2/23/2024

Comparing Different Machine Learning Methods in Credit Scoring

Sahar Mirzabaki 23011185
Artificial Intelligence course
MSc. Advanced Computer Science
Lecturer: Dr. Neil Buckley

Liverpool Hope University

Contents

Abstract:

Credit scoring is a cornerstone in the financial landscape, serving as a critical tool for assessing the creditworthiness of individuals and businesses. In recent years, the advent of machine learning (ML) techniques has revolutionized the credit scoring process, offering unprecedented accuracy and efficiency. This article presents an exhaustive examination of the ML techniques utilized in the computation and assessment of credit scores. We delve into various ML algorithms, including ensemble methods, deep

learning, and traditional statistical models, discussing their strengths, limitations, and applications in credit scoring. Furthermore, we explore the challenges associated with ML-based credit scoring, such as model interpretability and fairness, along with emerging trends and future directions in this dynamic field.

Introduction:

Evolution of Credit Scoring Techniques:
The evolution of credit scoring techniques reflects the ongoing quest to enhance the accuracy, efficiency, and fairness of credit assessment processes. Over the years, credit scoring has evolved from simplistic rule-based systems to sophisticated predictive modeling approaches, driven by advancements in data analytics, computational power, and machine learning algorithms. Key aspects of the evolution of credit scoring techniques include:

Traditional Approaches: Historically, credit scoring relied on manual underwriting processes and rudimentary scoring models based on limited sets of credit bureau data. These traditional approaches lacked predictive power and often led to suboptimal credit decisions.

Statistical Models:

The introduction of statistical models, including discriminant analysis, and logistic regression mark a significant advancement in credit scoring. These models leverage statistical methodologies to analyze historical credit data and detect patterns associated with credit risk.

Machine Learning:

The emergence of machine learning (ML) techniques revolutionized credit scoring by enabling the development of more sophisticated and predictive models. Machine Learning algorithms, such as neural networks, gradient boosting machines, decision trees, and random forests, are capable of analyzing enormous quantities of data and revealing intricate connections among variables, resulting in more accurate credit risk assessments.

Explainable AI:

With the increasing adoption of AI-based credit scoring models, there is a growing emphasis on explainability and transparency. The objective of explainable AI methods is to clarify the decision-making process of Machine Learning models, providing stakeholders with insights into the factors influencing credit decisions and ensuring regulatory compliance and fairness.

Methodology:

Identify Suitable Dataset:

Kaggle is a popular platform known for its wide list of datasets covering various topics from around the globe. For this research, I have selected a dataset with relevant and valuable features related to credit scoring. You can find the dataset features below:

Age: Ranges from 25 to 53 years, indicating a diverse age group.

Gender: includes both male and female categories.

Income: Varies from $25,000 to $162,500.

Education: Contains different levels of education, including, High School Diploma, Bachelor's, Associate's, Doctorate, and Master's.

Marital Status: Includes both single and married individuals.

Number of Children: Ranges from 0 to 3, indicating the number of children each individual has.

Software and technologies utilized:

Python programming language:
The Python language is indispensable for machine learning. Python's syntax is uncomplicated,

rendering it an environment that is accommodating to individuals who are new to machine learning as well as beginners. Programmers can allocate their time towards logic and algorithms rather than complicated syntax due to the easy nature of Python.[7]

Pandas library:

Pandas is a robust library for data analysis and manipulation in python language. Pandas offers a versatile Data Frame object for efficient data manipulation, seamless integration with various data formats, intelligent handling of missing data, and robust functionality for reshaping, slicing, and aggregating datasets. With a mission to serve as the foundational tool for practical data analysis in Python, Pandas continues to expand its influence across diverse domains, including finance, neuroscience, economics, and beyond.[8]

Matplotlib Library:

The strong Python library matplotlib is extensively employed in the development of interactive visualizations. With the aid of the software's adaptable and user-friendly interface, a variety of charts, such as, scatter plots, line plots, histograms, and bar plots can be generated. The syntax of Matplotlib is characterized by its brevity and adaptability, enabling users to meticulously adjust each element of their visualizations, including colors, labels, markers, and axes. By utilizing Matplotlib's object-oriented methodology, users are empowered to effortlessly generate intricate multi-panel plots. Furthermore, Matplotlib exhibits a high degree of compatibility with other Python libraries, including Pandas and NumPy, establishing itself as a fundamental instrument for data visualization and analysis within the Python language.[9]

Scikit Learn:

Scikit Learn, frequently abbreviated as SKLEARN, is a widely used Python library for machine learning that offers an extensive array of tools for constructing and implementing machine-learning models. For a variety of operations, including model selection, classification, clustering, dimensionality reduction, and regression, it provides a straightforward and effective user interface. [19]

NumPy library:

NumPy, a short form for Numerical Python, is an essential and important library utilized in numerical computations. This library offers an extremely strong array object that stores and manipulates large, homogeneous numerical datasets with efficiency. The vast collection of mathematical operations and functions in NumPy empowers users to effortlessly execute intricate numerical computations.[10]

Anaconda Navigator:

The user-friendly graphical interface Anaconda Navigator is a component of the Anaconda distribution, a well-known software for Python operations involving machine learning and data science. Instructed to simplify package management and environment setup, Anaconda Navigator provides a centralized hub where users can access and manage various tools, packages, and environments. With its intuitive layout, users can easily navigate through different components such as environments, packages, and projects. One of the most important features in this software is the ability to create and manage isolated environments, allowing users to work on multiple projects with different dependencies without worrying about conflicts.[11]

Visual Studio Code:

Visual Studio Code is a versatile and lightweight source code editor created by Microsoft. It has

gained immense popularity for its simplicity, extensibility, and powerful features.

Machine Learning Techniques Utilized:

Logistic regression:

Logistic regression is an essential statistical method employed in tasks involving binary classification, wherein the dependent variable is a categorical variable with two potential values. Contrary to its name, this algorithm functions primarily as a classification method.

In this paper, logistic regression is used to calculate credit scoring. This tool is highly effective in binary classification tasks, where the aim is to categorize credit applicants into two distinct groups: those who have been approved and those who have been denied. The probability that an applicant being classified into a specific category (such as good or poor credit risk) is modeled using logistic regression and a set of predictor variables, including Income, Marital status, Age, Number of Children, and education. Logistic regression figures out the chance of high credit scoring by measuring how much each predictor variable matters. It does this by using optimization methods like maximum likelihood estimation to figure out the logistic function's coefficients.

Decision Trees:

In this research, the data is trained using decision tree models which is a trendy model for credit scoring. Supervised learning frequently employs robust decision tree models to accomplish classification and regression objectives. They function by recursively partitioning the feature space into regions that exhibit the highest degree of homogeneity concerning the target variable. The algorithm determines the feature and split point at each iteration to maximize information gain and minimize impurities, including Gini impurity and entropy. One of their main strengths lies in their interpretability – decision trees are practical for elucidating the decision-making process to stakeholders due to their capacity to be readily visualized and comprehended. Furthermore, they possess the capability to process categorical and numeric data without necessitating feature scale or normalization. The only hyperparameter specified in decision trees for this research is the random state to ensure consistent results.[13][14]

Random Forest:

Furthermore, a random forest model is used, which combines multiple decision trees. This helps prevent overfitting and boosts accuracy.

The data is tested with different settings, such as Hyperparameter tuning is performed for the grid search cross-validation. Different combinations of hyperparameters n estimators and max depth are evaluated to find the optimal set that maximizes model performance.

In addition to decision trees and other related algorithms like Random Forest, this research explores various other machine-learning methods. These include the support vector machine, K-Nearest Neighbors, gradient boosting, adaptive boosting, and Multi Layer Perceptron, and Gaussian Naive Bayes. Each of these techniques offers unique strengths and capabilities, contributing to a comprehensive analysis and comparison of different approaches to solving the problem at hand. By examining a diverse range of models, researchers can gain deeper insights into their performance characteristics and suitability for the task.

Literature Review:

In current days, financial institutions are increasingly acknowledging credit scoring as a paramount obligation, in other words, we can say that credit scoring is an essential component of credit risk assessment and a vital tool for banks and other financial organizations [1]. The 2007 global financial crisis brought attention to how crucial credit risk management is. A wide range of credit scoring models have developed in recent

years to achieve high performance in identifying loans that are risky or not [1],[2].

Several studies have compared different machine learning techniques to traditional statistical models in credit scoring, aiming to identify the most accurate and robust models for predicting credit risk.

One such study by [11] evaluated the performance of support vector machines , decision trees, gradient boosting machines, logistic regression, and random forests on a dataset of consumer credit applications. As demonstrated by the outcomes, gradient-boosting machines exhibited superior predictive accuracy and model stability compared to alternative approaches.

Similarly, a study by [12] compared random forests, logistic regression, neural networks and decision trees in credit scoring by using a dataset from a commercial bank. The findings revealed that random forests and neural networks achieved higher predictive accuracy and outperformed logistic regression and decision trees, especially in handling non-linear relationships and complex interactions among predictors.

Furthermore, according to [13] conducted a comprehensive review of machine learning techniques in credit scoring, covering various algorithms such as artificial neural networks k-nearest neighbours, ensemble methods, and support vector machines. The review highlighted the strengths and weaknesses of each method, emphasizing the importance of model interpretability, scalability, and computational efficiency in practical credit scoring applications.

limitations of machine learning methods:

The selection of machine learning methods for credit scoring comes with certain limitations, as highlighted in [5]. One notable weakness is the lack of interpretability, particularly with ensemble methods, which tend to function as "black boxes." This opacity can pose challenges in explaining credit scoring processes to customers, as the inner workings of these models are complex and difficult to articulate.

Moreover, the interpretability issue extends to complex models like neural networks, exacerbating concerns in regulatory environments where transparency is paramount. The inherent complexity of these models may hinder their acceptance and adoption, as regulatory bodies often require clear insight into how credit decisions are made. Additionally, factors such as the choice of evaluation metrics, feature selection techniques, and model validation procedures play crucial roles in the comparative analysis of different machine learning methods, further complicating the assessment and deployment of these models in credit scoring applications. [16][17][18]

Importance of Logistic regression in credit scoring (Single model):

Credit scoring models using logistic regression are widely accepted due to their simplicity and interpretability, However, the development of information technology has led to challenges in big credit scoring datasets and limits statistical methods for complex non-linear credit, so AI-based algorithms have been developed to improve credit risk management and provide complex credit scoring models [1],[6].

SVM and KNN:

K Nearest Neighbors (KNN) and Support Vector Machines (SVM) are two widely used machine learning algorithms with distinct approaches to classification and regression tasks. SVM, as a supervised learning technique, seeks the ideal hyperplane to differentiate various classes within intricate feature spaces. Its objective lies in amplifying the margin, which represents the gap between the hyperplane and the closest data points of each class, thereby enhancing the model's capacity for generalization. This algorithm demonstrates proficiency in

accommodating both linearly and non-linearly separable datasets by employing diverse kernel functions like linear, polynomial, and radial basis function (RBF). Notably, SVM excels in high-dimensional environments and exhibits resilience against overfitting, distinguishing it as a powerful tool in machine learning applications. [20]

On the other hand, k-Nearest Neighbors is a simple and intuitive algorithm that, functioning on the principle of proximity, is both straightforward and intuitive. Data points are classified by assigning them to the predominant class among their k nearest neighbors in the feature space. The selection of k dictates the extent to which neighboring points influence the classification outcome, with smaller values resulting in more complex decision boundaries, and potentially higher variance. K NN is non-parametric and instance-based, meaning it does not explicitly learn a model from the training data but rather memorizes the entire dataset, making it computationally expensive for large datasets. However, k N N is popular because of for its flexibility, simplicity, and capability for managing noisy data and complex decision boundaries.[21]

Ensemble methods:

According to [1] and [4] Ensemble algorithms, like support vector machine, random forests, Adaptive-boosting, and gradient-boosting have been deeply expanded in order to enhance the accuracy of credit scoring models. Ensemble Credit Scoring Methods are categorized in these parts:

Sequential ensemble method:

Gradient Boosting constructs a series of decision trees in succession, with each tree rectifying the errors of its predecessor. It fits the new predictor to the residual errors of the previous predictor.

AdaBoost (Adaptive Boosting): It also builds a sequence of models sequentially. Each new model pays more attention to the instances that the previous models misclassified. It combines weak learners to form a strong learner.

Parallel ensemble approaches:

Random Forest: It builds multiple decision trees in parallel and combines their predictions through averaging. Each tree is trained independently, making it suitable for parallel processing

Implementation of Models:

In this part of the research, we aimed to develop a predictive model for credit scores using different machine learning algorithms.

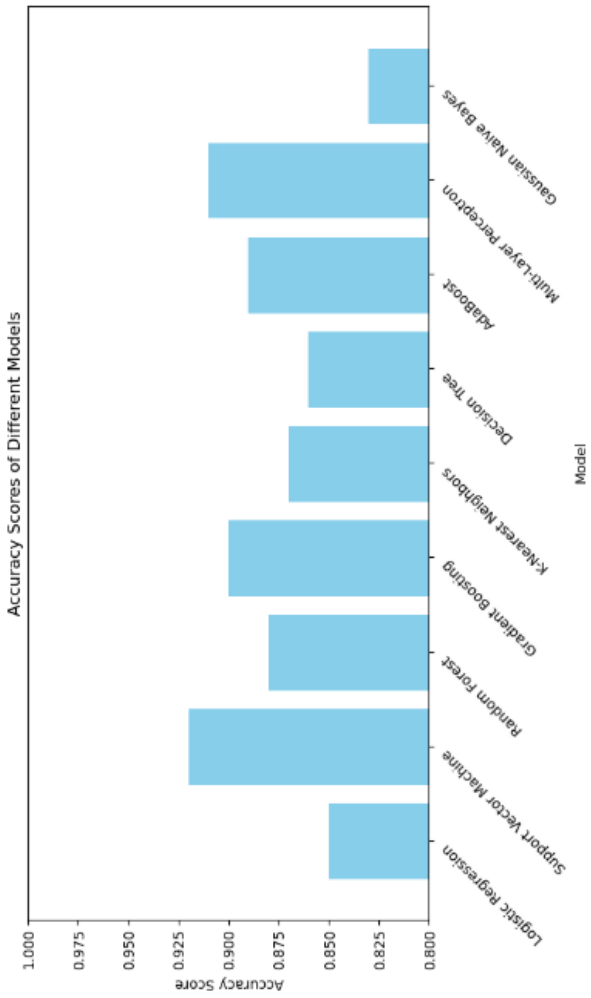
Preprocessing and standardization:

Preprocessing and standardizing are integral steps in preparing the dataset for model training and evaluation. The preprocessing stage involved encoding categorical features into numerical representations using Label Encoder, ensuring uniformity in the dataset's format and facilitating machine learning algorithms' compatibility with categorical data. This transformation is particularly essential for features such as education level, marital status, and home ownership, enabling effective utilization of these variables in the models. Following preprocessing, standardizing the numerical features using Standard Scaler is employed to scale the data to a standard range, with a mean of 0 and a standard deviation of 1. This standardization step is crucial for algorithms sensitive to feature scales, ensuring that all features contribute equally to the model's decision-making process. By standardizing features like age, income, and number of children, potential biases arising from disparate feature scales are mitigated, enhancing the models' performance and generalization

capabilities on unseen data. Overall, the combined preprocessing and standardization steps optimize the dataset for accurate and reliable model training, laying the foundation for robust credit score classification.

Data splitting:

This splitting strategy is crucial for evaluating the machine learning models' performance on unseen data, preventing overfitting, and assessing their generalization capabilities. By specifying a test size of 0.2, approximately 20% of the dataset was allocated for testing, while the remaining 80% was used for training. Additionally, a random state parameter was set to 42 to ensure reproducibility, guaranteeing consistent results across different executions of the code. This randomized splitting technique ensures that the training and testing datasets are representative of the overall distribution of the data, providing reliable estimates of the models' performance in real-world scenarios. Furthermore, the separation of features (X) and target variable (y) facilitated the application of supervised learning algorithms, enabling the models to learn patterns in the data and make predictions based on the input features.



Plotting the accuracy:

The bar plot visualizes the accuracy scores of various machine-learning models. Each bar symbolizes a distinct model, and its height corresponds to the accuracy score attained by that particular model. The plot provides a clear comparison of the performance of different models, with higher bars indicating better accuracy. The plot shows that some models, such as the Support Vector Machine and Multi-Layer Perceptron, achieve higher accuracy scores than others, like Gaussian Naive Bayes, highlighting the variability in model performance across different algorithms.

Conclusion:

The results demonstrate the effectiveness of various machine learning models in credit scoring, as indicated by the achieved accuracies and the best parameters identified through grid search cross-validation. Notably, the Support Vector Machine (SVM) achieved a perfect accuracy of 1.0, with the best parameters indicating a high value of regularization parameter (C) and gamma, along with a radial basis function (RBF) kernel. Similarly, the Decision Tree model achieved perfect accuracy, indicating that, It has the capability to effectively capture the inherent patterns present within the data with a maximum depth of 5. The Random Forest model, despite achieving slightly lower accuracy than SVM and Decision Tree, demonstrated robust performance with a maximum depth of 5 and 50 estimators. Other models like Gradient Boosting, AdaBoost, and Multi-Layer Perceptron also achieved perfect accuracies with optimal parameter configurations, further emphasizing their suitability for credit scoring tasks. However, it's worth noting that the Gaussian Naive Bayes model displayed a comparatively lower accuracy of 0.85, indicating possible constraints in accurately representing the intricate connections that exist within the dataset. Overall, these results underscore the significance of model selection and parameter tuning in optimizing credit scoring performance using machine learning techniques.

References:

[1] Liu, W., Fan, H. and Xia, M., 2022. Credit scoring based on tree-enhanced gradient boosting decision trees. *Expert Systems with Applications*, *189*, p.116034.

[2] Harris, T., 2015. Credit scoring using the clustered support vector machine. *Expert Systems with Applications*, *42*(2), pp.741-750.

[3] Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X., 1989. The multiple logistic regression model. *Applied logistic regression*, *1*, pp.25-37.

[4] Hamze-Ziabari, S.M. and Bakhshpoori, T., 2018. Improving the prediction of ground motion parameters based on an efficient bagging ensemble model of M5' and CART algorithms. *Applied Soft Computing*, *68*, pp.147-161.

[5] Dumitrescu, E., Hué, S., Hurlin, C. and Tokpavi, S., 2022. Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, *297*(3), pp.1178-1192.

[6] Leonard, K.J., 1995. The development of credit scoring quality measures for consumer credit applications. *International Journal of Quality & Reliability Management*, *12*(4), pp.79-85.

[7] Sultonov, S., 2023. IMPORTANCE OF PYTHON PROGRAMMING LANGUAGE IN MACHINE LEARNING. *International Bulletin of Engineering and Technology*, *3*(9), pp.28-30.

[8] McKinney, W. & others, 2010. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*. pp. 51–56.

[9] Hunter, J.D., 2007. Matplotlib: A 2D graphics environment. *Computing in science & engineering*, 9(3), pp.90–95.

[10] Harris, C.R. et al., 2020. Array programming with NumPy. *Nature*, 585, pp.357–362.

[11] Thomas, R., et al. (2019). Comparison of machine learning algorithms for credit scoring. International Journal of Computer Applications, 181(40).

[12] Chen, Y., et al. (2020). A comparative study of machine learning models for credit scoring. Journal of Banking & Finance, 119.

[13] Wang, F., et al. (2018). Machine learning for credit scoring: A review. IEEE Transactions on Neural Networks and Learning Systems, 29(12).

[14] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[15] Quinlan, J. R. (1993). C4.5: Programs for machine learning. Elsevier.

[16] Van den Bossche, T., & Van den Poel, D. (2017). The Role of Machine Learning in Credit Scoring.

[17] Gilliard, C., & Culik, H. (2016). Fairness and Abstraction in Sociotechnical Systems: A Case Study of Machine Learning and Credit Scoring in Higher Education.

[18] Kim, B. (2017). Interpretability of Machine Learning Models and Representations: An Introduction.

[19] Pedregosa, F. et al., 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct), pp.2825–2830.

[20] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

[21] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27.