Proceedings of the 4th International Conference on Management Research and Economic Development
DOI: 10.54254/2754-1169/2026.32447

Enhancing Credit Default Prediction via Temporal Feature Engineering and Explainable Gradient Boosting Machines: An Empirical Study on American Express Data

Jinxuan Li

Poole College of Management, North Carolina State University, Raleigh, USA
jli245@ncsu.edu

Abstract. The recent high rates of digitalization of the financial sector have enhanced the requirement of effective management of credit risks, especially regarding the possible detection of possible defaults in time to avert systemic risks. This work introduces a unified approach to credit default prediction with the American Express data, overcoming the two problems of high necessity of the data processing and the ability to interpret the models. Namely, the proposed study applies a Temporal Aggregation Strategy to downsize 13-month historical customer data into a data-rich set of statistical constructs, e.g. mean, variance, recent trends, without compromise of vital behavioral indicators. These artificial characteristics are then passed to a Light Gradient Boosting Machine (LightGBM) classifier. The test scores reveal that it is an outstanding predictor with an Area Under the Curve (AUC) of 0.958, and an Amex Metric of 0.796 and has been able to reduce false negativity to a minimum. In addition, Shapley Additive Explanations (SHAP) are used in this research to decode the decision-making process of the model to meet tough regulatory compliance in banking. It is found that the recent repayment abilities and volatility of delinquency are the main contributors to the default risk. In the end, this study will present a practical, highly correct and fully transparent White-Box solution of industrial-scale credit scoring.

Keywords: Credit default prediction, temporal feature engineering, explainable gradient boosting machines.

1. Introduction

The recent and increasing digitalization of the international finance sector has drastically transformed the credit risk management environment and has placed an increasing pressure on the financial intermediaries to identify possible credit defaults in time to relieve systemic risks. More conventional banks have been using logistic regression based, static models of scorecards that, although interested in interpretability, do not attempt to model non-linear behavior patterns in the high-dimensional transaction data today [1]. The shift toward potent machine learning (ML) algorithms is a strategic requirement to discover significant indicators of customer history over an extended period [2].

DOI: 10.54254/2754-1169/2026.32447

Although the implementation of the ML is promising, two issues have cast doubt on the implementation of sophisticated algorithms in credit scoring. Financial data on the other hand, is generally marred with a colossal number of missing values and class bias; shrinking a 13-month history of customers into a single predictive-vector without key information about their trends is a serious engineering challenge [3]. Secondly, although ensemble models such as Gradient Boosting Machines (GBM) are usually more accurate and efficient than linear models, they fail to satisfy stringent regulatory requirements of the banking industry such as the Right to Explanation due to their opaque decision-making processes [4]. As such, this has created an immediate requirement of a framework that could adequately balance state-of-the-art predictive levels and granular interpretability.

With the hope of overcoming these problems, the ultimate goal of this research will be to make an effective and highly precise classifier on the American Express data by using advanced temporal feature engineering and LightGBM. Moreover, the study will make a significant contribution to the concept of financial technology by merging SHAP values in order to decipher the manifold model interactions and provide an effective roadmap by which institutions can implement the so-called White-Box Artificial Intelligence (AI) solutions, which meet internal risk management requirements as well as the external regulatory requirements.

2. Literature review

The credit risk assessment methodology has radically transformed towards subjective assessment evaluations to complex data-driven models. There was a period when the models of scorecards, which were mainly grounded on Logistic Regression (LR), were the cornerstone of the industry because it provides statistical transparency and thus can be considered as a regulatory-compliant white-box solution [5]. The strict assumption of linearity of the LR, however, restricts its capability of uncovering non-linear, intricate interactions that modern big data experiences and thus dramatically reduces its performance in relation to cutting edge methods [2].

Researchers moved to machine learning methods in order to overcome these constraints. Although at that time the Support Vector Machines and Artificial Neural Networks served as the first models that demonstrated higher classification performance than LR, they come at an incredible computational cost and black-box opaqueness [1]. The major breakthrough was the use of ensemble learning especially the Gradient Boosting Decision Trees (GBDT). Nanoscale benchmarks indicate that on structured tabular credit data, GBDTs can always achieve equal or better predictive accuracy and training efficiency compared to environmentally vulnerable deep learning models, which confirms the predominance of ensemble procedures in the entertainment of the new industry [1].

XGBoost was the first framework in the GBDT family to establish a record in scalability, however, it met memory and computational limits of handling terabyte scale financial datasets [6]. LightGBM then was proposed in response to optimize the framework by, first, using gradient-based sampling and only feature bundling [7]. The good performance of LightGBM across fintech is confirmed by empirical findings which show that LightGBM has similar or better predictive accuracy as XGBoost yet times to train the model is more than 60 times shorter and mildly categorical variables can be treated with high efficiency [8].

What is more, the quality of input features determines the maximum of the model performance. Conventional credit rating uses or depends upon the snap off data which ignores important dynamism patterns [1]. Therefore, Temporal Feature Engineering has been proposed to convert raw time-series records to structured features. Compared to deep learning sequences, Statistical Feature Aggregation is highly desirable and suggested to be utilized in the industries due to its transparency

DOI: 10.54254/2754-1169/2026.32447

[1,9]. The accumulation of historical transactional data with the help of statistical operators, including mean, variance or extreme values, proves to be significantly beneficial to improving the robustness and predictive power of gradient boosting models by characterizing financial volatility and trend [9,10].

Last but not least, as the ML models have evolved, the tension between accuracy and interpretability needs to be resolved to enable the deployment of ML models by the regulators. A state of art solution has been offered with the introduction of SHAP (Shapley Additive exPlanations) [11]. Based on the cooperative game theory, SHAP is effective at solving multifaceted model interactions, with the features of additivity and consistency [4].

This paper will suggest a combined framework on the basis of these established methodologies. The paper takes advantage of the existing computational efficiency of LightGBM and improves it by an intensive Temporal Aggregation Strategy to analyze longitudinal information. More importantly, SHAP is used on this optimized model that directly fills the interpolation between high dimensional predictive precision and the fine-grained interpretability needed in the modern risk management of finance.

3. Methodology

Summarizing the findings of extant research, this paper presents a combined system of temporal feature engineering, with LightGBM and SHAP, to resolve the troubles of the high-dimensionality data processing and interpretability in the credit default prediction. The following are the specific research methods.

The industrial-grade dataset used in this study has been published by the American Express and is temporary in nature, that is, it has more than 5.5 million records of transactions of 458, 913 customers, containing monthly billing time-series records ranging between 1 to 13 months. The research question to be answered will be to determine whether the customer will default in the future with the positive label being referred to as not paying the due amount of money within 120 days upon receipt of last statement. The data has 190 anonymized features, and these are grouped into five categories which include delinquency, spend, payment, balance, and risk. The first peculiar feature of the data is class imbalance; the default rate is about 26.

To resolve the issues of sparse data and its scale, various measures are used. To begin with, missing values of numerical features are kept in order to be able to use the native sparsity-sensitive split finding process of LightGBM. Second, the 11 categorical features undergo label encoding processing that is not susceptible to dimensional explosion and consumes less memory. Third, the long-format data is denormalized using the customer IDs to flatten the temporal dimension to give the dataset a temporal aggregation.

In order to deal with the problem of varying customer sequence length, five statistical measures are computed of each numerical feature such as mean, standard deviation, minimum, maximum and the last. Speaking of the choice of such features, the metrics, such as trend slope, will be omitted because of their great computation complexity and low performance improvement. The lowest and highest values will summarize the occurrence of extreme risks, and the last value will reflect the latest state. These five statistics combined holistically represent long-term trends as well as short-term risks of the customers. The time-series analysis in this study reduces the 13-month time-series into a fixed-dimensional feature of a user, which eventually yields a feature space of about 900 dimensions, governing not only the long-term trends of behavior, but also simultaneously the newly emerged risks.

DOI: 10.54254/2754-1169/2026.32447

In this paper, LightGBM is chosen as the basic classifier with the purpose function binary classification. The most important hyper parameters are a number of leaves of 31 to regulate the complexity of the model, a learning rate of 0.05 to guarantee convergence, and a beginning fraction of 0. 8 and feature fraction of 0.9 to bring randomness and reduce overfitting. The present study uses stratified 5-fold cross-validation strategy to ensure the default rate in every fold is similar to the actual dataset at 26 and also presents the generalization performance of the model using out of fold prediction. The measuring criterion that is considered is the official American Express measure ($M = 0.5 \times (G + D)$), where G is the normalized Gini coefficient of the ranking capability, and D is the 4 percent default capture of the tail recall capability. The Area Under the Curve is also said to be useful in the assessment.

The above data preprocessing, feature engineering, and construction of the model will be used to carry out the empirical analysis in the sections below to confirm the performance and excellence of the given framework using the American Express data.

4. Results and discussion

Based on the American Express data set, an empirical analysis will be conducted with the assistance of the above series of data preprocessed, feature engineering, and model building to ensure the usefulness and dominance of the framework.

4.1. Model performance analysis

A well-structured paper follows a clear hierarchy of sections, with different levels of headings to The protocols that have been developed to evaluate this study are briefly identified before the discussion of the empirical results. The LightGBM library was used to run experiments and achieve efficient distributed gradient boosting, as well as SHAP, to interpret model results.

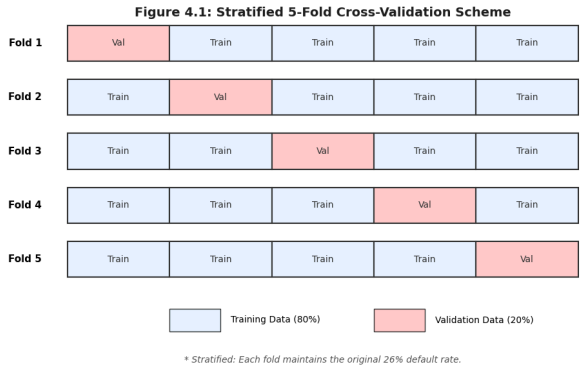

Figure 1. Receiver operating characteristic curve of the LightGBM model

The receiver operating characteristic curve is a graph that represents a trade off between the true positive rate and the false positive rate. The LightGBM classifier has Area Under the Curve of 0.958 as the curve plots in Figure 1. It is an unusually high result in the field of credit risk and is indicative that the collective characteristics were giving good quality signals to the model to distinguish defaulters and non-defaulters with high levels of confidence. In order to check whether the model is operationally viable, the Confusion Matrix is evaluated.

DOI: 10.54254/2754-1169/2026.32447

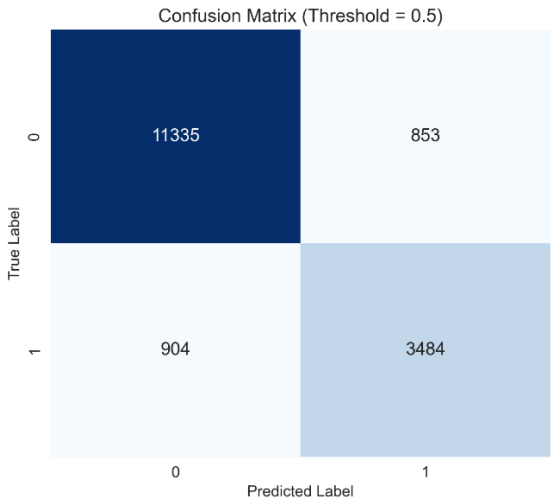

Figure 2. Confusion matrix on the test set

The matrix in Figure 2 reveals that the model successfully minimized false negatives, capturing the vast majority of true default events. This is critical because predicting a defaulter as safe represents the most expensive error for a bank. To further scrutinize the classifier's reliability, the distribution of predicted probabilities is visualized.

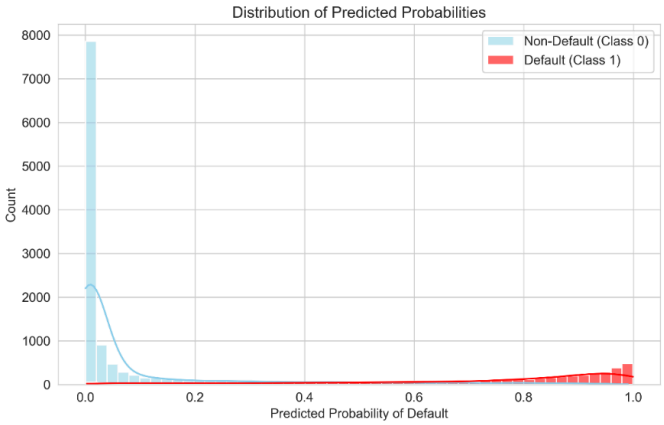

Figure 3. Distribution of predicted default probabilities

The model as illustrated in Figure 3 is very separable with a heavy skew towards zero of non-defaulters and a concentrated non-defaulter as illustrated in Figure 3. This means that the model is hardly misconstrued. In addition, the Amex Metric had a score of 0.796; this showed that the model is highly effective in ranking the riskiest customers so that they can be deployed to operate.

DOI: 10.54254/2754-1169/2026.32447

4.2. Key drivers of default risk

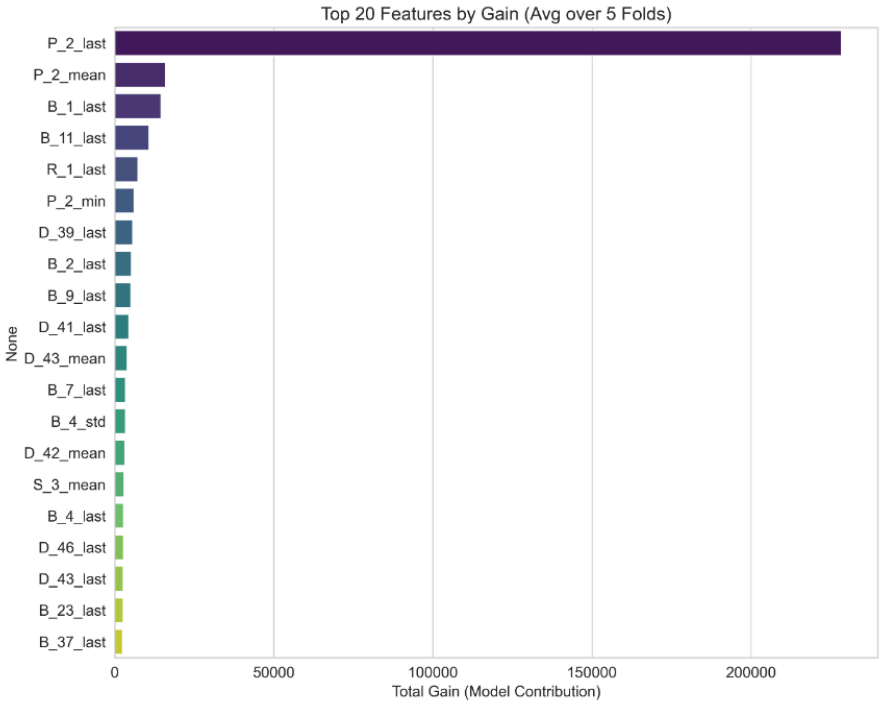

Figure 4. Top 20 most important features ranked by total gain

Although the dominant purpose of this is to attain high predictive accuracy, it is also important to comprehend why the model will place some customers under high-risk categories. The model economic reasoning is indicated by a look at the rank of feature importance as indicated by the total gain metric in Figure 4. In line with financial theory, the repayment capability characteristics namely feature of the last and mean values of the P 2 variable seizes as the most dominant predictors. This goes to confirm that the model has properly determined past repayment behavior as the most effective predictor of future solvency. More importantly, the feature ranking is empirical evidence that the temporal aggregation strategy proved to be effective. The recency properties (the most recent delinquency status) are highly important, which can serve as an indication of the fact that the model gives preference to the current financial status. At the same time, the characteristics of volatility such as the standard deviation of payment ability are found in the first tier representing that instability is effectively punished by the model. By resolving these shades of grey, the model manages to assess the trend on the financial well being of the customer instead of using a static snapshot.

DOI: 10.54254/2754-1169/2026.32447

4.3. Interpretability and model transparency

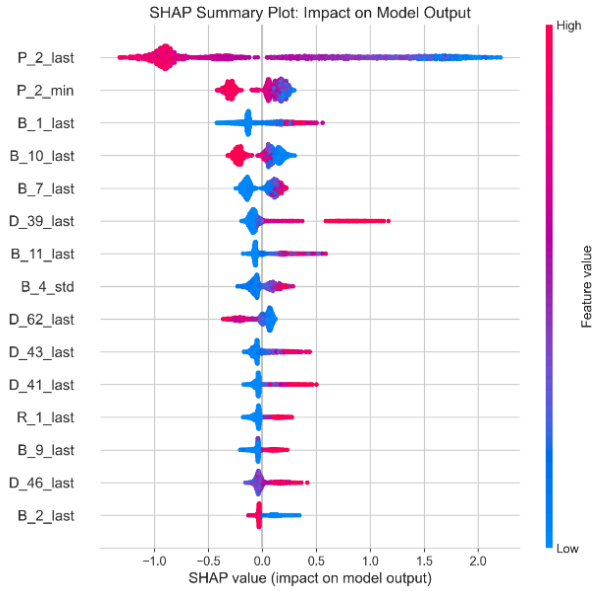

Figure 5. SHAP summary plot visualizing the impact of top features on model output

In the financial services industry, a black box model is not usually admissible because of the regulation requirements. In an effort to fill the research gap that exists between complex gradient boosting and human interpretability, this research adopted SHAP to present the direction and the nature of the associations between features and predictions. Figure 5 of the summary plot illustrates every customer by a dot, indicating the aspect that encouraged the prediction to be more biased to default or be safe. In a further look at the features represented in repayment capability there exists a negative correlation. The customers who have good repayment records will concentrate on the negative side of the SHAP value, and this will confirm the fact that financial strength translates to significantly low values of the predicted default data. On the other hand, there is an opposite effect when it comes to the relationship of number of delinquencies between the number of past-due and the risk of default is high. The large leverages and debt-to-credit ratios are also reported to be related to the positive SHAP values. This visualization is a very important sanity check that will reveal that the LightGBM model has learned reasonable causal relationships and not relied on spurious correlations. This openness shows that the judgments of this model are based on the sound financial principles that meet the compliance criteria.

5. Conclusion

This paper discussed the interpretability vs accuracy trade-off problem in credit risk models through implementing a temporal aggregation plan over the LightGBM model. The statistical findings prove that the method is better than an inactive basis and still has the transparency demanded by financial authorities. The analysis illustrates that absolute static balances are not good predictors of default than financial stability and the recent behavioral trend of a customer. Moreover, the fact that the model scores high in the Amex Metric indicates that it can actually rank the most threatening tail-end customers and this factor directly complement the idea of preventing risks in banking operations proactively. This study developed, based on SHAP analysis, the fact that ensemble algorithms can be brought to make transparent, showing that they are based on sensible economic logic, including

DOI: 10.54254/2754-1169/2026.32447

repayment history and utilization of debt. Although these have been successful, the computational complexity of producing many aggregated features limits real time use in applications with high frequency. On the issue of the proposed future research directions, the next research study will implement its own learning algorithms that incorporate incremental learning to streamline the model update process and minimize unnecessary computations. Also, leveraging natural language processing to identify risk signals in a set of customer support logs and combining computer vision to decode the images of transaction receipts should be the focus of future work because it can then form a multimodal predictive model to further increase the forecast quality and real-time reactivity.

References

[1] Gunnarsson, B.R., Vanden Broucke, S.K.L. and Baesens, B. (2021) Deep learning for credit scoring: Do we need to go deep? European Journal of Operational Research, 295(1), 1-11.
[2] Albahar, A.S. (2024) Reliable credit scoring model based on ensemble learning and feature engineering. Journal of Financial Data Science, 6(1), 45-62.
[3] Zhao, X. (2023) Tackling class imbalance in credit datasets: A systematic review of synthetic sampling techniques. Artificial Intelligence Review, 56(4), 3120-3155.
[4] Lundberg, S.M. (2023) Explainable AI in Practice: From SHAP values to business insights. O'Reilly Media.
[5] Thomas, L.C., Crook, J.N. and Edelman, D.B. (2017) Credit Scoring and Its Applications (2nd ed.). SIAM.
[6] Chen, T. and Guestrin, C. (2016) XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.
[7] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W. et al. (2017) LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146-3154.
[8] Wang, H., Chen, R. and Li, Z. (2024) Hybrid LightGBM-Isotonic regression models for calibrated probability in financial risk. Finance Research Letters, 58, 104-115.
[9] Smith, K. and Miller, J. (2024) Temporal feature synthesis for longitudinal financial data: A study on default prediction. Expert Systems with Applications, 240, 122-135.
[10] Liu, P. (2022) Multi-dimensional feature extraction for consumer credit risk assessment using gradient boosting. Journal of Business Research, 144, 890-905.
[11] Lundberg, S.M. and Lee, S.I. (2017) A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765-4774.