International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 12 Issue 5 ǁ May 2024 ǁ PP. 51-65

Handing high-dimensional credit data in real personal credit loans: A combination of machine learning and feature selection

Shanjie He
*[*1] School of Mathematics and Statistics, Chongqing Technology and Business University, Chongqing, China, 400060*

Abstract: *In this study, we examine the impact of feature selection on machine learning models for credit risk prediction using high-dimensional data, which often includes irrelevant features detrimental to model performance and efficiency. Through empirical analysis of eight machine learning algorithms applied to real-world credit data, this research investigates the influence of five feature selection techniques on model effectiveness. We introduce a novel evaluation metric that balances accuracy with the relevance of features to the financial sector, termed as the Correlation of Features to the Financial Domain (CFFD). Utilizing a dataset comprising 3552 real-world personal credit records, our findings underscore the dual potential of feature selection to either enhance or impair model performance, contingent on the algorithm employed. Notably, Recursive Feature Elimination emerges as the most suitable feature selection method, while the Random Forest algorithm is identified as the most effective for predicting credit risk.*

Keywords: *Feature selection, Credit risk, Correlation, Comprehensive evaluation*

Date of Submission: 28-04-2024 Date of acceptance: 07-05-2024

I. Introduction

Credit risk is the potential for financial losses that banks and other entities face due to uncertainties in their credit operations, as highlighted by Arora (2019). A specific subset, credit default risk, arises when borrowers fail to meet their repayment obligations, necessitating thorough credit assessments by banks before loan issuance. The advent of machine learning has introduced numerous algorithms aimed at enhancing credit risk predictions, offering promising avenues for research (Trivedi, 2020). Despite these advancements, financial professionals often encounter challenges in selecting the optimal model for precise risk evaluation, underscoring the need for continued investigation into the most effective machine learning strategies for credit risk assessment.

Feature selection is crucial in enhancing machine learning models by minimizing credit data complexity and removing non-essential features, thereby boosting model accuracy and dataset quality (Trivedi, 2020; Alaka et al., 2017; Li et al., 2016). It focuses on retaining only the most pertinent features, simplifying model training and revealing underlying data patterns. However, the application of feature selection techniques must be cautious, as inappropriate methods can negatively impact algorithm performance (Olu-Ajayi et al., 2023).

Espinosa et al. (2023) categorize feature selection methodologies into three distinct paradigms: filter, wrapper, and embedded methods. Filter methods prioritize features or subsets thereof based solely on intrinsic dataset characteristics, eschewing dependence on any classifier, as elucidated by Roffo et al. (2020). Conversely, wrapper methods assess features or their combinations through the lens of specific estimators, enhancing the efficacy of particular classifiers, a technique detailed by Jiang et al. (2019). Embedded methods, as Wang and Zhu (2018) describe, integrate feature selection directly into the learner's training phase, rendering it inseparable from the algorithm itself. This tripartite classification underscores the nuanced approaches to feature selection, each with implications for algorithmic efficiency and application specificity. Notably, Trivedi (2020) illustrates the application of these methods within the realm of credit risk prediction model development, highlighting their practical relevance and potential to contribute to more accurate and robust predictive analytics.

Despite the diversity of feature selection methods available for credit risk prediction, there exists a notable deficiency in comparative analyses and evaluations of these methods' efficacy. A prevailing limitation in extant literature is the reliance solely on accuracy metrics for assessing feature selection techniques, which significantly undermines the interpretability of results. Moreover, prior research in credit risk assessment frequently neglects to amalgamate feature selection methodologies with model learning processes, thus missing out on a holistic evaluation framework. This oversight becomes critical when considering that different feature selection methods can exert varying degrees of influence—both positive and negative—on a single classification

model's performance. Such a scenario underscores the imperative for rigorous empirical investigation. Consequently, this study is poised to fill these lacunae by identifying the most appropriate feature selection technique and the most efficacious machine learning model for credit risk prediction. It endeavors to conduct a thorough evaluation of the feature selection process, thereby contributing to a more nuanced understanding and application of these methodologies in the context of financial risk assessment.

This research undertook an extensive evaluation of eight prominent machine learning classification algorithms for the development of credit prediction models. These algorithms include Logistic Regression, Decision Tree, K Nearest Neighbors, Support Vector Machine, Random Forest, Extreme Gradient Boosting, Naive Bayes, and Deep Neural Networks. The empirical analysis is grounded on a dataset comprising real-world personal credit data from a banking institution, encompassing a total of 3552 processed samples. To enhance the predictive accuracy and relevance of the models to the financial domain, the study employed a variety of feature selection techniques. These encompassed filter-based methods such as Variance Threshold and Mutual Information, wrapper-based methods like Recursive Feature Elimination, and embedded methods including LASSO and Random Forest. Furthermore, this study introduces a comprehensive evaluation methodology for feature selection aimed at identifying features that not only enhance model accuracy but are also significantly relevant to the financial sector. The objective is to conduct an impartial comparison of different feature selection methods and machine learning algorithms to ascertain the most efficacious combination for credit risk prediction.

This study's contributions to the field of credit risk prediction are threefold: (i) It identifies the optimal feature selection technique and the most efficacious predictive model for credit risk assessment. (ii) It introduces a robust evaluation framework for analyzing feature selection methodologies. (iii) It offers guidance and recommendations for subsequent research endeavors focused on feature selection in credit risk prediction.

The subsequent structure of this article is as follows: Section 2 reviews existing research on credit risk and feature selection. Section 3 describes the background of the research question. Section 4 describes the data preprocessing process, feature selection methods, and model development and evaluation measures. Section 5 discusses performance results and findings and presents the theoretical and practical significance of this article. Section 6 provides the conclusion of the article and points out its shortcomings, providing future suggestions.

II. Literature review

2.1 Classification in Credit Risk Prediction

Machine learning classification models have emerged as pivotal tools in elucidating the nexus between requisite loan attributes (e.g., historical records, customer account numbers, revenue) and the likelihood of defaults, garnering widespread application in the domain of credit risk prediction (Zhang and Yu, 2023). Notably, a spectrum of studies has leveraged these models for credit risk assessment (Liu et al., 2020; García-Céspedes and Moreno, 2022; Shi et al., 2011; Baser et al., 2023; Aksakalli and Malekipirbazari, 2015). Liu et al. (2020) introduced a two-stage hybrid model aimed at augmenting the predictive accuracy of credit risk evaluations. García-Céspedes and Moreno (2022) investigated the efficacy of Machine Learning (ML) technologies in emulating and refining the model outputs derived from Vasicek's (1987) credit risk framework. Through an empirical study on bad debt recognition, Shi et al. (2011) underscored the proficiency of the Random Forest algorithm as a robust credit evaluation classifier. Baser et al. (2023) devised a Clustering-Based Fuzzy Classification (CBFC) methodology for credit risk appraisal, achieving commendable classification outcomes. Furthermore, Aksakalli and Malekipirbazari (2015) conducted a comparative analysis of Random Forests, Support Vector Machines, Logistic Regression, and KNN across metrics such as AUC and accuracy, deducing that Random Forests excel in identifying superior borrowers.

2.2 Feature selection based on credit risk

Feature selection is recognized as a pivotal data preprocessing strategy in the realm of machine learning, particularly for managing high-dimensional data (Li et al., 2016). It serves to streamline dimensions and mitigate the risk of overfitting by discerning the correlation between each feature and the output label. Moreover, feature selection facilitates the elimination of superfluous features within the dataset, thereby enhancing computational efficiency (Yu et al., 2020). This technique assumes a vital function in the prediction of credit risk, with numerous studies incorporating feature selection methodologies for this purpose (Yao et al., 2022; Rtayli and Enneya, 2021; Nali et al., 2020; Cui et al., 2021; Lappas and Yannacopoulos, 2021). Yao et al. (2022) introduced a Sequential Backward Feature Selection algorithm based on Ranking Information (SBFS-RI) alongside an Integrated Feature Selection method based on Multiple Sorting Information (FS-MRI), demonstrating that FS-MRI surpasses nine other feature selection techniques in yielding a more efficacious and robust feature subset. Rtayli and Enneya (2021) developed an Enhanced Credit Card Risk Identification (CCRI) methodology, utilizing a Random Forest classifier and Support Vector Machine feature selection algorithm for fraud risk detection, and concluded its superior classification performance over local anomaly factors, isolation

forests, and decision trees in extensive datasets. Nali et al. (2020) unveiled a novel hybrid data mining model, amalgamating various feature selection and ensemble learning classification algorithms with an innovative voting method, dubbed if_any, which outshone all competing voting techniques and individual feature selection algorithms in supporting decision-making processes. The hybrid model, integrating the if_any voting mechanism with a GLM+DT model, was found to excel beyond all other combined and singular classifier models in performance. Cui et al. (2021) devised a Multi-Structure Interaction Elastic Network (MSIEN) model for feature selection, employing a regularization model that merges an interaction matrix with an elastic network for feature subset selection. Lappas and Yannacopoulos (2021) proposed a strategy that synergizes soft computing approaches with expert knowledge, validating its effectiveness through test cases on a standard credit dataset.

In the context of this research, where credit data encompasses borrower attributes and credit records, the employment of machine learning algorithms for feature selection is posited to facilitate the identification of the most pertinent features. Presently, feature selection can be categorized into three principal methodologies: filter, wrapper, and embedded methods (Maldonado and Weber, 2009). This study employs a methodology comprising two filter methods, one wrapper method, and two embedded methods. The efficacy and applicability of these five distinct feature selection strategies are compared and analyzed across various machine learning classification algorithms.

III. Problem Description

In contemporary society, the burgeoning credit business, while marking significant progress, concurrently unveils challenges, notably elevated credit risk. The domain of financial lending is typified by information asymmetry, historically characterized by borrowers possessing a comprehensive insight into their financial status, repayment capacity, and intent, in stark contrast to financial institutions' limited or delayed grasp of the borrowers' risk levels. This disparity in information can precipitate financial losses for financial institutions during the lending process, owing to discrepancies between anticipated risks and actual outcomes, thereby substantially influencing their profitability.

Nevertheless, the advent of technological advancements has ushered in methodologies that amalgamate machine learning with credit risk prediction. This fusion facilitates the aggregation of diverse data sources, enabling the preemptive prediction and evaluation of customer risk levels, and informing credit decisions predicated on these assessments. Credit risk prediction inherently constitutes a binary classification challenge, with machine learning techniques for addressing such problems demonstrating efficacy.

This investigation will harness various machine learning algorithms, in conjunction with feature selection techniques, to forecast credit risk. Furthermore, it will undertake a comparative analysis of disparate algorithms to ascertain the most efficacious model for credit risk prediction.

IV. Data and methodology

This paper analyzes the influence of several feature selection methods on ML algorithm in credit risk prediction. Our research will further explore the ML classification algorithm for the most suitable collocation of feature selection methods. Therefore, we present the question: can all feature selection methods have a positive effect on ML model performance? The data set used in this study to develop all ML models is real-world credit data. In this section, we describe and merge the data in part 3.1, and then Data pre-procession it in part 3.2. Feature selection methods are described in detail in section 3.3, and features are analyzed for correlation in section 3.4. The adopted ML classification algorithm is described in section 3.5 for model training and testing, and finally, in section 3.6, we present the comprehensive evaluation index and evaluate our trained model. The overall flow chart of this study is shown in Fig. 1.

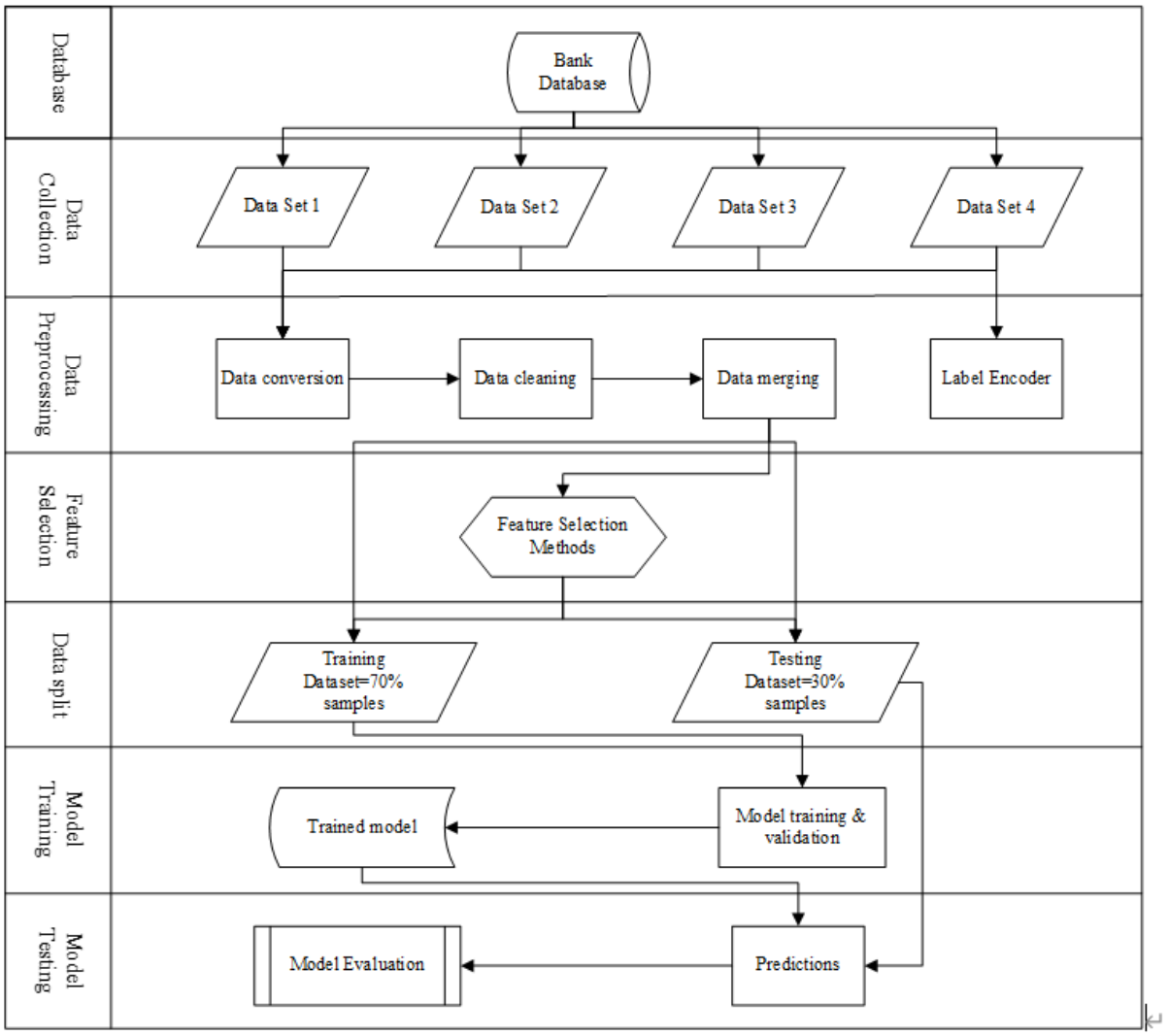
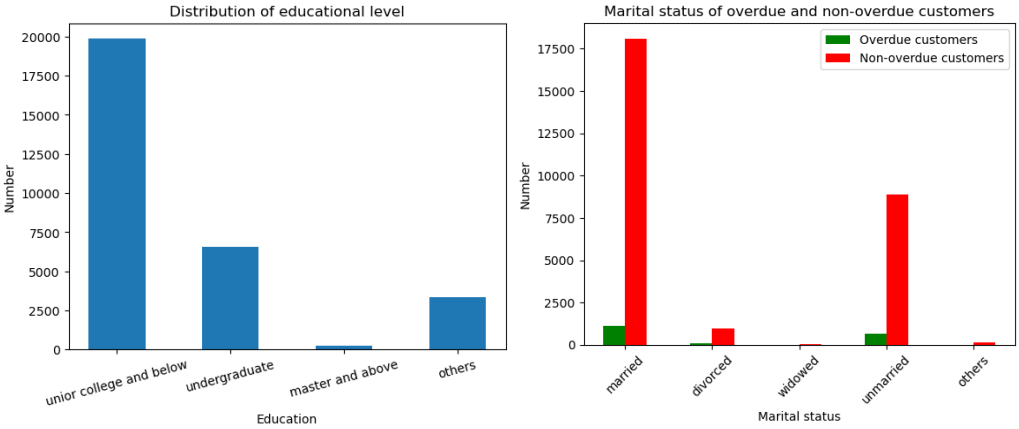


Fig. 1. The framework of the study.

4.1 Data description

The data set used in this study is collected from a real lending platform in Shanghai in China. The available training data include user basic information data set (Data Set 1), credit cue data set (Data Set 2), the unsold credit card or outstanding loan data set (Data Set 3) and the overdue information summary data set (Data Set 4). The number of samples in these data sets is different, so only the samples that are common to the four data sets are selected as the study objects. Some features of these data sets are considered to be related to credit prediction and are often used to build credit prediction models (Yu et al., 2022). What we should do is to select the most relevant features for credit risk prediction, improve the performance of classification model and reduce training costs. These data sets collect information about personal loans and credit cards as well as past records.

Data Set 1: Dataset 1 consists mainly of basic information about the user. These include report ids (for associating other data sets), marital status, income level, education level and so on. These features are often used to predict credit risk (Machado and Karray, 2022). We collected basic information about 30000 users. Fig. 2a shows the distribution of user education levels in the dataset. Education level is divided into junior college and below, undergraduate, master and above, others. In addition, the label variable of this prediction, namely, the repayment status of the user's latest payment, is also included in Data Set 1, which is a discrete binary variable for overdue and non-overdue customers respectively. Fig. 2b shows the repayment status of people with different marital status. Marital status is classified into five categories: married, divorced, widowed, unmarried, and others.



2a **2b**

Fig. 2a. User education level distribution.
Fig. 2b. Marital status of overdue and non-overdue customers.

Data Set 2: The credit tip data set mainly collects the number of loans or the number of credit or quasi-credit card accounts of the lender. Features such as the number of loans and the number of credit card accounts have been used to build a credit risk prediction model (Kruppa et al., 2013). The dataset has a total of 39970 samples with 11-dimensional features, including report ids.

Data Set 3: The data set of unsold credit card or outstanding loans includes the features of loan type, number of loan entities and number of loan institutions, which are considered as one of the key variables in credit risk prediction (Oreski and Oreski, 2013). In order to facilitate follow-up processing, we split the data set into three sub-data sets by type of loan (there are three categories, namely the summary of outstanding loan information, the summary of outstanding credit card information and the summary of quasi-outstanding credit card information). Each of these contains report ids that is used to associate other data sets.

Data Set 4: The data set of overdue information has 6-dimension features including report ids, which involves the number of overdue loans, the number of overdue months, etc.. Similarly, it breaks down into three data sets by type of loan (loan overdue, credit card overdraft of more than 60 days). Fig. 3 shows all the features from the four Data Sets.

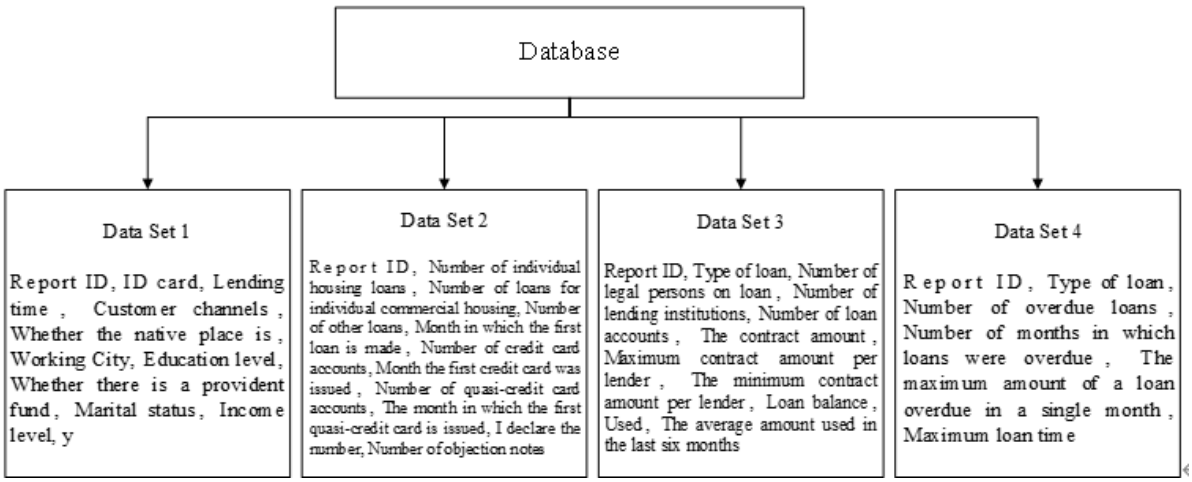


Fig. 3. The feature name is displayed.

In order to avoid the bias caused by unbalanced label classification, this paper uses a variety of evaluation indicators, such as accuracy, F1-score, AUC (Area Under Curve) and the comprehensive evaluation indicators we propose below. Finally, we combined the four data sets according to the report id to get a total data set.

4.2 Data pre-processing

Data pre-processing can be used to improve the performance of ML models (Mishra et al., 2020). Since data mining requires high quality data that must be accurate, complete and consistent, Data pre-processing is a necessary step in data mining (Tegunov and Cramer). The Data pre-processing includes feature scaling, missing

value processing, outlier processing, data conversion and so on. Some factors, such as missing values, noise, and non-normalization, can have a negative impact on modeling if inadequately preprocessed data is used directly (Esteban et al., 2020). Therefore, this paper makes the following pre-processing to the credit data.

Data Merging: We used the pandas library merge function in python 3.10 for data consolidation. The merge function takes on as Report ID and how as inner, and defaults to other parameters. After collation, we got a total of 16288 samples in 41 dimensions, Where Report ID and ID card do not participate in modeling.

Missing value handing: A total of seven features in the dataset contained missing values, including customer channels, income levels and the first quasi-credit card issue month, which was missing by more than 60% and therefore deleted directly. There are only a few missing data for the city of work, provident fund, first credit card month and first loan month. We use the mode to fill in the missing data.

Label encoder: Use the tag encoder in python's scikit-learn library to convert data into a machine-readable digital format. For example, convert the variable list for label y to 0 and 1, respectively, where 0 represents a non-overdue customer and 1 represents an overdue customer.

Normalization: In this paper, all the data used for modeling are normalized.

Unbalanced data processing: Fig. 4 shows the distribution of label categories in the combined data, showing that the number of overdue and non-overdue customers is extremely unbalanced, with a ratio of 888:15400. In order to reduce the loss of useful information and avoid the over-fitting problem caused by the large number of repeated samples, we adopt the method of under-sampling to process the data. Specifically, we randomly delete the samples labeled as non-overdue and ratio of non-overdue and overdue to 3:1, a total of 3552 samples were used for feature selection and modeling.

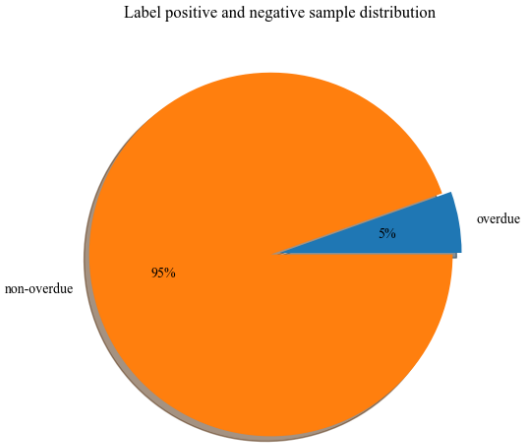


Fig. 4. Label positive and negative sample distribution.

4.3 Feature selection methods

In this study, 5 feature selection methods were used to explore the most relevant features for credit risk prediction. These methods include filtered Variance Threshold method and Mutual Information method, wrapped Recursive Feature Elimination method, and Embedded Lasso method and Embedded Random Forest method.

4.3.1 Filters

Variance Threshold: This is a univariate filter-based technique that uses variance to search the separability of each feature between classes (Yeoh et al., 2017). The feature is filtered by calculating the variance of the feature itself, and when the variance of a feature is small, indicating that the sample is essentially no different on that feature, then the feature has no effect on sample discrimination.

Mutual Information: This is a measure of the interdependence between variables, is Battiti (1994) proposed. The significance of mutual information is to amplify the correlation between features and tags, and to reduce the redundancy of selected features. When the two features are exactly the same, mutual information is the largest, so the greater the mutual information, the stronger the feature correlation.

4.3.2 Wrapper

Recursive Feature Elimination: This is a multi-variable packaging technology, it works by giving an external estimator that can be weighted with features (for example, the correlation coefficient of a linear model), and REF recursively selects features by considering smaller and smaller sets of features. After that, removes the least important feature from the current feature set (Hidalgo-Munoz et al., 2013). This step is repeated repeatedly over the set of features until the desired number of features is reached.

4.3.3 Embedded

Embedded LASSO: Lasso regression restricts model coefficients by adding L1 regularization, and the effect of L1 regularization is to restrict model coefficients to 0 as much as possible, that is, feature selection can be made by this point (Cui et al., 2021).
Embedded Random Forest: This method embeds a random forest algorithm for feature selection. The importance of each feature is measured by the Gini index, which measures the contribution of each decision tree in a random forest, and then averages it.
We used these five feature selection methods to select the top 15-dimensional features, as shown in Table 1. Lending time, Working City, Credit card contract amount, Used, and so on are some of the higher-ranking features. These are all important lending features, which have great influence on credit risk prediction.

Table 1
Rank of each feature selected using various selection methods.

Features	Variance Threshold	Mutual Information	Recursive Feature Elimination	Embedded LASSO	Embedded Random Forest
Lending time					
Whether the native place is					
Working City					
Education level					
Marital status					
Whether there is a provident fund	√	√	√	√	√
Number of individual housing loans			√		
Number of individual commercial housing loans	√	√	√	√	√
Number of other loans		√			
Month in which the first loan is made					
Number of credit card accounts					
Month the first credit card was issued					
Number of quasi-credit card accounts	√		√	√	√
Number of legal persons on loan			√		√
Number of lending institutions		√	√		√
Number of loan accounts		√	√		√
Loan contract amount					
Loan balance					
The average amount of loans used (6 months)			√		
Number of credit card holders	√	√		√	√
Number of credit card institution	√			√	√
Credit card contract amount	√	√	√	√	√
Maximum contract amount per lender		√	√		
The minimum contract amount per lender		√	√		
Used	√	√		√	√
Average credit card usage (6 months)	√			√	√
Number of overdue loans	√			√	√
Number of months in which loans were overdue	√	√		√	√
Maximum amount of overdue loan (one month)	√	√		√	√
Maximum loan time			√		
Number of credit card overdue accounts	√		√		
Credit card overdue (number of months)	√	√		√	√
Maximum amount of credit card overdue		√			
Credit card maximum loan length	√		√	√	
Number of quasi-credit card overdue accounts					
Quasi-credit card overdue (number of months)					
Maximum amount of quasi-credit card overdue					
Quasi-credit card maximum loan length					

4.4 Correlation

After feature selection, we regard the features selected by the five feature selection methods as the features of relatively high importance, and we also analyze the correlation between these features and target, an evaluation of this correlation is shown in Fig. 5 below. The correlation thermogram shows the correlation

between the feature and the label. As shown in Fig. 5, a higher correlation value indicates a stronger correlation between the two features. Therefore, Used and Credit card contract amount have a strong correlation. There is a strong correlation between the Credit card contract amount and Average credit card usage (6 months). In addition, there is a strong correlation between number of credit card accounts and number of credit card institution.

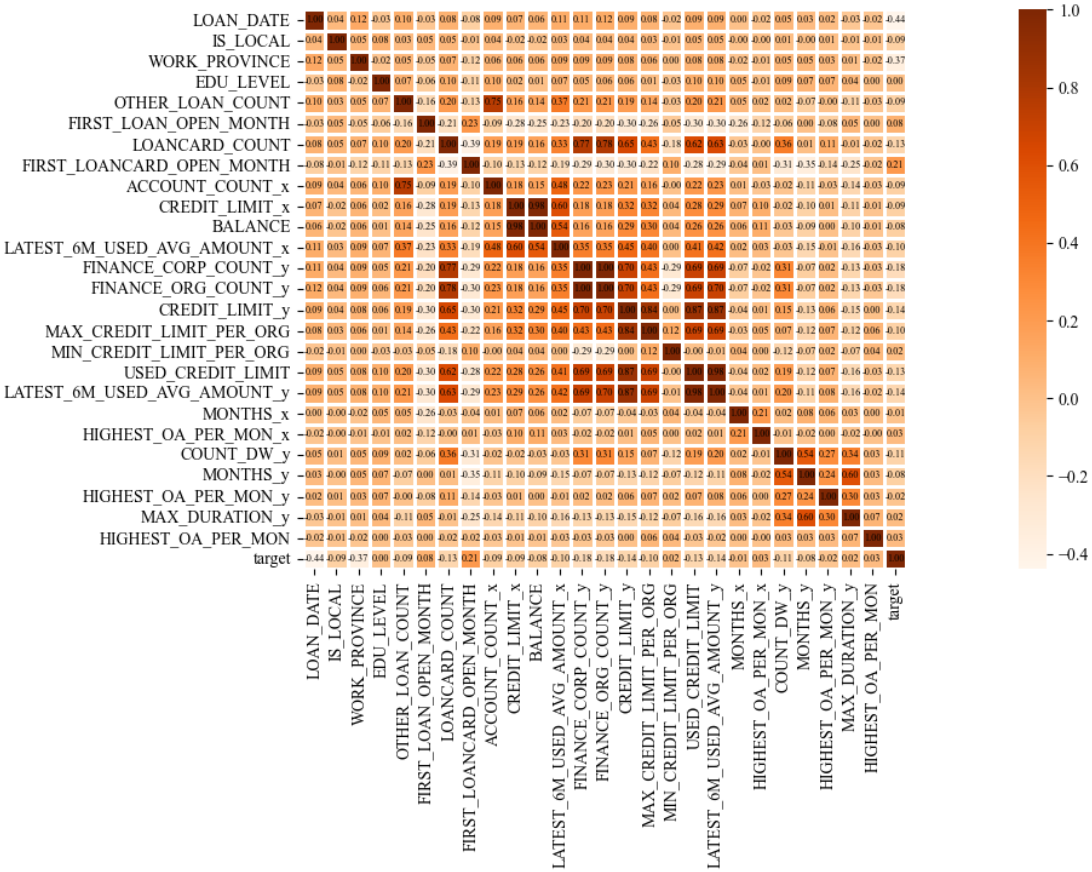


Fig. 5. Correlation between variables.

4.5 Model development

Credit risk prediction is a kind of supervised binary classification task. The overdue and non-overdue customers identified by banks were used as the class labels in this study. This study used python 3 programming language for data pre-processing, feature selection, model training, and test evaluation. We have developed a total of eight credit forecasting models, as follows:

Logistic Regression (LR)**:** This is a broad linear classifier for binary classification problems and belongs to supervised learning in machine learning (Li et al., 2023). The LR model was developed with 100 iterations and L1, L2 regularization was added.

Decision Tree (DT)**:** Tree model is one of the most important models in machine learning, and it is also the most commonly used basic classifier in ensemble learning. Decision tree is a kind of tree model, which uses tree flow chart to divide data into several groups. This is a general method and can be improved as the sample size increases (He et al., 2021). The principle of decision tree is simple and easy to understand, with high computational efficiency and excellent interpretability, it can provide a clear and intuitive decision path. The model parameters used for development DT are the 'Gini' and the 'best' splitter parameter.

K Nearest Neighbors (KNN)**:** KNN is one of the simplest algorithms in classification technology, which uses similarity or distance function to estimate the results based on the k closest training samples in feature space (Vommi and Battula, 2023). The parameters used to develop the KNN model were 5 neighbors, 30 blades, and size-uniform weights.

Support Vector Machines (SVM)**:** This is a kind of generalized linear classifier for binary classification of data based on supervised learning. Its decision boundary is the maximum margin hyperplane for learning samples. In addition, SVM is considered to be one of the most accurate data mining algorithms (Aram et al., 2023). The parameters used to develop the SVM model are Gauss kernels and the penalty coefficient

'C'=1.0.

Random Forest (RF)**:** RF is an algorithm that integrates multiple trees by the idea of ensemble learning. It can realize the learning of simple and complex problems (Sun et al., 2020). The RF model consists of 25 estimators.

Extreme Gradient Boosting (EGB)**:** This is the machine learning model developed by (Chen et al., 2016), which is widely used in various ML competitions, and has achieved good results. EGB can be used to classification and regression problems. EGB model parameters: (Learning rate=0.3, Loss function=Logistic, estimators=100).

Naïve Bayes (NB)**:** This is a classical ML classification algorithm based on probability theory. The NB principle is simple and easy to implement. It can be used for both binary classification task and multivariate classification problem. The default parameters were used in the development of the NB model.

Deep Neural Networks (DNN)**:** This is a machine learning technique that is a fully connected neural network with multiple hidden layers. The developed DNN model consists of 4 layers (1 input layer, 2 hidden layers and 1output layer).

The data used in this study to develop eight models is from real world bank personal lending data. After processing, there are a total of 3552 samples.

4.6 Model evaluation

This paper presents a comprehensive index for evaluating feature selection method. Most of the existing researches take accuracy, AUC and F1 score as the evaluation indexes of feature selection methods and model performance. However, if the feature selection method is considered to have good performance just because of high accuracy, it is not convincing enough and lacks certain interpretability. A model with high accuracy is not necessarily reliable (Ribeiro et al., 2016). To address this, we defined the correlation between features and the financial domain (CFFD), and weighed CFFD against accuracy. The aim is to select the feature set which can improve the performance of the model and is closely related to the financial field, which increases the post interpretability of the model from the feature perspective (Murdoch et al., 2019).

In this section, we first define what a CFFD is, and then describe in detail our proposed method for weighing accuracy against CFFD.

4.6.1 Defining CFFD

First, we interview the banking experts and select the feature subset 1 which is closely related to the financial field. Table 2 shows the feature set selected based on expert experience. Then, we use ML technology for feature selection and get feature subset 2. We define CFFD as the deviation between feature subset 1 and feature subset 2. The smaller the deviation, the larger the CFFD. Fig. 6 shows the definition process.

Table 2
Feature subset based on expert experience.

Working City	Marital status	Number of other loans	Number of credit card accounts	Number of loan accounts
Loan contract amount	Loan balance	Credit card contract amount	Number of overdue loans	Number of months in which loans were overdue
Maximum amount of overdue loan (one month)	Number of credit card overdue accounts	Credit card overdue (number of months)	Number of quasi-credit card overdue accounts	Maximum amount of quasi-credit card overdue

Fig. 6. A flowchart of the definition.

4.6.2 The introduction of the evaluation index of classification

The evaluation indexes of the classification are all based on the confusion matrix, which is shown in Table 3.

Table 3
Confusion matrix.

Confusion Matrix		True Value	
		Positive	Negative
Predicted Value	Positive	TP	FP
	Negative	FN	TN

Accuracy: This is one of the most commonly used evaluation indicators for classification tasks. It is a calculation of the exact match between the estimated value and the actual value. The formula for accuracy is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

F1 score: This is the result of taking a weighted average of the precision and recall over a range of (0, 1), with the larger the score the better. F1 score formula is:

$$\Pr ecision = \frac{TP}{(TP + FP)} \parallel \operatorname{Re} call = \frac{TP}{(TP + FN)}$$

$$F1 = 2 * \frac{(\Pr ecision * \operatorname{Re} call)}{(\Pr ecision + \operatorname{Re} call)}$$

ROC AUC: This is a model classification index, is the receiver operating feature curve under the area. It is an important index to measure the performance of unbalanced data classifiers and is widely used. When the ROC AUC value is 1, it represents the best but lowest ROC AUC score of 0.5 (Carrington et al., 2022).

4.6.3 CFFD quantification and trade off with accuracy

Quantification: Because the CFFD describes the deviation between two sets, we use the Jaccard similarity coefficient (also known as the Jaccard Index) to quantify it (Eelbode et al., 2020), write it down as $J(F, G(f)) \in [0,1]$. Namely:

$$J(F, G(f)) = \frac{| F \cap G(f) |}{| F \cup G(f) |}$$

Among them, F is the feature subset selected based on expert experience. G stands for a class of potential feature selection methods, f is a method in G, and $G(f)$ is a feature subset selected by f method.

When the sets F and $G(f)$ are empty sets, $J(F, G(f))$ is defined as 1, and the smaller the deviation between the two sets, the greater the $J(F, G(f))$.

Trade off: The J-A score is defined as a weighted average of Jaccard index and Accuracy to evaluate the performance of feature selection methods and models. It is a comprehensive evaluation index. The formula is as follows:

$$J - A = \frac{J(F, G(f)) + Acc(f)}{2}$$

The J-A score has a range of [0, 1], and the greater the value, the better.

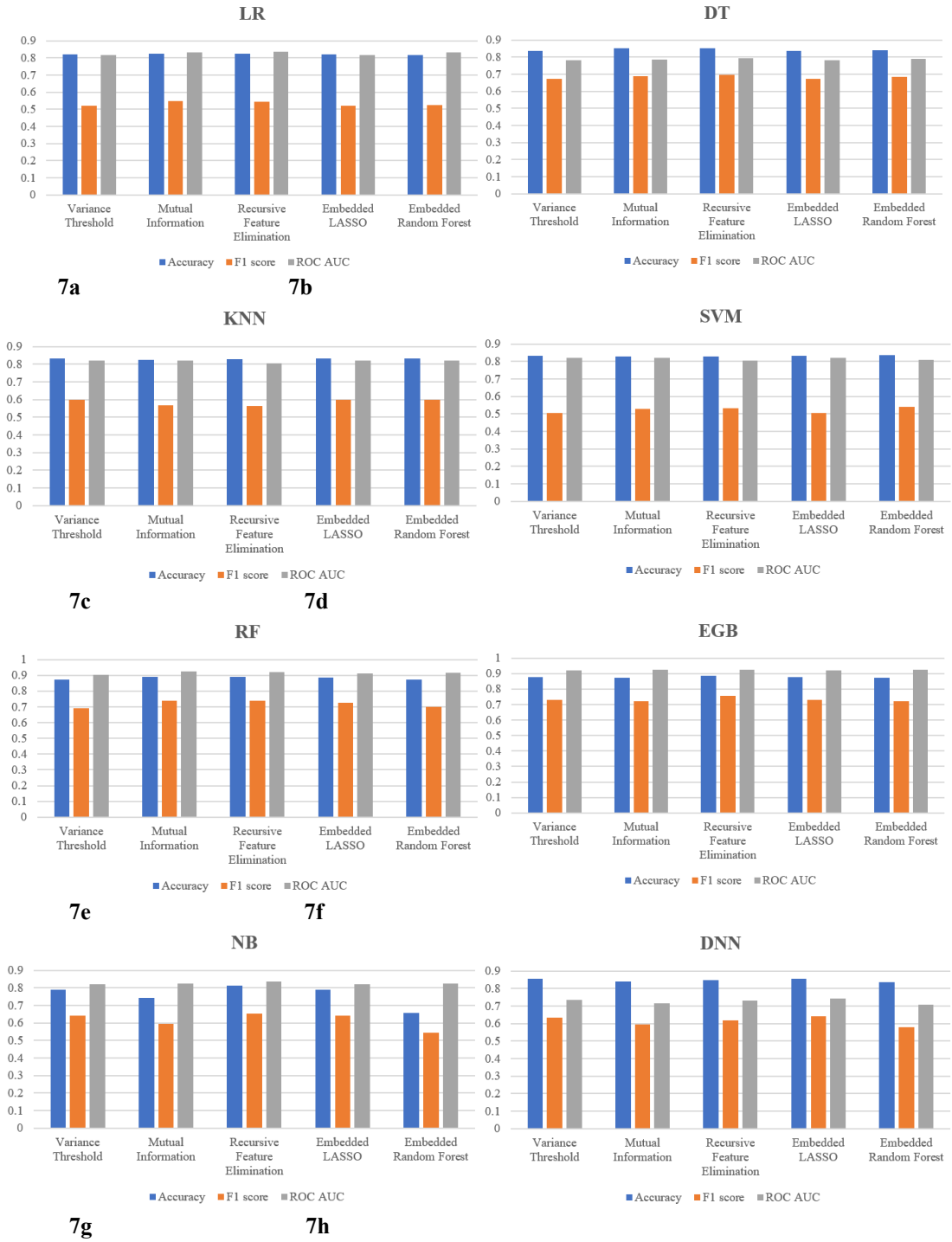
V. Result and discussion

This research delves into the impact of various feature selection methodologies on the performance of predictive models, utilizing real credit data. Five distinct feature selection methods were employed to identify the fifteen most pertinent features, which were then used to construct eight machine learning classification models. Subsequent analysis aimed to pinpoint the most efficacious feature selection techniques and to highlight the features of paramount importance for predicting credit risk. The models' predictive performance, depicted in Figures 7a to 7h, varies according to the feature selection methods applied.

Regarding accuracy, the Random Forest (RF) model demonstrates superior outcomes when employing Mutual Information and Recursive Feature Elimination as feature selection techniques. Each model attains its peak accuracy with distinct feature selection methods. For instance, Logistic Regression (LR) and Decision Tree (DT) models achieve commendable accuracies of 0.826 and 0.852, respectively, through the application of Mutual Information and Recursive Feature Elimination. Similarly, both RF and Extreme Gradient Boosting (EGB) models record their highest accuracy using the same feature selection methods. Additionally, the K-Nearest Neighbors (KNN) model exhibits a notable prediction accuracy of 0.834, leveraging Variance Threshold, LASSO, and Random Forest methods. The Support Vector Machine (SVM) model achieves an accuracy of 0.835 with Random Forest, while the Naive Bayes (NB) model reaches an accuracy of 0.812 using

Recursive Feature Elimination. The Deep Neural Networks (DNN) model attains its highest accuracy with LASSO.

This variability in model performance across different feature selection methods underscores the necessity of tailoring the feature selection approach to the specific model employed for credit risk prediction.



Figs. 7a~7h. Models performance results using various FS methods.

To address the research question concerning the potential positive impact of feature selection methods on predictive models, and to effectively employ the proposed comprehensive evaluation index, it is imperative to incorporate two sets of controlled experiments. These experiments entail modeling using a feature subset derived from expert experience and utilizing all features, respectively. Figure 8 is designed to provide a lucid comparison of the influence—whether positive or negative—of five feature selection methods on the predictive accuracy of eight machine learning (ML) models. Additionally, Table 4 compiles the prediction accuracy of

each model, delineating results obtained with and without the application of feature selection. This structured approach facilitates a nuanced understanding of the efficacy of feature selection methods in enhancing model performance, thereby enabling a data-driven determination of their value in the context of credit risk prediction.

Table 4
Performance result for each model with and without feature selection methods.

	Variance Threshold	Mutual Information	Recursive Feature Elimination	Embedded LASSO	Embedded Random Forest	Expert Experience	Without FS
LR	0.820	0.826	0.826	0.820	0.818	0.767	0.836
DT	0.837	0.852	0.852	0.837	0.841	0.743	0.848
KNN	0.834	0.826	0.829	0.834	0.834	0.750	0.807
SVM	0.832	0.828	0.829	0.832	0.835	0.751	0.831
RF	0.873	0.889	0.889	0.887	0.875	0.774	0.879
EGB	0.878	0.874	0.887	0.878	0.872	0.780	0.879
NB	0.790	0.743	0.812	0.790	0.659	0.583	0.607
DNN	0.856	0.841	0.847	0.857	0.836	0.765	0.826

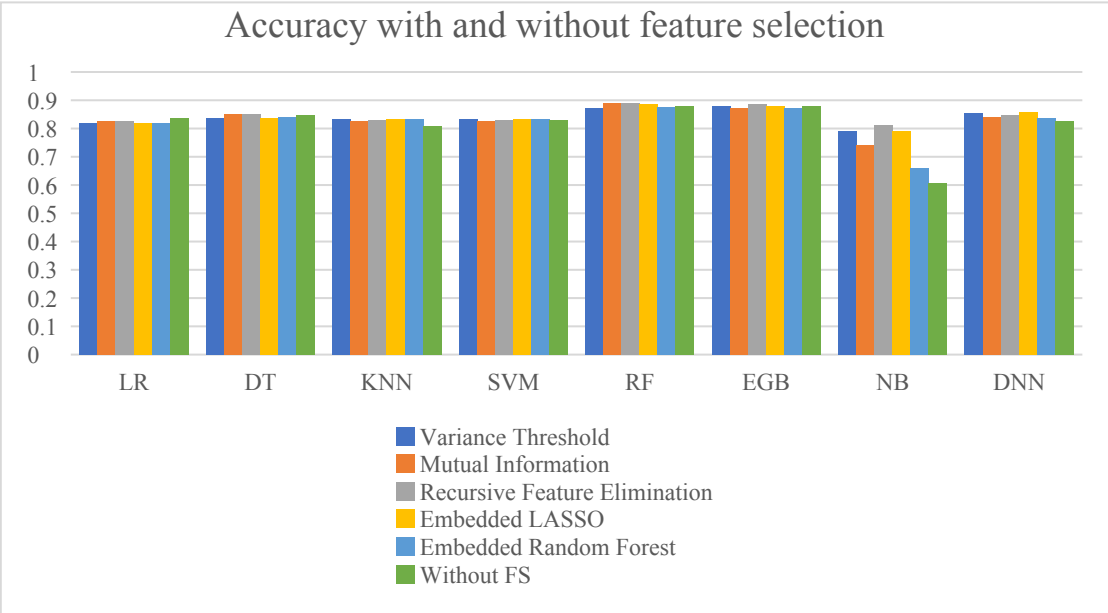


Fig. 8. Comparison of the accuracy of with and without feature selection.

While the Random Forest (RF) and Extreme Gradient Boosting (EGB) models showcased commendable performance even in the absence of feature selection, registering accuracies of 0.879 each, their performance was further enhanced to 0.889 and 0.887, respectively, through the implementation of feature selection techniques. However, it's noteworthy that feature selection methods do not universally benefit all models. For instance, the Logistic Regression (LR) model experienced a decrease in accuracy from 0.836 without feature selection to a lower value post-feature selection. In contrast, aside from LR, the accuracy of most models improved after feature selection, with the Naive Bayes (NB) model witnessing a substantial increase from an accuracy of 0.607 without feature selection to 0.812 with it.

Despite the relevance of expert experience-based feature selection to the financial sector, the resultant model accuracy did not meet expectations. To offer a more holistic evaluation of feature selection, a balance was sought between the Comprehensive Credit Fraud Detection (CCFD) and accuracy. This led to the adoption of the J-A score for evaluating feature selection methods. Table 5 presents the J-A scores for five feature selection methods across eight ML models, thereby providing a quantified measure of their impact on model efficacy.

Table 5
Comprehensive Evaluation of feature selection methods.

	Variance Threshold	Mutual Information	Recursive Feature Elimination	Embedded LASSO	Embedded Random Forest
LR	0.592	0.513	0.595	0.592	0.534
DT	0.601	0.526	0.608	0.601	0.546
KNN	0.599	0.513	0.597	0.599	0.542
SVM	0.598	0.514	0.597	0.598	0.543
RF	0.619	0.545	0.627	0.626	0.563
EGB	0.621	0.537	0.626	0.621	0.561
NB	0.577	0.472	0.588	0.577	0.455
DNN	0.610	0.521	0.606	0.611	0.508

The Random Forest (RF) model achieved the pinnacle of efficiency with a J-A score of 0.627 when employing Recursive Feature Elimination for feature selection. Despite achieving an accuracy of 0.889 with both Mutual Information and Recursive Feature Elimination methods, the features selected through Recursive Feature Elimination were found to be more pertinent to the financial sector. Figure 9 elucidates the importance ranking of features within a Random Forest model developed using the feature set derived from Recursive Feature Elimination. Notably, 'Lending Time' emerged as the feature with the highest importance, scoring 0.412. This was followed by 'Working City,' which secured a feature importance score of 0.17, while the importance scores of all other features remained below 0.1, indicating a significant disparity in the impact of different features on the model's predictive performance.

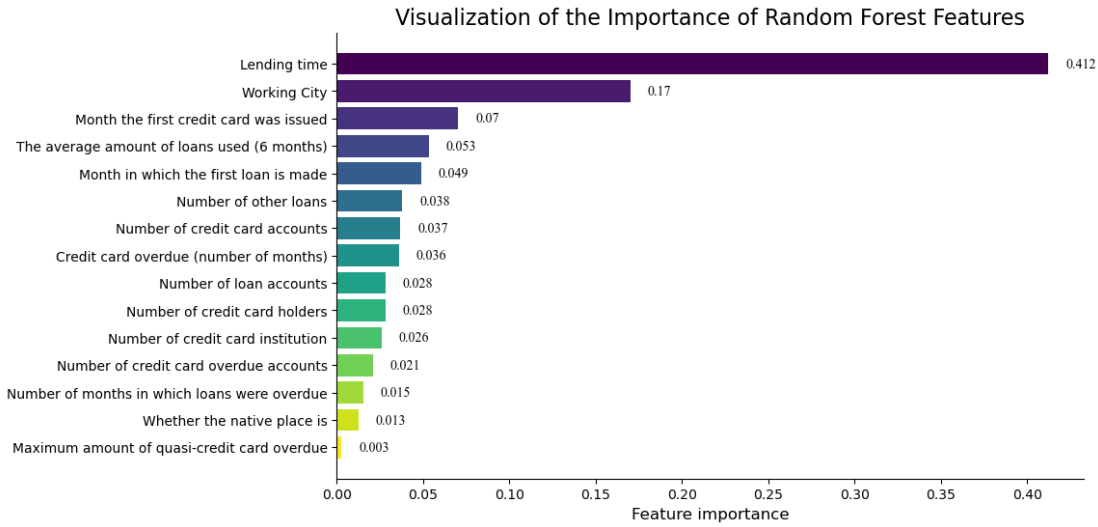


Fig. 9. Random Forest Feature Importance Ranking.

In addition, this part puts forward the theoretical and practical implications of this study.

Theoretical Implications: This research comprehensively compares the impact of five feature selection methodologies on the performance of eight machine learning (ML) classification algorithms, elucidating the nuanced effects—both positive and negative—of feature selection on prediction accuracy contingent upon the algorithm utilized, as corroborated by Olu-Ajayi et al. (2023). Our empirical findings reinforce the assertion that the outcome of feature selection significantly varies across different algorithms; notably, the Naive Bayes (NB) model witnessed a substantial accuracy enhancement of 0.214 through the application of Recursive Feature Elimination, whereas all five feature selection methods detrimentally affected the prediction accuracy of the Logistic Regression (LR) model.

Contrary to the predominant focus on prediction accuracy in prior research as the sole metric for evaluating feature selection methods (Prasetiyowati et al., 2021), this study advocates for a more holistic assessment approach. By integrating accuracy with the Comprehensive Credit Fraud Detection (CCFD) metric, this research not only amplifies the financial domain relevance of the selected features but also enhances the interpretability of the models in terms of feature significance (Murdoch et al., 2019). This dual-faceted evaluation methodology underscores the imperative of considering both predictive efficacy and domain-specific interpretability in the appraisal of feature selection techniques, thereby contributing to a more nuanced understanding of their strategic value in machine learning applications within the financial sector.

Practical Implications: Identifying the most pertinent features for credit risk prediction is paramount

for several reasons. Primarily, in the realm of credit risk model development, leveraging only the relevant features can significantly enhance the prediction accuracy of the model. Moreover, minimizing feature dimensions serves to simplify the model's complexity, thereby mitigating the risk of overfitting. This streamlining also yields benefits in reducing both the data collection workload and the computational demands of the model. Furthermore, by focusing on features with high relevance, banking professionals can devise targeted strategies to curtail credit risk effectively. In essence, selecting the optimal feature selection technique and forecasting model is crucial for augmenting the accuracy of credit risk predictions. This study has identified the Random Forest (RF) model as the most efficacious predictor of credit risk, with Recursive Feature Elimination emerging as the most suitable feature selection methodology when applied to actual credit data.

VI. Conclusion and recommendation

This research investigates the impact of five feature selection methodologies on the efficacy of eight machine learning (ML) algorithms. While the prevailing consensus, as noted by Alaka et al. (2018), suggests that feature selection inherently enhances model accuracy, our findings reveal a more nuanced reality. Specifically, the influence of feature selection on model performance is dual-faceted, manifesting both positive and negative outcomes. This dichotomy underscores the critical importance of appropriately aligning feature selection techniques with predictive models, illustrating that the effectiveness of feature selection is contingent upon the strategic pairing of these methodologies with suitable prediction algorithms.

In our experimental analysis, the Naive Bayes (NB) model initially demonstrated an accuracy of 0.607 without the application of feature selection techniques. However, upon implementing Recursive Feature Elimination, a notable increase in accuracy to 0.812 was observed, marking an improvement of 0.205. This finding indicates that the other four feature selection methods similarly enhanced the accuracy of the NB model to varying extents. Furthermore, the utilization of feature selection techniques also resulted in accuracy enhancements for both the K-Nearest Neighbors (KNN) and Deep Neural Networks (DNN) models.

Conversely, the Logistic Regression (LR) model exhibited a general decrease in accuracy following the application of all five feature selection methods, compared to scenarios where feature selection was not employed. It is also significant to note that the Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (EGB) models exhibited variable accuracy outcomes—experiencing both improvements and declines—contingent upon the choice of feature selection method. For instance, the accuracy of the DT model diminished subsequent to the application of the Variance Threshold method but increased following the use of Mutual Information and Recursive Feature Elimination.

The empirical outcomes of this study suggest that, among the eight scrutinized machine learning algorithms, the Random Forest model emerges as the most efficacious for credit risk prediction. Moreover, in alignment with our proposed comprehensive evaluation index for feature selection methodologies, Recursive Feature Elimination is identified as the most appropriate feature selection technique for synergizing with the Random Forest model, thereby optimizing predictive performance.

Given the plethora of feature selection techniques and machine learning (ML) algorithms available, this study's scope was necessarily limited to a select number of feature selection methods and ML algorithms. This limitation implies that the conclusions drawn—namely, the efficacy of Random Forest (RF) and Recursive Feature Elimination for credit risk prediction—while significant, cannot be deemed exhaustive. Consequently, there is a compelling need for future research to extend beyond the confines of this investigation. Future studies should endeavor to explore a broader array of feature selection methods and ML algorithms. This expanded investigation would not only provide a more comprehensive understanding of the landscape but also facilitate a nuanced comparison between these additional methodologies and those examined in the current study, such as RF and Recursive Feature Elimination. Such comparative analyses are vital for refining predictive accuracy and advancing the domain of credit risk assessment.

References

[1]. Aksakalli V, Malekipirbazari M. Risk assessment in social lending via random forests. Expert Systems with Applications, 2015, 42(10): 4621-4631.
[2]. Alaka H A, Oyedele L O, Owolabi H A, et al. Systematic Review of Bankruptcy Prediction Models: Towards A Framework for Tool Selection. Expert Systems with Applications, 2017, 94(MAR.):164-184.
[3]. Aram K Y, Lam S S, Khasawneh M T. Cost-sensitive max-margin feature selection for SVM using alternated sorting method genetic algorithm. Knowledge-based systems, 2023, 267, 110421.
[4]. Arora N, Kaur P D. A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. Applied Soft Computing, 2019, 86: 105936.
[5]. Baser F, Koc O, Selcuk-Kestel A S. Credit risk evaluation using clustering based fuzzy classification method. Expert Systems with Applications, 2023, 223, 119882.
[6]. Battiti R. Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on Neural Networks, 1994, 5, 537-550.

[7]. Carrington A M, Manuel D G, Fieguth P W et al. Deep ROC Analysis and AUC as Balanced Average Accuracy, for Improved Classifier Selection, Audit and Explanation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(1): 329-341.
[8]. Chen T, He T, Benesty M. XGboost: Extreme Gradient Boosting. 2016.
[9]. Cui L, Bai L, Wang Y, et al. Fused Lasso for Feature Selection using Structural Information. Pattern Recognition, 2021, 119, 108058.
[10]. Cui L, Bai L, Wang Y, Jin X, Hancock E R. Internet financing credit risk evaluation using multiple structural interacting elastic net feature selection. Pattern Recognition, 2021, 114, 107835.
[11]. Eelbode T, Bertels J, Berman M et al. Optimization for Medical Image Segmentation: Theory and Practice when evaluating with Dice Score or Jaccard Index. IEEE Transactions on Medical Imaging, 2020, 39(11): 3679-3690.
[12]. Espinosa R, Jimenez F, Palma J. Surrogate-Assisted and Filter-Based Multiobjective Evolutionary Feature Selection for Deep Learning. IEEE Transactions on Neural Networks and Learning Systems, 2023, DOI: 10.1109/TNNLS.2023.3234629.
[13]. Esteban O, Ciric R, Finc K, et al. Analysis of task-based functional MRI data preprocessed with fMRIPrep. Nature Protocols, 2020, 15, 2186-2202.
[14]. Gamba A, Saretto A. Growth Options and Credit Risk. Management Science, 2020, 66(9): 3387.
[15]. García-Céspedes R, Moreno M. The generalized Vasicek credit risk model: A Machine Learning approach. Finance Research Letters, 2022, 47, 102669.
[16]. He Z, Wu Z, Xu G, Liu Y, Zou Q. Decision Tree for Sequences. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(1): 251-263.
[17]. Hidalgo-Munoz A R, Lopez M M, Santos I M, et al. Application of SVM-RFE on EEG signals for detecting the most relevant scalp regions linked to affective valence processing. Expert Systems with Applications, 2013, 40(6):2102-2108.
[18]. Jiang L, Kong G, Li C. Wrapper Framework for Test-Cost-Sensitive Feature Selection. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2019, 51(3): 1747-1756.
[19]. Kruppa J, Schwarz A, Arminger G, Ziegler A. Consumer credit risk: Individual probability estimates using machine learning. Expert Systems with Applications, 2013, 40(13):5125-5131.
[20]. Lappas P Z, Yannacopoulos A N. A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment. Applied Soft Computing, 2021, 107, 107391.
[21]. Li J, Cheng K, Wang S, et al. Feature Selection: A Data Perspective. ACM Computing Surveys, 2016, 50(6): 3136625.
[22]. Li W, Li C, Jiang L. Learning from crowds with robust logistic regression. Information Sciences, 2023, 639, 119010.
[23]. Liu J, Zhang S, Fan H. A two-stage hybrid credit risk prediction model based on XGBoost and graph-based deep neural network. Expert Systems with Applications, 2022, 195, 116624.
[24]. Machado M R, Karray S. Assessing credit risk of commercial customers using hybrid machine learning algorithms. Expert Systems with Applications, 2022, 200:116889-.
[25]. Maldonado S, Weber R. A wrapper method for feature selection using support vector machines. Information Sciences, 2009, 179(13):2208-2217.
[26]. Mishra P, Biancolillo A, Roger J M, Marini F, Rutlrdge D N. New data preprocessing trends based on ensemble of multiple preprocessing techniques. Trends in Analytical Chemistry, 2020, 132, 116045.
[27]. Murdoch W J, Singh C, Kumbier K, et al. Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 2019, 116(44):201900654.
[28]. Nali J, Martinovi G, Agar D. New hybrid data mining model for credit scoring based on feature selection algorithm and ensemble classifiers. Advanced Engineering Informatics, 2020, 45, 101130.
[29]. Olu-Ajayi R, Alaka H, Sulaimon I, Balogun H, Wusu G, Yusuf W, Adegoke M. Building energy performance prediction: A reliability analysis and evaluation of feature selection methods. Expert Systems with Applications, 2023, 225: 120109.
[30]. Oreski S, Oreski G. Genetic algorithm-based heuristic for feature selection in credit risk assessment. Expert Systems with Applications, 2013, 41(4): 2052-2064.
[31]. Prasetiyowati M I, Maulidevi N U, Surendro K. Determining threshold value on information gain feature selection to increase speed and prediction accuracy of random forest. Journal of Big Data, 2021, 8(1).
[32]. Ribeiro M T, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. ACM, 2016.
[33]. Roffo G, Melzi S, Castellani U, Vinciarelli A, Cristani M. Infinite Feature Selection: A Graph-based Feature Filtering Approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(12): 4396-4410.
[34]. Rtayli N, Enneya N. Selection Features and Support Vector Machine for Credit Card Risk Identification. International Conference Interdisciplinarity in Engineering, 2021.
[35]. Shi L, Liu Y, Ma X. Credit Assessment with Random Forests. International Conference on Artificial Intelligence and Computational Intelligence, 2011.
[36]. Sun J, Yu H, Zhong G, Dong J, Zhang S, Yu H. Random Shapley Forests: Cooperative Game-Based Random Forests with Consistency. IEEE Transactions on Cybernetics, 2020, 52(1): 205-214.
[37]. Tegunov D, Cramer P. Real-time cryo-electron microscopy data preprocessing with Warp. Nature Methods, 2019, 16, 1146-1152.
[38]. Trivedi S K. A study on credit scoring modeling with different feature selection and machine learning approaches. Technology in Society, 2020, 63: 101413.
[39]. Vommi A M, Battula T K. A hybrid filter-wrapper feature selection using Fuzzy KNN based on Bonferroni mean for medical datasets classification: A COVID-19 case study. Expert Systems with Applications, 2023, 218:119612-.
[40]. Wang S, Zhu W. Sparse Graph Embedding Unsupervised Feature Selection. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2018, 48(3): 329-341.
[41]. Yao G, Hu X, Wang G. A novel ensemble feature selection method by integrating multiple ranking information combined with an SVM ensemble model for enterprise credit risk prediction in the supply chain. Expert Systems with Applications, 2022(Aug.):200.
[42]. Yeoh T W, Daolio F, Aguirre H E, Tanaka K. On the effectiveness of feature selection methods for gait classification under different covariate factors. Applied Soft Computing, 2017, 61, 42-57.
[43]. Yu K, Guo X, Liu L, Li J, Wang H, Ling Z, Wu X. Causality-based Feature Selection. ACM Computing Survey, 2020, 53(5): 1-36.
[44]. Yu L, Zhang X, Yin H. An extreme learning machine based virtual sample generation method with feature engineering for credit risk assessment with data scarcity. Expert Systems with Applications, 2022(Sep.):202.
[45]. Zhang X, Yu L. Consumer credit risk assessment: A review from the state-of-the-art classification algorithms, data traits, and leaning methods. Expert Systems with Applications, 2023, 237, 121484.