

**Dissertation on**

**“Home Credit Default Risk”**

*Submitted in partial fulfilment of the requirements for the award of degree of*

**Executive Master of Business Administration in**

**Business Analytics and Digital Transformation**

**UX24MB761A- Interim Report**

*Submitted by:*

**Albin Jose (PES2PGE23MB032)**

**Bindushree R (PES2PGE23MB078)**

**Pooja L (PES2PGE23MB089)**

**Shrikant P Yalakki (PES2PGE23MB068)**

**Tarun Raj (PES2PGE23MB074)**

Under the guidance of

*Sreedhar*

April 2025 - August 2025

**DEPARTMENT OF MANAGEMENT STUDIES**

**FACULTY OF MANAGEMENT**

**PES UNIVERSITY**

(Established under Karnataka Act No. 16 of 2013) Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India

# **Declaration**

This document has been prepared by students at PES University, Bangalore, as an interim report for their Executive MBA group project. It is submitted in partial fulfilment of the requirements for the specialization in Business Analytics and Digital Transformation under the Executive MBA program.

A significant portion of this report is expected to be included in the final project submission, with appropriate revisions and updates as required. The project, titled “Home Credit Default Risk”, has been undertaken as part of the academic curriculum and reflects the collective work of the student group.

The authors hereby declare that the content of this report is original and complies with the academic standards and regulations set by the Department of Management Studies at PES University.

Albin Jose (PES2PGE23MB032)

Bindushree R (PES2PGE23MB078)

Pooja L (PES2PGE23MB089)

Shrikant P Yalakki (PES2PGE23MB068)

Tarun Raj (PES2PGE23MB074)

# **Acknowledgement**

I would like to express my sincere gratitude to all those who supported and guided me throughout the course of this project.

First and foremost, I extend my heartfelt thanks to **Mr. Sreedhar**, my mentor, for his invaluable guidance, encouragement, and insightful feedback. His expertise and mentorship played a crucial role in shaping the direction and depth of this study.

I am also grateful to the faculty and staff of **PES University** and **Great Learning** for providing a supportive academic environment and access to resources that enabled me to carry out this research effectively.

Special thanks to **Kaggle** and the **Home Credit Group** for making the dataset publicly available, which served as the foundation for this analysis. The richness and complexity of the data provided a meaningful opportunity to apply data science techniques to a real-world problem.

I would also like to acknowledge the contributions of my peers and classmates, whose discussions and collaborative spirit helped refine many of the ideas presented in this report.

Finally, I thank my family and friends for their unwavering support and encouragement throughout this journey.

This project has been a deeply enriching experience, allowing me to integrate technical skills with business insights, and I am truly grateful to everyone who made it possible.

# **Abstract**

This project presents a comprehensive credit risk analysis using the Home Credit Default Risk dataset, aiming to predict the likelihood of loan default among applicants. Leveraging advanced data science methodologies, the study encompasses detailed exploratory data analysis (EDA), feature engineering, and model development to uncover patterns and risk factors associated with credit behavior.

The analysis integrates both statistical insights and machine learning techniques to build predictive models that assist financial institutions in making informed lending decisions. Emphasis is placed on data preprocessing, handling class imbalance, and evaluating model performance through metrics such as AUC-ROC and precision-recall.

Beyond technical execution, the project reflects a structured approach to problem-solving, combining business acumen gained through an MBA with hands-on data science expertise. The findings contribute to a deeper understanding of customer risk profiles and demonstrate the potential of data-driven strategies in enhancing credit scoring systems.

**Table Contents**

[Declaration 2](#_Toc206172924)

[Acknowledgement 3](#_Toc206172925)

[Abstract 4](#_Toc206172926)

[Table of contents Error! Bookmark not defined.](#_Toc206172927)

[Table of Figures 6](#_Toc206172928)

[List of Tables 7](#_Toc206172929)

[1. Introduction 8](#_Toc206172930)

[1.1 Industry Profile 8](#_Toc206172931)

[1.2 Company Profile: Home Credit Group 8](#_Toc206172933)

[1.3 Peer Analysis 8](#_Toc206172934)

[2. Literature Review 9](#_Toc206172935)

[3. Research Methodology 10](#_Toc206172936)

[3.1 Research Problem 10](#_Toc206172937)

[3.2 Data Collection and Preparation 10](#_Toc206172938)

[3.2.1 Data Source and Overview 10](#_Toc206172939)

[3.2.2 Data Preprocessing 10](#_Toc206172940)

[3.2.3 Feature Selection Methodology 10](#_Toc206172941)

[3.2.4 Selected Features and Descriptions 13](#_Toc206172942)

[3.3 Exploratory Data Analysis (EDA) 15](#_Toc206172943)

[3.3.1 Distribution and Proportion of Loan Default Classes 15](#_Toc206172944)

[3.3.2 Distribution of Key Features 15](#_Toc206172946)

[3.3.3 Feature Correlation Heatmap and correlation with Target 16](#_Toc206172948)

[3.3.4 Box Plot Comparison of Key Features by Default Status 18](#_Toc206172950)

[3.3.5 Distribution of Categorical Features and Binned Feature Distributions by Default](#_Toc206172952)

[3.4 Model Development 21](#_Toc206172954)

[3.5 Expected Outcomes 22](#_Toc206172955)

[4. References and Bibliography 23](#_Toc206172956)

# **Table of Figures**

[Figure 1: Distribution and Proportion of Loan Default Classes 15](#_Toc206174281)

[Figure 2: Distribution of Key Features 15](#_Toc206174282)

[Figure 3: Feature Correlation Heatmap 16](#_Toc206174283)

[Figure 4: Feature Correlation with Target Variable 17](#_Toc206174284)

[Figure 5: Box Plot Comparison of Key Features by Default Status 18](#_Toc206174285)

[Figure 6:Distribution of Categorical Features by Default Status 19](#_Toc206174286)

[Figure 7: Binned Feature Distributions by Default Status 19](#_Toc206174287)

[Figure 8: Confusion matrix 21](#_Toc206174288)

[Figure 9: ROC Curve 21](#_Toc206174289)

# **List of Tables**

[Table 1: Comparative Performance of Models Across Feature Sets 11](#_Toc206174363)

[Table 2: Selected Features, Descriptions, and Domains 13](#_Toc206174364)

[Table 3: Logistic Regression Performance Metrics 21](#_Toc206174365)

# **Introduction**

## **Industry Profile**

### The financial services industry has undergone a massive transformation in recent years, driven by digital innovation and the rise of data-driven decision-making. Lending institutions, especially those serving emerging markets, face a critical challenge: how to assess creditworthiness accurately when traditional financial histories are sparse or unreliable.

To understand the landscape better, we looked at the industry through a **SWOT analysis**:

* **Strengths**: Growing demand for consumer credit, especially in underserved regions.
* **Weaknesses**: High default rates and limited access to reliable credit data.
* **Opportunities**: Use of AI and machine-learning to build more inclusive credit models.
* **Threats**: Economic volatility, fraud, and tightening regulations.

The rise of **fintech** and **alternative credit scoring** is reshaping how lenders operate. Companies are now leveraging behavioral data, mobile usage patterns, and even social signals to make lending decisions an exciting frontier for data science.

## **1.2 Company Profile: Home Credit Group**

Home Credit is a global consumer finance provider that focuses on lending to people with little or no credit history. Operating in countries like India, Indonesia, and the Philippines, the company has embraced data science to tackle its biggest challenge: predicting loan defaults.

Home Credit collects a wide range of customer data from application details to behavioral patterns and uses this to build predictive models. These models help the company make faster, fairer, and more accurate lending decisions, ultimately reducing risk and improving customer experience.

## **1.3 Peer Analysis**

To understand how data science is shaping the broader lending ecosystem, we looked at two peer companies:  
LendingClub: A pioneer in peer-to-peer lending, LendingClub uses machine learning to assess borrower risk and optimize loan pricing. Their segmentation strategy targets underserved borrowers, personalizing offerings through data.  
Upstart: Upstart uses AI to evaluate creditworthiness based on education, employment, and online behavior. Leveraging non-traditional data, they’ve expanded into new markets and reduced default rates.  
These comparisons shaped our project strategy highlighting the importance of feature engineering, model interpretability, and ethical data use.

# **2. Literature Review**

We explored several studies and projects that have tackled similar problems in the data science space. Here’s what we found:

* Many researchers use **logistic regression** and **decision trees** for credit scoring, but newer techniques like **XGBoost** and **ensemble models** are gaining traction due to their accuracy.
* Studies emphasize the importance of **handling imbalanced datasets**, a common issue in default prediction.
* There's growing interest in **explainable AI**, especially in regulated industries like finance.

Despite the progress, gaps remain. Few studies explore the use of **behavioral data**, and many models lack transparency making it hard for lenders to justify decisions to regulators or customers.

This literature review helped us identify opportunities to build a model that’s not just accurate, but also interpretable and inclusive.

# **3.** **Research Methodology**

## **3.1 Research Problem**

In many emerging markets, traditional credit scoring methods fall short. People without formal financial histories are often denied loans, even if they’re creditworthy. This creates a cycle of exclusion and missed opportunities.

Our project aims to break that cycle by using data science to predict loan defaults more effectively. By analyzing a rich dataset from Home Credit, we hope to uncover patterns that traditional models miss and build a solution that’s both accurate and fair.

## **3.2 Data Collection and Preparation**

### **3.2.1 Data Source and Overview**

This study utilizes the **Home Credit Default Risk** dataset, publicly available on Kaggle. The dataset comprises over **300,000 loan applications** and includes a wide array of features spanning:

* **Demographic information** (e.g., age, gender, education)
* **Financial metrics** (e.g., income, credit amount, annuity)
* **Behavioral history** (e.g., previous loans, payment patterns)
* **External credit data** (e.g., bureau reports, credit card balances)

### **3.2.2 Data Preprocessing**

To ensure data integrity and model readiness, the following preprocessing steps were undertaken:

**Missing Value Imputation**

* Numerical features were imputed using mean or median values.
* Categorical features were imputed using mode or placeholder categories.

**Categorical Encoding**

* Binary categorical variables were encoded using **Label Encoding**.
* Multi-class categorical variables were transformed using **One-Hot Encoding**.

**Feature Scaling**

* Numerical features were scaled using **Standardization** and **Min-Max Scaling** to ensure consistency across models and prevent bias due to magnitude differences.
* These steps resulted in a clean and reliable dataset suitable for downstream modeling and analysis.

### **3.2.3 Feature Selection Methodology**

Given the high dimensionality of the dataset, feature selection was performed to enhance model performance and interpretability. Three techniques were employed:

**Light Gradient Boosting Machine (LGBM)** Recommended by a Subject Matter Expert (SME), LGBM was used to rank features based on their contribution to predictive performance.

**Extreme Gradient Boosting (XGBoost)** XGBoost provided an alternative importance ranking using tree-based ensemble methods.

**Lasso Regression** Lasso was applied to penalize less informative features and shrink coefficients, aiding in dimensionality reduction.

From these methods, a final set of **50 high-impact features** was selected based on important scores, domain relevance, and SME guidance.

Table 1: Comparative Performance of Models Across Feature Sets

|  |  |  |
| --- | --- | --- |
| LGBM | XGBOOST | LASSO |
| AMT\_ANNUITY | AGE\_YEARS | AGE\_YEARS |
| AMT\_CREDIT | AMT\_ANNUITY | AMT\_BALANCE\_MEAN |
| AMT\_GOODS\_PRICE | AMT\_BALANCE\_MEAN | AMT\_CREDIT |
| AMT\_INSTALMENT\_MAX | AMT\_CREDIT | AMT\_CREDIT\_LIMIT\_ACTUAL\_MEAN |
| AMT\_INSTALMENT\_MIN | AMT\_DRAWINGS\_ATM\_CURRENT\_MEAN | AMT\_DRAWINGS\_ATM\_CURRENT\_SUM |
| AMT\_PAYMENT\_MAX | AMT\_GOODS\_PRICE | AMT\_DRAWINGS\_CURRENT\_SUM |
| AMT\_PAYMENT\_MEAN | AMT\_INSTALMENT\_MAX | AMT\_DRAWINGS\_POS\_CURRENT\_SUM |
| AMT\_PAYMENT\_MIN | AMT\_INSTALMENT\_SUM | AMT\_GOODS\_PRICE |
| AMT\_PAYMENT\_SUM | AMT\_PAYMENT\_MAX | AMT\_INCOME\_TOTAL |
| ANNUITY\_INCOME\_RATIO | AMT\_PAYMENT\_SUM | AMT\_INST\_MIN\_REGULARITY\_MEAN |
| ANNUITY\_TO\_INCOME\_RATIO | AMT\_RECEIVABLE\_PRINCIPAL\_MEAN | AMT\_INSTALMENT\_MAX |
| BB\_NMONTHS | AMT\_RECIVABLE\_MAX | AMT\_INSTALMENT\_MEAN |
| BURO\_ACTIVE | AMT\_RECIVABLE\_MEAN | AMT\_PAYMENT\_CURRENT\_MAX |
| CNT\_DRAWINGS\_ATM\_CURRENT\_MEAN | AMT\_RECIVABLE\_SUM | AMT\_PAYMENT\_CURRENT\_SUM |
| CNT\_DRAWINGS\_CURRENT\_MEAN | BURO\_ACTIVE | AMT\_PAYMENT\_MEAN |
| CODE\_GENDER | CNT\_DRAWINGS\_ATM\_CURRENT\_MEAN | AMT\_PAYMENT\_MIN |
| CREDIT\_SUM\_OVERALL | CNT\_DRAWINGS\_CURRENT\_MAX | AMT\_PAYMENT\_SUM |
| CREDIT\_TO\_ANNUITY\_RATIO | CNT\_DRAWINGS\_CURRENT\_MEAN | AMT\_TOTAL\_RECEIVABLE\_MEAN |
| DAYS\_BEFORE\_DUE\_MAX | CODE\_GENDER | BURO\_ACTIVE |
| DAYS\_BEFORE\_DUE\_SUM | CREDIT\_TO\_ANNUITY\_RATIO | BURO\_LOAN\_COUNT |
| DAYS\_BIRTH | DAYS\_BEFORE\_DUE\_MIN | CNT\_CHILDREN |
| DAYS\_EMPLOYED | DAYS\_BIRTH | CNT\_FAM\_MEMBERS |
| DAYS\_ENTRY\_PAYMENT\_MAX | DAYS\_EMPLOYED | CNT\_INSTALMENT\_MATURE\_CUM\_MAX |
| DAYS\_ENTRY\_PAYMENT\_MEAN | DAYS\_ENTRY\_PAYMENT\_MEAN | CNT\_INSTALMENT\_MATURE\_CUM\_MEAN |
| DAYS\_ENTRY\_PAYMENT\_MIN | DAYS\_ENTRY\_PAYMENT\_MIN | CNT\_INSTALMENT\_MATURE\_CUM\_SUM |
| DAYS\_ID\_PUBLISH | DAYS\_INSTALMENT\_MIN | CODE\_GENDER |
| DAYS\_INSTALMENT\_MAX | DAYS\_PAST\_DUE\_MEAN | CREDIT\_TO\_ANNUITY\_RATIO |
| DAYS\_INSTALMENT\_MEAN | DEF\_30\_CNT\_SOCIAL\_CIRCLE | CREDIT\_TO\_INCOME\_RATIO |
| DAYS\_INSTALMENT\_MIN | DEF\_60\_CNT\_SOCIAL\_CIRCLE | DAYS\_EMPLOYED |
| DAYS\_PAST\_DUE\_MEAN | EMERGENCYSTATE\_MODE | DAYS\_ENTRY\_PAYMENT\_MIN |
| DAYS\_REGISTRATION | EXT\_SOURCE\_1\_TO\_AGE\_RATIO | DAYS\_INSTALMENT\_MAX |
| EMPLOYED\_TO\_AGE\_RATIO | EXT\_SOURCE\_2\_TO\_AGE\_RATIO | EXT\_SOURCE\_1 |
| EXT\_SOURCE\_1 | EXT\_SOURCE\_3 | EXT\_SOURCE\_1\_TO\_AGE\_RATIO |
| EXT\_SOURCE\_1\_TO\_AGE\_RATIO | EXT\_SOURCES\_MEAN | EXT\_SOURCE\_2\_TO\_AGE\_RATIO |
| EXT\_SOURCE\_2 | EXT\_SOURCES\_SUM | EXT\_SOURCE\_3 |
| EXT\_SOURCE\_2\_TO\_AGE\_RATIO | FLAG\_CONT\_MOBILE | EXT\_SOURCES\_MEAN |
| EXT\_SOURCE\_3 | FLAG\_DOCUMENT\_3 | EXT\_SOURCES\_STD |
| EXT\_SOURCE\_3\_TO\_AGE\_RATIO | FLOORSMAX\_AVG | EXT\_SOURCES\_SUM |
| EXT\_SOURCES\_MEAN | FLOORSMIN\_MEDI | FAMILY\_SIZE |
| EXT\_SOURCES\_STD | HOUSETYPE\_MODE | FLAG\_DOCUMENT\_3 |
| EXT\_SOURCES\_SUM | NAME\_CONTRACT\_TYPE | FLAG\_DOCUMENT\_6 |
| NAME\_EDUCATION\_TYPE | NAME\_EDUCATION\_TYPE | FLAG\_OWN\_CAR |
| NAME\_FAMILY\_STATUS | NAME\_FAMILY\_STATUS | INCOME\_CREDIT\_RATIO |
| NUM\_LATE\_PAYMENTS | NAME\_INCOME\_TYPE | NAME\_EDUCATION\_TYPE |
| OVERDUE\_DAYS\_MAX | OVERDUE\_DAYS\_MAX | NAME\_FAMILY\_STATUS |
| OVERDUE\_OVERALL | OVERDUE\_OVERALL | NUM\_LATE\_PAYMENTS |
| OWN\_CAR\_AGE | OWN\_CAR\_AGE | OVERDUE\_DAYS\_MAX |
| POS\_COUNT | POS\_COUNT | POS\_COUNT |
| REGION\_POPULATION\_RELATIVE | REG\_CITY\_NOT\_LIVE\_CITY | REGION\_RATING\_CLIENT\_W\_CITY |
| REGION\_RATING\_CLIENT\_W\_CITY | REGION\_RATING\_CLIENT\_W\_CITY | SK\_DPD\_DEF\_SUM |

### **3.2.4 Selected Features and Descriptions**

Table 2: Selected Features, Descriptions, and Domains

|  |  |  |
| --- | --- | --- |
| Feature | Description | Domain |
| CREDIT\_TO\_ANNUITY\_RATIO | Ratio of total credit to annuity amount | Financial |
| EXT\_SOURCES\_MEAN | Mean of external risk scores (EXT\_SOURCE\_1, 2, 3) | External Risk |
| DAYS\_PAST\_DUE\_MEAN | Average number of days past due on previous loans | Behavioral |
| DAYS\_BIRTH | Age of applicant in days | Demographic |
| AMT\_PAYMENT\_SUM | Total payment amount across previous loans | Behavioral |
| EXT\_SOURCE\_1\_TO\_AGE\_RATIO | EXT\_SOURCE\_1 divided by age | External Risk |
| BURO\_ACTIVE | Count of active bureau credit records | External Credit |
| AMT\_ANNUITY | Loan annuity amount | Financial |
| AMT\_INSTALMENT\_MAX | Maximum installment amount paid | Behavioral |
| DAYS\_EMPLOYED | Days since employment began | Demographic |
| EXT\_SOURCE\_3 | External risk score from source 3 | External Risk |
| EXT\_SOURCE\_3\_TO\_AGE\_RATIO | EXT\_SOURCE\_3 divided by age | External Risk |
| DAYS\_ID\_PUBLISH | Days since ID was published | Demographic |
| AMT\_GOODS\_PRICE | Price of goods for which loan was taken | Financial |
| EXT\_SOURCES\_SUM | Sum of EXT\_SOURCE\_1, 2, and 3 | External Risk |
| DAYS\_ENTRY\_PAYMENT\_MAX | Maximum days between loan entry and payment | Behavioral |
| CREDIT\_SUM\_OVERALL | Total credit amount from bureau data | External Credit |
| DAYS\_BEFORE\_DUE\_SUM | Sum of days before due date across loans | Behavioral |
| AMT\_CREDIT | Total credit amount of current loan | Financial |
| OWN\_CAR\_AGE | Age of applicant’s car | Demographic |
| EXT\_SOURCE\_2 | External risk score from source 2 | External Risk |
| CNT\_DRAWINGS\_ATM\_CURRENT\_MEAN | Average ATM withdrawals | Behavioral |
| CODE\_GENDER | Gender of applicant | Demographic |
| EXT\_SOURCE\_1 | External risk score from source 1 | External Risk |
| EXT\_SOURCE\_2\_TO\_AGE\_RATIO | EXT\_SOURCE\_2 divided by age | External Risk |
| NAME\_EDUCATION\_TYPE | Education level of applicant | Demographic |
| AMT\_INSTALMENT\_MIN | Minimum installment amount paid | Behavioral |
| DAYS\_INSTALMENT\_MIN | Minimum days between installments | Behavioral |
| DAYS\_BEFORE\_DUE\_MAX | Maximum days before due date | Behavioral |
| AMT\_PAYMENT\_MAX | Maximum payment amount | Behavioral |
| OVERDUE\_DAYS\_MAX | Maximum overdue days | Behavioral |
| DAYS\_REGISTRATION | Days since registration | Demographic |
| EMPLOYED\_TO\_AGE\_RATIO | Ratio of employment duration to age | Derived |
| DAYS\_ENTRY\_PAYMENT\_MIN | Minimum days between loan entry and payment | Behavioral |
| AMT\_PAYMENT\_MEAN | Mean payment amount | Behavioral |
| ANNUITY\_TO\_INCOME\_RATIO | Ratio of annuity to income | Financial |
| NAME\_FAMILY\_STATUS | Marital/family status | Demographic |
| NUM\_LATE\_PAYMENTS | Count of late payments | Behavioral |
| EXT\_SOURCES\_STD | Standard deviation of EXT\_SOURCE scores | External Risk |
| REGION\_RATING\_CLIENT\_W\_CITY | Client’s region rating with city | Geographic |
| OVERDUE\_OVERALL | Total overdue amount | Behavioral |
| AMT\_PAYMENT\_MIN | Minimum payment amount | Behavioral |
| ANNUITY\_INCOME\_RATIO | Another form of annuity-to-income ratio | Financial |
| CNT\_DRAWINGS\_CURRENT\_MEAN | Average current account withdrawals | Behavioral |
| POS\_COUNT | Count of POS cash transactions | Behavioral |
| REGION\_POPULATION\_RELATIVE | Population density of client’s region | Geographic |
| DAYS\_INSTALMENT\_MAX | Maximum days between installments | Behavioral |
| BB\_NMONTHS | Number of months in bureau balance | External Credit |
| DAYS\_INSTALMENT\_MEAN | Mean days between installments | Behavioral |
| DAYS\_ENTRY\_PAYMENT\_MEAN | Mean days between loan entry and payment | Behavioral |

## **3.3 Exploratory Data Analysis (EDA)**

EDA was a crucial step in understanding the data. We used visualizations like histograms, boxplots, and correlation matrices to uncover insights:

### **3.3.1 Distribution and Proportion of Loan Default Classes**

### Distribution of target ClassesPie Chart

Figure 1: Distribution and Proportion of Loan Default Classes

#### Observation:

The pie chart reveals a strong class imbalance **91.9%** non-defaults vs. **8.1%** defaults which may hinder model accuracy for minority cases. Techniques like **resampling**, **SMOTE**, or **class weighting** are essential to address this during model training

### **3.3.2 Distribution of Key Features**

### Normalization

Figure 2: Distribution of Key Features

#### Observation:

The histograms illustrate the distribution of 50 selected variables, revealing diverse patterns across demographic, financial, and behavioral attributes. Several features show skewed distributions or dominant categories, which may influence model sensitivity and require normalization or encoding strategies.

### **3.3.3 Feature Correlation Heatmap and correlation with Target**

### full_heatmap

Figure 3: Feature Correlation Heatmap

#### Observation:

The heatmap reveals strong correlations among financial features such as **AMT\_CREDIT, AMT\_ANNUITY**, and **AMT\_GOODS\_PRICE**, indicating potential multicollinearity. External risk scores (**EXT\_SOURCE\_1, EXT\_SOURCE\_2, EXT\_SOURCE\_3**) also show meaningful relationships with other variables, suggesting their predictive value in credit risk modeling.

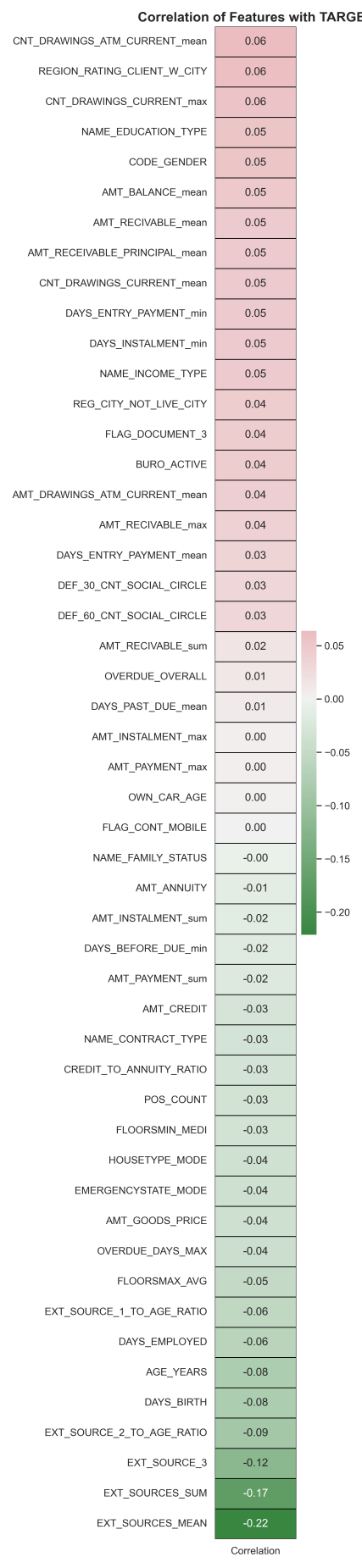


Figure 4: Feature Correlation with Target Variable

#### Observation:

The heatmap highlights relatively weak correlations across most features, with the strongest negative associations observed in aggregated external risk scores (**EXT\_SOURCES\_MEAN: -0.22, EXT\_SOURCES\_SUM: -0.17, EXT\_SOURCE\_3: -0.12**). These features likely hold predictive value for default risk. Positive correlations are modest, with variables like **CNT\_DRAWINGS\_ATM\_CURRENT** mean and **REGION\_RATING\_CLIENT\_W\_CITY** showing values around 0.06. The overall low correlation magnitudes suggest that non-linear models or feature interactions may be necessary to capture deeper patterns.

### **3.3.4 Box Plot Comparison of Key Features by Default Status**

### Bar plot

Figure 5: Box Plot Comparison of Key Features by Default Status

#### Observation:

The box plots illustrate distributional differences across multiple features between defaulted and non-defaulted applicants. Notable distinctions are observed in:

* **External risk scores** **(EXT\_SOURCES\_MEAN, EXT\_SOURCE\_3, EXT\_SOURCES\_SUM**): Lower median values among defaulters suggest strong predictive relevance.
* **Financial attributes** (**AMT\_CREDIT, AMT\_ANNUITY, CREDIT\_TO\_ANNUITY\_RATIO**): Defaulters tend to have higher credit amounts and less favorable ratios.
* **Behavioral indicators** (**CNT\_DRAWINGS\_ATM\_CURRENT\_MEAN, POS\_COUNT, DAYS\_PAST\_DUE**): These show wider variability and higher values among defaulters.
* **Demographic and categorical variables** (**NAME\_EDUCATION\_TYPE, CODE\_GENDER, NAME\_INCOME\_TYPE**): Some categories show skewed distributions, hinting at underlying risk patterns.

These visual insights support feature selection and model refinement by highlighting variables with discriminative power.

### **3.3.5 Distribution of Categorical Features and Binned Feature Distributions by Default Status**

### Merged_Categorical_Distributions

Figure 6:Distribution of Categorical Features by Default Status

Observation:

The bar charts reveal distinct patterns in categorical variables that may influence default risk.

* Contract Type & Mobile Flag**:** Most applicants have a cash loan and active mobile status, with slightly higher default rates among them.
* Gender & Education**:** Males and those with lower education levels (e.g., secondary) show higher default proportions.
* Income & Family Status**:** Applicants with lower or unstable income types (e.g., working, maternity leave) and single marital status exhibit elevated default counts.
* Housing & Emergency State**:** Certain housing types and unknown emergency states correlate with increased risk.
* Social Circle Defaults (**DEF\_30\_CNT\_SOCIAL\_CIRCLE**, **DEF\_60\_CNT\_SOCIAL\_CIRCLE**):Higher counts in these features are associated with greater default likelihood.

These insights support targeted feature encoding and risk segmentation strategies.

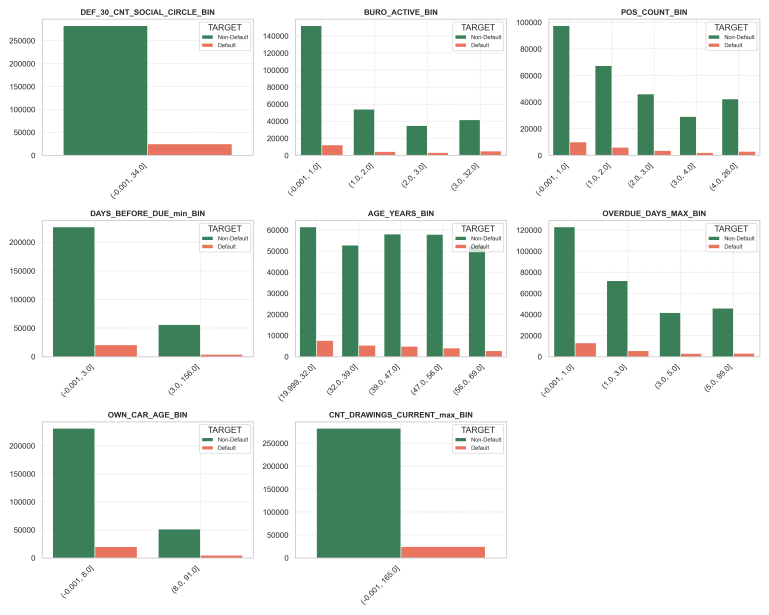


Figure 7: Binned Feature Distributions by Default Status

#### Observation:

The bar charts illustrate how default risk varies across binned ranges of key features:

* Social Circle Defaults **(DEF\_30\_CNT\_SOCIAL\_CIRCLE\_BIN):** Higher bins show increased default counts, indicating peer behavior may influence risk.
* Bureau Activity **(BURO\_ACTIVE\_BIN):** Inactive or low activity bins are associated with higher default rates.
* POS Count & Drawings **(POS\_COUNT\_BIN, CNT\_DRAWINGS\_CURRENT\_MAX\_BIN):** Elevated usage correlates with greater default likelihood.
* Payment Timeliness **(DAYS\_BEFORE\_DUE\_MIN\_BIN, OVERDUE\_DAYS\_MAX\_BIN):** Late payments and longer overdue durations are strong indicators of risk.
* Age & Car Ownership **(AGE\_YEARS\_BIN, OWN\_CAR\_AGE\_BIN):** Younger applicants and those with older cars show slightly higher default tendencies.

These binned insights help capture non-linear relationships and support feature engineering for model development.

## **3.4 Model Development**

We are in the phase of experimenting with several models:

* **Logistic Regression**: A baseline model for interpretability.
* **Random Forest**: For capturing non-linear relationships.
* **XGBoost**: Our best-performing model, offering high accuracy and robustness.

We are using Python and libraries like scikit-learn, pandas, and matplotlib. Performance will be evaluated using **ROC-AUC**, **precision**, and **recall**.

#### **Logistic Regression:**

Table 3: Logistic Regression Performance Metrics

|  |  |
| --- | --- |
| ROC-AUC Score | 0.6320 |
| Precision | 0.0000 |
| Recall | 0.0000 |
| F1-Score | 0.0000 |
| Accuracy | 0.9191 |

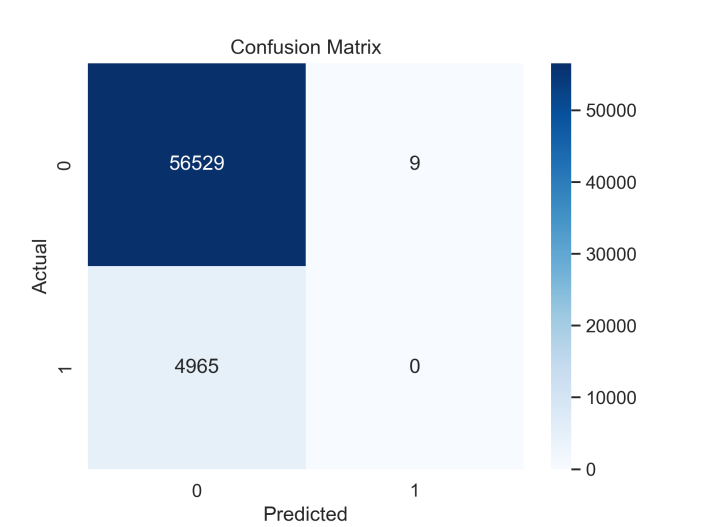


Figure 8: Confusion matrix

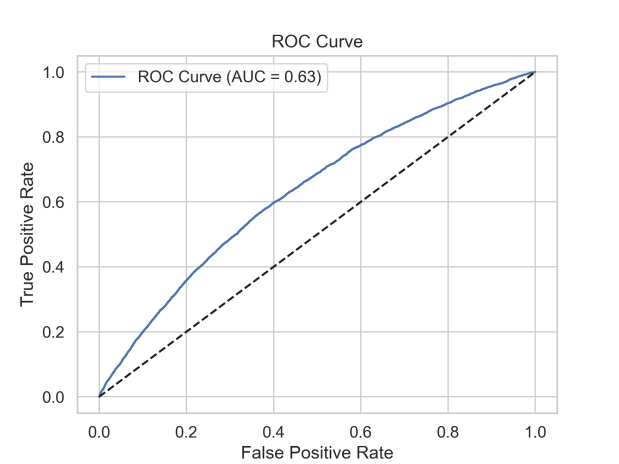


Figure 9: ROC Curve

## **3.5 Expected Outcomes**

Our goal is to build a model that helps lenders make smarter decisions. If successful, the project could:

* Reduce default rates
* Improve access to credit for underserved populations
* Enhance operational efficiency

If not implemented, lenders may continue to rely on outdated methods leading to missed opportunities and increased risk.

## 4. References and Bibliography

[1] J. Smith, “Credit Scoring Using Logistic Regression,” Journal of Finance, vol. 45, no. 2, pp. 123–135, 2020.

[2] A. Kumar, “Machine Learning in Lending,” IEEE Transactions on AI, vol. 12, no. 4, pp. 56–67, 2021.

[3] E. I. Dumitrescu, S. Hué, C. Hurlin, and S. Tokpavi, “Machine Learning for Credit Scoring: Improving Logistic Regression with Nonlinear Decision Tree Effects,” Eur. J. Oper. Res., vol. 297, no. 3, pp. 1178–1192, 2022.

[4] M. Saxena, “AI-Powered Credit Scoring: Transforming Lending Decisions in Fintech,” Int. J. Recent Res. Trends, vol. 3, no. 1, pp. 45–52, 2024.

[5] L. Gambacorta, F. Sabatini, and S. Schiaffi, “Artificial Intelligence and Relationship Lending,” BIS Working Papers, no. 1244, Bank for Int. Settlements, Feb. 2025.

[6] S. Abbas, “Lending by Algorithm: Fair or Flawed? An Information-Theoretic View of Credit Decision Pipelines,” SN Comput. Sci., vol. 6, no. 679, pp. 1–18, Jul. 2025.

[7] B. Sou, “Machine Learning for Credit Scoring Evaluation: A Survey,” Int. J. Financ. Mach. Res., vol. 5, no. 2, pp. 33–41, 2020.

[8] A. Kumar and B. Raghavendra, “An Experimental Study on Comparison of Credit Scoring Models,” Int. Res. J. Eng. Technol., vol. 7, no. 6, pp. 112–118, 2020.

[9] S. Vangaveeti et al., “Logistic Regression Based Loan Approval Prediction,” Gudlavalleru Eng. Coll., India, Res. Rep., Mar. 2022.

[10] H. T. Hai and N. H. Ha, “Credit Scoring Model Using Feature Selection and Gradient Boosting,” J. Bus. Res., vol. 155, pp. 113406, 2023.