# Applications of ML Algorithms to Assess Credit Risk – Experiments With Missing Data

Author: Giovanni Herbert
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands

**ABSTRACT,**
**Machine learning algorithms are increasingly more often used in credit risk predictions by banks and financial institutions. One of the biggest problems when working with machine learning algorithms is the quality of the data that is used by the algorithm. Therefore, this thesis looks at sparse and bad quality datasets and aims to find the best way to replace the missing data in these datasets. In the literature review, the relevance is presented as there are few, if any, studies researching this particular topic. Finding the best method of data replacement is done by comparing the methods of replacing by the mean, the median, and the mode of the variables with missing data, in addition, replacing by zero is also compared to these methods. These comparisons are done by replacing the missing data in the German credit dataset and implementing a Random Forest machine learning algorithm on these datasets. The comparisons are judged by comparing the feature importance of the algorithms and several accuracy metrics of said algorithms. The result of the experiment is that replacing by zero scores a combined first place, along with replacing by the mean of the available data in the variables, at the accuracy comparison and an absolute first place for the feature importance test. This means that replacing by zero is the preferred option for replacing the missing data in sparse and bad quality datasets when making consumer credit risk predictions using a machine learning algorithm.**

**Graduation Committee members:**
Dr. Machado, M. (Marcos)
Dr. Osterrieder, J.R.O. (Jörg)

**Keywords**
Random Forest, Machine Learning, Missing Data, Consumer Credit Risk Prediction, Accuracy, Feature Importance

# 1. INTRODUCTION

Credit risk refers to the risk that the commitments of the counterparty in a financial transaction are not met, this can either be due to inability or unwillingness. As credit risk is involved in every financial transaction, it is one of the biggest risks for financial institutions (Spuchľáková et al., 2015). This means that for banks and financial institutions, assessing their customers' credit risk is one of the most important steps of the due diligence during a loan application process.
In the last years, the outstanding credit for financial institutions has been increasing year after year[1], being higher than ever before (The Federal Reserve, 2023b). For example, during the fourth quarter of 2022, the total outstanding consumer credit in the USA was 4.79 trillion dollars and the outstanding credit from credit unions were 636.8 billion dollars (The Federal Reserve, 2023a).

Figure 1 shows the delinquency rate of credit cards in the United States of America between 2016 and 2022 in comparison to the delinquency rate for all the loans (including credit card delinquencies) during that period.

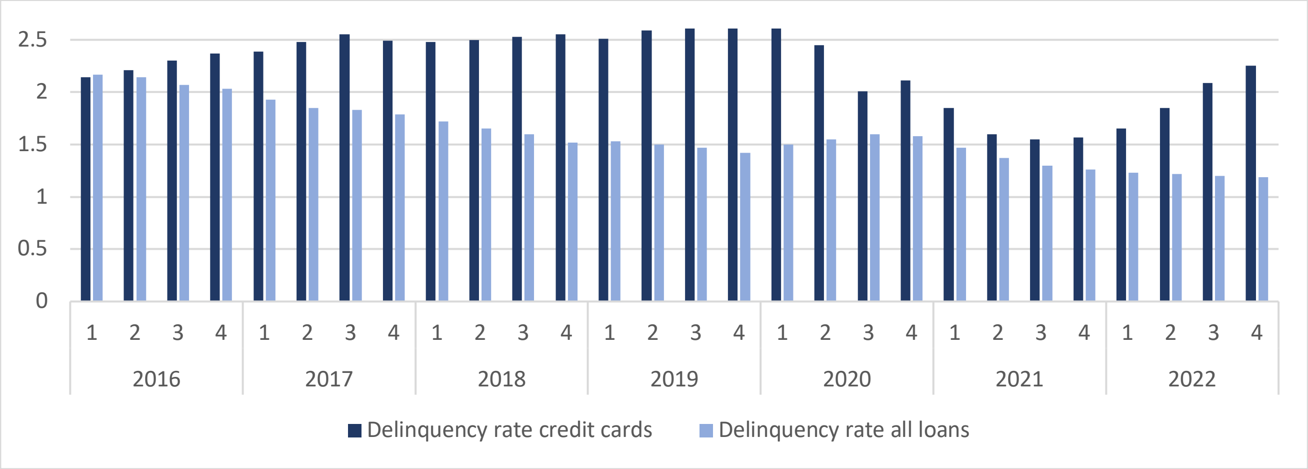

**Figure 1. Default rates in the USA between 2016-2022 per quartile (in %)**(FFIEC, 2023)

Figure 1 also shows that the delinquency rates for credit cards are, except for the first quarter of 2016, higher than the delinquency rates across all loans. On average during this period, the credit card delinquency rate was 2.25%, while for all loans, including credit cards, this was 1.60%. In addition, consumer credit all over the world has had a rapid increase in the last few decades (Rona-Tas & Guseva, 2018).

The delinquency rate on consumer credit means that during the fourth quarter of 2022, 14.3 billion dollars were lost due to defaults. With the higher rate of delinquencies in consumer credit, reliable and accurate models are important for financial institutions and banks to minimise losses and cases of fraud, as well as increase the efficiency of the consumer credit application process (Gupta et al., 2020). However, with the ever-rising consumer credit, this becomes increasingly more work for financial institutions, requiring more and more personnel to develop, implement, and analyse these risk assessment predictions.

Machine learning (ML) might be a solution to automate the consumer credit risk predictions. However, for machine learning to take over this task, it must be able to accurately and reliably predict whether a consumer is able to pay back the credit issued.

With the amount of outstanding credit, and the fact that machine learning models can help on reducing losses (Wei et al., 2023) and can help with increasing the efficiency and the speed of the application process (Galindo & Tamayo, 2000). Financial institutions have been implementing machine learning in their consumer credit risk predictions (Donepudi, 2017).

However, according to Clintworth et al., missing data is an important problem in financial modelling (Clintworth et al., 2023).

Therefore, there is room in the literature to determine the best way to pre-process sparse and bad-quality datasets, with the aim of cleaning up the missing data to be able to make accurate and reliable predictions for consumer credit risk using a machine learning algorithm. Thus, this thesis aims to compare some of the different ways to clean up a sparse and bad-quality dataset for consumer credit risk predictions. This exercise will compare the methods based on their accuracy and feature importance.
Thus, the main- and sub-research questions that this thesis aims to answer are:

**Main research question:**

- How should the missing data in sparse and bad quality datasets be replaced for implementation in a machine learning model for consumer credit risk predictions?

**Sub research questions:**

- What techniques can be used to pre-process sparse and bad quality datasets?
- What ways of analysis and comparison can be used to determine the best way to handle missing values and bad quality datasets?

# 2. LITERATURE RESEARCH

To understand the problem that this thesis will solve, it is important to have some knowledge about important topics and theories that will be covered by this research, these will be presented and subsequently explained in this chapter. The topics

[1] Except 2020 when there was a 0,3% decrease in total outstanding credit (The Federal Reserve, 2023b).

and theories include, but are not limited to, machine learning, reliability and explainability, and sparse and bad-quality datasets. In addition, an overview of similar studies will be presented in Table 1.

## 2.1 Machine learning

Machine learning acts in a similar way to how humans would experiment when learning and improving their analyses by using algorithms. Machine learning algorithms recognize patterns in large datasets and by this recognition, it learns and trains the algorithm to make decisions or predictions. This method tries to improve itself by judging its own prediction results against already known outcomes, and then adjusting until the accuracy of the predictions by the model is maximized (Helm et al., 2020).

There are multiple categories of machine learning algorithms (Ayodele, 2010), these include the following with some techniques given between paratheses:

- *Supervised learning*; this type creates a function that shows how outputs are calculated based on the inputs. This type of algorithm uses labels to categorise inputs. (Random Forest, logistic regression, and neural network (Aitha & Jathanna, 2019))
- *Unsupervised learning*; this type does not use labels, instead, it has to determine itself a way to predict an outcome from the inputs it has. (Deep learning and data clustering (Usama et al., 2019))
- *Semi-supervised learning*; this type is a combination of supervised and unsupervised learning as it uses both labelled and unlabelled inputs to predict an outcome. (Graph based method and manifold methods (van Engelen & Hoos, 2020))

Currently, machine learning is already being used in credit risk predictions (Shi et al., 2022). In addition, there are currently multiple research projects being conducted on credit risk machine learning models (Lestari et al., 2023; Niu et al., 2023; Rudin & Shaposhnik, 2023; Wei et al., 2023; Xia et al., 2023). Aitha et al. has found that for consumer credit risk predictions, the Random Forest approach resulted in the most accurate predictions.

Random Forest is a type of supervised machine learning. This type of algorithm makes its predictions based on creating and combining multiple decision trees to get accurate results (Aitha & Jathanna, 2019)

The different inputs for the Random Forest machine learning algorithm can have varying relevancies, some inputs might be crucial for the prediction, while other inputs have barely, if any, influence over the prediction that the algorithm makes. These inputs, however, do have an influence over the speed of the algorithm and can even lead to overfitting issues for the predictions. Input or feature selection reduces the number of inputs to prevent these issues and make the algorithm more efficient (Yu & Liu, 2004).

## 2.2 Reliability and explainability of machine learning models

To implement a machine learning algorithm in consumer credit risk predictions, it is important that the consumers trust that the predictions that the algorithm makes, are likely true. This can be proven by the financial institution by their claims of reliability of their machine learning algorithm. These claims have value for consumer trust because, on average, it is expected that the algorithm predicts with a similar reliability as during the validation process (Nicora et al., 2022) When a machine learning algorithm predicts the outcome of a consumer credit risk prediction, it does so by stating a certain probability that the consumer is able to pay back their requested loan, if that probability crosses a certain threshold, the algorithm can decide that the loan will be paid back or will not be paid back. However, if the uncertainty about the prediction is high, the algorithm should abstain from making a decision to avoid risk of delinquency (Nicora et al., 2022)

In addition to the reliability of the algorithm, the explainability is also critical in consumer credit risk predictions. Došilović et al. found that there are problems with the implications, both ethical and quality of life, of using artificial intelligence in real world scenarios in the current stage of the development of this technology (Došilović et al., 2018).

One specific problem that was mentioned by Došilović et al. was that machine learning models often are not clearly explaining why they came to a certain decision. When machine learning is used in practice for determining the credit risk and implications such as not issuing loans based on these predictions, the people who were not given loans might ask for the reason why they were not able to get a loan. This means that for the implementation of a machine learning model for consumer credit risk predictions, it is crucial to have the ability to clearly explain why certain decisions have been made.

## 2.3 Sparse and bad-quality datasets

Even the performance of reliable and explainable machine learning models is severely influenced by the quality of data that it uses. Bad quality datasets, such as datasets with missing data, can result in unreliability in the predictions that the algorithm make (Jain et al., 2020).

Clintworth et al. suggests three approaches to deal with missing data: (Clintworth et al., 2023)

1. *List-wise deletion*; an approach where incomplete observations are discarded.
2. *Omitted variable*; an approach where the covariates with missing values are removed from the dataset.
3. *Data imputation*; this approach replaces the missing values by new values according to a certain "rule", this can be the mean of the other values, the median of the other variables, or a completely different rule. This bachelor thesis will focus on this option.

Table 1 shows a number of studies conducted with machine learning to determine the consumer credit risk. The aim or target of the study is presented along with the (machine learning) techniques and analysis metrics that were used in each respective study. Among the studies in Table 1, a number of comparison studies are included that compare between the different kinds of machine learning algorithms that can be used for credit risk predictions, the techniques that were found to be the best are written in bold letters.

Table 1 shows that decision trees, and in particular Random Forest, are among the best techniques for credit risk purposes.

For the metrics of analysis, Accuracy is the most used metric, followed by Area Under the Curve and Precision.

In addition to this information, the last three columns show the literature review subsection that this study relates to most.

The last row of the Table 1 shows this thesis with the aim, machine learning model, and metrics that are used. This shows the theoretical relevance of this thesis as this thesis will build on existing knowledge to determine the best way to pre-process a sparse and bad quality dataset for consumer credit risk predictions. This is relevant as the other studies focus on discovering different types of knowledge in this subject area.

**Table 1. Literature review of past studies on machine learning for (consumer) credit risk predictions**

| Author/Study | Target | (Machine learning) techniques | Metrics | 2.1 | 2.2 | 2.3 |
|---|---|---|---|---|---|---|
| (Ribeiro et al., 2016) | Increasing the human trust in an algorithm by explaining individual predictions. | Linear models and Lime | Explainability to (non-) expert users | | ■ | |
| (Twala, 2010) | Comparing (combinations of) ML models on credit risk predictions to determine the best performing technique. | Artificial neural network, Naive Bayes classifier, and Decision trees | Accuracy (Excess error and Error rate) | ■ | | |
| (Brodley & Friedl, 1999) | Determining the best filter technique to identify mislabelled data. | Decision trees, K-Nearest Neighbour, and Linear Machine | Accuracy | | | ■ |
| (Veras et al., 2020) | Develop a sparse linear regression model that is more suited to work with incomplete datasets. | Forward Stagewise Regression (for incomplete datasets with Gaussian Mixture Modelling) | Average Mean Square Error and Euclidean Distance to the Baseline Model | | | ■ |
| (Shi et al., 2022) | Comparing ML models in credit risk assessment applications. | **Random Forest**, **Bagging**, Linear Regression, Artificial Neural Network, and others | Accuracy and Area Under the Curve | ■ | | |
| (Davis et al., 2022) | Providing ways to explain the output of ML models when forecasting home equity credit risk. | Random Forest, Neural Network, Inductive Logical Programming, and Optimal Classification Trees | Accuracy, Area Under the Curve, False Positive Rate, and False Negative Rate | | ■ | |
| (Nicora et al., 2022) | Present approaches to identify unreliable predictions. | Random Forest | Accuracy, Area Under the Curve, Precision Recall Curve, and Correlation Coefficient | | ■ | |
| (Aitha & Jathanna, 2019) | Comparing different methods of ML algorithms for credit risk assessment purposes. | Support Vector Network, Neural Network, Logistic Regression, Naive Bayes, **Random Forest**, and Classification and Regression Trees | F-Measure, Specificity, Accuracy, Sensitivity, Error-rate, and Precision | ■ | | |
| (Khandani et al., 2010) | Apply a ML algorithm to predict consumer credit risk | Classification and Regression Trees | F-Measure, Precision, Mean Absolute Error, and Root Mean Squared Error | ■ | | |
| (Gupta et al., 2020) | Apply a ML model to create a bank loan prediction system | Random Forest and Logistic Regression | Correlation between parameters | ■ | | |
| This thesis | Compare data pre-processing methods for consumer credit risk predictions | Random Forest | Accuracy metrics and Feature Importance | ■ | ■ | ■ |

# 3. METHODLOGY

## 3.1 Random Forest

In the literature review, the Random Forest machine learning algorithm was found to be the most accurate approach in the comparisons between the different kinds of machine learning models for credit risk predictions (Aitha & Jathanna, 2019; Shi et al., 2022). As the comparisons have found that this algorithm would be the most suitable for the application of this thesis, the Random Forest algorithm will be used for the experiment as will be described in the next chapter.

A Random Forest algorithm is known as a classification and regression method that can be applied to many different use cases and industries. This algorithm works by combining multiple randomised decision trees and creating a prediction by aggregating the results of those randomised decision trees (Biau & Scornet, 2016). Even though Random Forest algorithms are used on a widespread basis, the basic mathematical properties on which it works are not well understood. This lack of knowledge, in combination with the difficulty of analysing this algorithm, means that there is still a gap between the theoretical knowledge and the practical performance of these Random Forest models (Biau & Scornet, 2016).

## 3.2 Metrics of validation

In the literature review, the most used metric that was used to evaluate the results was prediction accuracy. Accuracy is a metric to assess the performance of a machine learning model, such as a Random Forest algorithm.

The metric calculates the accuracy of the model by dividing all the predictions where the algorithm predicted the correct outcome by the total number of predictions done (Dinga et al., 2019).

This calculation outputs a percentage of accuracy, which then could be used to compare to the other data clean-up methods to determine the most accurate way to clean up the data for consumer credit risk predictions. In addition to the “normal” accuracy measure, a balanced accuracy measure can be used. This balanced accuracy is calculated by averaging the sensitivity (percentage of true positives predictions) and specificity (percentage of true negative predictions) of the predictions made by the Random Forest algorithm (Dinga et al., 2019). Below, the formulas of the different metrics are given:

$$Accuracy = \frac{Correct\ predictions}{Total\ number\ of\ predictions}$$

$$Sensitivity = \frac{True\ postive}{Total\ positive}$$

$$Specificity = \frac{True\ negative}{Total\ negative}$$

$$Balanced\ accuracy = \frac{Sensitivity + Specificity}{2}$$

### 3.3 Feature importance

In addition to accuracy, another way to compare the different methods of data clean-up with each other is by seeing what variables or features the machine learning model will use to make its prediction. This is important as it can show the influence of the variables where the missing data is replaced. For example, if the variable has little to no influence on the final prediction, the data clean-up method might be less suitable than a clean-up method where those replaced variables have a high importance for the prediction.

The Random Forest algorithm can rank the importance of these variables via two methods: Mean Decrease Impurity (MDI) and Mean Decrease Accuracy (MDA) (Biau & Scornet, 2016). MDA is a method that links back to the accuracy measurement, it is based on the idea that if a variable is less important, it has a lower influence on the accuracy of the prediction (Biau & Scornet, 2016).

## 4. EXPERIMENT DESIGN

The aim of this thesis and therefore, this experiment, is to determine the best way to pre-process a dataset of sparse quality and with missing data. This dataset would then be used for consumer credit risk predictions.

The dataset that will be used for this experiment is the (shortened) German credit risk dataset (Hofmann, 1994). This dataset contains 1000 observations with 9 variables. In addition, a classification variable was included from the original dataset, as this was not included in the shortened dataset but is required for the fitting of the algorithm. During the fitting, this classification variable tells the algorithm whether the observation would have a high or low consumer credit risk.

In this dataset, some variables need to be encoded for the Random Forest machine learning model to be able to predict and classify whether the consumer has a high (bad) or a low (good) credit risk. These encodings are done using Ordinal Encoding, according to Provenzano et al. (2020), this is the most common method for encoding. Despite this, it does have a downside as non-ordinal variables will have non-existent ordinal relations assigned to them. The encodings of this dataset are displayed in Table 2.

**Table 2. Ordinal encoding applied to categorical variables**

| Variables | Object name | Numeric value |
|---|---|---|
| Sex | Male | 1 |
| | Female | 2 |
| Housing | Rent | 1 |
| | Free | 2 |
| | Own | 3 |
| Savings | Little | 1 |
| | Moderate | 2 |
| | Quite rich | 3 |
| | Rich | 4 |
| Checking | Little | 1 |
| | Moderate | 2 |
| | Rich | 3 |
| Purpose | Radio/TV | 1 |
| | Education | 2 |
| | Furniture/equipment | 3 |
| | Car | 4 |
| | Business | 5 |
| | Domestic appliances | 6 |
| | Repairs | 7 |
| | Vacation/others | 8 |
| Classification | Bad | 0 |
| | Good | 1 |

In addition to this encoding, the dataset will be pre-processed by replacing the missing datapoints with:

- the mean of the available datapoints in each respective variable;
- the median of the available datapoints in each respective variable;
- the mode of the available datapoints in each respective variable;
- zeros.

Replacing the missing values will result in four separate datasets, which will then be randomly separated in a training and a test dataset. This separation will happen on an 80/20 ratio, meaning that 80 per cent of the dataset will be in the training dataset, while 20 per cent will be in the test dataset. This split ratio will provide enough training data for the classification task of the Random Forest machine learning algorithm (Rácz et al., 2021).

After the algorithm has been fitted on each of the four individual training datasets, it will be run on their respective test datasets to determine the accuracy metrics, as discussed in the methodology chapter, and feature importance. The results will then be presented and the differences in results between the four methods of data pre-processing will be discussed. All the findings from this thesis will then be concluded.

Table 3 shows a descriptive overview of the dataset and Figure 2 shows an overview of the research framework for this thesis, with a step-by-step overview of the different steps of the experiment as described in this chapter.

**Table 3. Descriptive overview of the German credit dataset** (Hofmann, 1994)

| **Variables** | **Number of missing datapoints** | |
|---|---|---|
| Savings | 183 | |
| Checking | 409 | |
| **Variables** | **Mean values** | |
| Age (in years) | 35.546 | |
| Credit amount (in DM or Deutsch Mark) | 3271.258 | |
| Duration (in months) | 20.903 | |
| **Variables** | **Encoded values (Original object names)** | **Count** |
| Sex | 1 (Male) | 690 |
| | 2 (Female) | 310 |
| Job level | 0 (unskilled and non-resident) | 22 |
| | 1 (unskilled and resident) | 200 |
| | 2 (skilled) | 630 |
| | 3 (highly skilled) | 148 |
| Housing | 1 (Rent) | 179 |
| | 2 (Free) | 108 |
| | 3 (Own) | 713 |
| Savings | 1 (Little (less than 100 DM)) | 603 |
| | 2 (Moderate (between 100 and 500 DM)) | 103 |
| | 3 (Quite rich (between 500 and 1000 DM)) | 63 |
| | 4 (Rich (more than 1000 DM)) | 48 |
| Checking | 1 (Little (less than 0 DM)) | 274 |
| | 2 (Moderate (between 0 and 200 DM)) | 269 |
| | 3 (Rich (more than 200 DM)) | 48 |
| Purpose | Car | 337 |
| | Radio/TV | 280 |
| | Furniture/equipment | 181 |
| | Business | 97 |
| | Education | 59 |
| | Repairs | 22 |
| | Domestic appliances | 12 |
| | Vacation/others | 12 |
| Classification | 0 (Bad) | 300 |
| | 1 (Good) | 700 |

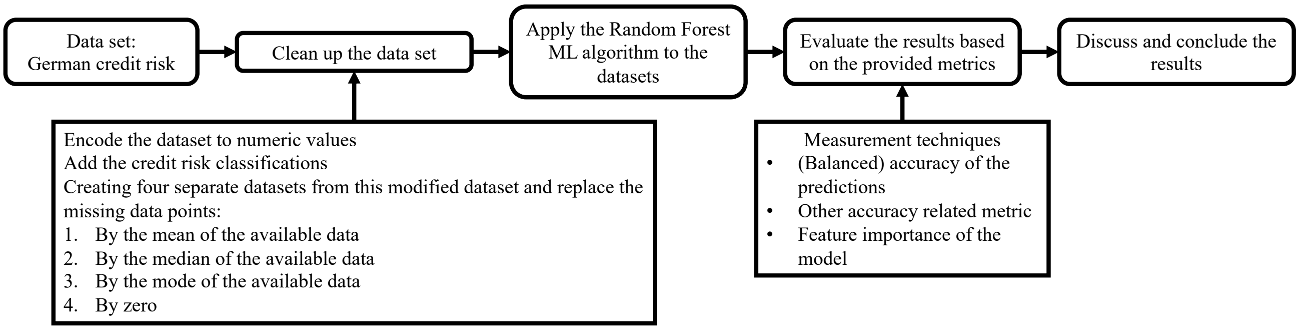
Data set:
German credit risk
Clean up the data set
Apply the Random Forest ML algorithm to the datasets
Evaluate the results based on the provided metrics
Discuss and conclude the results
Encode the dataset to numeric values
Add the credit risk classifications
Creating four separate datasets from this modified dataset and replace the missing data points:
1. By the mean of the available data
2. By the median of the available data
3. By the mode of the available data
4. By zero
Measurement techniques
• (Balanced) accuracy of the predictions
• Other accuracy related metric
• Feature importance of the model

**Figure 2. Research framework describing the successive steps in the experiments**

# 5. RESULTS & DISCUSSION

When the data is split into the training and test datasets, the splitting order is randomised. This means that every time the dataset is split into these training and test datasets, different consumers are assigned to these two different datasets until the 80/20 split is reached. This randomisation leads to different outcomes of the accuracy calculations. Thus, the accuracy of an algorithm cannot be assessed based on one split of the dataset into the training and test data. Therefore, the datasets were split ten times, leading to ten trials, and each time, the accuracy of the four datasets was measured. These measurements are presented in Table 4 and in a box and whisker plot in Figure 3. In addition to the four “standard” replacement methods that will be used in this experiment, a fifth method is added to both Table 4 and Figure 3 to show the effects on the accuracy of deleting the rows with missing values entirely, as proposed by Clintworth et al. (2023).

**Table 4. Accuracy measurements of replacement methods**

| **Trials** | **Replacement method for missing data** | | | | |
|---|---|---|---|---|---|
| | **Mean (%)** | **Median (%)** | **Mode (%)** | **Zero (%)** | **Delete (%)** |
| 1 | 74 | 73.5 | 74.5 | 72.5 | 73 |
| 2 | 77 | 76 | 75.5 | 77 | 77.5 |
| 3 | 72 | 76 | 71.5 | 73.5 | 72 |
| 4 | 71 | 68 | 68.5 | 70 | 71. |
| 5 | 74 | 74.5 | 72.5 | 72.5 | 73.5 |
| 6 | 75 | 75 | 76 | 77 | 76 |
| 7 | **79** | **79.5** | **78** | **82** | **79** |
| 8 | 70.5 | 74.5 | 71 | 71.5 | 70 |
| 9 | 74.5 | 73 | 75.5 | 74 | 75.5 |
| 10 | 73.5 | 70.5 | 73.5 | 72 | 72.5 |
| Mean accuracy (%) | 74.1 | 74.05 | 73.65 | **74.2** | 74.05 |
| Median accuracy (%) | 74 | **74.5** | 74 | 73 | 73.25 |
| Lowest value (%) | **70.5** | 68 | 68.5 | 70 | 70 |
| Highest value (%) | 79 | 79.5 | 78 | **82** | 79 |
| Spread (pp) | **8.5** | 11.5 | 9.5 | 12 | 9 |

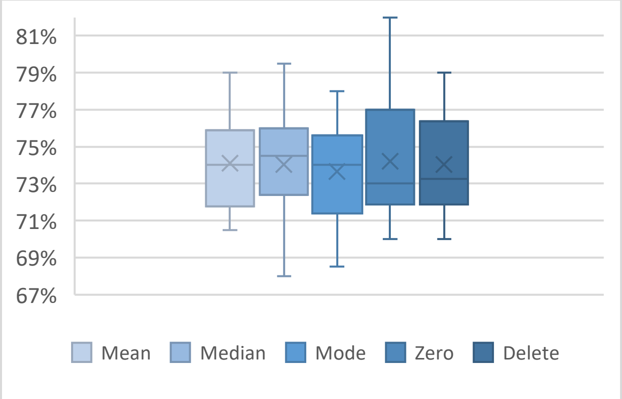

**Figure 3. Box and whisker plot of accuracy measurements**

Table 4 shows that there is no clear "winning" pre-processing method based purely on the (standard) accuracy metric. This can be seen as the pre-processing method of replacing by zero has the highest average accuracy but also the lowest median accuracy and the highest spread between the measurements of all pre-processing methods.

This spread between the highest and lowest measurements shows a degree of variation and therefore, uncertainty. A low variation between the highest and lowest measurements is preferred as this means that the uncertainty of the accuracy of the pre-processing method is also lower.

Table 4 shows that deleting the rows results in comparable results as the other methods, however, it does not outperform the other replacement methods. In addition, deleting the missing values is not possible for individual predictions as the algorithm that was used only worked if all variables had values.

Overall, in Table 4, the pre-processing method of replacing the missing values by the mean of the available data is performing the best as it has the second highest mean and median accuracy and the lowest spread.

Figure 3 shows that the top whisker of the pre-processing method of replacing by zero is the biggest, meaning that the mean of this method is positively influenced by the outlier-like best performance measured, while the biggest bottom whisker is from the pre-processing method of replacing by the median, negatively influencing the average accuracy measurement of this method.

In addition to the accuracy comparison, there were also further performance metrics of the pre-processing methods that were tested. This happened by performing the test five times and selecting the results from the algorithm that performed with the highest accuracy. Next, the results of these Random Forest algorithms on the four individual datasets will be presented and discussed.

| | | | |
|---|---|---|---|
| Actual value | 0 | 20 | 35 |
| | 1 | 11 | 134 |
| | | 0 | 1 |
| | | Predicted value | |

**Figure 4. Confusion matrix of predictions method Mean**

The confusion matrix in Figure 4 shows the classification values that the Random Forest algorithm predicted in the test dataset that was pre-processed by replacing the missing values by the mean of the available data and comparing these results to the actual classification values from the original dataset.

Figure 4 shows that most of the 200 predicted values (the testing dataset was 20 per cent of the original dataset of 1000 consumers) are correctly predicted to be positive, in addition, a minority of predictions correctly predicted the negative classification values, leading to 77% of the predictions to be accurate, as can be seen in Table 5. The remaining predictions incorrectly classified the consumers as either a high credit risk when they are a low risk (false negative) or a low credit risk when they should have been classified as a high credit risk (false positive) with the majority of the incorrect predictions being the latter. Preferably for financial institutions, the false negative should be more prominent than false positive. This would reduce the consumer credit risk because when all people who have a high credit risk are rejected and some people with a low credit risk are also rejected, the total credit risk is lower than if all these people are accepted.

In Table 5, along with previously explained metrics of validation, there are two new metrics of validation that were outputted by the Random Forest machine learning algorithm, these are Recall and Precision (Fränti & Mariescu-Istodor, 2023):

$$Recall = \frac{True\ positive\ predictions}{(True\ positive +\ False\ negative\ predictions)}$$

$$Precision = Sensitivity = \frac{True\ negative}{Total\ negative}$$

Due to the high relative number of True positive and the low relative number of  False negative predictions, the Recall values is considerably higher than the general accuracy metric, while the precision (or sensitivity) metric has a more similar value to the accuracy measured of the algorithm.

The specificity metric is considerably lower than the other metrics, leading to a balanced accuracy metric of 71.9% for the pre-processing method of replacing the missing values by the mean of the available data.

Table 5 shows an overview of the four pre-processing method that were used to replace the missing data.

**Table 5. Metrics of validation**

| Metrics of validation | Mean (%) | Median (%) | Mode (%) | Zero (%) |
|---|---|---|---|---|
| Accuracy | **77.00** | 74.00 | 70.50 | 76.38 |
| Recall | 92.41 | 92.14 | **100.00** | 91.16 |
| Precision/Sensitivity | 79.29 | 75.88 | 70.35 | **79.76** |
| Specificity | 64.52 | 63.33 | **100.00** | 58.06 |
| Balanced accuracy | 71.90 | 69.61 | **85.18** | 68.91 |

Figure 5 shows the feature importance that was outputted by the Random Forest algorithm for the pre-processing method of replacing by the mean. Figure 5 shows that the most important variables in this dataset, according to the Random Forest algorithm, are credit (amount), age, duration (of the loan), and (the amount of money in the consumer's) checking account. These four variables make up for 72.2% of the most influential features in this model. In addition, the variables where the missing data was present, and later replaced, account for 19.95% of the feature importance.

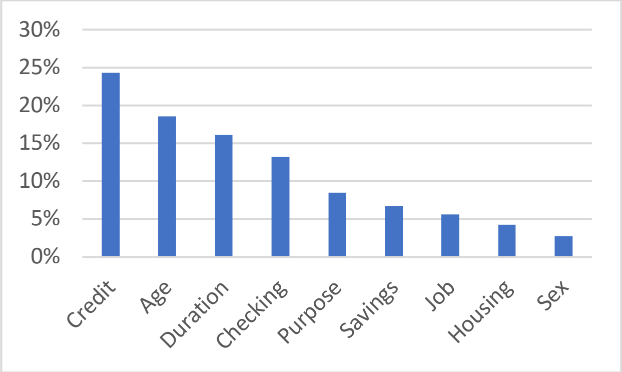

**Figure 5. Feature importance for method Mean**

| | | | |
|---|---|---|---|
| Actual value | 0 | 19 | 41 |
| | 1 | 11 | 129 |
| | | 0 | 1 |
| | | Predicted value | |

**Figure 6. Confusion matrix of predictions method Median**

The confusion matrix in Figure 6 shows that the way of pre-processing the dataset by replacing the missing values with the median of the variables is yields less correct results, both in true positive as in true negative, than replacing by the mean of the variables.
This can be seen as the correctly predicted values for the consumer credit risk has a total of 148 out of the 200 predictions in comparison to the 154 of the 200 predictions when replacing with the mean of the variables. However, the difference between these techniques cannot be accurately derived from Figure 6 as the values for both pre-processing techniques are similar with small relative deviations.
Relatively, replacing by the median of the variables also is more prone to false positive predictions, 41/52 compared to 35/46 in Figure 4. This means that the pre-processing technique from Figure 6 will approve relatively more consumer credit applications for people whose credit risk is higher than the algorithm predicted. This increases the general consumer credit risk for banks and other financial institutions.

The median column in Table 5 shows how that in every single measure of validation, replacing the missing data by the median yields lower results on the metrics than replacing by the mean as can be seen when comparing the results for replacing by the mean and median in Table 5. As these metrics of validation were the highest scores from the five trials that were run, the pre-processing method of replacing by the mean of the available data is the preferred option of these two methods when comparing accuracy metrics of validation.

Figure 7 shows the feature importance that was outputted by the Random Forest algorithm for the pre-processing method of replacing by the median. Figure 5 and Figure 7 have the same order of feature importance, the differences arise when comparing the percentual importance of each individual feature. The top four variables make up for 73.3% of the most influential features in this model. The individual feature importance for the first three values in the top four most influential features in Figure 7 are higher than the features in Figure 5, however, the fourth variable, Checking, where missing data was replaced, scores roughly 3.5 percent point lower than in Figure 5.
From this can be argued that replacing the missing values by the mean of the variables is more effective than replacing by the median. This is because the newly generated variables have a higher influence on the prediction of the consumer credit risk, and in combination with the higher accuracy metrics for replacing by the mean than by the median of the available variables, replacing by the mean has a higher influence on more accurate predictions than replacing by the median has.
In addition, the variables where some data was missing, and therefore replaced, account for 14.14% of the feature importance. This is lower than the 19.95% from Figure 5 meaning that the variables where the data was replaced by the median, had less relative importance for the predictions by the Random Forest machine learning algorithm.
Therefore, the pre-processing method of replacing missing values by the mean is also the preferred technique according to the feature importance comparison between Figure 5 and Figure 7.

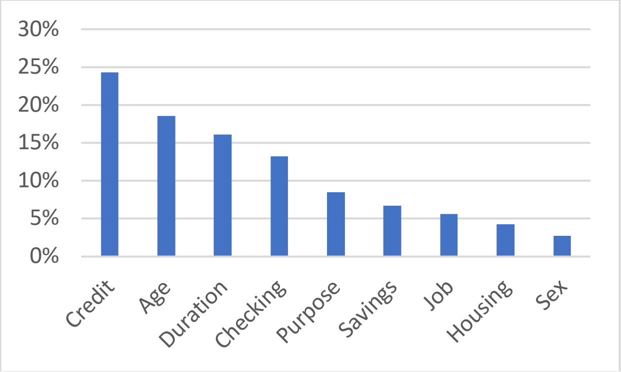

**Figure 7. Feature importances for method Median**

| | | | |
|---|---|---|---|
| Actual value | 0 | 1 | 59 |
| | 1 | 0 | 140 |
| | | 0 | 1 |
| | | Predicted value | |

**Figure 8. Confusion matrix of predictions method Mode**

As the pre-processing method of replacing the missing values by the mean of the variables has scored better than replacing by the median of the variables, this will be the benchmark to compare against for the method of replacing by the mode of the variables. Figure 8 shows that replacing by the mode results in an absolute decrease in the number of correct predictions. Although this pre-processing method correctly identified all consumers with a good (low) credit risk, it identified one of the consumers with a bad (high) credit risk.

As the mode column in Table 5 shows, the ratio of correct predictions for this pre-processing method means that the recall and specificity of this algorithm are 100 per cent, while general accuracy and precision/sensitivity are closer to 70 per cent. Due to the high specificity, the balanced accuracy is 85.18 per cent, making it the best scoring pre-processing method based on balanced accuracy.

However, the way that this algorithm reached such a high balanced accuracy was by approving, except for one, all the credit risk applications.

This approval rate means that the financial institution would have a great exposure to consumer credit risk. This greater risk is due to the possible upside of a loan or other type of credit is the interest, while the possible downside of a loan is the entire credit amount. In the period between January 2010 and February 2023, the average interest rate on consumer credit was 10.17% (Federal Reserve Bank of St. Louis, 2023). This interest rate was for a period of 24 months.

If standard compound interest is assumed (Ovaska & Sumell, 2017), the formula for the possible upside (or profit) for financial institutions would be:

$$Profit = FV - PV = (PV \times (1+i)^{t}) - PV$$

When the averages of the German credit dataset and the average interest rate since 2010 are inputted into this formula, the average upside for the financial institution would be:

$$(3271.26 \times (1 + 10.17\%)^{1.74}) - 3271.26 = 600.47\ DM$$

This means that the financial institution can profit 600.47 Deutsche Mark for every true positive consumer with low credit risk, while it loses 3271.26 Deutsche Mark for every false positive consumer credit risk prediction (given that the false positive applicant immediately defaults on their payments).

This means that the preferred Random Forest algorithm would reduce the credit risk more by minimising false positive rather than reducing false negative, as reducing false negative has a smaller revenue increase than false positive has a revenue decrease. The pre-processing method of replacing by the mean of the variables has a better ratio for predicting true negative and false positives (20/35) compared to the ratio of replacing by the mode (1/59). As this ratio is better for the pre-processing method of replacing by the mean, this is the preferred method, according to the comparison between Figure 4 and Figure 8.

As the highest balanced accuracy in the mode column in Table 5 is caused by the lack of false positives, this metric of validation is misleading for the overall performance of the pre-processing method and performance of the Random Forest algorithm.

Figure 9 shows the feature importance of the pre-processing method of replacing the missing values by the mode of the variables. This figure shows that the savings and checking variables have little influence on the overall prediction from the Random Forest machine learning algorithm, respectively 3.59 per cent and 2.74 per cent. In addition to this, the features before the variables where the missing data was replaced, account for 90.75 per cent of the feature importance.

The combined feature importance of the variables with, originally, missing data is 6.33 per cent meaning that these variables do not have a lot of influence on the predictions of the algorithm and also not a lot of correlation with the classification on whether someone has a low or high credit risk. However, higher amounts of money in the checking and savings accounts can be an indicator of financial responsible behaviour, and therefore financial capability (Xiao et al., 2014). Financial capability could be argued to have a correlation with consumer credit risk as consumers who are capable to manage their finances seem less likely to default on their loans and other forms of consumer credit. Therefore, it can be concluded that according to Figure 9, the pre-processing method of replacing the missing data by the mode of the variables is not a suitable method of data pre-processing, and replacing by the mean continues to be the preferred method.

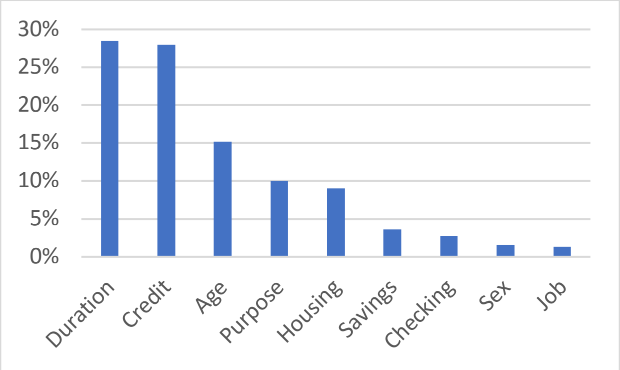

**Figure 9. Feature importances for method Mode**

| | | | |
|---|---|---|---|
| Actual value | 0 | 18 | 35 |
| | 1 | 13 | 134 |
| | | 0 | 1 |
| | | Predicted value | |

**Figure 10. Confusion matrix of predictions method Zero**

Figure 10 shows the final confusion matrix. This confusion matrix shows the predicted values for the pre-processing method of replacing the missing values with zero. As with replacing with the median and mode of the available data, these results will be compared against the results of the preferred method so far, replacing the missing values with the mean of the variables, as can be seen in Figure 4.

The positive predictions in Figure 10 have the same ratio as in Figure 4 with 134 true positive and 35 false positive predictions, while the negative predictions score lower in Figure 10 than in Figure 4 with an increase of 2 false negative and a decrease of 2 true negative predictions.

In addition to the comparison between Figure 4 and Figure 10, the replacing by zero column in Table 5 gives a similar image when compared to replacing by the mean. The scores for the latter are similar but on every metric a bit higher than replacing by zero, this is due to the differences in the negative predictions for the two pre-processing methods.

As the results for these two pre-processing methods are similar after the five trials that were run, it is possible that replacing by zero might score higher on some occasions.

Figure 11 shows the feature importance for the pre-processing method of replacing the missing values by zero. This figure gives a higher importance for the variables where the missing data was replaced (combined 25.48 per cent for replacing by zero compared to combined 19.95 per cent for replacing by the mean of the variables), suggesting that when the correlation between these variables and the classification is higher when replaced by zero than by replacing by the mean.

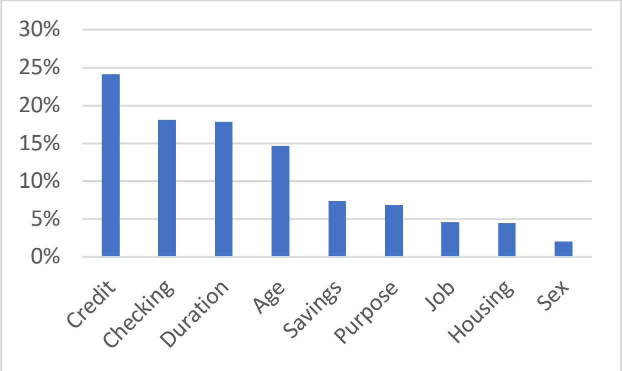

**Figure 11. Feature importance for method Zero**

## 6. CONCLUSION

The aim of this thesis is to discover the best way to replace the missing data in sparse and bad quality datasets for implementation in a machine learning model for consumer credit risk predictions.

The replacement methods of replacing the missing values by the mean, median, and mode of the variables and replacing by zero were applied to the German credit dataset. A Random Forest machine learning algorithm was applied to these pre-processed datasets and the results of the predictions were then evaluated based on accuracy metrics by the use of a confusion matrix, in addition, these pre-processing methods were evaluated on the feature importance that the algorithm presented.

The comparison between the accuracy metrics on the four pre-processing methods, in addition to deleting the missing values, shows that replacing by the mean and replacing by zero have the best results. When the four pre-processing methods were compared based on their feature importance, replacing by zero scored the best. Therefore, replacing by zero is the preferred replacement method for missing data in sparse and bad quality datasets, if it is used in combination with the encodings from Table 2.

The replacement method of replacing by zero, in collaboration with the encoding that was used, makes the Random Forest algorithm assume that the consumer does not have any/very little money in their savings and/or checking account. This is because the output of the replacement method is lower than the lowest possible value in the Savings and Checking variables. The replacement method of replacing by zero is low in consumer credit risk exposure for banks and other financial institutions due to a low risk of false positives.

The method of replacing by zero was the best method when applied to this specific case, however, this does not mean that this is the best method when applied to other cases.

Therefore, in future experiments, additional missing data replacement methods could be identified and tested to determine if there are more accurate ways to work with sparse and bad quality datasets. Additionally, one hot encoding can be used to present the non-ordinal variables more accurately to the Random Forest machine learning model. Cross validation could also be used in future research for the accuracy assessment of the different Random Forest algorithms.

## 7. REFERENCES

1. Aitha, V., & Jathanna, R. D. (2019). Credit risk assessment using machine learning techniques. *International Journal of Innovative Technology and Exploring Engineering*, 4. https://doi.org/10.35940/ijitee.A4936.119119
2. Ayodele, T. (2010). *Types of Machine Learning Algorithms*. https://doi.org/10.5772/9385
3. Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7
4. Brodley, C. E., & Friedl, M. A. (1999). Identifying Mislabeled Training Data. *Journal of Artificial Intelligence Research*, *11*, 131–167. https://doi.org/10.1613/jair.606
5. Clintworth, M., Lyridis, D., & Boulougouris, E. (2023). Financial risk assessment in shipping: a holistic machine learning based methodology. *Maritime Economics & Logistics*, *25*(1), 90–121. https://doi.org/10.1057/s41278-020-00183-2
6. Davis, R., Lo, A. W., Mishra, S., Nourian, A., Singh, M., Wu, N., & Zhang, R. (2022). Explainable Machine Learning Models of Consumer Credit Risk. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4006840
7. Dinga, R., Penninx, B. W. J. H., Veltman, D. J., Schmaal, L., & Marquand, A. F. (2019). Beyond accuracy: Measures for assessing machine learning models, pitfalls and guidelines. *BioRxiv*, 743138. https://doi.org/10.1101/743138
8. Donepudi, P. K. (2017). Machine Learning and Artificial Intelligence in Banking. *Engineering International*, *5*(2), 83–86. https://doi.org/10.18034/ei.v5i2.490
9. Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 210–215. https://doi.org/10.23919/MIPRO.2018.8400040
10. Federal Reserve Bank of St. Louis. (2023). *Board of Governors of the Federal Reserve System (US), Finance Rate on Personal Loans at Commercial Banks, 24 Month Loan*. https://fred.stlouisfed.org/graph/?id=TERMCBPER24NS,
11. FFIEC. (2023). *Charge-Off and Delinquency Rates on Loans and Leases at Commercial Banks* (F. F. I. E. Council, Ed.). https://www.federalreserve.gov/releases/chargeoff/delallsa.htm
12. Fränti, P., & Mariescu-Istodor, R. (2023). Soft precision and recall. *Pattern Recognition Letters*, *167*, 115–121. https://doi.org/https://doi.org/10.1016/j.patrec.2023.02.005
13. Galindo, J., & Tamayo, P. (2000). Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications. *Computational Economics*, *15*(1), 107–143. https://doi.org/10.1023/A:1008699112516
14. Gupta, A., Pant, V., Kumar, S., & Bansal, P. K. (2020). Bank Loan Prediction System using Machine Learning. *2020 9th International Conference System Modeling and Advancement in Research Trends (SMART)*, 423–426. https://doi.org/10.1109/SMART50582.2020.9336801
15. Helm, J. M., Swiergosz, A. M., Haeberle, H. S., Karnuta, J. M., Schaffer, J. L., Krebs, V. E., Spitzer, A. I., & Ramkumar, P. N. (2020). Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions. *Current Reviews in Musculoskeletal Medicine*, *13*(1), 69–76. https://doi.org/10.1007/s12178-020-09600-8
16. Hofmann, H. (1994). *German Credit risk*. https://www.kaggle.com/datasets/uciml/german-credit
17. Jain, A., Patel, H., Nagalapatti, L., Gupta, N., Mehta, S., Guttula, S., Mujumdar, S., Afzal, S., Mittal, R. S., & Munigala, V. (2020). Overview and Importance of Data Quality for Machine Learning Tasks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 3561–3562). Association for Computing Machinery. https://doi.org/10.1145/3394486.3406477
18. Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*(11), 2767–2787. https://doi.org/10.1016/j.jbankfin.2010.06.001
19. Lestari, N. I., Hussain, W., Merigo, J. M., & Bekhit, M. (2023). A Survey of Trendy Financial Sector Applications of Machine and Deep Learning. In M. A. Jan & F. Khan (Eds.), *Application of Big Data, Blockchain, and Internet of Things for Education Informatization* (pp. 619–633). Springer Nature Switzerland.
20. Nicora, G., Rios, M., Abu-Hanna, A., & Bellazzi, R. (2022). Evaluating pointwise reliability of machine learning prediction. *Journal of Biomedical Informatics*, *127*, 103996. https://doi.org/https://doi.org/10.1016/j.jbi.2022.103996
21. Niu, M., Wang, Y., Zhang, K., & Zhao, C. (2023). Comparison of different individual credit risk assessment models. In S. Jin & W. Dai (Eds.), *Second International Conference on Statistics, Applied Mathematics, and Computing Science (CSAMCS 2022)* (p. 120). SPIE. https://doi.org/10.1117/12.2672657
22. Ovaska, T., & Sumell, A. (2017). *Journal of Economics and Finance Education: "Increase Interest In Compound Interest: Economic Growth and Personal Finance." 16*, 85–97.
23. Provenzano, A. R., Trifirò, D., Datteo, A., Giada, L., Jean, N., Riciputi, A., Le Pera, G., Spadaccino, M., Massaron, L., & Nordio, C. (2020). *Machine Learning approach for Credit Scoring*.
24. Rácz, A., Bajusz, D., & Héberger, K. (2021). Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification. *Molecules*, *26*(4), 1111. https://doi.org/10.3390/molecules26041111
25. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778
26. Rona-Tas, A., & Guseva, A. (2018). Consumer Credit in Comparative Perspective. *Annual Review of Sociology*, *44*(1), 55–75. https://doi.org/10.1146/annurev-soc-060116-053653
27. Rudin, C., & Shaposhnik, Y. (2023). Globally-Consistent Rule-Based Summary-Explanations for Machine Learning Models: Application to Credit-Risk Evaluation. *Journal of Machine Learning Research*, *24*(16), 1–44.
28. Shi, S., Tse, R., Luo, W., D'Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: a systemic review. *Neural Computing and Applications*, *34*(17), 14327–14339. https://doi.org/10.1007/s00521-022-07472-2
29. Spuchľáková, E., Valašková, K., & Adamko, P. (2015). The Credit Risk and its Measurement, Hedging and Monitoring. *Procedia Economics and Finance*, *24*, 675–681. https://doi.org/https://doi.org/10.1016/S2212-5671(15)00671-1
30. The Federal Reserve. (2023a). *Consumer Credit Current - G.19*. https://www.federalreserve.gov/releases/g19/current/

31. The Federal Reserve. (2023b). *Consumer Credit Historical - G.19*. https://www.federalreserve.gov/releases/g19/HIST/cc_hist_mh_levels.html
32. Twala, B. (2010). Multiple classifier application to credit risk assessment. *Expert Systems with Applications*, *37*(4), 3326–3336. https://doi.org/https://doi.org/10.1016/j.eswa.2009.10.018
33. Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K. A., Elkhatib, Y., Hussain, A., & Al-Fuqaha, A. (2019). Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges. *IEEE Access*, *7*, 65579–65615. https://doi.org/10.1109/ACCESS.2019.2916648
34. van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, *109*(2), 373–440. https://doi.org/10.1007/s10994-019-05855-6
35. Veras, M. B. A., Mesquita, D. P. P., Mattos, C. L. C., & Gomes, J. P. P. (2020). A sparse linear regression model for incomplete datasets. *Pattern Analysis and Applications*, *23*(3), 1293–1303. https://doi.org/10.1007/s10044-019-00859-3
36. Wei, Y., Kirkulak-Uludag, B., ZHU, D., & Zhou, Z. (2023). Stacking Ensemble Method for Personal Credit Risk Assessment in P2P Lending. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4318348
37. Xia, Y., Xu, T., Wei, M.-X., Wei, Z.-K., & Tang, L.-J. (2023). Predicting Chain's Manufacturing SME Credit Risk in Supply Chain Finance Based on Machine Learning Methods. *Sustainability*, *15*(2), 1087. https://doi.org/10.3390/su15021087
38. Xiao, J. J., Chen, C., & Chen, F. (2014). Consumer Financial Capability and Financial Satisfaction. *Social Indicators Research*, *118*(1), 415–432. https://doi.org/10.1007/s11205-013-0414-8
39. Yu, L., & Liu, H. (2004). Efficient Feature Selection via Analysis of Relevance and Redundancy. *J. Mach. Learn. Res.*, *5*, 1205–1224.