SN Computer Science (2023) 4:96
https://doi.org/10.1007/s42979-022-01500-7

SURVEY ARTICLE

The Most Effective Strategy for Incorporating Feature Selection into Credit Risk Assessment

Dalia Atif[1] · Mabrouka Salmi[2,3]

Received: 29 June 2022 / Accepted: 5 November 2022 / Published online: 17 December 2022

Abstract

This paper aims to identify the most effective strategy for incorporating feature selection (FS) into credit risk classification, employing three classifiers: Logistic regression (Logreg), Random Forests (RF), and Support Vector Machine (SVM) with the linear kernel through various embedded and wrapper strategies existing in the literature. We performed a comparative analysis on the German Credit dataset using three criteria: classification error rate, stability of selection, and calculation time. According to the Welsh t-test, RFE-RF (Recursive Feature Elimination for RF) outperformed RFE-SVM and penalized Logistic regression, with no significant difference in F1-score for RFE-SVM and suffers from the long-running computation. Conversely, RFE-SVM offers the best stability of 71% with a significantly shorter computation time. Furthermore, the paper intends to introduce a new classification of feature selection strategies in credit risk assessment in light of recent developments. Based on this new classification, a comparison with related literature reveals that the one-stage FS (RFE-RF and RFE-SVM) provides roughly the same accuracy as the two-stage FS and the two-stage classification model and, in some cases, outperforms.

Keywords Credit risk classification · One-stage FS · Two-stage FS · Two-stage classification model · Stability

Introduction

Credit risk classification is an old discipline on which banks rely to measure the risk incurred when granting credit. The question then becomes whether a borrower can repay his loan without causing the bank significant losses. Thus, the 5C rule guides the bank's search for borrower features: Credit history, repayment Capacity, Capital, Collateral, and loan Condition [4, 13]. Because of the increased volume of data, banks have begun to automate decision-making through ML tools such as Random Forests (RF), Support Vector Machine (SVM), and Naive Bayes (NB) [32, 34, 41], with variable performance forcing researchers to improve the predictive qualities of classifiers constantly. As a result, we are now witnessing the emergence of classification by an ensemble of predictors [35, 45] or by hybrid systems [28, 33]; The second strategy differs from the first in that it uses a single predictor to classify data [52]. Today, this field of research has expanded with the problem of feature selection, the choice of the optimal subset being essential in the prediction quality; it remains an NP-Hard problem, solved by heuristics, and does not guarantee an optimal solution. Three strata-filter, wrapper, and embedded-are used to categorize various coping mechanism [22]. The first is still the most commonly used in the related fields [11, 50]. However, two drawbacks lurk behind this strategy: first, the inability to judge the prediction quality resulting from the subset selection [19]. Second, because noisy features and redundancies

This article is part of the topical collection "Recent Trends on Machine Learning & Intelligent Systems" guest edited by Akram Bennour, Tolga Ensari and Abdel-Badeeh Salem.

Study conception and design: Dalia Atif; analysis and interpretation of results: Dalia Atif; draft manuscript preparation: Dalia Atif, Mabrouka Salmi. All authors reviewed the results and approved the final version of the manuscript.

✉ Dalia Atif
atif.dalia@cu-tipaza.dz

Mabrouka Salmi
salmi.mabrouka@enssea.net

[1] Economics, University of Tipaza, 42000 Tipaza, Algeria

[2] Applied Statistics, National School of Statistics and Applied Economics, 42003 Kolea, Algeria

[3] Advanced Computing, University of Cordoba, 14014 Cordoba, Spain

swamp the data, feature screening can cause the optimization to be trapped at a local minimum [4]. As a result, the wrapper and embedded strategies propose incorporating the FS into the learning step [47]. The feature selection aims to use an assessment criterion $C(J)$ to reduce dimensionality and find the optimal subset J over 2^p possibilities [37]. This criterion can be, for example, the correlation-based feature (CFS) or Fisher score in the case of the screening feature, the prediction error on a hold-back sample or by using cross-validation, or the Out of bag error (OOB) in the case of RF [8]. The embedded strategy incorporates the FS into the optimization; it can be intrinsic to the model, as with decision trees, or imputed to the model by a penalty of the L1-norm type [30], as in the case of linear lasso (Least Absolute Shrinkage and Selection Operator) regression [49] or penalized SVM [54]. When features are correlated, the Lasso regression tends to select one feature per cluster [19]; a more elaborate version using the bootstrap (Bootstrap-lasso) stabilizes the selection [5]. On the other hand, wrapper selection is a more common strategy; it uses a classifier (RF, SVM, or NB) and applies recursive elimination based on classification accuracy. Many other wrapper strategies have been recently proposed that rely on an evolutionary search with the accuracy of a classifier as a fitness function.

Jadhav et al. [28] compare the accuracy of three classifiers as fitness functions: KNN, SVM, and NB, and Lappas et al. [33] compare two classifiers: KNN and NB. Other researchers have proposed a modified version of the RFE-RF, Mariammal et al. [36], which involve shuffling the data matrix, and Mustaqeem et al. [38], which propose the addition of feature replicas. The most commonly used dataset in this field is the German credit dataset [3, 28, 58], where recently researchers have used hybrid systems; Arora and Kaur [3] used Bolasso, and Zhou et al. [58] used MARS (multivariate adaptive regression splines) as a screening stage and penalized logistic regression-based FS, Jadhav et al. [28] compare the accuracy of three classifiers: SVM, KNN, and NB, in three configurations: as a baseline, as a wrapper FS (GA+Classifier), and as a two-stage FS (Filter then GA+Classifier), they get a clear accuracy improvement in the case of the wrapper strategy and a derisory improvement in the case of the two-stage FS only for SVM and NB. If the two-stage FS can offer a pitiful improvement in accuracy, it can also introduce bias in selection and increase complexity and computation time [33]. Thus, this work aims to compare different classifiers used in credit assessment (SVM, Logreg, and RF) and how they integrate the FS within learning using different strategies (embedded or wrapper). This work is an extension of a conference paper [4] in which a comparison was made between only two algorithms: penalized Logreg and NRFE-RF (Non Recursive Feature Elimination for RF), and according to a single criterion: classification-based measures. In this work, we seek to designate the most practical strategy for banks in terms of accuracy, stability, and computation time and which integrates the FS within the learning process to avoid selection bias. The dataset used in practical experiments is the standard German credit dataset, intending to be able to compare the results obtained with other strategies proposed in the related literature. Another goal of this research is to propose a new classification of FS strategies in credit risk evaluation.

The remainder of the paper will be structured as follows: In "Background" section, we present a new classification of the feature selection in credit assessment. "Methodology" section introduces the algorithms used in the comparison and the tools used for it, and "Practical Experiments" section describes the stages of experimentation. "Results and Discussion" section compares the results, and "Conclusion and Future Works" section concludes the research.

Background

Feature selection has emerged as the primary research topic in credit assessment. It has been the subject of many studies over the last decade, some of which introduce new heuristics [23, 33, 40], but most question the best combination amid the feature selection strategies and the choice of the classifier; among these practical researches, we can cite [3, 11, 28, 50, 58]. Other researchers assess credit risk without using feature selection either by hyperparameter tuning [31] or an ensemble model [35, 45] to avoid selection bias. Oreski and Oreski [40] used different filter feature selections and prior knowledge as potential chromosomes in the initialization of the genetic algorithm, which is especially beneficial because it allows the algorithm to converge faster by narrowing the space of solutions to more attractive points. Ha and Nguyen [23] employed a hybridization between RF and deep learning and selected the most important features using a score based on cross-validation accuracy. Dahiya et al. [11] utilized the PCA/chi-square test to select features in a screening-based selection, and the top-ranked features are advanced to a bagging ensemble model with C4.5/MLP as weak learners. Jadhav et al. [28] operated information gain to reduce the genetic algorithm's search space from 2^p to $2^{p'}$. The fitness functions used are Naive Bayes, KNN, and SVM, and the results show that the SVM classifier offers the best accuracy due to its robustness. Tripathi et al. [50] proposed a more sophisticated credit assessment model in which a heterogeneous deep learning ensemble model makes the prediction, and a ranking does the placement of the classifiers in the layers based on several criteria. Arora and Kaur [3] introduce applied research that aims to determine the best FS strategy and the best classifier in the case of credit scoring; they use four real datasets based on three criteria: accuracy, computation time, and selection stability and conclude that the best combination is Bolasso-based feature selection and RF prediction. In the same vein, Zhou et al. [58] compare different credit dataset situations (high dimension, multicollinearity...) with different FS+classifier configurations in an empirical study and thus seek to designate the most robust classifier to different FS strategies. Shen et al. [45] address the issue of unbalanced classes more precisely by proposing an improved version of SMOTE; once the classes are rebalanced, an adaptive homogeneous ensemble model with recurrent ANN as the base classifier is used for prediction. Liu et al. [35] embedded feature augmentation into gradient boosting using a multi-grained procedure to introduce feature variousness inspired by bagging. The gradient boosting algorithm based on the minimization of the loss function is therefore improved by reducing its variance and, consequently, improving its generalization power. Lappas and Yannacopoulos [33] proposed several scenarios using clustering feature subsets; then, a prior knowledge expert serves AHP pairwise decision-making as a filtering phase. All scenarios are then forwarded to genetic algorithm-based feature selection to find the optimal subset.

While the hybridization of feature selection strategies has resulted in the publication of many papers as a solution to an NP-Hard problem [3, 23, 28, 33, 40, 58], the use of homogeneous or heterogeneous ensemble models aims to achieve prediction robustness. If Guyon and Elisseeff's [22] classification in three strata remains relevant, a new classification in four strategies must be proposed in light of the latest advances:

1. **Filter Feature Selection**: Any selection whose purpose is to reduce the dimension as a first step and model the data as a second step is referred to as a filter feature selection. It takes two forms:
   - Pre-selection by feature ranking measures: like correlation or mutual information.
   - Pre-selection by ensemble feature ranking: accomplish the selection through the union, intersection, or aggregation of several subsets derived from feature ranking measures. [10, 50]
2. **One-stage feature selection**: incorporating the FS into the learning process can be done via embedded methods such as decision trees or penalized classifiers or a wrapper strategy that uses an optimization algorithm to generate several subsets (multiple possible solutions) and selects based on classifier accuracy or recursive feature elimination algorithms.
3. **Two-stage feature selection**: this strategy entails carrying out the selection in two stages: the first, aimed at feature screening, is performed using either a ranking measure or ensemble ranking measures. The wrapper FS performs the second step; many recent works rely on this strategy. [1, 28, 33, 40]
4. **Two-stage classification model**: feature screening is performed using an embedded or wrapper FS such as RFE-SVM, lasso, or Bolasso, and prediction is made in the second step using another classifier or ensemble model [3, 23, 36, 58].

Table 1 lists the works in the related field that have been discussed and classified using the above classification.

Methodology

"Feature Selection Strategies" section describes the feature selection strategies used to conduct the experiments, namely recursive feature elimination/penalized SVM, stepwise/penalized logistic regression, and RFE/NRFE random forests. "Classification-Based Measures" section and "Stability Measure" section describe the tools used in our comparison, namely classification-based measures, and stability measure.

Feature Selection Strategies

Consider our dataset $\Psi = \left\{(x_1, y_1), \ldots, (x_n, y_n)\right\}$ with $x_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$, in which each individual $\psi \in \Psi$ is associated with a value $X(\psi) = (X_1(\psi), X_2(\psi), \ldots, X_p(\psi))$. The goal of any feature selection strategy is to reduce the training space (feature set) $A = \left\{X_1, \ldots, X_j, \ldots, X_p\right\}$ of any classification model $\varphi(x)$ to $\mathbb{R}^J$. The section's final goal is to highlight the feature selection strategies used.

SVM-based feature selection

The SVM was first proposed in Vapnik's work [51] as part of the discrimination of instances that are linearly separable by a hyperplane; however, the existence of a plethora of potentially discriminating separators led to the intuition behind this machine learning tool, which involves identifying the hyperplane that allows for the greatest interclass margin. In this way equates to looking for a function $\varphi(x)$ with equation $\beta^T x + \beta_0 = 0$ satisfying the conditions:

$$\begin{aligned} \beta^T x_i + \beta_0 \geqslant 1 \rightarrow y_i = +1 \\ \beta^T x_i + \beta_0 \leqslant -1 \rightarrow y_i = -1 \end{aligned} \tag{1}$$

The two equations can be put together into a single equation:

$$y_i(\beta^T x_i + \beta_0) \geqslant +1 \tag{2}$$

Table 1 Review of related literature

Classifier	FS	FS-type	Datasets	Evaluation metrics	References
HGA-NN	Information gain Gain ratio Gini index Correlation Earlier experience	Two-stage FS	Croatian, German datasets	Accuracy, t-test, Wilcoxon test	[40]
FS approach based deep learning	Rndom forests	Two-stage Classification model	German, Australian datasets	Accuracy, time calculation	[23]
FS-HB with MLP and C4.5 as base classifiers	Chi-square, PCA	Filter FS	German dataset	Percentage error, Type I error, type II error, AUC	[11]
IGDFS[1]	IG[2]	Two-stage FS	German, Australian, Taiwan datasets	Accuracy, AUC	[28]
Multilayer ensemble classification	Ensemble feature selection	Filter FS	Australian, Japanese, German datasets	Accuracy, sensitivity, specificity, G-measure	[50]
Bolasso-RF	Bolasso	Two-stage classification model	German, LC, kaggle's bank loan	Accuracy, AUC,time calculation, stability	[3]
Adaboost with LTSM Network[3] as base classifiers	WFS[4]	Ensemble model WFS	German, Taiwan datasets	AUC, KS[5], Wilcoxon test	[45]
SVM, CART, KNN, Logreg	MARS Lasso	Two-stage classification model	German, Australian, Japanese, Chinese datasets	Accuracy, AUC, type I error, type II error	[58]
mg-GBDT[6]	WFS	Ensemble model WFS	German, Taiwan, Japanese Australian, LC,WE datasets	Accuracy, AUC,Brier score,recall, precision	[35]
GA+ NB/KNN	Expert knowledge	Two-stage FS	German, dataset	AUC, Gini Coefficient,time calculation	[33]

[1] Information gain directed feature selection algorithm

[2] Information gain

[3] Long-short-term memory network

[4] Tuning classifier parameters without FS

[5] Kolmogorov–Smirnov

[6] Multi-grained augmented gradient boosting decision trees

And the parallel hyperplanes passing through the support vectors satisfy the equations $\varphi(x) = +1$, $\varphi(x) = -1$ (see Fig. 1). Estimating the vector of weights β and the bias β_0 is equivalent to solving a convex optimization problem [42, 54].

$$\underset{\beta_0,\beta}{max}\frac{1}{\|\beta\|_2^2} \tag{3}$$

$$\text{Subject to } y_i(\beta^T x_i + \beta_0) \geqslant +1 \tag{4}$$

The primal formulation of this problem can be rewritten using the L2 norm of the form (loss+penalty) [54].

$$\underset{\beta_0,\beta}{min}\sum_{i=1}^{n}\left[1 - y_i(\beta^T x_i + \beta_0)\right]_+ + \frac{\lambda}{2}\|\beta\|_2^2 \tag{5}$$

Which amounts to minimizing the loss under the shrinkage constraint of the hyperplane weights. Thus solving this problem leads to a ridge-type optimization and, in the case of high dimension, requires a feature preselection step. Several researchers [6, 15, 39, 42, 54] have investigated the use

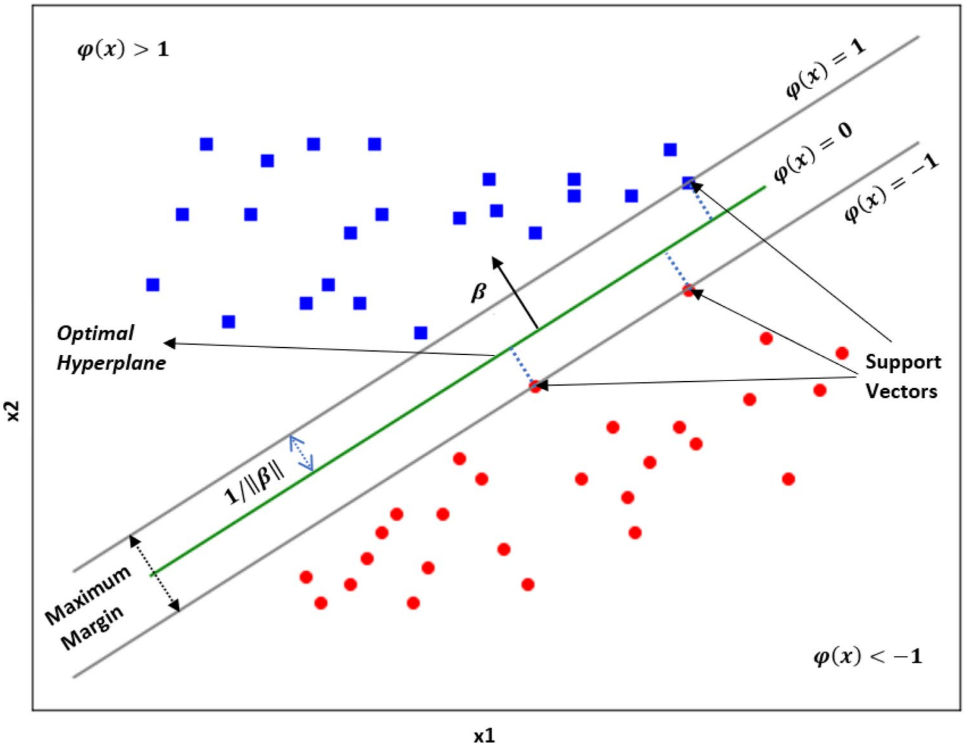


Fig. 1 Linear SVM's optimal hyperplane and support vectors

of the L1 norm to produce sparse-type solutions; the problem is then written as follows:

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} \left[1 - y_i(\beta^T x_i + \beta_0)\right]_+ + \lambda\|\beta\|_1 \tag{6}$$

Furthermore, the optimization is driven by the fast Newton method for linear programming of SVM [15] and is available in the lpsvm function of the penalizedSVM package in R [6]. The main drawback of the linear L1-SVM is its linearity in the shrinkage of the hyperplane weights, which introduces a bias in the selection, and this defect is managed by nonconvex penalties such as smooth clipped absolute deviation (SCAD), minimax concave penalty(MCP), and log-sum penalty (LSP) [20].

Another way to include feature selection in the learning phase via a sequential-type algorithm is Recursive Feature Elimination-SVM; features are eliminated backward after sorting them from most important to least important based on the feature weight magnitude in the hyperplane β_j^2. As described Guyon et al. [21], the procedure can be divided into four stages:

1. Train a model, then sort features in a ranked list based on the magnitude of the weights.
2. Eliminate the least important feature.
3. Evaluate the model trained on a feature subset by using leave one out and calculating the mean error of $\varphi(x)_{k-1}$.[1]
4. Repeat steps 1-3 until the last feature is left, then the lowest mean error is the model selection criteria among: $\varphi_1 \subset \ldots \subset \varphi_J \ldots \subset \varphi_p$

Algorithm 1 shows the pseudo-code of the above procedure. One algorithm drawback is the cost of computation time in high-dimension cases, but it can be sped up by removing several features at each iteration [21]. Thus, the procedure initially designed for linear kernels can be extended to nonlinear SVMs, as demonstrated by Refs. [44, 57]; however, this procedure may be applied to nonlinearly separable data and yield satisfactory results [21]. It has been widely used in the biomedical literature and has produced excellent classification accuracy results [12, 26, 56].

[1] A model trained on a subset of $k - 1$ features.

Algorithm 1: Pseudo-code of RFE-SVM algorithm

```
input  : $\Psi$ : Dataset
         $p$: Number of features $A$ : Feature set
output: $S_J$: Optimal feature subset
         SVM model.
repeat
    for $S_k \subseteq A$ do
        Train SVM model $\varphi_k$ on $S_k$
        Sort features based on the magnitude of estimated feature
          weight
        Eliminate the least important feature with
          $X_j = argmin_{X_j \in S_k}(\beta_j^2)$
        Evaluate the trained SVM model $\varphi_{k-1}$ using the leave-one-out
        Calculate the mean error of $\varphi_{k-1}$
    end
    Update $S_k = S_k - X_j$
    Update $k = k - 1$
until the last feature is left
Select SVM model with the lowest mean error among:
  $\varphi_1, \varphi_2, ..., \varphi_J, ..., \varphi_p$, the Recursive Feature Elimination-SVM model
Set $S_J$: optimal feature subset = the feature subset of the Recursive
  Feature Elimination-SVM
```

Logistic Regression-Based Feature Selection

Logistic regression is widely used to classify credit risk and is still the reference in this field [32, 55, 58]. Borrowers are classified by computing the posterior probability $\pi(\psi) = p(Y(\psi) = 1/X(\psi))$; the model parameters are then estimated using Newton's method by optimizing a convex function, which amounts to minimizing the negative log-likelihood.

$$LL = \sum_{\psi} y(\psi) \ln \pi(\psi) + (1 - y(\psi)) \ln(1 - \pi(\psi)) \tag{7}$$

The model relies on the power of statistical tests to validate an individual's score; however, it has a significant drawback in handling the high-dimensional problem in the presence of feature correlations and redundancies; a lasso-type optimization is therefore required, by acting on the bias-variance trade-off, i.e., by increasing the model bias and decreasing its variance; now consider the tuning parameter λ, the goal is to shrink the model parameters to arrive at a sparse solution, the problem formulation then turns into:

$$\hat{\beta} = \min_{\beta}\left[-LL + \lambda \sum_{j=1}^{p} |\beta_j|\right] \tag{8}$$

The coordinate descent algorithm is used for optimization [14], and the R implementation is found in the glmnet package. Despite significant drawbacks summarized in Refs. [24, 46], integrating feature selection within the learning phase of a stepwise regression type is widely used in the literature [53]. The most significant of these drawbacks are the lack of power of statistical tests in high dimensions, the exclusion of potentially discriminating features due to their individual non-significance (pvalue $>$ 0.05), the high variance of selected features in generalization [46], and the selection of non optimal feature subset. In a comparative analysis, Hastie et al. [25] show that lasso regression outperforms stepwise regression. The main algorithm steps are as follows:

- Processing: selecting a statistical criterion for adding or removing a feature.
- Stopping criterion: continue processing until no statistically significant improvement is observed.

Stepwise regression can be performed forward or backward, with the former being more appropriate when there are numerous features and correlations and thus being favored in our study; the related pseudo-code is presented in Algorithm 2.

Algorithm 2: Pseudo-code of forward stepwise feature selection algorithm

```
   input  : $A$ : original feature set
   output: Optimal feature subset
1  Create an empty feature set $S_k = \{\varnothing\}$, $k = 0$
2  Select a statistical criterion for adding features with a threshold $c_0$
3  Add feature $X_j$ only if $C(X_j) > c_0$, for example
4  Select best remaining feature: $X_j^+ = argmax_{X_j \in A \setminus S_k}(C(X_j^+))$ and
     $C(X_j^+) > c_0$
5  if no feature with $C(X_j^+) > c_0$ then
6      Stop the process
7      Set Optimal feature subset $= S_k$
   end
9  Update $S_k = S_k + X_j^+$
10 Update $k = k + 1$
11 Go back to step 3
   In other words (from line 5 to 11):
   while the statistical improvement $I$ is significant    /* not stopping
    criterion */
    do
       if no feature with $C(X_j^+) > c_0$ then
           the statistical improvement is not significant
       end
       Select best remaining feature: $X_j^+ = argmax_{X_j^+ \in A \setminus S_k}(C(X_j^+))$ and
         $C(X_j^+) > c_0$
   end
```

Random Forests-Based Feature Selection

Breiman [8] developed random forests to deal with decision tree instability and to develop classifiers with low variability; RF are based on two fundamental principles: bagging [7] with aggregation of 'a family of decision tree type classifiers $\left\{\varphi(x, \Phi_b), b = 1, ..., B\right\}$ built on B bootstrap samples; and the principle of randomization where Φ_b are iid vectors that allow the classifiers to be decorrelated; It should also be noted that the trees are grown without being pruned, which enables RF to overcome the bias-variance trade-off. As for SVMs, random forests allow feature selection to be integrated within the learning phase using a wrapper strategy; while theoretically, the selection can be made forward or backward, the majority of works use a backward elimination [19, 29] relying on the feature importance by permutation (IP). Many strategies exist in the literature [36, 38]; Gregorutti [19] has divided them into two large families based on whether the feature ranking is static or dynamic (respectively NRFE and RFE). Despite comparison studies between the two strategies, contradictory results emerge [19, 48]; however, theoretically, the ranking update allows for the study of feature importance variability within different subsets [18], whereas NRFE may not be able to meet the model generalization goal. Beyond the ranking criterion, the selection criterion is also fundamental. It is based on the classification error and can be executed either by OOB error or by resampling techniques such as cross-validation. The former can lead to overfitting because it is too optimistic, while the latter, while increasing computational complexity, may offer a probabilistic view of feature ranking.

Typically, the random forests-based feature selection algorithm proceeds in three steps:

1. Ranking process: rank the features in order of importance using the IP criterion.
2. Backward elimination: remove the least important feature.
3. Sort the final subset S_J: select the subset with the lowest classification error $C(J)$.

The main difference between the two procedures (RFE and NRFE, respectively) is that the first strategy requires you to repeat steps 1–2 until you reach step 3, whereas the second strategy requires you to repeat step 2 until you reach step 3 [19]. Since the first strategy is more prevalent, it will be summarized in Fig. 2, but both will be tested further in our research.

Classification-Based Measures

We relied on classification-based measures computed on a hold-back sample (HB) in our comparative analysis; this concept is fundamental in machine learning because it allows us to estimate a model's generalization error; this empirical estimate converges in probability to the generalization error as n_{HB} approaches infinity [2]. The confusion matrix concept and related notions must be introduced before stating the measures used (see Table 2).

- **TP**: Positive class examples are correctly classified.
- **FP**: Examples in the positive category are misclassified.
- **TN**: Negative examples are correctly classified.
- **FN**: Examples in the negative category are misclassified.

We will only present the measures that reflect a classifier's generalization power in the case of imbalanced learning, the most common scenario with credit risk classification datasets.

- **Recall**: represents the proportion of true positives in the positive class sample.

Table 2 Confusion matrix

Classified	Observed	
	Negative	Positive
Negative	TN	FN
Positive	FP	TP

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

- **Precision**: represents the proportion of true positives in a sample predicted to be positive.

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

- **F1-score**: is a harmonic mean that combines the two previous measures.

$$F1 - score = \frac{2Precision.Recall}{Precision + Recall} \tag{11}$$

Stability Measure

Aside from model generalization estimation, another fundamental concept is stability measure, intending to investigate the stability of the feature selection algorithm against sampling fluctuations. The measure used is pairwise similarity (Jaccard Index) across all feature selection subsets J_m [3, 43].

$$stability = \frac{2\sum_{m=1}^{M-1}\sum_{r=m+1}^{M} Sim(J_m, J_r)}{M(M-1)} \tag{12}$$

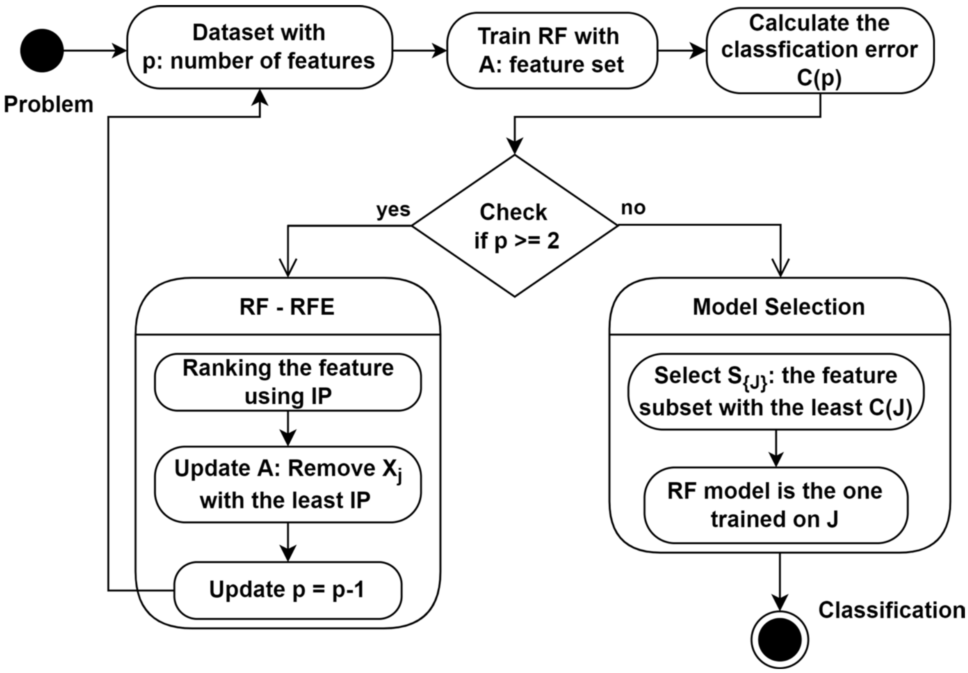


Fig. 2 Recursive feature elimination for random forests diagram

Table 3 The German credit dataset description

Features	Description	Range	Levels
C1	Current account status	–	4
N2	Length in months	[4;72]	–
C3	Credit report	–	5
C4	Credit motivation	–	10
N5	Total credit load	[250; 18424]	–
C6	Deposits account	–	5
C7	Time since present employment	–	5
N8	Installment rate as a proportion of available funds	[1;4]	–
C9	Civil state	–	5
C10	Co-applicant or guarantor	–	3
N11	The length of time spent at the current address	[1;4]	–
C12	Owns real estate or other property	–	4
N13	Age	[19;75]	–
C14	Existence or absence of additional payment arrangements	–	3
C15	Owns or rents a house	–	3
N16	Total credits earned	[1;4]	–
C17	Job	–	4
N18	Maintenance	[1;2]	–
C19	Telephone	–	2
C20	The status of a foreign worker or not	–	2
Target	Borrower status	–	2

$$Sim(J_m, J_r) = \frac{|J_m \cap J_r|}{|J_m \cup J_r|} \tag{13}$$

Because the scaled measure ranges between 0 and 1, it is possible to compare the stability of classifiers of various types.

Practical Experiments

The experiment was divided into three sections: data pre-processing, data transformation, and data processing.

Table 4 Data pre-processing

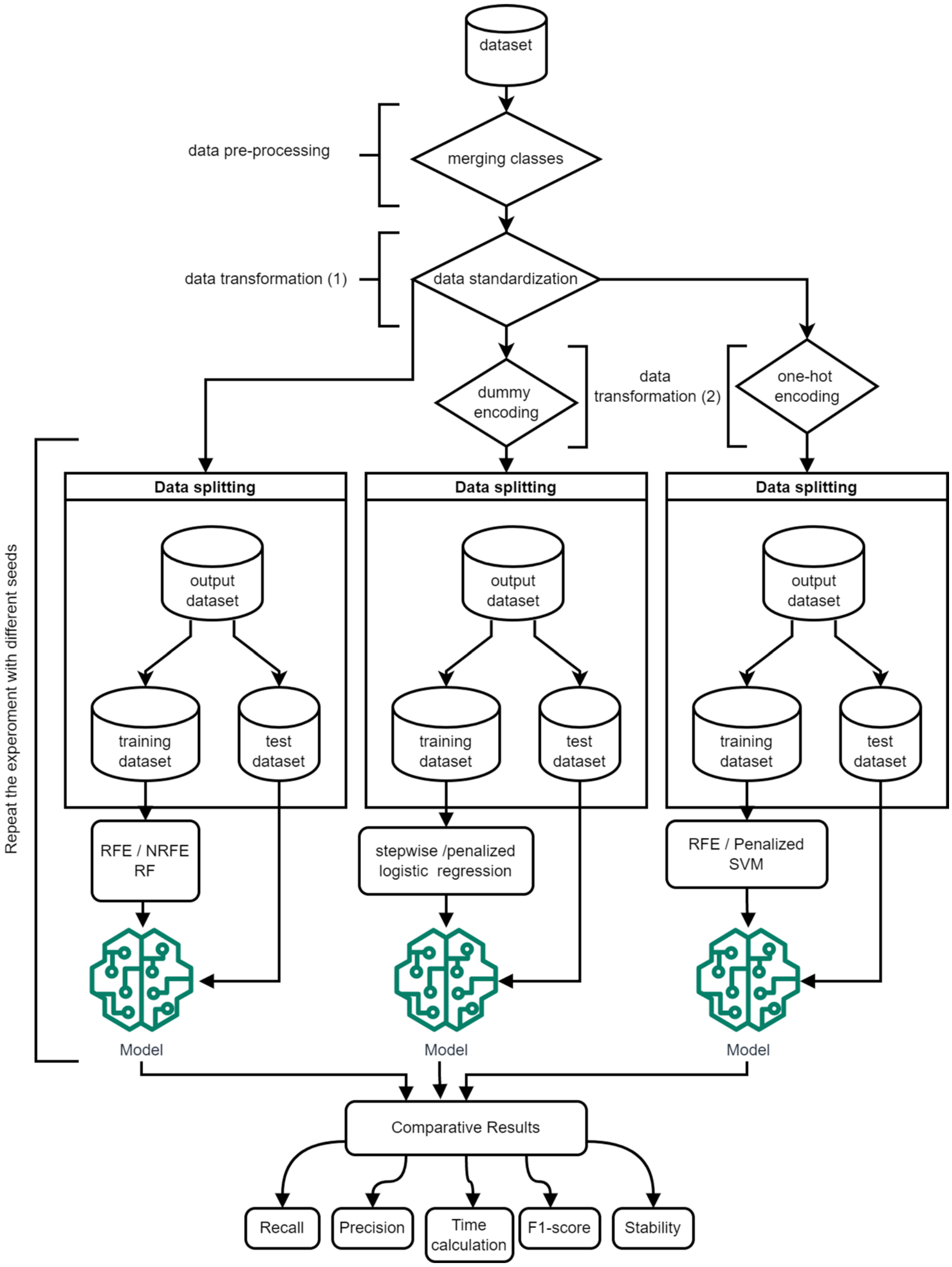
Feature	Initial levels	Merged levels

C7	Unemployed;
	... < 1;
	1 ⩽ ... < 4;
	4 ⩽ ... < 7;
	... ⩾ 7
		0 ⩽ ... < 1;
		1 ⩽ ... < 4;⩾ 4

C6	... < 100;
	100 ⩽ ... < 500;
	500 ⩽ ... < 1000;
	... ⩾ 1000
	no savings account
		⋯ < 100 ;
		100 ⩽ ... < 500;
		⩾ 500 ; no savings account

C9	Male: divorced/separated; female: divorced/separated/married; male: single; male: married/widowed; female: single	Male: married/divorced; male: single; female
C17	Unemployed/unskilled; unskilled - resident; skilled employee; officer	Skilled; unskilled
C4	Business; car (new); car (used); domestic appliances; education; others; television; repairs; retraining; furniture/equipment	Business; car; domestic appliances; others
C3	All credits paid back duly; all credits at this bank paid back duly; existing credits paid back duly till now; delay in paying off in the past; critical account	All credits paid back duly; existing credits paid back duly till now; delay in paying off in the past; critical account

Fig. 3 The workflow diagram



German Credit Dataset

The dataset is widely used in specialized literature [3, 9, 31, 33]. It contains 1000 records divided into 700 good and 300 bad borrowers; 20 predictors describe each borrower, 13 categorical and seven numerical. The dataset was rebalanced for the negative class and thus aligns with the imbalanced dataset; In Table 3, we provide a brief description.

Data Pre-processing

Because categorical features dominate the dataset, the features must have roughly equal numbers of categories hence no feature is favored during selection. The chi-square test is used as the criterion in a bottom-up merging strategy; two categories are merged if the difference in their distributions is insignificant. The operation is detailed in Table 4.

Table 5 Comparative results of SVM-based feature selection

	Penalized SVM			RFE-SVM		
Model	Recall	Precision	F1-score	Recall	Precision	F1-score
1	0.9598	0.7439	0.8382	0.8894	0.7750	0.8283
2	0.8454	0.7778	0.8102	0.8167	0.7722	0.7939
3	0.9135	0.7819	0.8426	0.8685	0.7974	0.8315
4	0.8857	0.7815	0.8304	0.8167	0.7722	0.7939
5	0.8551	0.7763	0.8138	0.8551	0.7937	0.8233
6	0.9268	0.7692	0.8407	0.9292	0.7829	0.8499
7	0.8821	0.7679	0.8210	0.8889	0.8402	0.8643
8	0.8857	0.7815	0.8304	0.8679	0.8598	0.8638
9	0.8502	0.7788	0.8129	0.8737	0.8160	0.8439
10	0.8888	0.7796	0.8307	0.9242	0.8133	0.8652
Mean	0.8879	0.7736	0.8269	0.8715	0.8012	0.8349
Low 95% CI	0.8634	0.7652	0.8184	0.8450	0.7809	0.8159
High 95% CI	0.9138	0.7823	0.8355	0.8997	0.8227	0.8548

Data Standardization

Data standardization entails reducing numeric features to a range of 0–1 with a single goal: do not favor features with a wide range of variation during selection. The z-score is the metric used for this.

Feature Encoding

Two encoding types were used for categorical features, dummy encoding in logistic regression and one-hot encoding in SVM. Feature encoding converts categorical features of l categories into $l - 1$ binary vectors in dummy encoding and l binary vectors in one-hot encoding. Because only one category is used as a reference in odds ratio interpretation, we select the most common class as the reference category; the consequences of its selection have already been studied [27], but this is beyond the scope of our investigation.

Data Splitting

The data splitting goal is to estimate the model's generalization error. With a proportionate stratification strategy [4], the ratio used is 70:30, corresponding to training the model on $n_{TR} = 700$ records and testing the model on the remaining $n_{HB} = 300$ records. To calculate the model's stability, we repeat this operation ten times ($M = 10$) with different seeds for each model.

Therefore, we pre-processed our dataset by merging classes; once we obtained approximately the same number of classes, we went to the data transformation (1) "data standardization" type for all models. Next, the data transformation (2) consists of dummy encoding the categorical features for logistic regression models (penalized-Logreg, stepwise-Logreg) and one-hot encoding the categorical features for SVMs (penalized SVM, RFE-SVM) but not for RF models (RFE-RF, NRFE-RF). Finally, the datasets obtained were 70: 30 split ten times with different seeds, and each time we built the six models on n_{TR} and tested them on n_{HB} to obtain the classification-based measures (recall, precision, F1-score). The repetition of the process ten times allows us to calculate the stability of all models. We schematize the workflow in Fig. 3.

Results and Discussion

SVM-Based Feature Selection Results

For this purpose, two algorithms were used: the RFE-SVM for the wrapper selection strategy with a linear kernel and the embedded strategy by penalized-SVM. For the former, the feature subset was chosen based on the classification error criterion by cross-validation (CV) with CV=30. We repeated the operation on ten different training sets; the results obtained are summarized in Fig. 4; we then tested each model on its corresponding hold-back sample, and summarized performances across several metrics in Table 5.

The second algorithm's kernel is also linear, and the hyperparameters were tuned using grid search with CV=30. The algorithm is implemented in the penalizedSVM package in R, and the test sample results are summarized in Table 5.

In Table 6, we extend our comparison to other criteria: "stability of selection" and "calculation time," and we use the Welsh t-test to see if there is a significant differences between the classification-based measures. Where the execution time for the two algorithms remains substantial, the results show that the RFE-SVM is more precise and steady against sampling fluctuations (see Fig. 5).

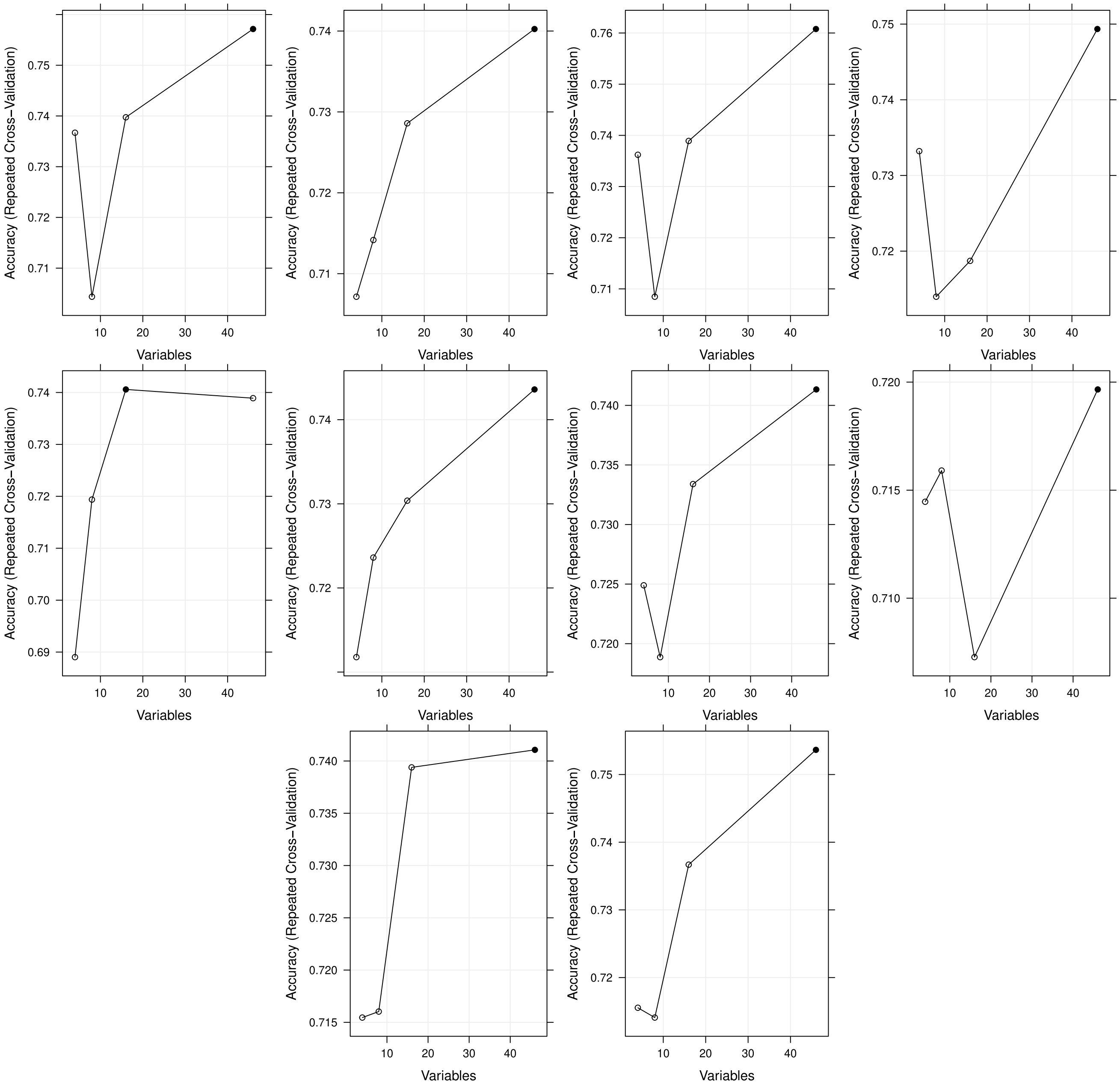


Fig. 4 The feature selection subsets of RFE-SVM for various training sets

Logistic Regression-Based Feature Selection Results

The Akaike criterion was used for the stepwise selection, Whereas the penalized strategy used cross-validation to select the tuning parameter lambda. We choose the value of one standard error of the loglikelihood minimum [4], referred to as lambda.1se, purely for parsimony; the algorithm is implemented in the glmnet package in R. As stated in the practical experiments section, we applied these two algorithms to ten different training sets. The results are shown in Table 7.

Because there is no significant difference in the F1 score, the two strategies produce the same results in terms of classification, and we also noticed this for the execution time; however, the logistic lasso regression provides better stability (see Table 8 and Fig. 6 for more details)

Random Forests-Based Feature Selection Results

We wanted to compare the RFE-RF and NRFE-RF strategies incorporating FS in random forests construction. The former is implemented in the Caret package. It uses the IP criterion

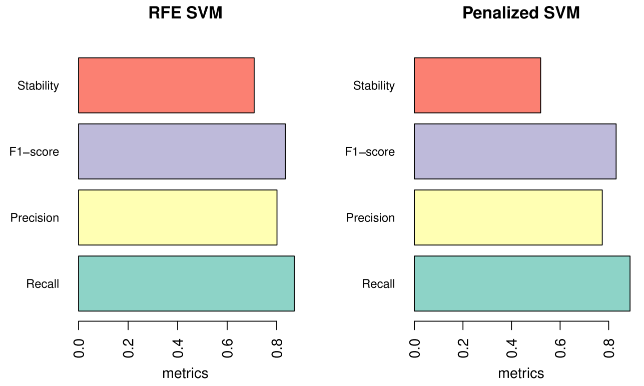


Fig. 5 SVM-based feature selection comparative metrics

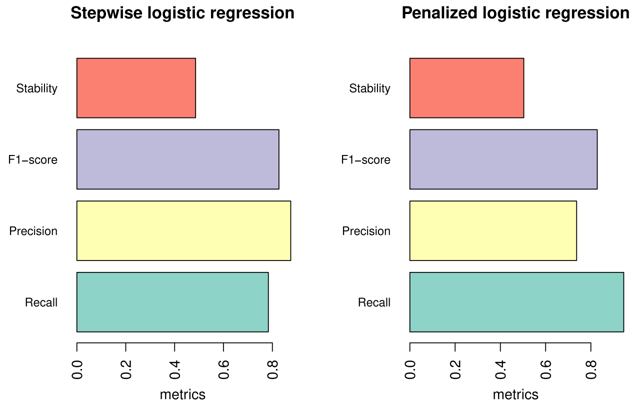


Fig. 6 Logistic regression-based feature selection comparative metrics

for ranking and cross-validation error for feature selection with a CV=30 setting, a strategy to be paralleled with the RFE-SVM algorithm. In contrast to the RFE-RF algorithm, the second algorithm can be misleading because of a rigid ranking tainted with feature correlations and redundancies. We used the Genuer et al. [16] strategy in the "VSURF" package [17]. The strategy principles are:

- Reduce selection bias using dense forests (≈ 2000).
- Increase selection stability by repeating the random forests construction process (≈ 50).

The OOB error is the selection criterion in this case. (For more information, see Ref. [16]).

We used the two strategies on ten training sets and tested the models on the corresponding hold-backs (see Table 9 and Fig. 7 for results).

Table 10 compares the results and shows that the RFE-RF outperforms the NRFE-RF algorithm on classification-based measures and selection stability (Fig. 8). Nonetheless, the execution time for the two algorithms remains significant compared to SVM and logistic regression-based feature selection.

Comparative Analysis

Table 11 used the Welsh t-test to compare the outperformed classifiers: RFE-RF, penalized logistic regression, and RFE-SVM based on the F1 score. The results show that RFE-RF outperforms RFE-SVM and Penalized Logistic regression, with no significant difference for RFE-SVM.

Table 12 compares the results from the German credit dataset and shows that training models without feature selection appear suitable for ensemble model such as AdaBoost or Gradient boosting but not for decision trees. The two-stage classification model's [4] main results improve accuracy for some classifiers but not others. Arora and Kaur [3] used the Bolasso as a screening stage and observed an improvement in accuracy for RF; however, no improvement for KNNs. On the other hand, Zhou et al. [58] observed no improvement in logistic regression, CART, and NN accuracy in the Lasso screening case and a decrease in the MARS screening case compared to baseline; however, they observed a clear improvement in the SVM classifier accuracy for the two screening type. For Jadhav et al. [28], the two-stage FS provides an improvement in accuracy for SVM but no improvement for NB and a decrease for KNN. It appears that one-stage FS (RFE-SVM and RFE-RF) produces roughly the same results in terms of accuracy as two-stage FS (IG then GA+KNN/NB) and two-stage classification model (Bolasso

Table 6 SVM-based Feature selection performances

Model	Criterion				
	Stability	Time [s]	Recall	Precision	F1-score
Penalized-SVM	0.52	690	$t = -0.9832$	$t = 2.4037$	$t = 0.9348$
RFE-SVM	0.71	348	df = 17.94	df = 16.7	df = 12.45
			sig = 0.3386	sig = 0.02	sig = 0.367

Table 7 Comparative results of Logistic regression-based feature selection

	Stepwise logreg			Penalized-logreg		
Model	Recall	Precision	F1-score	Recall	Precision	F1-score
1	0.7490	0.8447	0.7940	0.9502	0.7234	0.8215
2	0.8052	0.8732	0.8378	0.9323	0.7394	0.8248
3	0.7875	0.9130	0.8456	0.9346	0.7722	0.8457
4	0.7869	0.8807	0.8312	0.9205	0.7725	0.8401
5	0.8149	0.8851	0.8486	0.9420	0.7169	0.8142
6	0.7222	0.8792	0.7930	0.9598	0.7465	0.8398
7	0.7939	0.8809	0.8352	0.9727	0.7304	0.8343
8	0.7637	0.8744	0.8153	0.9739	0.6849	0.8043
9	0.7975	0.8772	0.8355	0.9409	0.7262	0.8197
10	0.8281	0.8472	0.8375	0.9300	0.7686	0.8416
Mean	0.7836	0.8751	0.8269	0.9453	0.7371	0.8283
Low 95% CI	0.7610	0.8616	0.8126	0.9326	0.7174	0.8186
High 95% CI	0.8077	0.8891	0.8417	0.9584	0.7579	0.8383

Table 8 Logistic regression-based feature selection performances

Model	Criterion				
	Stability	Time [s]	Recall	Precision	F1-score
Penalized-logreg	0.504	3	$t = 13.889$	$t = 12.804$	$t = 0.211$
Stepwise logreg	0.486	3	df = 14.302	df = 15.95	df = 16.04
			sig< 0.0001	sig< 0.0001	sig=0.8355

Table 9 Comparative results of Random Forests-based feature selection

	RFE-RF			NRFE-RF		
Model	Recall	Precision	F1-score	Recall	Precision	F1-score
1	0.9005	0.8190	0.8578	0.8150	0.7617	0.7874
2	0.9054	0.8340	0.8683	0.8486	0.7773	0.8114
3	0.8679	0.7830	0.8233	0.8856	0.7876	0.8337
4	0.9420	0.7677	0.8460	0.9078	0.7602	0.8274
5	0.8607	0.8047	0.8317	0.9447	0.7455	0.8333
6	0.9245	0.7903	0.8522	0.9196	0.7409	0.8206
7	0.9040	0.8100	0.8544	0.8178	0.8065	0.8121
8	0.9541	0.8189	0.8814	0.9551	0.7203	0.8213
9	0.9174	0.8511	0.8830	0.8848	0.7619	0.8188
10	0.8934	0.7243	0.8000	0.9151	0.8017	0.8546
Mean	0.9061	0.7987	0.8490	0.8869	0.7654	0.8217
Low 95% CI	0.8855	0.7728	0.8306	0.8525	0.7463	0.8092
High 95% CI	0.9276	0.8264	0.8683	0.9241	0.7855	0.8345

then NB or SVM) and (Lasso then CART/Logreg/NN), but more research is needed.

Conclusion and Future Works

The filtering strategy has prevailed in machine learning and credit risk classification, causing significant losses for banks because of the selection bias it introduces; as an alternative, several authors proposed integrating the FS during the learning phase. We have thus attempted to compare three classifiers: logistic regression, random forests, and SVMs that integrate the FS within the modeling, each using two strategies: penalized/stepwise logistic regression, RFE/NRFE RF, and penalized/RFE SVM. Three criteria were used to compare: classification-based measures, stability of selection, and calculation time. The results show that the RFE-RF outperforms in classification with an F1 score of 85% and reasonable stability of 64% but has a drawback in execution time. The other algorithms have

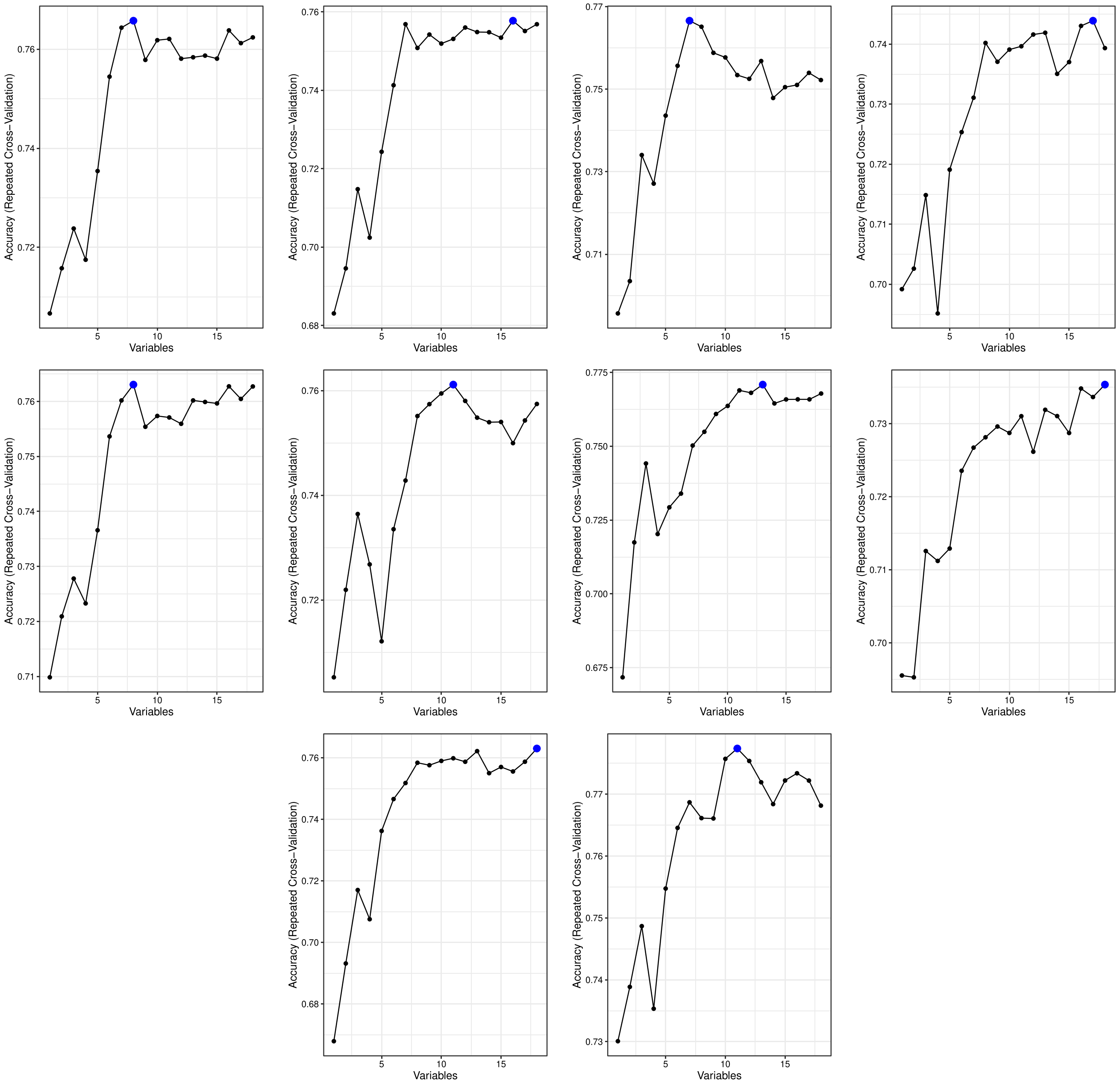


Fig. 7 The feature selection subsets of RFE-RF for various training sets

Table 10 Random Forests-based Feature selection performances

Model	Criterion				
	Stability	Time [s]	Recall	Precision	F1-score
RFE-RF	0.642	813	$t = -0.9723$	$t = 2.403$	$t = 2.8$
NRFE-RF	0.503	700	df = 14	df = 16.7	df = 15.81
			sig= 0.346	sig= 0.02	sig = 0.01

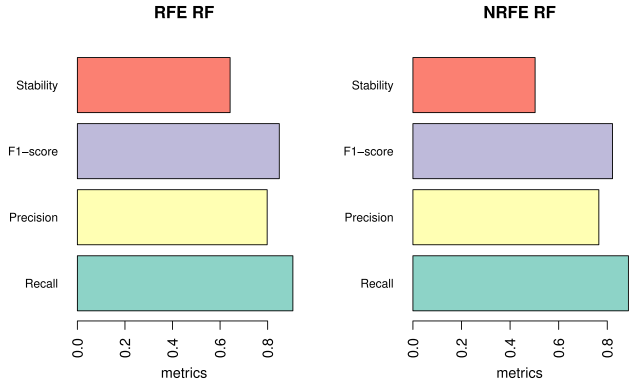


Fig. 8 Random forests-based feature selection comparative metrics

an F1 score of roughly 83% with mediocre stability, except for the RFE-SVM, having the best stability of 71%. Even though its execution time is significantly lower than the RFE-RF, it remains significant and can be disabling if the dimension becomes very large. A comparison with related literature revealed that RFE-RF and RFE-SVM produce roughly the same accuracy as two-stage Feature selection and, in some cases, outperform. However, because the comparison tools are not identical, the study will need to be expanded in the future. Another issue that appears to be essential to us is developing a performance evaluating consensus of credit risk classifiers.

Table 11 Comparative analysis based on F1 score

Classifier 1	Classifier 2		
	RFE-RF	RFE-SVM	Penalized logreg

RFE-RF	–	$t = -1.1932$
		df = 17.98
		sig = 0.2483
			$t = -2.168$
			df = 16.88
			sig = 0.04

RFE-SVM	–	–	$t = -0.794$
			df = 16.64
			sig = 0.438

Penalized Logreg	–	–	–

Table 12 Comparative performance analysis using German credit dataset

Classifier	Accuracy	FS	References
RF	0.771	RFE-RF	[this work]
SVM	0.755	RFE-SVM	[this work]
Logreg	0.73	Penalized Logreg	[this work]
C5.0	0.467	WFS	[31]
CART	0.591	WFS	[31]
AdaBoost	0.743	WFS	[35]
mg-GBDT	0.771	WFS	[35]
RF	0.84	Bolasso	[3]
KNN	0.748	Bolasso	[3]
SVM	0.76	Bolasso	[3]
NB	0.76	Bolasso	[3]
GA+SVM	0.8	WFS	[28]
GA+SVM	0.82	IG	[28]
GA+KNN	0.758	WFS	[28]
GA+KNN	0.768	IG	[28]
GA+NB	0.768	WFS	[28]
GA+NB	0.77	IG	[28]
GA+NN	0.789	WFS	[40]
HGA+NN	0.785	Gain ratio Gini index Correlation Earlier experience	[40]
Logreg	0.7603	WFS	[58]
Logreg	0.7347	MARS	[58]
Logreg	0.7689	Lasso	[58]
CART	0.7796	WFS	[58]
CART	0.7753	MARS	[58]
CART	0.7767	Lasso	[58]
NN	0.7380	WFS	[58]
NN	0.7332	MARS	[58]
NN	0.7532	Lasso	[58]
SVM	0.7700	WFS	[58]
SVM	0.8417	MARS	[58]
SVM	0.9979	Lasso	[58]

Funding Not applicable.

Availability of Data and Materials Data used in the practical experiments "Practical Experiments" section is available in the scorecard package, and all R packages referenced are freely available online. The code used in the data preprocessing "Data Pre-processing" section to merge classes will be provided upon request. The seeds used to split data were generated randomly.

Code Availability Not applicable.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Abdulrauf Sharifai G, Zainol Z. Feature selection for high-dimensional and imbalanced biomedical data based on robust correlation based redundancy and binary grasshopper optimization algorithm. Genes. 2020;11(7):717.
2. Amini MR. Principes de base en apprentissage supervisé. Eyrolles. Machine Learning. 2020; hal-03049016.
3. Arora N, Kaur PD. A Bolasso based consistent feature selection enabled random forest classification algorithm: an application to credit risk assessment. Appl Soft Comput. 2020;86: 105936.
4. Atif D, Salmi M. Feature selection for credit risk classification. In: Bennour A, Ensari T, Kessentini Y, Eom S, editors. Intelligent systems and pattern recognition. ISPR 2022. Communications in computer and Information science, vol. 1589. Cham: Springer; 2022. https://doi.org/10.1007/978-3-031-08277-1_14.
5. Bach FR. Bolasso: model consistent lasso estimation through the bootstrap. In: Proceedings of the 25th international conference on Machine learning; 2008. p. 33-40.
6. Becker N, Werft W, Toedt G, Lichter P, Benner A. penalizedSVM: a R-package for feature selection SVM classification. Bioinformatics. 2009;25(13):1711–2.
7. Breiman L. Bagging predictors. Mach Learn. 1996;24(2):123–40.
8. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.

9. Çetiner E, Koçak T, Güngör VÇ. Credit risk analysis based on hybrid classification: case studies on German and Turkish credit datasets. In: 2018 26th signal processing and communications applications conference (SIU). IEEE; 2018. p. 1–4.
10. Chaurasia V, Pal S. Stacking-based ensemble framework and feature selection technique for the detection of breast cancer. SN Comput Sci. 2021;2(2):1–13.
11. Dahiya S, Handa SS, Singh NP. A feature selection enabled hybrid-bagging algorithm for credit risk evaluation. Expert Syst. 2017;34(6): e12217.
12. Das P, Roychowdhury A, Das S, Roychoudhury S, Tripathy S. sigFeature: novel significant feature selection method for classification of gene expression data using support vector machine and t statistic. Front Genetics. 2020;11:247.
13. Fan S, Shen Y, Peng S. Improved ML-based technique for credit card scoring in internet financial risk control. Complexity. 2020;2020:8706285.
14. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1.
15. Fung GM, Mangasarian OL. A feature selection Newton method for support vector machine classification. Comput Optimiz Appl. 2004;28(2):185–202.
16. Genuer R, Poggi JM, Tuleau-Malot C. Variable selection using random forests. Pattern Recogn Lett. 2010;31(14):2225–36.
17. Genuer R, Poggi JM, Tuleau-Malot C. Vsurf: an r package for variable selection using random forests. R J. 2015;7(2):19–33.
18. Granitto PM, Furlanello C, Biasioli F, Gasperi F. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemom Intell Lab Syst. 2006;83(2):83–90.
19. Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests. Stat Comput. 2017;27(3):659–78.
20. Guan L, Qiao L, Li D, Sun T, Ge K, Lu X. An efficient ADMM-based algorithm to nonconvex penalized support vector machines. In: 2018 IEEE international conference on data mining workshops (ICDMW). IEEE; 2018. p. 1209–16.
21. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46(1):389–422.
22. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157–82.
23. Ha VS, Nguyen HN. Credit scoring with a feature selection approach based deep learning. In: MATEC web of conferences, vol. 54. EDP Sciences; 2016. p. 05004
24. Harrell FE. Regression modeling strategies: with applications to linear models. logistic and ordinal regression, and survival analysis, vol. 3. New York: Springer; 2015.
25. Hastie T, Tibshirani R, Tibshirani R. Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons. Stat Sci. 2020;35(4):579–92.
26. Huang ML, Hung YH, Lee WM, Li RK, Jiang BR. SVM-RFE based feature selection and Taguchi parameters optimization for multiclass SVM classifier. Sci World J. 2014;2014: 795624.
27. Huang Y, Montoya A. Lack of robustness of lasso and group lasso with categorical predictors: impact of coding strategy on variable selection and prediction. arXiv:40b200z6 [Preprint]. 2020. Available from: arXiv:40b200z6
28. Jadhav S, He H, Jenkins K. Information gain directed genetic algorithm wrapper feature selection for credit rating. Appl Soft Comput. 2018;69:541–53.
29. Jiang H, Deng Y, Chen HS, Tao L, Sha Q, Chen J, et al. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. BMC Bioinform. 2004;5(1):1–12.
30. Jović A, Brkić K, Bogunović N. A review of feature selection methods with applications. In: 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO). IEEE; 2015. p. 1200–05.
31. Khoraskani MM, Kheradmand F, Khamseh AA. Application and comparison of neural network, C5. 0, and classification and regression trees algorithms in the credit risk evaluation problem (case study: a standard German credit dataset). Int J Knowl Eng Data Min. 2017;4(3–4):259–76.
32. Kruppa J, Schwarz A, Arminger G, Ziegler A. Consumer credit risk: individual probability estimates using machine learning. Exp Syst Appl. 2013;40(13):5125–31.
33. Lappas PZ, Yannacopoulos AN. A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment. Appl Soft Comput. 2021;107: 107391.
34. Lessmann S, Baesens B, Seow HV, Thomas LC. Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research. Eur J Oper Res. 2015;247(1):124–36.
35. Liu W, Fan H, Xia M. Step-wise multi-grained augmented gradient boosting decision trees for credit scoring. Eng Appl Artif Intell. 2021;97: 104036.
36. Mariammal G, Suruliandi A, Raja SP, Poongothai E. Prediction of land suitability for crop cultivation based on soil and environmental characteristics using modified recursive feature elimination technique with various classifiers. IEEE Trans Comput Soc Syst. 2021;8(5):1132–42.
37. Molina LC, Belanche L, Nebot A. Feature selection algorithms: a survey and experimental evaluation. In: 2002 IEEE international conference on data mining. Proceedings. IEEE; 2002. p. 306–13.
38. Mustaqeem A, Anwar SM, Majid M, Khan AR. Wrapper method for feature selection to classify cardiac arrhythmia. In: 2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE; 2017. p. 3656–59.
39. Nazih W, Hifny Y, Elkilani W, Abdelkader T, Faheem H. Efficient detection of attacks in SIP based VoIP networks using linear L1-SVM classifier. Int J Comput Commun Control. 2019;14(4):518–29.
40. Oreski S, Oreski G. Genetic algorithm-based heuristic for feature selection in credit risk assessment. Exp Syst Appl. 2014;41(4):2052–64.
41. Pandey TN, Jagadev AK, Mohapatra SK, Dehuri S. Credit risk analysis using machine learning classifiers. In: 2017 International conference on energy, communication, data analytics and soft computing (ICECDS). IEEE; 2017. p. 1850–4.
42. Reeves DM, Jacyna GM. Support vector machine regularization. Wiley Interdiscip Rev: Comput Stat. 2011;3(3):204–15.
43. Saeys Y, Abeel T, Peer YVD. Robust feature selection using ensemble feature selection techniques. In: Joint European conference on machine learning and knowledge discovery in databases. Berlin, Heidelberg: Springer. 2008. p. 313–25.
44. Sanz H, Valim C, Vegas E, Oller JM, Reverter F. SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. BMC Bioinform. 2018;19(1):1–18.
45. Shen F, Zhao X, Kou G, Alsaadi FE. A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique. Appl Soft Comput. 2021;98: 106852.
46. Smith G. Step away from stepwise. J Big Data. 2020;5(1):1–12.
47. Somol P, Baesens B, Pudil P, Vanthienen J. Filter-versus wrapper-based feature selection for credit scoring. Int J Intell Syst. 2005;20(10):985–99.
48. Svetnik V, Liaw A, Tong C, Wang T. Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. In: International workshop on multiple Classifier systems. Berlin, Heidelberg: Springer; 2004. p. 334–43.
49. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc: Ser B (Methodol). 1996;58(1):267–88.

50. Tripathi D, Edla DR, Cheruku R, Kuppili V. A novel hybrid credit scoring model based on ensemble feature selection and multilayer ensemble classification. Comput Intell. 2019;35(2):371–94.
51. Vapnik V. The nature of statistical learning theory. Springer science & business media; 1995.
52. Verikas A, Kalsyte Z, Bacauskiene M, Gelzinis A. Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey. Soft Comput. 2010;14(9):995–1010.
53. Walter S, Tiemeier H. Variable selection: current practice in epidemiological studies. Eur J Epidemiol. 2009;24(12):733–6.
54. Wang L, Zhu J, Zou H. The doubly regularized support vector machine. Statistica Sinica. 2006;16(2):589–615.
55. Wang H, Xu Q, Zhou L. Large unbalanced credit scoring using lasso-logistic regression ensemble. PLoS ONE. 2018;10(2):e0117844.
56. Xia J, Sun L, Xu S, Xiang Q, Zhao J, Xiong W, et al. A model using support vector machines recursive feature elimination (SVM-RFE) algorithm to classify whether COPD patients have been continuously managed according to GOLD guidelines. Int J Chronic Obstr Pulm Dis. 2020;15:2779.
57. Xue Y, Zhang L, Wang B, Zhang Z, Li F. Nonlinear feature selection using Gaussian kernel SVM-RFE for fault diagnosis. Appl Intell. 2018;48(10):3306–31.
58. Zhou Y, Uddin MS, Habib T, Chi G, Yuan K. Feature selection in credit risk modeling: an international evidence. Econ Res-Ekonomska Istraživanja. 2021;34(1):3064–91.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.