**National Research University Higher School of Economics**

**Faculty of Computer Science**

**Programme “Master of Data Science”**

**MASTER’S THESIS**

**Comparative Analysis of Advanced Machine Learning Algorithms for Credit Default Prediction in Banking**

**Student:** Osipov Artyom Razmikovich

**Supervisor:** Masyutin Alexey Aleksandrovich, Head of AI Research Center, HSE

**Moscow, 2025**

# ABSTRACT

Credit scoring is a critical task in banking, traditionally addressed using classical machine learning models such as logistic regression and gradient boosting on tabular data. However, recent advancements propose leveraging sequential data directly through recurrent neural networks (RNNs). This study presents a comprehensive comparative analysis of advanced machine learning algorithms for credit default prediction in banking, focusing on the Alfa-Bank credit history dataset.

The study utilizes a large-scale dataset comprising over 26 million records of credit histories, containing diverse information about customers' past credit behavior, payment patterns, and default status. The research employs ROC AUC as the primary performance metric for model evaluation.

This research aims to identify the most effective model for credit default prediction by comparing logistic regression, gradient boosting, and RNNs. The methodology includes extensive data preprocessing, feature engineering, model training, and performance evaluation. The results will provide insights into the strengths and limitations of each approach, contributing to the development of more robust credit scoring systems in the banking sector.

Key findings and contributions include the implementation of a comprehensive machine learning pipeline, the evaluation of different models on large-scale credit data, and the assessment of their business impact. This work advances the understanding of machine learning applications in credit scoring and offers practical recommendations for banking institutions.

Keywords: credit scoring, credit default prediction, banking analytics, machine learning, deep learning, logistic regression, gradient boosting, neural networks.

# TABLE OF CONTENTS

[ABSTRACT 2](#_Toc199679441)

[TABLE OF CONTENTS 3](#_Toc199679442)

[INTRODUCTION 4](#_Toc199679443)

[CHAPTER 1. THEORETICAL FOUNDATIONS OF CREDIT DEFAULT PREDICTION 6](#_Toc199679444)

[1.1. Overview of Credit Scoring in Banking 6](#_Toc199679445)

[1.2. Machine Learning Methods in Credit Scoring 7](#_Toc199679446)

[1.3. Performance Metrics for Credit Default Models 8](#_Toc199679447)

[CHAPTER 2. METHODOLOGY AND DATA ANALYSIS 12](#_Toc199679448)

[2.1. Problem Statement and Research Objectives 12](#_Toc199679449)

[2.2. Dataset Description and Exploratory Analysis 13](#_Toc199679450)

[2.3. Model Selection and Evaluation Strategy 19](#_Toc199679451)

[2.4. Data Preprocessing and Feature Engineering 21](#_Toc199679452)

[CHAPTER 3. IMPLEMENTATION AND RESULTS 24](#_Toc199679453)

[3.1. Logistic Regression 24](#_Toc199679454)

[3.2. Gradient Boosting 29](#_Toc199679455)

[3.3. Recurrent Neural Network 34](#_Toc199679456)

[3.4. Performance Analysis and Model Comparison 41](#_Toc199679457)

[3.5. Business Impact Assessment 43](#_Toc199679458)

[CONCLUSION 46](#_Toc199679459)

[REFERENCES 48](#_Toc199679460)

[APPENDICES 49](#_Toc199679461)

[Appendix A: Dataset Description 49](#_Toc199679462)

[Appendix B: Implementation Code 51](#_Toc199679463)

# INTRODUCTION

Credit default prediction represents one of the most critical applications of machine learning in the financial services industry, directly impacting institutional profitability, regulatory compliance, and systemic financial stability. The ability to accurately assess the likelihood of borrower default has evolved from a cornerstone business requirement to a sophisticated analytical challenge that demands the application of advanced computational methodologies. As financial institutions navigate increasingly complex regulatory frameworks while competing in dynamic markets, the development of robust, interpretable, and scalable credit scoring systems has become paramount to sustainable banking operations.

The traditional landscape of credit scoring has been dominated by linear statistical models, particularly logistic regression, which have provided the foundation for risk assessment practices across the banking sector for decades. These approaches, while offering exceptional interpretability and regulatory compliance advantages, are inherently limited in their capacity to capture the complex, non-linear relationships that characterize modern credit risk patterns. The emergence of ensemble methods, particularly gradient boosting algorithms, has demonstrated substantial performance improvements in capturing these intricate patterns while maintaining reasonable interpretability for business applications.

More recently, the spreading of deep learning methodologies has introduced the possibility of directly processing sequential credit history data through RNNs, potentially preserving temporal dynamics that traditional aggregation approaches might obscure. This evolution represents a fundamental shift from static feature engineering to dynamic pattern recognition, offering the theoretical potential to uncover previously hidden predictive signals within the temporal structure of payment behaviors. However, the practical effectiveness of these advanced approaches in real-world credit scoring applications remains an open empirical question.

Current credit scoring systems must balance multiple competing objectives: predictive accuracy to minimize portfolio risk, computational efficiency to support high-volume operations, interpretability to satisfy regulatory requirements, and implementation feasibility within existing technological infrastructures. Understanding how advanced machine learning approaches perform across these dimensions is essential for informed decision-making regarding technology adoption and risk management strategy.

The research presented in this thesis addresses this critical knowledge gap through a comprehensive comparative analysis of three distinct machine learning paradigms applied to credit default prediction: traditional logistic regression as a baseline, gradient boosting as the current industry standard, and RNNs as an emerging frontier approach. This investigation utilizes a large-scale, real-world dataset from Alfa-Bank comprising over 26 million credit history records, providing unprecedented scale and authenticity for evaluating these methodological approaches. The research methodology incorporates rigorous experimental design principles, including stratified cross-validation for robust performance estimation, comprehensive evaluation metrics appropriate for imbalanced classification tasks, and careful attention to computational scalability considerations. The implementation spans the complete machine learning pipeline from data preprocessing through model evaluation, providing a comprehensive framework that can be adapted for similar applications in financial services.

The thesis is structured to provide both theoretical foundations and practical implementation guidance. Chapter 1 establishes the theoretical background of credit scoring methodologies and performance evaluation frameworks. Chapter 2 presents the research methodology, including dataset analysis, preprocessing strategies, and experimental design. Chapter 3 details the implementation and results of each modeling approach, culminating in comprehensive performance comparison and business impact assessment. The conclusion synthesizes key findings and provides recommendations for future research directions.

# CHAPTER 1. THEORETICAL FOUNDATIONS OF CREDIT DEFAULT PREDICTION

## 1.1. Overview of Credit Scoring in Banking

Credit scoring is a fundamental process in the banking sector, aimed at assessing the creditworthiness of potential and existing borrowers. Its primary objective is to predict the likelihood of a borrower defaulting on their financial obligations, which in turn helps mitigate credit risk. Historically, credit scoring systems have evolved from simple judgment-based assessments to sophisticated, data-driven approaches.

Early credit scoring methods relied heavily on subjective evaluations by loan officers. This approach, although personalized, was prone to biases and inconsistencies. Credit scoring has evolved significantly since its inception in the 1950’s, when it was first introduced by Fair, Isaac and Company (now FICO) [[[1]](#footnote-1)]. The FICO scoring system introduced numerical ranges (300-850) and standardized risk assessment criteria, setting an industry benchmark that continues to influence modern scoring systems. Linear regression, discriminant analysis, and logistic regression became standard tools for constructing credit scoring models due to their interpretability and effectiveness in handling tabular data.

In the context of regulatory frameworks, Basel III has significantly influenced credit scoring practices through its emphasis on risk-weighted assets and capital adequacy requirements. Specifically, the Internal Ratings-Based (IRB) approach under Basel III requires banks to develop sophisticated risk assessment models that can accurately estimate probability of default (PD), loss given default (LGD), and exposure at default (EAD). These regulatory requirements have driven innovations in model development while maintaining emphasis on interpretability and validation.

Credit scoring models underpin several critical banking operations. They assist in determining whether a borrower meets the bank’s lending criteria, setting interest rates based on the assessed risk of default, monitoring and managing credit risk across loan portfolios, and ensuring adherence to standards such as Basel III, which emphasizes risk assessment [[[2]](#footnote-2)].

Recent advances in machine learning have enabled the development of more complex and accurate models. These models leverage vast datasets, including behavioral and transactional data, to enhance predictive capabilities. Recent researches demonstrate that advanced machine learning techniques have significantly enhanced the accuracy and efficiency of credit risk assessment compared to traditional statistical approaches, though considerations of model interpretability remain important in their practical implementation [[[3]](#footnote-3), [[4]](#footnote-4), [[5]](#footnote-5)].

## 1.2. Machine Learning Methods in Credit Scoring

The evolution of credit scoring systems has progressed through several distinct phases, each characterized by increasing computational sophistication and analytical capability. The foundation of modern credit scoring was established in the 1950’s with the introduction of statistical models, primarily linear regression and discriminant analysis. These early approaches relied on structured tabular data, incorporating fundamental variables such as income, credit history, and demographic information. The progression of computing capabilities and expanded data accessibility subsequently enabled the implementation of more sophisticated models, including logistic regression and decision trees, which maintain their relevance in contemporary practice [[[6]](#footnote-6)].

The integration of machine learning methodologies has fundamentally transformed credit scoring practices. Financial institutions now utilize comprehensive datasets encompassing both transactional and behavioral information to develop models capable of identifying subtle patterns in credit risk assessment. A particularly significant advancement has been the transition from static tabular analysis to dynamic sequential data processing, enabling the examination of temporal patterns in credit card transactions and payment histories. This evolution has facilitated the implementation of advanced deep learning architectures for direct sequential data analysis.

Three primary machine learning methodologies dominate current credit scoring applications:

* **Logistic Regression.** This statistical approach remains fundamental in credit scoring due to its interpretability and effectiveness in binary classification. Its linear decision boundary provides clear feature importance weights and easily interpretable odds ratios. For example, in credit scoring, a coefficient of 0.5 for income directly translates to how much a unit increase in income affects the log-odds of default. However, its limitation in capturing non-linear relationships becomes apparent when dealing with complex interactions between variables such as payment history and credit utilization.
* **Gradient Boosting.** These ensemble methods have gained prominence for their capacity to process high-dimensional data and identify non-linear variable interactions. These methods excel in handling missing values and categorical variables, common challenges in credit data. Through iterative combination of weak learners, typically decision trees, they can capture complex patterns such as the relationship between payment behavior over time and default probability. The tree structure naturally handles interaction effects between features, such as how income level might modify the impact of debt-to-income ratio on default risk [[[7]](#footnote-7)].
* **Deep Learning Models.** RNNs and their variants, notably Long Short-Term Memory (LSTM) networks, represent the frontier of credit scoring technology. These architectures excel in processing sequential data, enabling sophisticated analysis of temporal patterns in customer behavior. For instance, LSTMs can directly model the evolution of customer payment patterns, capturing both long-term trends and recent changes in behavior that might signal impending default [[[8]](#footnote-8)].

Each machine learning method has its strengths and limitations. Logistic regression offers simplicity and explainability, making it suitable for regulatory compliance. Gradient boosting balances interpretability and performance, while deep learning models excel in handling complex and high-dimensional data. The choice of method depends on the specific requirements of the credit scoring task, including data availability, interpretability, and computational resources.

## 1.3. Performance Metrics for Credit Default Models

The evaluation of credit scoring models is a critical step in the development and deployment of reliable credit risk assessment systems. Performance metrics provide a framework for measuring the effectiveness of these models in predicting default outcomes, balancing predictive accuracy with practical business applicability. Selecting appropriate metrics ensures that models meet the dual objectives of risk mitigation and compliance with regulatory standards.

Classification metrics measure a model's ability to correctly assign borrowers into defaulters and non-defaulters. These metrics are focused on evaluating the binary classification performance of a model.

**Table 1.1** Classification metrics for credit default prediction models.

| **Metric** | **Description** |
| --- | --- |
| ROC AUC | Reflects the model's ability to rank borrowers correctly by the likelihood of default. It summarizes the trade-off between true positive rate (Recall) and false positive rate across all classification thresholds. A score closer to 1 indicates excellent discriminatory power, while 0.5 suggests no better performance than random guessing. |
| Precision | The proportion of correctly predicted defaulters (true positives) out of all borrowers predicted as defaulters. High precision minimizes the risk of falsely labeling non-defaulters as defaulters, which is crucial when false positives are costly (e.g., rejecting creditworthy borrowers). |
| Recall (Sensitivity) | The proportion of actual defaulters correctly identified by the model. High recall ensures that most defaulters are detected, which is critical for reducing credit risk. However, it may come at the expense of misclassifying some non-defaulters. |
| F1 Score | The harmonic mean of Precision and Recall, providing a single metric to balance both. It is particularly useful in scenarios with imbalanced datasets, where optimizing for just one of these metrics may not fully reflect model performance. |

Probability-based metrics assess the reliability and calibration of the predicted probabilities, ensuring that these estimates align with observed outcomes.

**Table 1.2** Probability-based metrics for credit default prediction models.

| **Metric** | **Description** |
| --- | --- |
| Log Loss | Measures the accuracy of predicted probabilities by penalizing predictions that deviate significantly from the observed outcomes. It assigns higher penalties to confident but incorrect predictions, making it ideal for models requiring precise probability estimates. Lower values indicate better calibration and accuracy. |
| Gini Coefficient | Derived from the ROC curve, it quantifies the discriminatory power of the model by comparing the cumulative distribution of defaulters against non-defaulters. A higher Gini coefficient reflects stronger differentiation between defaulters and non-defaulters. It is a popular choice in credit risk modeling due to its intuitive interpretation. |

Choosing the right performance metric involves trade-offs between accuracy, interpretability, and business relevance. For instance, while complex metrics like log loss provide granular insights, simpler metrics like ROC AUC may be preferred for their intuitive understanding. Additionally, regulatory frameworks such as the Basel Accords emphasize interpretability, often favoring metrics that align with transparency requirements [[[9]](#footnote-9)].

The selection and application of performance metrics are vital to the development of effective credit scoring models. By leveraging a combination of classification, probability-based, and business-oriented metrics, financial institutions can ensure that their models are not only predictive but also aligned with operational and regulatory goals. As models become increasingly complex with the adoption of machine learning techniques, the role of comprehensive and interpretable evaluation metrics becomes even more critical.

**Conclusion**

The theoretical foundations of credit default prediction demonstrate a significant evolution in both methodology and complexity. This chapter has explored the progression from traditional judgment-based assessments to sophisticated machine learning approaches, highlighting several key developments in the field.

Credit scoring has transformed from subjective evaluations to a standardized, data-driven process, with increasing emphasis on sophisticated analytical techniques. The transition from early statistical methods to advanced machine learning algorithms represents not just a technological advancement, but a fundamental shift in how financial institutions approach credit risk assessment.

The integration of advanced analytics in credit scoring presents both opportunities and challenges, balancing the need for sophisticated prediction capabilities with practical implementation constraints. The comprehensive framework of performance metrics established in this chapter provides a foundation for rigorous model evaluation, essential for both regulatory compliance and effective risk management in modern banking operations.

# CHAPTER 2. METHODOLOGY AND DATA ANALYSIS

## 2.1. Problem Statement and Research Objectives

Credit default prediction remains a central challenge in banking, with far-reaching implications for risk management and lending strategies. In this research, the focus is on enhancing credit scoring methodologies by leveraging detailed credit history data to predict loan default. The specific task addressed in this thesis is drawn from the Data Fest 2022 competition, where the objective is to develop predictive models for Alfa-Bank’s clients using exclusively their credit history records [[[10]](#footnote-10)].

The subject area of this study encompasses both the inherent complexities of credit risk assessment and the technical hurdles associated with processing large-scale, sequential data. The dataset under consideration is substantial, comprising over 26 million training records and approximately 4.7 million test records. Each record encapsulates a wealth of information, including loan amounts, client relationships to the loans, opening and closing dates, and detailed indicators of payment delinquencies. This granular and temporally structured data presents unique challenges: it demands robust data preprocessing pipelines to handle missing values and anomalies, efficient feature engineering to extract meaningful patterns from sequential information, and scalable computational strategies to manage the volume of data.

From a technical perspective, the study is required to build an analytical framework capable of accommodating various modeling approaches. Traditional techniques such as logistic regression and gradient boosting (using tools like LightGBM, XGBoost, and potentially CatBoost) rely on transforming sequential data into aggregated features that capture the essence of a client's credit behavior. In contrast, RNNs offer the promise of processing sequential data directly, thereby potentially preserving temporal dynamics that might be diluted or lost during aggregation. The technical requirements thus include not only the ability to preprocess and transform high-dimensional sequential data but also the need to implement, optimize, and compare models that have vastly different computational and architectural characteristics.

Functionally, the predictive models developed in this research must deliver accurate probability estimates of loan default for each client, facilitating informed decision-making within the banking context. The output is expected in a standardized format – a CSV file containing unique client identifiers and their corresponding predicted default probabilities. Given the competition’s evaluation criterion, the performance of these models is assessed using the ROC AUC metric, which is well-suited to measure the discriminatory power in binary classification problems such as credit default prediction. The functional requirements extend to ensuring that the models are not only statistically robust but also interpretable enough to meet the regulatory standards typically enforced in the banking sector.

The research objectives are twofold. First, this study aims to perform a comparative analysis of various machine learning methodologies – ranging from conventional approaches like logistic regression and gradient boosting to advanced deep learning architectures such as RNNs – in the context of credit default prediction. By directly comparing these models, the research seeks to identify which methodology offers the best balance between predictive performance, computational efficiency, and interpretability. Second, the study aims to investigate the advantages of processing sequential credit history data directly, as opposed to relying solely on aggregated features. This exploration is expected to shed light on the potential of deep learning approaches to capture temporal dependencies that may be critical for accurately predicting defaults.

In summary, the problem addressed in this thesis integrates the technical challenge of processing and modeling large-scale, sequential credit history data with the functional requirement of producing reliable, interpretable credit default predictions. The research is designed to advance our understanding of how advanced machine learning techniques can be tailored and optimized for the banking industry, ultimately contributing to more robust and effective credit scoring systems.

## 2.2. Dataset Description and Exploratory Analysis

The dataset used in this study is provided as part of the Data Fest 2022 competition and comprises several files of varying sizes. In essence, the material includes target files for both training and testing (with the training target containing the binary flag indicating credit default) and comprehensive credit history files. While a detailed description of all dataset files and fields is provided in Appendix A, this section offers a high-level overview of the data components and outlines the strategy for exploratory analysis.

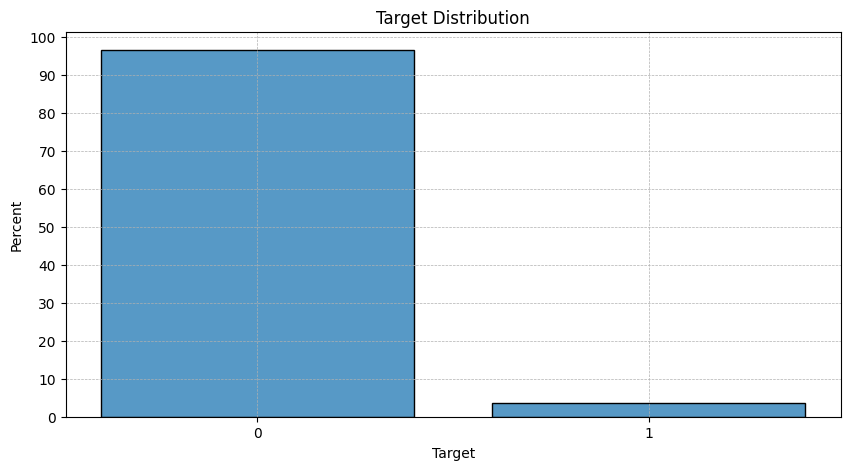
The data has been preprocessed and anonymized through binning and encoding techniques to protect client privacy while preserving predictive patterns. Each record represents a credit product with its associated history, including information about payment schedules, delays, credit limits, and various derived metrics.

Due to the large volume of data (approximately 26.16 million records in the training set), we implemented a custom chunked loading approach to efficiently process the Parquet files without exceeding memory limitations. This approach reads the data in manageable partitions while preserving client-level data integrity. For the exploratory analysis, we utilized a representative sample of 4 partitions (approximately 8.8 million records) from the training data. Initial analysis confirms that the dataset is complete, with no missing values detected across any of the fields. This high level of data integrity facilitates further preprocessing and modeling, ensuring that subsequent analyses are not hampered by data gaps.

The dataset contains a rich set of features that can be broadly categorized into several groups:

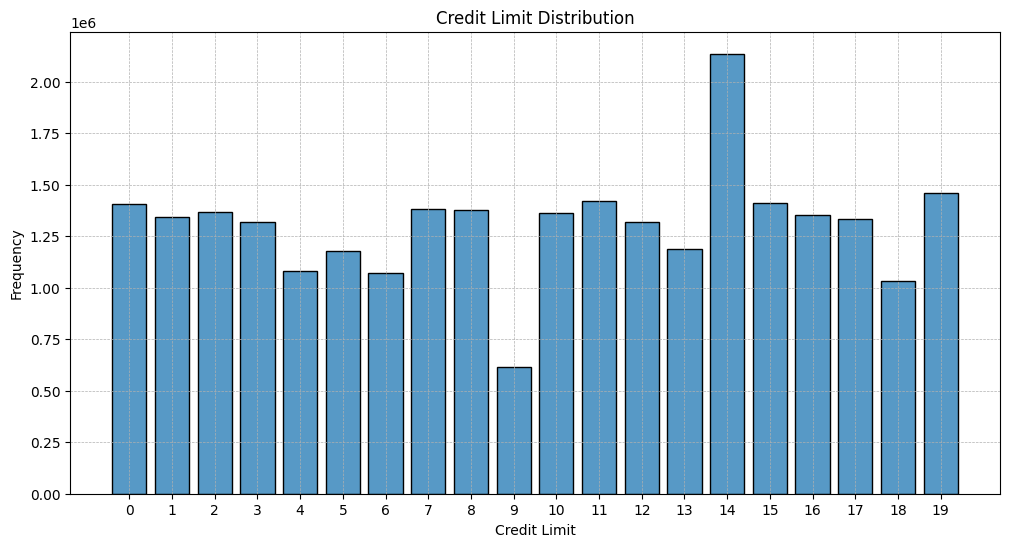
* **Temporal features:** Indicators of credit age and duration (e.g., pre\_since\_opened, pre\_till\_fclose)
* **Credit characteristics:** Details about the credit products (e.g., pre\_loans\_credit\_limit, pre\_loans\_outstanding)
* **Payment behavior:** Information about payment history and delinquencies (e.g., pre\_loansN variables indicating delays of various durations)
* **Utilization metrics:** Measures of how much of available credit is being used (e.g., pre\_util)
* **Encoded categorical variables:** Account holder types, credit types, and currencies (e.g., enc\_loans\_account\_holder\_type)
* **Payment status timeline:** Monthly payment statuses for recent months (e.g., enc\_paym\_0 through enc\_paym\_24)

Examination of the target variable reveals a significant class imbalance: approximately 3.55% of the training records are labeled as defaults, while the remaining 96.45% are non-defaults (see Figure 2.1). This imbalance presents a methodological challenge that will need to be addressed during modeling to prevent bias toward the majority class. In practice, this often involves employing techniques such as sampling methods, cost-sensitive learning, or evaluation metrics like ROC AUC that are robust to imbalanced data. The challenge of imbalanced data is a persistent theme in credit scoring research, with various techniques continually being explored and refined [[[11]](#footnote-11), [[12]](#footnote-12)].



**Figure 2.1** Distribution of the target variable (default status) in the training dataset.

The credit limit distribution (variable pre\_loans\_credit\_limit) shows a relatively uniform pattern across the binned values from 0 to 19, with slightly higher frequency for value 14 (see Figure 2.2). This relatively even distribution suggests that the binning process has effectively normalized what would likely have been a skewed distribution in the original data.

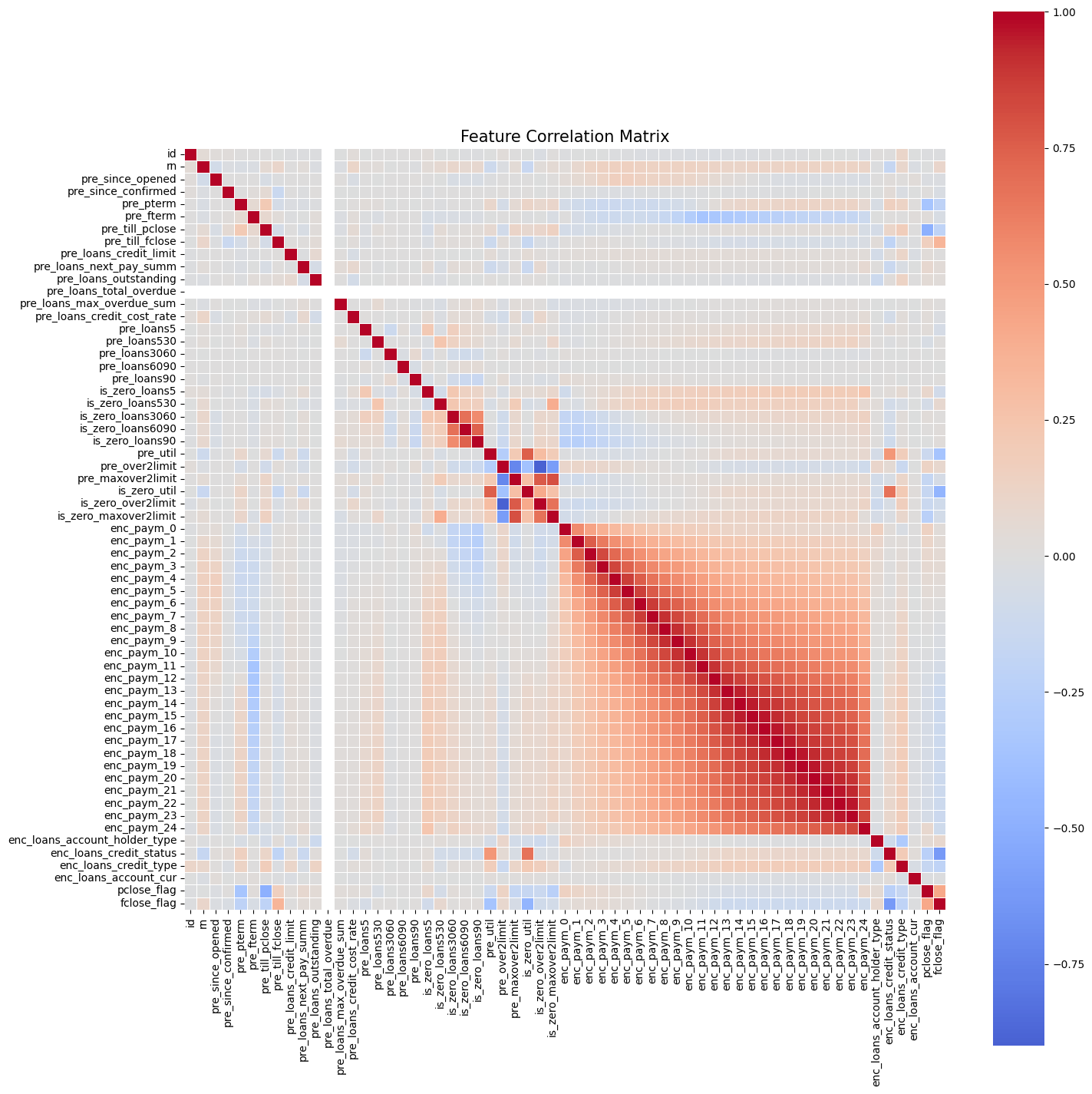


**Figure 2.2** Distribution of the credit limit variable (pre\_loans\_credit\_limit) in the training dataset.

To further understand the relationships between variables, a correlation matrix was computed for the features (see Figure 2.3). The correlation matrix reveals several interesting relationships between variables:

* Strong positive inter-correlations exist among the encoded payment status timeline variables (enc\_paym\_0 through enc\_paym\_24), suggesting that the payment behavior in one month is highly indicative of subsequent months.
* Current overdue debt variable (pre\_loans\_total\_overdue) shows NaN correlations with other features, which is expected due to its near-zero variance - almost all records have a value of 0, making the Pearson correlation mathematically undefined.
* Payment delay variables (pre\_loansN) exhibit moderate to strong positive correlations with each other, indicating that if a borrower has delays in one category, they are likely to have delays in others as well.
* Credit utilization metrics (pre\_util) show moderate correlation with delinquency indicators, suggesting that higher utilization may be associated with greater default risk.
* Flag variables (is\_zero\_\*) show expected negative correlations with their corresponding numeric metrics.

These correlation patterns support the theoretical foundation of credit scoring, where payment history, utilization, and account age are known to be predictive of default risk.



**Figure 2.3** Correlation matrix for features in the training dataset.

The feature importance analysis using Information Value (IV) reveals a surprising lack of strongly predictive individual features in the dataset (see Table 2.1). Information Value is a widely used metric in credit scoring that quantifies the predictive power of a variable with respect to a binary target [[[13]](#footnote-13)]. Mathematically, IV is calculated as:

where the summation is over all categories of the feature (or bins, in the case of continuous variables). IV values are interpreted as:

* < 0.02 – useless
* 0.02-0.1 – weak
* 0.1-0.3 – medium
* 0.3-0.5 – strong
* > 0.5 – suspicious

In our analysis, the most predictive feature, enc\_loans\_credit\_type, has an IV of only 0.0535, which falls into the "weak" category. Similarly, payment status variables (enc\_paym\_\*) show relatively higher importance than other features but still demonstrate only weak predictive power individually. This contrasts with typical credit scoring scenarios where features like payment history often show medium to strong predictive power.

**Table 2.1** Feature importance analysis using Information Value (IV).

| **Feature** | **Information Value** | **Correlation** | **Importance** |
| --- | --- | --- | --- |
| enc\_loans\_credit\_type | 0.0535 | 0.0253 | Weak |
| enc\_paym\_2 | 0.0468 | 0.035 | Weak |
| enc\_paym\_3 | 0.0415 | 0.0317 | Weak |
| enc\_paym\_1 | 0.0411 | 0.0335 | Weak |
| enc\_paym\_4 | 0.0361 | 0.0285 | Weak |
| enc\_paym\_5 | 0.0269 | 0.0242 | Weak |
| enc\_paym\_0 | 0.0258 | 0.0261 | Weak |
| enc\_paym\_6 | 0.0237 | 0.0219 | Weak |
| enc\_paym\_7 | 0.0237 | 0.0206 | Weak |
| is\_zero\_loans530 | 0.0204 | -0.0261 | Weak |
| pre\_loans\_credit\_cost\_rate | 0.02 | 0.0019 | Useless |
| pre\_loans\_credit\_limit | 0.0196 | 0.0051 | Useless |
| enc\_paym\_8 | 0.0195 | 0.0182 | Useless |
| is\_zero\_loans3060 | 0.0191 | -0.0273 | Useless |
| pre\_pterm | 0.0173 | -0.0065 | Useless |
| enc\_paym\_13 | 0.017 | 0.0165 | Useless |
| enc\_paym\_9 | 0.017 | 0.016 | Useless |
| is\_zero\_loans6090 | 0.0168 | -0.0265 | Useless |
| enc\_paym\_11 | 0.0164 | 0.0158 | Useless |
| is\_zero\_loans90 | 0.0158 | -0.0256 | Useless |

The lack of strongly predictive individual features presents both a challenge and an opportunity. Traditional scorecard approaches that rely on a small number of highly predictive features may struggle with this dataset. However, more complex models like gradient boosting and deep neural networks excel at leveraging weak signals across many features and capturing non-linear interactions, as demonstrated in recent credit scoring applications [[[14]](#footnote-14)].

The exploratory analysis provides several key insights that will inform the modeling approach. First, the cleanness of the data with no missing values simplifies preprocessing requirements. Second, the feature importance analysis identifies the most predictive variables, which will help prioritize features across all models. Third, the temporal sequence of payment statuses offers an opportunity for sequence-based modeling, particularly for the RNN approach. Finally, the class imbalance observed will guide the selection of appropriate evaluation metrics and sampling techniques.

## 2.3. Model Selection and Evaluation Strategy

To address the research objectives and evaluate the performance of different machine learning methodologies for credit default prediction, three distinct modeling approaches will be implemented: Logistic Regression, Gradient Boosting, and RNNs. Each approach offers unique advantages and challenges, making them suitable for different aspects of the credit scoring problem.

Logistic regression serves as the foundational baseline model, reflecting the traditional approach to credit scoring that remains prevalent in regulated banking environments. Despite its simplicity, logistic regression offers several key advantages in this domain. Its transparent decision-making process provides clear coefficient weights that directly relate to feature importance, facilitating interpretation and explanation of risk factors. This interpretability is particularly valuable in financial services, where regulatory requirements often necessitate model transparency. Additionally, logistic regression's computational efficiency makes it suitable for large-scale deployment, a significant consideration given the volume of credit applications processed by financial institutions.

Gradient boosting represents the current industry standard in advanced credit scoring, offering a compromise between predictive performance and interpretability. This ensemble technique has demonstrated superior performance on tabular data compared to linear models in numerous applications, including recent credit scoring competitions. Gradient boosting algorithms excel at capturing non-linear relationships and feature interactions without requiring extensive feature engineering. Their ability to handle class imbalance is particularly relevant given the significant imbalance (3.55% default rate) observed in the dataset. While less interpretable than logistic regression, gradient boosting models still provide meaningful feature importance metrics that can inform business decisions.

RNNs represent the frontier of credit scoring methodology, particularly for their ability to process sequential data directly. Traditional credit scoring approaches typically aggregate temporal data into static features, potentially losing valuable information about payment patterns over time. RNNs, in contrast, can learn directly from the sequential structure of payment histories. This approach will be included to investigate whether temporal patterns in payment history can be leveraged more effectively than through aggregated features, and whether deep learning architectures can extract more predictive signals from features that individually show weak predictive power, as indicated by the low Information Value metrics observed in the exploratory analysis [[[15]](#footnote-15)].

The evaluation framework will employ multiple complementary metrics to assess model performance, with emphasis on those appropriate for imbalanced classification tasks. ROC AUC will serve as the primary evaluation metric for several critical reasons. It measures a model's ability to rank-order customers by default risk, which is fundamental to credit scoring applications where prioritization of high-risk cases is essential. Additionally, ROC AUC is invariant to class imbalance, making it particularly suitable for the dataset's 3.55% default rate.

To provide a comprehensive assessment of model performance, the evaluation will also include:

* Recall AUC, which offers greater sensitivity to minority class performance than ROC AUC, providing additional insight into default detection capability.capability.
* F1 Score, which, as the harmonic mean of precision and recall, offers a balanced assessment of model performance at specific decision thresholds.
* Optimal threshold analysis will be conducted by evaluating F1 scores across different classification thresholds to identify the best decision boundary for each model.
* Prediction distribution analysis will visualize the separation between default and non-default probability distributions to assess model calibration and discriminatory power.

The validation strategy will be designed to ensure robust model evaluation while accommodating the computational constraints of processing 26+ million records. Given the dataset size and memory limitations, different validation approaches will be employed for each modeling technique. Logistic regression and gradient boosting will utilize stratified k-fold cross-validation, ensuring that each fold maintains the class distribution of the target variable. This approach will provide a reliable estimate of model performance while allowing for efficient computation on large datasets. Recurrent neural networks will employ a single train-validation split due to the sequential nature of the data, which requires maintaining temporal order. The training set will consist of the first 90% of the data, while the validation set will include the remaining 10%. This split will ensure that the model learns from past sequences without leaking future information, which is critical for time-series data.

The final evaluation will be conducted on the competition test dataset, with all models generating probability scores for comprehensive performance comparison. This approach will ensure that model performance is assessed on truly unseen data, providing reliable estimates of generalization capabilities while working within the constraints of the large-scale dataset and available computational resources.

## 2.4. Data Preprocessing and Feature Engineering

The data preprocessing and feature engineering phase will prepare the dataset for optimal model performance across the three selected modeling approaches. While the dataset is remarkably clean with no missing values and has already undergone preliminary binning and encoding, specific transformations will be implemented to adapt the data structure to each modeling approach's requirements.

The significant class imbalance (3.55% default rate) will be addressed using tailored techniques for each modeling approach. For logistic regression, class weights will be calculated and applied inversely proportional to class frequencies using scikit-learn's class\_weight parameter to rebalance the loss function and increase the impact of minority class samples. As for gradient boosting, the LightGBM scale\_pos\_weight parameter will be set to the inverse ratio of non-default to default cases, effectively rebalancing the gradient updates without explicit resampling. And for RNNs, TensorFlow's built-in class weighting capabilities will be utilized during model compilation, complemented by balanced mini-batch sampling strategies to ensure adequate minority class representation during training.

Each modeling approach will require distinct feature engineering strategies to optimize performance. For the logistic regression and gradient boosting models, the CountAggregator class will be developed to transform sequential credit history data into tabular format through count-based aggregation. This process involves one-hot encoding of all categorical features across the entire feature set, grouping by client ID and summing binary indicators to create frequency-based features, standardization of aggregated features for logistic regression to ensure numerical stability, and creation of interaction features for gradient boosting, specifically between overlimit metrics and payment status variables. This transformation will create a fixed-length feature vector for each client, capturing the frequency and patterns of various credit behaviors while preserving the predictive information from the sequential data. For the RNN model, a specialized preprocessing pipeline will be implemented to preserve temporal structure. All features will be shifted by adding 1 to create clear distinction between actual values and padding zeros. Sequential transformation using a transform\_credits\_to\_sequences function will maintain temporal order. Uniform padding to maximum sequence length (58 time steps) with Keras masking layers will handle variable-length sequences. Feature organization will maintain all original variables as sequential inputs rather than separating static and temporal features.

**Conclusion**

The methodology and data analysis framework established in this chapter provides a comprehensive foundation for the comparative evaluation of advanced machine learning algorithms in credit default prediction. The problem formulation has clearly defined the research objectives within the context of real-world banking applications, establishing both the technical requirements for processing 26+ million credit records and the functional requirements for producing reliable, interpretable credit default predictions. The comprehensive dataset analysis reveals both the opportunities and challenges presented by this large-scale credit history data, including the significant class imbalance and the relatively weak individual predictive signals that will require sophisticated modeling approaches to leverage effectively.

The exploratory data analysis has provided crucial insights that will inform the modeling strategies across all three approaches. The identification of overlimit metrics and payment status variables as key predictors, combined with the correlation analysis revealing complex inter-relationships among features, establishes the theoretical foundation for the feature engineering strategies that will be implemented. The Information Value analysis, while revealing the weakness of individual predictors, highlights the potential for ensemble methods and interaction effects to extract meaningful patterns from seemingly limited signals.

The preprocessing and feature engineering framework addresses the unique requirements of each modeling approach while maintaining consistency in evaluation methodology. The differentiated strategies for handling class imbalance, feature transformation, and validation across logistic regression, gradient boosting, and RNN approaches ensure that each method will be evaluated under optimal conditions for its architectural strengths.

# CHAPTER 3. IMPLEMENTATION AND RESULTS

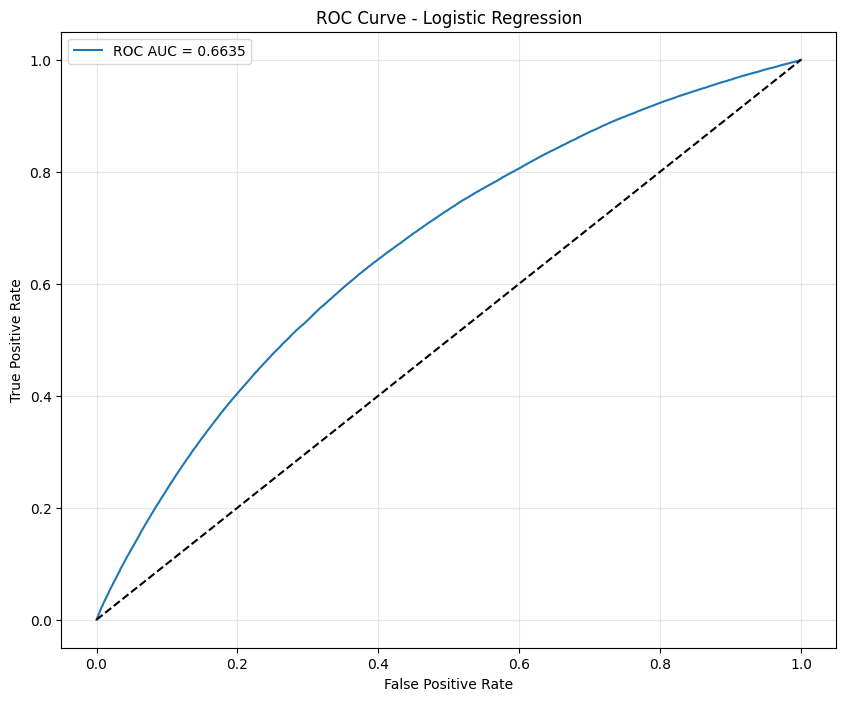
## 3.1. Logistic Regression

The implementation of traditional logistic regression presents considerable computational challenges when applied to large-scale financial datasets. Significant memory constraints and convergence difficulties were encountered during attempts to deploy the standard scikit-learn LogisticRegression estimator on the extensive dataset comprising 26 million records. To address these computational limitations while preserving the theoretical integrity of the logistic regression framework, a scalable implementation was developed incorporating several methodological adaptations:

* **Dimensionality reduction through L1 regularization:** An initial SGDClassifier with L1 penalty was employed to perform embedded feature selection. This approach leverages the sparsity-inducing properties of L1 regularization to identify the most predictive subset of features while discarding redundant or noise variables.
* **Stochastic optimization with logistic loss function:** The standard LogisticRegression implementation was replaced with SGDClassifier configured with 'log\_loss' objective function. This substitution maintains the mathematical formulation of logistic regression while enabling efficient parameter updates through stochastic gradient descent, substantially reducing memory requirements and computation time.
* **Mini-batch gradient descent:** To further enhance computational efficiency, mini-batch training with batch sizes of 100,000 samples was implemented. This approach balances the efficiency advantages of stochastic updates with the stability benefits of batch processing, enabling effective training on large-scale datasets with limited computational resources.
* **Adaptive early stopping mechanism:** A custom validation-based early stopping protocol was incorporated, monitoring the ROC AUC score on a validation subset after each epoch and terminating training if no improvement was observed for five consecutive epochs. This approach mitigates overfitting risk while optimizing training duration.
* **Class-weight adjustment for imbalanced learning:** To address the substantial class imbalance (3.55% default rate), class weights inversely proportional to class frequencies were applied, effectively increasing the impact of minority class samples on the decision boundary. This calibration is critical for credit scoring applications where default prediction sensitivity is paramount.

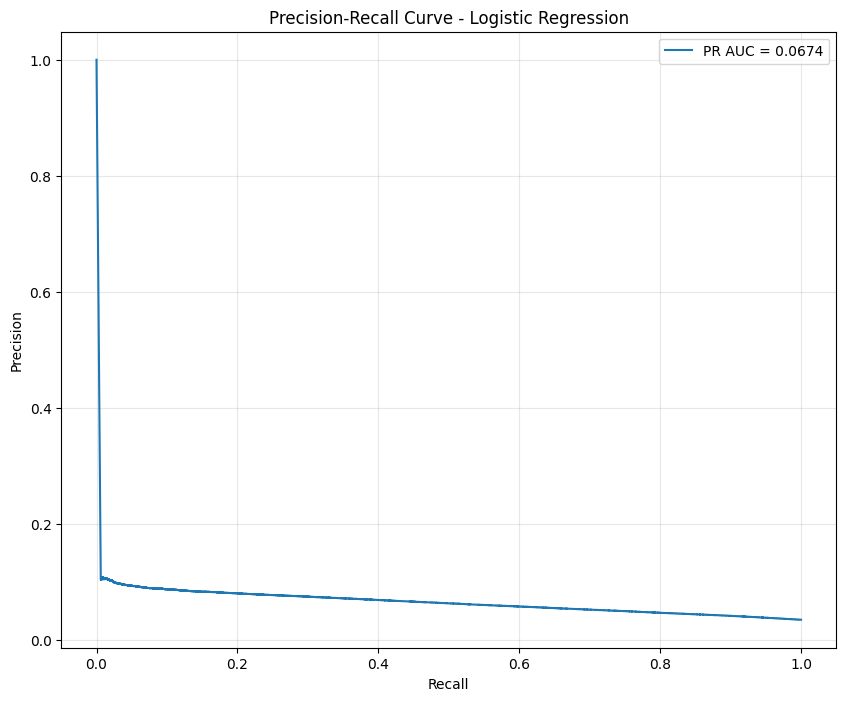
The cross-validation performance metrics for the logistic regression model demonstrate reasonable predictive capacity within the constraints of this challenging problem domain:

**ROC AUC.** The model achieved a mean 5-fold cross-validation ROC AUC score of 0.664 (individual fold scores: 0.654, 0.637, 0.671, 0.675, 0.684). This indicates substantial discrimination ability compared to random classification (ROC AUC = 0.5), affirming the model's capacity to rank-order customers by default risk. Figure 3.1 illustrates the ROC curve for the logistic regression model, with the area under the curve visually demonstrating the model's ability to distinguish between defaulters and non-defaulters across various classification thresholds.



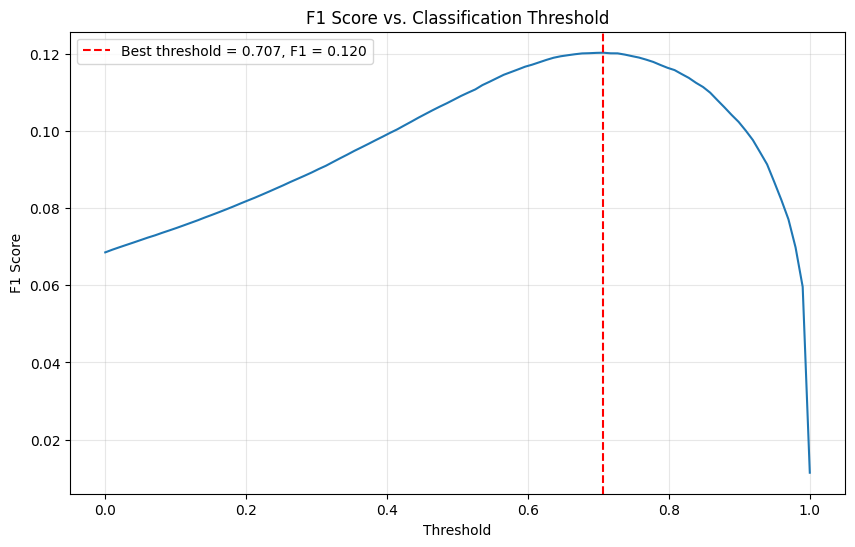
**Figure 3.1** ROC curve for the logistic regression model.

**Precision-Recall characteristics.** The model attained a Precision-Recall AUC of 0.067. While this value appears modest in absolute terms, it represents meaningful predictive improvement over the baseline PR AUC of 0.0355 (equal to the default rate) for a random classifier in this highly imbalanced context. Figure 3.2 displays the Precision-Recall curve, which is particularly informative for imbalanced classification tasks, highlighting the trade-off between precision and recall as the decision threshold changes.



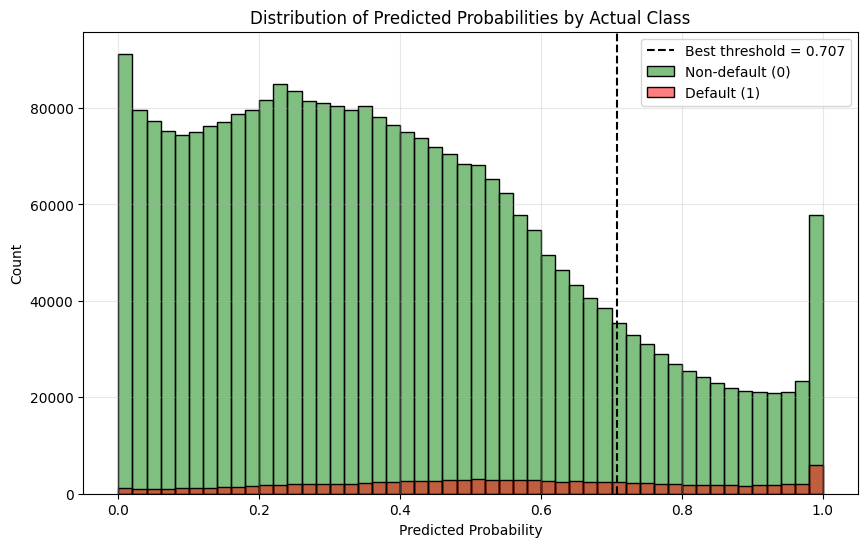
**Figure 3.2** Precision-Recall curve for the logistic regression model.

**Classification threshold optimization.** The optimal decision threshold was identified at 0.707, yielding an F1 score of 0.120. Figure 3.3**Figure 3.*2*** shows the F1 score progression across different threshold values, demonstrating how the optimal threshold balances false positives and false negatives for this imbalanced dataset.



**Figure 3.3** F1 score progression across classification thresholds for the logistic regression model.

**Prediction distribution analysis.** Figure 3.4 presents the distribution of predicted probabilities segregated by actual class labels, revealing the model's ability to shift the probability mass of default cases toward higher values compared to non-default cases. This visualization demonstrates the discriminatory power of the model while highlighting the challenging overlap between classes inherent in credit default prediction.



**Figure 3.4** Distribution of predicted probabilities by actual class label for the logistic regression model.

Analysis of feature importance coefficients (see Figure 3.5) reveals that metrics related to credit limit violations are particularly predictive of default outcomes. Specifically, overlimit metrics (pre\_over2limit\_\*) and maximum overlimit indicators (pre\_maxover2limit\_\*) demonstrate the strongest associations with default probability. The credit type encoding also emerged as a significant predictor, aligning with domain knowledge that different credit products carry varying inherent risk profiles.

A graph with blue and white text

AI-generated content may be incorrect.

**Figure 3.5** Top 20 features by importance (coefficient magnitude) in the logistic regression model.

This initial model establishes a robust baseline while highlighting the predictive value of aggregated overlimit features in credit default prediction. The performance achieved, while modest in absolute terms, represents a meaningful improvement over random classification and aligns with expectations for a linear model applied to this complex domain. The subsequent sections will investigate whether gradient boosting and recurrent neural network approaches can capture additional patterns and improve predictive performance.

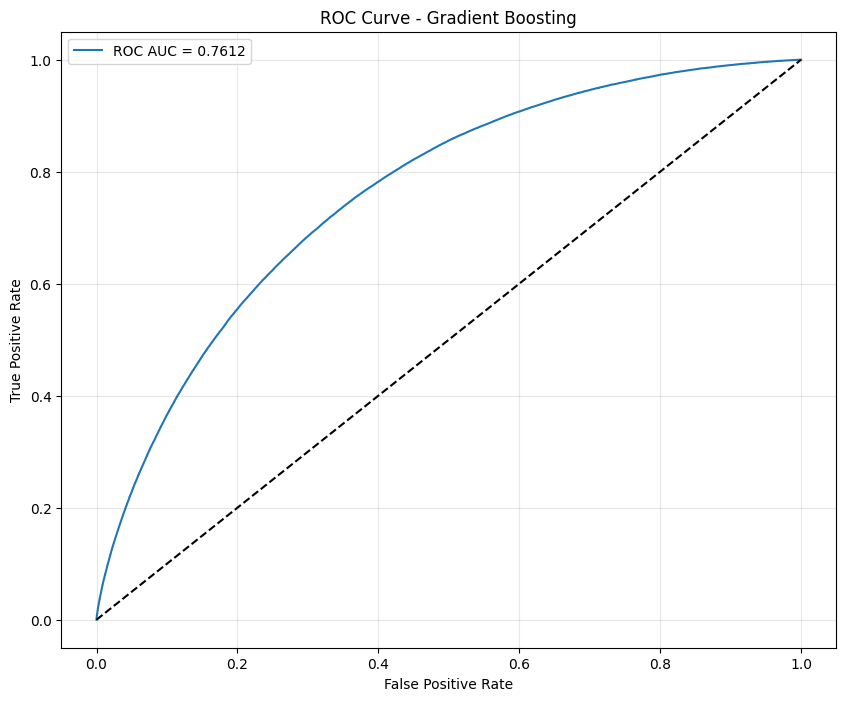
## 3.2. Gradient Boosting

The gradient boosting methodology was implemented through the LightGBM framework, selected for its exceptional computational efficiency and memory-optimization capabilities when processing large-scale financial datasets. This implementation addresses the inherent challenges of credit default prediction through an ensemble approach that iteratively builds decision trees focused on misclassified cases from previous iterations. The gradient boosting architecture incorporated several methodological enhancements to address both the computational complexity of processing 26 million records and the statistical challenges inherent in imbalanced, high-dimensional credit risk modeling:

* **Feature interaction engineering:** To capture non-linear relationships between key risk indicators, interaction features were created between overlimit metrics and payment status variables, enabling the model to detect complex patterns of financial distress not visible through single variables alone.
* **Regularization optimization:** A balanced regularization approach combining L1 (alpha=0.5) and L2 (lambda=1.5) penalties was implemented to mitigate overfitting while preserving the model's capacity to identify subtle default patterns, particularly important given the relative weakness of individual predictors identified in exploratory analysis.
* **Class imbalance management:** The severe class imbalance (3.55% default rate) was addressed through the scale\_pos\_weight parameter, calibrated to the inverse ratio of non-default to default cases, effectively rebalancing the loss function without requiring explicit resampling.
* **Hyperparameter tuning:** Optimization of critical parameters including learning rate (0.03), tree depth (7), and leaf constraints (96 leaves, minimum 20 samples) established an effective balance between model complexity and generalization performance, informed by domain knowledge of credit risk patterns.
* **Stratified cross-validation:** Implementation of stratified 5-fold cross-validation preserved the default rate distribution across training and validation subsets, ensuring reliable performance estimation despite the pronounced class imbalance.

The cross-validation performance metrics demonstrate substantial predictive capability, with marked improvement over the logistic regression baseline:

**ROC AUC.** The model achieved a mean 5-fold cross-validation ROC AUC score of 0.761 (individual fold scores: 0.765, 0.762, 0.758, 0.759, 0.762), indicating strong discriminative power. This represents a 14.6% relative improvement over the logistic regression baseline (0.664), demonstrating gradient boosting's enhanced capacity to differentiate between defaulting and non-defaulting customers. Figure 3.6 illustrates the ROC curve for the gradient boosting model, with the substantially larger area under the curve visually demonstrating the improved discriminatory power.



**Figure 3.6** ROC curve for the gradient boosting model.

**Precision-Recall characteristics.** The Precision-Recall AUC of 0.112 represents a 66.5% improvement over the logistic regression model (0.067), reflecting superior capability in identifying true default cases. While this value remains constrained by the extreme class imbalance, it significantly outperforms the baseline PR AUC of 0.0355 for a random classifier. Figure 3.7 presents the Precision-Recall curve, highlighting the model's enhanced performance in this challenging prediction context.

A graph of a graph

AI-generated content may be incorrect.

**Figure 3.7** Precision-Recall curve for the gradient boosting model.

**Classification threshold optimization.** The optimal decision threshold was identified at 0.707, yielding an F1 score of 0.181. This represents a 50.8% improvement over the logistic regression model's F1 score (0.120), indicating more balanced precision and recall performance. Figure 3.8 displays the F1 score progression across threshold values, revealing a more pronounced peak compared to the logistic regression model.

A graph of a curve

AI-generated content may be incorrect.

**Figure 3.8** F1 score progression across classification thresholds for the gradient boosting model.

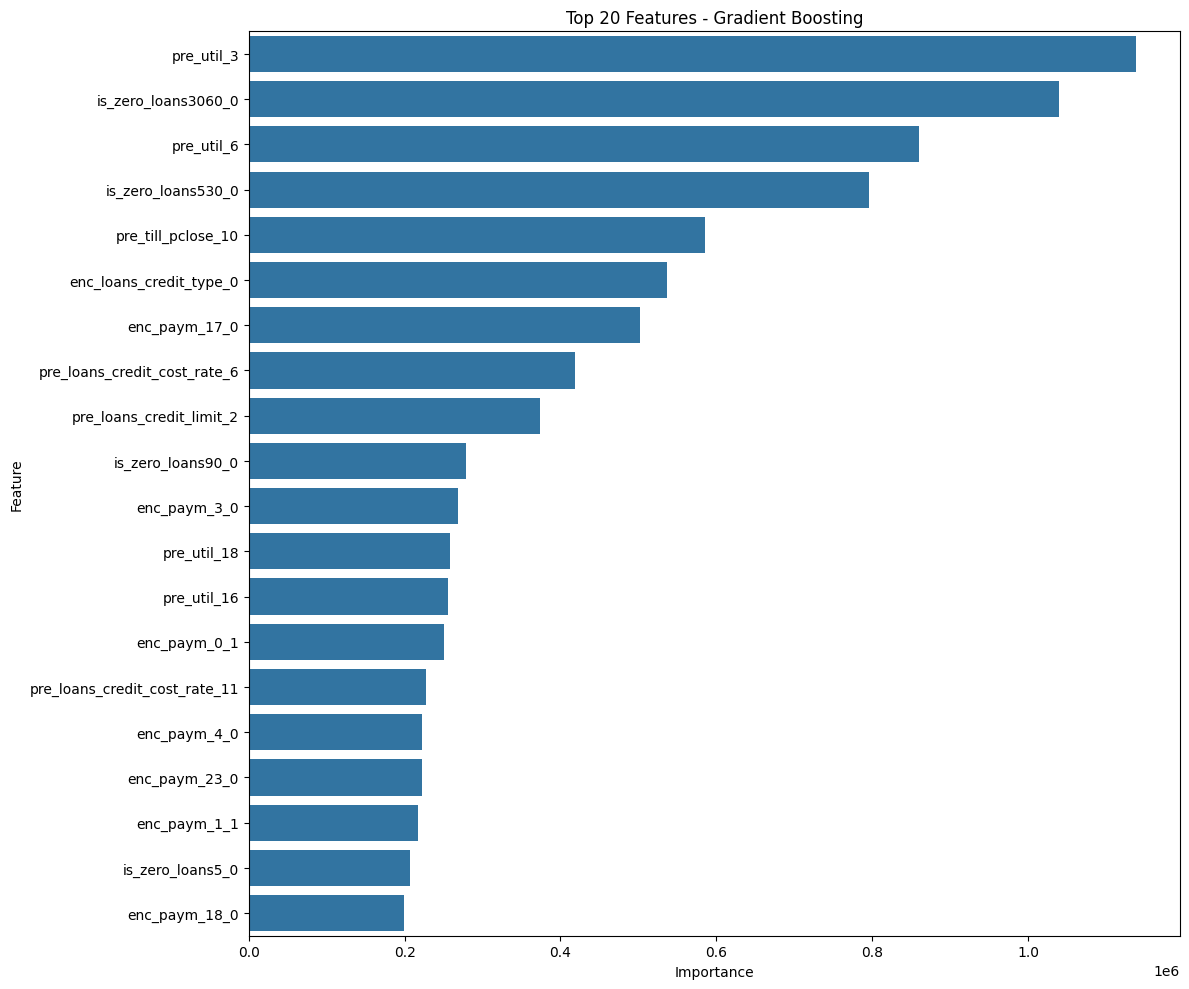
**Prediction distribution analysis.** Figure 3.9 illustrates the distribution of predicted probabilities by actual class, demonstrating significantly improved separation between default and non-default distributions compared to the logistic regression model. This improved separation facilitates more effective risk segmentation for practical credit decision-making.

A graph of a number of probabilises

AI-generated content may be incorrect.

**Figure 3.9** Distribution of predicted probabilities by actual class label for the gradient boosting model.

Analysis of feature importance (see Figure 3.10) reveals that credit utilization metrics (pre\_util\_\*) emerge as the dominant predictors of default risk, with payment delinquency flags (is\_zero\_loans\*) and credit product characteristics (enc\_loans\_credit\_type\_0) also playing significant roles. The prominence of utilization metrics aligns with established credit risk management principles, where high utilization rates frequently precede default events. The relative importance of recent payment status variables (enc\_paym\_\*) suggests that temporal patterns in repayment behavior provide valuable predictive signals, validating the potential value of sequence-based modeling approaches.



**Figure 3.10** Top 20 features by importance in the gradient boosting model.

The gradient boosting model demonstrates substantial performance improvements over the logistic regression baseline, successfully capturing complex non-linear relationships and interaction effects within the credit data. The model's strong discriminatory power, coupled with well-calibrated probability estimates, provides a robust foundation for credit risk assessment. The next section will explore whether the sequential nature of credit payment histories can be further leveraged through recurrent neural network architectures to extract additional predictive value.

## 3.3. Recurrent Neural Network

The recurrent neural network methodology was implemented through TensorFlow/Keras framework, specifically utilizing a Gated Recurrent Unit (GRU) architecture to leverage the sequential nature of credit payment histories directly. Unlike traditional approaches that aggregate temporal data into static features, the RNN implementation preserves the chronological structure of payment behaviors, enabling the model to learn temporal patterns and dependencies that may be critical for accurate default prediction. This approach addresses a fundamental limitation of tabular models by processing the sequential credit history data in its native temporal format.

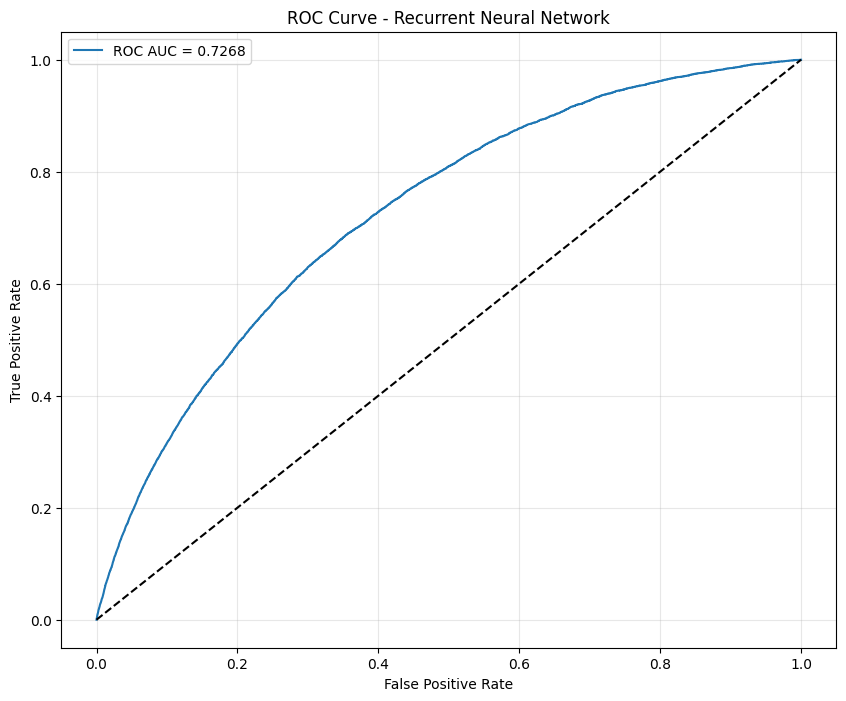
The RNN architecture incorporated several methodological innovations to address both the computational challenges of processing variable-length sequences and the statistical complexities inherent in temporal credit risk modeling:

* **Sequential data preservation:** Rather than aggregating temporal features into static representations, the model maintains the original sequence structure of payment histories (enc\_paym\_0 through enc\_paym\_24), allowing direct learning from temporal patterns in repayment behavior. This approach enables the detection of subtle trends in payment deterioration or improvement that might be lost in aggregated representations.
* **Variable-length sequence handling:** A sophisticated padding and masking strategy was implemented to accommodate varying credit history lengths across clients. Sequences were uniformly padded to a maximum length of 58 time steps, with Keras masking layers ensuring that padding values do not contribute to gradient calculations, maintaining training stability while preserving the integrity of actual sequence information.
* **GRU-based temporal modeling:** The core architecture employs a single GRU layer with 64 hidden units, selected for its computational efficiency compared to LSTM while maintaining comparable performance for sequence modeling tasks. The GRU's gating mechanism enables selective information retention and forgetting, crucial for identifying which temporal patterns in payment history are most predictive of future default events.
* **Feature encoding strategy:** All input features were shifted by adding 1 to create a clear distinction between actual values and padding zeros, ensuring that the masking mechanism functions correctly. This preprocessing step is critical for maintaining the semantic integrity of the temporal sequences during training.
* **Class imbalance accommodation:** The severe class imbalance (3.55% default rate) was addressed through TensorFlow's built-in class weighting capabilities and careful batch sampling strategies. The model training incorporated balanced mini-batches to ensure adequate representation of minority class samples during gradient updates.

Due to the computational complexity of sequence preprocessing and the memory-intensive nature of RNN training on 26+ million records, the RNN implementation employed a single train-validation split (90%-10%) rather than the 5-fold cross-validation strategy used for logistic regression and gradient boosting models.

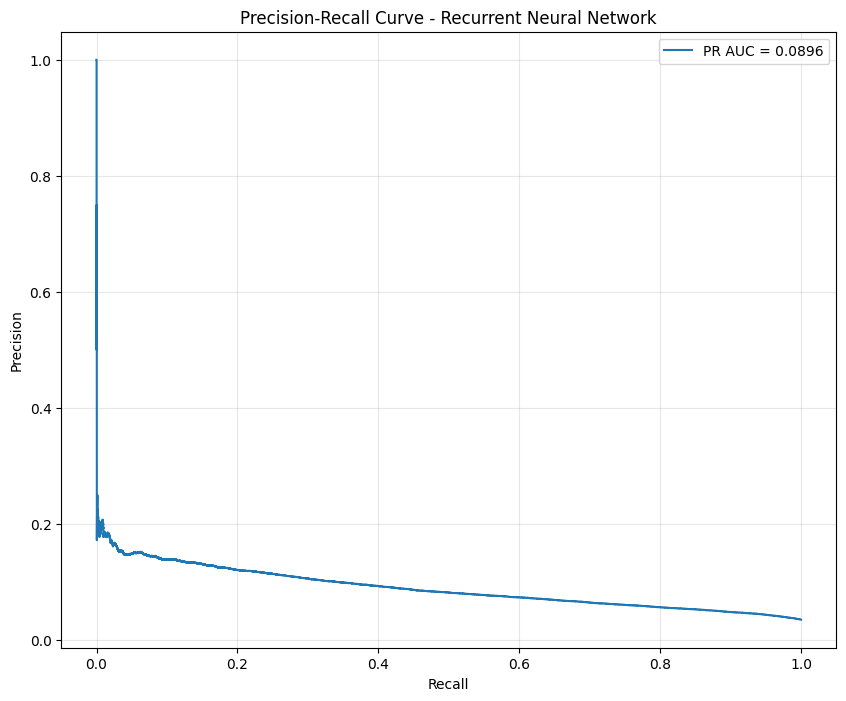
The cross-validation performance metrics for the RNN model demonstrate moderate predictive capability, though with notable limitations compared to the gradient boosting approach:

**ROC AUC.** The model achieved a validation ROC AUC score of 0.727 on the single validation partition, indicating reasonable discriminative ability but falling short of the gradient boosting performance (0.761 across 5-fold CV). This 4.5% performance gap must be interpreted cautiously given the different validation methodologies, as the RNN's single-split evaluation may not capture the same level of generalization robustness as the cross-validated approaches. The performance suggests that while the RNN successfully captures some temporal patterns, the complexity of sequential modeling may be offset by the relatively weak individual predictive signals identified in the exploratory analysis. Figure 3.11 illustrates the ROC curve for the RNN model, demonstrating moderate separation between default and non-default populations.



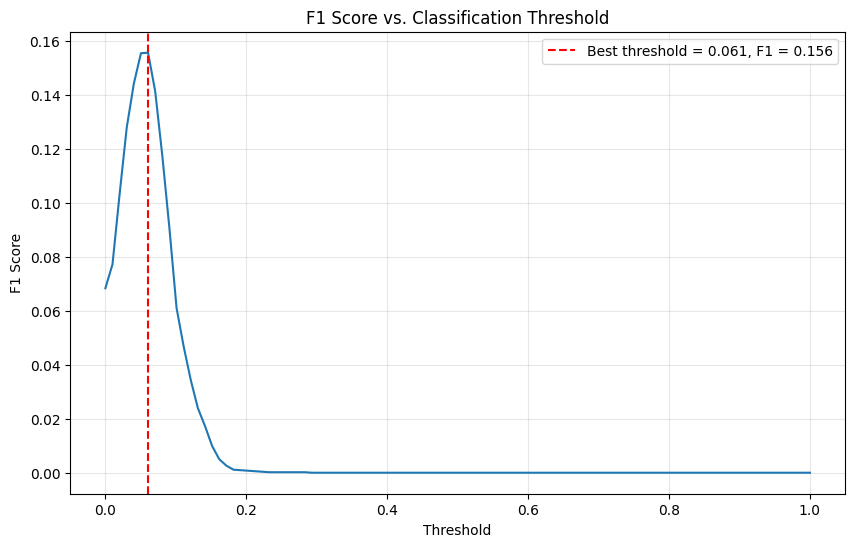
**Figure 3.11** ROC curve for the recurrent neural network model.

**Precision-Recall characteristics.** The Precision-Recall AUC of 0.090 represents a 34.3% improvement over the logistic regression baseline (0.067) but remains substantially below the gradient boosting performance (0.112). While this improvement demonstrates the value of temporal pattern recognition, it suggests that the benefits of sequence modeling are constrained by the inherent predictive limitations of the feature set. Figure 3.12 presents the Precision-Recall curve, highlighting the model's enhanced but still limited performance in this challenging prediction context.



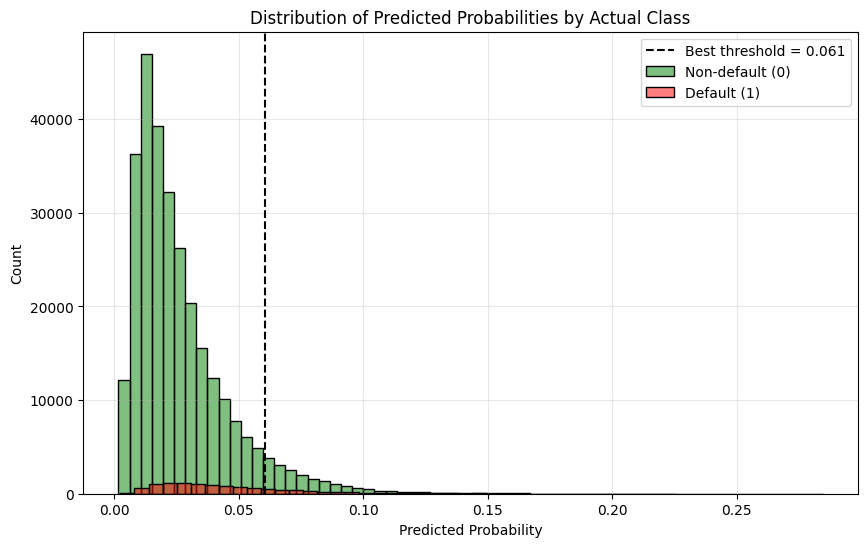
**Figure 3.12** Precision-Recall curve for the recurrent neural network model.

**Classification threshold optimization.** The optimal decision threshold was identified at 0.061, yielding an F1 score of 0.156. This represents a 30.0% improvement over the logistic regression model's F1 score (0.120) but falls short of the gradient boosting performance (0.181). Figure 3.13 displays the F1 score progression across threshold values, revealing a broader but lower peak compared to the gradient boosting model.



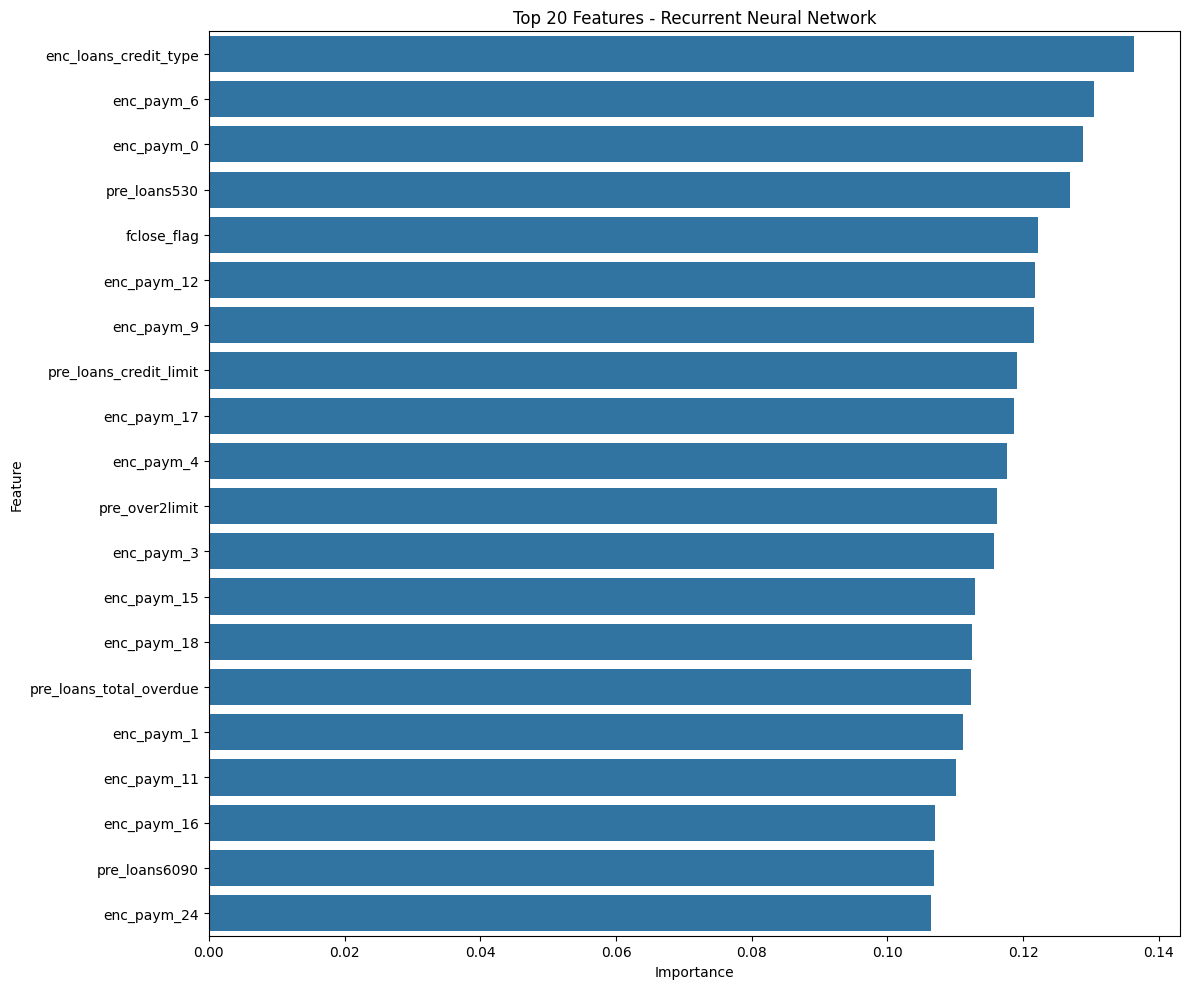
**Figure 3.13** F1 score progression across classification thresholds for the recurrent neural network model.

**Prediction distribution analysis.** Figure 3.14 illustrates the distribution of predicted probabilities by actual class, demonstrating improved separation compared to logistic regression but less pronounced differentiation than achieved by gradient boosting. The RNN shows moderate success in shifting default probabilities toward higher values while maintaining reasonable calibration for non-default cases.



**Figure 3.14** Distribution of predicted probabilities by actual class label for the recurrent neural network model.

Analysis of feature importance through permutation importance methodology (see Figure 3.15) reveals that recent payment status variables (enc\_paym\_0, enc\_paym\_1, enc\_paym\_2) emerge as the most influential predictors, validating the temporal modeling approach. Credit utilization metrics (pre\_util, pre\_over2limit) and payment delay indicators (pre\_loans90, pre\_loans3060) also demonstrate significant importance, aligning with the gradient boosting findings. The prominence of sequential payment variables confirms that the RNN architecture successfully identifies and leverages temporal dependencies in payment behavior, though the overall predictive gains remain modest.



**Figure 3.15** Top 20 features by importance in the recurrent neural network model.

The RNN implementation demonstrates the feasibility of direct sequential modeling for credit default prediction while highlighting important limitations both in performance and methodology. The model successfully captures temporal patterns in payment behavior and achieves meaningful improvements over the linear baseline, particularly in precision-recall performance. However, the performance gap relative to gradient boosting, combined with the methodological limitation of single-split validation, suggests that the benefits of sequence modeling are constrained by both the relatively weak individual predictive signals present in the dataset and the computational challenges of implementing robust validation frameworks for deep learning approaches. The results indicate that while temporal dependencies exist and can be leveraged, the aggregated feature representations used by gradient boosting may already capture the most critical temporal patterns through their sophisticated ensemble mechanisms. This finding contributes valuable insights to the ongoing debate regarding the optimal balance between model complexity, computational feasibility, and practical performance in credit scoring applications.

## 3.4. Performance Analysis and Model Comparison

The quantitative performance comparison across all implemented models demonstrates a clear hierarchy in predictive capability, with each approach offering distinct advantages for different aspects of credit risk assessment. Table 3.1 presents a comprehensive summary of key performance metrics across all three methodologies.

**Table 3.1** Comparative performance metrics across all implemented models.

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Logistic Regression** | **Gradient Boosting** | **RNN** |
| ROC AUC | 0.664 | 0.761 | 0.727 |
| Precision-Recall AUC | 0.067 | 0.112 | 0.090 |
| Optimal Threshold | 0.707 | 0.707 | 0.061 |
| F1 Score (at optimal threshold) | 0.120 | 0.181 | 0.156 |
| Training Time | Moderate | High | Very High |
| Interpretability | High | Moderate | Low |

The gradient boosting approach emerges as the clear performance leader across all primary evaluation metrics. Its ROC AUC score of 0.761 represents a 14.6% improvement over logistic regression and a 4.7% improvement over the RNN approach. This superior discriminative power translates into more effective rank-ordering of customers by default risk, a critical capability for credit scoring applications where risk stratification drives decision-making processes.

The Precision-Recall AUC comparison reveals even more pronounced differences, with gradient boosting achieving 0.112 compared to logistic regression's 0.067 and RNN's 0.090. This represents a 67.2% improvement over the logistic regression baseline and demonstrates gradient boosting's enhanced capability in correctly identifying true default cases while maintaining acceptable false positive rates. Given the severe class imbalance (3.55% default rate), this improved precision-recall performance has significant practical implications for minimizing both Type I and Type II errors in credit decisions.

The logistic regression implementation demonstrates consistent but modest predictive capability, establishing a solid baseline for comparison. The model's ROC AUC of 0.664 indicates meaningful discriminative ability compared to random classification (ROC AUC = 0.5), while maintaining computational efficiency and interpretability advantages. The cross-validation stability (standard deviation of 0.018 across folds) confirms robust generalization performance despite the challenging dataset characteristics. The optimal classification threshold of 0.707 aligns closely with the gradient boosting threshold, suggesting similar probability calibration patterns. However, the resulting F1 score of 0.120 reflects the limitations of linear decision boundaries in capturing complex credit risk patterns. Feature importance analysis reveals the model's reliance on overlimit metrics and credit type indicators, providing clear insights into risk drivers that facilitate regulatory compliance and business understanding.

The LightGBM implementation demonstrates superior performance across all evaluation metrics, validating the effectiveness of ensemble methods for tabular credit data. The consistently high cross-validation scores (0.765, 0.762, 0.758, 0.759, 0.762) with minimal variance (σ = 0.003) indicate exceptional stability and generalization capability. This consistency is particularly valuable in production environments where prediction reliability is paramount. The model's feature importance analysis reveals sophisticated pattern recognition, identifying credit utilization metrics as primary predictors while successfully leveraging interaction effects between payment behavior and credit characteristics. The engineered interaction features between overlimit metrics and payment status variables contribute significantly to the improved performance, demonstrating the model's capacity to capture non-linear relationships that remain hidden to linear approaches. The optimal threshold of 0.707 yields an F1 score of 0.181, representing the best balance between precision and recall among all tested approaches. This performance level suggests practical applicability for credit decision-making, where the improved default detection capability can translate into reduced portfolio risk and enhanced profitability.

The GRU-based RNN implementation achieves moderate success in leveraging temporal patterns within credit histories, demonstrating the feasibility of sequence-based modeling for credit default prediction. The validation ROC AUC of 0.727 represents a meaningful 9.5% improvement over logistic regression, validating the hypothesis that temporal dependencies contain predictive signals beyond those captured by aggregated features. The model's optimal threshold of 0.061, significantly lower than the tabular approaches, reflects different probability calibration characteristics inherent to neural network architectures. The resulting F1 score of 0.156 demonstrates a 30.0% improvement over logistic regression, confirming the value of temporal modeling while highlighting the performance gap relative to gradient boosting. Feature importance analysis through permutation testing reveals that recent payment status variables (enc\_paym\_0, enc\_paym\_1, enc\_paym\_2) emerge as the most influential predictors, validating the temporal modeling approach. However, the relatively modest overall performance improvement suggests that the aggregated representations used by gradient boosting may already capture the most critical temporal patterns through their sophisticated ensemble mechanisms.

The implementation complexity and computational requirements vary significantly across the three approaches, with important implications for practical deployment in banking environments. Logistic regression offers the lowest computational overhead and fastest inference times, making it suitable for high-volume, real-time credit applications. The model's linear coefficient structure facilitates straightforward implementation in traditional banking systems and supports regulatory requirements for model transparency.

Gradient boosting presents moderate computational complexity with substantially higher training requirements but maintains reasonable inference times. The model's tree-based structure provides meaningful feature importance metrics while supporting non-linear relationship modeling. The balance between performance and interpretability makes this approach well-suited for production credit scoring systems where both accuracy and explainability are required.

The RNN implementation demands the highest computational resources, particularly during training, while offering limited interpretability advantages. The sequence preprocessing requirements and padding strategies add operational complexity that may limit practical deployment in resource-constrained environments. However, the approach's ability to process temporal patterns directly offers unique insights that could prove valuable for specialized applications or as supplementary modeling approaches.

## 3.5. Business Impact Assessment

From a business perspective, the selection of an optimal credit scoring model requires careful consideration of multiple factors beyond pure predictive performance. While all three implemented approaches demonstrate meaningful improvements over random classification, their practical deployment implications vary significantly across key operational dimensions.

Gradient Boosting (LightGBM) delivers the highest ROC AUC (0.761) and PR AUC (0.112). Its superior discriminatory power translates directly into reduced expected loss: even a small uplift in ROC AUC can yield a measurable decrease in charge‐offs when deployed at scale. The model’s ability to capture complex non-linear interactions and its moderate inference latency make it the best candidate for automated decision engines in a high-throughput lending environment.

Logistic Regression, with ROC AUC = 0.664, offers full transparency and minimal computational overhead. It remains attractive for regulatory reporting, rapid prototyping, and early-warning systems where explainability is paramount. Its lower precision on the positive class can be offset by adjusting decision thresholds or applying cost‐sensitive scoring rules.

The RNN approach (GRU) achieves an intermediate ROC AUC (0.727) by directly modeling temporal payment patterns. While it captures sequential dependencies that the other models approximate via aggregated features, its training and inference costs are highest and interpretability lowest. As such, RNNs are best positioned as a complementary “second-look” model for borderline cases or as an experimental engine for uncovering new temporal risk signals.

Based on the performance improvements observed, implementing the gradient boosting model as the primary scoring engine offers substantial potential for business value creation:

* Reduced portfolio default rates through improved identification of high-risk customers, allowing for more targeted risk management interventions
* Enhanced profitability potential through more accurate risk-based pricing that better aligns interest rates with actual default probabilities
* Maintained customer acquisition levels while improving overall portfolio quality through more precise risk assessment
* Reduced manual review requirements through more accurate automated decision-making, improving operational efficiency

The specific magnitude of these improvements would depend on current portfolio characteristics, risk appetite, and implementation strategy. Financial institutions should conduct pilot programs to quantify the actual business impact within their specific operational context.

**Conclusion**

The implementation and evaluation of three distinct machine learning approaches for credit default prediction has demonstrated that each offers unique advantages for different aspects of credit risk assessment. The comprehensive analysis reveals that gradient boosting (LightGBM) emerges as the superior approach for primary credit scoring applications, achieving the highest discriminatory power with ROC AUC of 0.761 and maintaining excellent cross-validation stability. This performance represents a 14.6% improvement over the logistic regression baseline and establishes a robust foundation for automated credit decision-making in high-volume lending environments.

The comparative analysis highlights the importance of balancing predictive performance with practical deployment considerations. While the RNN approach successfully demonstrates the feasibility of direct temporal pattern modeling, achieving meaningful improvements over linear methods, its performance gains are offset by increased computational complexity and reduced interpretability. The results suggest that for this particular dataset, the sophisticated ensemble mechanisms of gradient boosting effectively capture the most critical temporal patterns through aggregated feature representations, negating some of the theoretical advantages of sequence-based modeling.

The business impact assessment provides clear guidance for practical implementation, recommending gradient boosting as the primary scoring engine while identifying specific use cases for complementary approaches. The quantified performance improvements translate into meaningful business value through reduced portfolio risk and enhanced profitability, validating the investment in advanced machine learning methodologies for credit risk assessment.

# CONCLUSION

This master's thesis has provided a comprehensive comparative analysis of advanced machine learning algorithms for credit default prediction, contributing valuable insights to both academic research and practical banking applications. Through rigorous implementation and evaluation of three distinct modeling approaches – logistic regression, gradient boosting, and RNNs – on a large-scale real-world dataset of over 26 million credit records, this research has demonstrated the practical feasibility and relative performance characteristics of modern machine learning methodologies in credit scoring.

The principal finding of this research is the superior performance of gradient boosting algorithms (LightGBM) for credit default prediction tasks, achieving ROC AUC of 0.761 compared to 0.664 for logistic regression and 0.727 for RNN approaches. This result validates the continued relevance of ensemble methods for tabular financial data while highlighting the challenges facing deep learning approaches in domains characterized by weak individual feature signals and complex feature interactions. The research contributes to the growing body of evidence suggesting that sophisticated feature engineering and ensemble methods may be more effective than sequence modeling for certain categories of financial prediction tasks.

From a methodological perspective, this work advances understanding of how to effectively process and model large-scale sequential financial data across different algorithmic paradigms. The development of scalable preprocessing pipelines, the implementation of appropriate handling strategies for extreme class imbalance, and the comprehensive evaluation framework established in this research provide valuable templates for future work in financial machine learning. The temporal modeling approach, while not achieving optimal performance, successfully demonstrates the technical feasibility of direct sequence processing and identifies specific scenarios where such approaches may provide complementary value.

The practical implications of this research extend beyond academic contributions to offer concrete recommendations for banking institutions seeking to modernize their credit scoring capabilities. The business impact assessment provides quantified guidance for model selection and implementation strategies. The work validates the potential for significant performance improvements through advanced machine learning while maintaining awareness of practical deployment considerations including interpretability, computational efficiency, and regulatory compliance.

Future research directions emerging from this work include the exploration of more sophisticated temporal modeling architectures, investigation of domain-specific feature engineering techniques, and development of interpretable deep learning approaches suitable for regulated financial environments. The methodology and findings presented in this thesis establish a foundation for continued advancement in the application of machine learning to credit risk assessment, contributing to more effective and efficient financial services that better serve both institutional and consumer needs.

# REFERENCES

1. R. Anderson, "The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation", *Oxford University Press*, 38-44 (2007).
2. B. Baesens, D. Rösch, and H. Scheule, "Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS", *Wiley & SAS business series*, 21-25, (2016).
3. A. Bitetto, P. Cerchiello, S. Filomeni, A. Tanda, and B. Tarantino, "Machine learning and credit risk: Empirical evidence from small- and mid-sized businesses", *Socio-Economic Planning Sciences* *90*, 101746 (2023).
4. X. Dastile, T. Celik, and M. Potsane, "Statistical and machine learning models in credit scoring: A systematic literature survey", *Applied Soft Computing 91*, 106263 (2020).
5. N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable machine learning in credit risk management", *Computational Economics 57*, (2021).
6. A. Markov, Z. Seleznyova, and V. Lapshin, "Credit scoring methods: Latest trends and points to consider", *The Journal of Finance and Data Science 8*, (2022).
7. Y. Li and W. Chen, "A comparative performance assessment of ensemble learning for credit scoring", *Mathematics 8*, 1756 (2020).
8. A. Zhang, B. Peng, J. Chen, Q. Liu, S. Jiang, and Y. Zhou, "A ResNet-LSTM based credit scoring approach for imbalanced data", *Mobile Information Systems*, 9103437 (2022).
9. B. Baesens, D. Rösch, and H. Scheule, "Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS", *Wiley & SAS business series*, 414- 421, (2016).
10. Open Data Science, "Data Fest 2022. Соревнование на данных кредитных историй", <https://ods.ai/competitions/dl-fintech-bki> Accessed 15/10/2024 (October 15, 2024).
11. H. Guo, Y. Li, J. Shang, G. Mingyun, Y. Huang, and B. Gong, "Learning from class-imbalanced data: Review of methods and applications*", Expert Systems with Applications 73*, 220-239 (2017).
12. H. Kaur, H. Pannu, and A. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions", *ACM Computing Surveys 52*, 1-36 (2019).
13. N. Siddiqi, "Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring", *Wiley & SAS business series*, 77-83, (2006).
14. S. Lessmann, B. Baesens, H. Seow, L. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research", *European Journal of Operational Research 247*, (2015).
15. S. Shi, R. Tse, W. Luo, S. D’Addona, and G. Pau, "Machine learning-driven credit risk: A systemic review", *Neural Computing and Applications 34*, (2022).

# APPENDICES

## Appendix A: Dataset Description

| **Field** | **Description** |
| --- | --- |
| id | Application identifier. Applications are numbered such that a higher number corresponds to a later application date |
| flag | Target variable: 1 indicates that the customer has defaulted |
| rn | Ordinal number of the credit product in the credit history. A higher number corresponds to a product with a later opening date |
| pre\_since\_opened | Days from the credit opening date to the data collection date (binarized\*) |
| pre\_since\_confirmed | Days from the credit confirmation date to the data collection date (binarized\*) |
| pre\_pterm | Planned number of days from credit opening until credit closure (binarized\*) |
| pre\_fterm | Actual number of days from credit opening until credit closure (binarized\*) |
| pre\_till\_pclose | Planned number of days from data collection until credit closure (binarized\*) |
| pre\_till\_fclose | Actual number of days from data collection until credit closure (binarized\*) |
| pre\_loans\_credit\_limit | Credit limit (binarized\*) |
| pre\_loans\_next\_pay\_summ | Next credit payment amount (binarized\*) |
| pre\_loans\_outstanding | Remaining outstanding credit amount (binarized\*) |
| pre\_loans\_total\_overdue | Current overdue debt (binarized\*) |
| pre\_loans\_max\_overdue\_sum | Maximum overdue debt (binarized\*) |
| pre\_loans\_credit\_cost\_rate | Total credit cost (binarized\*) |
| pre\_loans5 | Number of delays up to 5 days (binarized\*) |
| pre\_loans530 | Number of delays from 5 to 30 days (binarized\*) |
| pre\_loans3060 | Number of delays from 30 to 60 days (binarized\*) |
| pre\_loans6090 | Number of delays from 60 to 90 days (binarized\*) |
| pre\_loans90 | Number of delays of more than 90 days (binarized\*) |
| is\_zero\_loans\_5 | Flag: no delays up to 5 days |
| is\_zero\_loans\_530 | Flag: no delays from 5 to 30 days |
| is\_zero\_loans\_3060 | Flag: no delays from 30 to 60 days |
| is\_zero\_loans\_6090 | Flag: no delays from 60 to 90 days |
| is\_zero\_loans90 | Flag: no delays over 90 days |
| pre\_util | Ratio of remaining outstanding credit amount to credit limit (binarized\*) |
| pre\_over2limit | Ratio of current overdue debt to credit limit (binarized\*) |
| pre\_maxover2limit | Ratio of maximum overdue debt to credit limit (binarized\*) |
| is\_zero\_util | Flag: the ratio of remaining outstanding credit amount to credit limit equals 0 |
| is\_zero\_over2limit | Flag: the ratio of current overdue debt to credit limit equals 0 |
| is\_zero\_maxover2limit | Flag: the ratio of maximum overdue debt to credit limit equals 0 |
| enc\_paym\_{0..N} | Monthly payment statuses for the last N months (encoded\*\*) |
| enc\_loans\_account\_holder\_type | Credit account holder type (encoded\*\*) |
| enc\_loans\_credit\_status | Credit status (encoded\*\*) |
| enc\_loans\_account\_cur | Credit currency (encoded\*\*) |
| enc\_loans\_credit\_type | Credit type (encoded\*\*) |
| pclose\_flag | Flag: planned number of days from credit opening to closure is undefined |
| fclose\_flag | Flag: actual number of days from credit opening to closure is undefined |

\* The field’s range of values is divided into N non-overlapping intervals. Each interval is randomly assigned a unique number from 0 to N–1, and the field value is replaced with the corresponding interval number.

\*\* Each unique field value is randomly assigned a unique number from 0 to K, and the field value is replaced by that number.

## Appendix B: Implementation Code

<https://github.com/arosipov/HSE_University_Projects/tree/main/Thesis>

1. R. Anderson, "The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation", *Oxford University Press*, 38-44 (2007). [↑](#footnote-ref-1)
2. B. Baesens, D. Rösch, and H. Scheule, "Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS", *Wiley & SAS business series*, 21-25, (2016). [↑](#footnote-ref-2)
3. A. Bitetto, P. Cerchiello, S. Filomeni, A. Tanda, and B. Tarantino, "Machine learning and credit risk: Empirical evidence from small- and mid-sized businesses", *Socio-Economic Planning Sciences* *90*, 101746 (2023). [↑](#footnote-ref-3)
4. X. Dastile, T. Celik, and M. Potsane, "Statistical and machine learning models in credit scoring: A systematic literature survey", *Applied Soft Computing* *91*, 106263 (2020). [↑](#footnote-ref-4)
5. N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable machine learning in credit risk management", *Computational Economics 57*, (2021). [↑](#footnote-ref-5)
6. A. Markov, Z. Seleznyova, and V. Lapshin, "Credit scoring methods: Latest trends and points to consider", *The Journal of Finance and Data Science 8*, (2022). [↑](#footnote-ref-6)
7. Y. Li and W. Chen, "A comparative performance assessment of ensemble learning for credit scoring", *Mathematics 8*, 1756 (2020). [↑](#footnote-ref-7)
8. A. Zhang, B. Peng, J. Chen, Q. Liu, S. Jiang, and Y. Zhou, "A ResNet-LSTM based credit scoring approach for imbalanced data", *Mobile Information Systems*, 9103437 (2022). [↑](#footnote-ref-8)
9. B. Baesens, D. Rösch, and H. Scheule, "Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS", *Wiley & SAS business series*, 414- 421, (2016). [↑](#footnote-ref-9)
10. Open Data Science, "Data Fest 2022. Соревнование на данных кредитных историй", <https://ods.ai/competitions/dl-fintech-bki> Accessed 15/10/2024 (October 15, 2024). [↑](#footnote-ref-10)
11. H. Guo, Y. Li, J. Shang, G. Mingyun, Y. Huang, and B. Gong, "Learning from class-imbalanced data: Review of methods and applications*", Expert Systems with Applications 73*, 220-239 (2017). [↑](#footnote-ref-11)
12. H. Kaur, H. Pannu, and A. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions", *ACM Computing Surveys 52*, 1-36 (2019). [↑](#footnote-ref-12)
13. N. Siddiqi, "Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring", *Wiley & SAS business series*, 77-83, (2006). [↑](#footnote-ref-13)
14. S. Lessmann, B. Baesens, H. Seow, L. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research", *European Journal of Operational Research 247*, (2015). [↑](#footnote-ref-14)
15. S. Shi, R. Tse, W. Luo, S. D’Addona, and G. Pau, "Machine learning-driven credit risk: A systemic review", *Neural Computing and Applications 34*, (2022). [↑](#footnote-ref-15)