Evaluating Credit Risk Using Interpretable Machine Learning Models

Abstract -- To ensure transparency and auditability in credit risk assessment, balancing predictive accuracy with regulatory requirements is essential for modern credit scoring models. With such a goal, this research explores strategies for enhancing the predictive accuracy of simple and transparent models such as logistic regression and support vector machines (SVMs), attempting to achieve performance comparable to advanced machine learning techniques without sacrificing transparency. In this work, we conducted experiments on real-world credit datasets which demonstrated 71% accuracy for logistic regression and 72% accuracy for the SVM model highlighting their baseline effectiveness. To address the accuracy gap compared to advanced model, we worked on developing an ensemble approach that improves predictive performance while maintaining interpretability using frameworks like Locally Interpretable Model-Agnostic Explanations (LIME). The results highlight the feasibility of designing credit scoring models that simultaneously achieve high accuracy and explainability for regulatory compliance and decision-making in financial applications.

Keywords— Fintech, Banking, Credit Risk, Black Box, Ensemble.

# Introduction

The growing reliance on predictive models for decision-making in financial services, such as credit risk assessment has significantly been a key player in the advancement of interpretability in the field of machine learning [1][2]. Credit scoring models are required to provide a delicate balance between maximally accurate risk prediction and regulators demand for these models to be transparent and auditable. These principles of transparency and interpretability are critical in today's financial environment as more and more credit scoring models depend on very complex algorithms and machine learning techniques. In the United States, the Fair Credit reporting Act (FCRA) mandates that clients must understand the parameters that affect their credit scores. Additionally, we have the General Data Protection Regulation (GDPR) based in the European Union which allows consumers to clearly understand algorithms and machine learning decisions that can impact them, therefore enhancing transparency. Despite the availability of advanced machine learning techniques, simpler models like logistic regression and decision trees remain widely used in credit scoring due to their inherent interpretability. However, these simpler predictive models often fail to capture the full predictive potential of recent datasets. On the other hand, cutting edge methods like gradient boosting and neural networks perform with a significantly higher accuracy but are often questioned and termed as “black boxes”, limiting their applicability in regulated environments [1].

In this research, we explore the predictive accuracy of interpretable models such as logistic regression and SVM for credit risk assessment, compare their performance against ensemble methods, and explore frameworks such as LIME to bridge the gap between accuracy and interpretability. Our main goal is to contribute to the advancement of credit risk assessment techniques by offering efficient solutions to enhance transparency without undermining performance. Therefore, this research focuses on addressing the following key questions:

1. Which transparent models are most effective for assessing credit risk?
2. Does adding model complexity improve the performance of the credit risk assessment model?
3. How can complex models be used for credit risk assessment while ensuring interpretability?

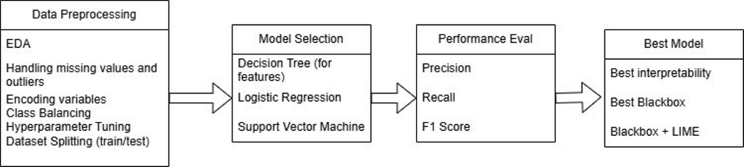
# Literature review

Through our literature review, we identified several promising machine learning models that can potentially assess credit risk with great accuracy. One such class of models is ensemble model, such as Deep Multiple Kernel Classifier, LASSO-logistic Regression Ensemble, Correlated-adjusted Decision Forest and many more. In their review paper [3], authors explored these models in-depth. The authors compared these models to each other, noting their strengths and weaknesses whenever necessary. Another such model is XGBoost, which was studied in detail by authors in [4]. Ensemble models work well with financial data due to their robustness and prediction accuracy. They improve upon base learners such as decision trees and logistic regression models. These models are also generally resilient against overfitting; thus, they are good candidate models for working with financial data.

In [5] authors explore different machine learning algorithms that are currently employed for assessing credit risk. Authors also studied a novel algorithm called Factorization Machine (FM), a supervised machine learning algorithm that combines the strengths of SVM and matrix factorization. This unique combination increases a model’s usability with sparse data. It also improves optimization time and makes the training process uncomplicated. The article also discusses data preprocessing for encoding categorical data and fixing missing values and outliers. The author draws parallels between FM and SVM, KNN and ANN (artificial neural networks). The article contains valuable trove of information that will be indispensable when we formulate our own research methodology.

Another method for assessing credit risk is through back propagation algorithms such as Multi-Layer Perceptron (MLP). MLP falls under deep learning algorithms and is not strictly within the conventional machine learning domain. The application of MLP for credit risk assessment was explored by the authors in [6]. In this article, authors found that MLP could potentially assess credit risk with an accuracy of 87%. They trained their model using German credit dataset. The model had 6 to 39 hidden layers, which enables the model to work with the vast German credit data. This model has superior predictive power compared to many black-box models. A point to be noted here though, is the high computational power requirements of deep learning algorithms. Thus, the scope of this research will not include developing deep learning models.

Most of the models discussed in this section and too complex and are not transparent. To address this, we focus on developing simpler, transparent models and optimizing them to achieve maximum accuracy while maintaining their transparency.



1. Steps for designing the credit risk assessment model.

# Methodology

In this study, we have implemented and assessed machine learning models for credit risk prediction. The aim of this study is to explore simple machine learning algorithms that will also enhance interpretability for us to maintain regulatory compliance with the financial sector. At the same time achieving and selecting a model with the highest performance (accuracy). Figure 1 represents the flowchart of steps that we took to achieve our goals.

## Dataset

In this study we used the dataset from [7]. The dataset provides a classification of people based on different profile features as either good or bad credit. The dataset contains information on 1000 individuals through 21 continuous and categorical features. The positive (good) class has a distribution of 70% and negative (bad) class of 30% of the dataset. A visual of the dataset in a two-dimensional space is shown in Figure 2. This dimensionality reduction was conducted using principal component analysis for the sole purpose of giving a glimpse of how the data distribution is on a two-dimensional screen. For more comparative distribution of the data on the two-dimensional feature space, an SVM decision boundary with the distribution of the data is fitted during model building and training of the SVM gaussian kernel.

## Data Preprocessing

Data preprocessing is a vital phase when implementing machine learning algorithms. Preprocessing ensures there is no noise (variability) and biasness that could cause generalization error making our model perform decimally.

A diagram of a credit risk data

Description automatically generated

1. Distribution of the data in a 2D feature space.

We performed exploratory data analysis to identify the distributions of different features in the dataset. We then checked for missing values and inputted them with the mean for numerical features and used a function transformer for categorical features to help reduce cardinality and impute them with mode.

After cleaning the data, to select the feature we fitted the data to a decision tree model to determine the importance ranking of the features in predicting the target variable. The importance ranking for the features is shown in Fig.3. From the figure, we can see that features like “Credit Amount” and “Credit History” are highly important for predicting the target variable. We select only the features that has importance > 0 for constructing our models. This step of feature selection is crucial in balancing variability and bias error to ensure generalization of the machine learning model by reducing the risk of overfitting. A decision tree model is used since it is insensitive to noises in the data.

## Model Building & Training

After the data preprocessing stage, we then split the dataset into a training and testing set. Before using the training set to train our models, we scaled the data for faster convergence. Scaling the dataset after the split also ensures that there was no leakage between the training and test data.

To get more descriptive and accurate picture of the distribution of the data between the two classes we construct a SVM model using linear kernel. The decision boundary in a two-dimension features space is shown in fig. 4. This distribution shows that it is impossible to classify this data accurately using a linear kernel.

A graph with blue bars

Description automatically generated

1. Chart demonstrating feature importance.

A collage of images of red and blue spheres

Description automatically generated

1. Decision Boundary for different C (regularization parameter) values for Gaussian SVM. Top row: C = 0.01, 0.1, 1 (left to right). Bottom row: C = 10, 100 (left to right)

We have also experimented with gaussian kernel. We vary the regularization parameter C for the gaussian kernel and show the effects of parameter on the decision boundary in Fig. 5. We have experimented with different C values ranging from 0.01, 0.1, 1, 10, 100. From the figure, it can be seen that an increase in C parameter beyond an optimized threshold results in an overfitting of the model as shown below.

# Model Evaluation

In this section we discuss the evaluation metrics used for the study, performance comparison of different machine learning models, and how we can use black box ensemble models for interpretability.

## Evaluation Metrics

In this study, we evaluate the models using the accuracy, recall and f1-score. These metrices are defined below.

Accuracy is measured as the ratio of correctly predicted classes to the total number of classes and can be calculated using the equation (1)

*accuracy =* (1)

Recall is the proportion of true examples that was classified as true positive and is calculated using equation (2).

A diagram of a training set

Description automatically generated

1. Decision Boundary for the SVM model.

*recall =* (2)

The f1-score balances the trade-off between precision and recall and is the harmonic mean of precision and recall. The f1-score is computed using equation (3).

1. Comparative Performance Analysis of Logistic Regression, SVM and Random Forest

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Kernel** | **Hyperparameter Tuning** | **Accuracy** | **Recall** | **f1-score** |
| *Logistic Regression* | None | No | 66% | 66% | 66% |
| Yes | **71%** | **71%** | **69%** |
| *SVM* | Linear | No | 62% | 62% | 62% |
| Yes | **68%** | **68%** | **68%** |
| Gaussian | No | 66% | 66% | 66% |
| Yes | **72%** | **72%** | **69%** |
| *Random Forest* |  | No | **76%** | **76%** | **76%** |

*f1-score* = (3)

## Comparative Study

In this work, we have experimented with logistic regression model, SVM model and a random forest model. The results for the evaluation metrices are shown in Table I.

For the SVM model we experimented with different kernels: linear and gaussian. We also experimented with the regularization hyperparameter for these SVM models. With hyperparameter tuning, it can be seen that both SVM models (with linear and gaussian kernel) performs better with than the no-tuned version. From the table, it can be also observed that the tuned logistic regression model performs better than the SVM model with linear kernel, however the SVM with gaussian kernel performs slightly better and has better accuracy.

Although, SVM models with gaussian kernel performs the best among the three models, the logistic regression model and SVM model with linear kernel are much simpler, provide better interpretability and are faster to train. Hence, from an interpretability perspective, we chose the logistic regression model over the SVM model for credit risk assessment.

We have also experimented with blackbox models like random forest model. The random forest model is not tuned. However, even without tuning, the performance of the random forest model is better than all other models. However, such blackbox model is less interpretable compared to a logistic regression model. To make such blackbox models interpretable we experiment with Local Interpretable Model-Agnostic Explanations (LIME) framework [8]. The results are discussed in the next subsection.31

1. Data for a sample loan application

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Duration** | **Purpose of credit** | **Credit Amount** | **Present Employment** | **Installment rate** | **Personal status** | **Age** | **Account**  **status** |
| 125 | 12 | New car | 2121 | 1<=X<4 | 4 | Male single | 30 | good |

## Interpreting the outputs of Ensemble using LIME

Our main goal being on providing an interpretable yet accurate model for credit assessment, we decided to focus on making the ensemble model (random forest) interpretable instead of focusing on tuning it further. To make the random forest model interpretable, we combined the model with a LIME instance to demonstrate how an ensemble could be interpreted hence providing both accuracy and interpretability.

LIME is a framework designed to make machine learning models more interpretable, especially complex ensemble or black box models like random forest. LIME functions by making small changes in the data around a particular prediction and observing how the model’s output responds to those changes. Then a simpler model like a linear regression is used to explain which features had the most effect on that specific prediction. In our case, we built our LIME framework on top of the random forest model to show how it works.

As an example of how LIME can be used to interpret the results of the random forest model, we chosen an instance of a 30-year-old single male requesting a 2,121 loan for a new car to be repaid in a 12-month duration with a skilled employment of less than 4 years. This sample datapoint is shown in Table II. After passing this datapoint through the random forest model and the LIME framework, we see that the probability of a good class is 83%, while the bad class has the probability of 17%. The model has therefore predicted good for credit assessment. The features with the biggest effect on the decision is the duration in months he is taking to pay the loan which has increased his chances significantly followed closely by the purpose of the credit and the credit amount. We can also classify the features in either the good or bad classes based for prediction.

A screenshot of a graph

Description automatically generated

1. Output from the LIME instance demonstrating the contributing features and their contribution in the decision.

# Conclusion

Our aim of this research was to use interpretable machine learning models to assess credit risk. We found the accuracy and precision for simple transparent models like logistic regression and SVM with linear kernel to assess credit risk was inadequate. To improve the performance further we experimented with SVM model with non-linear kernels like gaussian kernel. We also implemented a simple ensemble model like random forest and found that the performance can be further improved by adding model complexity which is consistent with existing result. However, this comes at a cost of reduced interpretability. To improve the interpretability of the random forest we integrated it with the LIME framework. We demonstrate that by integrating such framework we can design both interpretable and accurate models for credit risk assessment.

##### References

1. M. Bücker, G. Szepannek, A. Gosiewska, and P. Biecek, “Transparency, auditability, and explainability of machine learning models in credit scoring,” Journal of the Operational Research Society, vol. 73, no. 1, pp. 70–90, 2022. DOI: [10.1080/01605682.2021.1922098](https://doi.org/10.1080/01605682.2021.1922098).
2. O. A. Bello, “Machine Learning Algorithms for Credit Risk Assessment: An Economic and Financial Analysis,” International Journal of Management Technology, vol. 10, no. 1, pp. 109–133, 2023.
3. A. A. Montevechi, R. de Carvalho Miranda, A. L. Medeiros, and J. A. B. Montevechi, “Advancing Credit Risk Modelling with Machine Learning: A Comprehensive Review of the State-of-the-Art,” Engineering Applications of Artificial Intelligence, vol. 137, p. 109082, 2024. DOI: 10.1016/j.engappai.2023.109082.
4. N. Suhadolnik, J. Ueyama, and S. Da Silva, “Machine learning for enhanced credit risk assessment: An empirical approach,” Journal of Risk and Financial Management, vol. 16, no. 12, 2023. DOI: [10.3390/jrfm16120496](https://doi.org/10.3390/jrfm16120496).
5. J. Quan and X. Sun, “Credit risk assessment using the factorization machine model with feature interactions,” Humanities and Social Sciences Communications, vol. 11, p. 234, 2024. DOI: [10.1057/s41599-024-02700-7](https://doi.org/10.1057/s41599-024-02700-7).
6. Z. Zhao, S. Xu, B. H. Kang, M. M. J. Kabir, Y. Liu, and R. Wasinger, “Investigation and improvement of multi-layer perceptron neural networks for credit scoring,” Expert Systems with Applications, vol. 42, no. 7, pp. 3508–3516, 2015. DOI: [10.1016/j.eswa.2014.12.006](https://doi.org/10.1016/j.eswa.2014.12.006).
7. S. Rai, “Credit risk dataset,” Kaggle.com, 2024. [Online]. Available: <https://www.kaggle.com/datasets/satyajeetrai/credit-risk-dataset>.
8. M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should I trust you?’ Explaining the predictions of any classifier,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.