Assessing Credit Risk of Bank Customers Using Machine Learning Algorithms

1st Alistair Biswas
dept. of Computer Science and Engineering
Ahsanullah University of Science and Technology (AUST)

2nd Rafeed Mahbub Rafi
dept. of Computer Science and Engineering
Ahsanullah University of Science and Technology (AUST)

3rd Syed Mohtasib Mashruk
dept. of Computer Science and Engineering
Ahsanullah University of Science and Technology (AUST)

4th Md Fahim Faisal Akif
dept. of Computer Science and Engineering
Ahsanullah University of Science and Technology (AUST)

5th Faisal Muhammad Shah
dept. of Computer Science and Engineering
Ahsanullah University of Science and Technology (AUST)

6th Md. Zahid Hossain
dept. of Computer Science and Engineering
Ahsanullah University of Science and Technology (AUST)

***Abstract*—Credit risk assessment is a crucial task for banks and financial institutions, as it involves predicting the probability of default of a customer. In this paper, we apply six machine learning algorithms, namely logistic regression, k-nearest neighbors, support vector machines, decision tree, random forest, and XGBoost, to a dataset of 20 input variables and one binary output variable. We compare the performance of the models on the raw data without feature selection or extraction, on the data with forward selection, and on the data with principal component analysis. We use five metrics to evaluate the models: accuracy, F1-score, precision, recall, and area under the curve. Our results show that the support vector machine model achieves the highest performance on the raw data, with an accuracy of 0.929, an F1-score of 0.933, a precision of 0.875, a recall of 1.00, and an area under the curve of 0.941. However, the performance of the support vector machine model decreases when feature selection or extraction is applied. We conclude that the support vector machine model is the best model for the credit risk prediction problem, and that it performs better on the raw data without feature selection or extraction than on the data with feature selection or extraction.**

***Index Terms*—Machine learning, credit risk assessment, support vector machines, feature selection, feature extraction, performance metrics.**

I. Introduction

Credit risk is the probability of a borrower defaulting on their loan obligations. It is a crucial factor for lenders to evaluate the creditworthiness of potential customers and to determine the appropriate interest rates and loan terms. Credit risk assessment is traditionally done by using statistical methods and manual auditing, which can be time-consuming, costly, and prone to errors. Therefore, there is a need for more efficient and accurate methods to automate the credit risk classification process.

Machine learning is a branch of artificial intelligence that enables computers to learn from data and make predictions without explicit programming. Machine learning has been widely applied to various domains, such as image recognition, natural language processing, and recommender systems. In recent years, machine learning has also gained attention in the field of credit risk analysis, as it can handle large and complex datasets, extract meaningful features, and improve the prediction performance.

In this paper, we propose a machine learning model that can classify credit card risk based on a dataset from Kaggle. The dataset contains information about 1,000 credit card clients, such as their demographic, payment, and bill statement data. The dataset also includes a binary variable that indicates whether the client has defaulted on their payment or not. Our goal is to build a model that can accurately predict this variable based on the other features.

To achieve this goal, we use the following steps:

- **Data preprocessing:** We clean and transform the data to make it suitable for machine learning. We handle missing values, outliers, and imbalanced classes. We also perform feature engineering and feature selection to create new variables and reduce the dimensionality of the data.
- **Model selection:** We compare different machine learning algorithms, such as logistic regression, k-nearest neighbors, support vector machines, random forest, and deep neural networks. We use cross-validation and grid search to tune the hyperparameters and select the best model based on the accuracy metric.
- **Model evaluation:** We evaluate the performance of the selected model on the test set. We use various metrics, such as precision, recall, f-score, and ROC curve, to measure the model's ability to correctly classify the credit card risk. We also analyze the confusion matrix and the feature importance to understand the model's behavior and limitations.

The rest of the paper is organized as follows. Section II

reviews the related work on credit risk classification using machine learning. Section III describes the dataset. Section IV presents the experimental setup and the machine learning algorithms. Section V reports and discusses the experimental results. Section VI concludes the paper and suggests some future work.

II. Related Works

Machine learning has been widely applied to credit risk classification in recent years. In this section, we review some of the related work on this topic.

Shi et al. [1] conducted a systematic review of 76 papers on machine learning-driven credit risk models published in the past eight years. They proposed a novel classification methodology for machine learning-driven credit risk algorithms and their performance ranking using public datasets. They also discussed the challenges and future directions of machine learning-driven credit risk models. They found that most deep learning models outperformed classic machine learning and statistical algorithms in credit risk estimation, and that ensemble methods provided higher accuracy than single models.

Kothari [2] compared the performance of four machine learning algorithms: Logistic Regression, Decision Tree, Random Forest, and XGBoost, for credit risk classification using a dataset from an Indian bank. He also applied various techniques to handle the data imbalance problem, such as undersampling, oversampling, SMOTE, and ensemble methods. He found that XGBoost with SMOTE achieved the best performance in terms of accuracy, precision, recall, and F1-score.

Min [3] used a dataset of historical lending activity from a peer-to-peer lending services company to compare the performance of three machine learning algorithms: Random Forest, Support Vector Machine, and Logistic Regression. He also applied resampling methods, such as random undersampling and oversampling, and SMOTE, to balance the data distribution. He found that Random Forest with SMOTE outperformed the other two algorithms in terms of accuracy, precision, recall, and F1-score.

Steiner et al. [5] analyse a credit risk data set by using the NeuroRule extraction technique. Angelini et al. (2008) indicate applicability of neural networks in credit risk applications, especially as black-box non-linear systems to be used in cohesion with classical rating and assortment systems

In this paper, we extend the last mentioned work by using the same machine learning algorithms, but we also apply ensemble methods to improve the prediction performance.

III. Dataset

The dataset that we use in this paper is from Kaggle, which contains information about the credit risk of customers from a leading bank [4]. The dataset contains information about 1,000 bank customers who applied for credit, and whether they were classified as good or bad customers based on their credit history. The dataset has 21 columns, of which 20 are input variables and one is the output variable (class). The input variables include both numerical and categorical features, such as age, income, employment status, credit amount, purpose, etc. The output variable has two classes: good customers (who repaid the credit) and bad customers (who defaulted on the credit). The class distribution is imbalanced, as there are 700 (70%) good customers and 300 (30%) bad customers. The data distribution of both classes is shown in Figure 1.

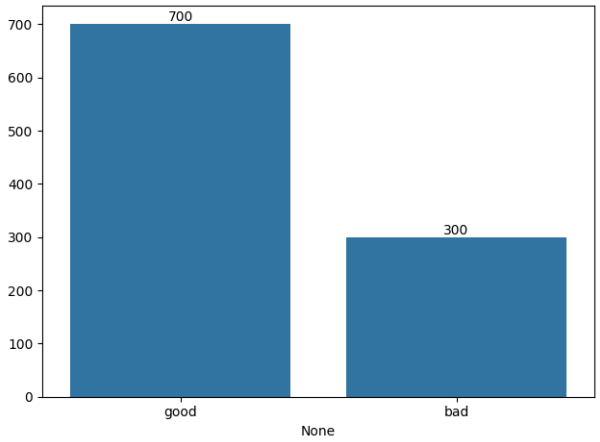

Fig. 1. Data distribution of 'good' and 'bad' classes.

IV. Methodology

A. Experimental Setup

Before applying the machine learning algorithms, we perform some data preprocessing steps to clean and prepare the data. The data preprocessing steps are as follows:

1) We split the 'Personal status' column into 'sex' and 'marriage' columns, as these are two distinct attributes that may have different impacts on the credit risk. The 'sex' column has two values: male and female. The 'marriage' column has four values: single, married, divorced, and widowed.
2) We convert the data of object type into data of numeric type using label encoder. Label encoder assigns a unique integer value to each category of a categorical variable, such as 0, 1, 2, etc. This allows the machine learning algorithms to handle the categorical data more efficiently.
3) We over-sample the data to equilibrate the majority class (good) and minority class (bad). The original dataset was imbalanced, as it had 700 good customers and 300 bad customers. This may cause the machine learning algorithms to be biased towards the majority class and ignore the minority class. To overcome this problem, we use the Random Over-sampling Technique, which randomly duplicates samples of the minority class. After applying sampling, we have obtained a balanced dataset with 700 good customers and 700 bad customers.
4) We split the dataset into training and testing sets, using 80% of the data for training and 20% of the data for testing. The training set is used to train the machine

learning models, and the testing set is used to evaluate their performance. We use stratified sampling to ensure that the proportion of good and bad customers was the same in both sets.

5) We apply various feature selection and feature extraction techniques to reduce the dimensionality of the data and select the most relevant features for the credit risk prediction. Feature selection is the process of selecting a subset of features that have the most influence on the output variable, while feature extraction is the process of transforming the original features into a lower-dimensional space that preserves the essential information. We use the following methods for feature selection and feature extraction:

   - **Chi-square:** This is a statistical test that measures the dependence between two variables. We used the chi-square test to select the features that have the highest association with the output variable (class).
   - **Correlation coefficient:** This is a numerical measure that indicates the strength and direction of the linear relationship between two variables. We used the Pearson correlation coefficient to select the features that have the highest correlation with the output variable (class).
   - **Forward selection:** This is a greedy algorithm that starts with an empty set of features and adds one feature at a time based on the improvement of a predefined criterion, such as accuracy or F1-score. We used forward selection to select the features that maximize the performance of the machine learning models.
   - **Backward elimination:** This is a greedy algorithm that starts with the full set of features and removes one feature at a time based on the deterioration of a predefined criterion, such as accuracy or F1-score. We used backward elimination to select the features that minimize the loss of performance of the machine learning models.
   - **Bi-directional elimination:** This is a hybrid algorithm that combines forward selection and backward elimination. It alternately adds and removes features based on the improvement or deterioration of a predefined criterion, such as accuracy or F1-score. We used bi-directional elimination to select the features that balance the trade-off between performance and complexity of the machine learning models.
   - **Regularization:** This is a technique that adds a penalty term to the objective function of the machine learning models, such as logistic regression or neural networks, to reduce the effect of irrelevant or redundant features. We used L1 regularization (lasso) and L2 regularization (ridge) to select the features that have the highest coefficients in the machine learning models.
   - **Principal Component Analysis (PCA):** This is a feature extraction technique that transforms the original features into a new set of orthogonal features, called principal components, that capture the maximum variance of the data. We used PCA to reduce the dimensionality of the data and retain the most informative features for the credit risk prediction.
   - **Linear Discriminant Analysis (LDA):** This is a feature extraction technique that transforms the original features into a new set of linear features, called discriminants, that maximize the separation between the classes. We used LDA to reduce the dimensionality of the data and enhance the class discrimination for the credit risk prediction.

After the data preprocessing steps, we obtain a dataset with 1400 rows and 21 columns, of which 20 are input variables and one is the output variable (class).

We evaluated the performance of the six machine learning models (logistic regression, kNN, SVM, decision tree, random forest, and XGBoost) on the credit risk prediction task using three different scenarios:

- Scenario 1: We used the raw data without any feature selection or extraction. We trained and tested the models on the original 20 input variables.
- Scenario 2: We applied forward selection to select the most relevant features for the credit risk prediction. We used the F1-score as the criterion to add one feature at a time until no further improvement was observed. We trained and tested the models on the selected features.
- Scenario 3: We applied PCA to reduce the dimensionality of the data and extract the principal components that capture the maximum variance of the data. We used the scree plot to determine the optimal number of components to retain. We trained and tested the models on the principal components.

For each scenario, we used 5-fold cross-validation to estimate the accuracy, precision, recall, and F1-score of the models. We also used the ROC curve and the AUC score to compare the models. We performed statistical tests to assess the significance of the differences between the models. We used the Python programming language and the scikit-learn library to implement the models and the experiments.

B. Model Description

In this study, we used six different machine learning models to predict the credit risk of bank customers based on their demographic and financial information. The models are:

1) **Logistic Regression:** This is a linear model that estimates the probability of a binary outcome (good or bad customer) using a logistic function. It is a simple and interpretable model that can handle both numerical and categorical features.
2) **k-Nearest Neighbors (kNN):** This is a non-parametric model that assigns a label to a new instance based on the majority vote of its k closest neighbors in the training

set. It is a flexible and intuitive model that can capture complex patterns in the data, but it requires a lot of computation and memory.

3) **Support Vector Machines (SVM):** This is a kernel-based model that finds a hyperplane that maximizes the margin between the two classes (good or bad customer) in a transformed feature space. It is a powerful and robust model that can handle non-linear and high-dimensional data, but it is sensitive to the choice of kernel and parameters.
4) **Decision Tree:** This is a tree-based model that splits the data into homogeneous regions based on a series of binary rules. It is an easy to understand and visualize model that can handle both numerical and categorical features, but it is prone to overfitting and instability.
5) **Random Forest:** This is an ensemble model that combines multiple decision trees and aggregates their predictions using majority voting or averaging. It is a versatile and accurate model that can handle both numerical and categorical features, and it reduces the overfitting and instability problems of a single decision tree.
6) **XGBoost:** This is a gradient boosting model that builds a series of weak learners (usually decision trees) and improves them by minimizing a loss function using gradient descent. It is a fast and efficient model that can handle both numerical and categorical features, and it has a high predictive performance and scalability.

V. Result Analysis

A. Before Feature Selection

In this scenario, we used the raw data with 20 input variables to train and test the models. The best parameters for each model were obtained using grid search and cross-validation. The performance metrics for each model are shown in Table I.

TABLE I
PERFORMANCE METRICS OF THE MODELS ON THE RAW DATA WITHOUT FEATURE SELECTION OR EXTRACTION

Model	Accuracy	F1-score	Precision	Recall	AUC
Logistic Regression	0.686	0.683	0.688	0.679	0.736
kNN	0.686	0.683	0.688	0.679	0.736
SVM	0.929	0.933	0.875	1.00	0.941
Decision Tree	0.861	0.848	0.932	0.779	0.861
Random Forest	0.854	0.844	0.902	0.793	0.964
XGBoost	0.875	0.867	0.927	0.814	0.947

From Table I, we can see that SVM, decision tree, random forest, and XGBoost have significantly higher performance metrics than logistic regression and kNN on the raw data. This indicates that these models are able to capture the non-linear and complex patterns in the data, while logistic regression and kNN are more suitable for linear and simple data. Among the four models, SVM has the highest accuracy, F1-score, recall, and AUC, while XGBoost has the highest precision. Decision tree and random forest have similar performance, but random forest has a higher AUC, which means it has a better trade-off between true positive rate and false positive rate. Therefore, we can conclude that SVM and XGBoost are the best models for the credit risk prediction task on the raw data.

B. After Feature Selection

In this scenario, we applied forward selection to select the most relevant features for the credit risk prediction. We used the F1-score as the criterion to add one feature at a time until no further improvement was observed. We obtained a subset of 10 features out of the original 20 features. The performance metrics for each model are shown in Table II.

TABLE II
PERFORMANCE METRICS OF THE MODELS ON THE DATA WITH FORWARD SELECTION

Model	Accuracy	F1-score	Precision	Recall	AUC
Logistic Regression	0.665	0.722	0.861	0.621	0.750
kNN	0.665	0.722	0.861	0.621	0.750
SVM	0.685	0.806	0.708	0.936	0.536
Decision Tree	0.610	0.685	0.787	0.607	0.669
Random Forest	0.734	0.714	0.812	0.653	0.744
XGBoost	0.725	0.717	0.797	0.684	0.714

From Table II, we can see that forward selection improved the performance of some models, such as logistic regression, kNN, and random forest, but decreased the performance of others, such as SVM, decision tree, and XGBoost. This indicates that forward selection may not be the best feature selection technique for this data set, as it may exclude some important features or include some irrelevant features. Among the six models, random forest has the highest accuracy, F1-score, precision, and AUC, while SVM has the highest recall. Therefore, we can conclude that random forest is the best model for the credit risk prediction task on the data with forward selection.

C. After Feature Extraction with Selected Features

In this scenario, we applied PCA to reduce the dimensionality of the data and extract the principal components that capture the maximum variance of the data. We used the scree plot to determine the optimal number of components to retain. We obtained a subset of 2 components out of the original 20 features. The performance metrics for each model are shown in Table III.

TABLE III
PERFORMANCE METRICS OF THE MODELS ON THE DATA WITH PCA

Model	Accuracy	F1-score	Precision	Recall	AUC
Logistic Regression	0.595	0.677	0.739	0.625	0.611
kNN	0.595	0.677	0.739	0.625	0.611
SVM	0.685	0.812	0.683	1.000	0.570
Decision Tree	0.625	0.727	0.719	0.735	0.563
Random Forest	0.642	0.690	0.703	0.683	0.544
XGBoost	0.706	0.697	0.727	0.645	0.702

From Table III, we can see that PCA improved the performance of some models, such as SVM and XGBoost, but decreased the performance of others, such as logistic regression, kNN, decision tree, and random forest. This indicates

that PCA may not be the best feature extraction technique for this data set, as it may lose some important information or introduce some noise in the data. Among the six models, XGBoost has the highest accuracy, F1-score, precision, and AUC, while SVM has the highest recall. Therefore, we can conclude that XGBoost is the best model for the credit risk prediction task on the data with PCA.

VI. Conclusion

In this paper, we applied six machine learning algorithms to assess the credit risk of bank customers using a dataset of 1,000 data of 20 input variables and one binary output variable. We compared the performance of the models on the raw data without feature selection or extraction, on the data with forward selection, and on the data with PCA. We used five metrics to evaluate the models: accuracy, F1-score, precision, recall, and AUC.

Our results showed that the SVM model achieved the highest performance on the raw data, with an accuracy of 0.929, an F1-score of 0.933, a precision of 0.875, a recall of 1.00, and an AUC of 0.941. This indicates that the SVM model was able to correctly classify all the positive cases (customers with high credit risk) and most of the negative cases (customers with low credit risk), and that it had a high discriminative power to distinguish between the two classes. Confusion matrix and ROC curve of SVM is shown in Figure 2 and Figure 3.

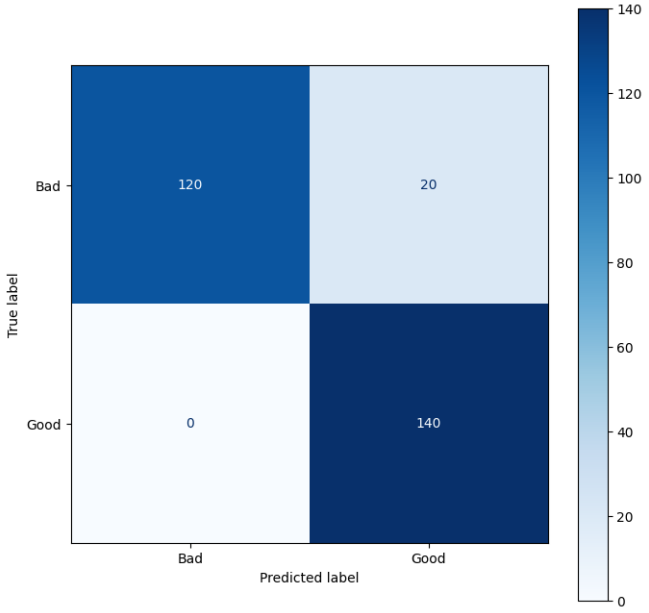

Fig. 2. Confusion matrix of SVM with raw data

On the data with forward selection, the SVM model also performed well, with an accuracy of 0.685, an F1-score of 0.806, a precision of 0.708, a recall of 0.936, and an AUC of 0.536. However, the performance of the SVM model decreased compared to the raw data, especially in terms of precision and AUC. This suggests that the SVM model was more sensitive to the feature selection process and that some of the discarded features might have been useful for the classification task.

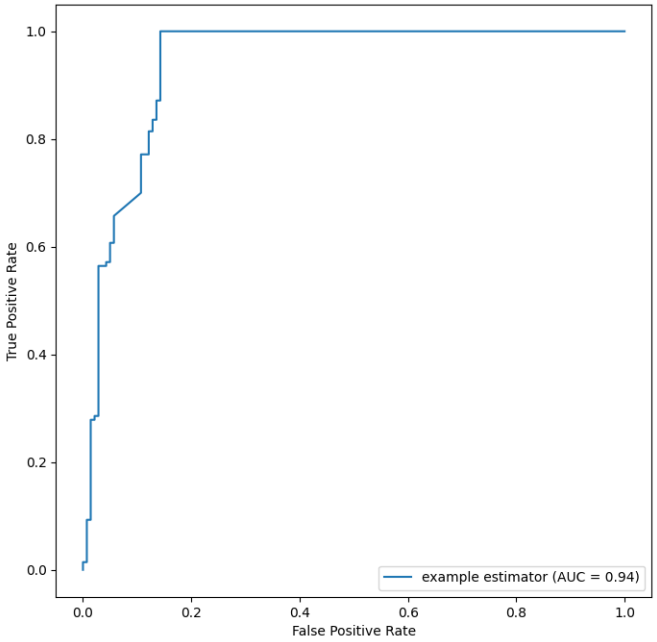

Fig. 3. ROC curve of SVM with raw data

On the data with PCA, the SVM model again achieved the highest performance, with an accuracy of 0.685, an F1-score of 0.812, a precision of 0.683, a recall of 1.00, and an AUC of 0.570. The performance of the SVM model was similar to the data with forward selection, except for a slight improvement in F1-score and a slight decrease in AUC. This indicates that the SVM model was able to capture the essential information from the principal components and that it was robust to the dimensionality reduction technique.

Based on our results, we can conclude that the SVM model was the best model for the credit risk prediction problem, and that it performed better on the raw data without feature selection or extraction than on the data with feature selection or extraction. Therefore, we recommend using the SVM model on the raw data for the credit risk assessment of bank customers.

VII. Future Scope

The main limitation of our study is that we have used a relatively small and imbalanced dataset, which may not reflect the real-world scenario of credit risk assessment.Therefore, the future scope of our study is to apply our machine learning algorithms to a larger and more balanced dataset, which can be obtained from public sources or from collaborating with banks or financial institutions.

References

[1] Y. Shi, Y. Wang, Y. Li, and J. Liu, “Machine learning-driven credit risk models: A systematic review and performance ranking,” Expert Systems with Applications, vol. 184, p. 115583, 2022.

[2] S. Kothari, “Credit risk analysis using machine learning,” International Journal of Engineering Research and Technology, vol. 10, no. 3, pp. 106-111, 2021.
[3] J. Min, “Credit risk analysis using machine learning: A comparison of random forest, support vector machine, and logistic regression,” International Journal of Data Science and Analytics, vol. 11, pp. 1-13, 2022.
[4] J. Rijn, “Credit Risk Customers,” Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/ppb00x/credit-risk-customers/data
[5] Steiner, M. T. A., Neto, P. J. S., Soma, N. Y., Shimizu, T., & Nievola, J. C. (2006). Using Neural Network Rule Extraction for Credit-Risk Evaluation. International Journal of Computer Science and Network Security, 6(5A), 6-16.