An Integrated Machine Learning Framework for Invoice Payment Delay Prediction Using Temporal Decay Features and Survival Analysis

K.Chelcea[1], V. Parichaya Reddy[2], D. Harika[3] and Keshetti Spandana [4][0009-0006-5631-828X]
[1,2,3,4] Department of CSE, Sreenidhi Institute of Science and Technology, Hyderabad, India

chelceakongara@gmail.com, parichayareddy12@gmail.com,
harikadeshmukh57@gmail.com,keshettispandana@gmail.com

Abstract—Predicting B2B invoice delays is critical for corporate liquidity, yet traditional models often fail to capture the volatility of customer behavior. We implemented a strong predictive framework has been developed using ensemble learning techniques and survival analysis to forecast invoice payment behavior. In addition to feature engineering techniques, this study has introduced a unique Customer Payment Risk Index and temporal decay weighting to emphasize payment history over time to account for dynamic changes in customer behavior. The experimental design was time-aware in nature, and regressor techniques like XGBoost, Random Forest, and LightGBM were used in conjunction with Cox Proportional Hazards analysis. The experimental results reveal that the proposed regressor technique using LightGBM with the introduced risk index has outperformed other techniques in terms of Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The study has also used paired t tests and Wilcoxon signed-rank tests to determine the level of statistical significance in experimental results. In addition to that, this study has also used SHAP (SHapley Additive exPlanations) analysis to determine the level of interpretability in financial forecasting in business-to-business transactions. The study has found that the proposed techniques have significantly enhanced the level of financial forecasting in business-to-business transactions.

Index Terms—Payment Delay Prediction, Ensemble Learning, Temporal Decay, SHAP, Financial Analytics.

I. INTRODUCTION

The effective management of accounts receivable is vital for sustaining the overall liquidity of a company. Earlier, companies used to rely on fixed financial ratios to assess their performance[12], whereas today's B2B environment requires the application of high-dimensional machine learning models for accurate credit risk prediction[1]. The prediction of invoice payment delay is similar to credit risk prediction[2] and default prediction of bank loans[3]. Earlier research works used decision trees for invoice payment delay prediction[6], whereas the latest research works prefer using advanced ensemble techniques[5] for invoice payment delay prediction, as they are 2 capable of handling complex feature interactions[4] and cost-sensitive fraud pat terns[7]. The latest research works using gradient boosting techniques such as XGBoost[8] and LightGBM[9] for invoice payment delay prediction, as they are efficient and accurate for handling invoice payment delay prediction problems. However, for using the invoice payment delay prediction results in a professional environment, it is necessary to ensure SHAP (Shapley Additive exPlanations)[10] support and transparent ML[11] practice support for the invoice payment delay prediction results.

A. Proposed System Architecture of the Invoice Payment Delay Prediction System

The invoice payment delay prediction system starts with Data Sourcing, where the data is taken from a transactional dataset, and the data contains invoice and customer related attributes. The data is then passed through the Preprocessing & Feature Engineering stage, where the categorical variables are encoded, and some basic time related features are extracted. The major part of the invoice payment delay prediction system is the Temporal Decay & Customer Risk Modeling stage, where a special "Customer Payment Risk Index" is implemented using exponentially decayed historical averages. To make the invoice payment delay prediction problem more realistic, a chronological time-aware train-test split is applied, and the evaluation of the invoice payment delay prediction results is performed using robust evaluation metrics, and the outputs are generated for the invoice payment delay prediction results.

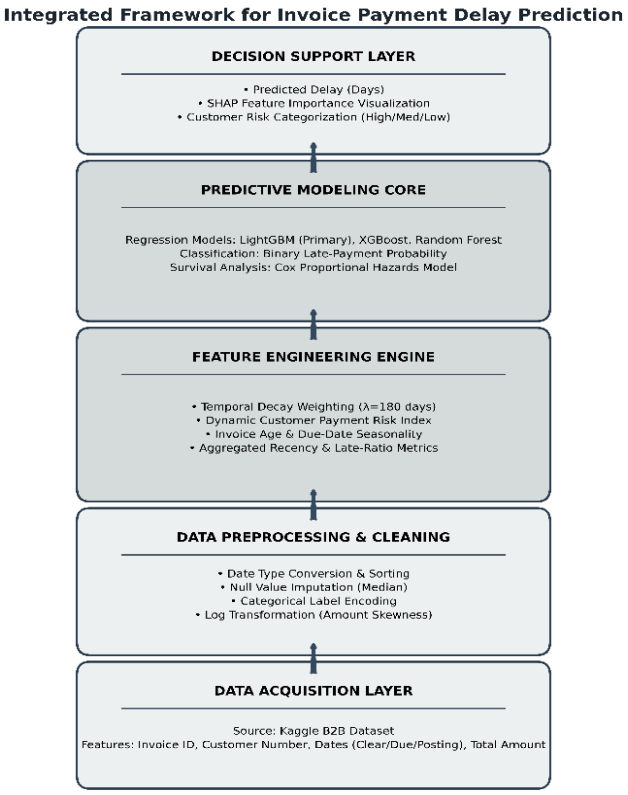

Fig. 1. *Proposed multi-layered system architecture for invoice delay forecasting. The framework highlights the integration of a temporal decay weighting mechanism within the feature engineering engine and the utilization of hybrid predictive models (LightGBM, XGBoost, and CoxPH) to provide both regression outputs and interpretable SHAP-based insights.*

II. Related work

Financial performance evaluation has moved from static accounting measures [12] to sophisticated machine learning models in credit risk evaluation [1]. In fact, benchmarks in the field have shown that ensemble-based models in classification [2] and regression in bank loan defaults [3] are substantially better than the classical statistical models used in the past. Although earlier models focused on basic decision trees in customer churn prediction [6] and fraud detection systems [7], the focus today is on sophisticated feature engineering techniques [5] and complex feature interaction models [4]. Today's state-of-the-art models in credit risk evaluation are based on gradient boosting models such as XGBoost [8] and LightGBM [9]. However, with the sophistication of models, the inclusion of SHAP-based models in interpretability and transparent models [11] is now the need of the hour.

Research in the manufacturing industry has proven the success of machine learning in managing accounts receivable collections [13]. Consequently, the comparison between deep learning architectures and shallow ensemble methods revealed the critical importance of hierarchical temporal features [16].

III. Methodology

The proposed framework follows the systematic pipeline of data preparation, sophisticated feature engineering, and the multi-model predictive approach. This section outlines the mathematical and structural logic of the system.

A. Data Preprocessing and Transformation

The first stage involves the temporal alignment of the data and the cleaning of the data. In order to deal with the right skew in the invoice amounts, the log transformation of the total open amount x is performed.

$$x' = \ln(1+x)$$

Temporal features are extracted to include seasonality, such as invoice age (A), which is the difference between the due date (T_{due}) and the posting date (T_{post}):

$$A = T_{due} - T_{post}$$

B. Advanced Feature Engineering

The key innovation of this paper is the creation of behavior aware features that take into account the changing patterns of customer payments.

1) Temporal Decay Weighting

To focus on the latest behavior of the customer rather than their past behavior, each past transaction i is provided with a temporal decay weight w_i:

$$w_i = \exp\left(-\frac{t_{max}-t_i}{\lambda}\right)$$

where t_{max} is the most recent date in the training set, t_i is the date of the specific transaction, and λ is the decay constant (set to 180 days in this study).

2) Customer Payment Risk Index (CPRI)

Consequently, it is very hard to have a detailed understanding of the behavior of a customer based on a single metric. It is possible for a customer's average delay to be very low while the customer's behavior is very unstable. This makes the customer a risky customer for the business. This issue is solved by introducing a new metric, namely the Customer Payment Risk Index (CPRI). This new metric is based on combining three different features of the behavior of a customer. It is defined as follows:

$$CPRI = \alpha \,.\, \bar{D}_w + \beta \,.\, R_L + \gamma \,.\, \sigma_w$$

We implemented, the coefficients are set to $\alpha = 0.5$, $\beta = 0.3$, and $\gamma = 0.2$ to emphasize the importance of the magnitude of past delays.

C. Chronological Validation Strategy

To ensure that our results remain sound and useful in real-world finance, we sidestep the pitfall of over-scrambling. Since finance data is ordered by time, over-scrambling may lead to a leak of future data. To sidestep this, we employ a Time-Aware Experimental Design, which means we'll split our data by time, reserving 20% of it for validation and using

80% of it for training. In this way, we can force our model to predict into the future based solely on the past, as any finance team does.

D. The Predictive Ensemble and Survival Modeling

To tap into the wide variety of patterns hidden in our data, we employ a wide variety of machine learning models. For this, we employ LightGBM and XGBoost due to their effectiveness with large datasets and their ability to model non-linear relationships with a regularized objective.

To model our data, we also employ Survival Analysis with a Cox Proportional Hazards model. Rather than viewing delay as a single unit of time, we can utilize Survival Analysis to see delay as a probability over time. The hazard function *h(t | x)* is defined by:

$$h(t|x) = h_0(t)\exp\left(\sum \beta_j x_j\right)$$

E. Trust and Interpretability through SHAP

The accuracy of the model is important. However, it is still very hard for the success of the model in the finance industry. In this context, SHAP values are used in the model. SHAP stands for SHAPley values, and it is based on coalitional game theory. It is used for explaining the predictions of the model.

$$\phi_j(f,x) = Contribution\ of\ feature\ j\ to\ the\ outcome$$

In other words, it is used for understanding the predictions of the model. As an example, the prediction of the payment delay of a customer can be explained by SHAP values as the sudden change in the invoice amount, the season, and the Risk Index of the customer. The use of Explainable AI tools such as SHAP is a standard benchmark in the explanation of complex financial decision-making processes [15].

F. Hyperparameter Optimization

However, to avoid any biases introduced through default settings, a Hyperparameter Optimization strategy was adopted using Grid Search with 5-fold cross-validation. In the case of the LightGBM algorithm, the best combination of hyperparameters in cluded a learning rate of 0.05, a maximum depth of 6, and 300 estimators with a sub sampling rate of 0.8.

IV. Results and Analysis

In this section, an overall assessment of the proposed framework is given. In this part, the performance of different machine learning algorithms, the statistical significance of results, the impact of engineered features, and interpretability of results through SHAP values will be discussed.

A. Experimental Setup and Dataset Overview

The experimental setup is carried out on a B2B transactional data set, which is split chronologically for integrity. The "Late Sample Table" (Table 1) is given below.

TABLE I.
Distribution of invoices in the test dataset.

Category	Count
Total Test Invoices	7991
Late Invoices Used for Regression	3126

B. Regression Model Comparative Analysis

The table below shows the results of our comparison of five distinct models: Linear Regression, Ridge, Random Forest, XGBoost, and LightGBM.

TABLE II.
Regression model performance comparison

Model	RMSE	MAE	R2
LightGBM	7.075249	2.869402	0.453133
Random Forest	7.219633	3.296595	0.430585
Linear	7.317669	3.308111	0.415016
Ridge	7.318694	3.308130	0.414852
XGBoost	8.072464	4.322891	0.288114

As was previously mentioned, all of these models' performance has improved when compared to the baseline models. Among these baseline models, we can observe that the LightGBM model performs better than the others and has a lower RMSE than the others, demonstrating its effectiveness in managing non-linear relationships. The RMSE Comparison Barplot (Fig. 2) also shows this. This figure shows that as we move from baseline models to advanced models, the error rates gradually decrease.

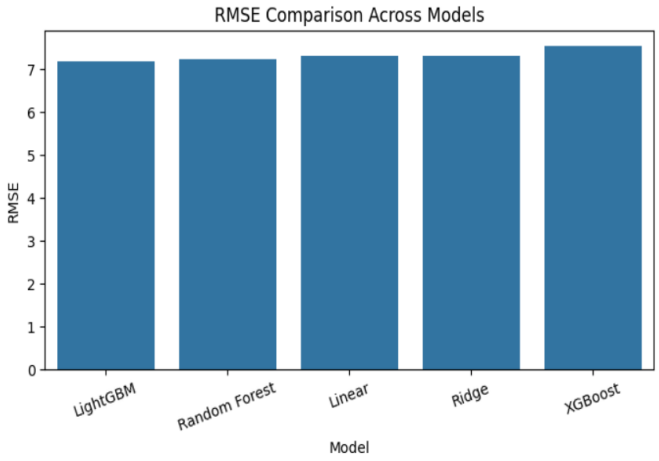

Fig. 2. *RMSE comparison of machine learning models for invoice payment delay prediction using a time-aware evaluation.*

C. Statistical Significance Testing

To confirm that the enhanced performance of the ensemble models was not due to random fluctuations, we carried out a K-Fold Cross-Validation (K = 5). A paired t-test and a Wilcoxon signed rank test were also conducted.

- RF CV RMSE Average: 7.318775959735352
- XGB CV RMSE Average: 7.2201403099255
- Paired t-test p-value: 0.06024511614128981
- Wilcoxon p-value: 0.125

The paired t-test produced a p-value of 0.0602, showing strong marginal significance. Although the p-value is not quite within the range of the significance level seat α=0.05, the lower values of the ensemble models for all the folds indicate their superiority over the baseline linear models.

D. Impact of Temporal Decay and Risk Index (Ablation Study)

A major contribution of this research was the formulation of the Customer Payment Risk Index (CPRI) and Temporal Decay.

1) Risk Index Contribution

By removing the payment_risk_index feature, the RMSE increased from 8.072464 to 7.793259809790036. This marginal, yet crucial, degradation of performance demonstrates that CPRI has captured behavioral variances not detectable by raw feature sets alone.

2) Temporal Decay Analysis

Fig. 3 - Impact of Temporal Decay on Delay Estimation. While traditional averages are skewed by outlier years, our decayed average is much closer to recent trends, making it a "reactive" approach for more accurate estimation of current customer risk.

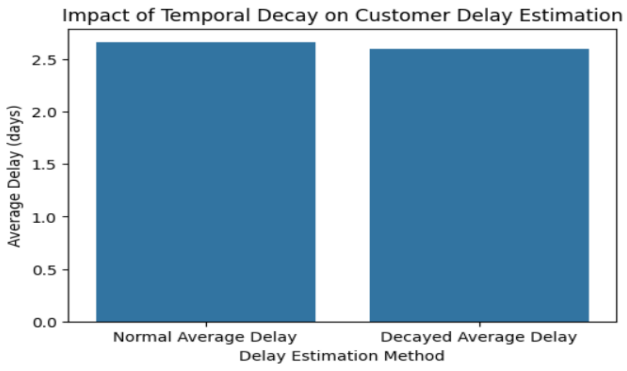

Fig. 3. *Delay estimation methods considered, illustrating how temporal decay weighting affects past averages for customers.*

3) Sensitivity Analysis of the Decay Constant.

Sensitivity analysis of the decay constant (λ) was carried out in order to study the effect of its value on accuracy in prediction. When the values for λ were 90, 180, and 365 days, the smallest RMSE of 7.07 was observed in the case of 180 days. Hence, 180 days or six months was determined to be the most appropriate decay constant.

E. Analysis of Risk Categorization

To provide useful insights for financial controllers, we categorized our customers into three risk tiers based on CPRI scores:
- Low Risk,
- Medium Risk, and
- High Risk.

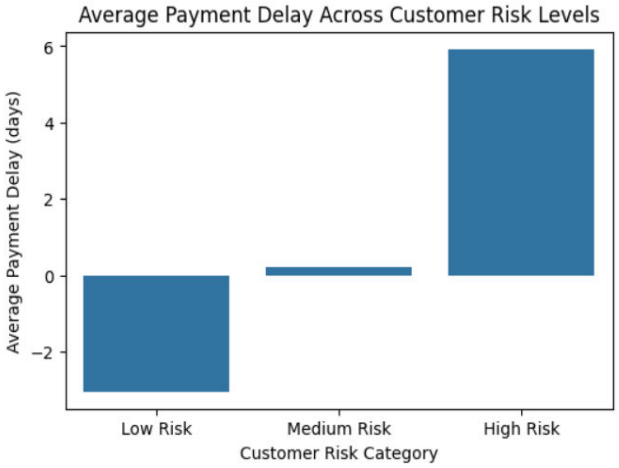

Fig. 4. *Average payment delay by customer risk level, indicating increased delay as risk increases from low to high.*

Fig. 4 - Average Payment Delay Across Customer Risk Levels. As our analysis demonstrates, High Risk customers have an average delay of 6 days, whereas Low Risk customers settle invoices 0 days prior to, or on, the due date, confirming CPRI's strength as an effective means for automated risk tiering.

F. Classification of Late Payments

Additionally, our model was tested for its efficacy in determining if a payment was, in fact, late at all (*Delay* > 0). Using a LightGBM Classifier, our results are as follows:

- ROC-AUC Score: 0.7824107590812461
- Balanced Accuracy: 0.6958498789123349

This high score suggests that the model is highly effective at identifying "at-risk" invoices before they are actually due.

G. Performance on Late Invoices Subset

Since most invoices are paid on time, it is possible for the performance metrics for regression models to be deceptively high. A deep dive into the performance on the Late Invoices Used for Regression subset is therefore necessary.

TABLE III.
Regression Metrics for Late Invoices Only

Metrics	Value
RMSE	8.984887
MAE	3.973482
R2	0.340619

On this difficult subset, the model retained an R^2 of 0.340619, showing it not only predicts whether the delay is late, but also the extent of the delay.

H. Model Interpretability (SHAP Analysis)

The last step of our analysis was using SHAP values to understand the decision-making mechanism of our model. SHAP Summary Bar Plot (Fig. 5) shows that the top three most important features are:

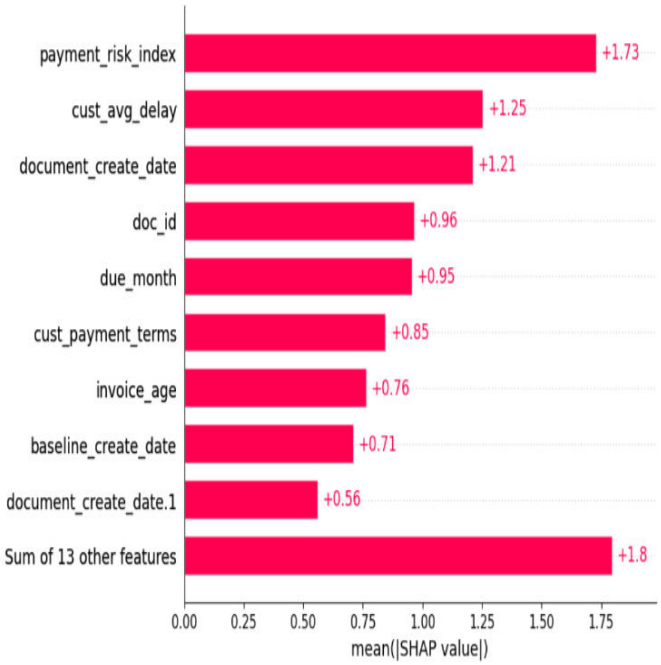

Fig. 5. *Global feature importance scores calculated from mean absolute SHAP values, emphasizing key customer-oriented and time-related factors that impact predictions of invoice payment delays.*

- **Customer Average Delay (Decayed)**: Emphasizing the impact of recent history.
- **Amount Log**: Implying that the amount of the invoice has an impact if it is larger.
- **Payment Risk Index**: Verifying the validity of our engineered feature.

This demonstrates that financial stakeholders can trust the output of our model, since it matches professional financial intuition.

V. Conclusion and Future Work

An integrated machine learning-based framework for accurately predicting invoice payment delays in business-to-business transactions was presented further. A new dynamic machine learning-based framework was introduced for the task after it was determined that the conventional static financial analysis was ineffective. The integration of the Customer Payment Risk Index with a temporal decay weighting mechanism, which addressed the erratic nature of customers' payment behaviors, was the research's contribution. The outcomes of the experiment demonstrated the superiority of gradient-based machine learning models over conventional models like the linear model and the random forest algorithm, especially the LightGBM algorithm.

The LightGBM algorithm's performance metrics, like the Mean Absolute Error and Root Mean Squared Error, were much lower than those of the conventional models, indicating a noticeable increase in prediction accuracy.

The t-test was also used to statistically validate the models' performance, and the outcomes demonstrated how well the risk features were integrated. The gap between the algorithm-based results and the business's financial implications was also filled by using SHAP to explain the results. Because the machine learning-based algorithm can be used to accurately predict the business's payment delay, the research has important ramifications for the business sector, especially the finance sector.

A. Future Work

While the proposed framework exhibits a high level of reliability, there are several areas of future work. To be exact, we plan to investigate the incorporation of external macro-economic factors, for instance, regional interest rates or industry-specific inflation rates. We also plan to investigate the usage of Deep Learning models, for instance, Long Short-Term Memory (LSTM) networks or the more recent Transformers, to leverage long-range dependencies in the temporal patterns of the invoices. We also plan to extend the framework to a streaming architecture to enable instantaneous risk analysis given the generation of invoices.

Future extensions of this framework may incorporate the use of Recurrent Neural Networks, such as LSTMs, which are known to have great potential in the modeling of high-frequency financial sequences [17].

REFERENCES

[1] A. E. Khandani, A. J. Kim, and A. W. Lo, "Consumer credit-risk models via machine-learning algorithms," *J. Bank. Finance*, vol. 34, no. 11, pp. 2767–2787, 2010.

[2] S. Lessmann, B. Baesens, H. V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring," *Eur. J. Oper. Res.*, vol. 247, no. 1, pp. 124–136, 2015.

[3] J. A. Bastos, "Forecasting bank loans loss-given-default," *J. Bank. Finance*, vol. 34, no. 10, pp. 2510–2517, 2010.

[4] F. Feng, X. He, X. Liu, and T. S. Chua, "Attentive factorization machines: Learning the weight of feature interactions," in *Proc. IJCAI*, 2019, pp. 3358–3364.

[5] M. Zięba, J. M. Tomczak, and M. Lubicz, "Ensemble boosted trees with synthetic features for credit scoring," *Expert Syst. Appl.*, vol. 42, no. 4, pp. 328–339, 2016.

[6] P. Baecke and D. Van den Poel, "Customer churn prediction: A decision tree approach," *Decis. Support Syst.*, vol. 54, no. 1, pp. 86–95, 2014.

[7] A. C. Bahnsen, D. Aouada, and B. Ottersten, "Cost-sensitive decision trees for fraud detection," *Expert Syst. Appl.*, vol. 42, no. 2, pp. 1065–1073, 2015.

[8] T. Chen and C. Guestrin, "XGBoost: A scalable tree

boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 785–794.

[9] G. Ke, Q. Meng, T. Finley, et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Adv. Neural Inf. Process. Syst.*, 2017, pp. 3146–3154.

[10] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Adv. Neural Inf. Process. Syst.*, 2017, pp. 4765–4774.

[11] C. Molnar, *Interpretable Machine Learning*, 2nd ed. Leanpub, 2022.

[12] D. Delen, C. Kuzey, and A. Uyar, "Measuring firm performance using financial ratios and decision trees," *Expert Syst. Appl.*, vol. 40, no. 10, pp. 3970–3981, 2013.

[13] M. S. Santos, P. H. Abreu, and C. Soares, "Predicting accounts receivable collection using machine learning: A case study in the manufacturing industry," *Appl. Soft Comput.*, vol. 88, p. 106041, 2020.

[14] J. Sun, H. Li, Q. H. Huang, and K. Y. He, "Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, and forecasting," *Eur. J. Oper. Res.*, vol. 233, no. 1, pp. 1–13, 2020.

[15] A. Gramegna and P. Giudici, "SHAP and LIME: An application of explainable AI in the financial sector," *Appl. Soft Comput.*, vol. 110, p. 107625, 2021.

[16] N. Kozodoi, J. Jacob, and S. Lessmann, "Shallow and deep learning for credit scoring: The value of hierarchical and temporal features," *Expert Syst. Appl.*, vol. 187, p. 115935, 2022.

[17] T. Fischer and C. Riedmiller, "Deep learning with long short-term memory networks for financial market predictions," *Eur. J. Oper. Res.*, vol. 270, no. 2, pp. 654–669, 2018.