# Vehicle Loan Prediction


GROUP 3

Mrunalini Devineni

Sara Sanchez

Aihan Liu

GITHUB LINK

# Introduction

- Financial institutions incur significant losses due to the default of Vehicle Loans. This has led to the constricting of vehicle loan underwriting and increased vehicle loan rejection rates
- A Credit Score: good or bad.
- Forecasted probability of default

# Want to get a car??

# Want to get a car??

Apply to car loan through the bank institution

+

Bank decides to give you a loan! But what factors affect their decision ?is it a YAY or NAY!!

→

You repay the loan over monthly installments including interest

Liquid Assets

Debt to Income Ratio

Credit score

Many more

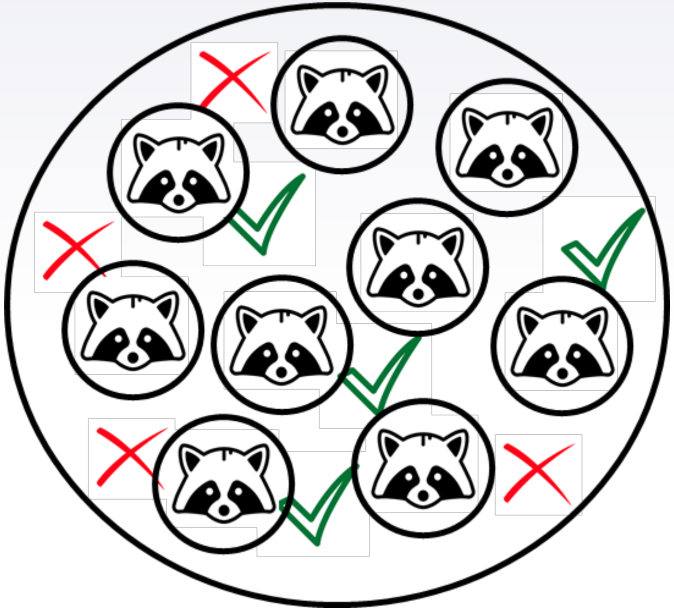
collateral

Income

# Score

Allows forecasting the probability of default based on the client's profile information.

**Goal:** Distinguish between good and bad profiles.



# Dataset

L&T company data set (kaggle)

- Loanee Information
  - (Demographic data like age, Identity proof etc.)
- Loan Information
  - (Disbursal details, loan to value ratio etc.)
- Bureau data & history
  - (Bureau score, number of active accounts, the status of other loans, credit history etc.)
- 40 variables
- 233 154 observations

SMART Question: What are the features that influence loan default based on customer’s profile information?

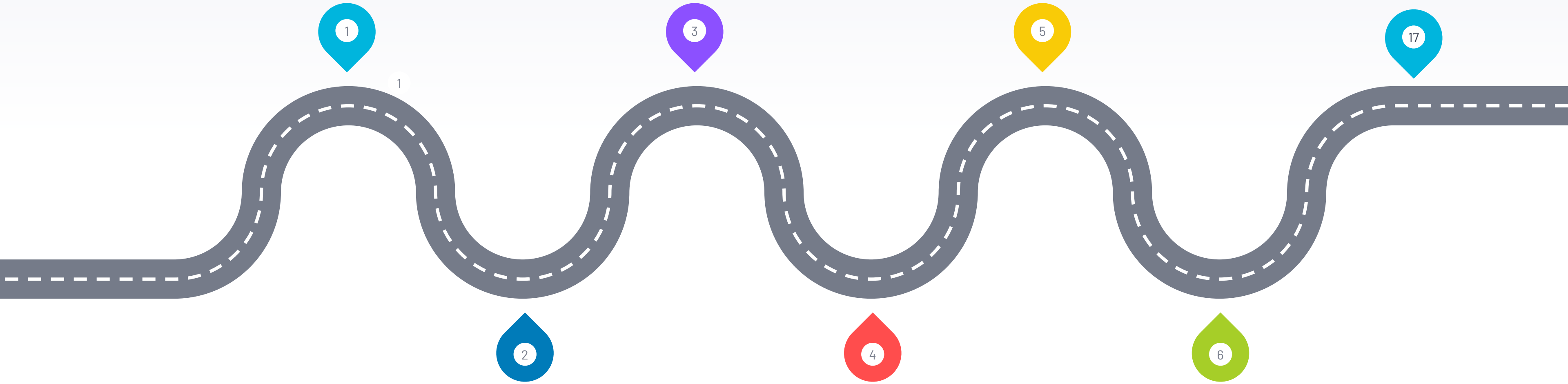
# Roadmap

1.collected the dataset from Kaggle

3.EDA(Exploratory Data Analysis)

5.Model Training and saving them using pickle

7.Integrate the models and EDA to the GUI (pyqt5)

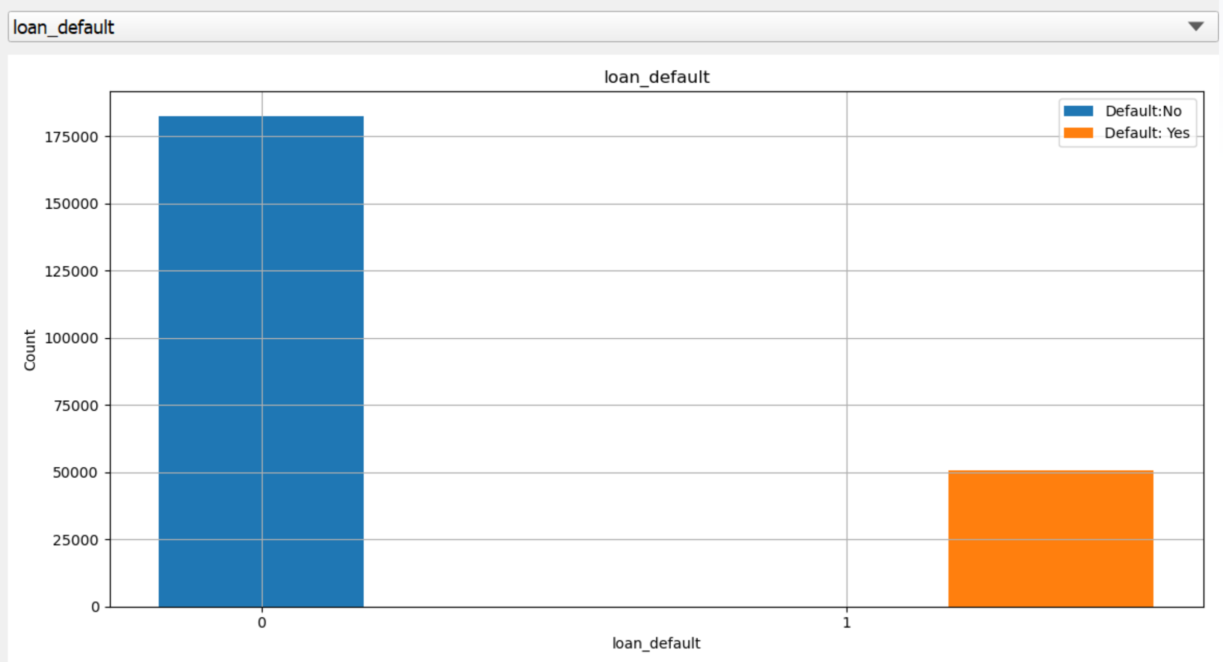



2.Loaded the dataset and performed necessary cleaning on it

4.Data Preprocessing-scale,Feature Engineering,Data Balancing

6.Prediction and evaluation of Target

# Exploratory Data Analysis (EDA)

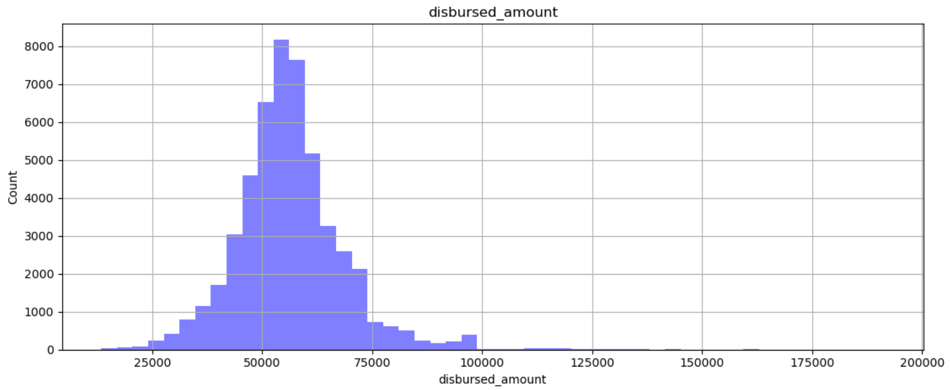



Clearly the graph proves the problem statement that the number of “No’s” to loan default is much higher than the “Yes”, the data is unbalance in this case.

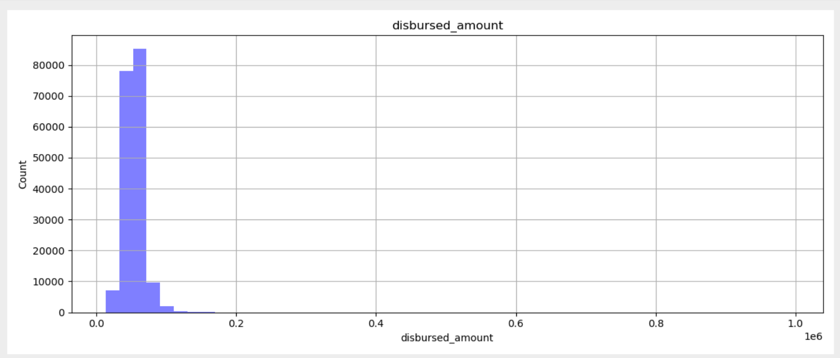
Apply SMOTE method to balance dataset.

# Disbursed Amount

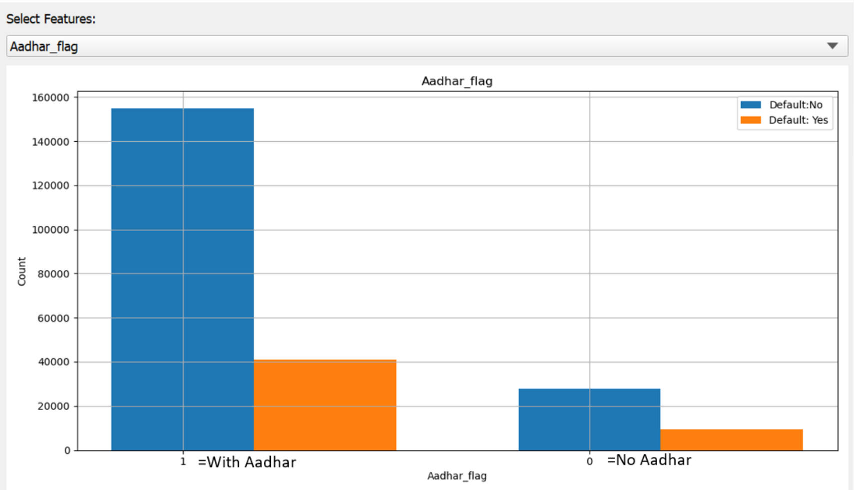
Loan default considered to be “yes”

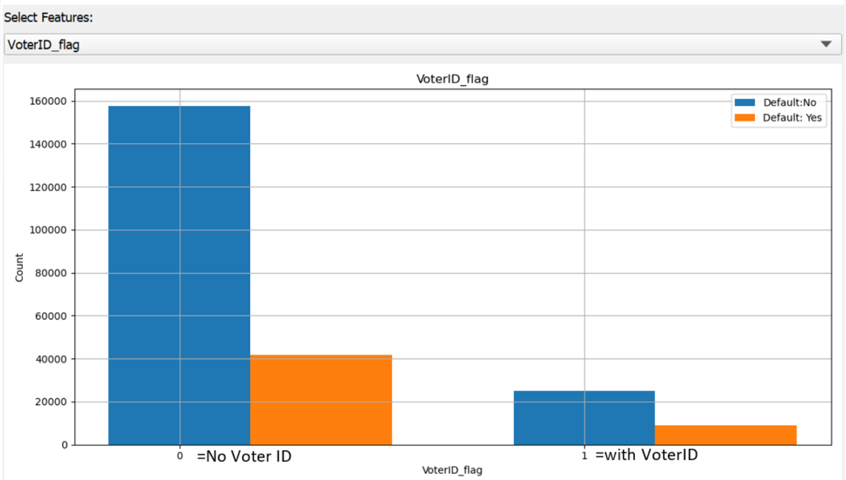



Loan default considered to be “No”



# Aadhar and Voter ID



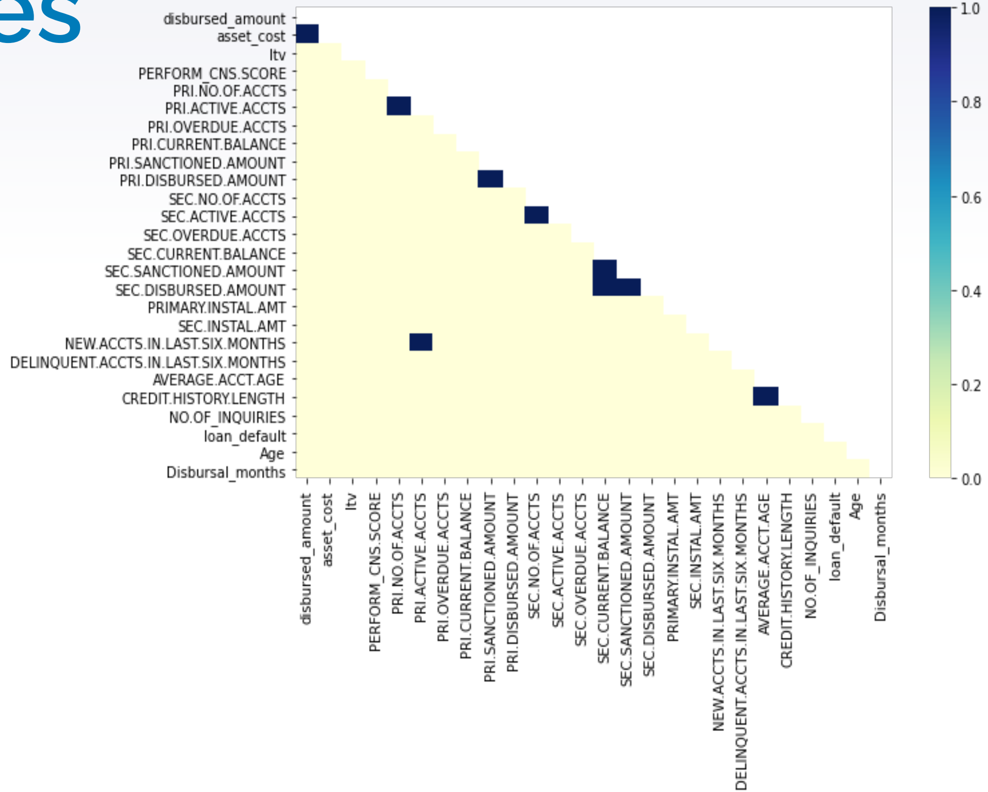

The count of people having Aadhar Card as address proof and not having it.

The count of people having Voter ID as address proof and not having it

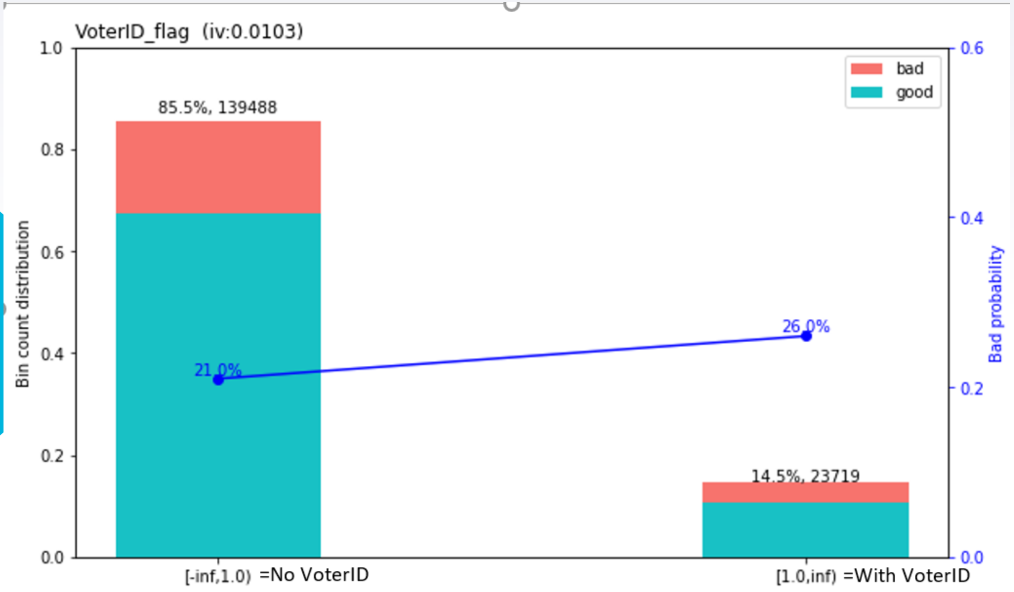
# Relationship between variables

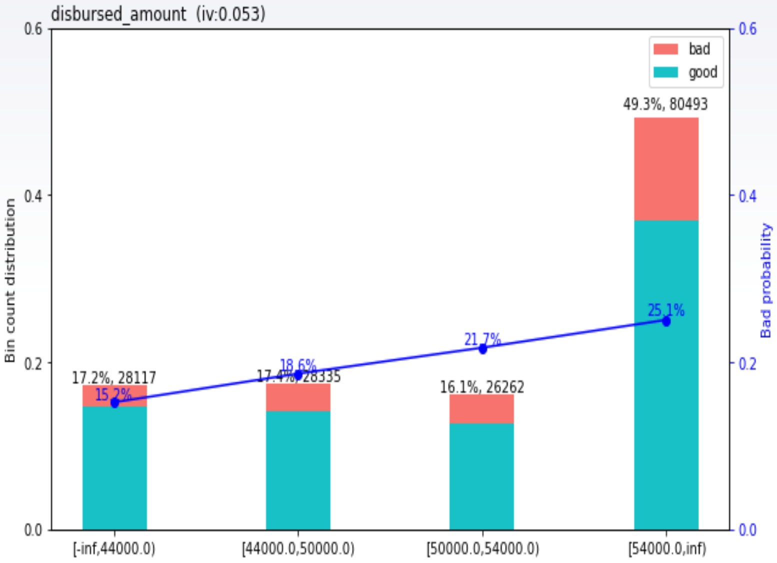
Correlated features:

- Asset cost<->Disbursed amount
- Sanctioned Amount<->Disbursed amount
- Credit history length<-> avg account age



# Weight of Evidence




These two variables can distinguish between profiles because, as we can see from the default rate, this is different for those who have or do not have a voter ID and those who have different disbursed amounts.

# Data Preprocessing

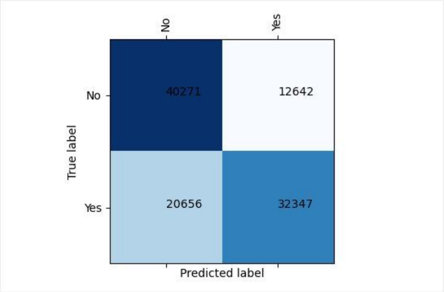
- Dropped the missing values
- Formatted date
- Calculated Age and Disbursal months from the existing features
- Feature elimination
- Feature Engineering
- Created buckets to reduce the number of categories
- Balanced the data with hyperparameter of each model

| EI | AVERAGE.ACCT.AGE | CREDIT.HISTORY.LENG |
|---|---|---|
| 0 | 0yrs 0mon | 0yrs 0mon |
| 1 | 1yrs 11mon | 1yrs 11mon |
| 0 | 0yrs 0mon | 0yrs 0mon |
| 0 | 0yrs 8mon | 1yrs 3mon |
| 0 | 0yrs 0mon | 0yrs 0mon |
| 0 | 1yrs 9mon | 2yrs 0mon |
| 0 | 0yrs 0mon | 0yrs 0mon |
| 0 | 0yrs 2mon | 0yrs 2mon |

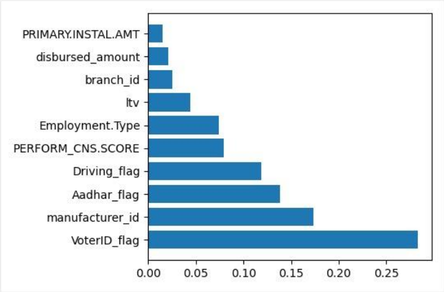
| Date.of.Birth | Employment.Type | DisbursalDate |
|---|---|---|
| 1/1/1984 | Salaried | 3/8/2018 |
| 31-07-85 | Self employed | 26-09-18 |
| 24-08-85 | Self employed | 1/8/2018 |
| 30-12-93 | Self employed | 26-10-18 |
| 9/12/1977 | Self employed | 26-09-18 |
| 8/9/1990 | Self employed | 19-09-18 |
| 1/6/1988 | Salaried | 23-09-18 |
| 4/10/1989 | Salaried | 16-09-18 |
| 15-11-91 | Self employed | 5/9/2018 |

# Decision Tree

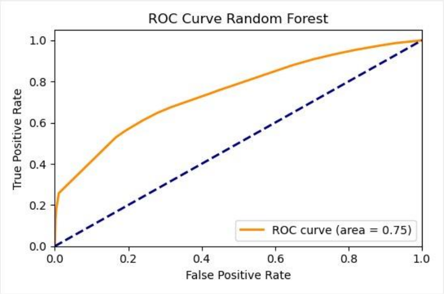
Confusion Matrix



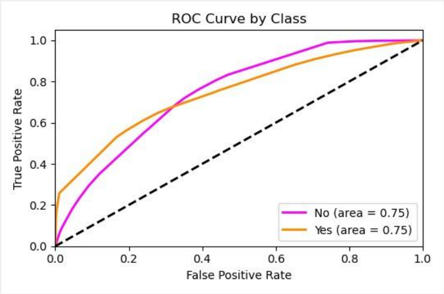

Importance of Features




ROC Curve




ROC Curve by Class



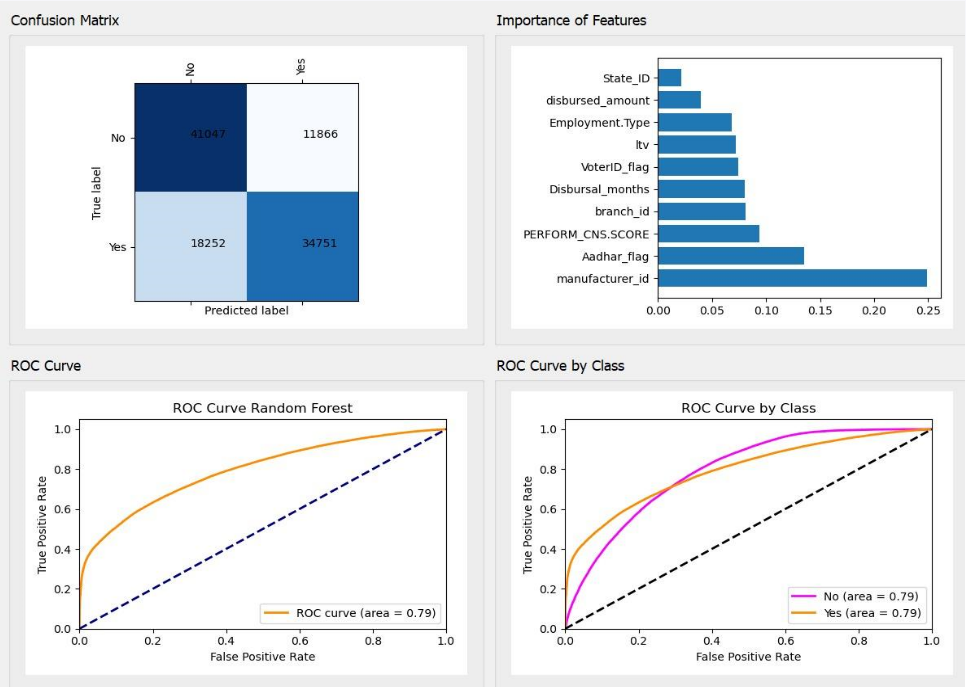

Measurements:

| | |
|---|---|
| Accuracy: | 68.56187922504627 |
| Precision: | 71.8997977283336 |
| Recall: | 61.02862102145161 |
| F1 Score: | 66.01967507551636 |

Other Models Accuracy:

| | |
|---|---|
| Logistic: | 66.60277956116167 |
| Random Forest: | 71.56425846897541 |
| Gradient Boosting: | 76.82408701234941 |

# Random Forest



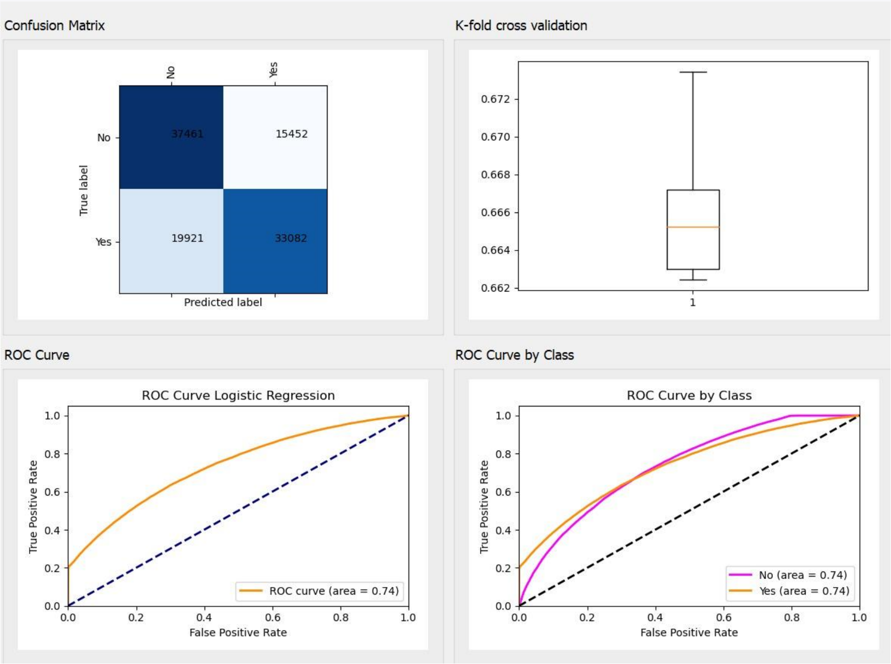

**Measurements:**

| | |
|---|---|
| Accuracy: | 71.56425846897541 |
| Precision: | 74.54576656584507 |
| Recall: | 65.56421334641436 |
| F1 Score: | 69.76711503714115 |

**Other Models Accuracy:**

| | |
|---|---|
| Logistic: | 66.60277956116167 |
| Gradient Boosting: | 76.82408701234941 |
| Decision tree: | 68.56187922504627 |

# Logistic Regression



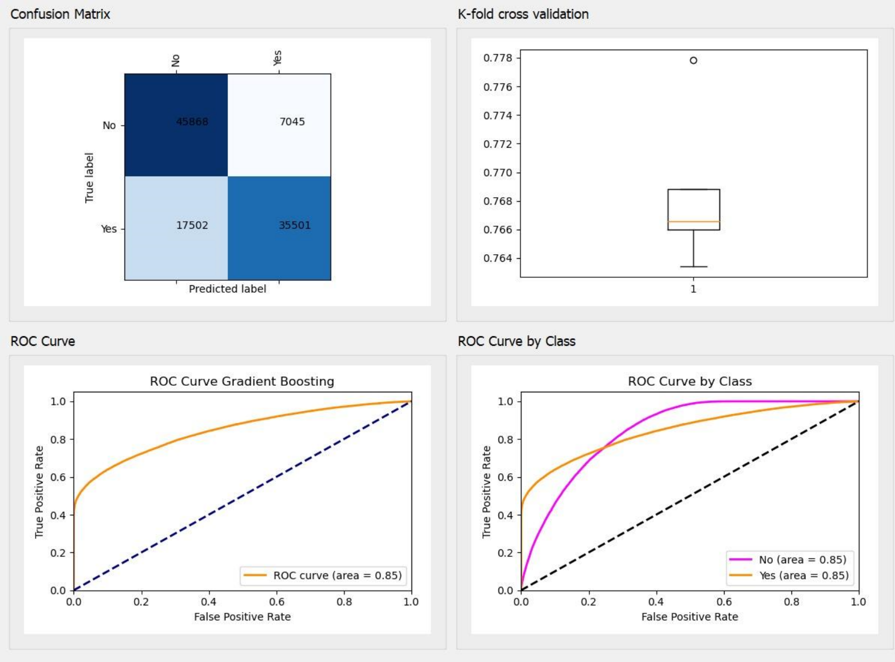

**Measurements:**

| | |
|---|---|
| Accuracy: | 66.60277956116167 |
| Precision: | 68.16252524003791 |
| Recall: | 62.415334981038804 |
| F1 Score: | 65.16245309591578 |

**Other Models Accuracy:**

| | |
|---|---|
| Decision Tree: | 68.56187922504627 |
| Gradient Boosting: | 76.82408701234941 |
| Random Forest: | 71.56425846897541 |

# Gradient Boosting




| Measurements: | | Other Models Accuracy: | |
|---|---|---|---|
| Accuracy: | 76.82408701234941 | Decision Tree: | 68.56187922504627 |
| Precision: | 83.4414516053213 | Logistic Regression: | 66.60277956116167 |
| Recall: | 66.97922759089107 | Random Forest: | 71.56425846897541 |
| F1 Score: | 74.30951658311442 | | |

# and table to compare Models

| Metrics | Random Forest | Logistic Regression | Gradient Boosting | Decision Tree |
|---|---|---|---|---|
| Accuracy | 71.56% | 66.60% | 76.82% | 68.56% |
| Precision | 74.54% | 68.16% | 83.44% | 71.89% |
| Recall | 65.56% | 62.41% | 66.97% | 61.02% |
| F-1 score | 69.76% | 65.16% | 74.30% | 66.01% |

# Conclusion:

The features that most affect the loan default are: Adahar_flag, voterID_flag, perform_cns.score, driving_flag, ltv, employer type and state_id.

The model we trained that have highest accuracy is Gradient boosting, however it has a lowest recall.

Other work that may improve the accuracy of this dataset is to apply PCA or other feature selection techniques. We can also use other ensemble method to get a better results. .

# THANKS!

***Any questions????***