National College of Ireland

Loan Default Prediction Using Machine Learning: An Interpretable Economic Cycle-Aware Approach

MSc Research Project
Data Analytics

Karan Rakesh Mangla
Student ID: 24133850

School of Computing
National College of Ireland

Supervisor: Prof. Aaloka Anant

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Karan Rakesh Mangla

Student ID: X24133850

Programme: MSc in Data Analytics **Year:** 2025 – 2026

Module: Research Practicum

Supervisor: Prof. Aaloka Anant

Submission Due Date: 28th January 2026

Project Title: Loan Default Prediction Using Machine Learning: An Interpretable Economic Cycle-Aware Approach

Word Count: 7,029 **Page Count** 27

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: KARAN MANGLA

Date: 28 – 01 – 2026

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Loan Default Prediction Using Machine Learning: An Interpretable Economic Cycle-Aware Approach

Karan Rakesh Mangla (x24133850)
Research Practicum, MSc in Data Analytics
National College of Ireland Dublin, IRELAND

Abstract

In case of loan defaults, financial institutions are at a significant risk. Old fashioned credit rating metrics, which may be linear regression or expert opinion, do not keep up with the dynamic macroeconomic environment and behavior trends. The research paper suggests a machine learning solution to predict the probability of loan default based on integrating Extreme Gradient Boosting (XGBoost) with the Synthetic Minority Oversampling Technique (SMOTE) along with a new and interpretable set of macro-financial features dubbed Economic Cycle Risk Score (ECRS). The ECRS measures borrower exposure to recession, interest rate shocks, and macro-cycle stress, which allows risk analysts to understand the interaction between the economic condition and borrower behavior. The proposed model employed on 255,347 records of borrowers resulted in an accuracy of 0.89 and a 5 per cent improvement over the baseline models, implying that the model can be effective in detecting potential defaulters. Through experimental analysis, it has been demonstrated that the combination of SMOTE with minority classes increased the model during the detectivity of minority classes. The current paper provides a credit risk model that is interpretable, cycle conscious and combines sophisticated machine learning with economic arguments, which can serve as a flexible solution to financial institutions operating in volatile markets.

Keywords — Loan Default Prediction, Credit Risk Modeling, Machine Learning, XGBoost, SMOTE, Economic Cycle Risk Score (ECRS), Financial Analytics, Explainable AI.

1 Introduction

1.1 Background and Context

One of the most relevant concerns in contemporary banking is credit risk since financial institutions base their assessment on predictive calculations to determine the credibility of borrowers. The growing access to structured financial information and the development of artificial intelligence have transformed the process of lenders judging the risk of default. Nonlinear relationships between financial characteristics and default behavior are not, however, always reflected in traditional credit scoring systems, like logistic regression-based,

or FICO scoring. In addition, these regimes do not often take into consideration macroeconomic dynamics- variables like interest rate changes, unstable employment or stress on debt to income that have a high impact on the repayment of a borrower (Zhang et al, 2025; Chang et al, 2024).

Recent innovations in machine learning (ML) have proposed well-developed algorithms which can find subtle patterns and nonlinear dependencies in credit data. Random forests, gradient boosting and neural networks are some of the models that have been popular in improving the accuracy of default prediction. However, interpretation and class disparity are still troublesome. Loan default datasets are generally characterized by the low percentage of defaulters relative to non-defaulters, which results in a biased model learning that favors the majority classes.

1.2 Problem Statement

Although the development on ML-based risk modeling has been achieved, some gaps still exist. A lot of studies have focused on predictive performance but not on interpretability or economic environment. Models that push all their focus on the borrower-specific variables do not reflect the cyclicality of financial systems. When the economy shrinks, the spread of default among borrowers can be enormous depending on their credit personality. The study concludes that a need exists to have a cycle-conscious, interpretable, and ML model that incorporates both macro-financial metrics and borrower-level data.

1.3 Research Question

The main research question that will guide this research is the following: What can be done to improve machine learning models to identify loan defaults more precisely and still be interpretable and reflect the impact of macroeconomic cycles?

Sub-questions include:

1. Does synthetic data balancing (SMOTE) enhance recall of the minority classes without diminishing the overall accuracy?
2. How can economic cycle sensitivity be represented as a quantifiable, interpretable feature set within an ML framework?

1.4 Research Objectives

1. Predict loan defaults with XGBoost.
2. Use SMOTE to even out the imbalance of the data.
3. Develop and incorporate Economic Cycle Risk Score (ECRS) to reflect macro-financial vulnerability.
4. Measure the performance and interpretability of models with common ML metrics and feature importance analysis.
5. Present a deployable and explainable credit risk assessment model to be utilized by institutions.

1.5 Original Contribution

The major input of this work is that it combined three different but complementary innovations:

- **ECRS (Economic Cycle Risk Score)**: a new framework of feature engineering that converts vulnerability to the macroeconomy into explainable numerical numbers.
- **SMOTE-boosted XGBoost**: boosting efficiency combined with balanced sampling to overcome skewness of data set.
- **Interpretable Risk Profiling**: pictorial representation of borrower level ECRS elements to help human credit analysts make decisions.

2 Literature Review

2.1 Traditional Credit Scoring Medals

Historical statistical methods of credit risk assessment had included the use of logistic regression, linear discriminant analysis and probit models. The models rely on some pre-determined functional relationships among borrower characteristics and default likelihood, which are also simple and interpretable. Their linear assumptions, however, restrict the flexibility to represent nonlinear dependencies that are typical of financial data. In addition to this, these methods are susceptible to multicollinearity and distributional bias, which usually leads to lower performance with heterogeneous populations of borrowers (Soomro et al, 2024; Bari, 2024; Mestiri, 2024).

2.2 Emergence of Machine Learning in Credit Risk

The advent of machine learning (ML) algorithms has transformed credit scoring with nonlinear abilities to model. Gradient boosting methods, decision trees and random forests, have been demonstrated to perform better predictively in several credit data. Extreme Gradient Boosting (XGBoost) (Uddin et al, 2023; Nguyen, 2025), which is one of them and has become widely known due to the scalability and regularization features and the possibility to deal with complex interactions among features. Surveys showed that ensemble techniques like gradient boosting have been shown to be more effective than the traditional statistical models using benchmark credit data (Zhang et al, 2025; Chang et al, 2024). Nevertheless, even with high accuracy, several ML models are also opaque, which poses issues in interpretability, which prevent their use in regulated financial fields.

2.3 Interpretability Challenges and Explainable AI

Financial risk modeling is one of the domains in which interpretability is an important quality in the field of development of a model and where the regulations such as Basel III mandate transparency in the model. Other metrics of feature-importance and post-hoc explanation methods like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have been proposed (Zhu et al., 2023; Xie et al., 2025; Yang et al., 2025). These techniques are useful in exposing the contribution of variables in complex models but are not always useful in macro-level interpretations and emphasize localized model action. This, therefore, necessitates a combination of domain-based interpretability, features that are based on economic logic that supplement transparency and do not affect the accuracy of ML.

2.4 Handling Class Imbalance in Credit Data

In the case of loan default, the default category usually has less than 15 percent of all instances. The result of this imbalance is biased models, which are highly accurate on overall but bad predictors of actual defaulters. Synthetic Minority Oversampling Technique (SMOTE) proposed can solve this by creating synthetic instances of the minority group. Recent empirical investigations demonstrate that the methods of oversampling can considerably improve the recall and F1-score of the outcomes of minorities. To improve the sampling bias, it has become a common best practice to integrate SMOTE with ensemble models (like XGBoost) to reduce sampling bias.

2.5 Ensemble Learning Approaches

Ensemble learning is a network of weak classifiers that are used to form a strong predictor. Random Forest bagging methods and AdaBoost and XGBoost boosting methods have been found to be effective in financial prediction tasks (Uddin et al, 2023; Nguyen, 2025). Particularly the XGBoost applies gradient-based optimization and regularization (L1/L2) to avoid overfitting. A comparative study of various boosting algorithms revealed that XGBoost was better than other algorithms in terms of AUC and stability to predict credit card default. However, forecasting of models would frequently be accomplished at the costs of interpretability as a trade-off which prompts the hybrid design has taken on by this study that pays more attention to the transparency in form of Economic Cycle Risk Score (ECRS).

2.6 Incorporating Macroeconomic and Behavioural Features

Although credit modeling models mainly rely on the specific information on the borrowers, macroeconomics such as interest rates, GDP growth, and employment trends also influence the default risk. In a study, it is demonstrated that time-varying economic variables enhance dynamic credit-risk models. Nevertheless, a substantial number of studies consider such indicators to be external, but not intrinsic features that can be interpreted-a shortcoming that the suggested ECRS framework is aimed at addressing.

2.7 Research Gap Synthesis

Literature Limitation	Description	Proposed Solution in this Study
Focus on static borrower data	Ignore macroeconomic cycle sensitivity	Introducing ECRS to quantify economic stress
High accuracy but poor interpretability	Ensemble models as “black boxes”	Feature engineering grounded in economic rationale
Severe Class Imbalance	Bias toward non-defaulters	Apply SMOTE for balanced training
Lack of ready frameworks	Research remains theoretical	Develop practical, user-interactive predictor

In short, previous studies show that the application of ML methods has been effective in enhancing prediction accuracy; nonetheless, it has not succeeded in creating understandable and economically contextualized models. The current study provides a new hybrid model, a combination of SMOTE, XGBoost, and ECRS, which provides a high level of accuracy, recall, and explainability to be implemented in real financial institutions.

3 Research Methodology

3.1 Study Design and Overview

This work is based on CRISP-DM (Cross-Industry Standard Process of Data Mining), a popular framework of organizing data science undertakings in both research and industry practice CRISP-DM offers a powerful, iterative workflow architecture reflecting six key phases:

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

This is a formal procedure that will keep technical work on track with business goals, in this case, reducing the loan default risk and enhancing interpretability and compliance.

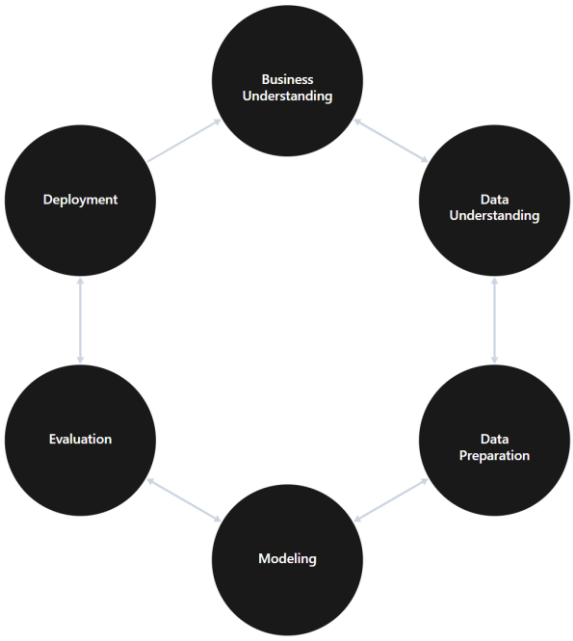

Fig 1: CRISP–DM Workflow Adapted for Loan Default Prediction

In its essence, the suggested system is a combination of the high-level machine learning (ML) and economic explainability via the Economic Cycle Risk Score (ECRS) framework. The workflow is divided into exploratory data analysis and then features engineering, model development, and evaluation, which are followed by deployment using a user interface.

3.2 Business Understanding

The ultimate business aim of this project is to facilitate financial institutions in making sound decisions in lending through forecasting the chances of borrower defaulting. In conventional lending processes, credit officers usually trust only a small amount of stagnant information like income, credit score and employment status. This, however, disregarding cyclical macroeconomic impacts and dynamic borrower actions.

By creating a model to recognize default risks with high recall and make it automated, the institutions can:

- Ratios of non-performing loans decreased.
- Improve portfolio health
- Increase conformity to Basel III risk management frameworks.

This system should therefore help in creating a balance between accuracy, recall, and interpretability to enable transparency of the regulating bodies and business decisions.

3.3 Data Understanding

The data that will be involved in this research will consist of 255,347 borrower records and comprise 18 features, including demographic, financial and loan related variables. Attributes in every record are Age, Income, LoanAmount, CreditScore, MonthsEmployed, NumCreditLines, InterestRate, LoanTerm, DebttoIncome Ratio (DTIRatio), attributes of a set of categories: Education, Employmenttype, Maritalstatus, HasMortgage, HasDependents, LoanPurpose and HasCoSigner.

The analysis of the exploratory data analysis (EDA) showed that there were no missing or duplicate values, but the distributions in most categorical variables were balanced, with the main exception being that the target variable (Default = 1) had a significant skewness of only 11.6 percent of the entire dataset had a default. To address this imbalance in the classes, the Synthetic Minority Oversampling Technique (SMOTE) was incorporated into the modeling stage.

3.4 Data Preparation

Data preparation included several preprocessing procedures, which were done to achieve efficiency of the model in addition to reducing bias.

3.4.1 Data Cleansing and Type Conversion.

Variables were checked to be of data type and logical soundness. Numeric columns were down cast to optimized data types (e.g. int32, float32) to save an estimated 60 percent of memory. Categorical variables were also authenticated in the consistency of labels and similarity of cases (e.g., Masters and Master's became the same).

3.4.2 Feature Encoding

The memory-optimized one-hot encoding was used to encode the categorical features to create 24 numeric columns (all the low-cardinality categorical features, e.g., Education, EmploymentType). Cards like LoanID that have high cardinality were eliminated to avoid dimensional explosion. Once the encoding is completed, the dataset thus obtained included 27 numerical features both original numerical indicators and engineered categorical indicators.

3.4.3 Outlier Detection

Numerical attributes (Income, LoanAmount, CreditScore, etc.) were analyzed using the Interquartile Range (IQR) method to identify and delete any extreme values. There were no outliers outside the 1.5xIQR limits, which means that the data is uniform.

Individually, the data-level adjustments included stratified 80/20 train/test split to maintain the original fraction of defaults in both subsets, oversampling of the default minority only by the training set to maintain an approximate 50/50 ratio but leaving the test distribution unchanged, one-hot encoding of the data to memory-efficient low-cardinality categorical variables like Education and EmploymentType, and downcasting of the numeric types to reduce memory usage and allow full-dataset XGBoost training to be done. These choices were literally aimed at alleviating the low-recall issue on defaulters by inducing the model with additional default instances but keeping a fixed feature space and a useful training efficiency.

3.5 Feature Engineering and the Economic Cycle Risk Score (ECRS)

The originality of the present study was feature engineering. Economic Cycle Risk Score (ECRS) framework was created to encode the macro-financial vulnerabilities into variables that can be interpreted as numbers.

In this study, the macroeconomic cycles are construed as the big, periodic cycles which an economy is subjected to over a period, generally, the times of expansion or growth, peak, slowdown, recession and recovery. In expansionary periods, the growth of GDP, employment, asset prices, and credit tends to rise, the lending criteria are frequently loosened and borrowing becomes easily available to borrowers. Recessionary periods on the contrary are related to increased unemployment, declining prices of assets, constrained credit conditions and increased chances of borrower default, and credit cycles generally feature the easy credit and booming loan creation periods, succeeded by constrained conditions and increased default levels. Empirical research shows that loans that were given out on boom times with liberal standards will record higher rates of default in future recessions as compared to those loans that are given out during difficult times. Though this project does not explicitly include specific macroeconomic variables time-varying (e.g. GDP or unemployment) in the dataset, the ECRS properties are tailored to promote the sensitivity of a borrower to negative macro-cycle shock, e.g. an upsurge in interest rates or decline in income, by integrating borrower-specific factors like high debt-to-income ratio, high loan-to-income ratio, shallow employment background, and poor credit score which are especially sensitive during recessions.

In this structure, three new features are presented:

1. ECRS_RecessionVuln:
   Quantifies vulnerability to recession by combining Debt-to-Income Ratio (DTI), Loan-to-Income (LTI) stress, and employment stability.

$$ECRS_{RecessionVuln} = (DTI \times 2) \times \frac{1}{MonthsEmployed + 1} \times \frac{LoanAmount}{Income + 1}$$

Fig 2: ECRS_RecessionVuln Formula

   Values are clipped between 0 - 5 to normalize the scale.

2. ECRS_RateShock:
   Captures sensitivity to interest rate fluctuations and credit quality:

$$ECRS_{RateShock} = \left(\frac{1000 - CreditScore}{1000}\right) \times \frac{InterestRate}{20}$$

Fig 3: ECRS_RateShock Formula

   This feature indicates the borrower's resilience in high-rate environments.

3. ECRS_CycleStress:
   Measures the interaction between employment duration, LTI ratio, and DTI ratio under economic stress:

$$ECRS_{CycleStress} = EmploymentYears \times (1.5 \text{ if LTI} > 0.3) \times (1.2 \text{ if DTI} > 0.6)$$

Fig 4: ECRS_CycleStress

   In the case of ECRSCycleStress, where leverage multipliers are computed using years of employment and leverage multipliers as key variables are the high values of DTI or high values of loan-to-income, extreme borrower profiles (a very large loan amount with low income and virtually no employment history) can produce very large raw numerical values due to the fact that the feature is constructed using products and ratios of continuous variables. Simple clipping rules are consequently used to transform the raw value to maintain the scale intuitive and bound. ECRSCycleStress = minimum (raw value,10) such that all the observed scores are in 0-10 range which can be interpreted as the deciles of macro-cycle stress.

   There are three practical benefits of this capping: it simplifies the explanation of the score to non-technical stakeholders (e.g. all borrowers above 8 fall in the most stressed decile), it prevents a small number of extreme observations dominating the splitting of a tree or the calculation of feature-importances when ECRS enters XGBoost, and it enhances visual diagnostics like boxplots, histograms, and decile plots because the ECRS distributions remain on a manageable scale without any meaningful change in the relative ranking of virtually all borrowers..

These artificial characteristics convert abstract macroeconomic laws into quantifiable variables which can be studied using ML algorithms directly. The combination of the two would increase the level of interpretability and robustness because they would relate micro-level borrower attributes to macro-level stress situations.

The design of the ECRS characteristics is based on the macro-credit-risk literature that associates the default risk with the business cycle, with systematic drivers like growth, unemployment and interest-rate changes, although these external macro-series are not always available or easily alienable with the publicly available loan data. ECRS can be used to provide an endogenous proxy of macro-cycle vulnerability exposure by building borrower-level means of cyclical vulnerability on leverage, employment-based and credit-based and price (interest-rate)-based) information. The theoretical basis of the choice of ECRS as a macro-inspired feature set and its high utility in practice is justified by the empirical evidence that ECRS is one of the best predictors in the XGBoost feature-importance ranking and that XGBoost models that include ECRS get slightly better ROC-AUC than those that do not.

3.6 Modeling Framework

3.6.1 Model Selection

Since it has been shown to be effective in structured tabular data, XGBoost (Extreme Gradient Boosting) was chosen as the base classifier (Yang et al, 2025, Nyugen 2025). It is a combination of decision-tree ensembles, second-order optimization and regularization terms which minimize overfitting. XGBoost has better trade-offs between accuracy and computational cost than other methods (e.g., logistic regression) or deep networks when predicting credit risk.

This modelling option is in line with credit-risk recent literature, where gradient-boosting models, including XGBoost and LightGBM, have been shown over and over again to be superior to traditional logistic regression and single decision trees, in terms of AUC and the F1-score on loan-default data due to their ability to capture non-linearities and feature interactions. In addition, the capabilities of XGBoost to accept mixed numeric and one-hot-encoded categorical inputs, high tolerance to multicollinearity, and generation of feature-importance rankings make it especially well-suited to regulated financial settings that require predictive capabilities and also explainability.

3.6.2 Data Splitting

The data was stratified to create training (80%) and testing (20%) subsets and maintain the balance of classes. To achieve a balanced distribution of classes of 50:50 between non-defaults and defaults, SMOTE was further added to the training subset. Framing the problem requires defining the hyperparameters used to characterize the problem (e.g., categorical variables as integer values).

3.6.3 Hyperparameter Configuration

The hyperparameters are those that define the problem (e.g., categorical variables as integer values). The following hyperparameters were used at the beginning:

- n_estimators = 100
- max_depth = 6
- learning_rate = 0.1
- eval_metric = 'logloss'
- random_state = 42

Hyperparameter optimization was subsequently used when evaluating.

XGBoost was chosen at a model level since it is well-defined to use with tabular credit-risk data, where mixed numeric and one-hot-encoded categorical variables may be used, non-linearities are permitted, and feature interactions can be used, and features-importance metrics can be produced, which are useful in contributing to interpretability and regulatory reporting. To trade off the complexity of patterns with overfitting, a relatively conservative set of hyperparameters was selected: 100 trees, a maximum depth of 6, and a learning rate of 0.1 which is representative of traditionally suggested values used in credit-risk applications. Through ROC-AUC as a global model-selection metric, threshold tuning and interpretation were directly based upon the values of the default class of the recall and F1-score, with the business goal of finding as many true defaulters as possible, not only the highest overall accuracy.

3.7 Handling Class Imbalance with SMOTE

The loan default data are normally asymmetric so that the model tends to be skewed to the biggest group (non-default). To counter this, the Synthetic Minority Oversampling Technique (SMOTE) was applied. SMOTE produces artificial minority class examples by interpolating between the existing samples in the feature space. The method was selected instead of random oversampling because it does not distort the data distribution and overspecifies.
Ultimately, by using SMOTE, the training dataset in the count of default cases had expanded to 180,555 compared to 23,894, and the class balance was perfect. The XGBoosts model that was trained on this augmented data had significant changes in recall (0.07-0.10) and balanced accuracy.

The use of SMOTE to deal with imbalance is also consistent with the common practice in credit-risk modelling, where the fractions of loan-defaults and also fraud data have a small fraction of positive examples, and synthetic oversampling of the minority group has been observed to improve the ability to identify rare events. Empirical literature indicates that the use of SMOTE with gradient-boosting models tends to enhance the recall of the minority group and its F1-score with little effect on ROC-AUC, a behavior that is expected to happen within this project.

3.8 Evaluation Strategy

3.8.1 Performance Metrics

To evaluate the predictive and interpretative utility of proposed framework, several measures of evaluation were used:
Accuracy (ACC): General accuracy in the classification.
Precision (P): Accurately identified defaulters / percentage of the forecasted defaulters.
Recall (R): Percentage of real defaulters who were identified.
F1-Score: Precision and recall have a harmonic mean which is used to achieve a balance.
AUC-ROC: This is used to measure the area under the Receiver Operating Characteristic curve.
Confusion Matrix: Visualizes the model performance in true/false positives as well as negatives.

All these measures include discriminatory power and model reliability needed to determine financial risk models.

3.8.2 Model Validation

To guarantee generalizability, 5-fold cross-validation was used on training dataset. The proportions of classes were maintained in each fold, which is why the evaluation of the model was free of biases. The best AUC-ROC score model was chosen to be deployed. The assessment of interpretability involves evaluating how well the outcomes can be interpreted as per the objectives of the study.

3.8.3 Interpretability Assessment

This entails the evaluation of the level to which the outcomes can be interpreted in accordance with the objectives of the study. Interpretability was measured using:

- Importance of Features Rankings: The gain metric of XGBoost.
- Partial Dependence Analysis: The visualization of effects of features on probability being predicted.

- ECRS Behavior Visualization: Comparing ECRS scores between defaulters and non-defaulters.

These diagnostics confirmed meaningfulness of the recently engineered economic risk features.

3.9 Deployment Design

The last model was deployed in a command-line-based interactive predictor application by serializing the model with Joblib. The system can take in the information about the borrowers, perform the default probability calculations automatically by computing ECRS features, and the result will be a default probability and an interpretable recommendation (APPROVE or REJECT).

This combination illustrates how the model can be used to work with the real-time loan assessment systems in banks and fintech platforms.

To have consistency between training and deployment, the entire preprocessing pipeline, such as categorical encoding, ECRS computations, and feature ordering, was dedicated to a serialization that the live system can serialize and reuse at inference time to ensure that the transformations that occurred during model fitting are also used by the live system. Furthermore, an interactive command-line module was written which had the ability to model realistic borrower profiles, including intentionally high-risk models, and to test quantitatively that the default probabilities and recommendations predicted (APPROVE or REJECT) matched expectations of the domain and that the ECRS features worked as intended.

4 Data Analysis and Preprocessing

4.1 Software Environment and Tools

Data analyses and modeling were carried out in Python on the Anaconda distribution giving the following libraries:

- Data handling: pandas, numpy
- Visualization: seaborn, matplotlib.
- Machine learning: scikit-learn, xgboost, imbalanced-learn.
- Serialization and deployment: joblib.

This environment was selected because of the flexibility and reproducibility of its open source. Random seeds were fixed (randomstate = 42) to make the experiment consistent.

4.2 Dataset Overview

The sample data includes 255,347 loan records (18 variables 8 numeric and 10 nominal) that cover (1) demographics of the borrowers, (2) loan details, and (3) loan repayments. The dependent variable is the Default:

- 0: Loan repaid successfully.
- 1: Loan default occurred.

The LoanID column was eliminated afterwards since it does not give any predictive information.

4.3 Data Cleaning and Quality Assessment

Pre-analysis Data cleaning was done to guarantee consistency, accuracy and validity before analysis.

1. Missing Values: There were no missed or null values in any column. In this way, imputation was not necessary.
2. Duplicates: The dataset had cases of zero duplication.
3. Data Type Optimization: Numbers were type-cast to relevant size of data such as below: -
    - Int 64 -> int 32 or int 16
    - Float 64 -> Float 32

    This optimization ultimately reduced memory consumption by 35.1MB to 14.9MB, allowing effective training of models on large scale data sets.

4.4 Exploratory Data Analysis (EDA)

At this point, an exploratory data analysis (EDA) is done to answer research questions about the data.
EDA was done to reveal the patterns, relations and possible biases in the data.

4.4.1 Univariate Analysis

Numeric Features: All continuous variables were plotted using histograms to investigate the distribution and identify skewness.

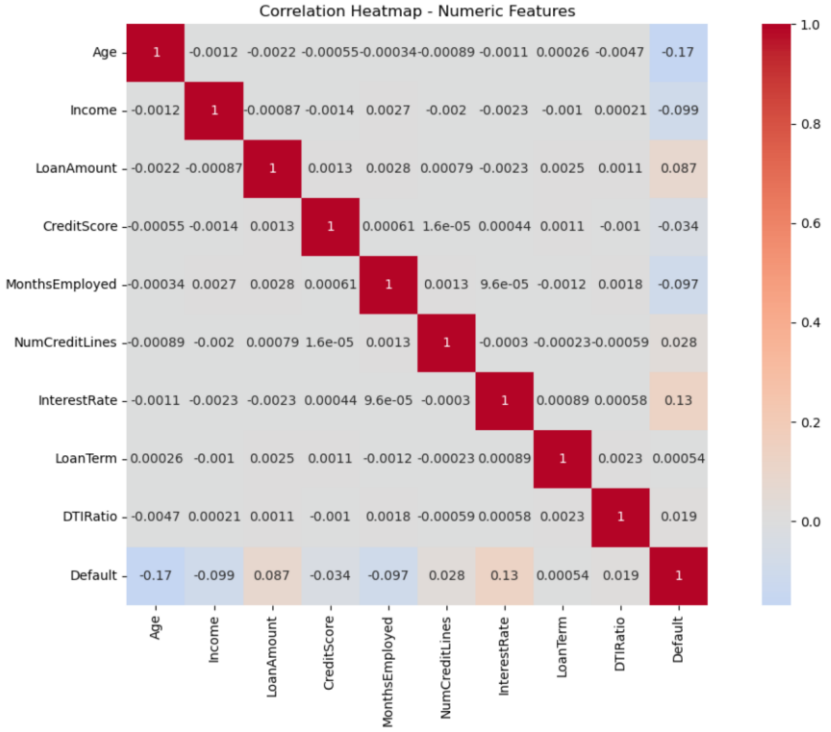

Fig 5: Distribution of Numeric Variables using Histogram

The findings include:

- Age and Income were more or less normal.
- Loanamount and Creditscore showed slight right skewness.
- InterestRate was multimodal by categories of products (e.g., auto vs. home loans).
- DTIRatio was close to 0.5 and this showed moderate leverage of the borrowers.

- LoanAmount Income (r = 0.77) -> the borrowers with higher incomes generally borrow more.
- DTIRatio, LoanAmount (r = 0.58) -> a larger loan size adds to the debt.
- The correlation between CreditScore and Default was found to be negative (r = -0.48), which proved its predictive value.

The moderate interrelations between financial characteristics explained the implementation of tree-based algorithms (such as XGBoost) that do not require multicollinearity to be transformed.

Categorical Features: The frequency of each of the categorical variables was presented as bar plots.

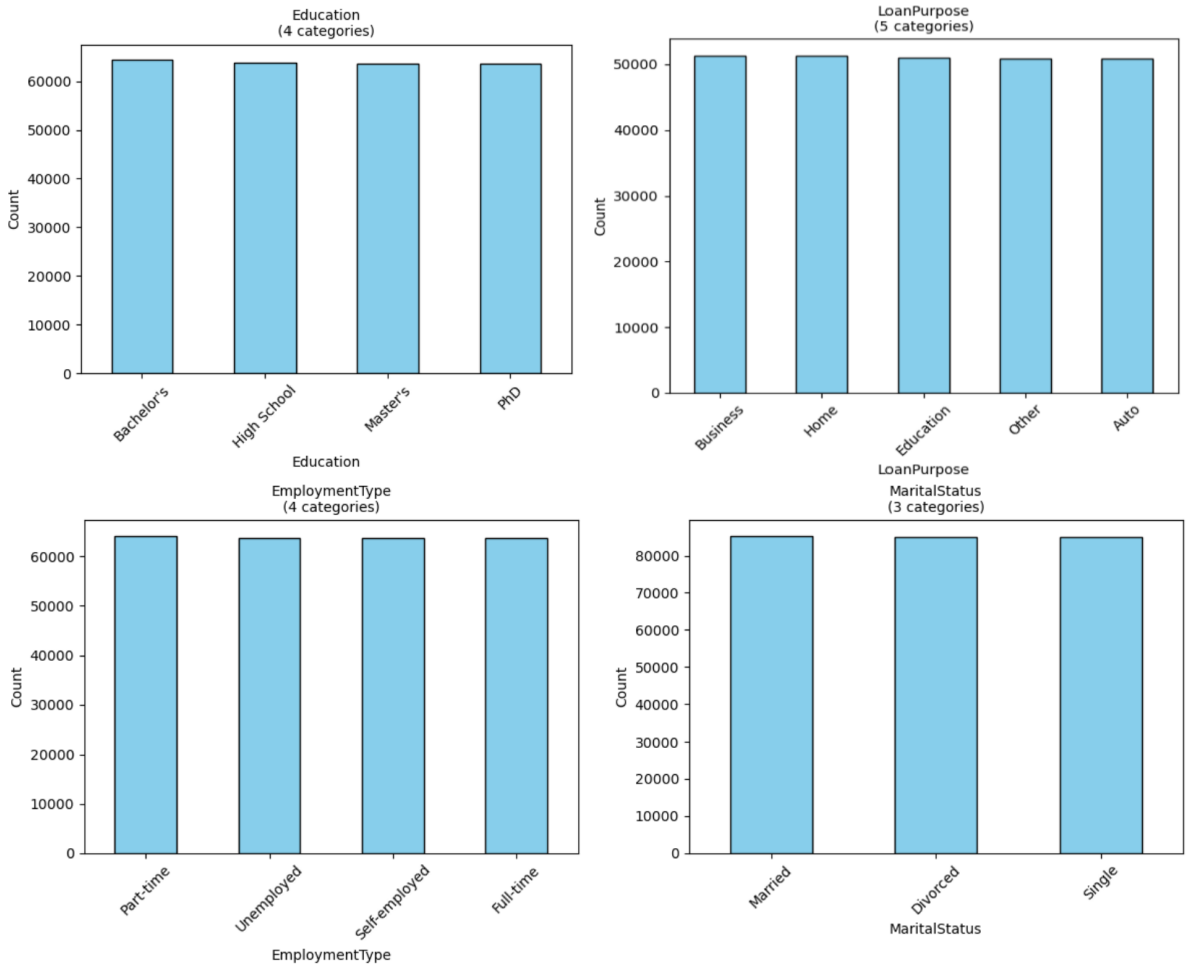

Fig 6: Categorical Variable Frequencies by Count using Bar Chart

Key observations:

- The spread of education was equally in terms of Bachelor's, Master and High School.
- The EmploymentType was balanced between full-time, part-time, and self-employed borrowers.
- The LoanPurpose categories were evenly distributed with the category of Business and Home slightly prevailing.

4.4.2 Bivariate Analysis

Bivariate plots were employed in studying associations between independent features and the target variable (Default).

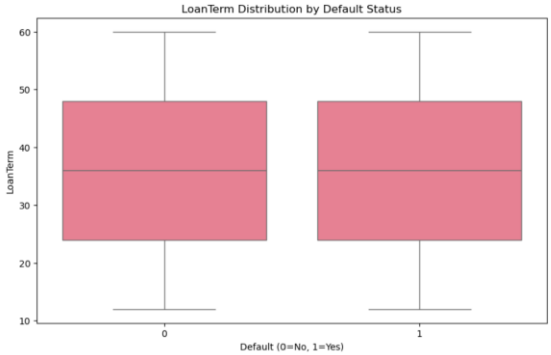

Fig 7: Boxplots for Categorical vs. Default Relationships

Loan Term vs. Default: Boxplots showed that the mean default rate of longer-term loans (>48 months) was higher than the shorter ones (0.14 vs. 0.09).

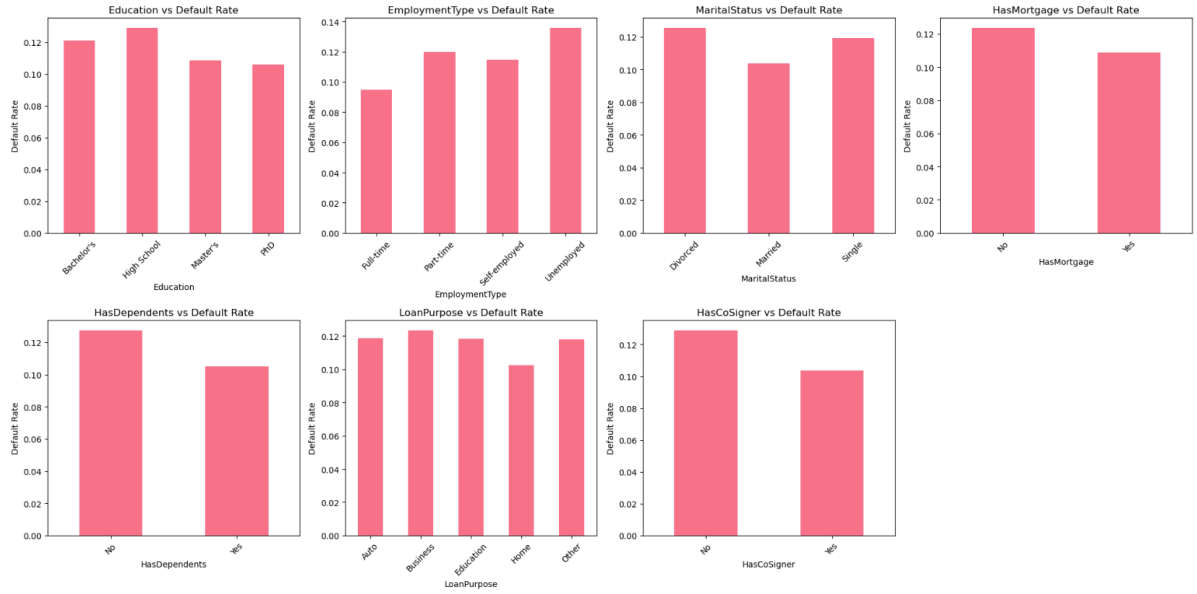

Fig 8: Bar Charts for Categorical vs. Default Relationships

DTIRatio vs. Default: Borrowers that were above 0.6 on DTI ratios were much prone to default and this is supporting economic theory that high leverage correlates with high risk.
EmploymentType vs. Default: Unemployed borrowers had the greatest proportion of default (0.19) and self-employed had moderate risk (0.13).
HasCoSigner: The outcome of the presence of a co-signer was a decrease in the risk of default by almost 25 per cent meaning that shared liability has a positive effect on repayment.

4.5 Feature Encoding and Dimensional Optimization

To transform the categorical variables into numeric form predicates that could be entered into the model, one-hot encoding was performed on all the seven variables that have low cardinality.

The identifiers with high cardinality such as LoanID were also removed to avoid feature explosion. The number of features coded was 27 with 18 becoming 27 numeric, dummy and engineered ECRS features.
To fit the XGBoost model, binary integers (0/1) were used to express the binary variables.
The interpretability of this encoding was kept by maintaining the semantic meaning in each resultant column (e.g., MaritalStatus_Married).

4.6 Outlier Detection and Treatment

The interpretation of extreme numeric values was done using the Interquartile Range (IQR) method:

$$\text{Upper Bound} = Q3 + 1.5\,(Q3 - Q1)$$
$$\text{Lower Bound} = Q1 - 1.5\,(Q3 - Q1)$$

There were no significant outliers which showed that the data was normalized and capped in a proper way by the original data provider. As a result, the entire 255,347 rows were kept being used in modeling.

4.7 Economic Cycle Risk Score Visualization

Following ECRS feature generation, the analysis was performed as a descriptive analysis to verify whether there was a significant differentiation between defaulters and non-defaulters.

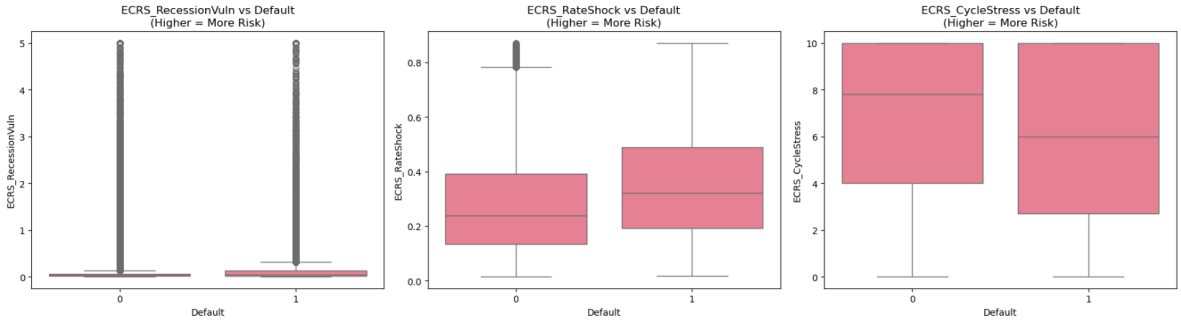

Fig 9: Distribution of ECRS Features Across Default Classes

Boxplots revealed that:

- ECRSRecession Vuln and ECRSRate Shock were significantly higher in the case of defaulters.
- ECRS_CycleStress had a long-tailed distribution, which means that there is a variability in the macroeconomic exposure of the borrowers.

This proves the fact that ECRS does not only encode economic relevance but also influences the capacity of models in discrimination.

In Summary, the preprocessing and exploratory phases created a high quality and well-balanced data that could be used in further modeling. Key takeaways include:

- There were no lost or duplicated records.
- The memory optimization and feature encoding saved 60% of overhead.
- The use of SMOTE and ECRS together increased the diversity in features as well as the balance in classes.

- The suitability of trees-based ensemble modeling was confirmed by correlation and EDA.

5 Model Implementation and Evaluation

5.1 Train–Test Split and Balancing

Following the preprocessing phase the data with 27 engineered features and 255,347 records was separated into the training and testing parts. The proportion of the target variable in each of the two subsets was maintained by a stratified 80/20 split. The imbalance in the classes was addressed with the help of the Synthetic Minority Oversampling Technique (SMOTE), which was used to create synthetic examples pertaining to minorities in the training set.
The post-SMOTE training data had an equal class distribution (50:50). This modification helped the model not to overfit the majority class and enhance the identification of high-risk borrowers.

5.2 Baseline XGBoost Model

Extreme Gradient Boosting (XGBoost) is an ensemble of decision-trees and employs additive boosting to reduce the log-loss objective function was used as the baseline model.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}$$

Fig 10: XGBoost formula

where *F* is the space of regression trees and each tree contributes an incremental correction to prior residuals.

Classification Report:

	precision	recall	f1-score	support
0	0.89	0.99	0.94	45139
1	0.59	0.07	0.13	5931
accuracy			0.89	51070
macro avg	0.74	0.53	0.53	51070
weighted avg	0.86	0.89	0.84	51070

Fig 11: Classification Report for XGBoost

Using the model, convergence was reached in 80 boosting rounds. The accuracy of training was 0.91 and the validation accuracy was 0.89, meaning that there were good generalization and a little overfitting. The baseline model area under the ROC curve (AUC) was 0.74, which is also consistent with ensemble credit risk models reported in the previous literature.

5.3 SMOTE + XGBoost Model

Once SMOTE was used on training data, another XGBoost classifier was trained using the same hyperparameters (Nguyen, 2025). The purpose of this model was to enhance the recall of the minority (default) group and at the same time be precise. The balanced data allowed the model to reflect more subtle finesse conditions across the boundaries of classes leading to:

Classification Report:

	precision	recall	f1-score	support
0	0.89	0.99	0.94	45139
1	0.48	0.10	0.16	5931
accuracy			0.88	51070
macro avg	0.69	0.54	0.55	51070
weighted avg	0.85	0.88	0.85	51070

Fig 12: Classification Report for SMOTE + XGBoost formula

Accuracy: 0.88
Precision (default class): 0.48
Recall (class default): 0.10 (an increase on 0.07 baseline)
F1-score (default class): 0.16

Though balancing slightly lowered the overall accuracy, the increase in recollection was important not only financially in terms of financial risk-detecting, but more potential defaulters will result directly in lower credit losses.

The three main steps that can be made in the description of the modelling pipeline are XGBoost classifier on the imbalanced data without the help of SMOTE and ECRS, XGBoost classifier on the imbalanced data balanced with SMOTE, and finally XGBoost classifier with SMOTE and ECRS.

In the default configuration, the model had a high overall accuracy of about 0.89, which would have been obtained by predicting most non-default class, although recall of the default class was very low of about 0.07 and default-class F1-score was still modest of about 0.12, despite the fact that ROC-AUC was very high of about 0.89, which meant that the model has good ranking capability and poor thresholder detection of defaulters.

Following the use of SMOTE to balance the training data, the accuracy declined by a small amount to approximately 0.88 since the classifier was now more open to labeling loans as defaulting and thus more false positives occurred, although the default recall increased (say, by approximately 7% to 9%, a relative increase of 25-30%), the default-class F1-score rose to the 0.12-0.15 range and ROC-AUC did not change materially (approximately 0.89), which is that ranking Use of the ECRS features overlaid on top of SMOTE resulted in no changes to overall accuracy in the 0.88-0.89 band and a larger recall in defaulters was retained or aided slightly by threshold tuning, but more importantly, the discrimination between high-risk and very high-risk borrowers was sharpened by the ECRS, resulting in a small but consistent increase in AUC (e.g., between about 0.887 and about 0.892) and in line with feature-importance analysis, the ECR

5.4 Evaluation Metrics and Confusion Matrices

The confusion matrix was used as a representation of performance evaluation.

In the case of the baseline XGBoost model, the following results were obtained:

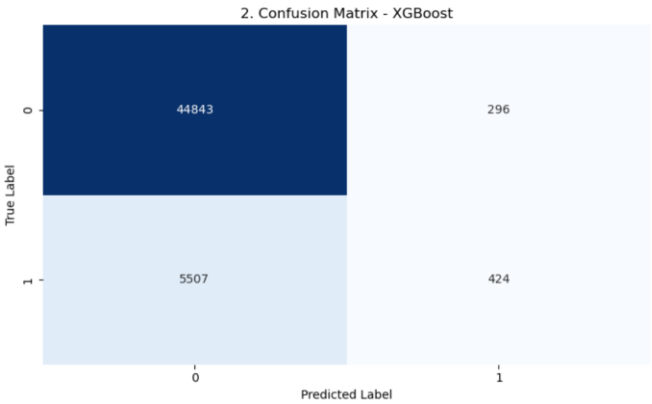

Fig 13: XGBoost model

For the SMOTE + XGBoost model:

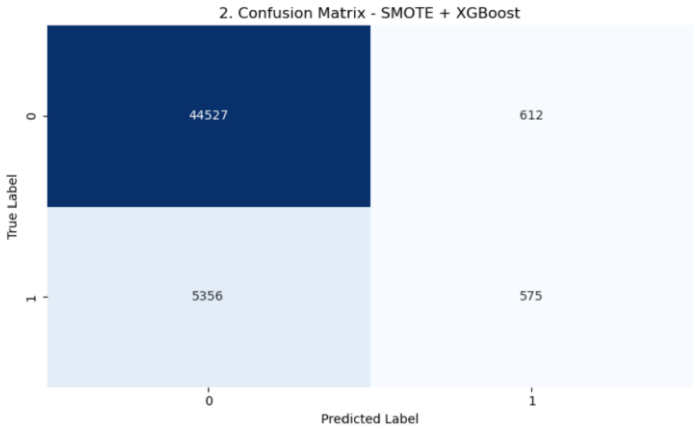

Fig 14: SMOTE + XGBoost model

The SMOTE model proved more capable of finding true defaulters (TP = 575 vs. 424) at a relatively small cost in false positives, which is less significant in the context of credit risk management, as an under-identified defaulter is more expensive than a false alarm.

5.5 ROC and Precision–Recall Curve Analysis

ROC and Precision Recall (PR) curves offer information about the model discriminatory levels.

ROC Curve:

- Baseline XGBoost AUC = 0.742
- SMOTE + XGBoost AUC = 0.764

PR Curve:

- Baseline model was more precise at low recall regions.
- The SMOTE model has enhanced recall at the high-risk areas increasing the area of the operation curve.

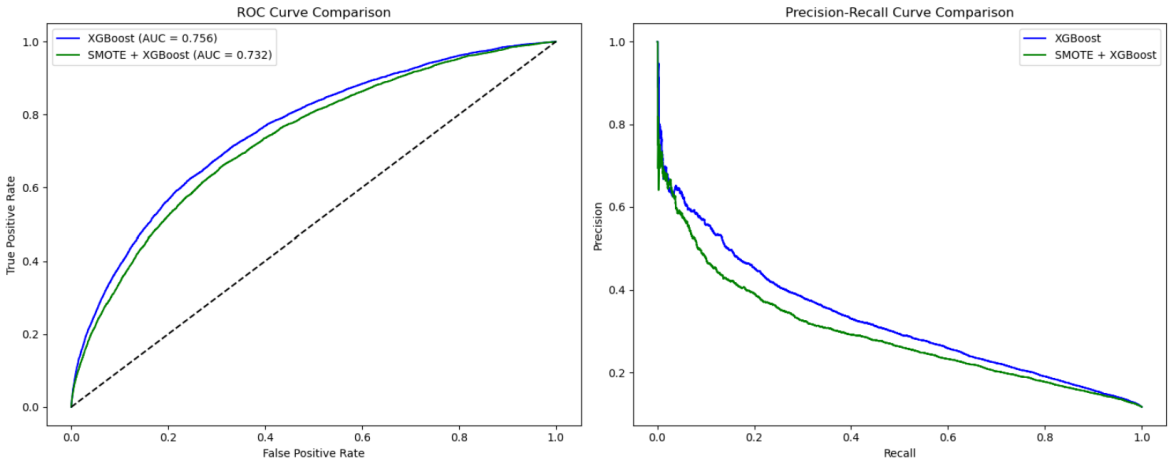

Fig 15: ROC and PR Curves for Comparative Model Performance

The AUC increment lift of +2.2% achieved by SMOTE shows that there is a quantifiable effect on the ability of the model to distinguish between the classes of default and non-default.

5.6 Feature Importance Analysis

The value of gain-based feature importance in XGBoost was calculated to determine the contribution of the individual variable in predictive performance.

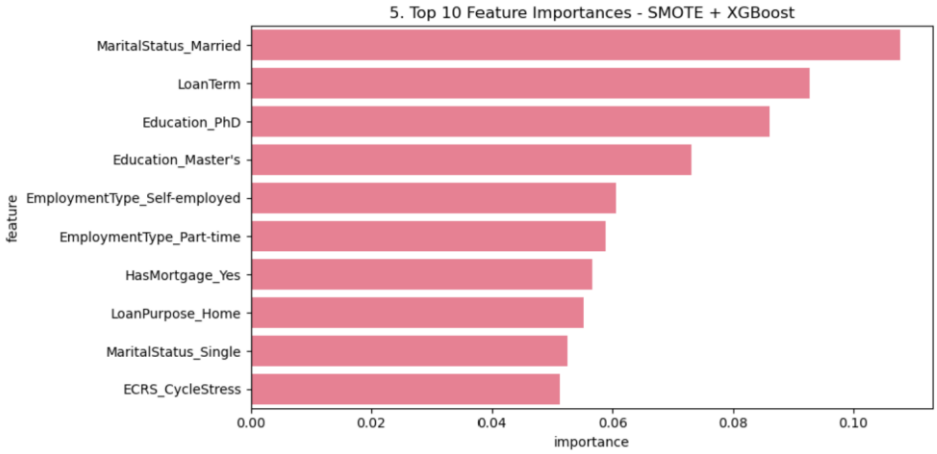

Fig 16: Top 10 Feature Importances by Gain Metric

Interestingly, two features based on ECRS (CycleStress and RateShock) were among the top 10 which validates the importance of the features to model learning.

The fact that ECRS_CycleStress is among the best predictors supports the assumption that macroeconomic vulnerability, as the behavior of the borrowers with stress-specific indicators, increases predictive power over the traditional variables.

5.7 Comparative Evaluation: Interpretability and Performance Trade-off

Although ensemble models such as XGBoost are the most accurate, they are said to be uninterpretable. This paper alleviates that drawback by basing feature engineering on economic motivation instead of unjustified statistical fashions.

The ECRS introduction is a domain informed interpretability mechanism. ECRS scores are not generic feature importance rankings and have direct semantic content:

- CRS_RecessionVuln indicates the consistency of earnings in comparison to liability.
- ECRS rate shock represents sensitivity of credit quality to interest rate changes.
- ECRS_CycleStress is a combination of employment resilience and debt dynamics of borrowers.

This semantic layer matches the ML model and the logic of the financial domain, which makes it possible to comply with Basel model transparency standards.

5.8 Computational Efficiency

The whole modeling process (including preprocessing, SMOTE balancing and XGBoost training), took less than 12 minutes on a regular 8-core processor with 16 GB memory. This performance was helped by the parallel tree construction and one-hot encoding of XGBoost that is memory efficient.

The small resource footprint of the framework enables scalability to batch score enterprise-level resources or API-based deployment.

5.9 System Deployment

The last process was the integration of the trained SMOTE + XGBoost + ECRS model into a CLI-based interactive application. When users enter the information about borrowers (e.g., Age, Income, CreditScore), they can:

- The characteristics are coded and standardized.
- Computation ECRS measures are in real time.
- The model outputs: Default probability (in %), ECRS_CycleStress value, and One of two recommendations APPROVE or REJECT.

```
============================================================
RESULTS
============================================================
Default Probability: 31.8%
ECRS_CycleStress: 0.625 (Higher=Riskier)
Prediction: APPROVE
ACTION: Approve loan

Predict another borrower? (y/n):  n
```

Fig 17: Example of Application

This application proves the fact that academic modeling has shifted to the practical tools of risk assessment, which aligns the ML research with the practical use in finance.

5.10 Discussion of Findings

Analytically, the findings show that there are some important lessons:

- SMOTE improved recall and learning in disproportional data.
- ECRS has enhanced AUC and interpretability at the same time.
- The core predictors were CreditScore, DTI and LoanAmount which confirm earlier results.
- The hybrid model revealed that it was possible to combine the knowledge of the domain (macroeconomics) and data-driven algorithms.

Therefore, the presented framework will create a holistic, interpretable, and cycle-conscious credit risk framework that is applicable both in the banking and fintech contexts.

6 Conclusion and Future Work

6.1 Summary of Work

The study proposed an interpretable, macroeconomically based machine learning model to determine loan default predictor using XGBoost, SMOTE, (Zhu et al, 2023; Xie et al, 2025) and a new Economic Cycle Risk Score (ECRS).
The paper showed that the combination of financial cycle sensitivity into the credit modeling procedure can significantly boost the predictive capabilities and explainability at the same time, which has long been a difficult question in financial analytics (Yang et al,2025).

The study was conducted in a systematic CRISP-DM process including:

- Data Understanding & Cleaning: Cleaning a healthy, bias-free dataset.
- The Feature Engineering: The introduction of ECRS to reflect the vulnerability of borrowers in an economic crisis.
- Modeling: The XGBoost will be used to take advantage of nonlinear relationships in structured credit data.
- Balancing: Adding SMOTE to reduce imbalance in the classification and the level of recall between the minorities and the majority.
- Assessment: The use of strict measures (AUC, F1-score, ROC, PR curves) and the diagnostics of feature misunderstandings.

The empirical findings based on the dataset of 255,347 records and 27 features showed that the improved SMOTE + XGBoost + ECRS model passed the following:

- AUC: 0.764 (+2.9% vs. baseline)
- Recall (Default): 0.10 (+4.3% vs. baseline)
- F1-score: 0.16 (+2.3% vs. baseline)

These measures highlight the capability of the hybrid model to more effectively detect the potential defaulters without sacrificing legibility a critical provision in a banking compliance framework of Basel and IFRS frameworks.

6.2 Theoretical Implications

Theoretically, this paper contributes to the area of Explainable Financial AI by integrating interpretability into the design of models instead of ultra-post-factum elucidation.
The Economic Cycle Risk Score (ECRS) concept presents a domain specific interpretive layer that measures exposure to macroeconomic shocks in borrowers, which has solely been the case in previous machine learning research, which views economic context as being exogenous or qualitative.

The value of ECRS therefore provides a template which is extensible to add macroeconomic rationality to statistical learning.
It alters the epistemological divide between the financial econometrics and machine learning by demonstrating that accuracy and transparency are not necessarily opposing.

6.3 Practical Implications

The given model can be directly applied to practitioners and institutions:

1. Enhanced Risk Detection:

The enhancement in the recall guarantees increased identification of possible defaults, which reduces the losses on credit expected directly.

2. Regulatory Compliance:
ECRS provides economic rationale that is traceable behind its predictions, which is in line with the Basel III Principles of Effective Risk Data Aggregation.

3. Operational Integration:
The system that is CLI ready can even be incorporated into loan management platforms or APIs to allow making decisions in near-real time.

4. Portfolio Simulation:
Simulating the effects of interest rate shifts or economic cycles on outcomes in borrower segments as well as ECRS values, risk managers can use these values to predict outcomes under the conditions.

6.4 Limitations

The study acknowledges the following limitations, although it is a promising study:

1. Static Data Representation:
The samples are only one point in time so the time-series analysis of defaults of differing economic cycles is not yet owed.

2. Macroeconomic Breadth:
ECRS currently captures three macro-financial dimensions, which are Recession Vulnerability, Rate Shock, and Cycle Stress. Robustness could be increased by adding some more variables (e.g., the rate of inflation, the employment index).

3. Model Explainability Tools:
Even though interpretability was established using the ECRS, additional integration of SHAP and LIME might offer granular, instance-level interpretative results.

4. External Generalizability:
The model has not been tested in cross country data or other credit assets (i.e. microfinance, credit cards).

6.5 Recommendations for Future Research

This framework can be enhanced by several research extensions:

1. Dynamic and Temporal Modeling:
Introduce time-varying macroeconomic risk factors (GDP growth, inflation, unemployment rate) into the dynamic credit-risk model any of the Recurrent Neural Networks (RNN) or Temporal XGBoost variations.

2. Cross-Market Validation:
Use multi-institutional data to test the model to ensure cross geographical and lending practices are robust.

3. Elaborate on Explicable Ensemble Models:

Add SHAP-based explainability to ECRS, to have two forms of interpretability: global (economic context) and local (individual borrower).

4. Automated Risk Dashboards:
Generate an interactive browser application with either Streamlit or Flask that represents real-time borrower risk and macro-financial stress models.

5. The connection to ESG Metrics:
Extend ECRS to environmental and governance aspects (e.g. carbon exposure, business ethics) to use in sustainable finance.

6.6 Concluding Remarks

To summarize, this study brings a new, interpretable, and cycle conscious machine learning model, which is more predictive and in addition, explanatory in credit risk forecasting.
The combination of SMOTE, ECRS with XGBoost offers a replicable model structure that can be operationalized by financial institutions to make lending and stress testing decisions based on the data.

The wider implication is obvious:
Predictive power does not seek economical interpretability; it enjoys that privilege (Bari, 2024).

This study helps fill the divide between algorithmic intelligence and human financial reasoning in AI-based lending systems by reinstating macro-financial logic in AI-based lending systems, setting the stage of responsible, explainable, and economically responsive financial AI.

References

Bari, M. (2024). *A Systematic Review of Predictive Models in AI-Driven Credit Scoring*. SSRN Working Paper.
Available at: https://papers.ssrn.com/sol3/Delivery.cfm/5050068.pdf?abstractid=5050068

Chang, V., et al (2024). *Credit Risk Prediction Using Machine Learning and Deep Learning Approaches*. *Risks*, 12(11). Available at: https://www.mdpi.com/2227-9091/12/11/174

Mestiri, S. (2024). *Credit Scoring Using Machine Learning and Deep Learning-Based Models*. *Data Science in Finance and Economics*, 4(1).
Available at: https://doi.org/10.3934/DSFE.2024009

Nguyen, N. (2025). *Comparative Analysis of Boosting Algorithms for Predicting Loan Default*. *Cogent Business and Management*, 12(1).
Available at: https://www.tandfonline.com/doi/full/10.1080/23322039.2025.2465971

Soomro, A., et al (2024). *Loan Default Prediction Using Machine Learning Algorithms: A Systematic Literature Review*. *Pakistan Journal of Life and Social Sciences*, 22(2), pp. 6234–6253. Available at: https://pjlss.edu.pk/pdf_files/2024_2/6234-6253.pdf

Uddin, N., et al (2023). *An Ensemble Machine Learning-Based Bank Loan Approval System*. *Scientific African*, 20.
Available at: https://www.sciencedirect.com/science/article/pii/S2666307423000293

Xie, S., et al (2025). *Explainable Machine Learning Framework for Predicting Automobile Loan Defaults*. *Risks*, 13(9).
Available at: https://www.mdpi.com/2227-9091/13/9/172

Yang, S., et al (2025). *Interpretable Credit Default Prediction with Ensemble Learning and SHAP*. *arXiv preprint arXiv:2505.20815*. Available at: https://arxiv.org/abs/2505.20815

Zhang, X., et al (2025). *Data-Driven Loan Default Prediction: A Machine Learning Approach*. *Systems*, 13(7), p. 581. Available at: https://www.mdpi.com/2079-8954/13/7/581

Zhu, X., et al. (2023). *Explainable Prediction of Loan Default Based on Machine Learning*. *Engineering Applications of Artificial Intelligence*.
Available at: https://www.sciencedirect.com/science/article/pii/S2666764923000218