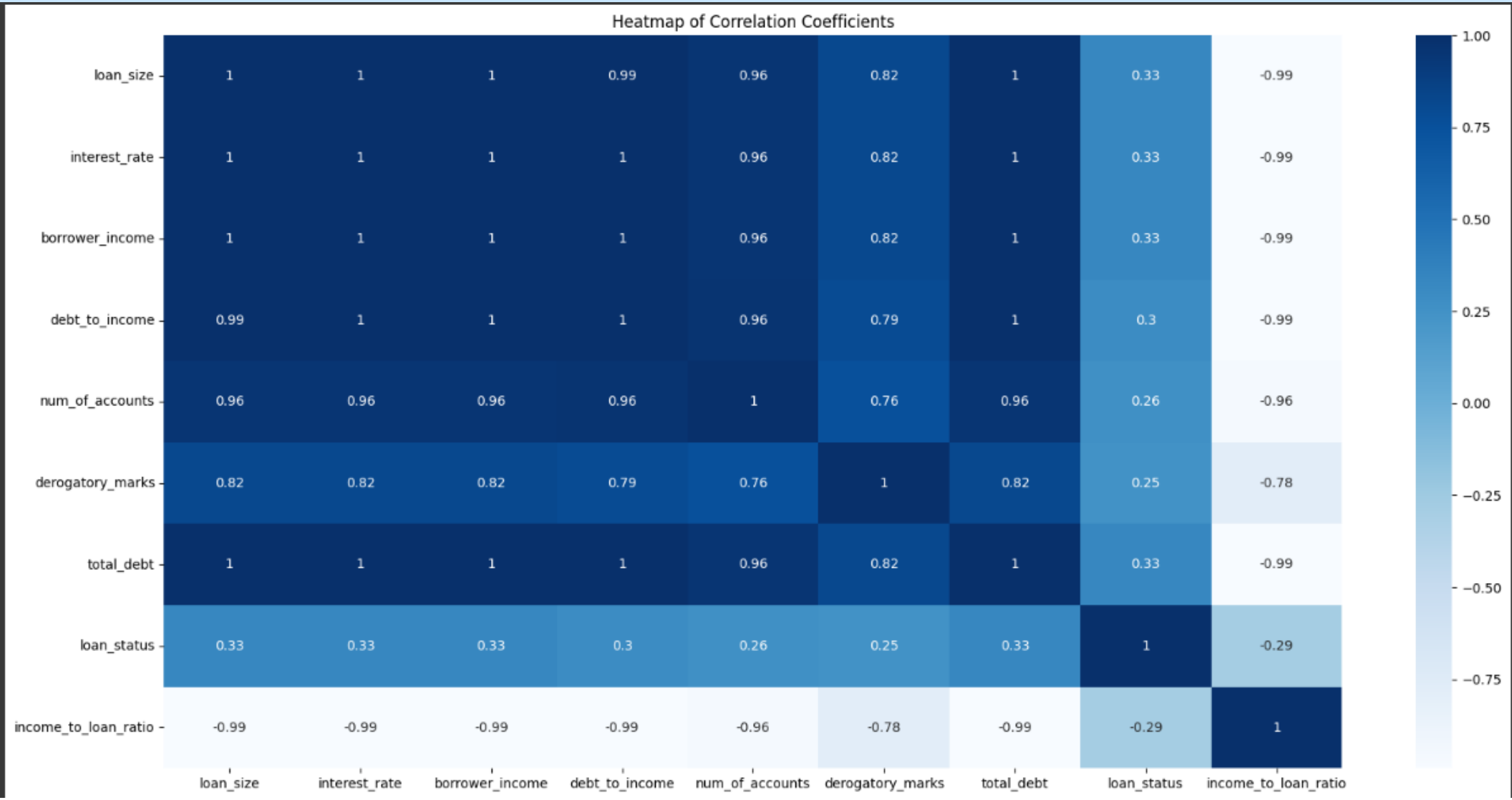
- Checking for the missing values
- Identifying the outliers by visual representation and handling the outliers.
- Implemented IQR method but, since our data is imbalanced out of 5% features from class 1 we are loosing 3% features.
- So, we used SMOTE library for imbalanced data which will help balance the class distribution.

```
After SMOTE on training data:
X_train_sm shape: (120056, 8)
y_train_sm distribution:
 loan_status
0    60028
1    60028
Name: count, dtype: int64
```

Exploratory Data Analysis

- Performed feature engineering and created a new feature with the help of existing features (borrower_income and loan_size).
- Visualizing each feature's distribution and its correlation with the loan_status to help understand feature importance and data patterns.
- Represented a correlation matrix on how each feature is correlated with the other features.

Correlation Matrix:



Model Development

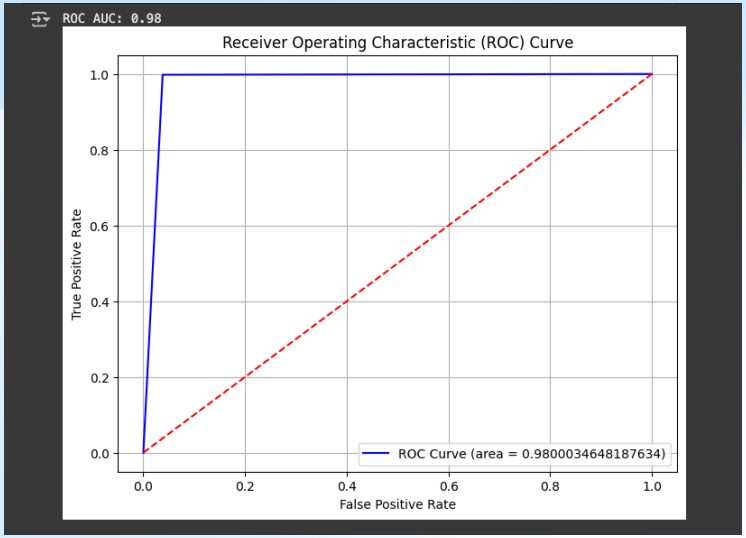
Random forest

- It captures class 1 extremely well (high recall) but often mislabels class 0 as class 1 (lower precision for class 1) and accuracy is 96%.

```
[[14438   570]
 [    1   499]]
              precision    recall  f1-score   support

           0       1.00      0.96      0.98     15008
           1       0.47      1.00      0.64       500

    accuracy                           0.96     15508
   macro avg       0.73      0.98      0.81     15508
weighted avg       0.98      0.96      0.97     15508

The  accuracy with the Random Forest model is 0.9631802940417848
```

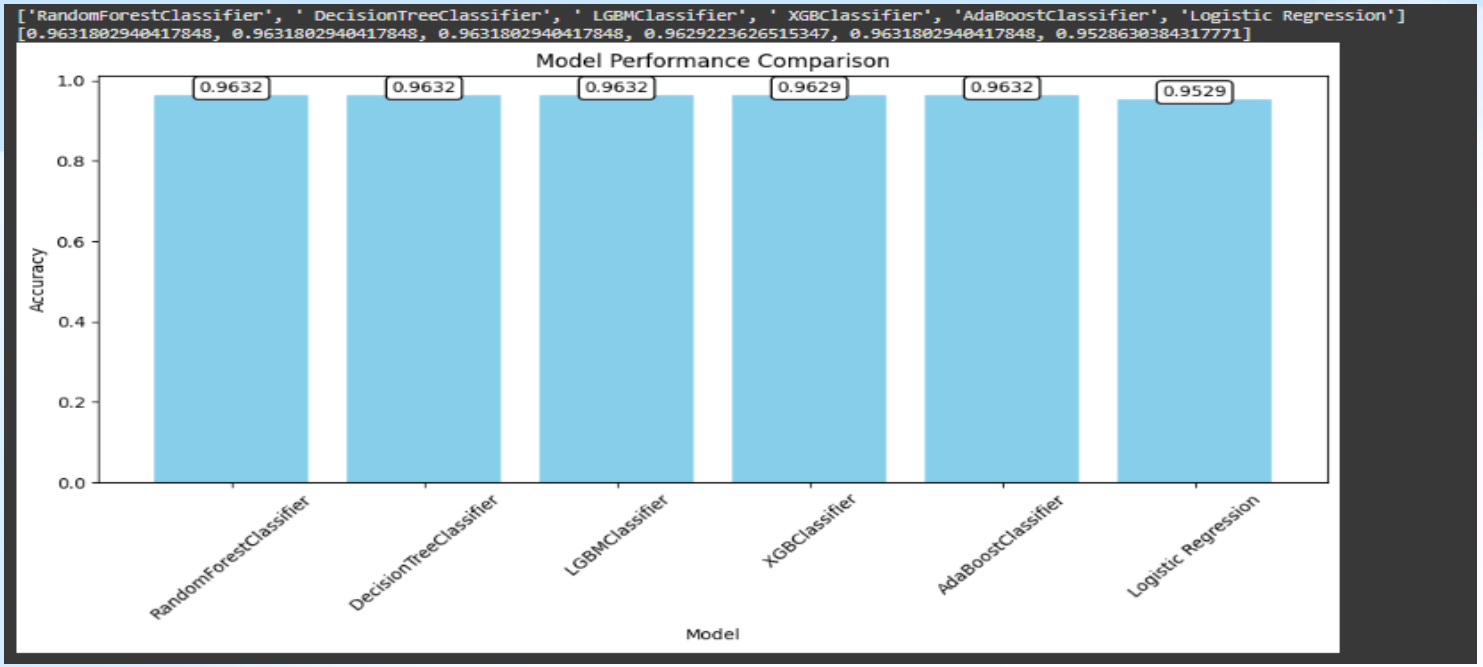
ROC Curve

- The ROC Curve is of 0.98 which is model indicate that your performs exceptionally well at distinguishing between the two classes.



Comparison of models with dataset

- After implementing different models by comparing them we can say that Random forest performed best among all of them.



Conclusion

- Comprehensive data preprocessing and feature engineering significantly enhanced model accuracy
- Addressing class imbalance using SMOTE improved the model's sensitivity to minority class predictions
- Multiple models were evaluated, with Random Forest delivering the best performance in terms of both accuracy (96%) and ROC AUC (0.98).


Thank You!