CREDIT SCORING IN DIGITAL AGE: BENCHMARKING TELECOM, CARD TRANSACTIONS AND CREDIT HISTORY DATA

Masuda Isaeva Utkirovna

Expert in Retail Business Department of "JSC Asakabank"

masudaisaeva@gmail.com

Abstract: In this study, we compare the predictive power of models based on credit history, card transactions, and telecom data. We train machine learning models on online microloan data and benchmark their performance against models proposed by telecom operators. Our results indicate that the model based solely on credit history data outperforms those based on card transaction and telecom data. Based on this evidence, we recommend prioritizing credit history data when developing credit scoring models.

Keywords: *card history, credit history, scoring models, hyperparameter optimization, imbalanced data.*

KREDIT SKORING: TELEKOMMUNIKATSIYA, KARTA TRANZAKSIYALARI VA KREDIT TARIXI MA'LUMOTLARINI SOLISHTIRISH

Masuda Isaeva O‘tkirovna

"Asakabank OAJ", Chakana biznes bo‘limi eksperti

masudaisaeva@gmail.com

Annotatsiya: Ushbu tadqiqod kredit tarixi, karta operatsiyalari va telekommunikatsiya ma’lumotlariga asoslangan skoring modellarning bashorat qilish qilish qobiliyatini solishtiradi. Biz onlayn mikroqarz ma’lumotlaridan foydalanib mashinani o‘rganish modellarini tuzdik va ularning natijalarini telekommunikatsiya operatorlari tomonidan taklif qilingan modellar bilan taqqosladik. Natijalar shuni ko‘rsatadiki, faqat kredit tarixi ma’lumotlariga asoslangan model karta transaksiyalari va telekommunikatsiya ma’lumotlariga asoslangan modellardan ko‘ra yaxshiroq aniqlikka ega.

Kalit so‘zlar: *karta tarixi, kredit tarixi, skoring modellari, giperparametr optimallashtirish, muvozanatlashmagan ma’lumotlar.*

КРЕДИТНЫЙ СКОРИНГ В ЦИФРОВУЮ ЭПОХУ: БЕНЧМАРКИНГ ДАННЫХ ТЕЛЕКОММУНИКАЦИОННЫХ КОМПАНИЙ, КАРТОЧНЫХ ТРАНЗАКЦИЙ И КРЕДИТНОЙ ИСТОРИИ

Масуда Исаева Уткировна

Эксперт в Департаменте рознисного бизнеса «ОАО Асака»

masudaisaeva@gmail.com

Аннотация: В данном исследовании мы сравниваем предсказательную силу моделей, основанных на кредитной истории, карточных транзакциях и данных телекоммуникационных компаний. Мы обучаем модели машинного обучения на реальных данных онлайн-микрозаймов и сравниваем их эффективность с моделями, предлагаемыми операторами связи. Наши результаты показывают, что модель, основанная исключительно на данных кредитной истории, превосходит модели, основанные исключительно на данных карточных транзакций и телекоммуникационных компаний.

Ключевые слова: *история карт, кредитная история, скоринговые модели, оптимизация гиперпараметров, несбалансированные данные.*

INTRODUCTION

Loans that do not generate interest or principal repayments for a minimum of ninety days are classified as non-performing loans (NPLs) or loan defaults [1]. Financial institutions are required to allocate a certain portion of their assets to cover potential losses, and this provisioning ratio increases with increasing loan defaults. Thus, loan defaults not only reduce profitability but also constrain banks' future operations and investment opportunities [2]. Hence, the proportion of loan defaults plays a critical role in the stability of the banking sector.

Predicting loan defaults has been extensively studied in both finance and data science and is often framed under research areas such as the probability of default and credit scoring. Banks and other financial institutions utilize credit scoring models to assess the creditworthiness of applicants. A scoring model is a tool designed to estimate the probability of loan default [3]. There are two primary types of scoring models: application scoring and behavioral scoring. The former is used to determine whether credit should be granted, while the latter is used for portfolio monitoring and initiating preventive actions upon early signs of default [4].

In Uzbekistan, retail banking products began gaining traction after 2020. This growth was fueled by the liberalization of the foreign exchange market, the expansion of government-backed lending programs and foreign debt, and improvements in financial infrastructure. The ease of access to credit led to a sharp rise in retail loan

uptake [5]. Due to the absence of prior loans or credit card usage (usage of only debit cards was common), most borrowers lacked credit histories, making it difficult for banks to assess creditworthiness [6] and retail loans were issued with fixed interest rates, often without collateral or insurance.

However, beginning in 2021, banks began encountering a growing problem of non-performing loans. According to Central Bank of Uzbekistan, in early 2020, the volume of loan defaults was 3.17 trillion UZS (1.5% of total loan portfolio), and by September 2021, this share had increased to 6.24%. Avo Bank recorded the highest default rate, with NPLs constituting 65% of its loan portfolio, followed by Octobank (37%) and Khalq Banki (24%). These values are significantly above the 5% regulatory ceiling for the NPL ratios [7]. Consequently, debt collection and credit scoring have become urgent priorities and banks quickly started establishing soft, hard, and legal collection departments. Although these measures helped reduce defaults after 2022, collection remains costly and inefficient.

The Credit Information Analysis Center (KATM) provides credit history data to financial institutions to support credit risk assessment. However, many financial institutions report that these data are of limited utility because a large share of potential borrowers lack formal credit histories and documented income [8]. Additionally, credit history data are often outdated, sometimes reflecting behavior from one or more years ago, making them less relevant for assessing current financial behavior [9]. As a result, financial institutions in Uzbekistan are exploring alternative data sources to better predict applicants' creditworthiness.

Researchers have suggested using alternative data such as digital footprints [10, 11], mobile phone usage [12], and e-commerce activity [13] to assess the creditworthiness of applicants. Transaction data, aggregated into categories such as gambling, groceries, utilities, transportation, insurance, entertainment, and telecom expenses, can predict repayment behavior of borrowers [14]. In fact, models trained on such transaction data have been shown to outperform those based on credit history - even when trained on unaggregated, raw transaction data over a 90-day window [15].

The motivation for this study is to examine the relative predictive value of credit history, card transaction, and telecom data in credit scoring and to assess the trade-offs of using each dataset. We develop a credit scoring model using the Extreme Gradient Boosting (XGBoost) algorithm and evaluate model performance using the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Our research addresses the following questions:

1. Which data source provides the highest predictive power for online microloan defaults?

2. Which features have the most significant impact on loan default risk?

3. When is telecom data an effective substitute to credit history data in developing scoring models?

This study contributes to the literature in several ways. First, to the best of our knowledge, this is the first empirical comparison of credit scoring models based on credit history, card transaction, and telecom data. Second, we train and evaluate models using a real dataset provided by a commercial bank in Uzbekistan, which consists of detailed transaction-level data on online microloans. Third, we interpret model predictions at both the global and individual levels by employing SHapley Additive exPlanations (SHAP) values. Fourth, we offer practical insights for financial institutions by comparing the performance of different data-driven scoring models.
The rest of this paper is structured as follows: Section 2 presents a literature review, Section 3 describes the data and methodology, Section 4 reports the empirical results, and Section 5 concludes with implications and future research directions.

LITERATURE REVIEW

Credit scoring models have the potential to support financial inclusion and serve as an efficient tool for promoting economic growth. It is critical that financial institutions and regulators collaborate to harness the benefits from scoring models while also managing their associated risks and challenges. Using machine learning algorithms to develop credit scoring models helps automate application processes, improve prediction accuracy, and ultimately enhance financial inclusion eliminate human bias, foster trust, build consumer confidence. However, to avoid unintended consequences, financial institutions must ensure fairness, transparency, interpretability, accountability, and privacy and security of personal data when deploying machine learning-based scoring models [2].

In data science, a scoring model is a supervised learning problem where the goal is to predict a borrower's future repayment behavior and classify borrowers as either good or bad (i.e., defaulters). The predictive performance of a scoring model is essential for maximizing the profitability of financial institutions, as even minor improvements can have major commercial implications and lead to significant cost savings. Research on credit scoring models can be broadly categorized into three groups: the first group focuses on proposing new models that improve accuracy over existing models, typically by integrating multiple models rather than relying on a single classifier; the second group addresses the problem of imbalanced data, and the third explores the predictive potential of various data sources for loan defaults.

Traditional machine learning algorithms are supervised learning models based on a single classifier and a combination of input variables to predict loan defaults. Logistic regression is the most commonly used model in credit scoring due to its ease of interpretation and implementation [16, 15]. It uses a maximum likelihood function

to classify observations as default or non-default. Its advantages include the use of statistical tools such as confidence intervals and offering easy interpretability [9]. While it is main limitations is that logistic regression cannot work with raw data and requires preprocessing steps such as encoding, scaling, or normalization [15]. In contrast, tree-based models are better suited for handling nonlinear data and identifying complex patterns in financial data. Consequently, algorithms such as neural networks (NNs), support vector machines (SVMs), and decision trees (DTs) are also have gained popularity in credit scoring. These models use network-like structures that mimic synaptic connections in the brain, adjusting weights iteratively to minimize prediction errors [17].

The use of ensemble classifiers in credit scoring has become a trend to overcome the limitations of single classifiers. Ensemble models combine multiple base learners using specific strategies and consistently outperform single-model approaches [18, 19]. For example, [20] developed a 16-layer hybrid ensemble model that achieved 97.39% accuracy and method called the Deep Genetic Hierarchical Network of Learners with 29 levels, which achieved 94.6% accuracy [21]. A three-stage learning algorithm based on unsupervised learning is proposed by [22], this model performs well in handling negative transfer learning problems. A dynamic scoring model of SurvXGBoost based on survival gradient boosting decision trees is developed by [3]. The results of this study indicate that proposed model outperformes others in both accuracy and misclassification cost. The APSO-XGBoost model developed by [17], which is based on adaptive particle swarm optimization, demonstrates superior performance compared with traditional machine learning and ensemble learning methods. Similarly, to these results, [9] suggests that that deep learning methods effectively manage highly correlated features and can automatically extract relevant features from raw data.

Data imbalance refers to the unequal distribution of samples across categories [23]. Addressing class imbalance in training datasets is essential to avoid bias toward the majority class. In credit scoring, this issue is significant because non-default loans typically dominate datasets, leading models to misclassify default cases (which are underrepresented) even when overall accuracy is high [24]. Common strategies to address this include data resampling techniques and cost-sensitive learning algorithms. Resampling methods, such as under-sampling, over-sampling, and Synthetic Minority Over-sampling Technique (SMOTE), are widely used [16, 25, 19, 26]. Comparison of the performance of Logistic Regression, Decision Tree, and Random Forest models on imbalanced datasets resampled with SMOTE for Nominal and Continuous (SMOTE-NC) and SMOTE-Edited Nearest Neighbors (SMOTE-ENN) shows that, all the models performed better on datasets resampled with SMOTE-ENN and therefore recommended its use for analyzing imbalanced data [27]. However, using resampling techniques can result in information loss, overfitting, changes to the original data

distribution, and high computational costs [24]. On the other hand, cost-sensitive learning involves incorporating misclassification costs directly into the model to minimize the overall cost and adjust classification thresholds [23, 28].

Popular datasets for building scoring models include Statlog from the UC Irvine Machine Learning Repository [20, 17, 29], Fannie Mae data [30], Australian FinTech lender data [14], and Lending Club data [18]. These datasets contain detailed information on borrower income, accounts, credit cards, and balances. However, models based on such datasets may be less effective in emerging markets where borrowers often lack credit histories or formal financial behaviors. To address these challenges and improve credit risk assessments, financial institutions are increasingly turning to nontraditional datasets.

By using alternative data sources, lenders aim to expand lending opportunities, better assess credit applications, and reduce decision times [22]. Using digital footprints is proposed by [10], which can match the predictive power of credit bureau scores and serve as complementary data. Studies demonstrated that email usage and psychometric variables can also yield accurate predictions [11, 29]. Using telecom data from mobile operators [12] and online purchase data [16] also offers high predictability of NPLs when no other data are available. Deep learning models trained solely on card transaction data can effectively predict future loan defaults for new customers [9, 15]. Categorizing transactions into groups, such as gambling, groceries, utilities, transportation, insurance, entertainment, and telecommunications, helps forecast borrower repayment behavior [14]. However, collecting such nontraditional data poses challenges stemming from the need to integrate data from multiple sources and their larger volume compared to traditional datasets [31]. Furthermore, reliance on digital footprints can oversimplify financial decision-making and may not account for the economic viability of credit scoring models [32]. Despite the surge in research on credit risk management in recent years, limited attention has been given to the use of such transaction-level data [9].

Although there is growing interest in using machine learning for credit scoring, effectively interpreting their predictions remains a significant challenge. Regulatory concerns over explainability and interpretability limit the deployment of complex black-box models in banking. Regulations often prohibit the use of opaque models, even when they outperform simpler ones [33]. The field of explainable AI focuses on tracing the decision-making process of machine learning models and identifying key features that drive predictions.

One widely adopted method is SHAP, a feature-based interpretability technique that integrates seamlessly with supervised models to increase their transparency and reliability [34]. SHAP can explain model outputs both globally and locally, such as by explaining why an individual loan application was rejected [15]. SHAP summary plots

show the overall impact of each feature, while force plots reveal the contribution of each feature value to a specific outcome, and stacked force plots highlight heterogeneity across values [35]. SHAP works well with tree-based models such as Random Forest and XGBoost, performs with some limitations in linear models, and does not support deep learning models [34].

Data

Credit data, as a form of financial data, typically suffers from class imbalance: there are far fewer default cases than non-default ones. Additionally, such datasets often include many features with complex interrelationships and many missing values. The data used in this study were provided by a commercial bank in Uzbekistan and it was fully anonymized to ensure customer confidentiality and prevent the identification of individuals and their financial relationships. Hence, there is no risk of private data leakage from this study.

This research follows a standard data preparation process. Initially, the dataset included 74,570 customers who obtained online microloans in January 2023. It tracked their repayment behavior up to July 2024, providing a performance window of 18 months after loan issuance and each borrower is labeled as a defaulter or non-defaulter. If a borrower delayed a payment for more than 90 days at least once during this period, then the loan was categorized as defaulted and the dependent variable was assigned a value of 0. Otherwise, if the borrower never delayed or delayed less than 90 days, then the loan was categorized as a good loan, with a dependent variable value of 1.

In the first step, the dataset contained 74,570 rows and 54 columns. In the second step, we removed two columns that had a single value in more than 99% of the rows and 19,991 rows that had missing values in more than 30% of their columns. We also analyzed the data for anomalies and outliers, removing an additional 9,821 rows. In the third step, we used a Random Forest Classifier to assess feature relevance and discarded 20 variables that had low or near-zero importance, following the method recommended by [26]. After cleaning, the final dataset consisted of 44,753 rows and 32 columns.

The dependent variable is a binary non-default indicator: it equals 1 for 38,946 non-default cases and 0 for 5,799 default cases, indicating an imbalanced dataset. In the final step, we split the dataset into card history and credit history subsets.

The first dataset represents card transaction histories. In Uzbekistan, two payment systems, Humo and Uzcard, issue deposit cards that are used for domestic transactions. Since 2023, these systems have made transaction history and balance data available to third parties via API for free of charge, with access limited to the 90 days preceding the application date. This dataset includes 14 independent variables calculated from transaction patterns such as incoming and outgoing transfers, ATM

withdrawals, spending on utilities and gambling, peer-to-peer transfers, and demographic data such as age and salary. Table 1 presents a summary.

Table 1

Summary of card history data

Variable name	*Description*	*Measurement*	*Min*	*Max*
avg_in_from_bank	Average monthly inflow from banks	Million Uzbek soums	0	7.66
max_amount_of_in	Maximum value of inbound transactions		0	0.2
avg_in_from_p2p	Average monthly peer-to-peer inflow		0	5.16
avg_in	Average monthly cumulative inflow		0	16.3
avg_out_p2p	Mean monthly outbound P2P transfers		0	5.86
avg_out	Average monthly outflows via all channels		0	16.4
avg_spend_gambling	Average monthly gambling-related outflows		0	3.74
avg_spend_util	Average monthly utility-related outflows		0	3.74
avg_cash_atm	Mean monthly ATM cash withdrawals		0	2.14
anomal_transactions	Maximum recorded anomalous transaction		0	3.9
ratio_of_avg_in_to_out	Ratio of P2P outflow to inflow	Integer	0	21,47
number_of_card_transactions	Average monthly number of transactions		0	2,65
number_of_salaries	Number of official salaries		20	86
age	Age of borrower at the time of application	Years	0	65
non-default indicator	Loan repayment status	Binary (0 = default, 1 = good)	0	1

All the variables are numeric. Ten variables are continuous and measured in millions of Uzbek soums (floats), and four are integers. The age variable ranges from 20 to 65 years. The number_of_salaries variable reflects the number of official salary payments received in the last 60 months. The maximum value of the variable equal to 86 indicates that the client had more than 1 official job.

The second dataset reflects the applicants' credit history, which was provided by KATM to financial institutions. It contains data on loan repayment behavior over the past five years, with six float variables (in million Uzbek soums) and twelve integer variables. Most variables have a minimum value of 0, except for scoring_grade, which has a minimum value of 98. This suggests that even borrowers with no prior loans or reported income are assigned a score based on alternative data, such as enforcement

bureau records, alimony payments, and utility bill payments over the last 12 months [5].

Table 2

Summary of credit history data

Variable name	*Description*	*Measurement*	*Min*	*Max*
total_outstand_debt_obl	Total outstanding debt obligations	*Million Uzbek soums*	0	1.85
avg_payment_outstand_debt_obl	Average monthly payment toward outstanding obligations		0	0.33
max_amoun_of_overdue_principals	Maximum overdue principal amount		0	375.2
total_accrued_interest_in_arrears	Total accrued interest in arrears		0	64.5
avg_monthly_payment	Average monthly debt obligations		0	27.7
actual_avg_monthly_payment	Observed average monthly loan repayment across all active credit accounts		0	506.1
count_of_closed_loans	Number of fully repaid loans	*Integer*	0	443
count_of_closed_loans_from_bank	Number of fully repaid bank loans		0	303
count_of_30_days_delays_in_3y	Number of 30+ days delinquency occurrences during the last three years		0	80
count_of_30_days_delays_in_1y	Number of 30+ days delinquency occurrences during the last year		0	59
scoring_grade	Credit score rating		98	487
number_of_contingent_liabilities	Number of contingent liabilities		0	366
max_days_of_overdue_interest	Maximum days of interest in arrears		0	1,83
max_day_of_overdue_principals	Maximum days of principal in arrears		0	1,68
number_of_loan_requests	Number of loan applications		2	3,38
number_of_overdue_principals	Number of delinquent principal obligations		0	589
number_of_claims_wt_contracts	Number of loan applications without contracts		0	1,75
number_of_salaries	Number of official salaries		0	86
Age	Age of borrower on application date	*Years*	20	65
non-default indicator	Loan repayment status	*Binary (0 = default, 1 = good)*	0	1

Uzbekistan has five telecom operators, three of which offer credit scoring services. For confidentiality reasons, we refer to them as Operator #1, Operator #2 and Operator #3. To evaluate their model performance, we provided to all operators anonymized data of applicants containing repayment outcomes for 25% of the borrowers to facilitate retro-testing However, due to the limited size of the retro-test dataset, the models returned results for different numbers of borrowers:

Operator #1 assessed 5,008 borrowers – Model 1

Operator #2 assessed 17,692 borrowers – Model 2

Operator #3 assessed 44,753 borrowers across three models: Model_3, Model_4, and Model_5.

We used a tree-based model, XGBoost, in which no normalization or feature scaling was applied, following the recommendation of [26] and the datasets were split into 70% training and 30% testing subsets for model training and evaluation.

METHODOLOGY

The reason for choosing XGBoost in this study is that it works well with raw and imbalanced data. This makes it particularly suitable for credit scoring model development [36]. Although XGBoost generally performs well, its optimal performance depends on proper hyperparameter tuning [17], therefore, we optimize the hyperparameters of XGBoost using the grid search method. The choice of using cost-sensitive learning and the Area Under the Curve (AUC) as evaluation metrics is motivated by the nature of the data as both techniques are widely recommended for handling imbalanced datasets.

XGBoost (Extreme Gradient Boosting) is an ensemble learning algorithm based on decision trees. It adopts regression trees as weak learners and then continuously adds new trees in successive iterations to fit the residuals, thereby improving model performance and efficiency [26]. XGBoost is relatively insensitive to data imbalance because it emphasizes the correct ranking of imbalanced data points, offers high flexibility, strong predictive power, strong generalization ability, scalability, high training efficiency, and robustness [23]. Since hyperparameters directly influence the model's structure and performance, appropriate tuning is crucial [17].

Hyperparameter tuning/optimization refers to the process of adjusting the hyperparameters of model to achieve the best performance [36]. The most commonly used techniques include grid search, random search, and Bayesian optimization [28]. Grid search is a method that systematically explores a specified range of hyperparameter combinations. It applies each combination to the model and evaluates the resulting performance. The combination yielding the best performance is selected as the optimal configuration [26].

The core idea of cost-sensitive learning is to assign higher weights to errors from underrepresented classes in the loss function. This ensures that defaulted loans receive more attention during model training [28]. By setting different misclassification costs for different classes, cost-sensitive learning offers a flexible approach for handling imbalanced data sets [23].

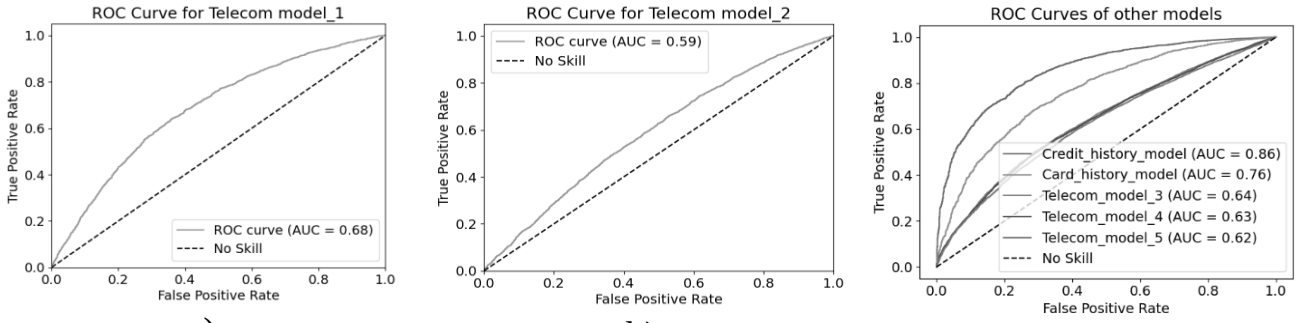
For imbalanced classification problems, the Area Under the Curve (AUC) is used as a key performance metric [28]. The AUC represents the area under the Receiver Operating Characteristic (ROC) curve, which plots the False Positive Rate (x-axis) against the True Positive Rate (y-axis) [37]. A higher AUC value indicates that the ROC curve is closer to the top-left corner, indicating better classification performance [23].

SHAP (SHapley Additive exPlanations) analysis is a feature-based interpretability method that provides both local and global explanations. It offers interpretable values and can be easily implemented using widely available packages [34]. SHAP values are model-agnostic, requiring only the model outputs and observed values, regardless of model type. They also capture the heterogeneity in feature effects across different outcomes. In addition to indicating feature importance, SHAP values reveal the direction (positive or negative) of each feature's impact on the outcome variable [38].

DISCUSSION AND RESULTS

The AUC (Area Under the Curve) is one of the most commonly used and intuitive indicators of a model's predictive ability. In credit scoring, even a small improvement in model performance can significantly reduce financial losses by helping institutions identify high-risk loan applicants more accurately.

Figure 1 presents the AUC curves for all the models. Since model_1 and model_2 have different sample sizes, we visualize them separately. Model_3, model_4, and model_5 which have the same sample size, are visualized together.

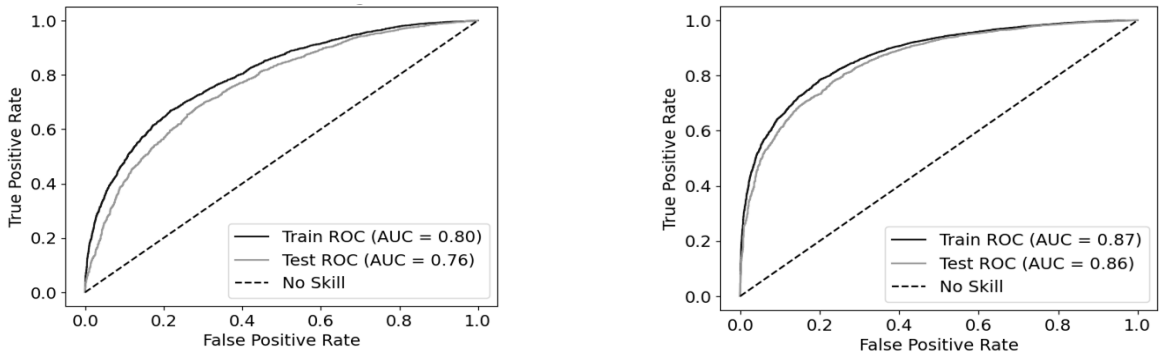


a) b) c)

Figure 1. AUC performance of models based on telecom (a, b), credit history data and card transaction data (c)

The figure shows that the model based on credit history data achieves the highest predictive performance. Model_1 predicts loan defaults with 68% accuracy (Figure 1a), and model_2 predicts loan defaults with 59% accuracy (Figure 1b) and model_3, model_4, model_5 have similar AUC values of approximately 0.63 (Figure_1c). This finding indicates that model_1 performs better than other models from telecom operators.

The model based on card transactions achieves the second-highest performance, with an AUC of 0.76, whereas the credit history-based model achieves the highest AUC of 0.86. Considering that the credit history dataset contains 19 independent variables and that the card transaction dataset contains only 14, the card transaction model demonstrates strong predictive power despite using fewer variables. In the next step, we combined the columns of credit history and the card transaction dataset to improve the predictive power of the model. However, this hybrid model had an AUC of 0.86, meaning that the performance of the hybrid model was equal to the performance of the model based solely on credit history data. This result indicates that credit history and card transaction data are alternative rather than complementary to each other. The high AUC values of both models raise concerns about potential overfitting, so we conducted an overfitting test, the results of which are presented in Figure 2.



a) b)

Figure 2. Overfitting check for models based on (a) card transaction history and (b) credit history data

In both cases, the difference between the AUC values for the training and testing datasets is less than 0.05. This suggests that there are no significant overfitting issues. Both the credit history and card transaction-based models demonstrate strong predictive power, indicating that, compared with telecom data, financial institutions can more effectively predict loan defaults using any of these data. However, the credit history-based model performs best and thus was selected for further analysis of the contributions of features on the outcome.

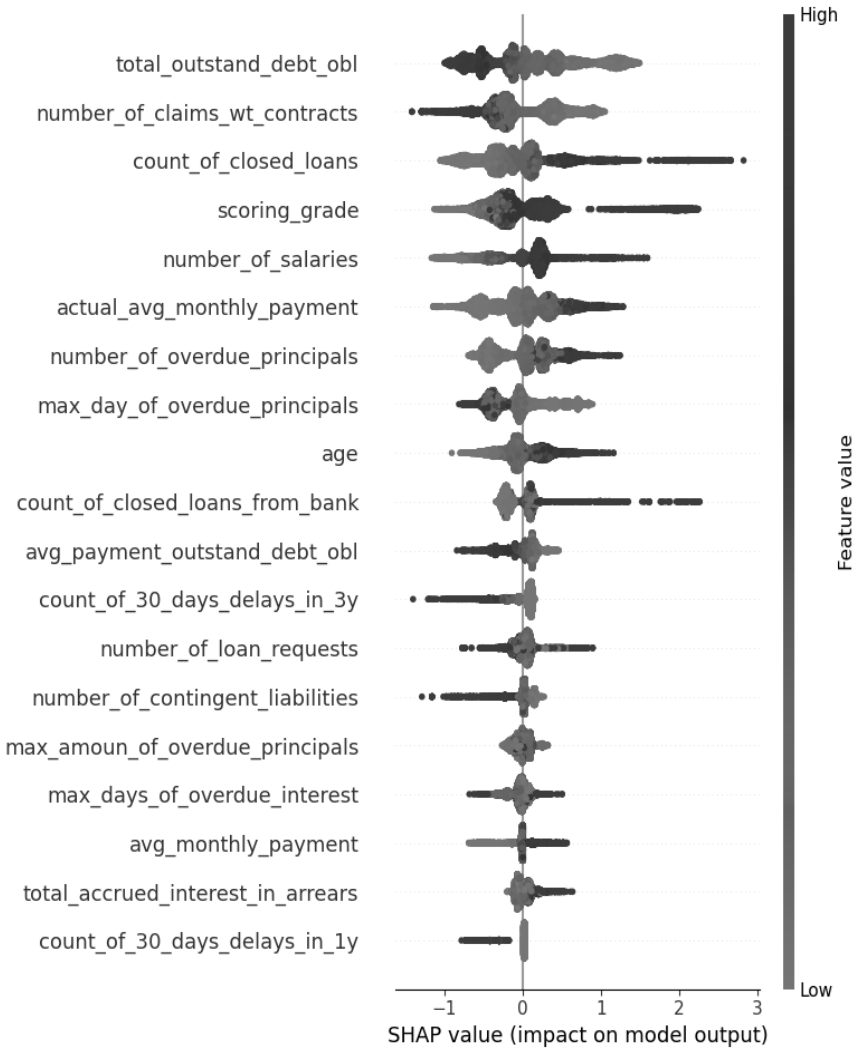


Figure 3. Summary plot of SHAP values for the model based on credit history data

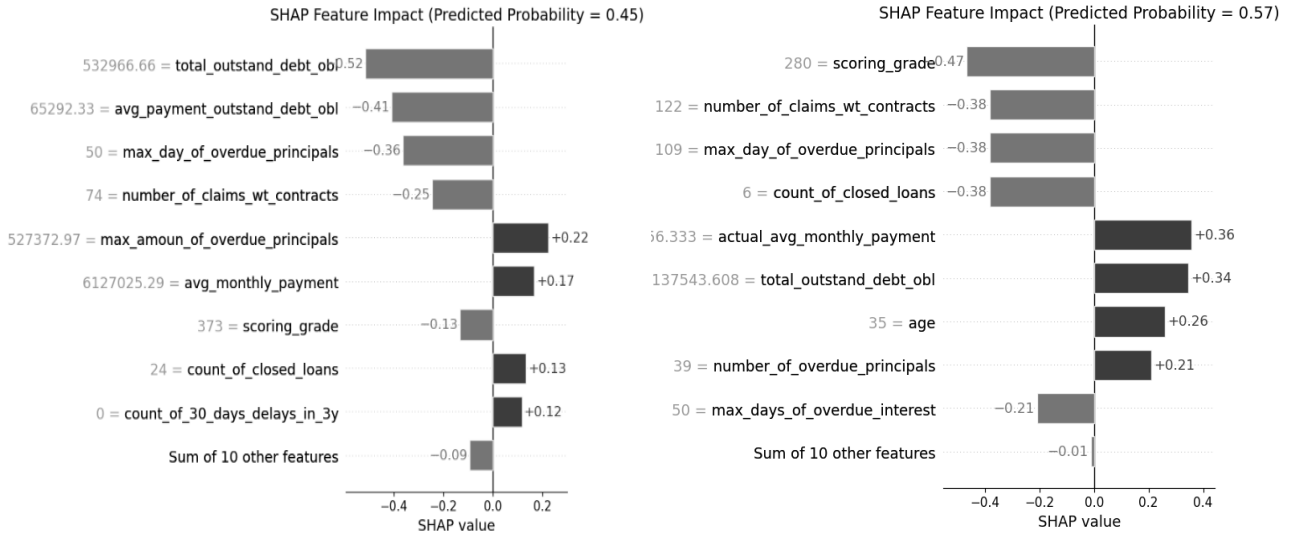
To interpret the impact of features, we apply SHAP values at both the global and individual levels. The SHAP summary plot helps us understand the relationship between features and model predictions (Figure 3).

In the plot, blue represents low feature values, and red represents high feature values. The values located on the left side have a negative effect on the model's output, whereas those on the right side have a positive effect. The top three features with the highest impact are total_outstand_debt_obl, number_of_claims_wt_contract, and count_of_closed_loans. The total_outstand_debt_obl and number_of_claims_wt_contract have a negative effect, whereas count_of_closed_loans have a positive effect on the model output. The three least important features are count_of_30_days_delays_in_1y, total_accrued_interest_in_arrears, and avg_monthly_payment. The effects of count_of_30_days_delays_in_1y are negative, whereas

total_accrued_interest_in_arrears and avg_monthly_payment have a positive effect on the model outcome.

The summary plot reveals that feature impacts are heterogeneous, i.e., different values of the same feature can have varying effects on the outcome. Figure 4 shows SHAP force plots that illustrate the contribution of individual feature values to positive and negative model predictions.

Figure 4a presents a case with a predicted probability of being a good loan = 0.57 (i.e., probability of default = 0.43). The features in blue negatively affect the prediction, whereas those in red increase it. For example, scoring_grade=280, number_of_claims_wt_contracts=122, max_day_of_overdue_principals=109, and count_of_closed_loans=6, along with 10 other features, reduce the predicted probability. In contrast, actual_avg_monthly_payment=475265 UZS, total_outstand_debt_obl=137543 UZS, age=35 and number_of_overdue_principals=39 increase the probability of being a good loan.



a) b)

Figure 4. SHAP force plots showing contributions of feature values to (a) positive and (b) negative model outcomes

Figure 4b shows a case with a predicted probability of being a good loan = 0.45 (i.e., probability of default = 0.55). Features such as total_outstand_debt_obl=532966 UZS, avg_payment_outstand_debt_obl=65292, max_day_of_overdue_principals=50, number_of_claims_wt_contracts = 74, scoring_grade=373 and the sum of the other 10 features reduce the probability. Meanwhile, max_amount_of_overdue_principals=527372 UZS, avg_monthly_payment=6127025 UZS, count_of_closed_loans=24 and count_of_30_days_delays_in_3y=0 increase it.

Importantly, SHAP analysis does not quantify the real-world importance of predictors. Instead, it indicates how features contribute to the model's predictions, not to actual observed outcomes.

Discussion

When issuing retail loans, the Central Bank of Uzbekistan stipulates that applicant should meet several requirements. First, applicants should not have actual delayed loan payments and due payments for utilities, court payments, or fines. KATM provides all of this information to financial institutions. When a borrower has no prior loans, a score is assigned based on alternative data, such as enforcement bureau records, alimony payments, and utility bill payments over the last 12 months [5].

In addition, the applicant must have an official income for at least six months, and the total monthly payments for all loans should not exceed 50% of their official income. Applicants are required to provide details of at least one bank card through which they receive their salary and the transactions of this card are analyzed to verify the applicant's income and repayment capacity.

As a result, financial institutions are obliged to use both credit history and card transaction data during the application evaluation process. Since they already incur costs to access these data, it is more practical and cost-effective to develop scoring models based on these sources rather than incurring additional expenses by using telecom data (models). The advantages and disadvantages of each data source are summarized in Table 3.

Table 3

Advantages of using credit history, card transaction history and telecom data

Data	*Advantages*	*Disadvantages*
Credit history	– High relevance to credit scoring – Does not require additional expenses – Allows features selections and data preprocessing – Allows modifying models and variables	– Requires technical infrastructure for data collection and processing – Requires hiring a data team (e.g., data scientist, analyst, and engineer) – Requires at least two years of client data for model development
Telecom models (data)	– Easy to deploy, i.e., no technical infrastructure and data needed – Fast implementation, i.e., no need to hire a data team	– Low relevance and model performance – Lack of transparency (i.e., telecom operators do not explain which features and algorithms are used) – Limited control (i.e., institutions cannot modify variables or models) – Additional cost – Data privacy concerns (e.g., requires sharing applicants' phone numbers with telecom operators)
Card transactions	– Moderately relevant to credit scoring	– Requires technical infrastructure – Requires hiring a data team

	– No additional expenses are needed – Allows feature selection and preprocessing – Requires collecting less data (up to 90 days)	– Requires at least two years of client data for robust model development

On the other hand, developing in-house scoring models requires collecting credit history and card transaction data from a large number of applicants. Model development typically starts one to two years after the performance window closes and requires hiring a data team composed of a data analyst, data engineer, and data scientist. Overall, the model development and deployment process can take up to two years. Not all banks have such resources. In Uzbekistan, there are currently 37 banks, 11 of which were established after 2020 and only began issuing loans in 2022 [7]. These newer banks may lack both the necessary data and infrastructure to build in-house scoring models. For such institutions, using a telecom-based model is better than not using any scoring model at all.

In conclusion, we recommend that financial institutions without the infrastructure or historical data to develop in-house models temporarily adopt telecom-based scoring models for one or two years. During this time, they should focus on collecting client data and building the necessary infrastructure to transition to more robust in-house models based on credit history or card transaction data.

CONCLUSION

In this study, we focused on developing credit scoring models using data from actual loan application cases, where telecom data, credit history, and card transaction history were collected as part of the credit assessment process. A commercial bank in Uzbekistan provided real-world data on online microloans, and two telecom operators contributed assessment results based on their own models. We employed the XGBoost algorithm, data balancing techniques, hyperparameter optimization, and the AUC-ROC metric for our analysis.

We found that the model based on credit history data was highly predictive of future loan defaults. This result suggests that credit history data have greater predictive power than telecom or card transaction data in credit scoring. This may be because card transaction data reflect recent financial behavior over the last 90 days, whereas telecom data tend to represent broader behavioral patterns. Therefore, we recommend using credit history data for developing scoring models rather than relying on telecom data. We also observed that card transaction data also have good predictive value for credit history data. However, when we integrated card transactions and the credit history dataset, the predictive power of this hybrid model was equal to the predictive power of

the model based on credit history data. This finding indicates that card transaction data may serve as a viable alternative rather than a complement to credit history data.

Our research adds to the ongoing debate about the optimal data sources for credit scoring by demonstrating that, for this specific task, models trained on credit history and card transaction data outperform those trained on telecom data. Our findings align with those of [32, 9, 15] and contrast with those of [12]. This discrepancy suggests that the effectiveness of telecom data may vary depending on country-specific factors, reinforcing the need for further research on model performance across national and institutional contexts. We analyze the advantages and disadvantages of using credit history, telecom, and card transaction data, considering the requirements of regulations.

Moreover, our study contributes to the field of data science by demonstrating the effectiveness of hyperparameter tuning and handling imbalanced datasets. Additionally, the results show that XGBoost is capable of automatically extracting meaningful features from raw credit history and card transaction data, eliminating the need for manual feature engineering. The SHAP analysis in our study enhances the understanding of both global and local explanations for model predictions. At the global level, we found that total outstanding debt obligations, the number of fully repaid loans, and the number of overdue principals were among the most influential predictors of loan default risk. These findings are consistent with those of traditional credit risk modeling studies, which frequently emphasize aggregate behavioral indicators. Our study extends this literature by applying SHAP to models trained on raw data, offering deeper insights into the model's behavior and its decision-making rationale.

The findings of this study are interesting for financial institutions and data scientists. The strong performance of the model trained on card transaction data is encouraging for incumbent banks. This finding demonstrates that even data limited to the 90 days prior to application can be highly predictive of future defaults. This highlights the value of such data for financial institutions lacking extensive credit history or telecom data. As the volume of card transaction data continues to grow, its potential to enhance the performance and profitability of credit scoring models is substantial. In addition, including card transaction history in application scoring models could streamline the application process, free up institutional resources, and enhance the customer experience during loan applications.

Finally, this study is based on data from a single commercial bank in Uzbekistan, which may limit the generalizability of the findings to other financial institutions or developing economies. Future research should consider data from a broader range of financial institutions to validate and expand on these findings. Additionally, further analysis could explore differences between online and offline loan applications or between various types of individual loans.

REFERENCES

1. M. Khan. (2019). “Determinants of non-performing loans in the banking sector in developing state”, Asian Journal of Accounting Research, pp. 135-145.

2. WB. (2019). “Credit scoring approaches guidelines”, The World Bank Group.

3. Y. Xia. (2021). “A dynamic credit scoring model based on survival gradient boosting decision tree approach”, Technological and Economic Development of Economy, p. 96–119.

4. R. Muñoz-Cancino. (2022). “On the dynamics of credit history and social interaction features, and their impact on creditworthiness assessment performance”, arXiv preprint arXiv.

5. S. Latipov. (2022). “The Central Bank announced the growth of the debt burden on the economy”.

6. E. Fazilbekov. (2022). “In December, banks got rid of 1.4 trillion soums of problem loans”.

7. CBU. (2025). Statistical report of Central Bank of Uzbekistan.

8. A. Abdukadirov. (2022). “The main problems of macroeconomic policy of the Central Bank”.

9. L. Hjelkrem. (2022). “The value of open banking data for application credit scoring: case study of a Norwegian bank”, Journal of Risk and Financial Management, 15(12), pp. 1-15.

10. T. Berg. (2020). “On the rise of fintechs: Credit scoring using digital footprints”, The Review of Financial Studies, pp. 2845-2897.

11. V. Djeundje. (2021). “Enhancing credit scoring with alternative data”, Expert Systems with Applications.

12. D. Björkegren. (2020). “Behavior revealed in mobile phone usage predicts credit repayment”, The World Bank Economic Review, pp. 618-634, 2020.

13. B. Rozo. (2023). “The role of web browsing in credit risk prediction”, Decision Support Systems.

14. M. Gao. (2023). “Consumer behaviour and credit supply: evidence from an Australian FinTech lender”, Finance Research Letters.

15. L. Hjelkrem. (2023). “Explaining deep learning models for credit scoring with SHAP: A case study using Open Banking Data”, Journal of Risk and Financial Management, 16(4), p. 221.

16. X. Dastile. (2020). “Statistical and machine learning models in credit scoring: A systematic literature survey”, Applied Soft Computing, p. 106263.

17. C. Qin. (2021). “XGBoost optimized by adaptive particle swarm optimization for credit scoring”, Mathematical Problems in Engineering.

18. Y. Li. (2020). “A comparative performance assessment of ensemble learning for credit scoring”, Mathematics, p. 1756.

19. F. She. (2021), “A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique”, Applied Soft Computing, p. 106852.

20. P. Pławiak. (2019). “Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring”, Applied Soft Computing, p. 105740.

21. P. Pławiak. (2020). “DGHNL: A new deep genetic hierarchical network of learners for prediction of credit scoring”, Information Sciences, pp. 401-418.

22. A. Ashofteh. (2021). “A conservative approach for online credit scoring”, Expert Systems with Applications, p. 114835.

23. P. Zhang. (2022). “Research and application of XGBoost in imbalanced data”, International Journal of Distributed Sensor Networks, 18(6), 2022.

24. P. Yotsawat. (2021). “A novel method for credit scoring based on cost-sensitive neural network ensemble”, IEEE Access, pp. 78521-78537.

25. Z. Jiang. (2021). “A new oversampling method based on the classification contribution degree”, Symmetry, p. 194.

26. B. Maulana. (2025). “Churn Prediction in Credit Customers Using Random Forest and XGBoost Methods”, Indonesian Journal of Data and Science, 6(1), pp. 81-90.

27. M. Isaeva. (2025). “Cross-Selling Car Loans to Remittance Recipients in Uzbekistan: A Machine Learning Approach Using SMOTE-ENN”, Proceedings of 27th International Conference on Advanced Communications Technology (ICACT2025).

28. S. He. (2021). “An effective cost-sensitive XGBoost method for malicious URLs detection in imbalanced dataset”, 9, pp. 93089-93098.

29. S. Trivedi. (2020). “A study on credit scoring modeling with different feature selection and machine learning approaches”, Technology in Society, p. 101413.

30. C. Chlebus. (2022). “The Advantage of Case-Tailored Information Metrics for the Development of Predictive Models, Calculated Profit in Credit Scoring”, Entropy.

31. C. Onay. (2018). “A review of credit scoring research in the age of Big Data”, Journal of Financial Regulation and Compliance.

32. A. Loutfi. (2022). “A framework for evaluating the business deployability of digital footprint-based models for consumer credit”, Journal of Business Research, pp. 473-486.

33. S. Serengil. (2022). “A Comparative Study of Machine Learning Approaches for Non-Performing Loan Prediction”, International Journal of Machine Learning and Computing, pp. 2008-2014.

34. A. Ponce-Bobadilla. (2024). “Practical guide to SHAP analysis: explaining supervised machine learning model predictions in drug development”, Clinical and Translational Science, pp. 1-15.

35. M. Isaeva. (2025). “Understanding the Heterogeneous Impact of Remittances on Saving Behaviour in Uzbekistan: A Machine Learning Approach”, Machine Learning and Applications an International Journal, pp. 15-32.

36. S. Putatunda. (2018). “A comparative analysis of hyperopt as against other approaches for hyper-parameter optimization of XGBoost”, Proceedings of 2018 international conference on signal processing and machine learning.

37. M. I. Utkirovna. (2022). “Determinants of loan prepayment and comparison of machine learning approaches”, Proceedings of 2022 IEEE World Conference on Applied Intelligence and Computing.

38. M. Isaeva. (2024). “Remittances and Financial Inclusion: Micro-Level Empirical Evidence from Uzbekistan”, Proceedings of 4th Interdisciplinary Conference on Electrics and Computer (INTCEC).