Journal of Computer and Communications, 2025, 13(6), 33-46
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Explore the Use of Prompt-Based LLM for Credit Risk Classification

Qizhao Chen

Graduate School of Information Science, University of Hyogo, Kobe, Japan
Email: af24o008@guh.u-hyogo.ac.jp

How to cite this paper: Chen, Q. (2025) Explore the Use of Prompt-Based LLM for Credit Risk Classification. *Journal of Computer and Communications*, **13**, 33-46.
https://doi.org/10.4236/jcc.2025.136003

Received: May 6, 2025
Accepted: June 10, 2025
Published: June 13, 2025

Abstract

Credit risk assessment plays an important role in financial services by estimating the chance of a borrower defaulting. Recently, although the Large Language Models (LLMs) have demonstrated superior performance in various tasks, especially in natural language processing, their effectiveness in credit risk evaluation remains unknown. Therefore, this study explores the use of prompt-based LLMs for credit risk classification using the "Give Me Some Credit" dataset. The performance of LLM is compared with traditional models, including XGBoost, Support Vector Machine (SVM), Random Forest, and Multi-Layer Perceptron (MLP). The results show that the LLM does not outperform these traditional models in prediction accuracy. However, the LLM offers clear reasoning that can help support decisions. Furthermore, SHAP value analysis highlights the most important features affecting model predictions. Adversarial training also shows that the LLM and XGBoost have similar robustness. These findings suggest that LLMs can be used alongside traditional models to improve transparency and support financial decision-making.

Keywords

Credit Risk Classification, LLM, Random Forest, SVM, XGBoost

1. Introduction

Credit risk assessment is crucial in the financial service sector, influencing loan approvals, interest rates, and investment strategies. Financial institutions aim to reduce defaults and improve returns, driving the need for precise and efficient credit risk evaluation techniques. Historically, credit risk has been analyzed using statistical models and machine learning algorithms, including logistic regression [1], XGBoost [2], Random Forest [3], and Support Vector Machine (SVM) [4]. These methods excel at processing structured tabular data, providing high predictive accuracy when correctly adjusted. Additionally, neural network models such as Multi-Layer Perceptrons (MLPs) have been used to identify intricate, nonlinear patterns within financial data [5]. Many deep learning models can outperform traditional machine learning models and statistical algorithms in credit risk evaluation [6].

Although these conventional methods can be effective, they often require significant effort in feature engineering, data preprocessing, and hyperparameter tuning to perform at their best. Additionally, their accuracy tends to drop in real-world scenarios, especially when data patterns change or only small amounts of labeled data are accessible. As financial organizations grapple with the need for scalable, flexible, and transparent solutions, many are turning to newer modeling techniques that require less manual adjustment while still generating clear and reliable results.

At the same time, the rise of Large Language Models (LLMs) such as LLaMA, GPT, and other transformer-based architectures [7] has transformed artificial intelligence [8]. Although these models were initially designed for tasks like understanding and generating text, they've shown impressive abilities in reasoning, learning from minimal examples, and even adapting to new domains with the right prompts [9] [10]. This adaptability leads to an interesting possibility: Could well-designed prompts enable LLMs to handle classification tasks in areas beyond typical language processing, such as structured financial data?

Recent studies have begun testing ways to adapt structured data for LLMs by turning numerical features into written descriptions that these models can process. Instead of retraining the model, this method takes advantage of LLMs' ability to understand context, letting them classify data based on carefully crafted prompts. If this works well, it could provide a fast, adaptable way to assess credit risk, particularly when labeled data is limited, features change frequently, or quick implementation is needed [11].

Despite the promise, significant gaps remain in our understanding of how well prompt-based LLMs perform on structured data tasks compared to traditional machine learning models. Financial datasets pose particular challenges: they are often imbalanced, sensitive to small numerical differences, and less naturally suited to textual interpretation. Therefore, empirical studies are needed to assess the feasibility, strengths and limitations of using LLMs for tasks such as credit risk classification.

This paper addresses these gaps by making the following contributions:

1) This paper proposes a prompt-based framework for applying LLMs to structured tabular data in the context of credit risk assessment, converting financial features into structured natural language prompts.

2) This paper conducts a systematic benchmarking study, comparing the performance of a prompt-based LLM with four widely used traditional models: XGBoost, Random Forest, SVM, and a Multi-Layer Perceptron (MLP).

3) This paper analyzes the practical implications of using prompt-based LLMs,

discussing their advantages, such as flexibility and transparency, as well as challenges, such as numerical precision.

This study aims to shed light on the emerging role of LLMs in structured data classification tasks and contribute to the growing research effort to broaden the applicability of LLMs beyond their traditional NLP domains.

2. Related Work

Credit risk classification has been a critical area of research in finance, with increasing interest in applying machine learning techniques to improve predictive accuracy. For example, Chang *et al.* [12] investigate credit risk prediction for credit card customers using machine learning models such as neural networks, logistic regression, AdaBoost, XGBoost, and LightGBM. Key findings indicate that XGBoost outperforms others with 99.4% accuracy, suggesting improved predictive accuracy for informed lending decisions. Emmanuel *et al.* [13] propose a stacked classifier approach with filter-based feature selection for efficient credit risk prediction. The model uses Random Forest, Gradient Boosting, and Extreme Gradient Boosting as base estimators, achieving high performance across multiple datasets. Compared to ANN, DT, and KNN, the stacked model shows superior AUC scores. Quan and Sun [14] apply the factorization machine model to credit risk assessment and use one-hot encoding for non-numerical features. Experimental results show that the factorization machine outperforms logistic regression, SVM, k-nearest neighbors, and ANN in accuracy and computational efficiency. Melese *et al.* [15] propose a hybrid CNN-SVM/RF/DT model for credit-risk prediction in banking. Four classifiers were developed: a fully connected CNN, and hybrid models with SVM, DT, and RF. Experimental results show prediction accuracies of 86.70%, 98.60%, 96.90%, and 95.50%, respectively, with the hybrid methods outperforming the fully connected CNN. Rahayu *et al.* [16] apply Kernel Logistic Regression (KLR) to credit risk classification, optimizing parameters via grid search with 5-fold cross-validation. Using UCI machine learning datasets, KLR demonstrates promising performance compared to other machine learning techniques. Putri *et al.* [17] analyze credit risk using machine learning, specifically Support Vector Machine (SVM) with polynomial kernel achieving the highest accuracy (0.9508) and AUC (0.9419). Variables include gender, income, job, and loan history.

With the rapid development of LLMs and the advantages of prompt-based approach, an increasing number of researchers are applying these approaches to financial tasks. For example, Chen [11] uses a prompt-based LLM approach with stock image as input to predict the stock movement. Yang *et al.* [18] propose an open-source LLM, FinGPT, for different financial tasks, such as sentiment analysis and stock forecasting. Chen [19] proposes a novel framework integrating LLM with Proximal Policy Optimization (PPO) to improve stock price predictions by incorporating risk-adjusted mechanisms. Xie *et al.* [20] introduce the first financial LLM, FinMA, fine-tuned from LLaMA with 136K instruction data samples. It includes an evaluation benchmark with five NLP tasks and one prediction task. The framework aims to advance open-source financial AI research by providing datasets, benchmarks, and model results. Teixeira *et al.* [21] propose a Labeled Guide Prompting (LGP) to improve Large Language Models' (LLMs) reliability in generating credit risk reports. LGP includes annotated examples and Bayesian network descriptions to guide LLMs. Using 100 credit applications, LGP-generated reports were preferred by human analysts 60-90% of the time and showed significantly more insightful responses.

This research addresses a gap in the existing literature by providing a direct comparison between prompt-based LLMs and traditional machine learning models. Although prior studies have primarily focused on either structured data with conventional models or unstructured text with LLMs, this study bridges the two by evaluating the effectiveness of prompt-based LLMs in a structured data classification task. This comparison offers valuable insights into the current capabilities and limitations of LLMs when applied beyond their typical natural language processing domains.

3. Methodology

3.1. Data

This study utilizes the publicly available "Give Me Some Credit"[1] dataset from Kaggle, which provides real-world financial and demographic data aimed at predicting an individual's likelihood of experiencing financial distress within the next two years. The dataset consists of 150,000 observations with 11 variables, including both input features and a binary target label. Table 1 presents the summary statistics of the dataset features. Below is a brief description of each variable.

Table 1. Summary statistics of credit-related features (rounded to two decimals).

	Serious Dlqin2yrs	Revolving Util	Age	30 - 59 Days Late	Debt Ratio	Monthly Income	Open Lines Loans	90 Days Late	Real Estate Loans	60 - 89 Days Late	Dependents
count	150000.00	150000.00	150000.00	150000.00	150000.00	150000.00	150000.00	150000.00	150000.00	150000.00	150000.00
mean	0.07	6.05	52.30	0.42	353.01	5348.14	8.45	0.27	1.02	0.24	0.74
std	0.25	249.76	14.77	4.19	2037.82	13152.06	5.15	4.17	1.13	4.16	1.11
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	0.00	0.03	41.00	0.00	0.18	1550.00	5.00	0.00	0.00	0.00	0.00
50%	0.00	0.15	52.00	0.00	0.37	4357.50	8.00	0.00	1.00	0.00	0.00
75%	0.00	0.56	63.00	0.00	0.87	7400.00	11.00	0.00	2.00	0.00	1.00
max	1.00	50708.00	109.00	98.00	329664.00	3008750.00	58.00	98.00	54.00	98.00	20.00

[1]https://www.kaggle.com/c/GiveMeSomeCredit.

The input features are as follows:

- **Revolving Utilization of Unsecured Lines**: Total balance on credit cards and personal lines of credit relative to total credit limits (expressed as a percentage).
- **Age**: Age of the individual in years.
- **Number of Time 30 - 59 Days Past Due Not Worse**: Number of times the individual has been 30 - 59 days past due on a payment but not worse, in the past two years.
- **Debt Ratio**: Monthly debt payments, alimony, living costs divided by monthly gross income.
- **Monthly Income**: Monthly income of the individual.
- **Number of Open Credit Lines and Loans**: Total number of open credit lines (e.g. credit cards, installment loans, etc.).
- **Number of Times 90 Days Late**: Number of times the individual has been 90 days or more past due.
- **Number Real Estate Loans or Lines**: Number of mortgage and real estate loans the individual holds.
- **Number of Time 60-89 Days Past Due Not Worse**: Number of times the individual has been 60-89 days past due but not worse.
- **Number of Dependents**: Number of dependents claimed by the individual (e.g. children, family members).

The target variable is:

SeriousDlqin2yrs: A binary indicator where 1 signifies that the individual experienced serious financial distress within two years (such as bankruptcy, foreclosure, or charge-offs), and 0 indicates no financial distress.

80% of the data are used for training and the remaining 20% of the data are used for testing.

Addressing Data Imbalance Using SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) is a widely used method for addressing class imbalance in classification tasks. The core idea behind SMOTE is to generate synthetic samples for the minority class by interpolating between existing minority class samples. Instead of merely duplicating the minority class instances, SMOTE creates new, unique data points by selecting a random neighbor from the same class and generating synthetic examples along the line connecting the selected sample and its neighbor. This process increases the diversity of the minority class, which can help improve the performance of classifiers by providing them with more balanced and varied data. In the context of the Give Me Some Credit dataset, where the number of defaulters (minority class) is significantly smaller than the non-defaulters (majority class), SMOTE helps mitigate the risk of the model being biased toward predicting the majority class. By generating new, plausible instances of the minority class, SMOTE allows the model to better learn the underlying patterns that differentiate defaulters from non-defaulters, leading to more accurate and generalizable predictions.

3.2. Prompt-Based LLM Approach

The LLM used in this study is DeepSeek-r1-distill-llama-70b. This version of DeepSeek model is from the American AI company Groq[2]. Although fine-tuning is not available, the DeepSeek-r1-distill-llama-70b has been proven to be able to generate competitive performance in stock price forecasting task without fine-tuning compared to models such as LLaMA and Gemma [22]. This model employs knowledge distillation, a technique where a compact model (student) learns by mimicking the outputs of a sophisticated, larger model (teacher). Here, the LLaMA-70B is used as the teacher model to train the student model DeepSeek to closely match its responses. The teacher produces detailed predictions that help the student learn more effectively. This approach maintains most of the teacher's capabilities while creating a streamlined model that can operate efficiently even with constrained computing power.

A structured prompt was designed to guide the LLM in generating classification results. Table 2 presents the structured prompt template. The relevant information from the "Give Me Some Credit" dataset was incorporated into the prompt to describe each applicant's financial profile.

Table 2. Prompt for credit risk classification.

You are a credit risk assessment expert.

An applicant has the following financial information:

- Revolving Utilization of Unsecured Lines: {row['RevolvingUtilizationOfUnsecuredLines']:.2f} (ratio between 0 and 1)
- Age: {row['age']} years
- Number of Times 30 - 59 Days Past Due (Not Worse) in the last 2 years: {row['NumberOfTime30-59DaysPastDueNotWorse']}
- Debt Ratio (monthly debt payments/monthly gross income): {row['DebtRatio']:.2f}
- Monthly Income: ${row['MonthlyIncome']}
- Number of Open Credit Lines and Loans: {row['NumberOfOpenCreditLinesAndLoans']}
- Number of Times 90 Days Late: {row['NumberOfTimes90DaysLate']}
- Number of Real Estate Loans or Lines: {row['NumberRealEstateLoansOrLines']}
- Number of Times 60 - 89 Days Past Due (Not Worse) in the last 2 years: {row['NumberOfTime60-89DaysPastDueNotWorse']}
- Number of Dependents: {row['NumberOfDependents']}

Based on this information, predict whether this applicant is at **High Risk** (likely to default within 2 years) or **Low Risk** (not likely to default within 2 years).

Only answer with **"High Risk"** or **"Low Risk"**.

3.3. Baseline Models

The baseline models chosen for comparison include XGBoost, SVM, Random Forest

[2]https://groq.com/.

and MLP. The following is a brief description of each model.

- **XGBoost** is a scalable and efficient gradient boosting framework widely recognized for its superior performance on structured tabular data. It builds an ensemble of decision trees sequentially, where each new tree corrects the errors made by the previous ones. With its ability to handle missing values, regularization techniques, and parallelized implementation, XGBoost often serves as a strong benchmark for classification tasks.
- **Random Forest** works by building hundreds of decision trees during training, then combining their predictions through majority voting. It introduces randomness in two ways: by using different data samples (bootstrapping) and selecting random features at each decision point. This dual randomness helps prevent overfitting while boosting the model's ability to generalize.
- **Support Vector Machine (SVM)** algorithm works by identifying the best possible dividing line (hyperplane) that creates the widest separation between different classes. What makes SVM powerful is its effectiveness with high-dimensional data and its smart use of kernel functions to navigate complex, nonlinear patterns. For our analysis, we first standardized all input features through scaling, a crucial step that significantly improved SVM's classification performance.
- **Multi-Layer Perceptron (MLP)** is a class of feed-forward neural networks capable of modeling complex nonlinear relationships. The MLP architecture used in this study consists of two hidden layers with 64 and 32 neurons respectively, using ReLU activation functions and trained with backpropagation.

3.4. Evaluation

All baseline models are evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC) along with classification reports, including precision, recall, and F1-score. The LLM is only evaluated using the classification reports because the prompt-based LLM directly outputs the class labels without predicted probabilities, which are necessary to calculate the AUC.

The AUC metric measures the ability of a model to distinguish between classes across all possible classification thresholds. An AUC of 1.0 indicates perfect classification, while an AUC of 0.5 suggests no discriminative ability, equivalent to random guessing. AUC is particularly valuable for imbalanced datasets, as it considers the trade-off between true positive and false positive rates.

Precision represents the proportion of correctly predicted positive observations out of all predicted positive observations. High precision indicates that the model makes few false positive errors, which is crucial when the cost of misclassifying a low-risk borrower as high-risk is high.

Recall, also known as the sensitivity or true positive rate, measures the proportion of actual positive cases that were correctly identified by the model. High recall ensures that most people at risk of financial distress are correctly detected, minimizing missed high-risk cases.

F1-score is the harmonic mean of precision and recall, providing a single metric that balances both concerns. It is especially useful when there is an uneven class distribution, ensuring that neither precision nor recall is ignored when evaluating model performance.

4. Results

Table 3 shows the performance of the credit risk classification of both the baseline models and LLM. Figure 1 visualizes the performance of all models.

The baseline models, including XGBoost, Random Forest, SVM, and MLP, all demonstrated strong and consistent performance in precision, recall, and F1 score, indicating their reliability in classifying credit risk. Among the baseline models, Random Forest has the best performance in terms of all evaluation metrics, followed by XGBoost. SVM has the worst performance. Furthermore, the prompt-based LLM approach cannot outperform the baseline models. Specifically, it achieved a high precision but significantly lower recall and F1-score. This means that while the LLM was accurate when predicting a high-risk case (low false positives), it failed to identify many actual high-risk cases (high false negatives). This performance could be attributed to the model's limited ability to generalize

Table 3. Performance comparison of baseline models and prompt-based LLM.

Model	AUC	Precision	Recall	F1-Score
XGBoost	0.95	0.88	0.88	0.88
Random Forest	0.97	0.91	0.91	0.91
SVM	0.82	0.74	0.74	0.74
MLP	0.93	0.85	0.85	0.85
Prompt-based LLM	-	0.91	0.45	0.58

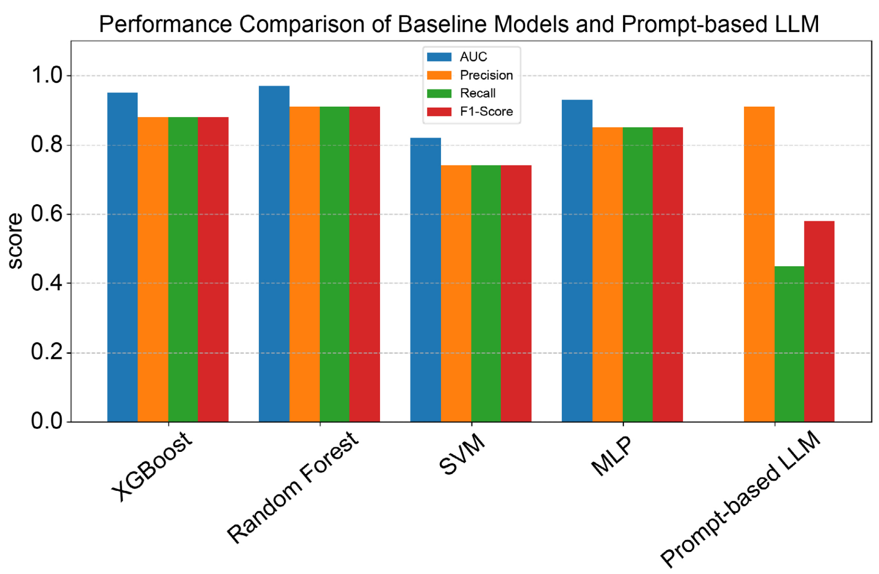

Figure 1. Performance comparison of baseline models and LLM.

from prompts crafted from tabular data. These results highlight the current limitations of prompt-based LLMs in structured data tasks, especially in scenarios where recall is critical for risk management.

To improve credit risk assessment, the prompt-based LLM approach can be used together with other machine learning models because the reasoning process provided by LLM can also be considered to improve the financial decision-making. For example, Table 4 is a sample of the LLM response.

Specifically, the LLM demonstrates a step-by-step, rule-based reasoning process by sequentially analyzing each financial variable and interpreting its implications for credit risk. It considers both quantitative thresholds (e.g. high revolving credit utilization at 77% and a debt ratio of 0.80) and qualitative context (e.g. viewing two late payments as moderate but notable). The model weighs both risk-enhancing factors (high debt obligations, multiple credit lines) and mitigating ones (absence of severe delinquencies, high income), synthesizing them into an overall judgment. Rather than relying on a black-box prediction, the LLM mimics human-like reasoning, integrating multiple dimensions of the applicant's financial profile to justify its final decision regarding default risk.

Table 4. LLM response for credit risk assessment.

LLM Response
Okay, so I need to assess the credit risk of this applicant. Let me go through each of the provided financial metrics one by one to understand the overall picture.
First, the Revolving Utilization of Unsecured Lines is 0.77. That means the applicant is using 77% of their available credit. High utilization can be a sign of financial stress because it indicates they might be relying heavily on credit, which could lead to difficulty.
Next, the applicant is 45 years old. Age can be a factor, but it's not as critical as other metrics. It might just indicate they've had time to build credit history, but without knowing the actual credit score, it's hard to judge.
The number of times they've been 30-59 days past due in the last two years is 2.0. Late payments, even if they're not more severe, can negatively affect credit scores. Two instances might not be too bad, but it's something to consider alongside other factors.
The Debt Ratio is 0.80, which is quite high. This means 80% of their monthly gross income goes towards debt payments. A high debt ratio can limit their ability to take on more debt or handle unexpected expenses, increasing the risk of default.
The Monthly Income is $9120.0, which seems substantial. However, with a high debt ratio, the actual disposable income might be low, which could strain their finances.
They have 13 open credit lines and loans. That's a lot, which could indicate they're managing multiple debts, possibly leading to higher risk if they can't keep up with payments.
There are zero instances of being 90 days late, which is a positive sign. More severe delinquencies are a bigger red flag, so this is good.
They have 6 real estate loans or lines, which might mean they have significant assets, but it also adds to their debt obligations. More debts mean more financial commitments each month.
There are zero instances of 60 - 89 days past due, which is another positive point. It shows they haven't had more severe late payments recently.
Lastly, they have 2 dependents. This could affect their financial situation as dependents can increase expenses, but it's not directly a credit risk factor unless it impacts their ability to repay debts.
Putting it all together, the high revolving utilization, high debt ratio, and multiple open credit lines are concerning. While the income is high, the debt ratio suggests they might be overextended. The few late payments add to the risk. Despite the lack of severe delinquencies, the overall picture leans towards a higher risk of default.

5. Further Discussion

5.1. Feature Importance and SHAP Values

To further investigate how different features contribute to credit risk prediction, SHAP (SHapley Additive exPlanations) values are used with the XGBoost model. SHAP values help explain the output of a machine learning model by showing how much each feature increases or decreases the prediction. This method provides clear insights into which features are most important in the model's decisions. In order to avoid misleading interpretations, the original data without SMOTE processing is used because SMOTE will generate synthetic data points that do not correspond to real-world observations.

Figure 2 is a SHAP summary plot, which helps us understand how each feature influences the model's prediction of credit risk. Each dot represents one individual, and its position on the x-axis shows how much that feature affected the model's output (SHAP value). If the SHAP value is positive, it means the feature pushes the prediction toward higher credit risk; if it is negative, it pushes the prediction toward lower risk. The color of each dot indicates the value of the feature: red means high, and blue means low.

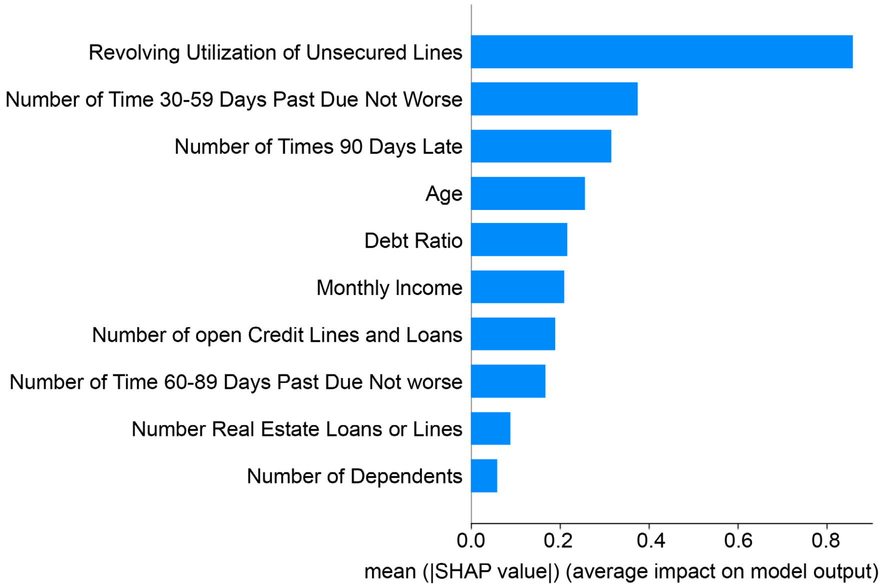

Figure 2. SHAP Values.

From the plot, we observe that **Revolving Utilization of Unsecured Lines** has the largest impact on the model's predictions. Higher values of this feature (shown in red) are associated with higher SHAP values, meaning they increase the predicted risk. Similarly, **Number of Time 30 - 59 Days Past Due Not Worse** and **Number of Times 90 Days Late** also have strong positive effects on risk when their values are high. This suggests that past payment delays are strong indicators of future credit problems. On the other hand, **age** shows the opposite trend: younger individuals (blue) tend to have higher predicted risk, while older individuals (red) tend to lower it.

Other features like **Debt Ratio**, **Monthly Income**, and **Number of Open Credit Lines and Loans** contribute to the prediction, but with less impact. Features such as **Number Real Estate Loans or Lines** and **Number of Dependents** appear to have a minimal influence overall. This analysis highlights that high credit utilization and frequent past delinquencies are the most important warning signs for credit risk in this dataset.

The feature importance can also be seen in Figure 3, which shows the average impact of each feature on the predictions of the model. Each bar represents a feature, and the length of the bar reflects the mean absolute SHAP value. That is, how much, on average, the feature changes the model's output. A longer bar means the feature has a bigger influence.

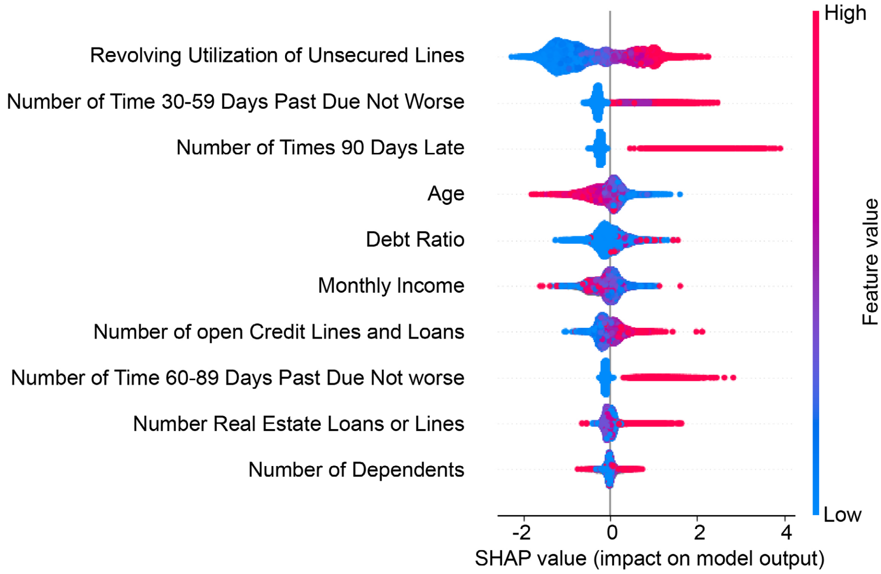

Figure 3. SHAP values, bar chart.

5.2. Adversarial Training

In order to further analyze the robustness of the LLM, adversarial training is applied. The performance of LLM is compared with XGBoost. Adversarial training involves adding small, intentional changes to the input data to test whether the model can still make correct predictions. By training both models on these adversarial examples, we can assess how well each model handles challenging or manipulated inputs.

The adversarial examples are generated by slightly modifying the most important input features identified through SHAP values. Specifically, we first compute the average absolute SHAP values across all training samples and select the top five most influential features.

A small perturbation is applied to these selected features in the training set. For each instance, we multiply the feature value by a random factor between 0.95 and

DOI: 10.4236/jcc.2025.136003

1.05 (*i.e.* ±5% change). This simulates subtle but realistic shifts in the input distribution. To maintain data validity, all perturbed values are clipped to the original range observed in the training data.

The perturbed dataset is then used to train the models under adversarial conditions. By comparing the performance of XGBoost and the LLM on both clean and adversarial data, we can assess which model better resists performance degradation when faced with small input disturbances.

Table 5 shows the model performance of the adversarial training for both LLM and XGBoost. The results show that the performances of both XGBoost and LLM under adversarial training are almost the same compared to when using the clean data (Table 3). This suggests that both models have comparable robustness to small perturbations in the input data. This indicates that the added complexity and capacity of the LLM do not necessarily translate into stronger resistance to adversarial changes, at least in this setting. It also implies that simpler models like XGBoost can perform competitively in terms of robustness, making them a practical choice when computational resources or interpretability are a concern.

This finding highlights the importance of not only focusing on model accuracy but also considering robustness when selecting models for real-world deployment, especially in sensitive domains.

Table 5. Adversarial training performance comparison.

Model	AUC	Precision	Recall	F1-Score
XGBoost	0.95	0.88	0.88	0.88
Prompt-based LLM	-	0.90	0.48	0.63

6. Conclusions

This paper compares the performance of a prompt-based LLM approach with traditional machine learning models, including XGBoost, SVM, Random Forest, and MLP for credit risk classification. Experimental results show that the LLM does not outperform traditional models in terms of classification accuracy. Among the baseline models, Random Forest has the best performance, while SVM has the worst performance. Furthermore, SHAP value analysis reveals that features related to credit utilization and past-due history have the greatest impact on model predictions. Furthermore, adversarial training using small perturbations on the top SHAP-ranked features demonstrates that both the LLM and XGBoost exhibit similar levels of robustness. This suggests that while the LLM offers enhanced interpretability through its reasoning process, traditional models like XGBoost remain competitive and efficient, especially in robustness and performance. Therefore, the LLM's output can serve as a complementary interpretive tool alongside traditional models to support more transparent and informed financial decision-making.

Future studies could explore how prompt design influences the performance of Large Language Models (LLMs). Additionally, various LLMs may be incorporated for comparative analysis.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Sharma, V., Singh, A., Saxena, A.K. and Saxena, V. (2023) A Logistic Regression Based Credit Risk Assessment Using Woe Bining and Enhanced Feature Engineering Approach ANOVA and Chi-Square. 2023 12*th International Conference on System Modeling & Advancement in Research Trends* (*SMART*), Moradabad, 22-23 December 2023, 499-507. https://doi.org/10.1109/smart59791.2023.10428399

[2] Wang, K., Li, M., Cheng, J., Zhou, X. and Li, G. (2022) Research on Personal Credit Risk Evaluation Based on XGBoost. *Procedia Computer Science*, **199**, 1128-1135. https://doi.org/10.1016/j.procs.2022.01.143

[3] Kurniawan, R. (2024) Application of Random Forest Algorithm on Credit Risk Analysis. *Procedia Computer Science*, **245**, 740-749. https://doi.org/10.1016/j.procs.2024.10.300

[4] Xu, W., Zhou, S., Duan, D. and Chen, Y. (2010) A Support Vector Machine Based Method for Credit Risk Assessment. 2010 *IEEE* 7*th International Conference on E-Business Engineering*, Shanghai, 10-12 November 2010, 50-55. https://doi.org/10.1109/icebe.2010.44

[5] Yu, Z. (2023) Prediction of Credit Card Loan Risk Based on Multilayer Perceptron Neural Network Model. *BCP Business & Management*, **38**, 126-134. https://doi.org/10.54691/bcpbm.v38i.3679

[6] Shi, S., Tse, R., Luo, W., D'Addona, S. and Pau, G. (2022) Machine Learning-Driven Credit Risk: A Systemic Review. *Neural Computing and Applications*, **34**, 14327-14339. https://doi.org/10.1007/s00521-022-07472-2

[7] Chen, Q. and Kawashima, H. (2025) A Novel Sentiment Correlation-Based Method with Dual Transformer Model for Stock Price Prediction. https://doi.org/10.21203/rs.3.rs-6479946/v1

[8] Chen, Q. (2025) Comparing Different Transformer Model Structures for Stock Prediction. arXiv: 2504.16361.

[9] Chen, Q. and Kawashima, H. (2024) Stock Price Prediction Using LLM-Based Sentiment Analysis. 2024 *IEEE International Conference on Big Data* (*Big Data*), Washington, 15-18 December 2024, 4846-4853. https://doi.org/10.1109/bigdata62323.2024.10825946

[10] Chen, Q. (2025) Comparing Vision-Instruct LLMs, Vision-Based Deep Learning, and Numeric Models for Stock Movement Prediction. *International Journal of Advanced Computer Science and Applications*, **16**, 11-18. https://doi.org/10.14569/ijacsa.2025.0160402

[11] Chen, Q. (2025) Image-Driven Stock Price Prediction with Llama: A Prompt-Based Approach. *International Journal of Modeling and Optimization*, **15**, 17-24. https://doi.org/10.7763/ijmo.2025.v15.867

[12] Chang, V., Sivakulasingam, S., Wang, H., Wong, S.T., Ganatra, M.A. and Luo, J. (2024) Credit Risk Prediction Using Machine Learning and Deep Learning: A Study on Credit Card Customers. *Risks*, **12**, Article 174. https://doi.org/10.3390/risks12110174

[13] Emmanuel, I., Sun, Y. and Wang, Z. (2024) A Machine Learning-Based Credit Risk Prediction Engine System Using a Stacked Classifier and a Filter-Based Feature Selection Method. *Journal of Big Data*, **11**, Article No. 23. https://doi.org/10.1186/s40537-024-00882-0

[14] Quan, J. and Sun, X. (2024) Credit Risk Assessment Using the Factorization Machine Model with Feature Interactions. *Humanities and Social Sciences Communications*, **11**, Article No. 234. https://doi.org/10.1057/s41599-024-02700-7

[15] Melese, T., Berhane, T., Mohammed, A. and Walelgn, A. (2023) Credit-Risk Prediction Model Using Hybrid Deep-Machine-Learning Based Algorithms. *Scientific Programming*, **2023**, 1-13. https://doi.org/10.1155/2023/6675425

[16] Rahayu, S.P., Zain, J.M., Embong, A. and Purnami, S.W. (2010) Credit Risk Classification Using Kernel Logistic Regression with Optimal Parameter. 10*th International Conference on Information Science*, *Signal Processing and their Applications* (*ISSPA* 2010), Kuala Lumpur, 10-13 May 2010, 602-605. https://doi.org/10.1109/isspa.2010.5605437

[17] Putri, N.H., Fatekurohman, M. and Tirta, I.M. (2021) Credit Risk Analysis Using Support Vector Machines Algorithm. *Journal of Physics*: *Conference Series*, **1836**, Article ID: 012039. https://doi.org/10.1088/1742-6596/1836/1/012039

[18] Yang, H., Liu, X. and Dan Wang, C. (2023) FinGPT: Open-Source Financial Large Language Models. *SSRN Electronic Journal*.

[19] Chen, Q. (2025) A Two-Stage Framework for Stock Price Prediction: LLM-Based Forecasting with Risk-Aware PPO Adjustment. *Journal of Computer and Communications*, **13**, 120-139. https://doi.org/10.4236/jcc.2025.134008

[20] Xie, Q.Q., Han, W.G., *et al*. (2023) PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance. arXiv: 2306.05443.

[21] Teixeira, A.C., Marar, V., Yazdanpanah, H., Pezente, A. and Ghassemi, M. (2023) Enhancing Credit Risk Reports Generation Using LLMs: An Integration of Bayesian Networks and Labeled Guide Prompting. 4*th ACM International Conference on AI in Finance*, Brooklyn, 27-29 November 2023, 340-348. https://doi.org/10.1145/3604237.3626902

[22] Chen, Q. (2025) Stock Price Change Prediction Using Prompt-Based LLMs with RL-Enhanced Post-Hoc Adjustments. In: *Advances in Intelligent Systems Research*, Atlantis Press International BV, 475-483. https://doi.org/10.2991/978-94-6463-742-7_46