

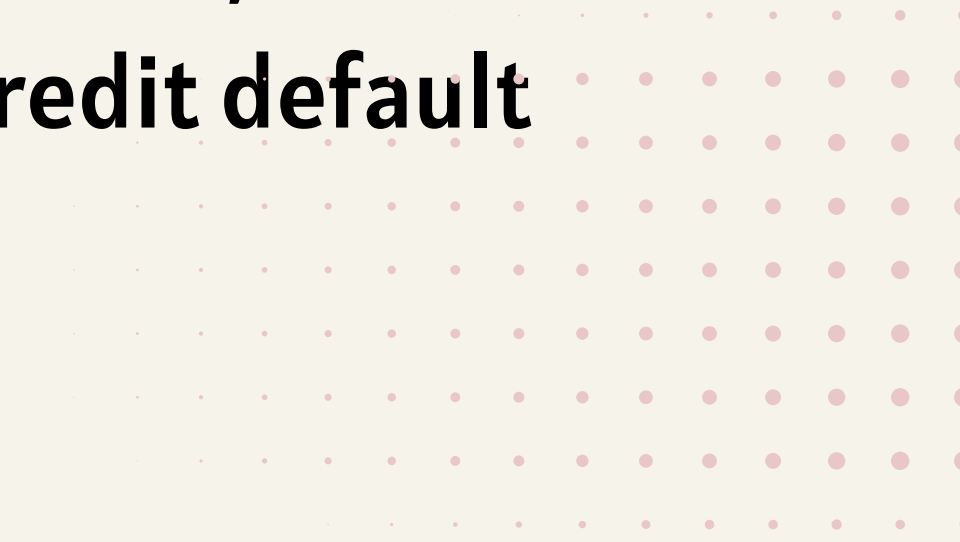
RISK ANALYSIS: HOUSE CREDIT DEFAULT

Mentor: Mr. Narendra Kumar



PROBLEM STATEMENT

The risk of default on home credit loans is a significant concern for financial institutions. This project focuses on performing exploratory data analysis (EDA) and building predictive models to assess the default risk for home credit applicants. By analyzing various features related to applicants' demographics, financial history, and loan attributes, the project aims to gain insights into the factors influencing credit default and develop models to predict the likelihood of default.



PLANNING

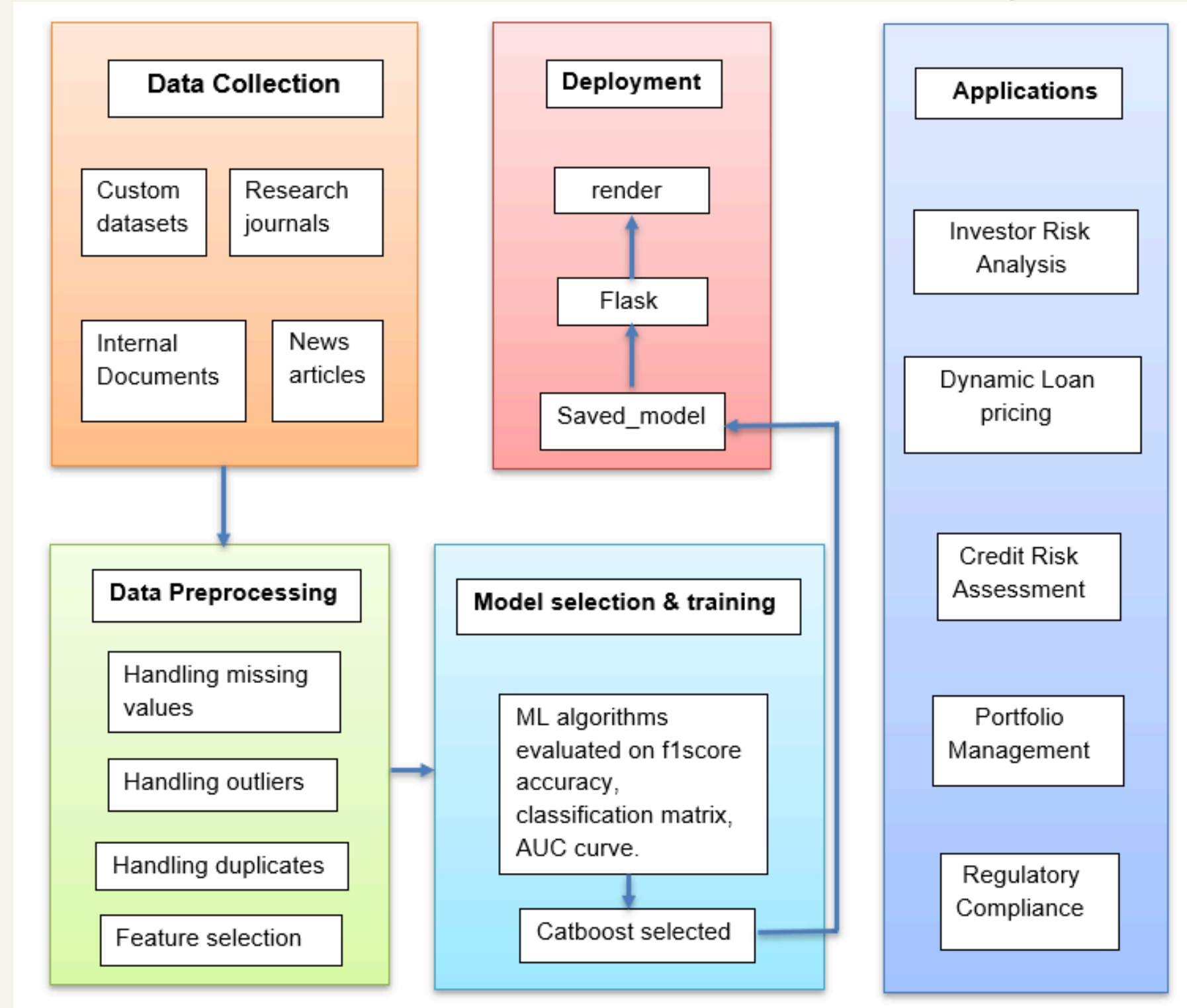


Fig 1- System workflow

Literature Survey

Research Highlights: Loan Default Prediction & Credit Risk Assessment Traditional Techniques:

Logistic Regression	Simple, interpretable, widely used for default prediction.
Decision Trees (e.g., C4.5)	Classify borrowers effectively with fewer attributes.
Support Vector Machines (SVMs)	High accuracy, often paired with Naïve Bayes.

Advanced Techniques

Random Forest	High accuracy, aggregates decision trees, outperforms individual trees.
XGBoost	Gradient boosting with state-of-the-art results; effective in feature selection.
Artificial Neural Networks (ANNs)	High accuracy in handling complex datasets.
Genetic Programming (GP)	Combines deep learning with evolutionary principles; surpasses traditional models.

Emerging Insights

Hybrid Models	Combining techniques improves predictive power.
Data Preprocessing	Cleaning and feature selection enhance model accuracy (e.g., SMOTE).
Challenges	Handling imbalanced datasets and optimizing algorithm choice remain critical.

Data Analysis

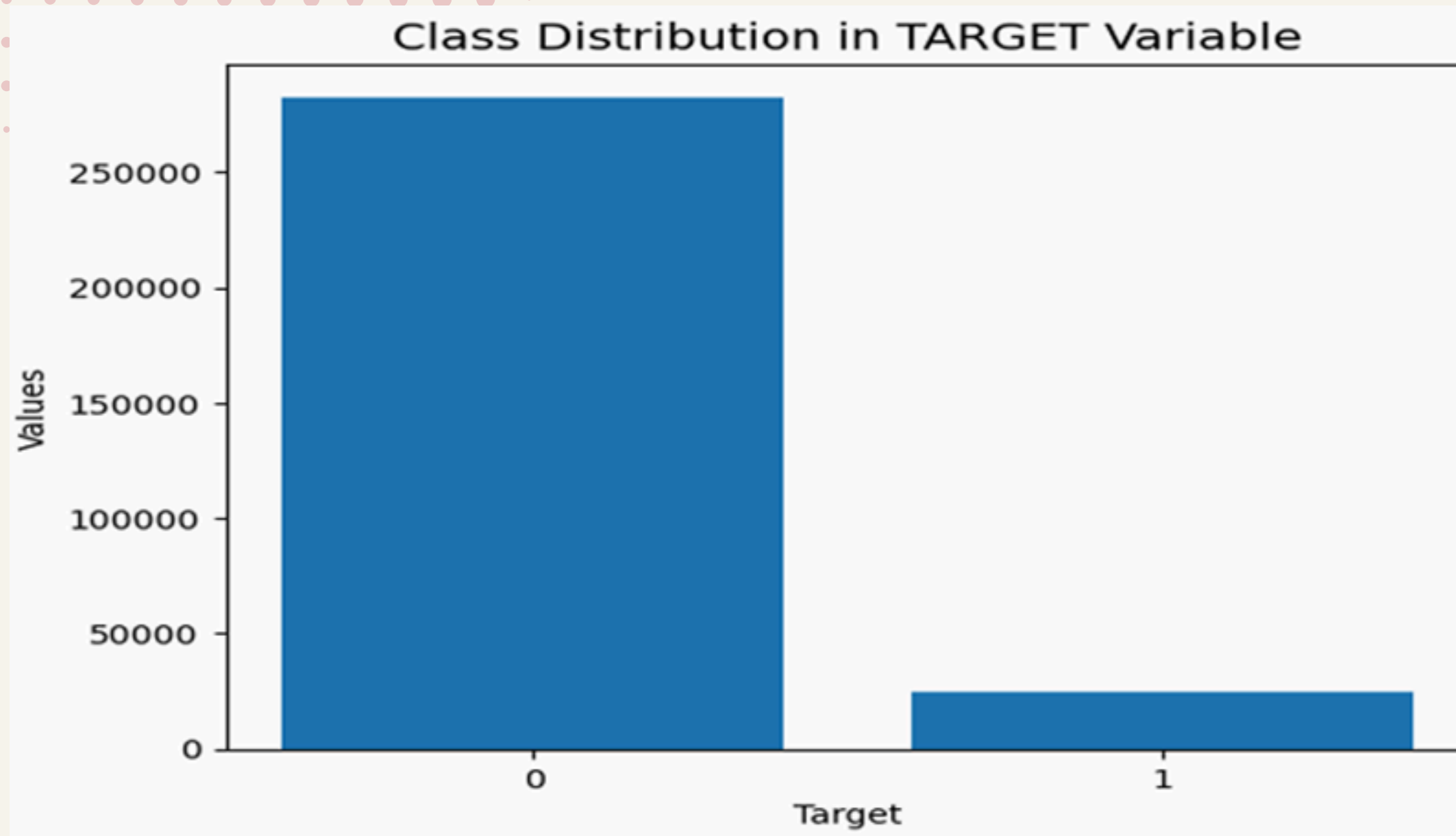


Fig 2- Imbalance of Data for TARGET variable 0 and 1 in Application Data

The distribution of the target variable reveals a significant imbalance, with only 8.07% of loans being defaulted (minority class) and 91.9% of loans being non-defaulted (majority class).

The Defaulters have been assigned a Target variable of 1 and Non-Defaulters have been assigned Target Variable 0.

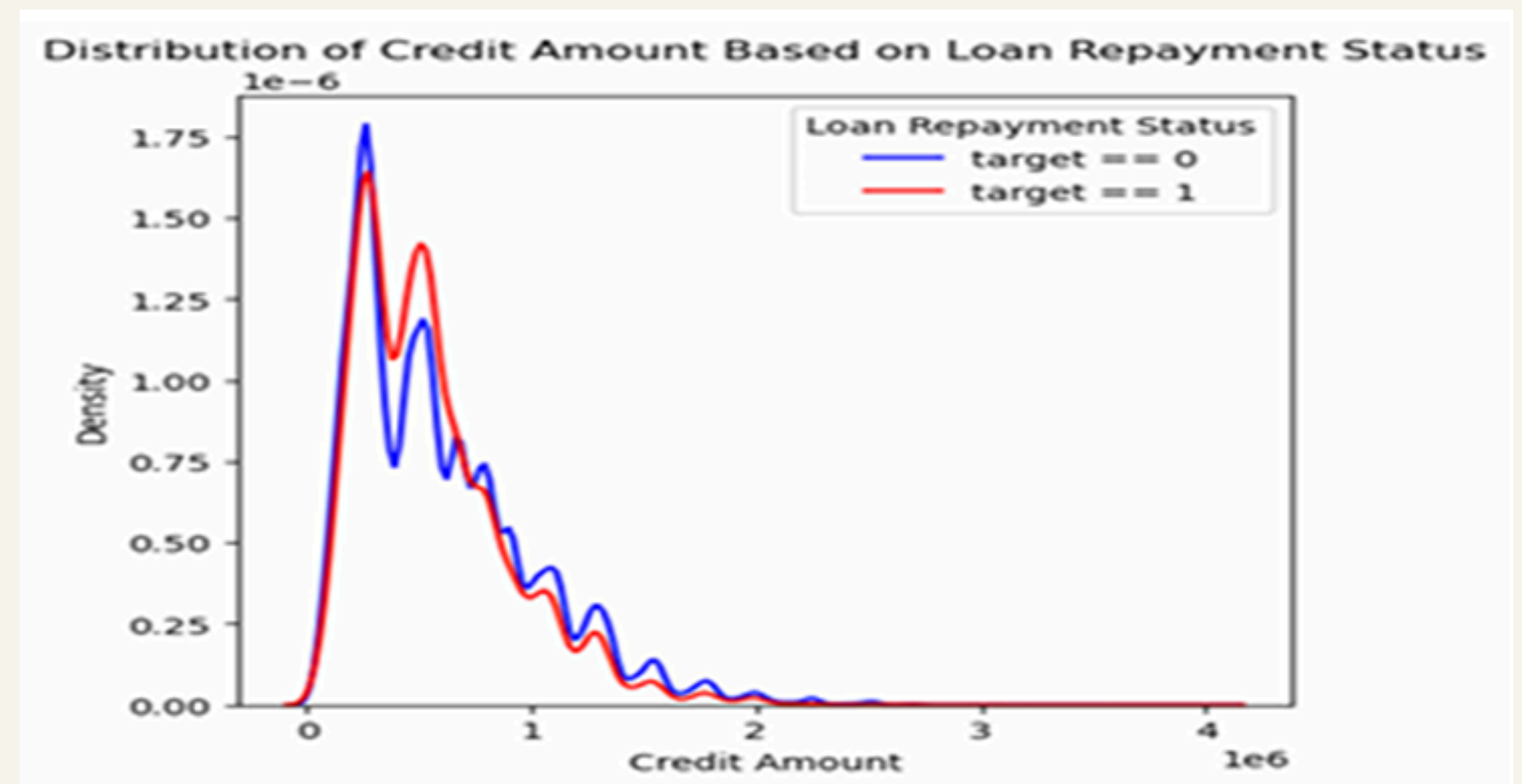
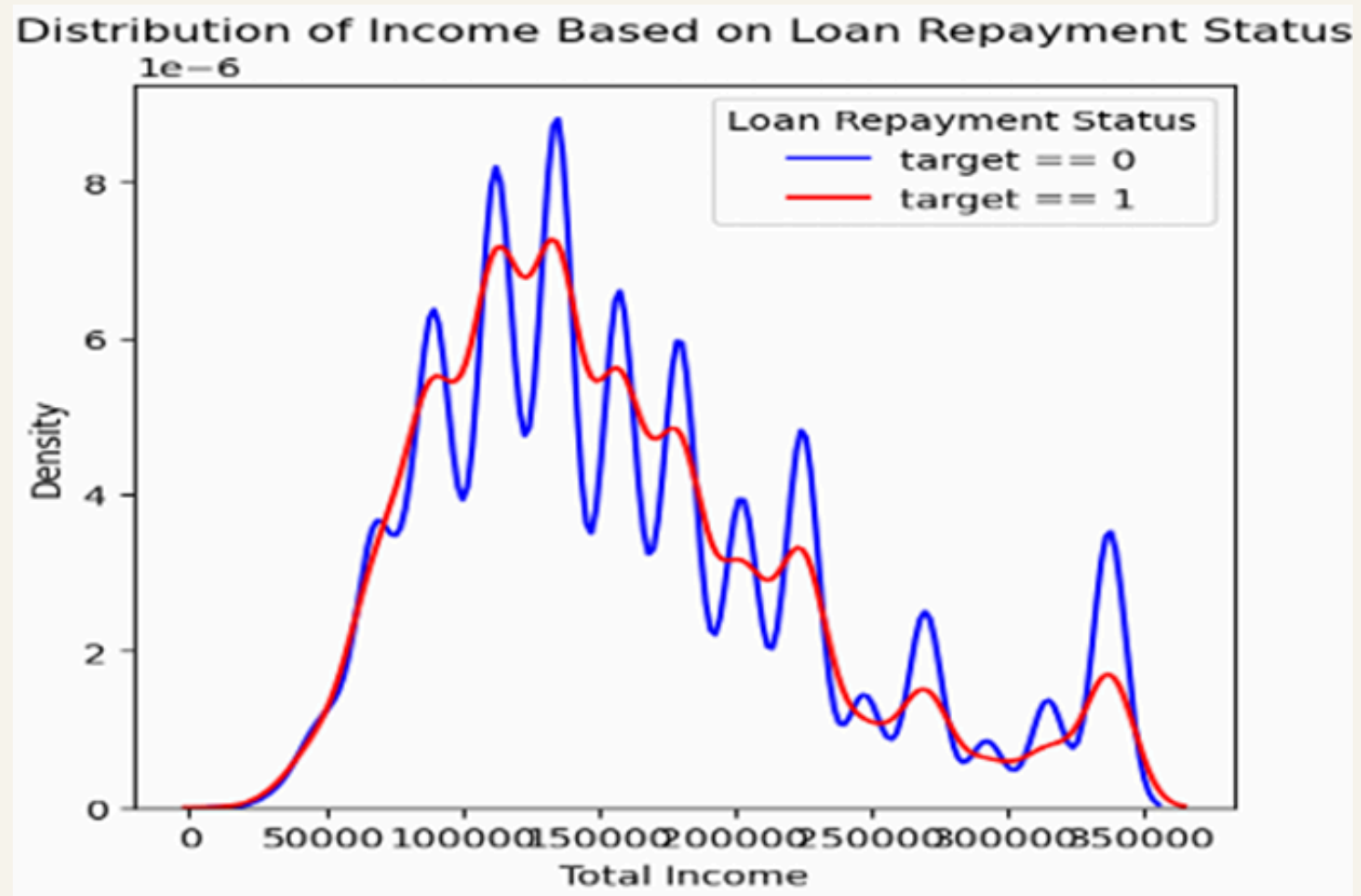
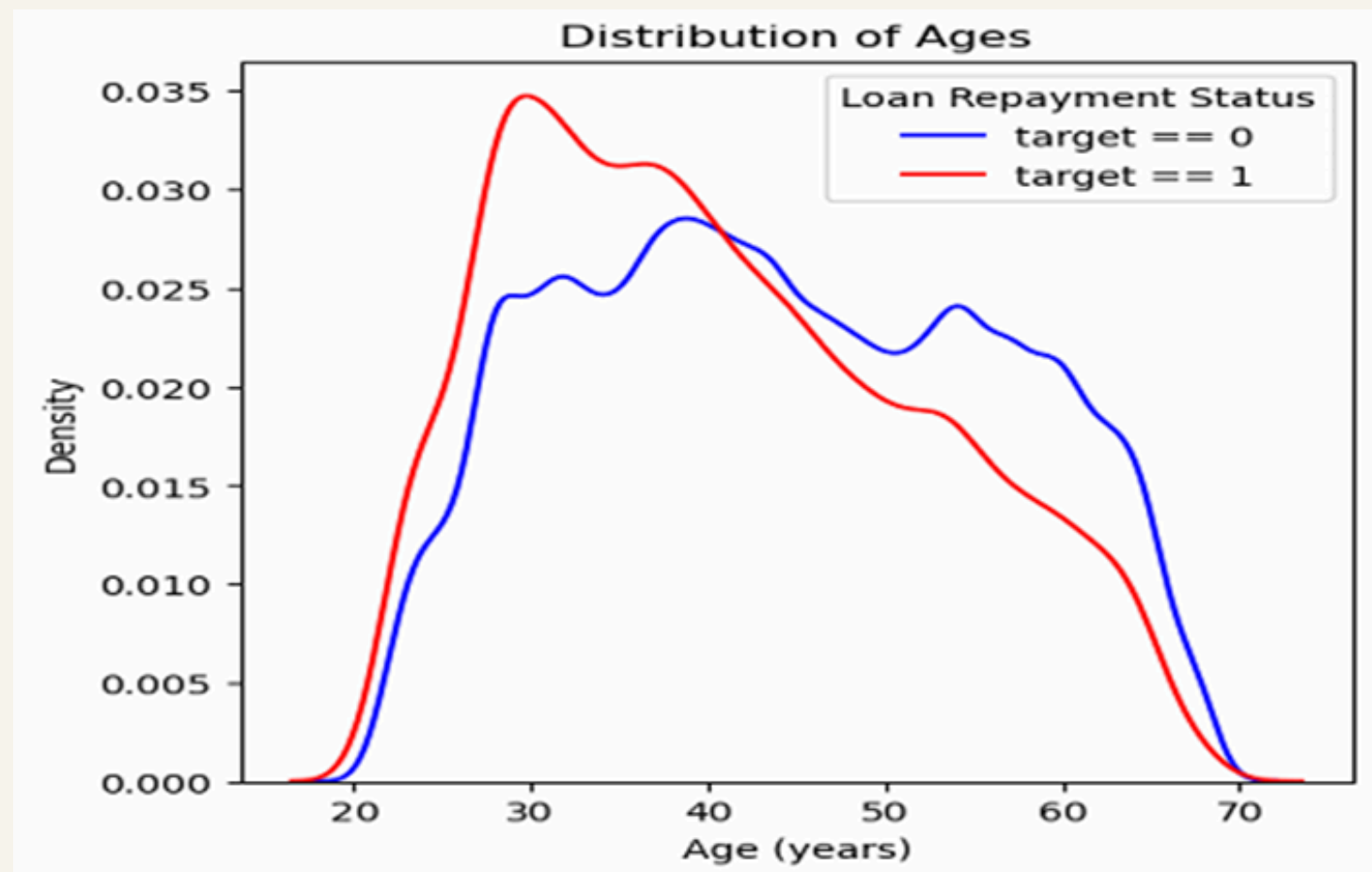


Fig 3- Univariate Analysis

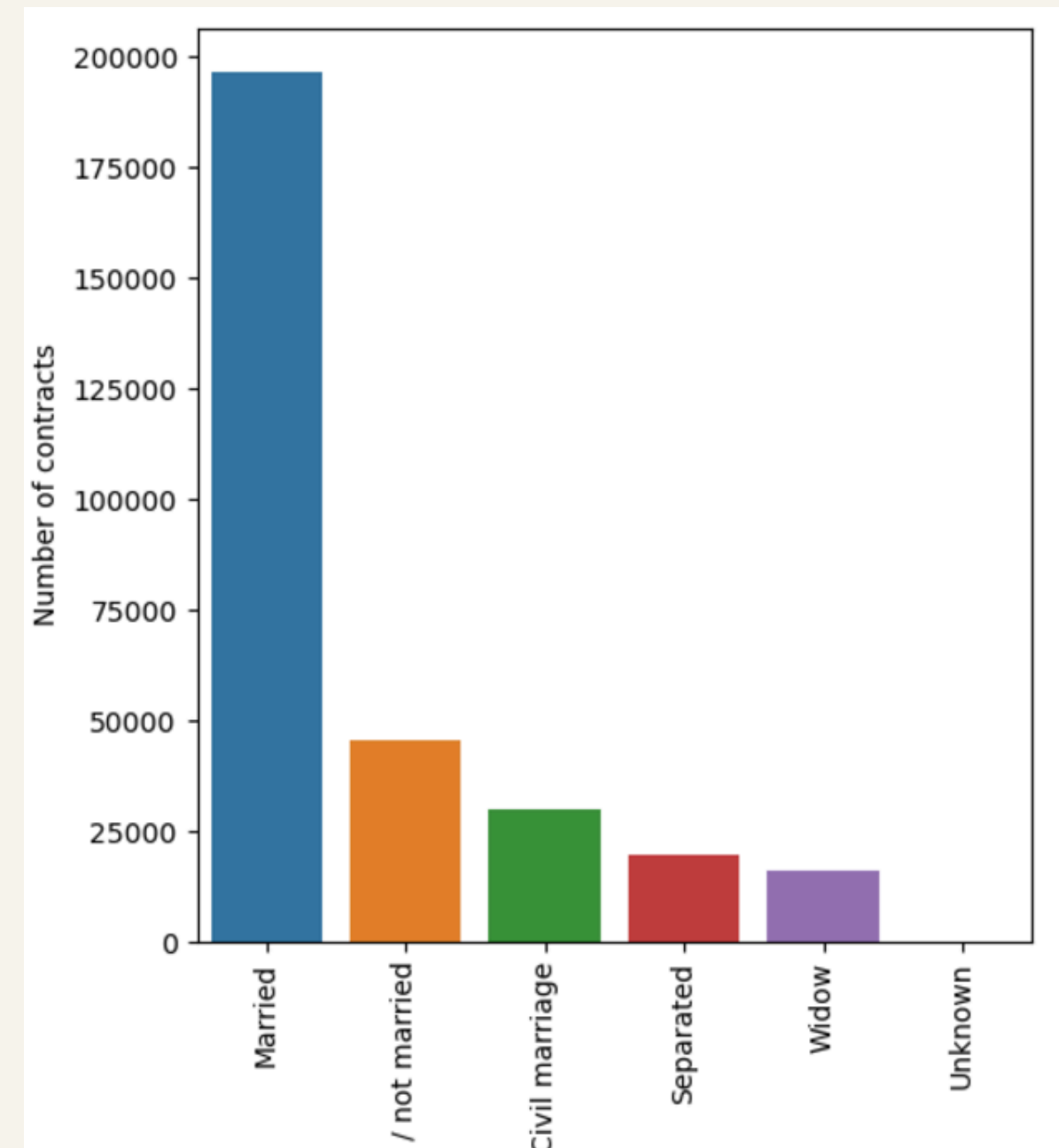
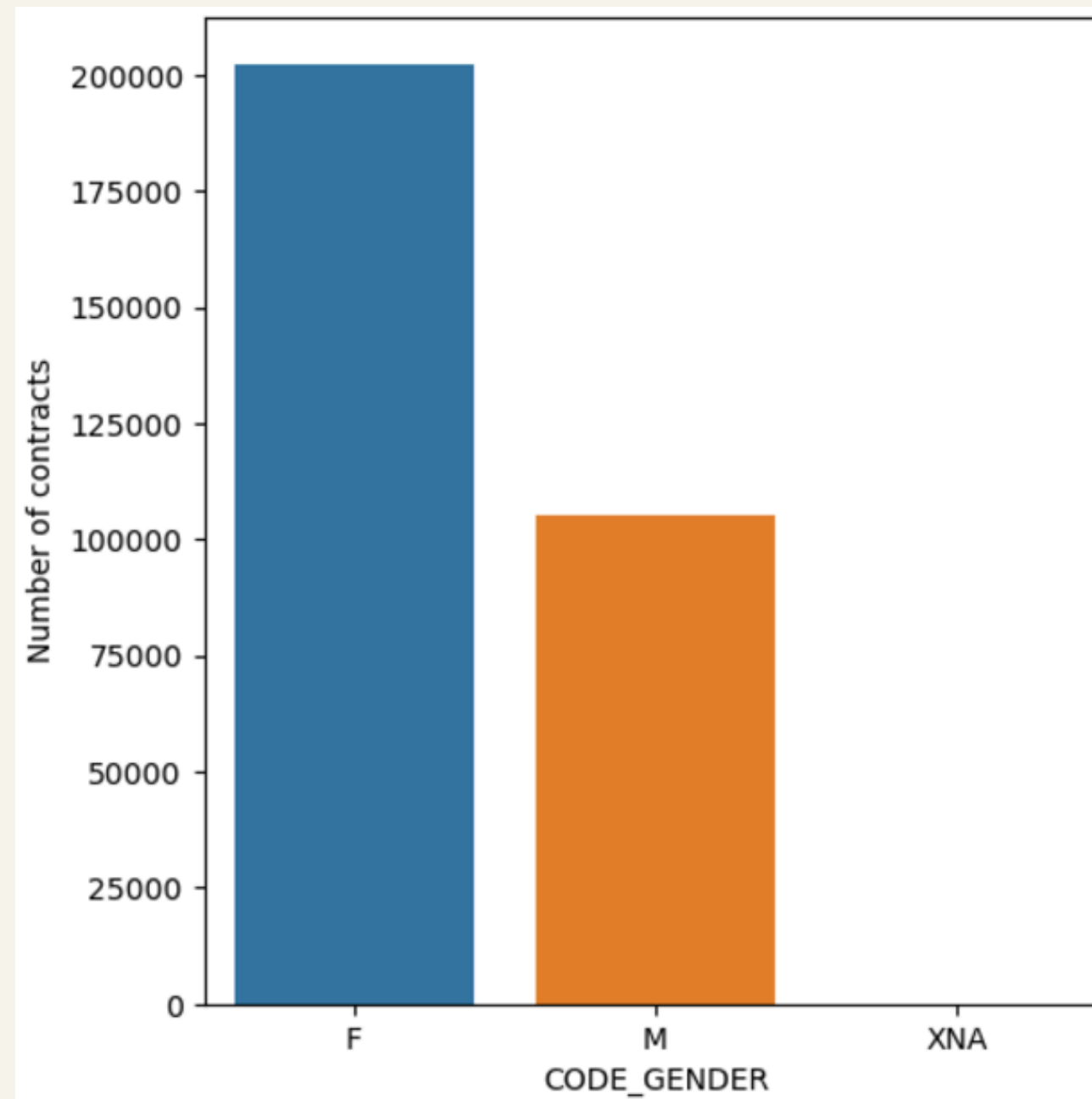
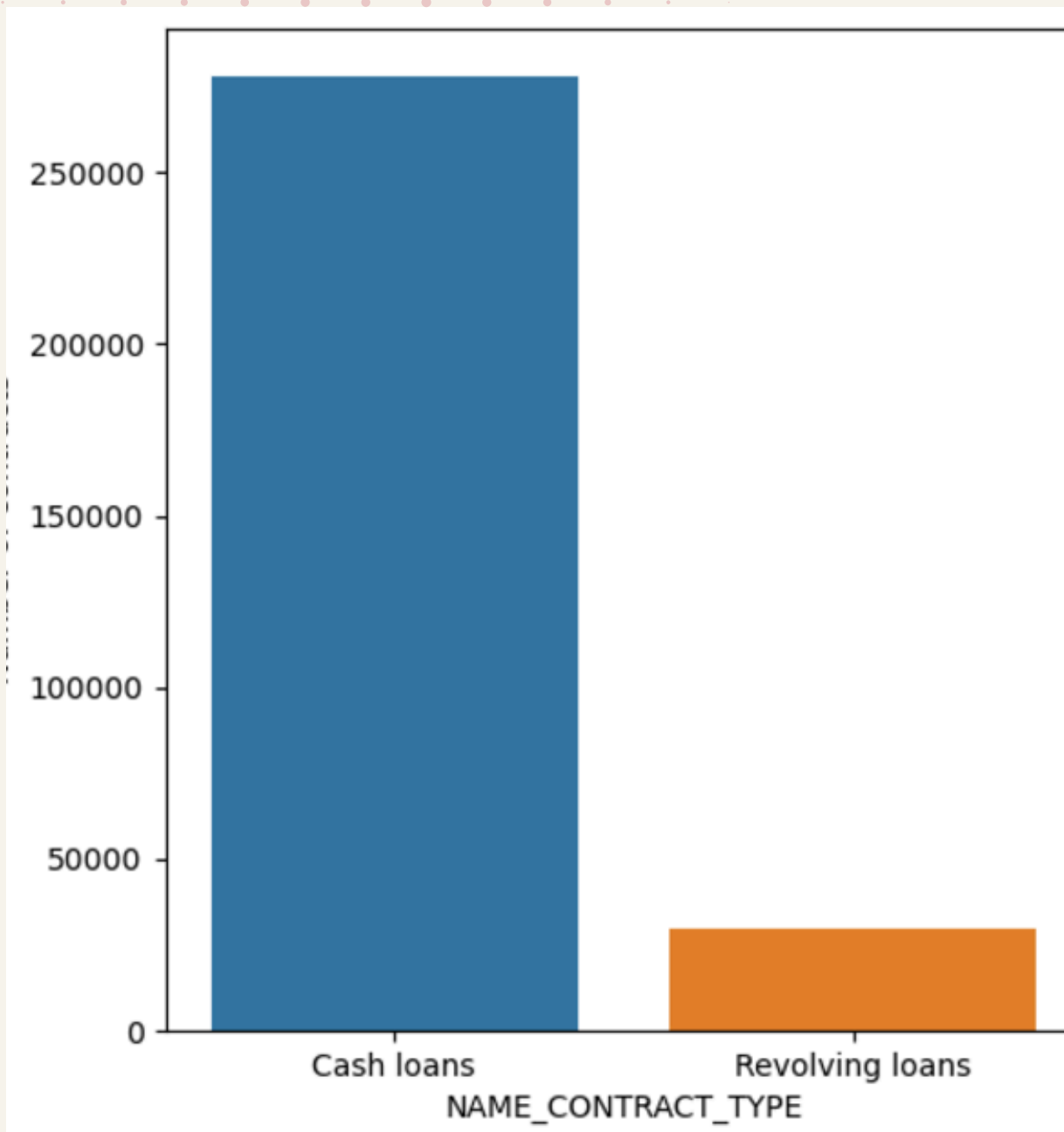


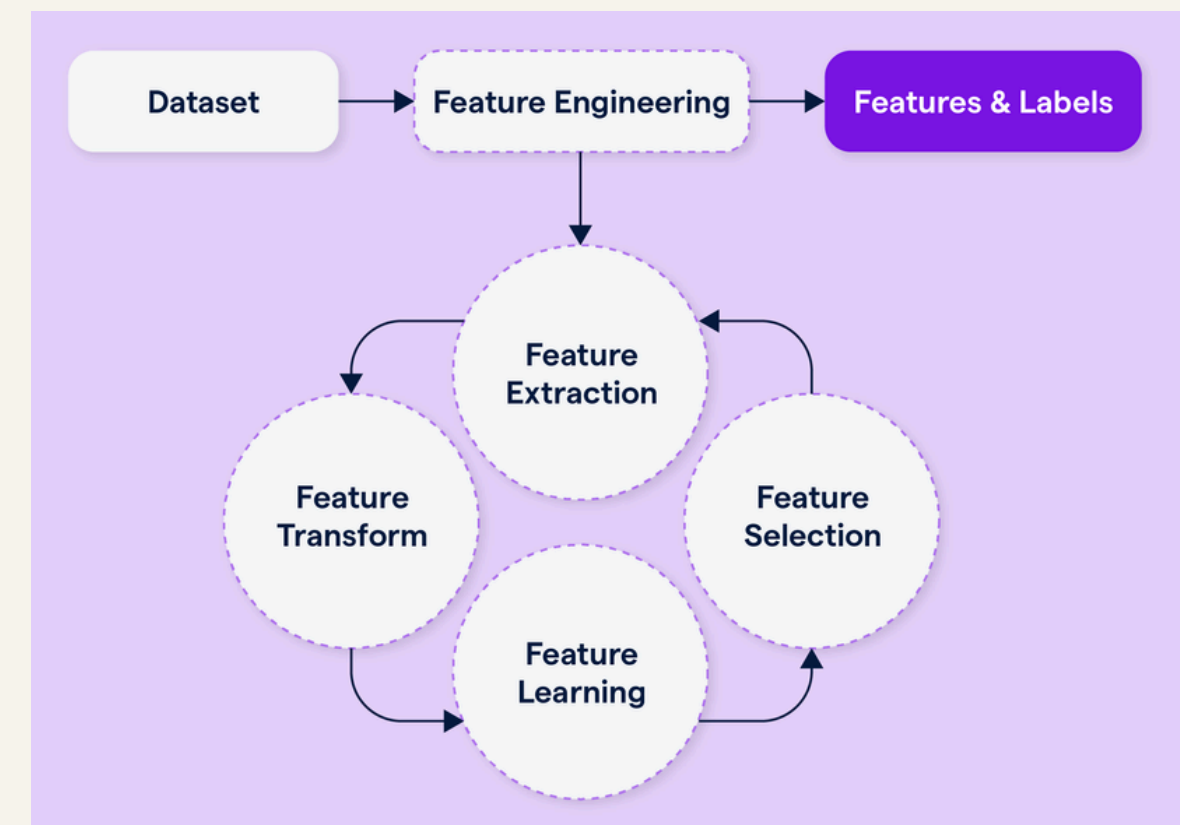
Fig 4- Bivariate Analysis

Feature engineering

Feature engineering involves transforming raw data into features that improve model performance. It ensures that the data is represented in a way that the algorithm can learn from effectively.

In this we have taken some of the inputs as the parameters they are:

- **AMT_BALANCE:** Represents the remaining balance on the applicant's previous credit.
- **AMT_ANNUITY:** The annual payment amount the applicant needs to make for the loan.
- **SK_DPD:** The number of days the applicant is past due on a previous credit.
- **CNT_CHILDREN:** The number of children the applicant has.
- **FLAG_OWN_CAR:** Indicates if the applicant owns a car.
- **CODE_GENDER:** The gender of the applicant.
- **NAME_FAMILY_STATUS:** The marital status of the applicant.
- **NAME_INCOME_TYPE:** The type of income the applicant earns.
- **NAME_HOUSING_TYPE:** The type of housing the applicant resides in, listed as 'House / apartment'.
- **NAME_CONTRACT_TYPE:** The type of loan contract, which is 'Cash loans' or revolving loans.
- **DAYS_CREDIT:** The number of days since the applicant's previous credit was opened, which is in negative (negative indicating the past).
- **DAYS_DECISION:** The number of days since the decision was made on a previous application.
- **AMT_PAYMENT:** The payment amount made by the applicant.
- **AMT_INSTALLMENT:** The installment amount the applicant is required to pay.
- **AMT_APPLICATION:** The amount requested by the applicant in the loan application.



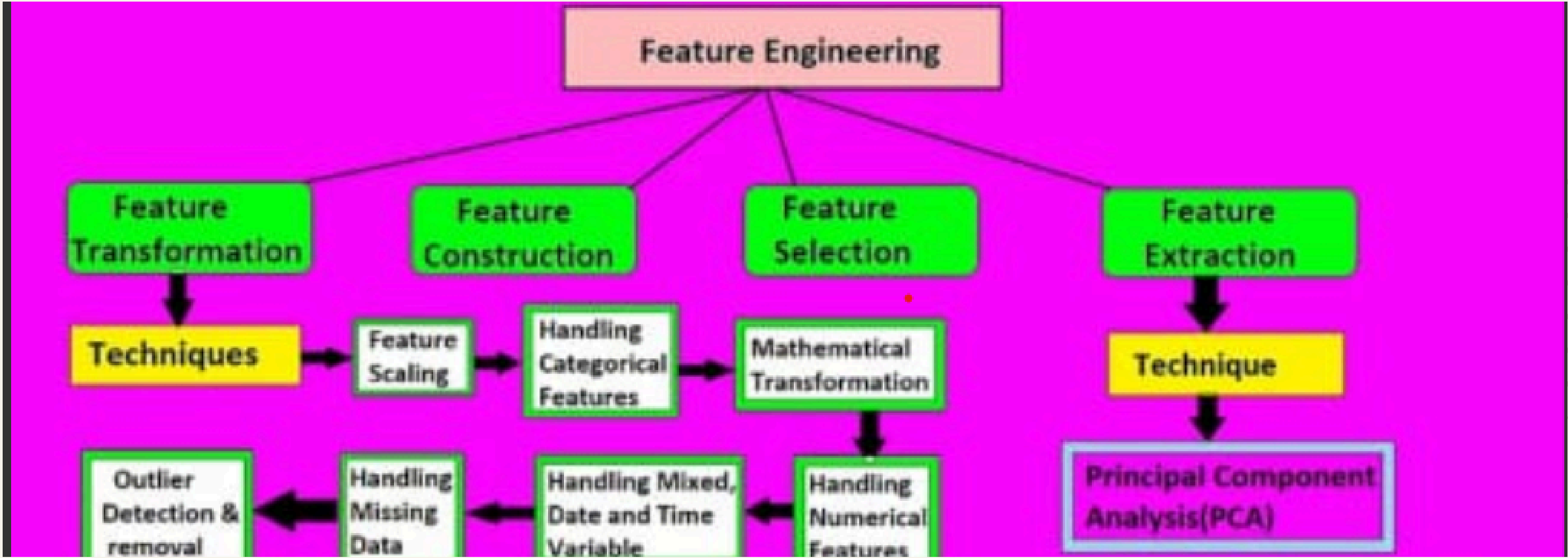
The output parameter will be the target variable which specifies whether a person is defaulter or not

Defining Numerical and Categorical Features: Features are classified into numerical (numerical_ features) and categorical (categorical_ features) groups to apply appropriate preprocessing techniques.

Numerical Feature Preprocessing: A pipeline is defined to handle numerical features: Missing values are imputed using the mean. The data is standardized using Standard Scaler, which scales features to have a mean of 0 and a standard deviation of 1.

Categorical Feature Encoding: Categorical features are one-hot encoded using One Hot Encoder. This converts categorical variables into a binary matrix suitable for machine learning algorithms.

Combining Preprocessing Steps: The Column Transformer combines separate preprocessing pipelines for numerical and categorical features, ensuring a unified transformation process.



MODEL TRAINING AND EVALUATION

ML Algorithm	F1 Score	Accuracy	AUC
Neural Network	0.1156	90.36	0.65
CatBoost	0.0496	92.75	0.75
Logistic Regression	0.0168	92.68	0.74
DNN	0	92.69	0.74
LightGBM	0.1085	92.75	0.72
Random Forest	0.0333	92.69	0.71
XGBoost	0.0896	92.31	0.71
AdaBoost	0.1037	92.38	0.71
Naive Bayes	0.1372	8.88	0.65
KNN	0.0451	92	0.63
SVM	0	92.69	0.61
Decision Tree	0.1707	87.15	0.55

Fig 5- Different Models Accuracy, F1 Score and AUC value

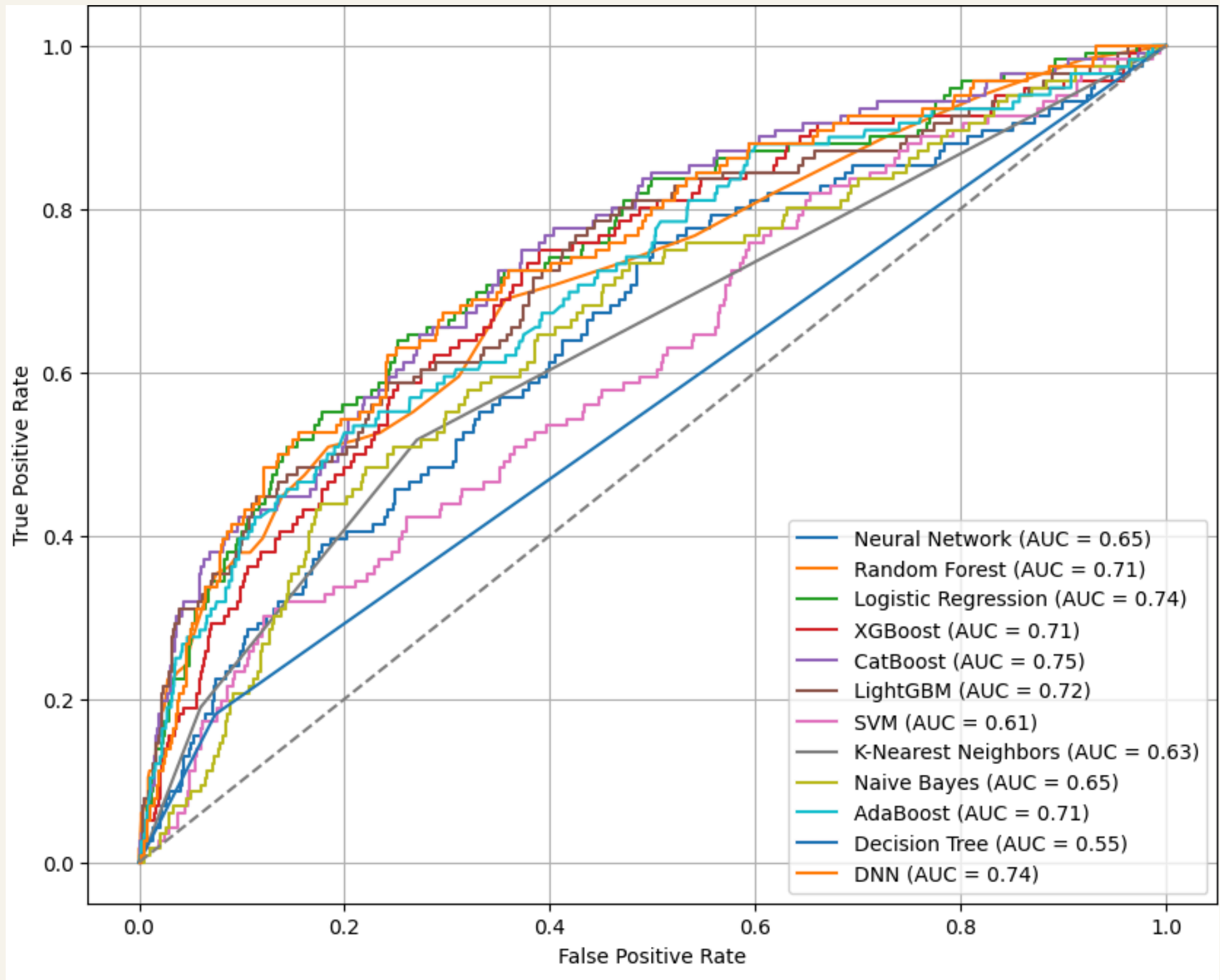


Fig 6- ROC Curve

MODELLING OF CHOOSEN ALGORITHM

CatBoostClassifier: A gradient boosting algorithm specifically designed to handle categorical features and missing values efficiently. Parameters used: **iterations**, **depth**, **learning_rate**, **loss_function** and **cat_features**

- **catboost_model** is saved to **catboost_model_pipeline.pkl**.

Implemented a robust pipeline for credit score calculations and default predictions.

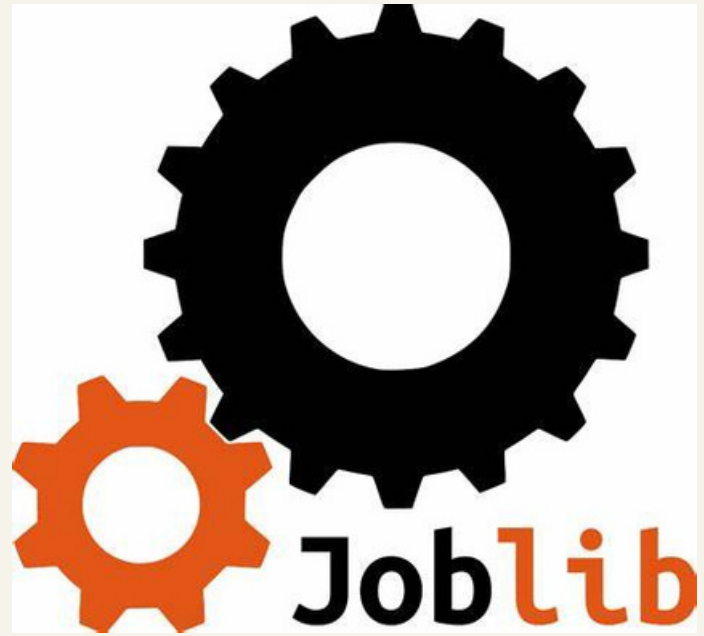
The goal was to calculate a **credit score** (scaled up to a maximum of 850) using multiple factors such as **payment history, credit utilization, credit history length, credit mix, and new credit inquiries**. Each factor contributes a specific weight to the overall score.

A FICO score ranges from **300 to 850** and is used by lenders to assess borrowers' creditworthiness. A FICO score is a type of credit score based on information in a borrower's credit report that lenders use to assess credit risk and determine whether to extend credit. The higher the FICO score, the more likely a borrower will repay their debts on time. Created by the **Fair Isaac Corporation**

FICO Score	Rating	What the Score Means
< 580	Poor	<ul style="list-style-type: none">• Well below average• Demonstrates to lenders that you're a risky borrower
580 – 669	Fair	<ul style="list-style-type: none">• Below average• Many lenders will approve loans
670 – 739	Good	<ul style="list-style-type: none">• Near or slightly above average• Most lenders consider this a good score
740 – 799	Very Good	<ul style="list-style-type: none">• Above average• Demonstrates to lenders you're a very dependable borrower
800+	Exceptional	<ul style="list-style-type: none">• Well above average• Demonstrates to lenders you're an exceptional borrower

Fig 7- FICO Range Table

DEPLOYMENT



Saved Model file
using Joblib



Flask



Render

the URL for our project is:- <https://house-credit-default-infosys-6.onrender.com/>

Payment Amount

Instalment Amount

Application Amount

Predict

Prediction Results

Credit Score: 312

FICO Range: Poor

Prediction: Defaulter

Payment Amount

Instalment Amount

Application Amount

Predict

Prediction Results

Credit Score: 793

FICO Range: Very Good

Prediction: Non-Defaulter

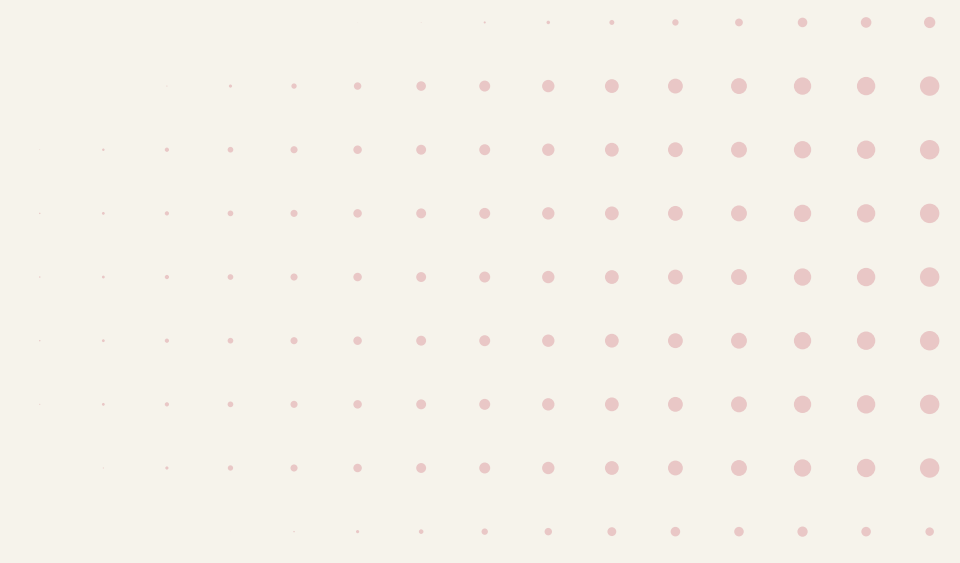
Fig 8- Our Model is able to Predict both the defaulters and non defaulter accurately. people who have poor or fair FICO range are found to be defaulters and people with good , very good and exceptional are found to be non defaulters.



RESULTS AND CONCLUSION

The Risk Analysis for Home Credit Default project establishes a solid foundation for building a predictive model to assess loan default likelihood. Through thorough exploratory data analysis (EDA), we identified key patterns and improved feature selection strategies by handling missing data, addressing outliers, and performing feature engineering. Visualizations uncovered important trends and relationships, aiding informed decision-making. In the next phase, machine learning algorithms will leverage these insights to predict default probabilities more accurately. The ultimate goal is to provide financial institutions with actionable recommendations to mitigate credit risk and support informed lending decisions.

Future Scopes

- Develop transparent dashboards for stakeholder insights.
 - Deploy the model on scalable cloud platforms.
 - Enable real-time predictions for instant loan decisions.
 - Integrate with financial systems via APIs.
 - Ensure GDPR compliance and data privacy.
- 

The background features three vertical stripes on the left: a wide pink stripe, a medium blue stripe, and a narrow beige stripe. The rest of the background is a light beige color with two rectangular areas of a pink dot pattern, one in the top right and one in the bottom right.

THANK YOU