INSTITUTE FOR ADVANCED

COMPUTING AND

SOFTWARE DEVELOPMENT

AKURDI, PUNE

Documentation On

**“NBFI Vehicle Loan Lending Prediction”**

PG-DBDA SEP 2023

*Submitted By:*

**Group No: 14**

**Amrut Amble Patil 230906**

**Ganesh Waman 230916**

# Mrs. Priti Take Mr. Rohit Puranik

# Project Guide Centre Coordinator

# 

# ACKNOWLEDGEMENT

We would like to express our sincere gratitude to everyone who has contributed to the completion of our project.

First and foremost, we would like to thank our project guide **Mrs Priti Take** Mam for their constant guidance and support throughout the project. We extend our sincere thanks to our respected centre coordinator, **Mr Rohit Puranik** for allowing us to use the facilities available.

We would also like to express our appreciation to the faculty members of our department for their constructive feedback and encouragement. Their insights and suggestions have helped us to refine our ideas and announce the quality of our work.

Furthermore, we would like to thank our families and friends for their unwavering support and encouragement throughout our academic journey, their love and support have been a constant source of motivation and inspiration for us.

Thank you for all your valuable contributions to our project.

**Table of Contents**

[ACKNOWLEDGEMENT 2](#_heading=h.3znysh7)

[INTRODUCTION 5](#_heading=h.2et92p0)

[1.1 PROBLEM STATEMENT 5](#_heading=h.tyjcwt)

[1.2 ABSTRACT 5](#_heading=h.3dy6vkm)

[1.3 PRODUCT SCOPE 6](#_heading=h.1t3h5sf)

[1.4 AIM & OBJECTIVES 7](#_heading=h.4d34og8)

[1.](#_heading=h.2s8eyo1) OVERALL DESCRIPTION 8

[1.1](#_heading=h.17dp8vu) WORKFLOW OF PROJECT 8

[2.2 DATA DESCRIPTION 8](#_heading=h.3rdcrjn)

[2.3 Importing Dataset. 9](#_heading=h.26in1rg)

[2.4 Project Initiation: 11](#_heading=h.lnxbz9)

[2.5 Data Acquisition: 11](#_heading=h.35nkun2)

[2.6 Data Exploration and Understanding: 12](#_heading=h.1ksv4uv)

[2.7 Data Pre-processing: 13](#_heading=h.44sinio)

[2.8 Feature Engineering: 18](#_heading=h.3whwml4)

[2.9 Model Development: 19](#_heading=h.3as4poj)

[2.10 Model Evaluation: 20](#_heading=h.49x2ik5)

[2.11 Model Interpretation: 22](#_heading=h.3o7alnk)

[2.12 Model Deployment: 22](#_heading=h.23ckvvd)

[2.13 Documentation and Reporting: 22](#_heading=h.ihv636)

[2.](#_heading=h.32hioqz) REQUIREMENTS SPECIFICATION 23

[3.1 Hardware Requirement: 23](#_heading=h.1hmsyys)

[3.2 Software Requirement: 23](#_heading=h.41mghml)

[3.](#_heading=h.2grqrue) CONCLUSION 24

**TABLE OF FIGURE:**

[Figure 1 WorkFlow of Project](about:blank) 8

[Figure 2 Calculating Default Rate 13](#_heading=h.2jxsxqh)

[Figure 3 Visualzation of Default Rate 13](#_heading=h.z337ya)

[Figure 4 Outlier Handling 14](#_heading=h.3j2qqm3)

[Figure 5 Client\_income after Outliers Handling 14](#_heading=h.1y810tw)

[Figure 6 Correlation Analysis 1 15](#_heading=h.4i7ojhp)

[Figure 7 Correlation Analysis 2 16](#_heading=h.2xcytpi)

[Figure 8 Correlation Matrix 17](#_heading=h.1ci93xb)

[Figure 9 Code For Balancing Dataset using SMOTE 18](#_heading=h.2bn6wsx)

[Figure 10 Count of Balanced Classes 18](#_heading=h.qsh70q)

[Figure 11 Confusion Matrix 19](#_heading=h.1pxezwc)

[Figure 12 ROC Curve 20](#_heading=h.2p2csry)

[Figure 13 Accuracy vs Sensitivity vs Specificity 21](#_heading=h.147n2zr)

# INTRODUCTION

## 1.1 PROBLEM STATEMENT

**NBFI Vehicle Loan Lending Prediction**

## 1.2 ABSTRACT

Financial institutions incur significant losses due to the default of vehicle loans. This has led to the tightening up of vehicle loan underwriting and increased vehicle loan rejection rates. The need for a better credit risk scoring model is also raised by these institutions. This warrants a study to estimate the determinants of vehicle loan default. A financial institution has hired you to accurately predict the probability of Borrower /borrower defaulting on a vehicle loan in the first EMI (Equated Monthly Instalments) on the due date. Following Information regarding the loan and Borrowers are provided in the datasets:

• Borrower’s Information (Demographic data like age, income, Identity proof etc.)

• Loan Information (Disbursal details, amount, EMI, loan to value ratio etc.)

• Bureau data & history (Bureau score, number of active accounts, the status of other loans, credit history etc.)

Doing so will ensure that clients capable of repayment are not rejected and important determinants can be identified which can be further used for minimising the default rates.

## 1.3 PRODUCT SCOPE

1. Data Collection and Integration:
   * Gather and integrate diverse data sources, including customer profiles, credit histories, income details, employment stability, and macroeconomic indicators.
   * Implement mechanisms for real-time data updates to ensure model accuracy.
2. Predictive Modelling:
   * Develop machine learning algorithms to predict the likelihood of vehicle loan approval.
   * Utilize various models such as logistic regression, decision trees, random forests, and gradient boosting to achieve accurate predictions.
   * Consider ensemble methods for combining the strengths of multiple models.
3. Feature Selection and Engineering:
   * Identify and prioritize key features that significantly impact loan approval decisions.
   * Implement feature engineering techniques to enhance model performance.
4. Model Interpretability and Explainability:
   * Ensure transparency in model predictions to build trust among stakeholders.
   * Implement methods for explaining model decisions to assist in understanding the rationale behind loan approval predictions.
5. Validation and Evaluation:
   * Conduct rigorous validation processes using cross-validation techniques.
   * Evaluate models based on performance metrics such as accuracy, precision, recall, and F1 score.

## 1.4 AIM & OBJECTIVES

The aim for NBFI Vehicle Loan Lending Prediction is to leverage advanced analytics and machine learning techniques to enhance the efficiency and accuracy of the lending process within Non-Banking Financial Institutions (NBFI). The primary objectives include:

1. Optimizing Loan Approval Decisions:
   * Develop predictive models to assess the creditworthiness of loan applicants for vehicle financing.
   * Improve the accuracy of loan approval decisions by leveraging historical data and advanced analytics.
2. Reducing Risk and Enhancing Portfolio Quality:
   * Identify and evaluate risk factors associated with vehicle loan lending.
   * Implement predictive models to proactively identify potential defaults or high-risk applicants, thereby improving the overall quality of the loan portfolio.
3. Increasing Operational Efficiency:
   * Streamline the loan approval process by automating and optimizing decision-making using predictive models.
   * Reduce manual effort and processing time, leading to quicker and more efficient loan approvals.
4. Enhancing Customer Experience:
   * Provide a faster and more responsive loan approval process, improving the overall experience for customers.
   * Implement fair and transparent lending practices to build trust and loyalty among customers.

By achieving these aims, NBFI institutions can create a more efficient, customer-centric, and risk-aware lending environment, ultimately contributing to the growth and sustainability of their vehicle loan portfolios.

# OVERALL DESCRIPTION

## WORKFLOW OF PROJECT



## 2.2 DATA DESCRIPTION

The Data Set contains:

A non-banking financial institution (NBFI) or non-bank financial company (NBFC) is a type of financial institution that is not authorized to operate as a bank or is not under the supervision of a banking regulatory agency at the national or international level. NBFCs provide financial services similar to those of banks, such as investment, risk pooling, contractual savings, and market brokering.

An NBFC is currently facing profitability issues due to an increase in defaults in the vehicle loan category. The company aims to assess the loan repayment ability of clients and determine the relative importance of each factor contributing to a borrower's ability to repay the loan.

The objective is to predict whether a client is likely to default on their vehicle loan payment or not. The task involves building a model using the Train\_Dataset and testing the model on the Test\_Dataset.

* *train.csv* contains the training data with details on loan as described in the last section
* *data\_dictionary.csv* contains a brief description on each variable provided in the training and test set.
* *test.csv* contains details of all customers and loans for which the participants are to submit probability of default.
* sample\_submission.csv contains the submission format for the predictions against the test set. A single csv needs to be submitted as a solution.

## 2.3 Importing Dataset.

Here we can see that we have categorical and continuous variables, we can also see that we have a lot of null values.

Here the explanation about the variables:

* ID: Client Loan application ID
* Client\_Income: Client Income
* Car\_Owned: Any Car owned by client before applying for the loan for another car (0 means No and 1 means otherwise)
* Bike\_Owned: Any bike owned by client (0 means No and 1 means otherwise)
* Active\_Loan: Any other active loan at the time of application of loan (0 means No and 1 means otherwise)
* House\_Own: Any house owned by client (0 means No and 1 means otherwise)
* Child\_Count: Number of children the client has
* Credit\_Amount: Credit amount of the loan
* Loan\_Annuity: Loan annuity
* Accompany\_Client Who accompanied the client when client applied for the loan
* Client\_Income\_Type: Clients income type
* Client\_Education: Highest level of education achieved by client
* Client\_Marital\_Status: Marital status of client (D- Divorced, S- Single, M- Married, W- Widowed)
* Client\_Gender: Gender of the Client
* Loan\_Contract\_Type: Loan Type (CL- Cash Loan, RL- Revolving Loan)
* Client\_Housing\_Type: Client Housing situation
* Population\_Region\_Relative: Relative population of the region where the client is living. Higher value means the client is living.
* Age\_Days: Age of the client at the time of application submission
* Employed\_Days: Days before the application, the client started earning
* Registration\_Days: Days before the loan application, the client changed his/her registration
* ID\_Days: Days before the loan application, the client changed his/her identity document with which the loan w...
* Own\_House\_Age: Age of Client's house in years
* Mobile\_Tag: Mobile Number provided by Client (1 means Yes and 0 means No)
* Homephone\_Tag: Homephone Number provided by Client (1 means Yes and 0 means No)
* Workphone\_Working: Was workphone number reachable (1 means Yes and 0 means No)
* Client\_Occupation: Client Occupation type
* Client\_Family\_Members: Number of family members does client have
* Cleint\_City\_Rating: Client city rating. 3 denotes best and 2 denotes good and 1 denotes average
* Application\_Process\_Day: Day of the week on which client applied for the loan (0-Sun, 1-Mon,2-Tues, 3-Wed, 4-Thrus,5-Fri, 6-S...
* Application\_Process\_Hour: hour of the day on which client applied for the loan
* Client\_Permanent\_Match\_Tag: Indication if client contact address does not match permanent address.
* Client\_Contact\_Work\_Tag: Indication if client work address does not match contact address.
* Type\_Organization: Type of organization where client works
* Score\_Source\_1: Score sourced from other source. This is a normalzied socre
* Score\_Source\_2: Score sourced from other source. This is a normalzied socre
* Score\_Source\_3: Score sourced from other source. This is a normalzied socre
* Social\_Circle\_Default: How many friends/family member of client defaulted on any loan payment in last 60 days
* Phone\_Change: How many days before the loan application, client changed his/her phone
* Credit\_Bureau: Total number of enquiries in last year
* Default: 1 means the client defaulted on loan payments and 0 means otherwise

## 2.4 Project Initiation:

Define the project scope, objectives, and deliverables.

Identify stakeholders and establish communication channels.

Set up project management tools and resources.

## 2.5 Data Acquisition:

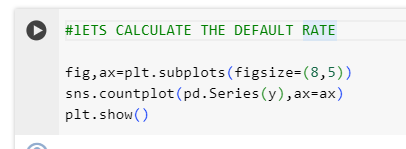
* Obtain the NBFI Vehicle Loan repayment dataset from the appropriate source, ensuring compliance with data privacy and security regulations.
* Understand the structure and format of the dataset, including the features, target variable, and any data dictionaries or documentation provided.

## 2.6 Data Exploration and Understanding:

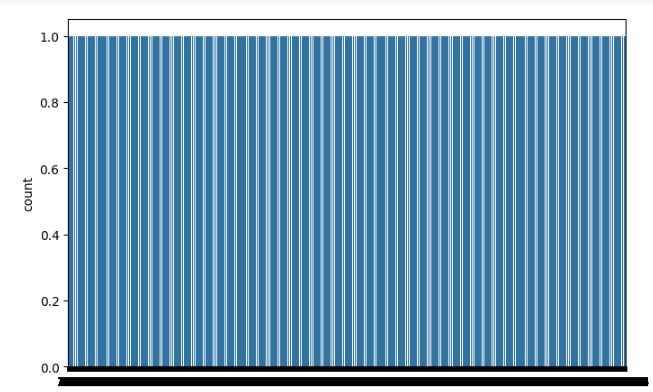
* Perform initial data exploration to gain insights into the dataset's characteristics, including the distribution of variables, missing values, outliers, and potential data quality issues.
* Visualize key features using descriptive statistics, histograms, box plots, and correlation matrices.

## 2.7 Data Pre-processing:

* Handle missing data by imputation, removal, or interpolation techniques.
* Address outliers through trimming, winsorization, or transformation methods.
* Encode categorical variables using techniques such as one-hot encoding or label encoding.
* Scale numerical features to a comparable range using standardization or normalization methods.

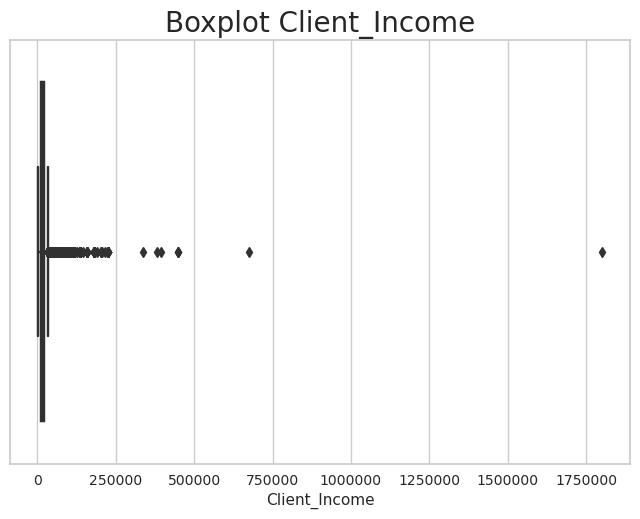


*Figure 2 Calculating Default Rate*

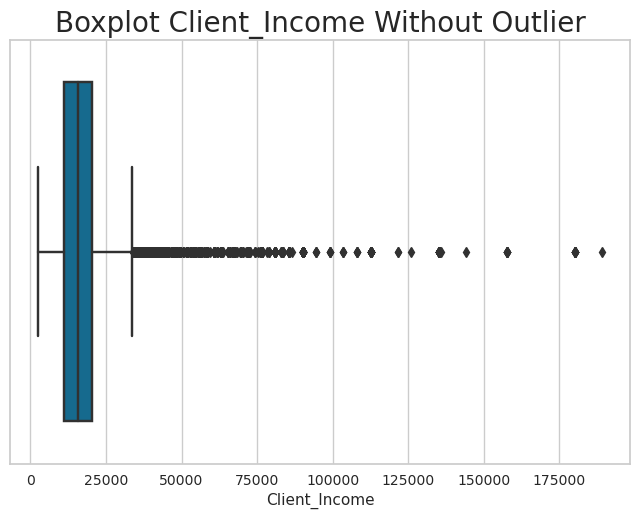


*Figure 3 Visualzation of Default Rate*

**Outlier Handling:**

****

*Figure 4 Outlier Handling*

****

*Figure 5 Client\_income after Outliers Handling*

**Correlation Analysis:**

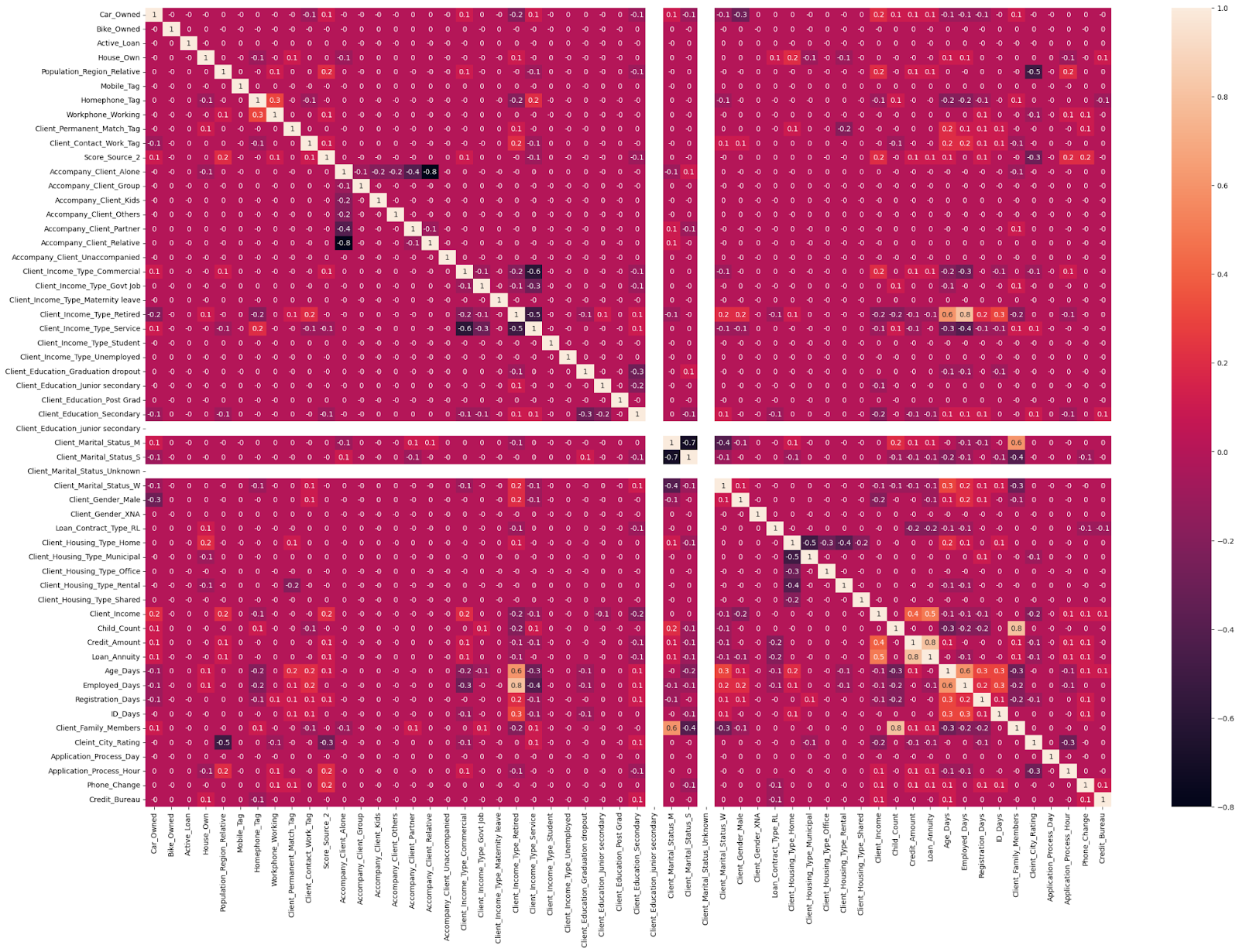
* Compute and visualize the correlation matrix to identify relationships between numerical features.
* High correlation between features might indicate multi-collinearity.



*Figure 6 Correlation Analysis 1*



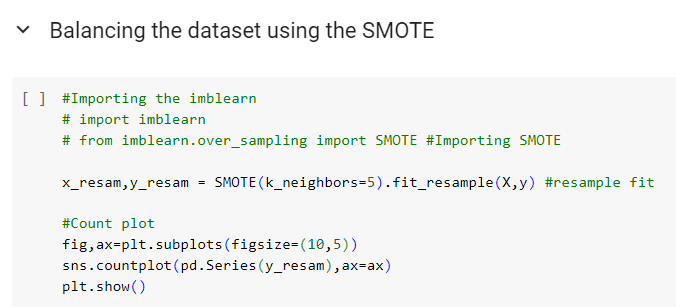
*Figure 7 Correlation Analysis 2*



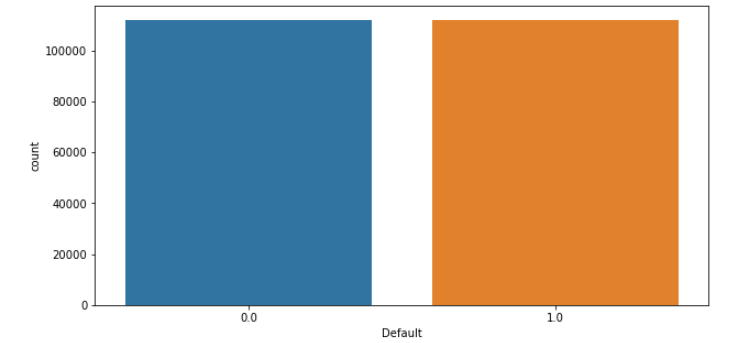
*Figure 8 Correlation Matrix*

## 2.8 Feature Engineering:

* Generate new features that may capture additional information relevant to loan repayment behaviour, such as debt-to-income ratio, loan-to-value ratio, or payment history indicators.
* Select or create relevant features based on domain knowledge and exploratory data analysis insights.



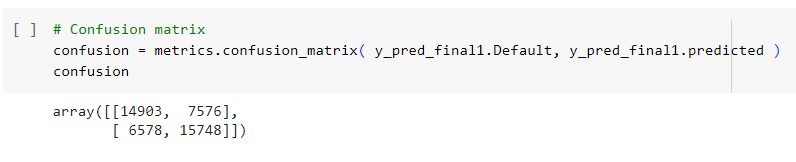
*Figure 9 Code For Balancing Dataset using SMOTE*



*Figure 10 Count of Balanced Classes*

## 2.9 Model Development:

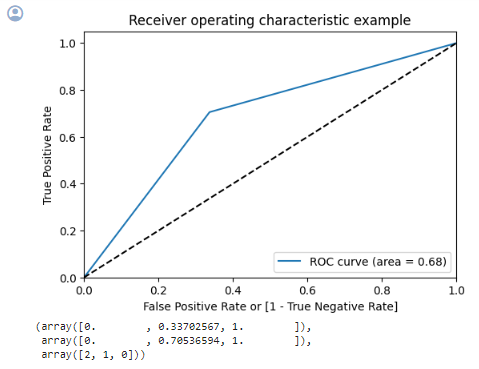
* Split the dataset into training, validation, and test sets to evaluate model performance effectively.
* Select appropriate machine learning algorithms for the predictive task, such as logistic regression, decision trees, random forests, gradient boosting, or neural networks.
* Train initial models using default hyperparameters and evaluate their performance using appropriate evaluation metrics.
* Perform hyperparameter tuning using techniques such as grid search, random search, or Bayesian optimization to optimize model performance.
* Experiment with ensemble methods or model stacking to further enhance predictive accuracy.



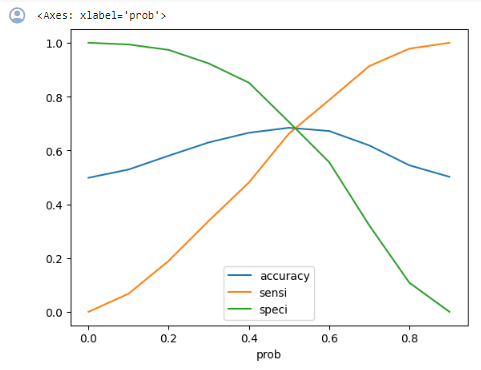
*Figure 11 Confusion Matrix*

## 2.10 Model Evaluation:

* Assess the performance of trained models using a comprehensive set of evaluation metrics, including accuracy, precision, recall, F1-score, ROC AUC, and calibration plots.
* Conduct sensitivity analysis to understand the robustness of the models across different scenarios and thresholds.
* Compare the performance of different models and select the best-performing one based on predefined criteria.



*Figure 12 ROC Curve*



*Figure 13 Accuracy vs Sensitivity vs Specificity*

## 2.11 Model Interpretation:

* Interpret model predictions and feature importance to understand the factors driving vehicle loan repayment behaviour.
* Communicate findings to stakeholders in a clear and interpretable manner, highlighting actionable insights and recommendations.

## 2.12 Model Deployment:

* Deploy the selected model into production environments, integrating it into the NBFI's loan approval process to assist in decision-making.
* Implement monitoring mechanisms to track model performance and recalibrate the model as needed over time.

## 2.13 Documentation and Reporting:

* Document the entire project workflow, including data pre-processing steps, model development process, evaluation results, and deployment procedures.
* Prepare a comprehensive report summarizing the project findings, insights.

# REQUIREMENTS SPECIFICATION

## 3.1 Hardware Requirement:

* 500 GB hard drive (Minimum requirement)
* 8 GB RAM (Minimum requirement)
* PC x64-bit CPU

## 3.2 Software Requirement:

* Windows/Mac/Linux
* Python-3.9.1
* VS Code/Anaconda/Spyder
* Python Extension for VS Code

**Libraries:**

* Numpy 1.18.2
* Pandas 1.2.1
* Matplotlib 3.3.3
* Scikit-learn 0.24.1

# CONCLUSION

#### We can see that we have numerical and continuous data, in our database we have a good amount of data both in rows and in columns, which theoretically was to facilitate our work, but when we look deeper into the data we can see that we have many null values and inconsistent columns that need to be treated, many values needed to be treated, we had to exclude some columns due to the number of null values, and in some we transformed the null values either by the median or by the mode, I used the criterion that made the most sense for me.

#### When we look at the correlation we can see that we only have a strong correlation between our variables, looking at the exploratory analysis we can see some patterns in the data that were commented on in the specific part, the big problem is when we look at our Target variable, we have a lot of data of one result and very little of the other, which makes it necessary to balance our data, when we look at the bivariate analysis with our target variable, we cannot find a very different pattern between the data.

#### Going deeper into the part of creating the Machine Learning models, in preparation I categorization of the Score Source 2 variable, which was the variable that most caught my attention in the exploratory analysis, I used the onehot encoder to transform the categorical variables into columns, after separating the data into training and testing, I applied SMOTE to our Target variable to get a balanced target variable.

#### Talking about the results of the models, when we look at the accuracy the best models we had was the Logistic Regression with 68.4%, however it does not meet our objective because it learned only to predict the negative result and considerably fails to predict the positive result.

# REFERENCES

* <https://www.kaggle.com/datasets/meastanmay/nbfi-vehicle-loan-repayment-dataset>
* <https://gemini.google.com>
* <https://www.w3schools.com>
* <https://www.kaggle.com/code/raphaelmarconato/vehicle-loan-eda-ml-and-balancing-65-90/notebook>
* <https://github.com/shriya1705/Big-Data-Analytics/blob/master/Data%20Analysis%20in%20HIVE.sql>