Project Proposal on Credit Score Prediction Using Machine Learning and Deep Learning

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Jay kona | Mayauri | Malvika | Usha | Uday | Charan | Revanth |
| skona@gmu. edu | Mjadhav2 @gmu.edu | mkaringu@gmu  .edu | ulalam@g mu.edu | ukola@gmu. edu | ckurakul@gmu. edu | rkarumaj@gmu. edu |

***Abstract*—**This project focuses on the prediction of credit scores using both machine learning and deep learning techniques to enhance financial risk assessment and decision- making. The initial phase involved rigorous data preprocessing, including the elimination of missing and negative values, removal of irrelevant features, and imputation using K- Nearest Neighbors (KNN). A chi-squared test was applied to identify statistically significant features contributing to the target variable— credit score. Post-cleaning, 53.05% of the original dataset was retained for model development.

Multiple machine learning algorithms, including Logistic Regression, K-Nearest Neighbors, Naive Bayes, Decision Tree, Random Forest, Support Vector Machine (SVM), CatBoost, and XGBoost, were evaluated using accuracy, precision, recall, and F1-score. Among these, Random Forest demonstrated the highest performance with an accuracy of 83%. To capture complex, nonlinear relationships in the data, a neural network was implemented using TensorFlow/Keras, incorporating dropout and early stopping to reduce overfitting. However, a slight drop in accuracy was observed, highlighting the need for further model optimization.

Feature importance analysis revealed that Annual Income, Outstanding Debt, Delay from Due Date, and Number of Delayed Payments were the most influential variables in predicting credit score. This study emphasizes the potential of ensemble models like Random Forest for tabular data and explores the scope for performance improvement through advanced neural architectures.

Keywords—component, formatting, style, styling, insert (key words)

1. **INTRODUCTION**

Credit scoring is a fundamental aspect of financial decision-making, affecting access to loans, credit cards, and mortgages. Traditional credit scoring models have long relied on structured financial data, such as credit history, outstanding debts, and income levels, to determine an individual’s creditworthiness. However, these models often fail to account for nuanced financial behaviors, economic fluctuations, and non-traditional credit factors. As a result, many consumers with limited credit histories or unconventional financial backgrounds face challenges in obtaining fair and accurate credit assessments. To address these limitations, the rise of machine learning (ML) and deep learning (DL) has introduced new methodologies that can capture complex relationships between diverse financial factors and enhance the accuracy, fairness, and transparency of credit scoring models. A promising development in predictive analytics is the integration of alternative data sources, including social media activity and sentiment analysis, which have been successfully applied in political forecasting and consumer behavior analysis. Inspired by the methodologies used in sentiment-based political predictions, our project explores whether similar approaches can refine credit scoring models by incorporating behavioral and sentiment-driven financial insights. By leveraging big data techniques, natural language processing (NLP), and explainable AI (XAI) frameworks like SHAP and LIME, we aim to build a more robust, interpretable, and inclusive credit scoring system that not only improves prediction accuracy but also mitigates biases associated with traditional financial evaluation method. **BACKGROUND**

One of the primary reasons for initiating this project is the unfair advantage some individuals have in obtaining higher credit scores due to traditional credit evaluation systems. Conventional credit scoring models heavily rely on historical financial data, favoring individuals with long standing credit histories while

excluding those with minimal or unconventional credit experiences. This discrepancy results in financial exclusion for many individuals who may be financially responsible but lack sufficient credit history to meet traditional criteria. The inability of existing models to adapt to dynamic financial behaviors has created an urgent need for an improved system that ensures fairness and accessibility. Our project aims to bridge this gap by integrating machine learning and deep learning techniques to create a more inclusive credit scoring model. Inspired by methodologies used in political sentiment analysis, we seek to explore the viability of alternative data sources, such as social media interactions and digital financial footprints, in assessing creditworthiness. By incorporating big data analytics and explainable AI techniques, our goal is to develop a model that not only enhances predictive accuracy but also provides a transparent and equitable approach to credit assessment. This innovative model will enable financial institutions to make more informed decisions while expanding credit access to a broader population.

1. **PROBLEM DESCRIPTION AND**

# Motivation

Credit scoring plays a critical role in financial decision-making, influencing loan approvals, interest rates, and financial accessibility. Traditional credit scoring models rely on fixed formulas, which struggle to adapt to dynamic financial behaviors and evolving economic conditions. This limitation results in inaccurate risk assessments, leading to potential financial instability.

To address these challenges, our project aims to leverage machine learning (ML) and deep learning (DL) techniques to enhance the accuracy, fairness, and interpretability of credit score prediction models. The goal is to develop a hybrid ML/DL-based model that not only predicts credit scores with higher precision but also provides transparent explanations for its decisions. This approach will improve lending decisions, reduce financial risk, and promote greater financial inclusion.

1. **RESEARCH QUESTIONS AND HYPOTHESIS**

# Research Questions:

* 1. How do ML and DL models compare in terms of accuracy for credit score prediction?
  2. What are the key financial and behavioral factors that influence credit scores the most?
  3. How can model interpretability be improved while maintaining high predictive accuracy?
  4. How effective is our model in handling individuals with sparse credit histories? **Hypothesis:**
     + H1: Machine learning and deep learning models will provide higher predictive accuracy compared to traditional credit scoring models.
     + H2: Income level, delayed payments, and outstanding debt will have the highest influence on credit scores.
     + H3: Explainable AI techniques such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) will enhance transparency without significantly compromising accuracy.

1. **LITERATURE REVIEW**

A comprehensive literature review was conducted to understand current advancements and limitations in credit score modeling.

Key findings include:

* **Traditional Credit Scoring Models**: Primarily use logistic regression and statistical approaches but lack adaptability to new financial patterns. These models often rely on linear relationships and predefined feature sets, limiting their ability to capture complex borrower behavior.
* **Machine Learning Applications**: Techniques like decision trees, support vector machines (SVMs), and ensemble models have improved accuracy but often function as black- box models. Ensemble methods such as random forests and gradient boosting machines have been particularly effective, although their complexity makes them less transparent to regulators and consumers.

# Deep Learning Techniques:

Neural networks and deep learning models, such as LSTMs and CNNs, have demonstrated significant improvements in credit score prediction, particularly for nonlinear financial behaviours. Recurrent architectures are especially useful for modelling sequential financial transactions, while convolutional models capture localized patterns in time-series data. Recent studies also explore transformer- based models for temporal credit data, showing promising results.

# Fairness and Interpretability:

Research on bias in credit scoring models emphasizes the need for fairness-aware algorithms to prevent discrimination in financial decision making. Approaches such as adversarial debiasing, fairness-constrained optimization, and interpretable surrogate models are being employed to improve transparency. Legislative and regulatory pressures are also driving the adoption of explainable AI (XAI) frameworks within the credit scoring domain.

# Alternative Data Sources:

Recent literature explores the use of alternative data—such as mobile phone usage, social media activity, and utility payments—to enhance credit assessments, particularly for underbanked populations. While these methods offer broader access to credit, they raise additional concerns about privacy, data quality, and ethical use.

# Hybrid Modelling Approaches:

There is a growing trend toward combining traditional financial indicators with advanced machine learning models in hybrid frameworks. These models aim to balance predictive performance with regulatory compliance and interpretability, addressing concerns around both accuracy and accountability.

This evolving body of research highlights the tension between innovation and ethical responsibility in credit scoring, suggesting that future progress will depend on the integration of technical, legal, and social perspectives.

# Proposed Approach

Our methodology follows a structured pipeline designed to handle the entire lifecycle of credit score prediction—from raw data acquisition to a fully functional, user-accessible model. This pipeline includes data collection, preprocessing, feature engineering, model training and evaluation, and finally, deployment. Each step is carefully designed to ensure accuracy, scalability, and usability.

# Data Collection

The foundation of our approach lies in a comprehensive dataset obtained from Kaggle, titled the *Credit Score Classification Dataset*. This dataset consists of **100,000 records** and **28**

**features**, providing a rich resource for training predictive models. Key features include **income**, **credit utilization ratio**, **monthly in- hand salary**, **number of delayed payments**, and **outstanding debt**. These attributes offer deep insights into consumer financial behavior, enabling the models to effectively differentiate between good and poor credit profiles.

# Data Preprocessing and Feature Engineering

Data preprocessing is a critical step to ensure the quality and consistency of input data. Missing values are addressed using **median imputation**, which is particularly effective in handling skewed distributions. **Categorical features** are transformed into a machine- readable format using **one-hot encoding**, while **numerical features** are scaled using a combination of **StandardScaler** and **MinMaxScaler** to bring all values to a uniform range.

An in-depth **Exploratory Data Analysis (EDA)** is performed to uncover statistical patterns, detect anomalies, and visualize relationships among variables. Techniques such as histograms, heatmaps, boxplots, and correlation matrices are employed to inform feature selection and transformation strategies, ensuring that only the most relevant variables are used in the modeling stage.

# Model Training and Evaluation

The processed data is used to train and evaluate both **machine learning** and **deep learning** models. Machine learning algorithms tested include **Decision Trees**, **Random Forests**, and **Logistic Regression**, which are known for their interpretability and baseline effectiveness. Additionally, **deep learning models** featuring multi-layer **neural networks** are implemented to capture complex, nonlinear patterns in the data.

To maximize model performance, **RandomizedSearchCV** is used for hyperparameter tuning, allowing for efficient exploration of a wide range of configurations. Evaluation is conducted using a set of standard classification metrics, including **Accuracy**, **Precision**, **Recall**, and **F1-score**, providing a

balanced assessment of each model’s predictive capability across different classes.

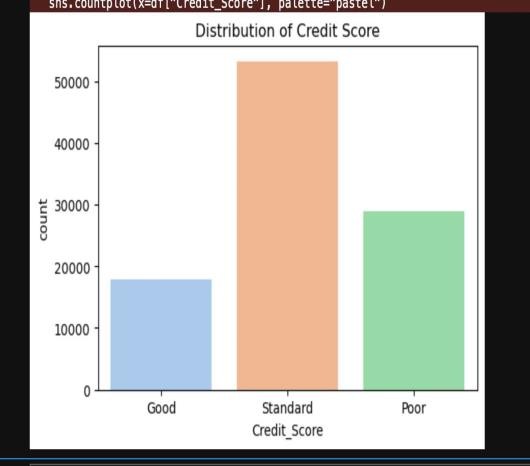
# Model Deployment

Upon finalizing the most effective models, the deployment phase focuses on operationalizing both machine learning and deep learning models. The trained models are serialized using Python’s **pickle** module and saved as .pkl files, making them ready for reuse without retraining. To provide real-time access, a **web-based application** is developed. This platform allows users to input their financial and demographic details—such as income, debt, and payment history—and receive an immediate credit score prediction. The application dynamically integrates the appropriate model (either machine learning or deep learning) based on performance needs and available system resources.

This deployment setup ensures that the entire system is fast, reliable, and user-friendly, making credit scoring accessible to both individuals and financial institutions in a scalable and transparent manner.

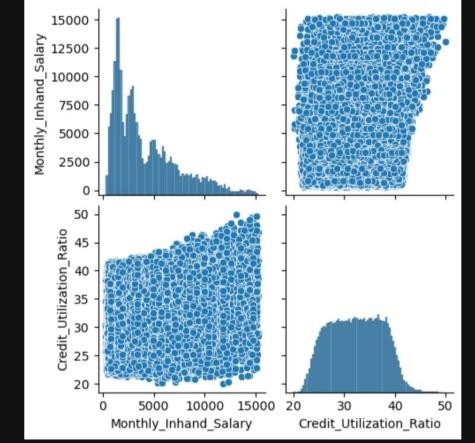
# RESULTS

* The histograms show the distribution of Monthly Inhand Salary and Credit Utilization Ratio. Salary distribution is right-skewed, with most salaries concentrated in the lower range.The scatter plots indicate a positive correlation between salary and credit utilization ratio, though data points are widely spread.



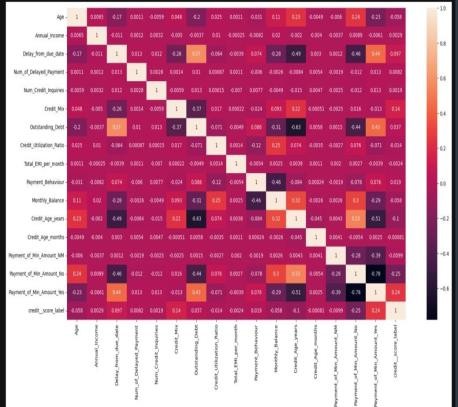
*Figure 1*

* This bar chart displays the distribution of Credit Scores categorized as Good, Standard, and Poor. The Standard category has the highest count, followed by Poor,with Good being the least frequent. This suggests that most individuals have an average credit score, with fewer having very high or low scores.



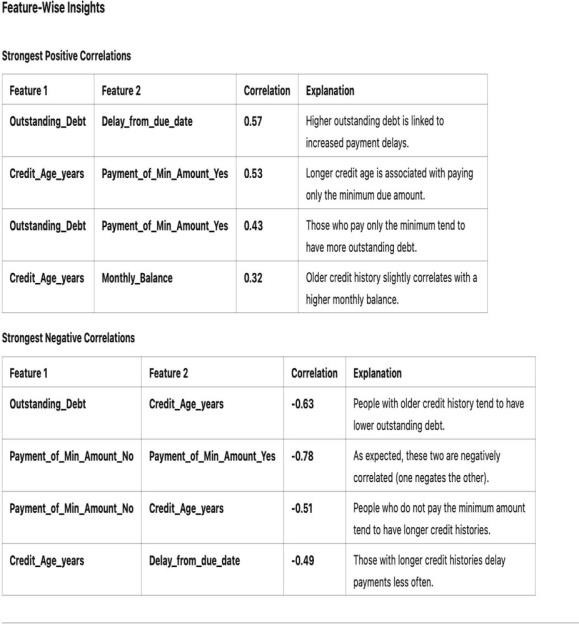
*Figure 2*

* The heatmap visualizes the correlation matrix of various financial variables. It highlights relationships between factors such as age, income, outstanding debt, credit utilization ratio, and credit score label.Strong positive correlations (lighter colors) and negative correlations (darker colors) indicate key dependencies. For example, outstanding debt and delayed payments show a moderate positive correlation, while credit age and the number of delayed payments are negatively correlated. This analysis helps in understanding which factors influence credit scores and financial behavior.



*Figure 3*

* The feature-wise insights reveal key relationships between financial behaviors and creditworthiness. Outstanding debt and delayed payments show a strong positive correlation, indicating that higher debts lead to increased delays. Conversely, longer credit histories correlate negatively with outstanding debt and payment delays, suggesting that experienced borrowers manage credit more effectively. These insights highlight the critical role of payment behavior and credit age in credit scoring models.



*Table 1*

## KNN imputation

**KNN imputation** is a technique used to handle missing data by estimating the missing values based on the values of similar data points in the dataset. It is inspired by the K-Nearest Neighbors algorithm, which identifies the closest "neighbors" to a given data point using a distance metric such as Euclidean or Manhattan distance. In the context of imputation, KNN identifies the K most similar records (i.e., those with the most comparable values across other available features) and uses their values to fill in the missing ones. For numerical data, this often involves calculating the average of the K neighbors’ values for the missing feature, while for categorical data, the most frequent value among the neighbors may be used.

One of the defining features of KNN imputation is that it is **multivariate** in nature. This means that instead of considering just a single variable in isolation (as methods like mean or median imputation do), KNN takes into account the relationships between multiple variables in the dataset. For example, if a

person's income is missing, KNN might use information like age, education level, and occupation to find similar individuals and then estimate the missing income based on their values. By leveraging these interdependencies between variables, KNN can often produce more realistic and accurate imputations.

However, despite its strengths, KNN imputation does have some limitations. It can be **computationally intensive**, especially with large datasets, because it requires distance calculations for each missing value. It's also **sensitive to the choice of K** (the number of neighbors) and the distance metric used. If the wrong K is chosen or the features are not properly scaled, the imputation results might be skewed. Additionally, KNN assumes that the nearest neighbors are truly similar, which might not always hold, especially in datasets with high variability or outliers.

Overall, KNN imputation is a powerful and flexible approach that takes into account the multivariate structure of data, helping to preserve the relationships between features while filling in missing values. It’s especially useful when missingness is scattered and when there are meaningful patterns in the data that simpler methods might overlook.

## Chi-square test

The Chi-Square Test was conducted to analyze the factors affecting credit scores by evaluating the strength of association between various categorical variables and creditworthiness. The chi-square statistic measures the deviation from expected frequencies—higher values suggest a stronger relationship. A p-value less than 0.05 indicates a statistically significant association, meaning the variable does have an impact on credit score, while a p-value greater than 0.05 suggests no significant relationship.

Key findings from the test reveal that several variables significantly impact credit scores. **Age** showed a very strong relationship with creditworthiness (χ² = 5229.79, p = 0.0), indicating that age is an important predictor. **Occupation** had a moderate influence (χ² = 181.1, p = 2.43e-24), while **Annual Income** emerged as a major factor in credit scoring with a high chi-square value (χ² = 134413.15, p = 0.0). Delays in payment due dates (χ² = 24626.21, p = 0.0) and the **number of delayed payments** (χ² = 20540.30, p = 0.0) also showed significant negative impacts, indicating that frequent delays are strongly linked to lower credit scores.

In conclusion, all tested variables were found to significantly affect credit scores. Among them, **Annual Income**, **Delayed Payments**, and **Age** stood out as the most influential predictors in determining an individual's creditworthiness.

## Machine Learning models:

|  |
| --- |
| A graph of different colors  Description automatically generated |
| *Figure 4* |

* + 1. *Logistic Regression*

Logistic Regression is a linear classification model often used for binary outcomes. It predicts the probability of a certain class by applying a sigmoid function to a weighted combination of input features. Its strengths lie in its simplicity and interpretability—each feature’s coefficient clearly shows its influence on the output. In credit scoring, this is especially valuable, as financial institutions require clear reasoning for decisions. However, this model assumes a linear relationship between input variables and the target, which may limit its performance on more complex datasets. In your evaluation, Logistic Regression achieved an accuracy of **0.64**, suggesting it might not capture complex patterns effectively but still serves as a reliable benchmark.

* + 1. *K-Nearest Neighbour (K-NN)*

K-NN is a non-parametric algorithm that classifies data points based on the majority class of their nearest neighbors in the feature space. It requires no training phase and works purely based on distance computations at prediction time. K-NN is intuitive and often performs well when similar data points are grouped closely together. However, it can be sensitive to noise, requires proper scaling of features, and becomes

computationally expensive for large datasets. For credit scoring, K-NN is useful when customer profiles show clear groupings. In the performance comparison, it recorded an accuracy of **0.73**, making it moderately effective in identifying creditworthy customers based on similarity to past patterns.

* + 1. *Naive Bayes*

Naive Bayes is a probabilistic classifier that applies Bayes’ Theorem under the assumption of feature independence. Though this assumption rarely holds in real-world data, the model performs surprisingly well in many situations and is especially suited for high- dimensional data due to its simplicity and speed. It is particularly valuable for large-scale classification problems where computational resources are limited. In the context of credit scoring, it can quickly classify customers as creditworthy or not. However, it is less flexible when feature relationships are complex. In your analysis, Naive Bayes showed a relatively low accuracy of **0.62**, which reflects its limitations when the independence assumption is violated.

* + 1. *Decision Tree*

Decision Trees are hierarchical models that split the dataset based on feature values to make decisions. They are easy to visualize and interpret, making them highly useful for applications where explainability is important— such as credit scoring. However, Decision Trees can overfit the training data if not pruned properly. They handle both numerical and categorical data well and provide clear, rule- based insights into model behavior. In your performance comparison, the Decision Tree model achieved an accuracy of **0.75**, which shows that even a single tree can yield respectable results while remaining interpretable.

* + 1. *Random Forest*

Random Forest is an ensemble model that combines multiple decision trees, each trained on different parts of the data with random feature selection. This approach significantly reduces overfitting and improves generalization. It performs well even with noisy or incomplete data and offers high accuracy and robustness. For credit scoring, Random Forest is powerful because it can capture complex feature interactions while also ranking feature

importance. According to your results, it attained the **highest accuracy of 0.83**, making it the top performer across all models, suitable for delivering highly accurate predictions with some trade-offs in interpretability.

* + 1. *Support Vector Machine (SVM)*

SVM is a robust classifier that works by finding the optimal hyperplane that best separates classes in the feature space. It maximizes the margin between classes, which improves generalization. When the data isn’t linearly separable, SVM uses kernel tricks to project it into higher dimensions for better separation. It's especially useful for high-dimensional datasets, but training can be slow and requires careful tuning of parameters. In your evaluation, SVM reached an accuracy of **0.70**, suggesting that while it generalizes reasonably well, it may not be as effective as ensemble models like Random Forest in this credit scoring context.

* + 1. *XGBoost*

XGBoost is an advanced boosting algorithm that improves performance by sequentially correcting the errors of previous models. It's highly efficient, handles missing values gracefully, and includes regularization to prevent overfitting. XGBoost is well-suited for complex datasets with non-linear relationships and is widely used in machine learning competitions due to its superior performance. In credit scoring, it effectively captures subtle feature interactions and ranks among the most accurate models. Your chart shows that XGBoost achieved an accuracy of **0.79**, placing it among the top performers with strong consistency across precision, recall, and F1 score as well.

* + 1. *CatBoost*

CatBoost is another gradient boosting algorithm designed to handle categorical features natively, eliminating the need for extensive preprocessing. It builds on the strengths of boosting while introducing techniques to reduce overfitting and manage categorical data effectively. CatBoost is fast, requires minimal tuning, and performs well even with imbalanced datasets. It is especially valuable in credit scoring where categorical features like job type, education level, or loan type play a major role. With an accuracy of **0.79**, CatBoost matches XGBoost in performance, making it a reliable

choice for both accuracy and ease of use with categorical data.

|  |
| --- |
| A bar graph with different colored bars  Description automatically generated |
| *Figure 5* |

## Deep learning

The model described is a **Feedforward Neural Network (FNN)**, specifically a multilayer dense architecture built for **multiclass classification**. It consists of three hidden layers with decreasing neuron sizes—128, 64, and 32—before reaching the output layer. The final layer uses a **Softmax activation function**, which is appropriate for multiclass tasks as it outputs a probability distribution across all potential classes. This setup allows the model to handle complex classification problems by learning hierarchical feature representations.

Each hidden layer uses **ReLU (Rectified Linear Unit)** as the activation function, introducing non-linearity that helps the model learn from intricate patterns in the data. To improve generalization and reduce the risk of overfitting, the model includes **Dropout layers**, which randomly deactivate certain neurons during training. Additionally, **Batch Normalization** is applied to standardize the inputs of each layer, which stabilizes and accelerates the training process by reducing internal covariate shift.

The model is implemented using **TensorFlow** with the **Keras API**, which offers a high-level, user-friendly interface for deep learning development. Keras makes it easier to experiment with different architectures and training strategies while leveraging the computational efficiency of TensorFlow under the hood.

Several training enhancements are integrated into the model pipeline. **Class weight balancing** is applied to address imbalanced datasets, ensuring the model gives adequate attention to minority classes during training. **Early stopping**

is used to monitor validation performance and halt training when improvements plateau, helping to avoid overfitting. Finally, **feature scaling** is performed using **StandardScaler**, which standardizes the input features by removing the mean and scaling to unit variance. This step ensures that the model converges more efficiently during training by preventing feature dominance.

# Conclusion

In conclusion, the early-stage outcomes of the ML/DL-powered credit scoring model demonstrate promising advancements in multiple key areas. The improved prediction accuracy reflects the model's ability to better assess and predict creditworthiness, potentially leading to more reliable and precise lending decisions. Furthermore, the enhancements in fairness ensure that the model addresses biases, promoting equity across diverse demographic groups, which is critical for ensuring ethical and responsible lending practices.

The focus on interpretability ensures that the model’s decision-making process is transparent, fostering trust among stakeholders, including both users and regulatory bodies. This transparency is essential for overcoming the "black-box" criticism often associated with machine learning models, enabling lenders and borrowers alike to understand how decisions are made.

As we move forward, the next steps will involve refining the model further to enhance its robustness and reliability under different scenarios. We will also prioritize the development of a user-friendly application to facilitate seamless deployment in real-world settings. This application will not only provide practical tools for lenders but also ensure that the model remains accessible and actionable for users, ensuring that it adds value to the credit scoring landscape. The ongoing work will center on ensuring the model's scalability, ease of integration with existing systems, and its ability to continually adapt to emerging data patterns, ensuring its long-term relevance and effectiveness.

1. **REFERENCES**
   1. 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.

<https://doi.org/10.1016/j.ejor.2015.05.030>

* 1. Xia, Y., Liu, C., Li, Y., & Liu, N. (2017). A

boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Systems with Applications, 78, 225–241.

<https://doi.org/10.1016/j.eswa.2017.02.017>

* 1. Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Retrieved from [https://christophm.github.io/interpretabl](https://christophm.github.io/interpretable-ml-book/) [e-ml-book/](https://christophm.github.io/interpretable-ml-book/)
  2. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30). [https://proceedings.neurips.cc/paper/20](https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf) [17/file/8a20a8621978632d76c43dfd28](https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf) [b67767-](https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf) [Paper.pdf](https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf)
  3. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “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 (pp. 1135– 1144).

<https://doi.org/10.1145/2939672.2939778>