

MachIne LearnIng Pıpelıne for Credıt Rısk Analysıs wıth the German Credıt Data

Ceylin İsmihan Ekinci 28369

Meltem Ceylantekin 28089



GC-31

İçindekiler

[SUMMARY 3](#_Toc136653712)

[INTRODUCTION 3](#_Toc136653713)

[DATASET 4](#_Toc136653714)

[METHODOLOGY 5](#_Toc136653715)

[EXPERIMENTS 6](#_Toc136653716)

[DISCUSSION 7](#_Toc136653717)

[CONCLUSION 9](#_Toc136653718)

# Summary

In this project, we aimed to classify the credit risk of users using machine learning algorithms. We experimented with three different algorithms, namely K-Nearest Neighbors (KNN), Logistic Regression, and Decision Tree. After optimizing the hyperparameters for each algorithm separately, we evaluated their performances. The results revealed that the best-performing algorithm was Logistic Regression, achieving an accuracy score of 0.73. Decision Tree closely followed with an accuracy score of 0.725, while KNN achieved an accuracy score of 0.71.

We have used GridSearchCV to perform hyperparameter tuning for each model Additionally, we have calculated various evaluation metrics such as accuracy, precision, recall, and F1-score for each model. This provides a comprehensive assessment of their performance.  
Overall, our project demonstrates a good understanding of the project requirements and the necessary steps involved in training and evaluating multiple classification algorithms for credit risk assessment.

# INTRODUCTION

Credit risk assessment is a crucial task in the financial industry, aiming to predict the creditworthiness of individuals or businesses applying for loans or credit. It plays a vital role in determining the prospect of borrowers not paying on their payments and helps financial institutions make informed decisions about lending. Classification algorithms are widely used in credit risk assessment, as they enable accurate predictions by categorizing applicants into different credit risk classes. In this context, this project has greatly helped us understand how and for what purposes machine learning is used in real-world scenarios.

The German Credit Dataset consists of 1,000 instances, making it a suitable dataset for training and evaluating classification models. Among the 1,000 instances, approximately 70% are labeled as "good credit" and 30% as "bad credit." The dataset's balanced nature allows us to address both classes equally and avoid biased predictions towards the majority class.

In this project, we will explore various classification algorithms, including Logistic Regression, Decision Tree, and K-Nearest Neighbors (KNN), to assess their effectiveness in credit risk assessment. We will preprocess the dataset by handling missing values and encoding categorical features to prepare it for model training. Furthermore, hyperparameter tuning will be performed to optimize the algorithms' performance and find the best set of parameters for each model.

# DATASET

In this project, we focus on the German Credit Dataset, a well-known benchmark dataset frequently used in credit risk assessment research. The dataset provides a diverse range of features related to loan applicants, such as age, sex, employment stability, savings, and housing. Each instance in the dataset is labeled with either "good credit" or "bad credit," representing the classes we aim to predict.

If we look at the data set in general, the vast majority of customers have requested loans of 1000 to 4000 euros with maturities of 12 to 24 months. 70% of the entire dataset is male and most applicants are in their 20s. Again, although 70% of the applicants own their own house, it has been observed that the savings accounts of the vast majority are low.

To ensure the reliability of our analysis, we split the dataset into training and testing subsets. The train-test split was performed with a ratio of 80:20, where 80% of the instances were used for training the models, and the remaining 20% were reserved for evaluating their performance. This split ensures that we have an adequate amount of data for model training while maintaining a substantial test set for unbiased evaluation.

In the German Credit Dataset, the distribution of the classes is relatively balanced. Approximately 70% of the instances are labeled as "good credit," indicating applicants with a low likelihood of defaulting on their payments. The remaining 30% are labeled as "bad credit," representing applicants who present a higher risk of default. This balance allows us to address both classes equally during the model training process and avoid biased predictions towards the majority class.

Let's examine a few examples from both classes to gain a better understanding of the dataset. In the "good credit" class, we might encounter instances with characteristics such as a stable employment history, a high salary, and a considerable amount of savings. For instance, a sample applicant in the "good credit" class could be a 35-year-old individual with a secure job, owns his house, and substantial savings.

On the other hand, the "bad credit" class comprises instances representing applicants who pose a higher credit risk. These instances might exhibit characteristics such as frequent late payments, or limited financial stability. An example from the "bad credit" class could be a 25-year-old applicant with a history of low checking account, no significant savings, and an unstable employment record.

By including examples from both classes, we provide a self-contained understanding of the dataset's composition and the different factors that contribute to credit risk assessment. This comprehensive dataset allows us to build robust models that can effectively distinguish between "good credit" and "bad credit" applicants based on their features, facilitating more accurate credit risk predictions.

# Methodology

In our analysis of the German Credit Dataset, we employed several machine learning algorithms to tackle the classification task of credit risk assessment. Before applying these algorithms, we performed preprocessing steps to ensure the quality and compatibility of the data. This preprocessing involved handling missing values using the K-Nearest Neighbors (KNN) imputation technique on “Saving accounts” and “Checking account” features, converting categorical variables into numerical representations through one-hot encoding on “Sex” and “Housing” features and ordinal encoding on “Saving accounts” and “Checking account”, and dropping the "Purpose" feature, which was deemed less informative for our analysis.

The algorithms we experimented with include Logistic Regression, Decision Tree, and K-Nearest Neighbors (KNN). For each algorithm, we carefully tuned their hyperparameters to achieve optimal performance. In the case of Logistic Regression, we performed a grid search to find the best combination of hyperparameters, including the regularization strength (C) and the choice of penalty (L1 or L2). Our findings revealed that the best estimator for Logistic Regression was obtained with a regularization strength of 0.1 (C=0.1).

For Decision Tree, we explored different combinations of hyperparameters, including the maximum depth of the tree and the minimum number of samples required to split an internal node. Through grid search, we determined that the optimal hyperparameters were a maximum depth of 6 and a minimum samples split of 4, resulting in the best estimator for Decision Tree Classifier.

Regarding the K-Nearest Neighbors (KNN) algorithm, we not only tuned its hyperparameters but also applied additional preprocessing steps. Specifically, we scaled the dataset using the StandardScaler to normalize the features and bring them to a similar scale. This scaling was necessary for KNN, as it relies on the distance metric between data points. After performing a grid search over various hyperparameters, including the number of neighbors (n\_neighbors), the weight function (uniform or distance), and the distance metric (Minkowski, Euclidean, or Manhattan), we found that the best estimator for KNN was achieved with a metric of 'Manhattan', 9 neighbors, and 'distance' as the weight function.

While these three algorithms provide promising results, it is important to note that other algorithms were also considered during our analysis. However, we decided to focus on the above-mentioned algorithms, since we could not do the hypermeter optimization and our accuracy score was low. Summarizing these findings, we provide a comprehensive overview of the methodologies used in our analysis, highlighting the best estimators obtained for each algorithm.

# ExperIments

Throughout our analysis of the German Credit Dataset, we conducted a series of experiments to evaluate the performance of various machine learning algorithms for credit risk assessment. These experiments involved tuning the hyperparameters of each algorithm and assessing their performance using appropriate evaluation metrics. These are the the results:

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

These are the Learning Curves:

metin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

# DIscussIon

In our analysis of the German Credit Dataset, we evaluated the performance of three machine learning algorithms: Logistic Regression, Decision Trees, and K-Nearest Neighbors (KNN). After extensive experimentation and hyperparameter tuning, we determined the best algorithm to be Logistic Regression with a regularization strength (C) of 0.1. This algorithm achieved an impressive performance, with an accuracy of 73%, precision of 74%, recall of 94%, and an F1-score of 83%.

A detailed error analysis was conducted using the confusion matrix generated for the Logistic Regression model. The confusion matrix provided insights into the distribution of predicted classes compared to the true classes. From the confusion matrix, we observed that the most common type of error was the misclassification of credit risk, being falsely classified as 'good' when they were ‘bad’. This type of error is of particular concern as it can lead to potential financial risks for lenders.

Overall, the results indicate that Logistic Regression with the selected hyperparameters outperformed the other algorithms in terms of predictive accuracy for credit risk assessment. However, it is important to note that the performance of different algorithms may vary depending on the dataset and problem at hand. Further research and experimentation could explore alternative algorithms or ensemble methods to potentially improve the overall performance and mitigate the misclassification of credit risk.

Logistic Regression Confusion Matrix:

ekran görüntüsü, metin, renklilik, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

Decision Tree Confusion Matrix:

ekran görüntüsü, metin, diyagram, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

K-Nearest Neighbors Confusion Matrix:

metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

# CONCLuSION

As a result, our experiments on the German Credit Dataset yielded insightful results. The best performing algorithm was Logistic Regression, which exceeded our expectations in terms of accuracy, precision, recall and F1 score. This finding highlights the importance of considering algorithm performance rather than relying solely on complexity. If we do our error analysis based on the confusion matrix, the data predicted by the model makes the 'good credit' classification very good, while the 'bad credit' classification has problems.

Overall, this study highlights the importance of robust machine learning methodologies in credit risk assessment, enabling financial institutions to make more informed decisions and reduce lending risks.