International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 09 | Sep 2024 www.irjet.net p-ISSN: 2395-0072

CREDIT CARD SCORING ANALYSIS USING MACHINE LEARNING AND DEEP LEARNING

Anmol K A, K D Sruthi

Anmol K A , Msc. Computer science, St.Thomas(Autonomous)College, Thrissur 680001,Kerala , India
K D Sruthi , Msc. Computer science, St.Thomas(Autonomous)College, Thrissur 680001,Kerala, India

---

Abstract - *In today's world, credit scores are essential to determine credit worthiness for lending institutions, and they impact everything from getting a mortgage to renting an apartment. This thesis addresses key challenges in credit scorecard development, focusing on three main contributions. Firstly, it evaluates the performance of supervised classification techniques on imbalanced credit scoring datasets. Secondly, it explores the low-default portfolio problem, a severe form of class imbalance in credit scoring. Thirdly, it quantifies differences in classifier performance across various implementations of a real-world behavioral scoring dataset. Additionally, the thesis demonstrates the use of artificial data to overcome challenges associated with real-world data, while acknowledging the limitations of artificial data in evaluating classification performance.*

Keywords: Credit scoring, Machine learning, Deep learning, FE-Transformer, Feature selection.

1.INTRODUCTION

Credit scoring models play a crucial role in the business landscape, providing a numerical assessment of an individual's creditworthiness based on diverse financial factors. Lenders and credit card companies heavily rely on these scores to make informed decisions on loan approvals or credit extensions. Typically ranging from 300 to 850, higher scores signify better creditworthiness. Scores above 700 are generally considered good, while those below 600 are seen as poor. Despite meticulous verification processes, there's no absolute assurance that credit cards are granted only to deserving candidates, emphasizing the ongoing importance of refining credit assessment strategies. Credit scoring serves as a conventional decision model, primarily focusing on risk assessment related to credit products like credit cards and loans.Financial institutions are increasingly embracing diverse risk assessment tools, including statistical analysis, to minimize potential risks. The utilization of deep learning algorithms, specifically transformers, based on online behavioral data has demonstrated superior performance compared to LSTM and traditional machine learning models. Our innovative end-to-end deep learning credit scoring framework incorporates both credit feature data and user behavioral data. The framework comprises a wide part and a deep part, enabling automatic learning from user data to enhance decision-making in credit granting.

1.1 Credit Scoring Process

Credit scoring which is a conventional decision model and it is mainly focusing on risk approximation approach associated with credit products such as credit card, loans, etc. and is estimated based on applicant's historical data which helps credit lenders in granting credit products.The following diagram illustrates the flow of data through the credit scoring process, from data collection to the final credit score output.

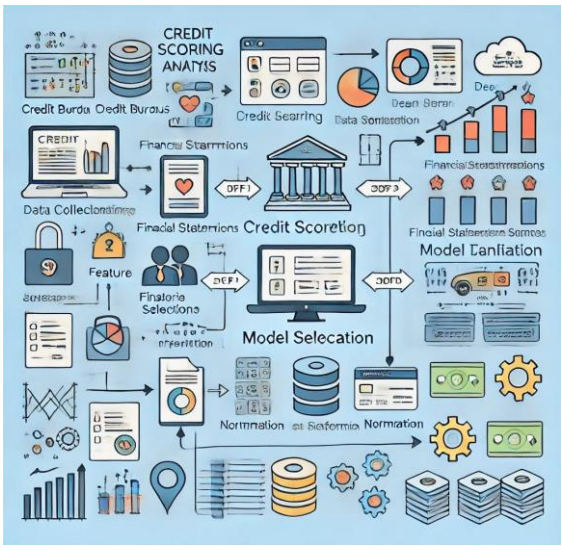

Fig-1: Data Flow Diagram

Probability of Default (PD) analysis is carried out for generating credit scores for individual customers to identify default when he visited a bank for a loan and check their credit score. The credit score is used by banks for credit bureaus.

1.2 Internal Rate Of Return (IRR)

IRR is a crucial method for assessing investment profitability. While there are alternative strategies to gauge expected returns, the internal rate of return (IRR) stands out as the widely employed calculation. This rate plays a pivotal role in determining the interest assigned to loans for borrowers. It's noteworthy that IRR was historically associated with negative training examples.[3]

1.3 Credit Risk Components

Credit risk scoring involves various components to assess a borrower's creditworthiness: Credit history is used to examining the borrower's past credit behavior, including payment history, defaults and delinquencies. Credit utilization is the ratio of current credit usage to the total available credit, indicating the borrower's reliance on credit. Debt-to-Income Ratio(DTI) is assessing the borrower's ability to manage additional debt based on their income compared to existing debt obligations [5]. Length of credit history is the duration of the borrower's credit accounts, as a longer credit history provides more data for evaluation. New Credit is used to monitoring recent credit applications, as multiple applications within a short period may signal financial stress. Types of credit in use in considering the variety of credit accounts, such as credit cards, mortgages and installment loans, in the borrower's portfolio. Public records for to identifying any bankruptcies, liens or legal judgements that may impact creditworthiness. Payment behavior is to assessing the consistency and timeliness of payments on existing credit accounts. Income stability is to evaluating the stability and reliability of the borrower's income source. Employment history be the stability and continuity of the borrower's employment can provide insights into their financial stability. Combining these components through statistical models, machine learning or deep learning techniques helps create a comprehensive credit risk score, aiding lenders in making informed decisions. Regular updates and refinements to the scoring model are crucial to adapt to changing economic conditions and borrower behaviors.

2. CREDIT SCORING SYSTEM

In credit scoring system two phases are being used to find out the best practices for filtering the customers' matches for a particular credit score. In the first phase, application scoring is an initial assessment made based on the information provided by the borrower in their credit application. Commonly used models in this stage include rule-based systems or simpler analysis models to quickly filter and categorize applicants. In the second phase, behavioral scoring is done after passing the initial stage, borrowers undergo a more in-depth evaluation based on their behavior and credit history. This stage involves more sophisticated models such as machine learning or neural networks, to analyze historical data and predict future credit behavior. Our research is to build a reliable predictive model for credit scoring which helps lenders to allocate funds in financial institutions based on the credit score. The accurate outcome is produced by the model even if the dataset is imbalanced and improves the feature selection process by adopting a deep neural network which makes a balanced dataset. In this scheme, the decision tree classifier is used to assign new weight for every class in the predictive model with respect to accuracy. The model is validated on different credit scoring dataset in real-world scenarios and which is capable of improving the effectiveness and accuracy for training data and ensures that the training data is balanced.

3. MACHINE LEARNING AND DEEP LEARNING

Machine learning algorithms have transformed credit scoring, offering enhanced accuracy in assessing creditworthiness. Trained on extensive datasets, these models excel in pattern recognition, outperforming traditional credit scoring methods. A key advantage of machine learning lies in its capacity to mitigate bias. Unlike traditional models, which may exhibit biases based on factors like race or gender, machine learning algorithms are designed to be unbiased, relying solely on data without incorporating preconceived biases. This contributes to fairer credit-scoring decisions. Moreover, machine learning excels in efficiency compared to traditional models. Rapid analysis of vast data sets enables near-instantaneous credit-scoring decisions, streamlining the lending process for both borrowers and lenders. In recent years, deep learning has demonstrated its efficacy across various applications such as text sentiment classification, image classification, and recommendation systems. Applying deep learning to credit scoring has shown promise, particularly in automatically learning features from data. The large volume, high dimension, and sequential nature of user online behavioral data pose challenges for traditional machine learning algorithms, motivating researchers to explore deep learning methods. Notably, Hidasi et al. enhanced recommendation systems using a recurrent neural network based on user online behavioral data, showcasing an improvement over existing

methods. Long Short-Term Memory network (LSTM) to analyze consumer activity on e-commerce websites, yielding favorable experimental results. Existing models, such as the LSTM, exhibit long-term dependence but lack parallelization capabilities, warranting further exploration of deep learning algorithms. Notably, there is a research gap in the development of end-to-end neural network models for credit scoring, combining both user behavioral data and feature data.

4. METHODS

4.1 LSTM

LSTM is widely used for processing sequential data like text classification and machine translation. It addresses long-term dependencies through gate structures, making it suitable for credit scoring analysis to capture temporal patterns. LSTMs find applications in various areas like time series prediction, handling variable-length sequences, feature representation, and default prediction. The forget gate, input gate, and output gate play key roles in managing information flow. However, LSTMs have limitations in parallel processing due to sequential calculations and may not entirely eliminate long-term dependency issues.

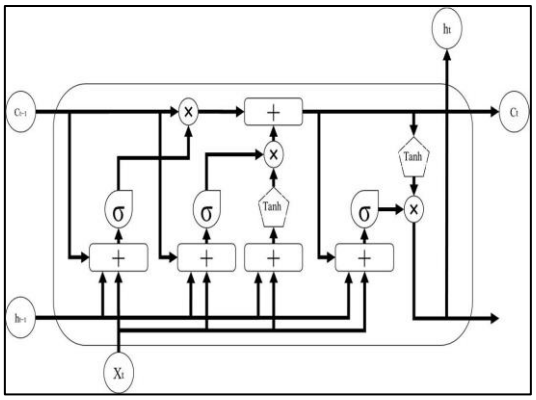

Fig-2: Structure of the LSTM model.

4.2 Transformer

The transformer model, initially employed by Google for machine translation, adopts an encoder-decoder structure with six layers in total—consisting of an encoder and a decoder. Unlike traditional models, it doesn't utilize a recurrent structure. Through a 6-layer encoder, input data progress to the decoder for attention calculation. The transformer comprises four modules: input, encoding, decoding, and output. Leveraging self-attention and parallel processing, it excels in machine translation, surpassing RNNs and CNNs as the current mainstream feature extractor. Addressing LSTM limitations, it utilizes attention to reduce sequence position distances and allows parallel computation, showcasing superior feature extraction capabilities in machine translation tasks. Consequently, attention-mechanism-based LSTM has transitioned to the transformer model's network structure [1,4].

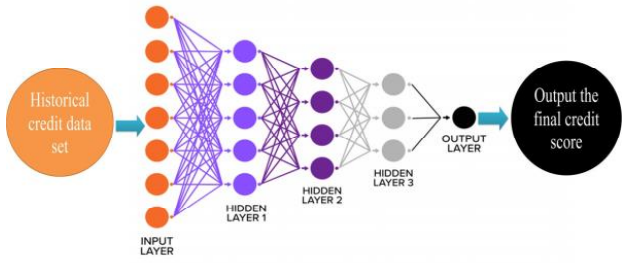

Fig-3: Deep neural network with hidden layers used in proposed approach.

The model effectively predicts new instances even with missing data, employing the ReLU activation function as a threshold mechanism. This boolean method classifies instances based on input and weight vectors, with a cost function minimizing squared error. Additionally, a decision tree contributes to classification, resulting in optimal predictions for credit rating. [3]

4.3 Feature Embedded Transformer

In this research study, we integrate a transformer into credit scoring, presenting an end-to-end deep learning credit scoring framework known as the Feature Embedded Transformer (FE-Transformer). A FE-Transformer in credit scoring would involve incorporating both the temporal aspects of a borrower's financial history and the static features into a unified model. The model contains joint representations for both sequential and static features, leveraging the strengths of transformers in capturing long-term dependencies and feature embedding in handling categorical or numerical aspects.

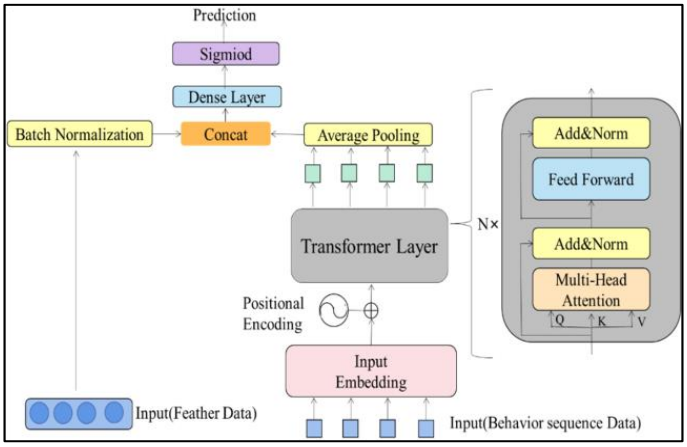

Fig-4: Network architecture of the FE-Transformer.

4.4 Input Data and Data Coding

The input data models consist of two types: feature data and behavioral data. Feature data containing gender, age, credit record, and other credit-related information, undergo processing before serving as model input. Behavioral data is inspired by NLP treats each behavior event as a word, forming a sequence resembling a sentence. Raw online operation behavior records are transformed into chronological event sequences. Through embedding and position encoding, behavioral data is encoded for model input. Events are converted to vectors using word embedding, and positional encoding is employed to capture event positions within the sequence. The resulting position vectors are added to the event vectors, creating the model input.

4.4 Transformer Encoding Layer

Layer consists of stacked encoders, each containing a multi-head attention layer and connected feed-forward layer. The embedding layer's output is processed through the multi-head attention and feed-forward layers in the encoder. The resulting encoding information matrix for all events in the behavior sequence is obtained after one or more encoders. The transformer architecture employs a self-attention mechanism, improving the model's ability to capture internal correlations without relying heavily on external information. Scaled dot-product attention is used in the attention layer, offering faster computation and space efficiency. The self-attention mechanism calculates the relatedness between events by projecting each event into query (Q), key (K), and value (V) vectors. These vectors are used to query candidate positions, and the dot products are scaled, normalized using softmax, and weighted to determine the final self-attention result.

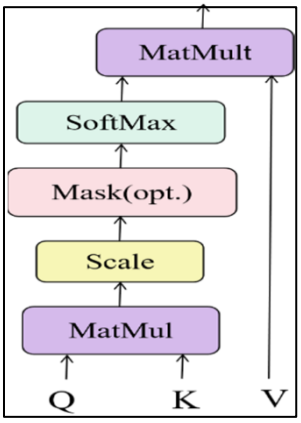

Fig-5: Scaled dot-product attention.

The multi-head self-attention mechanism in transformers allows the model to simultaneously capture diverse information from various positions and representation subspaces, enhancing its ability to understand the structure and relationships within sequences of events. This feature plays a crucial role in enriching the model's comprehension of both semantic and syntactic information.

4.5 Evaluation Metrics

In evaluating the model's performance for credit scoring, commonly used indicators include AUC and KS. AUC measures the model's ability to discriminate between defaulters and non-defaulters, with a higher value indicating better performance. KS assesses the maximum difference between cumulative distributions of good and bad credit applicants, emphasizing the model's ability to distinguish between defaulting and on-time borrowers. Precision and recall offer insights into positive predictions, and accuracy, while assessing overall correctness, may be misleading in imbalanced datasets.The table below provides a comparison of the different models based on these metrics, illustrating their effectiveness in predicting creditworthiness.

Table-1: Comparison of Model Performance Based on Evaluation Metrics

Model	AUC	KS
Logistic Regression (LR)	0.78	0.45
XGBoost	0.85	0.50
LSTM	0.88	0.53
AM-LSTM	0.90	0.55
FE-Transformer	0.92	0.58

This table highlights the performance differences between the models, with the FE-Transformer generally showing the best result in both AUC and KS, followed by AM-LSTM,LSTM,XGBoost, and finally Logistic Regression.At present, the ROC curve is often used to evaluate the predictive ability of models.The ROC curve, based on true positive rate (TPR) and false positive rate (FPR), visualizes model performance, with a curve closer to the upper left corner indicating better classification. The AUC value quantifies the ROC curve's proximity to perfection, offering a more precise evaluation [1,2].

5. DATA PREPARATION

The initial phase involves collecting and preparing data from diverse sources, including credit bureaus, financial statements, and loan applications. This study utilizes a dataset from an anonymous P2P lending company in China, encompassing 100,000 borrowers with feature and behavioral data. Labels (1 or 0) indicate borrower default. Sorting loans by date, the last month's data form the test set (20% of the dataset) to assess model stability. The remaining data constitute training sets for model development. User behavioral data, varying in sequence length, is standardized to fixed-length sequences through preprocessing and coding.

6. EXPERIMENTAL RESULTS

In the proposed credit scoring model, the FE-Transformer utilizes 2 transformer coding layers with 4 heads in the multi-head attention mechanism. To mitigate overfitting, a dropout of 0.3 is applied to neural units in the transformer coding layer. Model training utilizing adaptive motion estimation rules for parameter updates. The training process incorporates an early stopping strategy to address potential overfitting issues in deep learning model training.

Feature Selection is the method involves choosing a subset of the most pertinent features from the initial se, eliminating redundant, irrelevant, or noisy ones to enhance model efficiency and interpretability.

Table-2: Feature Importance Ranking Across Different Models

Feature	Logistic Regression (LR)	XGBoost	LSTM	AM-LSTM	FE-Transformer
Credit History	1	2	1	1	2
Debit-to-Income Ratio	2	1	2	2	1
Employment History	3	3	3	3	3
Income Stability	4	4	4	4	4

To demonstrate the performance superiority of the FE-Transformer approach suggested in this study, three types of experiments were conducted using different datasets. The first type focused on feature data only, employing traditional models like logistic regression and XGBoost due to the low dimensions unsuitable for training deep learning models. The second type used a dataset exclusively with behavioral data, where deep learning models directly utilized user event sequences as input, while traditional models (LR and XGBoost) required manual feature extraction. The third type utilized a dataset with all available data, employing LR, XGBoost, LSTM, AM-LSTM, and Feature Selection. Experimental results indicate that the FE-Transformer model outperforms LR, XGBoost, LSTM, and AM-LSTM in terms of AUC and KS. This suggests that the FE-Transformer deep learning model accurately predicts user default risk, contributing to reduced loan default rates and credit risk for financial enterprises, fostering their healthy and sustainable development, especially in scenarios where user behavior events may change post-APP upgrades.

7.CONCLUSIONS

The advancement of big data and artificial intelligence technology has shifted the research focus towards machine learning and deep learning models in credit scoring. This study delves into the credit scoring methods of financial enterprises, highlighting the significance of the FE-Transformer neural network model. Key findings include the innovative use of user online behavioral data as a credit scoring source, enhancing the effectiveness of user default analysis models. The FE-Transformer model outperforms other comparison methods, affirming its efficiency and feasibility in credit scoring. The model's output of user default probabilities serves as a foundation

for the loan approval decisions and risk solving, empowering financial institutions to enhance their credit risk management capabilities. As machine learning and deep learning continue to progress, the model's accuracy in risk analysis is expected to improve, along with enhanced interpretability.

For future research, several considerations emerge. Firstly, addressing the challenge of data acquisition, this experiment relies on the datasets of a single enterprise. Future work aims to explore datasets from various enterprises to enhance the generalizability of findings. Secondly, the credit scoring model in this study is static, and there is a growing interest in dynamic updates. Leveraging blockchain technology for secure, transparent, and efficient sharing of credit-related information among financial institutions is a potential avenue. This not only aids in reducing fraud but also contributes to improved model accuracy through dynamic updates. Additionally, there is a need to emphasize cybersecurity measures to safeguard sensitive credit-related data, given the rising sophistication of cyber threats in the financial sector.

REFERENCES

[1] DR D Shanti ,Pranava Bhattacharya Amity ,"Credit Score Prediction System using Deep Learning and K-Means Algorithms", August 2021

[2] C. Liberati and F. Camillo, "Personal values and credit scoring: new insights in the financial prediction, "Journal of the Operational Research Society, vol. 69, no. 12, pp. 1–21, 2018.

[3] Ashwani Kumar et al 2021 J. Phys.: Conf. Ser. 1998 012027, Credit Score Prediction System using Deep Learning and K-Means Algorithms

[4] TSAI, C. & WU, J. (2008). Using neural network ensembles for bankruptcy pre- diction and credit scoring. Expert Systems with Applications, 34, 2639–2649. 126

[5] WEST, D. (2000). Neural network credit scoring models. Computers & Operations Research, 27, 1131–1152. 5, 7, 105, 113, 126