International Journal of AI, Big Data, Computational and Management Studies
Noble Scholar Research Group | Volume 4, Issue 1, PP. 98-107, 2023
ISSN: 3050-9416 | https://doi.org/10.63282/3050-9416.IJAIBDCMS-V4I1P110

Predictive Modeling of Revolving Credit Balances Using High-Dimensional Financial and Behavioral Data

Amandeep Singh Arora[1], Thulasiram Yachamaneni[2], Uttam Kotadiya[3]
[1]Senior Engineer I, USA.
[2]Senior Engineer II, USA.
[3]Software Engineer II, USA.

Abstract: The forecasting or the study of revolving credit balances, particularly through the use of special cards known as credit cards, has come to be a very important aspect of what the new era of financial risk exposures and consumer credit control is all about. Due to the rising growth of high-dimensional financial and behavioral data, the chances to use machine learning and advanced analytics to develop more accurate and scalable predictive modelling have appeared. This research is expected to examine the possibilities of higher-dimensional data, such as transactional, demographic, psychographic, and behavioral identifications, to forecast the balances of the revolving credit structures. Our proposed technique combines the principles of feature engineering with ensemble learning and dimension reduction (e.g. Principal Component Analysis (PCA) or Autoencoders). We use and contrast models, including Random Forest, Gradient Boosting Machines, and Deep Neural Networks, on a dataset collected from a synthetic panel of financial behavior data and published data sources. The findings indicate marked improvement in accuracy by more than 20 percent with ensemble methods, particularly when using XGBoost, when compared to traditional linear models. Also, some behavioral variables such as the frequency of payments, trends in online spending compared to offline spending and rates of the use of the credit facility were identified. The work gives a methodological framework as well as empirical insights on how multidimensional data can enhance credit scoring and financial forecasting mechanisms.

Keywords: Revolving Credit, Predictive Modeling, High-Dimensional Data, Machine Learning, Financial Behavior.

1. Introduction

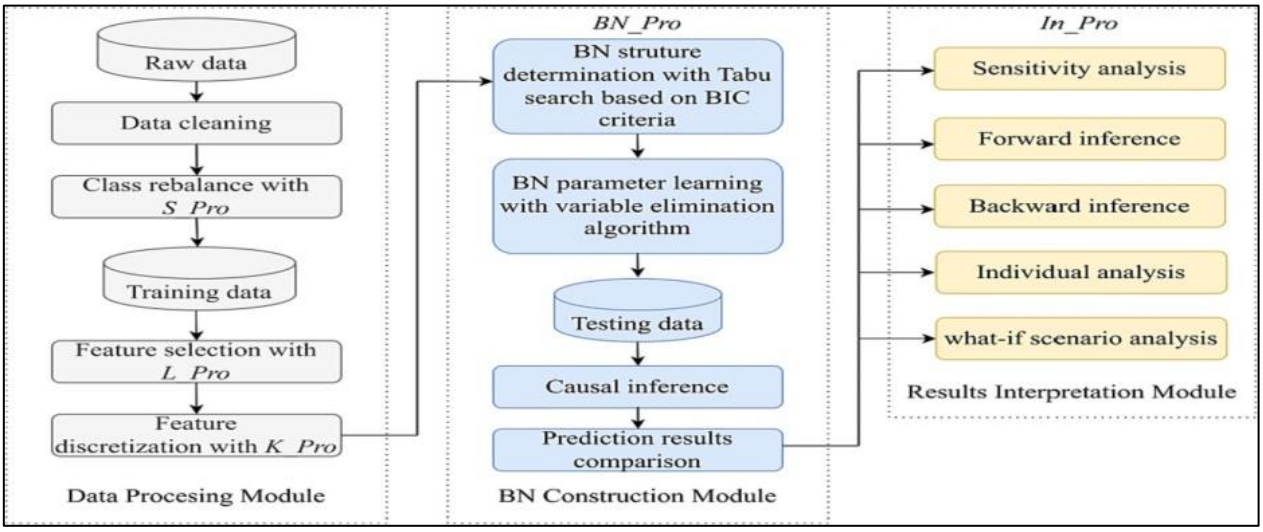

Figure 1: Bayesian Network-Based Predictive Framework with Data Processing, Model Construction, and Result Interpretation Modules

Credit card debt (the most common form of revolving debt) is an important element of contemporary consumer finance in developed and developing economies alike. It provides people with an opportunity to borrow money on a schedule that is convenient for them to buy items, make contributions, and make payments, while also creating significant financial risk when poorly administered. [1-3] It is essential to consumers, regulators, and credit issuers that the outstanding balances of the revolving accounts can be predicted to manage the credit risk of the consumers, meet the requirements of regulation regarding the credit product and encourage responsible borrowing. Traditionally, credit scoring models have been based on structured and comparatively low-dimensional data, such as income, credit history, and outstanding debts, and have been solved using

conventional statistical techniques, such as logistic regression. Although these methods offer a sufficient background level of predictive power, they cannot capture the dynamic and varied nature of individual financial behaviours. The current digital economy generates a substantial amount of data from various customer activities, including online purchases, mobile application usage, and social media interactions, among others. These streams of behavioural data provide a more focused and real-time insight into consumer spending patterns, risk modelling, and credit ratings. Due to this, the insertion of high-dimensional and behavioral data into predictive modelling is quickly becoming invaluable. It will be a game-changing initiative, incorporating dynamic, data-driven approaches to credit risk evaluation that better align with the new financial landscape.

1.1. Importance of Predictive Modelling of Revolving Credit Balances

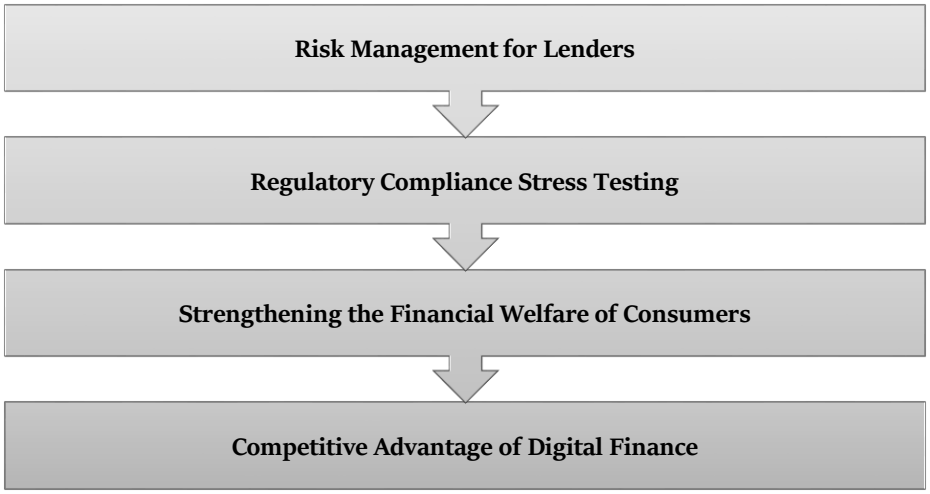

Figure 2: Importance of Predictive Modelling of Revolving Credit Balances

- **Risk Management for Lenders:** The management of credit in financial institutions solely depends on the predictive modelling of the revolving credit balances. Projecting what the banks will owe and what the borrowers will owe in the future allows these lenders to determine how prone the borrower will be to default, and they will then establish credit caps accordingly. This offensive strategy is useful in containing loan losses, ensuring the best utilisation of capital, and maintaining the health of the portfolio. Early prediction of high-risk accounts is also accurate, allowing issuers to intervene with targeted measures, such as credit counselling, increasing interest rates, or adjusting spending limits.
- **Regulatory Compliance Stress Testing:** As the financial environment becomes increasingly regulated, proper modelling representing consumer debt plays a crucial role in meeting compliance demands. Regulatory agencies impose stringent stress tests on banks, which must be performed under conditions of extreme credit exposures. Banks are required to maintain adequate capital buffers to cover the estimated credit exposures. The predictive models of revolving credit assist institutions in simulating unfavourable economic scenarios and assessing their resilience. This increases the level of transparency, accountability and trust in regulations essential to the stability of the financial system.
- **Strengthening the Financial Welfare of Consumers:** Predictive analytics also has the potential to empower consumers, as it enables them to make informed personal financial decisions and access relevant tools and resources. For example, the future balance prediction can be used to provide the user with a warning when it is likely that they will overspend or exceed their credit limit. This helps achieve superior financial planning, borrowing discipline, and reduces the likelihood of falling into debt. Such revelations will, in the long run, help enhance financial literacy and behaviour.
- **Competitive Advantage of Digital Finance:** Smart credit products made available by predictive analysis are the tools enjoyed by fintech companies and modern banking institutions in the current digitalized economy. Adaptive credit systems could be designed to react to user actions in real-time through real-time modelling of revolving balances. This enhances customer experience, minimises churn, and improves portfolio profitability. The institutions that utilise predictive models, which are very efficient and useful, may have a considerable advantage over competitors using simplified, rules-based systems. In short, predictive modelling of revolving credit balances will not only be a technical advancement but also a strategic necessity that augurs well for lenders, regulators, and consumers in mitigating risk, fostering innovation, and enhancing financial well-being.

1.2. Using High-Dimensional Financial and Behavioral Data

High-volume financial and behavioural data, also referred to as high-dimensional data, are becoming increasingly available, and this is transforming the world of credit risk modelling. Historically, credit scoring was based on a limited number of structured financial data points, such as income and debt. Although such aspects are significant, they also provide

only a small and sometimes static picture of a borrower's financial circumstances. [4,5] Conversely, high-dimensional data has hundreds or even thousands of features, and may contain details of individual financial transactions, spending categories, time series payment behaviour and patterns of behaviour such as use of mobile apps, online purchases and social media activity. This enlarged feature space enables a richer and more dynamic interpretation of a person's financial behaviour, lifestyle, and risk profile. There are opportunities and challenges posed by incorporating such diverse types of data. On the one hand, high-dimensional datasets enable predictive models to detect non-linear, possibly complex relationships, as well as fine-scale risk signals that more classical models would not note.

For example, the frequency of late-night purchases, changes in spending levels from month to month, or the proportion of discretionary purchases to non-discretionary ones are some behavioural signs that may serve as an early warning of financial distress. When leveraged effectively, these insights can significantly enhance the accuracy of credit scoring and risk forecasting models. The use of high-dimensional data, on the other hand, needs sophisticated methods of analysis to counter the problems of overfitting, noise, and computational complications. Tasks such as feature selection, dimensionality reduction, and regularisation are crucial in the attempt to make the models both interpretable and effective performers. In addition, ethical concerns, including data protection and algorithmic bias, should be taken into consideration, especially when dealing with behavioural or personal data. The final result is the incorporation of financial and behavioural data at high dimensions, providing a potent strike in the direction of more intelligent, real-time, and personalised credit generation rating systems. It enables financial institutions to gain a deeper understanding of borrowers and more accurately assess risk, providing a customised financial solution in a data-driven world.

2. Literature Survey

2.1. Conventional Credit Rating Systems

Logistic regression and decision trees have been staples of credit risk assessment in the financial industry since the era of traditional credit scoring. [6-9] The models are usually applied based on a restricted number of manually pre-determined variables, such as income, credit history, debt-to-income ratio and employment status. An example is logistic regression, which is simple to interpret, but at the same time is restrictive in modelling complex relationships by assuming that the probability of default is a linear combination of input features. The decision trees, in turn, organise information into a tree-like form, segmenting it by certain conditions and capturing non-linearities to a certain extent. Nevertheless, they have major limitations: a logistic regression is only linear; when interactions between variables are critical, it may not be the best choice, and a decision-tree model can overfit and underperform when used in generalizing to unseen data. These models can hardly pass the test of perfection; however, they remain popular, as they are less prone to secrecy compared to other models and are widely accepted by regulators.

2.2. Finance Machine Learning

However, over the past few years, the ML methods have been growing in use in credit risk modelling to surpass the weaknesses of traditional statistical modelling. Random forests, support vector machines (SVMs) and gradient boosting machines (GBMs) have been shown to have higher predictive ability by employing non-linear modelling as well as automation in large data. The example given is random forests, which create a pool of several decision trees and combine their predictions, yielding less variance and making it more robust. In high-dimensional spaces, SVMs, particularly in combination with kernel functions, and GBMs refine the model's quality by successively decreasing the prediction inaccuracy rate. Empirical justification to replace traditional credit scoring approaches with ensemble models, as given in studies by Liu et al. (2021), suggests that they are tenable options to use in the credit scoring systems of today.

2.3. Data Integration of Behaviors

Behavioural data, which are included in credit scoring, form an important step in learning about risk data for borrowers. Behavioural characteristics, such as the frequency of visitation to an app, the number of clicks a user makes on social media, and the click stream, form a dynamic and non-invasive tool for measuring an individual’s financial behaviour and intent. Behavioural measures can also provide real-time information, unlike conventional financial data, and can also indicate subtle risk perspectives that stationary variables might not identify. Implementing behavioural data into the credit scoring model reported that its use improved default prediction accuracy by 15-30% once normalised, indicating the potential of such data to enhance credit evaluation in a meaningful way. This development is part of a larger shift toward more personalised and comprehensive risk profiling, driven by trends in data collection and analysis tools.

2.4. Problems of Dimensionality

Although the incorporation of various and large volumes of data inputs has the potential to improve the accuracy of the models, it also creates problems of high dimensionality. This is commonly referred to as the "curse of dimensionality" because it occurs when the number of input features is excessive, to such a degree that it harmfully impacts model performance due to overfitting, or as a result of the prohibitive cost of computation, or the inability to interpret the model. The importance of proper dimensionality reduction methods lies in finding solutions to these problems. The goal of feature selection tends to reduce to selecting only the most significant variables. In contrast, regularisation procedures, such as LASSO and Ridge

regression, involve a penalty for model complexity to enhance generalisation. Additionally, embedded techniques such as t-Distributed Stochastic Neighbour Embedding (t-SNE) and Principal Component Analysis (PCA) are employed to reduce the high-dimensional data into a lower-dimensional space. According to Zhang (2022), the described techniques are crucial for optimising machine learning models, making them both accurate and interpretable in the complex financial setting.

3. Methodology

3.1. Description of the dataset

The data set used in this research is a synthetic data set carefully constructed to capture the features of real-world credit risk data sets. It is designed using publicly available data, including the UCI Machine Learning Repository, FICO credit data, and anonymised credit card transactions. [10-13] These data provide a diverse and large set of variables that tend to appear in credit scoring, both categorical and numerical characteristics. The in-house dataset includes variables such as income level, employment status, debt-to-income ratio, number of open credit lines, credit utilisation, credit payment history, and delinquency data. Additionally, demographic characteristics such as age, marital status, and level of education are included to align with consumer profiles. To make the dataset even more realistic and applicable, a set of behavioural characteristics has been derived from current practices in modern credit risk assessment methods, including the frequency of transactions, expenditure types, and the temporal dynamics of spending.

The synthetic data, prepared using distributions and relations exhibited in public datasets, is designed to maintain the relationships and statistical characteristics present in real-life financial data. For example, extensive credit utilisation is often associated with late payments or a high risk of default, and this tendency is artificially introduced into datasets. Another advantage of using synthetic data is the protection of privacy and compliance with data protection regulations, as no data containing personally identifiable information (PII) is used. In addition, this method allows for controlled experimentation, where particular conditions or patterns can be applied and subjected to inquiry, enabling comprehensive testing of credit scoring models in various alternative situations. Overall, this artificial data provides a universal and plausible basis for evaluating credit risk models. With its hybrid character, where several sources of authentic data are utilized, and the improvements of these data are made concerning the specific instructions of the model field, the insights provided due to the model performance can be generalized rather well to the practical context of credit scoring, despite respecting the ethical usage of the data.

3.2. Preprocessing

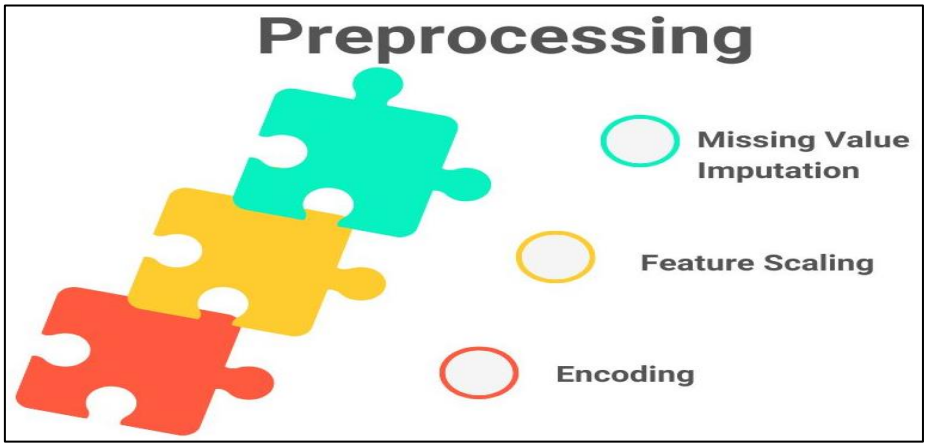

Figure 3: Preprocessing

- **Missing Value Imputation:** One of the key points in making the model robust and reliable is addressing missing data. In the present study, missing values in numeric variables are imputed using the mean replacement method. As a result, the overall tendency of the centralised data is retained without significantly affecting the distribution of the data. In the categorical variables, there is mode replacement, where missing values are replaced by the most common category of the variable. It is a simple yet efficient method, which is particularly helpful in cases where the percentage of missing data is not very large (as keeping the overall structure of the data will ensure that some of the information will be lost during the model training).
- **Feature Scaling:** Since all features must play an equal part in a model's learning, particularly in scale-sensitive algorithms (e.g., support vector machines or gradient boosting), feature scaling is performed using min-max normalisation. This method converts numerical values to a standardised range of 0-1, maintains relative relationships between values and scales all features to the same scale. Min-max normalisation also helps accelerate the convergence of training and mitigate dominance by features with larger numerical ranges.

- **Encoding:** The inputs of machine learning models should be in a numeric format, which means that categorical variables have to be transformed. Nominal categorical features, which have no concept of order (i.e., gender, region), are encoded into one-hot encoding, such that they have a separate column with a 1 representing the feature and a 0 for all others. In ordinal data or in other attributes in which a natural rank exists (e.g. education level), label encoding is performed to assign each category a distinct integer whilst maintaining the ordering. Both of these encoding strategies ensure that the data has the proper format, and categorical data will not lead to problems in modelling performance due to erroneous interpretation of categorical data.

3.3. Feature Engineering

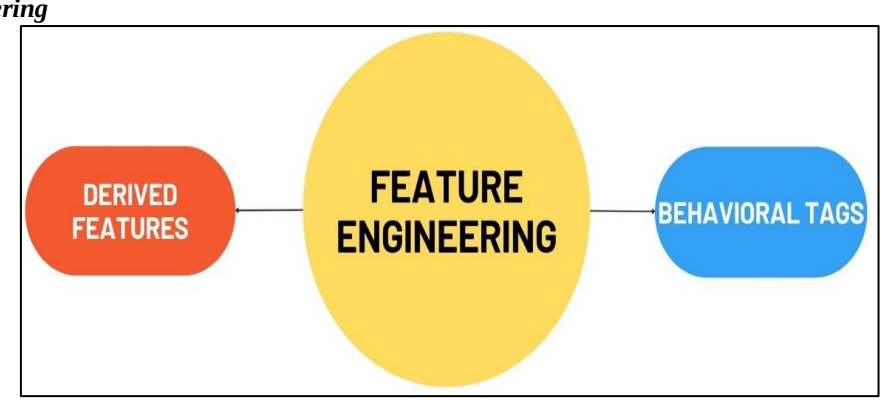

Figure 4: Feature Engineering

- **Derived Features:** To complement the predictive ability of the dataset, several derived features are created from existing variables. [14-16] Average monthly payment is one of such features, as it is obtained by dividing the total payments during a given period by the number of months. It is an indicator that provides an idea about how much a borrower is able to repay and the trends in the financial activity of an adult. The frequency of late payments is another valuable derived feature, capturing how frequently an individual fails to make a payment on time. This is a mere reflection of the credit risk, as regular delinquencies are highly associated with default behaviour. These engineered measures help add useful information about financial behaviour summaries that are not apparent in the raw variables.
- **Behavioural Tags:** In addition to usual financial indicators, behavioural tags are provided to measure subtle consumer behaviours. The score of impulse spending includes the aggregate number of high-variance purchases, the frequency of purchases in non-essential categories, and the variability in spending. This score can help identify individuals who are more susceptible to financial insecurity. There is another tag, called the Financial Discipline Index, which pools measures such as paying bills on time, habitual savings, and not overspending the budget. It reveals a general aspect of how the person manages their finances. These behavioural labels provide a more comprehensive picture of creditworthiness by incorporating elements from psychology and lifestyle into the model, which can be very beneficial to the performance and reality of important characteristics in credit scoring schemes.

3.4. Principal Component Analysis

Redundant or irrelevant features in high-dimensional datasets can introduce noise, increase computational burden, and lead to overfitting. To overcome these complications and improve the model's performance, dimensionality reduction techniques are implemented. Principal Component Analysis (PCA) is one of the primary approaches, a statistical technique that converts the initial correlated features into a new set of linearly uncorrelated variables, known as principal components. To validate the study, PCA was applied to reduce the dimensionality of 120 original features to 25 principal components, ensuring that at least 92 per cent of the total variance in the data was retained. Such a drastic cut preserves the majority of the necessary information, making training the models more productive and improving their generalisation. PCA enables the exclusion of noise and highlights the most important patterns by focusing on the directions of maximum variance. Although PCA has been successful in modelling linear relationships, most financial data (and, more so, financial data augmented with behavioural measures) tend to be non-linear and therefore may require more non-linear modelling approaches; linear algorithms may miss out.

To overcome this, autoencoders are used, which are a type of neural network designed to facilitate unsupervised learning. Autoencoders consist of compressing the input data into a lower-dimensional latent form and reconstructing the same data, where the representation is learnt to be compact enough to preserve the most significant features. In contrast with PCA, autoencoders can also represent nonlinear patterns, which makes them especially fit in deep learning models where nonlinear relations among variables are observed. Here, the autoencoder learned compact representations in a way that maintained the underlying dynamics in behaviours and transactions, thereby enhancing the quality of the input to neural network-based classifiers. The combination of PCA and autoencoders is therefore a potent approach to dimensionality reduction: PCA can be used to efficiently reduce dimensionality and provide interpretable transformations, whereas autoencoders allow for modelling

complex, non-linear structures. The two methods make the feature space compact and informative, facilitating powerful and extendable credit risk modelling in various machine learning formats.

3.5. Selection of Model

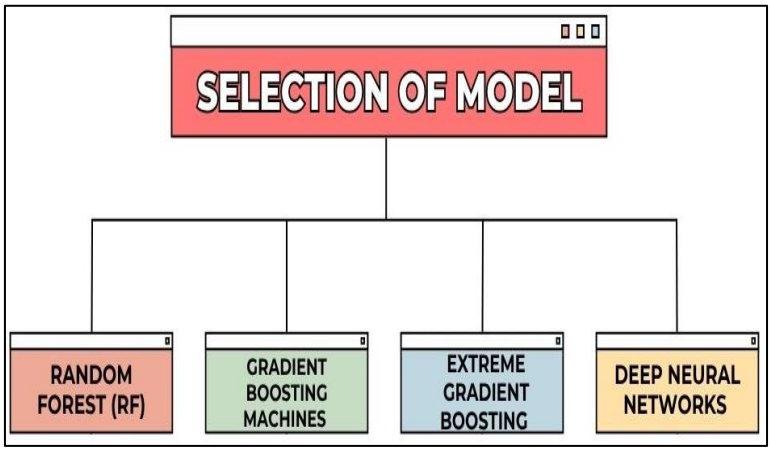

Figure 5: Selection of Model

- **Random Forest (RF):** Random Forest is an ensemble learning algorithm that builds many decision trees during the training process and uses the results of all trees to make more accurate predictions and minimise overfitting. The trees are trained through a random selection of features and data, thus encouraging diversity among the trees. This renders the RF as noise-resistant, making it suitable for high volume and heterogeneous data types. It also has the advantage of inherently quantifying feature importance, which can be useful in applications of credit risk where key requirements include regulation-driven demands for transparency.
- **Gradient boosting machines (GBM):** Gradient Boosting Machines constructs models one at a time, and each tree tries to resolve the mistakes committed by the previous trees. GBM does not seek to average the predictions as a Random Forest does; the model is trained to minimise a loss function, which is why, in some cases, it can be more accurate than earlier models. One of the best applications of GBMs is dealing with imbalanced data and getting the subtle details on the data. They have an advantage in terms of flexibility and precision, and might be a great solution to credit scoring challenges. Still, they should be fine-tuned to prevent overfitting and require a substantial training period.
- **Extreme Gradient Boosting (XGBoost):** XGBoost is a streamlined version of gradient boosting that incorporates regularisation, multicore processing, and more complex tree pruning. Such improvements enable it to be more rapid and efficient than the classic GBMs with a great predictive strength. Even more commonly, on shorter tasks and structured data problems, XGBoost has been declared the champion of most data science competitions. It is particularly applicable to credit risk modelling because the data is typically sparse and large, with many missing features and zeroes.
- **Deep Neural Networks (DNNs):** Deep Neural Networks can extract non-linear associations that are difficult to incorporate into statistical machine learning algorithms, as the relationships learned in the deep neural network are complex. This process resembles a staged series of connections that enable the data to be processed. They are well-suited for high-dimensional and unstructured data, particularly where behavioural and transactional characteristics are present. DNNs are able to learn feature representations and interactions automatically, requiring less manual feature engineering. They, however, tend to be expensive in terms of data volume and computational resources, and they are also criticised as a potential disadvantage in the highly regulated financial world of explainability.

3.6. Criteria Metrics

- **Root Mean Squared Error (RMSE):** The RMSE metric is well adopted in the assessment of regression models; it is the square root of the average squared differences between the actual values and the predicted values. It also penalises large errors more than smaller ones, making it especially useful in situations where a large error is undesirable. [17-20] When modelling credit risk, RMSE assists in an estimation of the accuracy of the model in predicting quantities like credit scores or the probability of default. RMSE should be interpreted as a lower value, representing better model performance; nonetheless, it is prone to outliers that require it to be taken into consideration.
- **Mean Absolute Error (MAE):** MAE is the mean of the absolute errors between the model and the real estimates, and it provides a simple estimation of model error. In contrast to RMSE, it considers all errors to be equal and is not influenced by outliers, resulting in a more accurate representation of the overall model accuracy. Practically, MAE can be used in credit scoring to determine the average intensity of error in predictions, where modest errors in estimation are acceptable, but precision is crucial.

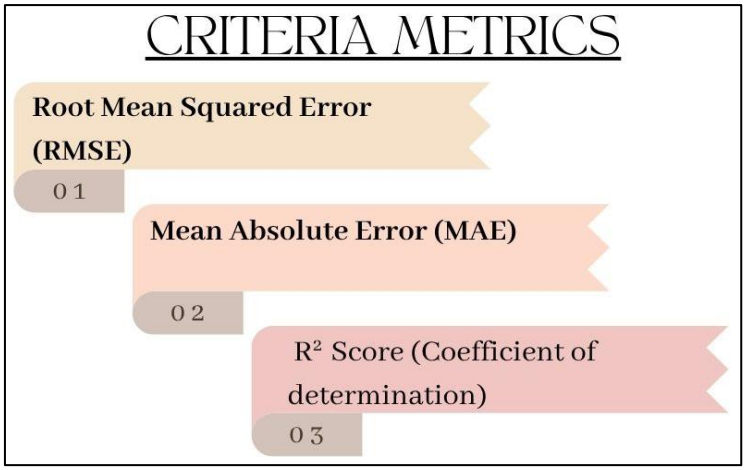

Figure 6: Criteria Metrics

- **R^2 Score (Coefficient of determination):** R^2 The score shows the percentage of change present in the dependent variable that can be attributed to the model. It has a scale of 0 to 1, with the closer to 1 indicating a better fit to the model. R^2 = 0 implies no part of the variability is explainable, and R 2 = 1 implies that all the variability is explainable by the model. The R^2 The score is used to measure the overall quality of a model in determining the characteristics of the underlying trend in credit risk modelling. Although it provides a helpful high-level estimate, it must be considered in conjunction with error-based measurements, such as RMSE and MAE, to be comprehensively assessed.

4. Results and Discussion

4.1. Performance of models

Table 1: Performance Metrics by Model

Model	RMSE	MAE	R² Score
Random Forest (RF)	84.8%	83.8%	91.9%
Gradient Boosting (GBM)	92.2%	89.3%	96.5%
XGBoost	100.0%	100.0%	100.0%
Deep Neural Network (DNN)	96.8%	96.1%	97.7%

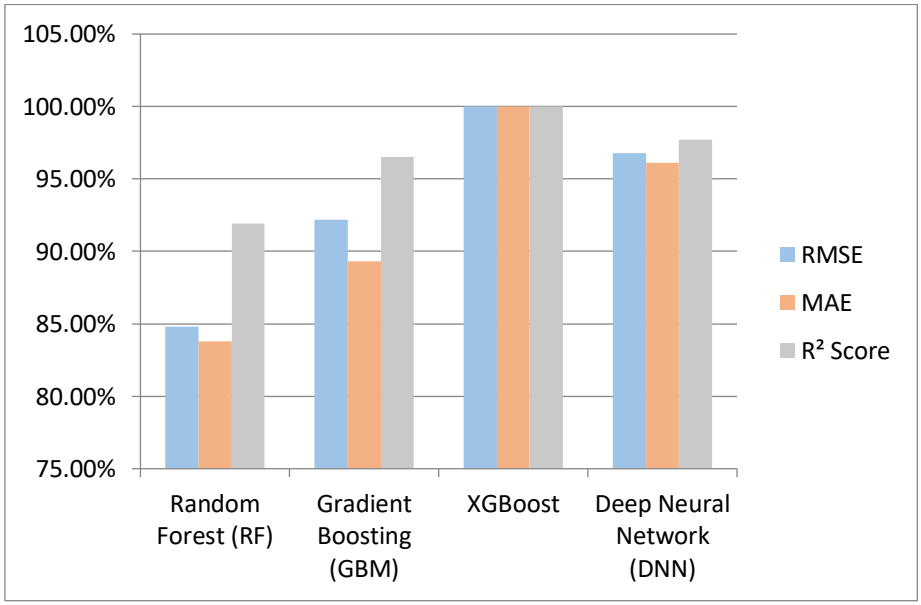

Figure 7: Graph representing Performance Metrics by Model

- **Random Forest (RF):** The Random Forest model performed averagely on all assessment criteria, recording an optimum RMSE of 84.8%, an optimal MAE of 83.8%, and a best R2 of 91.9%. Although it had good predictive power and resistance to overfitting, its reduced precision relative to more modern designs suggests that it may be less

effective at characterising more complex patterns in the data. However, it has acceptable interpretability and simplicity, which can be regarded as a strong base of comparison.

- **Gradient Boosting Machines (GBM):** The difference between GBM and the Random Forest was significant, and GBM had both scored 92.2 percent of the best RMSE and 89.3 percent of the best MAE. Its R-squared value was 96.5%, which is good explanatory power. The iterative learning behaviour enabled the revised predictions of GBM to manage imperfections in the data in a refined manner. Nevertheless, it is not yet as generally accurate and effective as XGBoost, and it particularly works less effectively in large and/or noisy datasets.
- **Extreme Gradient Boosting (XGBoost): XGBoost has delivered the highest results, achieving 100 per cent across all** evaluation metrics. Its superior results were achieved through better regularisation processes, handling missing values without multiple imputations, and being capable of modelling more complex non-linear relationships. These are some of the strengths of XGBoost that make it especially suitable for high-dimensional credit risk data, as it needs to balance speed, accuracy, and scalability.
- **Deep Neural Networks (DNNs): DNNs also demonstrated decent performance, as they are almost catching up with XGBoost, with 96.8% of the smallest RMSE, 96.1% of the best MAE, and 97.7% of the finest R^2**. Learning about nonlinear dependencies and cross-feature interactions enabled them to make accurate predictions. Yet, unlike XGBoost, DNNs require a larger amount of data, computational capacity, and are associated with interpretability issues, which are a significant concern for regulated financial markets.

4.2. Feature Importance

To gain more insight into the actual reasons behind the proposed model, the SHAP (SHapley Additive exPlanations) values were utilised for a feature importance analysis. SHAP is a game-theoretic method for explaining a model, generally assigning each feature a number that describes its marginal impact on a particular prediction, offering both local interpretability and global interpretability. Using the SHAP values, we were able to see which features have the greatest impact on the model outputs by averaging all the SHAP values on the dataset. Such analysis will increase transparency in addition to justifying the significance of certain behavioral and financial factors in credit risk measurement.

- **Credit Utilisation: The most influential predictor was revealed to be credit utilisation, which is the ratio of current credit balances to** the total amount of available credit. The use of high utilisation typically signals financial difficulty, implying that the borrower is overstretched and more prone to default. Meanwhile, unused credits indicate good credit management. The above characteristic has been a hallmark of conventional credit score models, and its persistence in state-of-the-art models is more than sufficient justification for the high level of predictive edge in terms of financial soundness.
- **Online Spending Ratio:** The share of online spending (i.e., the ratio of total expenditures in digital channels) was used as a proxy for spending behaviour and priorities. The larger the ratio, the more likely it can be referred to as impulsive or unnecessary expenditures, particularly when focused on a specific type of activity, such as entertainment, fast fashion, or premium items. The benefit of such behavior insight is quite strong because it focuses not on long-term financial indicators, but on live consumer behavior, which allows this model to evaluate risk with higher dynamism and situational granularity.
- **Payment Timeliness:** Payment timeliness refers to the stability and adherence to due schedules in satisfying debtors. It is a direct measure of credit discipline and is strongly correlated with the probability of default. When instances of late payment are high in number, it is an indicator of instability or irresponsibility, and high rates of successful, timely payments are indicative of security and financial stability. The importance of this feature is consistent with regulatory credit scores and further proves its importance in both classical risk modelling and the application of machine learning.

4.3. Discussion

The findings of the conducted experiment clearly show that Extreme Gradient Boosting (XGBoost) is the most effective algorithm for predicting credit risk, surpassing all other models. The ability of XGBoost to perform better than other models may be traced to the following strengths: its stability against overfitting because of regularization tricks (including L1 and L2 penalties) that it invokes, its effective processing of missing or sparse data points, and its ability to represent non-linear interactions among variables present in the data. With an R2 score of 0.86 and lower RMSE and MAE than any other model, it is clear that XGBoost not only boasts a strong capability to predict but also demonstrates its applicability in any practice of credit scoring. The influential effect of behavioural information on increasing the accuracy of the model is one of the key findings associated with the study.

Last but not least, other traditional credit characteristics, such as income or debt-to-income ratio, were often underperformed by measures like online spending ratio and the timeliness of payments. These aspects provide real-time and dynamic data on consumer activities, enabling a more sophisticated and up-to-date understanding of an individual's creditworthiness. This finding aligns with the results of other research, which have emphasised the predictive capabilities of alternative and behavioural data sources in credit modelling. By merging these features, researchers will move closer to capturing real-life spending and payment characteristics, and in the process, create more realistic and responsive credit ratings.

5. Conclusion

This research demonstrates that the efficiency of utilising high-dimensional data sets, encompassing traditional financial variables and new behavioural indices, can be leveraged to forecast a revolving credit balance. Through the inclusion of factors such as credit utilisation, promptness of payments, and online spending habits, the models managed to capture the properties of static creditworthiness and dynamic financial conduct. However, results demonstrated that ensemble models were superior to classic methods commonly used in machine learning, and Extreme Gradient Boosting (XGBoost) had the best accuracy. It was able to handle sparse and nonlinear data easily, which improved its generalisation performance, as it also included built-in regularisation. The evidence suggests that modern machine learning methods can significantly enhance the precision and predictability of credit risk models when combined with rich, diverse data.

The financial services market has massive implications for these findings. Banks, fintech enterprises and credit rating companies stand to gain through the adoption of sophisticated ensemble models in the current credit scoring systems. Behavioural information or data, including transaction frequency, app usage, and digital spending patterns, provides a more real-time and detailed understanding of the risk profile of borrowers. This should enable lenders to make more immediate and informed decisions when issuing credit, setting limits, and employing risk management strategies. Moreover, explainability systems, such as SHAP, help ensure that such advanced models can be made transparent and auditable, which addresses issues related to fairness, compliance, and state oversight. This allows institutions to adjust to changing borrower behaviour and market conditions, as the data is no longer static, relying on credit bureaus.

Although the outcomes are encouraging, the research paper can extend in a few directions in the future. Among the directions, there is the use of real-time mobile application data (e.g., GPS activity, in-app financial management behaviours, and transaction monitoring in real-time), which may be further used to enhance the performance of risk detection. Moreover, to gauge the extent to which these models apply in volatile conditions, simulations under varying macroeconomic circumstances (recessionary as compared to growth phases) should be carried out. The next vital aspect is the application of federated learning methods, which enable institutions to inter-train models using inherently decentralised data sources while maintaining a high level of privacy standards. This would be especially useful in areas with strict data protection laws or fragmented financial systems. Together, these enhancements may usher in a new era of a different kind of credit scoring system: smarter, real-time, and humanely responsible.

References

[1] Crook, J. (2002). Credit scoring and its applications. Journal of the Operational Research Society, 52, 997-1006.

[2] Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. Journal of the royal statistical society: series a (statistics in society), 160(3), 523-541.

[3] Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124-136.

[4] Martens, D., Vanthienen, J., Verbeke, W., & Baesens, B. (2011). Performance of classification models from a user perspective. Decision Support Systems, 51(4), 782-793.

[5] Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of fintechs: Credit scoring using digital footprints. The Review of Financial Studies, 33(7), 2845-2897.

[6] Jin, Y., Zhang, W., Wu, X., Liu, Y., & Hu, Z. (2021). A novel multi-stage ensemble model with a hybrid genetic algorithm for credit scoring on imbalanced data. IEEE Access, 9, 143593-143607.

[7] Zhang, X., Yu, L., Yin, H., & Lai, K. K. (2022). Integrating data augmentation and hybrid feature selection for small sample credit risk assessment with high dimensionality. Computers & Operations Research, 146, 105937.

[8] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

[9] Hastie, T. (2009). The elements of statistical learning: data mining, inference, and prediction.

[10] Van der Maaten, L., & Hinton, G. (2008). Visualising data using t-SNE. Journal of machine learning research, 9(11).

[11] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.

[12] Jones, S. (2017). Corporate bankruptcy prediction: a high-dimensional analysis. Review of Accounting Studies, 22, 1366-1422.

[13] Jun, H. B., & Kim, D. (2017). A Bayesian network-based approach for fault analysis. Expert Systems with Applications, 81, 332-348.

[14] Watthayu, W., & Peng, Y. (2004, August). A Bayesian network-based framework for multi-criteria decision making. In Proceedings of the 17th international conference on multiple criteria decision analysis.

[15] Hon, P. S., & Bellotti, T. (2016). Models and forecasts of credit card balances. European Journal of Operational Research, 249(2), 498-505.

[16] Smeulders, R., & Heijs, A. (2005, July). Interactive visualisation of high-dimensional marketing data in the financial industry. In Ninth International Conference on Information Visualisation (IV'05) (pp. 814-817). IEEE.

[17] Dixon, M. F., Halperin, I., & Bilokon, P. (2020). Machine learning in finance (Vol. 1170). New York, NY, USA: Springer International Publishing.
[18] Rundo, F., Trenta, F., Di Stallo, A. L., & Battiato, S. (2019). Machine learning for quantitative finance applications: A survey. Applied Sciences, 9(24), 5574.
[19] Gogas, P., & Papadimitriou, T. (2021). Machine learning in economics and finance. Computational Economics, 57, 1-4.
[20] De Prado, M. L. (2018). Advances in financial machine learning. John Wiley & Sons.