Credit Modeling Research Project

Replication Study: Machine Learning or Econometrics for Credit Scoring: Let's Get the Best of Both Worlds

(Elena, Dumitrescu , Sullivan, Hu´e , Christophe, Hurlin , Sessi, Tokpavi, 2021)

ECON 6693 Master Research Project

Prepared by:Brandon Arnold &Chao Deng

8/14/2025

Abstract

As the growth of big data continues to reshape the financial sector there is increasing pressure on institutions to adopt more sophisticated and accurate modeling techniques. These gains in predictive performance often come at the cost of interpretability. However, interpretability is an essential requirement in regulated domains such as credit risk assessment. Due to the opaque nature of many advanced models, their application in credit decisioning and risk evaluation remains constrained. This paper aims to review the standard practices in credit risk modeling and then explore the implementation of Penalized Logistic Tree Regression (PLTR), a hybrid approach that seeks to balance predictive accuracy with interpretability. Through empirical analysis, we assess the viability of PLTR as a competitive alternative to both traditional logistic regression and ensemble methods.

Introduction

Credit functions as a vital mechanism within an economy by facilitating the flow of funds from savers and lenders to borrowers who can use those funds for productive purposes. It allows households to smooth consumption over time, enabling them to purchase goods, services, or assets without having to wait until they have accumulated the full cost in savings. For businesses, credit provides the capital needed to invest in equipment, technology, and expansion. This drives economic growth and job creation. At the macroeconomic level, access to credit supports higher levels of aggregate demand, fuels entrepreneurship, and can amplify the effects of monetary policy. However, excessive or poorly allocated credit can lead to financial instability, asset bubbles, and debt crises. The development of regulation and prudent management of credit flows essential for sustainable economic health.

It's important to note that when lending out credit there is always a possibility of default, and this probability affects both who receives loans and the interest rate at which they can borrow. Higher perceived default risk typically results in stricter lending criteria, higher interest rates, or outright denial of credit. This underscores the importance of accurately assessing borrower risk when firms decide to extend credit, as poor risk evaluation can lead to financial losses, increased non-performing loans, and reduced profitability. To address this challenge, the credit industry relies heavily on econometric and statistical modeling to quantify and predict the likelihood of default. One of the most widely adopted techniques is logistic regression.The logistic regression model is widely used in credit scoring practice due to its strong interpretability of results (Runchi, ligou, qin, 2022). This transparency makes it a cornerstone in credit scoring systems, where both predictive performance and justifiable decision-making are critical As society progresses and our capacity for data collection continues to expand at an exponential rate, our ability to process and analyze information is also rapidly increasing. This evolution prompts questions about whether logistic regression should continue to serve as the gold standard in credit risk modeling.With advancements in computational power and the emergence of more sophisticated classification methods such as: random forests, gradient boosting, and neural networks; it becomes essential to reconsider if logistic regression remains the optimal choice. While logistic regression's interpretability and transparency remain key advantages, emerging methodologies promise enhanced predictive accuracy, potentially reshaping best practices within the credit risk industry.

However, it's apparent that these models will have to meet the same standards of clarity and interpretability that logistic regression clearly. In the United States, the main standard you must meet is SR 11-7: Guidance on Model Risk Management. This document states that models implemented must meet a reasonable threshold of transparency, and

interpretability.

"All model components—inputs, processing, outputs, and reports—should be subject to validation; this applies equally to models developed in-house and to those purchased from or developed by vendors or consultants." (Federal Reserve, 2011) The three criteria mentioned for validation are: "*Evaluation of Conceptual Soundness,Ongoing Monitoring and Outcomes Analysis"* (Federal Reserve, 2011) These act as a guideline for credit risk models here in America, and ensure opacity is at minimum.

The European Banking Authority guidelines state, " Understand the quality of data and inputs to the model and detect and prevent bias in the credit decision-making process, ensuring that appropriate safeguards are in place to provide confidentiality, integrity and availability of information and systems." (ERB, 2020)

The bank for international settlement states "Supervisors or external auditors should also assess the quality of a bank's own internal validation process where internal risk ratings and/or credit risk models are used. Supervisors should also review the results of any independent internal reviews of the credit granting and credit administration functions." (Bis,2025) Validation and interpretation are paramount when getting audited and for quality control. So, developing a transparent and interpretable model ensures that companies are staying within guidelines.

Literature Review

(Elena, Dumitrescu† , Sullivan, Hu´e‡ , Christophe, Hurlin§ , Sessi, Tokpavi, 2021)
In credit risk modeling, financial institutions must carefully balance two competing priorities: predictive accuracy and regulatory transparency. Logistic regression remains a widely used tool due to its interpretability and straightforward implementation, making it highly suitable for environments where decisions must be clearly justified to regulators and stakeholders. However, its linear structure limits its ability to capture complex, non-linear relationships present in borrower behavior. In contrast, machine learning models such as Random Forests and Gradient Boosting Machines offer superior predictive performance, but their lack of transparency makes them difficult to deploy in settings that demand clear model governance. To address this trade-off, the Penalized Logistic Tree Regression (PLTR) model was proposed as a hybrid solution designed to deliver both interpretability and performance.

PLTR integrates two modeling approaches: tree-based segmentation and penalized logistic regression. The model begins by using a decision tree to partition the dataset into distinct subgroups, or leaves, based on the most informative features. Each leaf represents a cluster of observations with similar characteristics. These leaves are then added to a penalized logistic regression framework as explanatory variables. They use a penalization framework to enhance generalizability and reduce the risk of overfitting, using regularization techniques such as L1 (Lasso) or L2 (Ridge) penalties. This structure allows PLTR to model complex, non-linear interactions within the data like the random forest, while preserving interpretability through the log-odd coefficient estimates that can be evaluated at the subgroup level.

The authors of the original PLTR study validated the model using three publicly available credit datasets: the German Credit dataset, the Australian Credit dataset, and the Taiwan Credit Card Default dataset. These datasets provided diverse testing environments with both categorical and continuous variables, and varying degrees of class imbalance. PLTR consistently outperformed traditional logistic regression in terms of AUC and accuracy, while remaining competitive with more complex machine learning models. Most importantly, the model maintained its interpretability, offering clear decision pathways and localized explanations. This positions PLTR as a valuable tool for credit risk practitioners who require robust performance without compromising transparency or compliance standards.

However, the original study did not provide comprehensive details regarding data preprocessing procedures or the precise specifications of the implemented models. This lack of transparency introduces uncertainty regarding the comparability of results, particularly with respect to whether the models under evaluation were constructed using identical input variables and feature sets. In the present replication study, our objective is to systematically document

the development of the baseline models and to establish a controlled experimental framework in which model comparisons are conducted under consistent conditions;therefore, ensuring that the evaluation constitutes a true "apples-to-apples" comparison.

Data Management

The data set that we'll be using will be the Taiwanese credit card data set. This is one of the dataset used within the original paper referenced. This panel data contains 30,000 credit card client observations spanning 6 months.

At first glance, the dataset offers a balanced blend of demographic features, such as age, sex, education level, and marital status, alongside more dynamic indicators of financial health and credit behavior. The average age of clients is around 35 years, and variables like SEX, EDUCATION, and MARRIAGE provide categorical insights into client profiles. However, it's the behavioral variables particularly the repayment status indicators (PAY_0 to PAY_6) that begin to tell a deeper story. These values range from 0 (on-time payment) to 8 (serious delay), with most clients paying on time, but a significant subset showing signs of financial distress.

Equally important are the six months' worth of bill amounts and payment amounts, which capture trends in spending and repayment. These variables display substantial variation. For example, some clients carry modest balances, while others face bills exceeding several hundred thousand. Likewise, payments range from zero to extraordinary amounts in the millions. Some BILL_AMT values are even negative, suggesting cases of overpayment or credit reversals.

While the raw data offers powerful insights, it also presents challenges: inconsistent encodings, and skewed distributions. These are not flaws, but rather realities of financial data and they inform the careful adjustments and transformations we'll introduce in the next stage of our modeling process.

Data Challenges & Preprocessing

When preprocessing the data we have to look at certain issues that may arise when implementing the model in its raw form. This includes figuring out tactics to reduce problems like incorrect functional form or issues like multicollinearity. Since this is panel data it's apparent we can't use the monthly variables as originally provided. This is due to multicollinearity issues that will cause our estimates to be off and our standard errors to be unreflective of reality.

The first issue to take care of is the multicolinearity issue. Given that the dataset is structured as panel data, it is evident that the monthly variables cannot be used in their original form. Their high correlation with each other would introduce severe multicollinearity. To mitigate this issue, the most effective approach is to aggregate these monthly variables by computing their averages. This transformation reduces redundancy, preserves the interpretability of the, and ensures that the resulting estimates are more statistically reliable. This led us to create *average balance(avg_balance), average payments (avg_payments) and average delay (avg_delay)*.

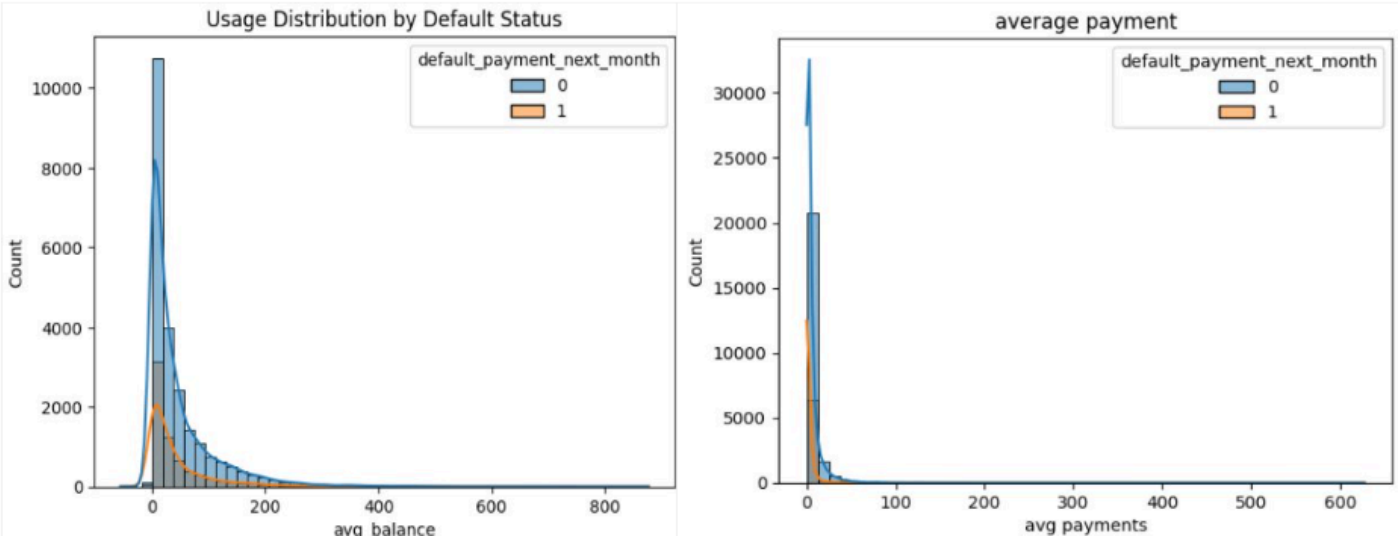

Figure 1: Distributions of balances and payments

The next consideration, after reviewing Figure 1, its that the average payment and average balance variables exhibit significant positive skewness. Such skewness can adversely affect model performance by violating the assumption of a correctly specified functional form and by disproportionately weighting extreme values. To address this, the variables are transformed using a natural logarithmic scale. This transformation reduces skewness, stabilizes variance, and improves the model's functional form, thereby enhancing both the efficiency and interpretability of the resulting estimates.

After correcting the raw data, we identified concerns with using the limit balance variable in its original form. While limit balance typically reflects the bank's assessment of an individual's creditworthiness, it does not directly capture the borrower's behavioral patterns. To address this limitation, we constructed a new variable designed to better reflect the influence of individual behavior on credit risk. The *usage* variable measures the proportion of available credit that an individual utilized on average over the six-month period and is calculated by dividing the average balance by the limit balance. This transformation provides a more behaviorally informative metric, enabling a richer assessment of credit risk beyond institutional credit limits.

This was complemented by the creation of a *payment-to-balance ratio*, which measures the proportion of an individual's outstanding balance that was paid off on average during the six-month period. This metric provides additional insight into repayment behavior, allowing for a more comprehensive evaluation of a borrower's credit risk profile.

The final constructed feature was the coefficient of variation for both payments and balances. This metric captures the relative volatility of these variables by measuring the standard deviation as a proportion of the mean. Higher values indicate greater variability, providing an additional dimension for assessing the stability and predictability of a borrower's financial behavior over the observed period.

Summary Statistics

Variable	Mean	Std Dev	Min	25%	50%	75%	max
LIMIT_BAL	167484.3	129747.7	10000	50000	140000	240000	1000000.00
SEX	.6	0.49	0	0	1	1	1.00
EDUCATION (cleaned)	0.82	0.38	0	1	1	1	1.00
MARRIAGE (cleaned)	0.46	0.5	0	0	0	1	1.00
AGE	35.49	9.22	21	28	34	41	79.00
avg_delay	0.28	0.6	0	0	0	0.33	6.00
avg_balance (NT$)	44976.95	63260.72	-56043.2	4781.33	21051.83	57104.42	877314.00
avg_payment (NT$)	5275.23	10137.95	0	1113.29	2397.17	5583.92	627344
usage	0.37	0.35	-0.23	0.03	0.28	0.66	5.36
CV balance	62.45	6402.00	-922.73	0.08	0.35	0.83	894427.00
Cv Payments	0.90	0.65	0.00	0.38	0.78	1.30	2.44
default payment next	0.22	0.42	0	0	0	0	1.00

Figure 2: Summary statistics

This table gives us an overview of the key characteristics in our cleaned dataset. On average, individuals had a credit limit of around 167,000, but there's a wide range from as low as 10,000 to as high as 1 million showing just how different people's financial situations can be. Most people in the dataset are around 35 years old, and about 22% of them ended up defaulting on their payments the next month.

We also created several new variables to better reflect behavior. For example, avg_delay captures how late someone usually is with payments, and usage reflects how much of their credit limit they're using. The large variation in payment and balance amounts shows how unpredictable financial activity can be, with some people paying nothing and others making very large payments. Overall, this summary helps us understand the financial habits and risk levels in our sample before modeling.

Empirical Methods

In this section, we review the four primary modeling methods implemented in the study, outlining their respective strengths and weaknesses. This discussion ensures transparency in the analytical process and establishes a solid foundation for understanding and interpreting the resulting model outputs. By clearly defining the capabilities and limitations of each approach, we provide the necessary context for evaluating their comparative performance in addressing the research objectives.

OLS

Ordinary Least Squares (OLS) regression is one of the most widely used statistical techniques for modeling relationships between variables, offering a clear framework for estimating the linear association between predictors and a continuous outcome variable. Its popularity stems from its mathematical simplicity, ease of interpretation, and strong theoretical grounding in classical regression analysis. OLS provides unbiased and efficient estimates under the Gauss–Markov assumptions, making it a valuable starting point for understanding the structure of a dataset and the

potential influence of different predictors. Because of its transparency and straightforward interpretation, OLS often serves as a foundational method in academic and applied research, allowing analysts to establish a baseline understanding of how variables interact before progressing to more complex models.

Despite its utility as a general modeling tool, OLS is not well-suited for credit risk prediction. Credit default is a binary outcome either a borrower defaults or they do not. which violates OLS assumptions regarding continuous dependent variables and normally distributed residuals. Using OLS in this context can lead to predictions that fall outside the [0,1] probability range and fail to properly account for the nonlinear relationship between predictors and the probability of default. For these reasons, logistic regression (logit models) is preferred in credit risk modeling, as it is specifically designed for binary classification problems, ensures outputs are valid probabilities, and aligns more closely with the statistical requirements of regulatory frameworks such as Basel III and EBA guidelines.

Logit Regression

The logit model is the most widely adopted statistical method for modeling binary outcomes such as credit default. Unlike OLS logistic regression is estimated using maximum likelihood estimation (MLE) rather than minimizing squared errors. This approach identifies the set of parameters that maximizes the likelihood of observing the actual default and non-default outcomes in the data. By modeling the log-odds of the dependent variable as a linear combination of the independent variables, the logit model ensures that predicted probabilities always fall within the [0,1] range. This makes it particularly well-suited for classification problems in credit risk, where probability predictions must be both bounded and interpretable.

One of the primary advantages of logistic regression in credit risk modeling is its interpretability. Coefficients can be directly translated into odds ratios, allowing lenders and regulators to understand how each predictor influences the likelihood of default. This transparency is critical under regulatory frameworks such as Basel III and the EBA's Internal Ratings-Based (IRB) requirements. which emphasize model explainability and auditability. Logistic regression can also incorporate non-linearities and interactions through variable transformations, making it adaptable to more complex datasets. However, it assumes a specific functional form and may underperform relative to advanced machine learning models when capturing highly non-linear relationships or complex variable interactions. Despite these limitations, its combination of predictive reliability, interpretability, and regulatory acceptance has made it the gold standard in credit risk modeling.

Random Forest

Random Forest is an ensemble learning method built upon the foundation of decision trees. Unlike a single decision tree, which can be prone to overfitting, Random Forest combines the predictions of many individual trees to produce a more robust and generalizable model. It employs a process known as bootstrapping. This is drawing random samples with replacement from the training dataset to build each tree. Ensuring that no single model is overly dependent on any specific subset of the data. Additionally, at each split within a tree, a random subset of predictor variables is considered, further increasing model diversity and reducing correlation between trees. This ensemble approach improves predictive stability and reduces variance, making Random Forest a powerful tool for classification problems, including credit default prediction.

One of the key advantages of Random Forest is its ability to naturally model non-linear effects and complex interactions between variables without requiring explicit functional form specification. This flexibility allows it to capture patterns in borrower behavior that might be missed by traditional linear models such as logistic regression. Furthermore, Random Forest is relatively robust to outliers and multicollinearity and it provides measures of variable

importance that can guide feature selection. However, it is often criticized for its reduced interpretability compared to simpler models, which can present challenges in regulatory environments such as credit risk management under Basel III and EBA requirements. Despite this, its high predictive accuracy and adaptability to large, complex datasets make Random Forest a valuable benchmark when evaluating model performance in credit scoring applications.

Model Sequence and Justification

When implementing our models, we follow a deliberate sequencing strategy to establish a strong foundation for comparison across all methods. To begin, we employ Ordinary Least Squares (OLS) regression to gain an initial understanding of how the explanatory variables relate to the dependent variable, providing a baseline assessment of variable interactions. We then proceed to the logistic regression (logit) model, which is more appropriate for binary classification and serves as the primary benchmark for credit risk prediction. Once we have identified the optimal set of explanatory variables, these same inputs are applied to both the Penalized Logistic Tree Regression (PLTR) model and the Random Forest model, ensuring a consistent basis for comparison. Finally, we present a comprehensive evaluation of model performance, highlighting relative strengths, weaknesses, and insights gained from the analysis.

OLS Regression Tables	Model 1 (Characteristic Variables)	Model 2 (Financial Variables)	Model 3 (Late Payment Variable)	Model 4 (Usage Variable)	Model 5 (Variation Variables)
Intercept	.2555 (.013***)	0.2239 (.013***)	.1634 (.012***)	.1256 (.012***)	0.105 (.013***)
Sex	-.0353 (.005***)	-0.0321 (.005***)	-.0199 (.005***)	-.0160 (.005***)	-0.0155 (.005***)
Education	-.0156 (.006**)	-.0088 (.006)	-.0016 (.006)	.0062 (.006)	0.0071 (.006)
Marriage	.0275 (.005***)	.0260 (.005***)	.0207 (.006***)	.0230 (.005***)	0.0235 (.005***)
Age	-.0004 (.0000)	-.0002 (.000)	.0002 (.000)	.0003 (.000)	0.0004 (.000)
Log average balance		.0135 (.001***)	-0.0058 (.001***)	-.0154 (.001***)	-0.0142 (.001***)
Log average payments		-.0303 (.001***)	-.0073 (.001***)	-.0037 (.001***)	-0.0059 (.001***)
payment balance ratio		.0001 (.0000)			
Average Delay			.2560 (.004***)	.2463 (.004***)	0.2424 (.004***)
Usage				.1226 (.008***)	0.1303 (.008**)
Cv Payments					0.0157 (.001***)
Cv Balance					0.0001 (.000*)
R^2	0.003	0.026	0.151	0.158	0.158
Adjusted R^2	0.003	0.026	0.151	0.157	0.158
Accuracy	0.49	0.54	0.71	0.76	0.77

Figure 3: OlS Regression Results Across Variable Group

To understand which variables contribute most to credit default prediction, we began by estimating five models, each grouped by a specific set of features. These included borrower characteristics (Model 1), financial status (Model 2), late payment behavior (Model 3), credit usage (Model 4), and overall variation across several financial behaviors (Model 5). By progressively adding more dynamic and behavior-based variables, we observed a clear pattern in model performance and explanatory power.

Model 1, which relied solely on demographic characteristics like sex, education, marriage, and age, provided minimal predictive power. Its R^2 was just 0.003, and accuracy hovered around 49%, barely better than random guessing. However, it serves as a good baseline to tell the impact on how much impact additional variables will have. Model 2 introduced financial averages such as log average balance and payments, resulting in a small performance improvement with an R^2 of 0.026 and accuracy increasing to 54%. However, these models still lacked the ability to capture more predictive behavioral dynamics.

Substantial gains emerged in Models 3 through 5, where behavioral variables such as average payment delay, usage, and credit variation metrics were introduced. Model 3, focusing on late payments, achieved a sharp jump in accuracy to 71%. Model 4, which emphasized credit usage behavior, performed even better with 76% accuracy, and Model 5 is the most comprehensive model. It achieved the highest accuracy at 77%.

The key takeaway from the model results is that the payment-to-balance ratio appears to provide little to no additional predictive value beyond the baseline characteristics. In contrast, the average delay, along with the coefficient of variation and usage variables, demonstrate strong predictive power when compared to traditional demographic and baseline financial variables. This suggests that behavioral and volatility-based measures may capture dimensions of credit risk that are not reflected in static borrower characteristics alone.

Logistic Regression Table Log-Odd Ratios	Model 1 (Characteristic Variables)	Model 2 (Financial Variables)	Model 3 (Late Payment Variable)	Model 4 (Usage Variable)	Model 5 (Variation Variables)
Intercept	-1.0656 (.074***)	-1.2477 (.077***)	-1.7060 (.084***)	-1.9462 (.086***)	-2.1289 (.093***)
Sex	-.2035 (.028***)	-0.1910 (.029***)	-0.1346 (.031***)	-.1133 (.031***)	-0.1082 (.031***)
Education	-.0890 (.0037**)	-.0572 (.037)	-.0084 (.040)	.0355 (.040)	0.0434 (.041)
Marriage	.1599 (.031***)	.1543 (.032***)	.1365 (.034***)	.1508 (.034***)	0.1560 (.034***)
Age	-.0021 (.0000)	-.0015 (.002)	-.0011 (.002)	.0021 (.002)	0.0024 (.002)
Log average balance		.0719 (.007***)	.0262 (.007***)	-.0733 (.008***)	-0.0644 (.008***)
Log average payments		-.1518 (.007***)	-.0612 (.008***)	-.0403 (.008***)	-0.0619 (.009***)
payment balance ratio		.0001 (.0000)			
Average Delay			.2560 (.004***)	1.3548 (.027***)	1.322 (.028***)
Usage				.6929 (.052***)	0.7663 (.054***)
Cv Balance					0.0000 (.000)
Cv payments					0.1369 (.027***)
Pesduo R^2	0.002	0.022	0.1293	0.1348	0.137
Brier Score	0.2213	0.2251	0.1931	0.192	0.1925
Accuracy	0.49	0.64	0.7	0.77	0.75

Figure 4: Logistic Regression Results Across Variable Group

To further analyze the predictors of credit default, we implemented five logistic regression models, each focusing on a different set of variables. These models estimate the log-odds of default and allow for direct interpretation of the effect size of each predictor. Model 1 included only borrower characteristics (sex, education, marriage, and age), while the subsequent models incorporated financial indicators (Model 2), late payment behavior (Model 3), usage data (Model 4), and broader financial variations (Model 5). As the models evolved from static demographics to behavior-based features, both predictive power and model accuracy improved significantly.

Model 1, relying solely on characteristics, performed poorly with a pseudo R^2 of 0.002 and an accuracy of just 49%. Although statistically significant, demographic variables such as sex and marital status only marginally improved the model's prediction power. Model 2 added financial averages like log average balance and payments, which increased the pseudo R^2 to 0.022 and accuracy to 64%. These results indicate that while financial status provides more useful signals than demographics, the model still fails to capture more dynamic behavioral patterns.

Performance notably improved in Models 3, 4, and 5. Model 3 incorporated average delay, a strong late-payment indicator, and achieved an accuracy of 70%. Model 4 further included credit usage behavior, showing that higher usage and average delay significantly increased the odds of default. This model reached the highest accuracy at 77% and had a lower Brier Score (0.192), indicating better calibrated probabilities. Model 5, which added variation metrics such as credit balance and payment fluctuations, achieved comparable results. Though slightly lower in performance, Model 5's added complexity did not yield meaningful gains over Model 4, suggesting potential overfitting. Ultimately, Model 4 emerged as the most efficient and interpretable, capturing key behavioral signals while maintaining strong generalizability.

Model 4 was selected as the preferred specification because it offers a strong balance between predictive performance and model interpretability. This balance of statistical robustness, behavioral insight, and predictive accuracy makes Model 4 the most suitable candidate for further application and comparison against more complex modeling techniques.

Metric	No Default	Default	Macro Avg	Weighted Avg
Precision	0.86	0.53	0.70	0.79
Recall	0.87	0.51	0.69	0.79
F1-Score	0.87	0.52	0.69	0.79
Support	4672	1328	6000	6000
Accuracy				**0.79**

Figure 5: Classification Metrics Summary – Random Forest Model Performance

The Random Forest model achieved a strong overall accuracy of 79%, reflecting its ability to distinguish between defaulters and non-defaulters with reasonable success. It performed especially well on the majority class (non-default), with high recall (0.87), indicating that the model is highly reliable in identifying borrowers who are likely to repay. However, its performance on the minority class (defaults) was notably weaker, with recall dropping to 0.51, suggesting

a substantial number of defaulters were misclassified. Despite this, the AUC-ROC of 0.77 indicates that the model still maintains a good overall balance between sensitivity and specificity.

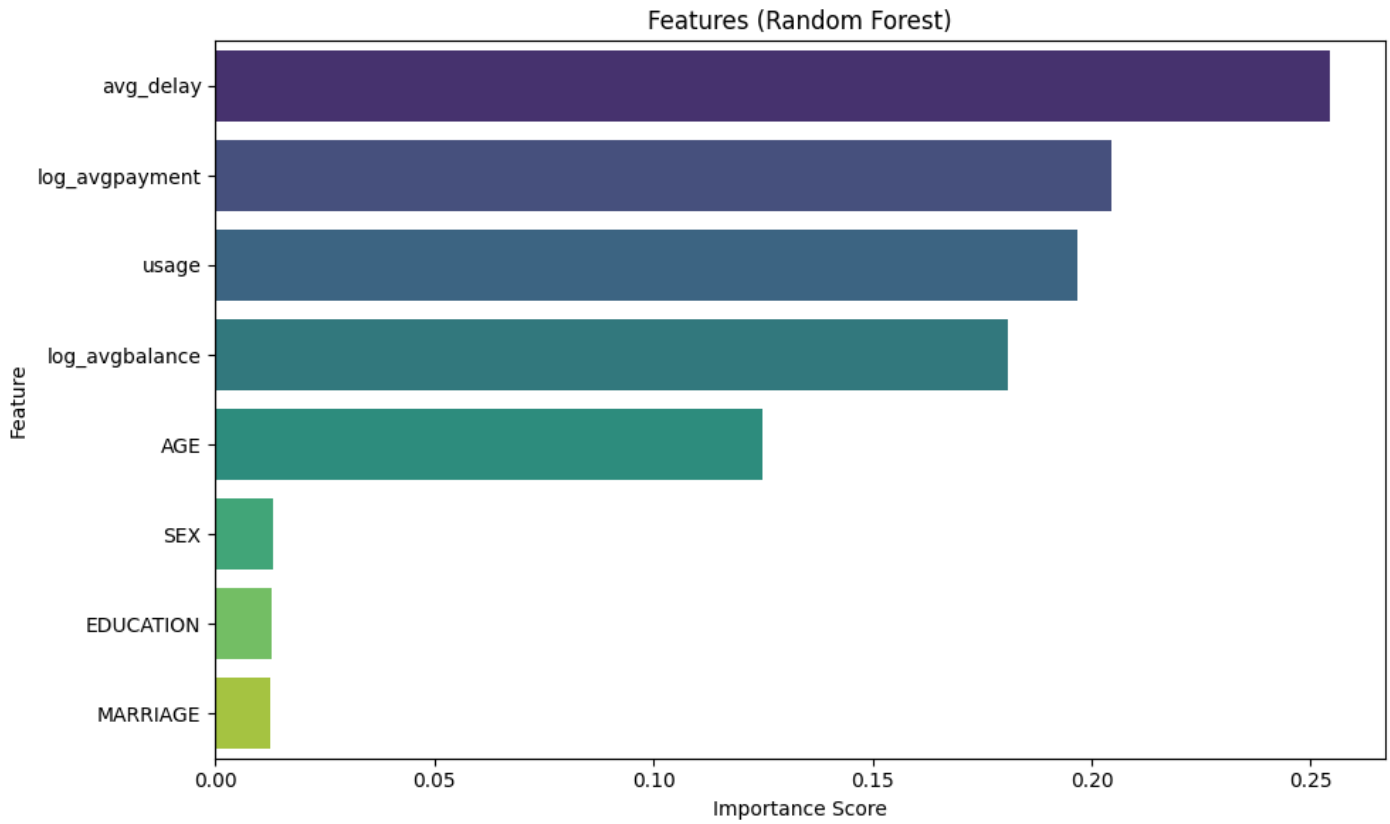

Figure 6: Features – Random Forest Model

To better understand the drivers of the model's predictions, we examined the feature importance scores generated by the Random Forest model. Feature importance measures how much each variable contributes to reducing prediction error across the ensemble of decision trees. For classification problems, importance scores are typically computed based on the average reduction in impurity achieved by splits involving a given variable, weighted by the number of observations those splits affect. Higher scores indicate variables that, when used for splitting, consistently improve the model's ability to correctly classify borrowers as default or non-default.

The most influential predictors in our Random Forest were behavioral: average payment delay, credit usage, average payments, and average balance consistently ranked at the top. These variables capture not only a borrower's financial capacity but also their real-time repayment habits, reinforcing our hypothesis that behavioral patterns are stronger indicators of default risk than static demographic or financial averages alone. Overall, the Random Forest model not only improved predictive performance but also highlighted the variables most relevant in practice, offering both a powerful predictive tool and valuable insight into credit risk dynamics. However, unlike logistic regression, Random Forest does not produce interpretable coefficients, limiting the ability to directly quantify the marginal effect of each predictor on default probability despite its high accuracy and robustness.

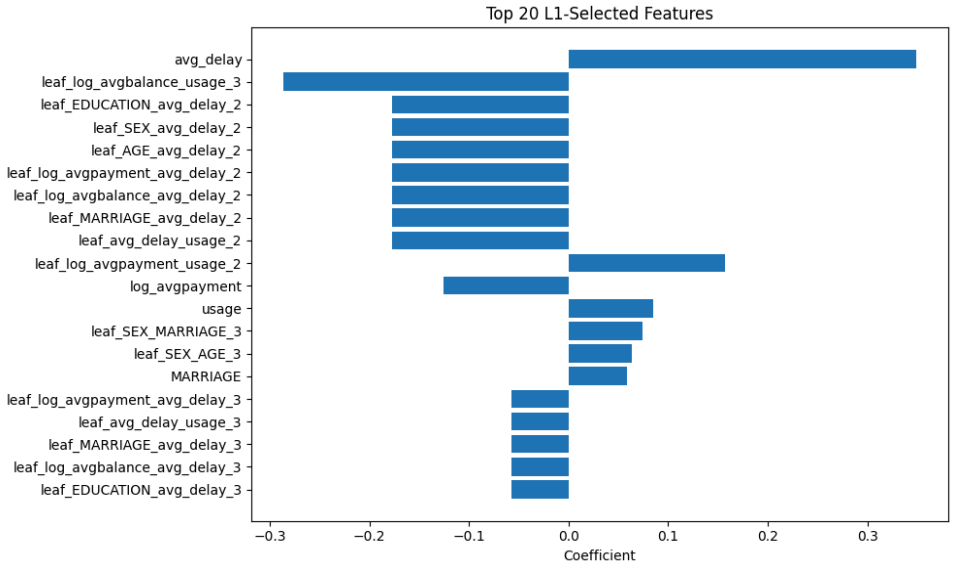

Figure 7: Top 20 Predictive Features – PLTR Model

Despite its conceptual appeal, our replicated PLTR model achieved an accuracy of 75% and an AUC of 0.71, both notably lower than the original paper's reported metrics (accuracy: 79%, AUC: 0.778). This performance gap may reflect challenges in replicating the precise penalty structure and tree segmentation logic described in the original study. However, the L1-regularized logistic regressions within each leaf did produce interpretable and sparse models. As shown in the figure above, the most influential predictor was avg_delay, followed by several interaction terms such as leaf_EDUCATION_avg_delay_2 and leaf_log_avgbalance_usage_3. These features reflect localized behavioral and demographic effects, showing that PLTR can still uncover meaningful structure even when performance metrics fall short of expectations.

Some of the features are combinations of other traits, like education and delay or balance and usage. These combinations help the model capture more specific behaviors. Such as: how delayed payments may affect different education levels or how balance levels interact with card usage. Not all important features increase the risk of default; some have negative values, meaning they actually reduce the risk. For instance, people who consistently make higher payments are less likely to default. Overall, the model suggests that both payment behavior and personal characteristics, especially when combined, play a major role in assessing credit risk.

Reflection

To assess the predictive performance of various modeling strategies in credit risk assessment, we conducted a comprehensive evaluation of three main approaches: Logistic Regression, Random Forest, and Penalized Logistic Tree Regression (PLTR). Each model was examined under both the original study's implementation and our replication. The goal of this section is to compare their effectiveness using key performance metrics such as AUC-ROC and accuracy, while also reflecting on their strengths, limitations, and suitability for real-world deployment. The following analysis

highlights the nuanced trade-offs between simplicity, interpretability, and predictive accuracy across these competing methods.

Model Performance Comparison

	AUC-ROC	Accuracy
Logistic Regression (Original Paper)	0.5963	0.7035
Logistic Regression (Replication)	0.771	0.77
Random Forest (Original Paper)	0.7722	0.8102
Random Forest (Replication)	0.7654	0.79
PLTR (Original Paper)	0.778	0.79
PLTR (Replication)	0.7117	0.75

Figure 8: Model Performance Comparison

We evaluated both the original results published in the paper and our replication performance, using key metrics such as AUC-ROC and Accuracy in our version. The goal is to assess how each model performs in terms of discrimination, calibration, and generalizability.

In the original paper, Non-linear Logistic Regression performed poorly, with an AUC-ROC of 0.5963 and accuracy of 70.35%, indicating limited predictive power. However, our replication significantly improved these results. By applying better preprocessing techniques including categorical encoding, log transformations for skewed variables, and new behavioral variables, we achieved an AUC of 0.7710 and 77% accuracy. This demonstrates that while Logistic Regression lacks flexibility in modeling non-linear interactions, it can still deliver competitive results when combined with thoughtful data preparation.

Random Forest delivered the highest accuracy of all models. In the original paper, it reached an AUC of 0.7722 and an accuracy of 81.02%. Our replication came very close, with an AUC of 0.7654 and accuracy of 79%. Despite minor variance, both results confirm Random Forest's strong generalizability and ability to model complex, non-linear relationships with minimal tuning. We optimized key hyperparameters such as the number of estimators (300), minimum leaf size (2), and feature selection strategy (sqrt) , and estimated optimal parameters using grid search to enhance model stability. These results make Random Forest the most reliable choice in terms of predictive power.

PLTR, a hybrid model combining tree partitioning and logistic regression with regularization, achieved the highest AUC-ROC of 0.7780 in the original study, along with a solid 79% accuracy. This made it especially attractive for applications requiring both accuracy and interpretability. However, our replication showed a noticeable drop in performance, with an AUC of 0.7117 and accuracy of 75%. This gap is likely due to differences in implementation, such as how leaves were defined or how penalized logistic regressions were fit to each subgroup. Unlike standard models, PLTR is highly sensitive to model design and regularization settings, which makes replication more difficult

Conclusion

Overall, our results closely replicate the core findings of the original study. Random Forest emerged as the most consistently strong model, achieving high accuracy and AUC with minimal sensitivity to preprocessing. PLTR stood out for its balance between interpretability and performance, though its complex structure made replication more challenging. Logistic Regression, while simple, benefited greatly from data transformation and rebalancing strategies. Together, these results highlight the trade-offs between model transparency, tuning complexity, and predictive power in credit scoring applications.

A key consideration in model selection for credit risk prediction is the trade-off between accuracy and interpretability. In our analysis, all models performed reasonably well, showing solid predictive power across the board. However, Random Forest consistently outperformed the others, achieving the highest accuracy (79%) and AUC-ROC in our replication. Its strength lies in capturing complex, non-linear interactions, but this comes at the cost of transparency, as it functions largely as a black box in regards to marginal effects

In contrast, both Logistic Regression and PLTR offered greater interpretability. Logistic models provide globally interpretable coefficients, while PLTR adds a layer of flexibility by allowing localized, leaf-level decision logic. Although PLTR did not match the original study's performance in our replication, it still represents a promising middle ground between traditional econometrics and modern machine learning.

---

8. References and Tables

academic references, original source papers, and methodologies.

Dumitrescu, E.-I., Hué, S., Hurlin, C., & Tokpavi, S. (2021). Machine learning or econometrics for credit scoring: Let's get the best of both worlds. *SSRN Electronic Journal.* https://ssrn.com/abstract=3553781 or http://dx.doi.org/10.2139/ssrn.3553781

Zhang, J., Jiang, C., Qiu, Y., & Dong, J. (2023).
Hybrid credit scoring model based on interpretable machine learning and statistical learning. *Expert Systems with Applications, 213*, 118991.
https://doi.org/10.1016/j.eswa.2022.118991
→ https://www.sciencedirect.com/science/article/abs/pii/S0957417422017511

Library of Economics and Liberty. (n.d.).
Bonds. *Econlib*.
Retrieved from https://www.econlib.org/library/Topics/Details/bonds.html

Board of Governors of the Federal Reserve System. (2011, April 5).
Supervisory Guidance on Model Risk Management (SR 11-7).
https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm

Autorité de Contrôle Prudentiel et de Résolution (ACPR). (2020, June).
Governance of Artificial Intelligence in Finance.
https://acpr.banque-france.fr/system/files/import/acpr/medias/documents/20200612_ai_governance_finance.pdf

European Banking Authority (EBA). (2020, May 29).
Guidelines on loan origination and monitoring (EBA/GL/2020/06).

https://www.eba.europa.eu/sites/default/files/document_library/Publications/Guidelines/2020/Guidelines%20on%20loan%20origination%20and%20monitoring/884283/EBA%20GL%202020%2006%20Final%20Report%20on%20GL%20on%20loan%20origination%20and%20monitoring.pdf

Basel Committee on Banking Supervision (BCBS). (2023, April).
Principles for the effective management and supervision of climate-related financial risks (BCBS d595).
https://www.bis.org/bcbs/publ/d595.pdf

- Include **appendix tables**: summary statistics, performance metrics, variable descriptions, model specifications.

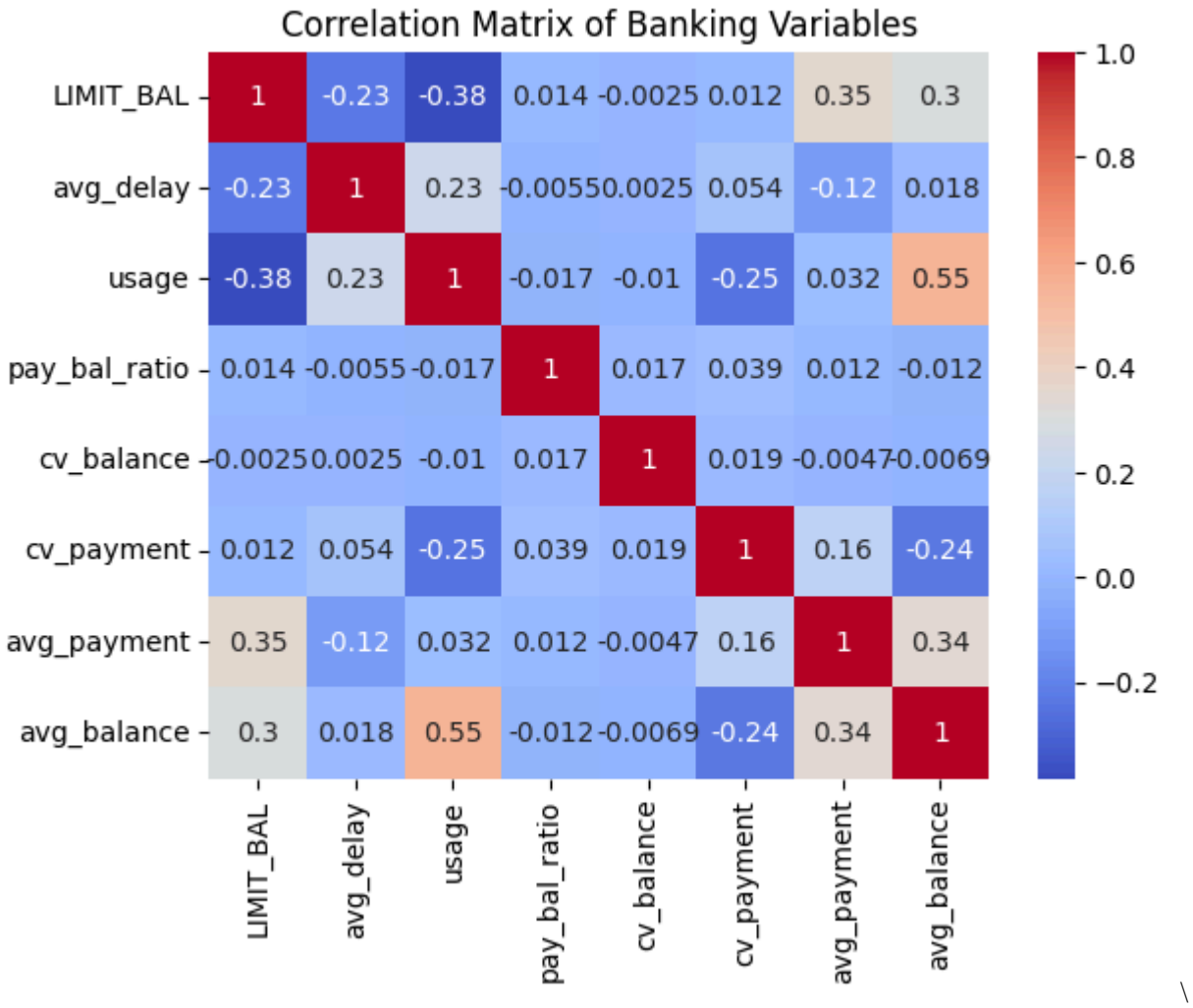

Figure: Correlation Matrix