CHARLES UNIVERSITY
FACULTY OF SOCIAL SCIENCES
Institute of Economic Studies

Binning numerical variables in credit risk models

Master's thesis

Author: Bc. Matyáš Mattanelli

Study program: Economics and finance

Supervisor: doc. PhDr. Jozef Baruník, Ph.D.

Year of defense: 2023

Declaration of Authorship

The author hereby declares that he or she compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain any other academic title.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis in whole or in part and agrees with the thesis being used for study and scientific purposes.

Prague, July 28, 2023

Bc. Matyáš Mattanelli

Abstract

This thesis investigates the effect of binning numerical variables on the performance of credit risk models. The differences are evaluated utilizing five publicly available data sets, six evaluation metrics, and a rigorous statistical test. The results suggest that the binning transformation has a positive and significant effect on the performance of logistic regression, feedforward artificial neural network, and the Naïve Bayes classifier. The most affected aspect of model performance appears to be its ability to differentiate between eligible and ineligible customers. The obtained evidence is particularly pronounced for moderately-sized data sets. In addition, the findings are robust to the inclusion of missing values, the elimination of outliers, and the exclusion of categorical features. No significant positive effect of the binning transformation was found for the decision tree algorithm and the Random Forest model.

JEL Classification	C18, C51, C58, C61, G21
Keywords	Credit risk, binning, machine learning, performance
Title	Binning numerical variables in credit risk models

Abstrakt

Tato práce zkoumá vliv diskretizace numerických proměnných na výkonnost modelů kreditního rizika. Rozdíly ve výkonnosti jsou vyhodnoceny s využitím pěti veřejně dostupných datových souborů, šesti indikátorů výkonnosti a statistického testu. Výsledky naznačují, že diskretizace má pozitivní a významný vliv na výkonnost logistické regrese, neuronové sítě a naivního Bayes klasifikátoru. Nejvíce ovlivněným aspektem výkonnosti modelu se zdá být jeho schopnost rozlišovat mezi dobrými a špatnými klienty. Výsledky jsou zvláště patrné pro středně velké datové soubory. Závěry jsou odolné vůči chybějícím hodnotám, eliminaci extrémních pozorování a vyloučení kategorických proměnných. Pro rozhodovací strom a náhodný les nebyl nalezen žádný významný pozitivní účinek diskretizace na výkonnost.

Klasifikace JEL	C18, C51, C58, C61, G21
Klíčová slova	Kreditní riziko, diskretizace, strojové učení, výkonnost
Název práce	Diskretizace numerických proměnných v modelech kreditního rizika

Acknowledgments

I would like to express my utmost gratitude to my supervisor, doc. PhDr. Jozef Baruník, Ph.D., for his valuable guidance and advice. I would also like to convey my profound appreciation to my family and significant other, without whose extensive moral support, the compilation of this thesis would not be possible.

Typeset in FSV LaTeX template with great thanks to prof. Zuzana Havrankova and prof. Tomas Havranek of Institute of Economic Studies, Faculty of Social Sciences, Charles University.

Bibliographic Record

Mattanelli, Matyáš: *Binning numerical variables in credit risk models.* Master's thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague. 2023, pages 103. Advisor: doc. PhDr. Jozef Baruník, Ph.D.

Contents

List of Tables

List of Figures

Acronyms

AUC Area Under Curver

BIS Bank for International Settlements

BS Brier Score

CART Classification and Regression Trees

DCCCT Default of Credit Card Clients in Taiwan

ER Event Rate

FPR False Positive Rate

GMSC Give Me Some Credit

HCDR Home Credit Default Risk

IV Information Value

KS Kolmogorov-Smirnov

NN Neural network

PGI Partial GINI Index

PD Probability of Default

RF Random Forest

ROC Receiver Operating Characteristic

SGCC South German Credit Card

SVM Support Vector Machines

TPR True Positive Rate

UCI MLR University of California Irvin Machine Learning Repository

WoE Weight of Evidence

Chapter 1

Introduction

Estimating the probability that a customer will default in the future is one of the most fundamental responsibilities of financial institutions. Since a correct evaluation of a customer's credibility may have a substantial impact on profitability and is required by the regulatory framework, considerable attention is devoted to the methodology of the development of credit risk models. While the estimation process is thoroughly scrutinized, the focus on data preprocessing is fairly neglected in the existing literature (Raymaekers *et al.* 2022). In addition, despite the rapid advancements in computational power and the resulting feasibility of more complex models, the performance improvements achieved by employing overly complicated methods do not appear to be substantial (Lessmann *et al.* 2015). As a result, the enhancement of data quality inspection and data preprocessing may constitute a more rewarding line of research.

Nevertheless, even the most basic machine learning methods have been shown to considerably enhance model performance as compared to the industry standard logistic regression (Baesens *et al.* 2003; Lessmann *et al.* 2015). And since even a minor improvement in the model's performance may yield substantial cost savings for the financial institution (Khandani *et al.* 2010; Lessmann *et al.* 2015), the utilization of machine learning methods is enticing. However, the regulatory framework requires credit risk models to be interpretable, and thus the employment of non-transparent "black-box" models is infeasible. As a result, various attempts were made to enhance the interpretability of complicated models. One of the suggested options is the binning of numerical variables and their subsequent Weight of Evidence transformation (Augasta & Kathirvalavakumar 2013; Raymaekers *et al.* 2022). While this procedure may increase

the model's interpretability, its effect on performance is yet to be thoroughly inspected.

Consequently, this thesis investigates the effect of binning numerical variables on the performance of common machine learning estimation methods. The considered algorithms are the industry standard logistic regression, the decision tree, the Random Forest model, the artificial neural network, and the Naïve Bayes classifier. In order to address the deficiencies in the existing literature, six evaluation metrics, as well as a rigorous statistical test, are utilized to investigate the differences in model performance. In addition, five publicly available data sets are considered for estimation to secure the robustness of the results. Moreover, the reliability of the findings is further verified by various supporting analyses, including the treatment of missing values, exclusion of categorical variables, and handling of outliers.

The results obtained in this thesis suggest that three of the five considered estimation methods seem to significantly benefit from utilizing the binning transformation of numerical features. Strong evidence was found in favor of the binning transformation for the feedforward artificial neural network and the Naïve Bayes classifier. In addition, the logistic regression appears to benefit from the binning transformation in certain aspects of model performance, especially its ability to differentiate between good and bad customers. The findings are particularly strong for moderately-sized data sets which represent the industry standard in credit risk modeling. On the other hand, the performance of two considered tree-based algorithms, the decision tree and the Random Forest model, does not seem to be improved by binning numerical variables.

The rest of the thesis is structured as follows. Section 2 provides an overview of the existing literature with regard to the effect of binning numerical variables on the performance of credit risk models. In addition, a discussion of the existing binning algorithms is presented, along with a brief outline of the utilization of machine learning models in credit risk. The chapter concludes by demonstrating the expected contribution of this thesis and enumerating the tested hypotheses. Section 3 offers a description of the utilized data sets, while Section 4 provides a thorough discussion of the employed methodology, including the depiction of the binning algorithm, the utilized methods, and, most importantly, the evaluation metrics. The results for each estimation method and the

supplementary analyses are examined in Section 5. The conclusion, along with policy recommendations and research limitations, is available in Section 6.

Chapter 2

Literature review

The literature review is structured as follows. The first part provides an overview of the extant evidence regarding the effect of binning numerical variables on the performance of classification models. Then, given the existence of various types of binning algorithms, the second subsection is concerned with their description and an inspection of their relative performance. The third part investigates the current state of the literature concerning the usage of machine learning algorithms in credit risk modeling, particularly emphasizing the methodology utilized for comparing classifiers' performance. The last two subsections demonstrate the expected contribution of this thesis to the extant literature and enumerate the proposed hypotheses, respectively.

2.1 Binning and performance

The issue of the effect of binning numerical variables on the performance of classification models is rather scarcely inspected in the existing literature. The lack of investigation may be partially attributed to the fact that only a handful of studies focus on postprocessing and preprocessing the data (Raymaekers *et al.* 2022). Nevertheless, several studies attempted to examine this topic.

Sharma (2011) found that the utilization of binning and a subsequent Weight of Evidence (WoE) transformation has a positive effect on the performance of a logistic regression model and an adverse impact on the predictive accuracy of a Random Forest (RF) model. The study utilized three credit risk-related datasets with a binary dependent variable. On the other hand, Lustgarten *et al.* (2008) showed that discretizing numerical features improves the performance

of three machine learning algorithms (Support Vector Machines, Naïve Bayes, and Random Forest). However, the data used for training and evaluation of the models were biomedical and thus likely bear different characteristics than credit risk data sets. Most importantly, the credit risk data sets tend to be heavily imbalanced (Addo *et al.* 2018; Gunnarsson *et al.* 2021) since they usually contain only a handful of defaulted clients. On the other hand, the class distribution in the biomedical data sets utilized in the study oscillates around 50%. Nevertheless, the results are still relevant for binary classification models in general.

Similarly, Dougherty *et al.* (1995) demonstrated that Naïve Bayes and a decision tree algorithm both benefit from pre-binning continuous features. In fact, the C4.5 decision tree algorithm showed the same or better performance on all 16 data sets covering various fields, including medicine and, most importantly, credit risk. Abraham *et al.* (2006) reached the same conclusion for the Naïve Bayes classifier. As an additional opposing view serves the study by Ventura & Martinez (1995), who showed that discretization deteriorates the accuracy of an ID3 decision tree algorithm.

As a result, it can be observed that the literature seems to support the performance improvement induced by binning for some estimation methods (SVM, Naïve Bayes) but is divided regarding the pre-binning of continuous features when it comes to tree-based algorithms. The question stands whether the local discretization inherent to the estimation process of all tree-based models is superior to binning the data prior to the estimation. While Dougherty *et al.* (1995) and Lavangnananda & Chattanachot (2017) show the dominance of global discretization for a decision tree, Sharma (2011) demonstrates the opposite for a Random Forest model. The RF model combines the predictions of many decision trees, which are, in turn, built by assessing variables' importance and a subsequent setting of an appropriate threshold that maximizes entropy (Breiman 2001). Since the RF model has been shown to outperform the decision tree classifier, especially on larger data sets (Ali *et al.* 2012; Esmaily *et al.* 2018; Prajwala 2015), it may be the case that while the lack of predictive power of a decision tree might be boosted by pre-binning the data, this might not hold for the RF classifier. This thesis will attempt to further elucidate this matter.

Even though the direct effect of binning on performance might be ambiguous,

there are some widely accepted advantages of discretization. For example, the ability to conveniently handle missing values by creating a separate bin, thus preventing information loss. In addition, the binning process allows for decreasing outliers' influence since they are put in a bin with regular observations (Leung *et al.* 2008; Verster 2018; Zeng 2014). Moreover, binning might decrease the nonlinearity in the data (Leung *et al.* 2008) and reduce the estimation variance since a slight change in the data does not significantly impact the results (Dougherty *et al.* 1995). Furthermore, certain classification algorithms can handle only categorical variables, and as a result, the discretization of continuous features becomes a necessity (Augasta & Kathirvalavakumar 2013; Ventura & Martinez 1995; Wójciak & Ĺupińska Dubicka 2018). Other classification algorithms may be designed to utilize continuous features but still perform better when dealing with categorical attributes, for example, the Naïve Bayes classifier (Wu *et al.* 2006). In particular, the Naïve Bayes classifier requires the estimation of frequencies which might prove cumbersome for continuous attributes with many distinct values. One solution is to assume that the continuous variables are normally distributed, but this may not always hold (Kotsiantis & Kanellopoulos 2005). Lastly, binning continuous variables and thus reducing the number of unique values significantly increases the speed and efficiency of classification algorithms (Augasta & Kathirvalavakumar 2013; Dash *et al.* 2011).

On the other hand, the possible disadvantages of binning need to be acknowledged as well. Most importantly, discretizing a continuous variable results in a loss of information, the so-called "discretization error" (Higham 2002). The effect of this phenomenon on the models' performance may be two-fold. Firstly, the reduced amount of information may decrease the model's predictive power. Conversely, if the data is noisy, the loss of information may prove beneficial since obsolete information can be disregarded and the data is represented in a more general way, which may help prevent overfitting (Augasta & Kathirvalavakumar 2013; Ventura & Martinez 1995). The extent of the discretization error heavily relies on the choice of an appropriate binning algorithm. This thesis's main aim is to evaluate one of these algorithms.

2.2 Binning algorithms

So far, the discussion covered only binning in general, i.e., converting a numerical variable into a categorical variable. However, there are many ways of achieving this goal. More specifically, the binning algorithms can be divided into several groups, as outlined in Augasta & Kathirvalavakumar (2013). Firstly, depending on whether the class label is utilized during the binning process, the algorithms are classified as either supervised (class information is considered) or unsupervised (class information is not considered). The superiority of either type is ambiguous in the extant literature since even though some studies suggest that supervised methods tend to outperform their unsupervised counterparts (Augasta & Kathirvalavakumar 2013; Dougherty *et al.* 1995; Kohavi & Sahami 1996), others show that the results heavily depend on the utilized data, and in some cases, unsupervised binning may perform equally well (Agre & Peev 2002; Dash *et al.* 2011; Wójciak & Ĺupińska Dubicka 2018). Nevertheless, the studies suggest that the unsupervised methods may group together values with distinct class labels, leading to the loss of class information and subsequent deterioration of a classifier's performance. On the other hand, supervised binning algorithms may put all values of a continuous variable into a single interval if the correlation with the dependent variable is close to zero, effectively disqualifying it for classification (Lustgarten *et al.* 2008). This may be viewed as an additional advantage of discretization since an appropriate supervised binning algorithm may be used for variable selection (Liu & Setiono 1997). Evidently, the feasibility of supervised algorithms depends on the availability of class labels. In addition, in favor of unsupervised algorithms speaks their efficiency since, due to their simplicity, they are significantly less computationally demanding (Augasta & Kathirvalavakumar 2013).

The supervised binning algorithms can be further divided into error-based, entropy-based, and statistics-based. The error-based algorithms attempt to minimize the prediction error on the training set resulting from classifying the observations using solely the given discretized variable (Kohavi & Sahami 1996). On the other hand, the entropy-based methods find the optimal bins by minimizing the entropy of the resulting intervals (Augasta & Kathirvalavakumar 2013). Lastly, statistics-based algorithms rely on statistical tests when searching for optimal interval division. An example is the ChiMerge algorithm (Kerber 1992) which initially assigns all unique values to separate intervals and subsequently iteratively merges adjacent intervals until all intervals are

statistically significant from each other.[1] Performance-wise comparison of the three types of algorithms seems to favor statistics-based binning (Augasta & Kathirvalavakumar 2013; Ventura & Martinez 1995). Kohavi & Sahami (1996) compared only error-based and entropy-based algorithms, and the latter appears to be superior. Nevertheless, the studies unanimously recognize that the performance depends on the underlying data and the classifier utilized.

Secondly, binning algorithms can be differentiated into dynamic and static based on whether or not the discretization procedure is intertwined with the classification process. Most binning algorithms are static since the data are transformed prior to classification. An example of dynamic binning would be the decision tree algorithm which recursively searches for optimal cut points as a part of the classification process (Breiman *et al.* 1984). Thirdly, global discretization methods perform binning on the entire data set, while local methods use only a subset of the data. As in the previous example, the decision tree algorithm may use only data subsets of a fixed size to find the optimal cut points and thus may represent an instance of a local discretization technique. This approach is common, especially for ensemble methods (Breiman 2001). Furthermore, binning algorithms such as ChiMerge, which start with smaller intervals and iteratively merge them, are called bottom-up. On the other hand, top-down methods begin with the whole range of data and search for optimal cut points (once more, the decision tree algorithm serves as an example of such a method). Moreover, direct methods require the number of intervals to be specified manually, whilst incremental methods seek optimal division based on a given criterion. Lastly, univariate binning algorithms differ from their multivariate counterparts by considering only a single variable at a time during the binning process rather than applying the algorithm on all variables simultaneously.

2.3 Machine learning in credit risk

Even though the extant credit risk literature does not seem to devote a lot of attention to data preprocessing, the utilization of machine learning models for predicting the probability of customers' default constitutes a popular research topic (Bhatore *et al.* 2020).[2] It is in banks' best interest to produce as accurate

[1] The significance is assessed using the χ^2 test.

[2] For a comprehensive review of the machine learning algorithms utilized in credit risk modeling see Breeden (2021).

predictions as possible since the estimated probabilities of default (PDs) enter the calculation of capital requirements under the Basel III regulatory framework (BIS 2011) and in case of application scoring also drive the decision of whether to provide or not to provide the given product. However, the stringency of the regulatory framework with respect to the methodology of the utilized models hinders the usage of machine learning algorithms since one of the conditions for the eligibility of a PD model is its interpretability. As a result, "black-box"[3] models are automatically disqualified, and more straightforward, transparent methods, such as logistic regression, remain the industry standard (Raymaekers *et al.* 2022).

Nevertheless, a large number of studies investigate the opportunities of utilizing machine learning algorithms for predicting default probability. One of the reasons is the existence of methods introducing interpretability even for complicated black-box models. As Raymaekers *et al.* (2022) indicate, one way to introduce interpretability to complex models is to discretize continuous variables and subsequently apply the WoE transformation.[4] In this way, the resulting discrete features attain a small number of unique values and have a straightforward relationship with the dependent variable. As a result, more complex models become feasible, including machine learning algorithms.

However, being more complex does not automatically imply better performance. As Addo *et al.* (2018) and Gunnarsson *et al.* (2021) demonstrate, deep learning models do not offer substantial performance gains over less complicated and computationally demanding models such as a single-layer neural network. A possible explanation might be that since credit risk data sets are usually small or medium-sized, the potential of deep learning models to uncover complicated relationships is difficult to exploit (LeCun *et al.* 2015). On the other hand, gradient-boosting algorithms such as XGBoost (Chen & Guestrin 2016) perform significantly better than the industry standard logistic regression. Such performance improvement might have a sizeable material importance for banks since even a slight improvement in the prediction accuracy may result in significant cost savings (Khandani *et al.* 2010; Lessmann *et al.* 2015). This is caused by the decrease in the probability of providing a loan to an ineligible customer.

[3] A model is classified as "black-box" if its complexity does not allow for a straightforward interpretation (Petch *et al.* 2022).

[4] The discussion of the methodology behind the WoE transformation will be provided in Section 4.1.

As mentioned above, the comparison of the performance of machine learning algorithms on credit risk data sets appears abundantly in the literature.[5] Among the most influential studies belongs Lessmann *et al.* (2015) who compared 41 classifiers across eight credit risk datasets. The authors extend and update the previous article by Baesens *et al.* (2003) in order to account for the current progress in the machine learning field. The study utilizes a rigorous methodology for classifier comparison, which will serve as an inspiration for this thesis. One of the most important methodological issues raised by the article is using an insufficient number of possibly inappropriate evaluation metrics. Most of the previously mentioned studies used accuracy as a performance indicator,[6] a suboptimal choice for substantially imbalanced credit risk data sets (Bekkar *et al.* 2013; De La Bourdonnaye & Daniel 2021). Similarly, as Hand (2009) and Powers (2012) showed, the very commonly utilized Area Under Curve (AUC) metric has several deficiencies. In addition, many studies employ only one or a few data sets and consequently cannot secure the robustness of the results. This thesis will attempt to avoid these deficiencies. A profound discussion is available in Section 3.

As for the results of Lessmann *et al.* (2015), the best-performing models appear to be heterogeneous ensembles that combine several distinct classifiers' predictions. The closest contender is the Random Forest model which represents a homogeneous ensemble. As mentioned above, more complex models do not consistently achieve better performance than their simpler counterparts. The results of the study support this claim since logistic regression manages to outperform a large number of models, including dynamic ensembles. Nevertheless, algorithms such as a one-layer neural network or the RF model still have the upper hand, demonstrating the advantage of machine learning utilization. The study does not consider any form of discretization of numerical features, which leaves an opportunity for the research of this thesis. It should be noted that the authors use two versions of each data set. In the first, categorical features are converted to dummy variables; in the second, they are encoded using WoE. The advantage of the latter approach is a substantially reduced dimensionality of the data (Raymaekers *et al.* 2022) which also enhances the efficiency of the classification process (Sharma 2011). Especially for variables with many unique values, the dummy variable approach creates a considerable number of

[5] See, for example, Khandani *et al.* (2010); Peng & Kou (2008); Putri *et al.* (2021).
[6] See Table 2.1.

features. In addition, it has been shown that using some form of encoding may positively affect performance (De La Bourdonnaye & Daniel 2021; Potdar *et al.* 2017).

2.4 Expected contribution

The issue of estimating customers' probability of default has been widely recognized as a very important topic. Consequently, a lot of attention in the literature is devoted to the perfection of the performance of the utilized models. However, while the estimation process tends to be thoroughly scrutinized, data preprocessing is fairly neglected. As a result, this thesis will attempt to evaluate the performance implications of utilizing a binning transformation of numerical features using rigorous methodology, several data sets, and multiple classifiers.

As shown above, various studies have attempted to investigate the effects, however, the utilized methodologies indicate numerous potential avenues for future research. As a further demonstration, Table 2.1 provides an overview of the existing articles and the evaluation metrics utilized in their analyses. As can be seen, all but a single study utilize merely the accuracy indicator which as discussed above may be a suboptimal choice for imbalanced data sets. Lustgarten *et al.* (2008) utilizes the Relative Classifier Information metric which is similar to AUC but is easily applicable also to multiclassification problems.[7] In addition, apart from Lustgarten *et al.* (2008),[8] none of the mentioned studies employ a rigorous statistical test to compare the performance of the classifiers. This thesis will attempt to improve upon the extant literature by employing several evaluation metrics each capturing a different aspect of classifiers' performance. Moreover, a statistical test will be conducted to obtain a significance level of the results.

Since the employment of machine learning algorithms in the credit risk industry is currently hindered by the stringent regulatory framework, new ways of increasing the feasibility of said algorithms need to be inspected. Improving the interpretability of complex models may lead to their increased utilization, which may result in substantial advancements in performance. As indicated

[7]See, for example, Statnikov *et al.* (2005).

[8]The study utilizes the Wilcoxon signed rank test and the t-test on the results from 24 datasets.

Table 2.1: Overview of the existing articles

Article	Metric	Classifier
Sharma (2011)	Accuracy	Logistic regression, RF
Lustgarten *et al.* (2008)	RCI	SVM, Naïve Bayes, RF
Dougherty *et al.* (1995)	Accuracy	Naïve Bayes, Decision tree (C4.5)
Abraham *et al.* (2006)	Accuracy	Naïve Bayes
Ventura & Martinez (1995)	Accuracy	Decision tree (ID3)
Lavangnananda & Chattanachot (2017)	Accuracy	NN, KNN, Naïve Bayes, Decision tree (ID3), SVM
Wu *et al.* (2006)	Accuracy	Naïve Bayes, Decision tree (ID3)
Augasta & Kathirvalavakumar (2013)	Accuracy	NN
Wójciak & Łupińska Dubicka (2018)	Accuracy	Bayesian network

Source: Author's review

above, even a minor performance enhancement may generate significant cost savings for financial institutions.

An additional contribution of this thesis is expected to be the evaluation of a novel binning algorithm widely utilized in practice. Moreover, various supplementary analyses will attempt to uncover the precise nature of the benefits of binning numerical variables, including the effect of missing values, outliers, encoding methods, and the involvement of categorical features.

2.5 Hypotheses

Based on the existing literature, the current thesis formulates the following hypotheses.

Hypothesis 1: *The binning of numerical variables improves the performance of the logistic regression model on credit risk data sets*

The first hypothesis concerns the most widely utilized method in the credit risk industry, the logistic regression. In line with the reviewed articles (Leung *et al.* 2008; Sharma 2011) and given the ability of the binning transformation to alleviate the effect of outliers as well as remove the noise from the data, the effect on performance is expected to be positive.

Secondly, a similar hypothesis is formulated for the decision tree.

Hypothesis 2: *The binning of numerical variables improves the performance of the decision tree algorithm on credit risk data sets*

The extant literature does not seem to have reached a consensus with regard to the implications of the binning transformation for the decision tree classifier.

This thesis argues that the advantages exceed the drawbacks, and thus the final effect on performance is positive.

Thirdly, the prevalent evidence in the reviewed articles is the deterioration of the performance of the Random Forest classifier upon the utilization of the binning transformation. As a result, the following hypothesis is formulated.

Hypothesis 3: *The binning of numerical variables does not improve the performance of the Random Forest classifier on credit risk data sets*

Fourthly, a very powerful machine learning algorithm is the neural network. The existing literature is not particularly dense with respect to the effect of binning of numerical variables on its performance. As a result, the following hypothesis is formulated.

Hypothesis 4: *The binning of numerical variables improves the performance of the feedforward artificial neural network classifier on credit risk data sets*

Lastly, it appears to be well-established that the Naïve Bayes algorithm benefits from the discretization of numerical variables. Consequently, the following hypothesis will be verified in this thesis:

Hypothesis 5: *The binning of numerical variables improves the performance of the Naïve Bayes classifier on credit risk data sets*

Chapter 3

Data description

The current chapter provides a description of the data utilized in this thesis. Following Lessmann *et al.* (2015), who suggested the usage of multiple data sets for classifier comparison to secure the robustness of the results, five publicly available data sets were acquired. The data originate from two primary sources. The first two data sets were provided by Kaggle, a Google LLC subsidiary and a public platform for publishing data sets for machine learning competitions. The three remaining data sets were obtained from the UCI Machine Learning Repository, which is a popular data source among researchers in the credit risk modeling field.[1] Each following subsection of the current chapter is devoted to one data set and its description. Given the quantitative nature of the analysis and the usage of multiple data sets, detailed characteristics of only the first data set are provided for conciseness. The last subsection presents an overall summary of all data sets.

3.1 Give Me Some Credit data set

The first data set called Give Me Some Credit (GMSC), acquired from Kaggle, was part of a competition whose goal was to develop a model predicting the probability that a bank's customer will default within two years from the reporting date (Cukierski 2011).

The data set contains 150 000 observations with ten independent variables and one dependent variable. Each observation represents a single borrower. All

[1] See, for example, Peng & Kou (2008); Potdar *et al.* (2017); Wójciak & Łupińska Dubicka (2018).

of the explanatory features are numerical, which serves well for the purpose of this thesis. The dependent variable is a binary indicator of whether the given customer was delinquent within two years from the reporting date. As is common in the credit risk industry, the data set is largely imbalanced. Merely approximately 6.68% of observations are flagged as defaulted. However, as Lessmann *et al.* (2015) argues, even though the class imbalance might bias the absolute performance of a classifier, the relative performance of two classifiers should remain comparable. And since this thesis aims to compare various methods rather than establish their absolute performance, no class-balancing techniques should be necessary.

The summary statistics of the data are available in Table 3.1. As can be seen, only two variables contain missing values. In case of *NumberOfDependents*, unavailable observations constitute approximately 2.61% of all observations, while for *MonthlyIncome*, the percentage is higher at approximately 19.82%. The treatment of missing values is described in Section 4.1.

As for the characteristics of the individual variables, the age of an average customer in the data set is 52, according to both the mean and the median. Therefore, the distribution does not appear to be substantially skewed. Since 41 and 63 are the 25th and 75th percentiles, respectively, the data set captures the older part of the population. The maximum age in the data set is 109, which seems improbable but still is not high enough to be disregarded. As discussed in Section 2.2, the binning algorithm is expected to handle outliers well by putting them in a bin with regular observations. Nevertheless, there appear to be only 13 customers over the age of 100, which is not expected to influence the results. On the other hand, the minimum value of zero is unreasonable, and thus the observation will be removed. After that, the minimum value rises to 21, which is above the minimum age of loan eligibility and is therefore deemed satisfactory. Commonly, credit risk data sets contain illogical or erroneous observations. As a result, data quality inspection is an essential part of the modeling process (Leung *et al.* 2008).

Moving on to the number of dependents, the maximum is twenty, which also appears unrealistic, especially considering that the 75th percentile is one. Nevertheless, following the reasoning from the previous paragraph, the observations will not be trimmed.

An additional variable of interest is the income. The distribution of the variable

is heavily right-skewed, meaning that the data set contains several abnormally wealthy individuals. While the raw skewed variable may distort the results since it appears to be far from the normal distribution, the binning algorithm may group together income groups with similar characteristics and improve the classification process. It might be argued that above a certain income level, the effect on the probability of default becomes constant since moderately rich people are not substantially more likely to default than extremely rich individuals.

Four variables are present in the data set to capture the credit information about a customer. Firstly, the debt ratio indicates the total monthly expenses, including debt payments and living costs, as a percentage of monthly income. Notably, approximately 23.42% of observations attain a value over 1, meaning that many clients have higher expenses than income. As a result, the variable's distribution is also heavily skewed. Secondly, the feature capturing credit card utilization has similar characteristics. However, in this case, only approximately 2.21% of observations surpass the value of 1, signaling higher credit card debt than the limit. Lastly, two additional variables represent the number of open credit lines and the number of mortgages, respectively. The distributions of the variables do not appear to be substantially skewed since the mean and median attain similar values. An average customer in the data set has eight open lines of credit and one mortgage.

The last group of features records the customers' historical delinquency. All variables show the number of times a customer was delinquent for a specific number of days. Particularly, between 30 and 59 days, between 60 and 89, and lastly, more than 90 days. As expected, most observations in the data set attain a value of 0. The utilization of these kinds of variables should be approached with caution since they are very closely related to the target variable. As a result, they tend to explain most of the variation during training but may cause poor performance on the testing set. In addition, for new applications, historical delinquency data may not be available. Furthermore, considering that they carry very similar type of information, they are very likely highly correlated. The measures employed to mitigate multicollinearity are discussed in Section 4.1.

Table 3.1: Give Me Some Credit data set - summary

Statistic	n	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Age	150,000	52.295	14.772	0	41	52	63	109
NumberOfDependents	146,076	0.757	1.115	0	0	0	1	20
MonthlyIncome	120,269	6,670.221	14,384.670	0	3,400	5,400	8,249	3,008,750
DebtRatio	150,000	353.005	2,037.819	0.000	0.175	0.367	0.868	329,664.000
NumberOfOpenCreditLinesAndLoans	150,000	8.453	5.146	0	5	8	11	58
NumberRealEstateLoansOrLines	150,000	1.018	1.130	0	0	1	2	54
RevolvingUtilizationOfUnsecuredLines	150,000	6.048	249.755	0.000	0.030	0.154	0.559	50,708.000
NumberOfTime30-59DaysPastDueNotWorse	150,000	0.421	4.193	0	0	0	0	98
NumberOfTime60-89DaysPastDueNotWorse	150,000	0.240	4.155	0	0	0	0	98
NumberOfTimes90DaysLate	150,000	0.266	4.169	0	0	0	0	98
SeriousDlqin2yrs	150,000	0.067	0.250	0	0	0	0	1

Source: Cukierski (2011)

3.2 Home Credit Default Risk data set

The second Kaggle's data set called Home Credit Default Risk (HCDR) was also published as a part of a competition (Anna Montoya 2018). The competition was announced by Home Credit, a Czech Republic-based company providing consumer loans. Its goal was to improve upon the company's currently employed credit risk models.

The HCDR data set is substantially larger than the previous one and is the most sizeable data set utilized in this thesis. It contains 307 511 observations and 121 variables. Out of all features, 68 are numerical and 52 categorical.[2] However, some of the numerical variables are simple transformations[3] of a single variable, and thus are likely highly correlated among each other and will have to be disregarded. Table B.1 shows the summary statistics of selected numerical variables. Given the large dimensions of the data, only one transformation of each numerical variable is presented[4] and categorical features are not shown for the sake of conciseness.

A detailed discussion of the summary statistics will not be provided for the sake of brevity. However, a comparison with the previous data set might be of interest. For example, after normalizing the age variable[5] from days to years, it can be seen that the value of mean and median oscillates around 43, which is substantially lower than in the previous data set. Overall, the customers represent a younger part of the population. Moreover, as in the previous case, the dependent variable is highly imbalanced. Approximately

[2]Excluding the dependent variable.
[3]Mean average, mode, median.
[4]The median and mode transformation are not shown.
[5]The name of the age variable in the data set is DAYS_BIRTH.

8.07% of observations are flagged as defaulted. Furthermore, unlike in the previous data set, several variables have a significant portion of missing values. This issue is addressed in Section 4.1.

3.3 Credit Approval data set

The second group of data sets originates from the UCI Machine Learning Repository (UCI MLR). The first data set from this group is the Credit Approval data set (Quinlan 2017),[6] which contains anonymized information about credit card applications.

Since the variables' names are unavailable, the summary statistics in Table B.2 do not convey useful information. Nevertheless, it is essential to note that unlike in the previous data sets, the current dependent variable is not imbalanced. Approximately 44.5% of the observations are flagged as delinquent. Apart from the target, the data set contains 690 observations, six numerical variables, and nine categorical features. Again, for conciseness, categorical data are not presented. With regard to missing values, two numerical variables seem to have a small portion of unavailable observations.

3.4 Default of Credit Card Clients in Taiwan data set

An additional data set from the UCI MLR is the Default of Credit Card Clients in Taiwan (DCCCT). The data set was initially published in Yeh & hui Lien (2009) and later was made available in the repository.[7] It contains information about customers' default payments in Taiwan. The number of observations is 30 000, the number of numerical features is 20, and the number of categorical variables is 3. The dependent variable is imbalanced since only approximately 22.12% of observations are flagged as defaulted. The data do not contain any missing values.

The set of independent variables is very similar to the previous data sets. The summary statistics are available in Table B.3. Interestingly, the average age of a customer is approximately 35, which is substantially lower than in the first

[6]Further referred to as CA.
[7]DCCCT (2016)

data set. Most of the observations (around 60%) represent women and less than half of the customers in the data set are married.

The remaining variables can be divided into three sets. The first one captures the customers' payment diligence. Each of the six variables[8] records the repayment status in a given month with values ranging from -1 (repaid on time), 1 (1-month delay), up to 9 (9 and more months delay). Each variable represents a different month. Given the similar nature of the variables, they are expected to have a very high pairwise correlation with each other.

The second group records how much money a given customer owes to the credit card company at the end of the month. Similarly to the previous group, each of the six variables stands for a different month. Finally, the third set of features indicates the amount of money the customer actually paid each month.

3.5 South German Credit Card data set

Finally, the last data set from the UCI MLR and the last one utilized in this thesis is the South German Credit Card data set (SGCC). There are two versions available in the repository. Grömping (2019) provides the latest one which is, as the authors claim, stripped of coding errors of the original data set.[9]

The data set consists of 1000 observations and 20 independent variables, of which three are numerical and 17 are categorical. The dependent variable is imbalanced, with 30% of the observations being flagged as defaulted. However, as the authors claim, the provided sample was obtained from the population using stratified sampling, and therefore, the minority class is oversampled. As for the independent variables, the average age is approximately 36 years. In addition, one of the variables captures the duration of the credit, which averages to about 21 months. The last numerical variable shows the amount of credit loaned. The summary statistics are available in Table B.4.

3.6 Summary

The summary of the main characteristics of each data set is presented in Table 3.2. It can be seen that the data sets obtained from Kaggle are notably more

[8] *PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6.*
[9] SGCC (2019)

sizable. In addition, the class distribution in the UCI MLR data sets is less imbalanced. Consequently, the Kaggle data sets are likely more representative of the credit risk industry. Nevertheless, the wide utilization of the UCI MLR in the academic literature supports the validity of the provided samples.

Table 3.2: Data sets summary (raw)

Data set	n	# of numerical features	# of categorical features	% of defaults
Give Me Some Credit	150 000	10	0	6.68
Home Credit Default Risk	307 511	69	52	8.07
Credit Approval	690	6	9	44.5
Default of Credit Card Clients in Taiwan	30 000	20	3	22.12
South German Credit Card Data Set	1 000	3	17	30.00

Source: Author's computations

Chapter 4

Methodology

The current chapter outlines the methodology utilized in the empirical analysis. The initial subsection discusses the data preprocessing steps taken prior to estimation, including the description of the binning algorithm. The second subsection describes the considered estimation methods. Moreover, the third part is devoted to the methodology regarding classifier performance and comparison, including a thorough discussion of the employed evaluation metrics. Finally, the last subsection presents an overview of the supplementary analyses.

4.1 Data preprocessing

As indicated in Section 3, several data issues must be addressed. Firstly, three out of the five utilized data sets contain missing values. While the binning algorithm can natively handle missing data by assigning them to a separate bin, unavailable observations in the raw form must be replaced or disregarded. The latter option is selected to maximize the models' comparability. As a result, any potential performance improvement caused by the binning algorithm cannot be attributed to the treatment of missing values. An alternative approach will be taken in the supplementary analyses to investigate the impact on the results. In addition, since, for example, the HCDR data set comprises many variables, some of which have a significantly low fill rate, the omission of missing values while retaining all the features would result in a low number of total observations. Consequently, for all data sets, only variables with at least 80% of non-missing observations are preserved for the analysis.

Secondly, the main aim of this thesis is to investigate the binning of numerical

variables. However, four of the utilized data sets contain categorical features as well. Since the relationship among the numerical and categorical variables might affect performance, they will be retained following Raymaekers *et al.* (2022) and Sharma (2011). However, for conciseness and to reduce the data's dimensionality, only categorical variables with at most two categories will be considered for the analysis. These variables will be dummy encoded.

Thirdly, given the large number of variables in some datasets, adverse statistical and numerical consequences of multicollinearity may arise if not appropriately addressed (Alin 2010). Therefore, following Lessmann *et al.* (2015), highly correlated features will be disregarded before estimation. The Pearson correlation coefficient will measure the pairwise relationship between the variables. Since some variables, including the dependent variable, are categorical, the point biserial correlation will assess the relationship between the numerical and dichotomous variables (Kornbrot 2005). The point biserial correlation is numerically equivalent to the Pearson correlation. For two highly correlated variables, the one with the higher correlation with the dependent variable will be retained. In addition, numerical features take precedence over categorical variables since they are the main focus of this thesis. The threshold for removal is set quite conservatively to 0.75. As a further rigorous verification of this approach, the Variance Inflation Factor (VIF) will be calculated for the remaining features. The VIF quantifies how much the behavior of an independent variable is affected by its interaction with the other explanatory features. For each variable, it is computed using the following formula:

$$VIF_j = \frac{1}{1 - R_j^2}, \tag{4.1}$$

where R_j^2 is the R-squared of a regression where the j-th independent variable is regressed on the remaining explanatory variables. Unlike pairwise correlation, the VIF can detect more complex relationships among multiple variables. As Marcoulides & Raykov (2018) suggest, values of VIF above ten may indicate an alarming degree of multicollinearity. Consequently, variables with values above this threshold will be disregarded.

After the aforementioned steps, one last data preprocessing stage remains, the binning of numerical variables. As discussed in Section 2, binning algorithms can be divided into multiple categories. For the purposes of this thesis, a supervised statistics-based algorithm will be utilized. As the literature suggests,

supervised learning algorithms usually tend to achieve better performance. In addition, since the class label is available during training, it is reasonable to exploit the information it carries. Within the supervised algorithms category, the statistics-based procedures appear superior to their entropy-based and error-based counterparts.[1] Furthermore, to the author's best knowledge, the mathematical programming formulation of the optimal binning problem introduced by Navas-Palencia (2020)[2] was not yet evaluated in the literature in the credit risk context, even though it is widely utilized in practice.[3]

A detailed description of the algorithm is beyond the scope of this text and is available in Navas-Palencia (2020). Nevertheless, a short overview will be provided. The algorithm consists of two steps. Firstly, the numerical variable is divided into n pre-bins. The split points are selected using the Classification and Regression Tree (CART) (Breiman *et al.* 1984), which iteratively searches for optimal split points until a maximum number of pre-bins is reached, or no additional splitting is feasible. The maximum number of pre-bins is arbitrary and, for the purpose of this thesis, will be set to twenty, as suggested by the authors of the algorithm. Too many bins may hinder the algorithm's ability to generalize the data, ultimately defeating the purpose of binning. On the other hand, a small number of bins may result in a loss of too much information. The pre-binning step substantially reduces the search space and thus significantly decreases the complexity of the optimization problem. Secondly, the optimal bins are found by iteratively merging adjacent pre-bins to maximize the given variable's Information Value (IV). IV quantifies the predictive power of an independent variable with respect to a binary dependent variable and is a widely utilized measure in credit risk analysis (Zeng 2013). Denoting g_i and b_i the number of observations for bin i where the dependent variable is equal to zero and one, respectively, the IV for a given variable is calculated as

$$IV = \sum_{i=1}^{n} \left(\frac{g_i}{g} - \frac{b_i}{b} \right) WoE_i, \tag{4.2}$$

where

$$WoE_i = \ln \left(\frac{\frac{g_i}{g}}{\frac{b_i}{b}} \right) \tag{4.3}$$

[1] Refer to Section 2.2 for a detailed review.

[2] The implementation of the binning algorithm is available in the python OptBinning library (Navas-Palencia 2023) and will be utilized for the purposes of this thesis.

[3] The OptBinning library was subjected to more than 200 000 downloads in June 2023 (Flynn 2023).

is the Weight of Evidence (WoE), which evaluates the ability of each bin i to differentiate between the two classes, g and b are the total number of observations where the dependent variable is equal to zero and one, respectively, and n is the number of bins.

Some additional constraints can be defined for the optimization problem. For example, since the models are required to be interpretable in the credit risk industry, it is essential to impose a monotonicity constraint on the bins (Raymaekers *et al.* 2022). In this way, a straightforward relationship with the dependent variable is secured. Let $ER_i = \frac{b_i}{b_i+e_i}$ be the Event Rate (ER). Then, the monotonicity constraint translates to consecutive bins having either ascending or descending value of ER. Moreover, the Z-test is utilized to assess whether the difference in ERs between adjacent bins is statistically significant. A constraint is set for the p-value to be lower or equal to 0.1 for the differences to be significant at the 10% significance level. Furthermore, as Siddiqi (2012) indicates, an appropriate binning algorithm needs to satisfy the following requirements:

- Missing observations have a dedicated bin
- The number of observations in each bin is larger or equal to 5% of all observations
- In every bin, there is at least one observation from each class

The first condition is relatively straightforward and constitutes one of the main strengths of binning algorithms in general. Nevertheless, it is not directly applicable to the primary analysis of this thesis since missing observations are disregarded. The second condition attempts to secure the representativeness of each sample bin of the total population. Without it, the algorithm could easily lead to overfitting. The third condition ensures the computational feasibility of the measures defined above.

The last step of data preprocessing is encoding the binned variables. One option is to use dummy encoding, which would entail creating a new variable for each bin. However, this approach substantially increases the dimensionality of the data and, in addition, the literature suggests that categorical variables, in general, tend to benefit from other forms of encoding, especially those that utilize the values of the dependent variable (De La Bourdonnaye & Daniel 2021; Potdar *et al.* 2017). As a result, following the standard credit scoring practice (Leung *et al.* 2008), the data will be encoded using WoE as defined in

(4.3). The WoE transformation does not inflate the data's dimensionality and utilizes the class label information. In addition, combined with the monotonicity constraint, the resulting feature has an interpretable relationship with the dependent variable.

4.2 Estimation methods

Several estimation methods will be utilized to assess the impact of binning on performance since, as discussed in Section 2, the effect does not appear to be homogeneous across algorithms. Firstly, the logistic regression represents the industry standard in credit risk modeling (Leung *et al.* 2008; Raymaekers *et al.* 2022) and, as such, should be considered for the analysis. In addition, since WoE is, in essence, a logit transformation, the methods are closely linked. Even though the logistic regression is considered interpretable, it is still common practice to utilize binning and a subsequent WoE transformation (Leung *et al.* 2008).

The logistic regression estimates a model of the following form (Wooldridge 2013):

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j=1}^{M} \beta_{ij} x_{ij} + \epsilon_i, \tag{4.4}$$

where β_0 is the intercept, $\beta_1, \ldots, \beta_M$ are the coefficients, M is the number of independent variables, ϵ_i represents the error term, and p_i is the probability of $y_i = 1$. The estimates of the coefficients are obtained using the Maximum Likelihood Estimator (MLE) such that the log-likelihood function is maximized.

For regression analysis in general, with an increasing number of variables, the chance of overfitting grows (Hawkins 2004). And since some of the utilized data sets contain many features, this potential issue must be addressed. Given the quantitative nature of the analysis, any form of advanced feature selection is infeasible. In addition, the comparability between the models with binned variables and with raw variables must be preserved, which could not be the case if the chosen variables differed across the models. Therefore, an alternative approach will be taken through regularization.

The most common types of regularization are the ℓ_1 (Lasso) and ℓ_2 (Ridge). These algorithms introduce a penalty for the size of coefficients into the loss function (Friedman *et al.* 2010). The penalty terms have the following form:

- ℓ_1: $\lambda_1 \sum_{j=1}^{M} |\beta_j|$
- ℓ_2: $\lambda_2 \sum_{j=1}^{M} \beta_j^2$,

where λ_1 and λ_2 control the regularization strength. Unlike Ridge, the Lasso regularization method allows for the coefficients to shrink to the value of zero, and as a result, effectively acts as a feature selection method (Bühlmann 2011). For the purposes of the analysis, both methods will be considered along with several values of λ_1 and λ_2, which will be subjected to hyperparameter boosting.[4] For completeness and also to accommodate data sets with a sufficiently low number of observations, the logistic regression without regularization will be considered as well.

Moving on to the machine learning field, the second employed estimation method is the Classification and Regression Tree (CART) introduced by Breiman *et al.* (1984). As discussed in Section 2, the impact of binning on tree-based methods appears to be ambiguous and therefore requires further investigation. In addition, the CART represents an interpretable machine learning model, making it an eligible candidate for credit risk modeling purposes (Khandani *et al.* 2010). The algorithm searches through all possible splits among all variables and selects the one that decreases a given criterion the most. The training data are partitioned according to the split, and a new search is initiated for each partition. Each partition is called a node, and a node that is not split further is called a leaf.

The commonly used criteria are either the Gini index or the Entropy criterion. The former is defined for binary classification for node T as

$$Gini(T) = N_T \sum_{k=0}^{1} p_T(k)[1 - p_T(k)] = 2N_T p_T(0)p_T(1), \tag{4.5}$$

where N_T is the number of observations for node T and $p_T(k)$ is the probability that an observation belonging to node T is equal to class k. On the other hand, the entropy criterion is defined as

$$Entropy(T) = -N_T \sum_{k=0}^{1} p_T(k) \log p_T(k), \tag{4.6}$$

Since neither of the criteria is universally superior, both will be considered for

[4] A discussion of hyperparameter boosting is available in Section 4.3.

estimation. Moreover, similarly to logistic regression, there are various methods to prevent overfitting a decision tree. For example, a threshold for the minimum number of observations demanded to perform a split can be set, the number of total leaves of the tree can be regulated, or the tree is allowed to only reach a certain depth. Alternatively, the tree can be pruned post-estimation by replacing some of the subtrees with leaves. For the hyperparameter optimization, two ways of regularization will be considered, maximum depth and maximum number of leaves. The former impacts the algorithm only by stopping when a certain depth is reached. The latter on the other hand, searches during each iteration for the best possible split among all existing nodes, which usually results in an assymetric tree structure.

An additional tree-based algorithm that will be utilized in the analysis is the Random Forest classifier.[5] As Lessmann *et al.* (2015) shows, this homogeneous ensemble achieves one of the best performances in credit risk modeling. With respect to pre-binning the variables, as discussed in Section 2, it appears to have an adverse effect on the classifier. However, since it belongs to the group of so-called "black-box" models, it is necessary to introduce some form of interpretability into the model to ensure its feasibility for credit risk modeling.

The Random Forest is an ensemble of decision trees whose individual predictions are averaged to form the final prediction. Each subtree is built in the following manner. The training data set is sampled with replacement, with the sample size being optional but usually set to the original data size. In addition, only a random subset of features is considered when looking for the best split. Apart from the optimizable parameters of the individual trees (maximum depth and maximum number of leaves), the RF classifier has an additional parameter, which is the number of trees. As a result, all the tree parameters will be candidates for hyperparameter tuning. While a too complex decision tree may overfit and provide unreliable predictions, the strength of the RF is in the ability to combine the predictions of many possibly overfitted trees, which results in superior performance (Ali *et al.* 2012).

Remaining in the area of black-box models but moving away from tree-based algorithms, the next highly popular machine learning method that will be considered is the neural network (NN). As shown in Section 2, the literature suggests that neural networks appear to induce a performance improvement compared

[5]Introduced by Breiman (2001).

to simpler methods such as logistic regression. On the other hand, exceedingly complicated deep learning models do not seem to outperform their single-layer counterparts. As a result, only a one-layer feedforward artificial neural network will be employed. Following Heaton (2008), the number of nodes of the hidden layer will be set to 2/3 of the number of independent variables. This rule of thumb is expected to prevent both underfitting and overfitting of the neural network. However, to ensure the optimal parameter setting, two additional options for the number of nodes will be considered: $\frac{K}{2}$ and K, where K is the number of independent variables. Moreover, to avoid the issue of overfitting even further, the ℓ_2 regularization will be employed to shrink the coefficients. The regularization parameter will be subjected to hyperparameter boosting. With regard to the activation function, three candidates will be considered. Firstly, the Rectified Linear Unit (ReLU) activation function, which is defined as $ReLU(x) = max(x, 0)$. The ReLU appears to be superior in performance to other common activation functions (Bircanoğlu & Arıca 2018; Krizhevsky *et al.* 2012), it does not suffer from the problem of vanishing gradients, and it secures high computational efficiency. Moreover, its formulation allows for disabling nodes by outputting the value of zero. However, for completeness, the well-know sigmoid and tanh activation functions will be considered as well. The former is defined as $\sigma(x) = \frac{1}{1+e^{-x}}$, while the latter as $tanh(x) = 2\sigma(2x) - 1$. As can be seen, the tanh activation function is a transformation of the sigmoid, which results in a symmetrical function with the derivation in zero being equal to one.

The last model included in the analysis is the Naïve Bayes classifier. This probabilistic model based on the Bayes theorem seems to greatly benefit from the binning of numerical features.[6] Notably, the Bernoulli Naïve Bayes is only able to handle binary variables, and thus the categorization of numerical variables is necessary for its feasibility. However, since, as introduced above, the data will be WoE encoded, the Gaussian Naïve Bayes (GaussNB) will be utilized instead. Apart from the essential assumption that the explanatory variables are independent of each other, the GaussNB algorithm also assumes that they are normally distributed. This constitutes a very strong assumption but is often adopted since the model seems to perform quite well even if it is violated (Soria *et al.* 2011).

[6] See Section 2.

4.3 Evaluation

An essential part of the empirical analysis is evaluating the models' performance and their subsequent comparison. As discussed in Section 2, the existing studies often suffer from methodological deficiencies, such as using inappropriate evaluation metrics or deriving conclusions based on comparing only a single performance indicator. Following Lessmann *et al.* (2015), these shortcomings will be addressed by employing six evaluation metrics, each capturing a different aspect of the model's performance. The metrics can be divided into three distinct categories.

Firstly, a fundamental characteristic of the model is its ability to differentiate between creditworthy and unreliable customers. The Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) is often utilized to quantify this ability. Given a threshold above which predicted probabilities are classified as defaulted, the ROC curve represents values of the True Positive Rate (TPR) and False Positive Rate (FPR) for all possible thresholds. The TPR is calculated as $\frac{TP}{TP+FN}$, where TP are observations correctly classified as defaulted and FN are observations falsely classified as non-defaulted. The FPR is equal to $1 - TPR$. The area under the ROC curve represents the probability that a default observation will receive a higher probability prediction than a non-defaulted observation. As a result, the AUC measure attains values between 0 and 1, with 1 indicating a perfect model.

Despite its wide popularity, the AUC measure suffers from several shortcomings (Hand 2009). Most importantly, comparing two classifiers can be misleading if their ROC curves cross. Therefore, Hand (2009) proposed a more appropriate alternative that considers a distribution of misclassification costs depending on the classification problem rather than the classifier. As a result, to ensure the robustness of the results, the H-measure will be used to quantify the classifiers' performance.

Another deficiency of the AUC metric is its focus on the entire distribution of the predicted probabilities. In the credit risk context, only observations with probabilities below a certain threshold might be accepted. Therefore, the investigation of the lower tail of the distribution is crucial. To address this issue Pundir & Seshadri (2012) proposed the usage of the Partial Gini Index

(PGI), which is calculated as

$$PGI = 2 * AUC(\beta) - 1 \tag{4.7}$$

The PGI rescales the AUC metric such that it has more intuitive properties. It ranges from -1 to 1, with 1 being a perfect model. In addition, the value of 0 corresponds to an AUC value of 0.5 which represents a 50% percent probability of assigning a higher probability to a default observation than to a non-default one. A model with such a value of AUC is equivalent to a coin toss and thus is deemed underperforming. In addition, values of AUC, which are lower than 0.5, correspond to negative values of PGI, which is likely more intuitive to the human eye. The term Partial stems from the fact that the AUC value in (4.7) is calculated only for observations with probabilities below a certain threshold β. Following Lessmann *et al.* (2015), β is set to 0.4.

The second group of metrics evaluates the accuracy of the probability predictions. The only member of this group is the Brier score[7] (BS), which is calculated as the mean-squared error between the predicted probabilities and the dependent variable (Hernandez-Orallo *et al.* 2011). As a result, unlike for the remaining metrics, the lower the BS, the better performance of the model. The BS was not utilized in any of the studies reviewed in Section 2 for assessing classifier performance regarding the binning of numerical variables and, therefore, might provide valuable insight into this issue. The calibration of the estimated probabilities is an inseparable part of the credit modeling process, and thus should be adequately evaluated.

Lastly, the third group of metrics captures the correctness of categorical predictions. The first member of this group is the widely utilized Kolmogorov-Smirnov statistic (KS).[8] It is obtained using the KS test, which quantifies the similarity between two distributions. In the credit risk modeling context, it is utilized to assess the distance between the distribution of estimated probabilities of defaulted observations and the distribution of estimated probabilities of non-defaulted observations. The higher the value of the KS statistic, the lower the p-value for the null hypothesis that the two distributions are identical. Therefore, higher KS statistic indicates better model performance.

The last employed evaluation metric is the F2-score. Generally, the correctness

[7]Brier (1950).
[8]Smirnov (1939).

of categorical predictions is often evaluated using accuracy.[9] However, given the imbalance of credit risk data sets, accuracy is not a suitable metric since it assigns equal weights to the majority class and the minority class (Branco *et al.* 2016). A model predicting that all customers are expected not to default would likely achieve an acceptable level of accuracy; however, it would result in significant problems for the bank employing the model. The F-beta score is often utilized to address this issue since it calculates a harmonic mean between the TPR[10] and the precision.[11] The formula for the F-beta score is defined as follows

$$F_\beta = \frac{(1+\beta^2) \times precision \times TPR}{\beta^2 \times precision + TPR} \tag{4.8}$$

As can be seen, TPR is β times more important in the formula. Since it is desirable to achieve as low Type II error as possible in the credit risk industry, β will be set to 2 to prioritize the TPR.[12]

Unlike all the previous metrics, the F2-score requires class membership predictions. These are derived from the estimated probabilities by setting an appropriate threshold. Following Lessmann *et al.* (2015), the threshold is selected such that the resulting percentage of defaults in the testing set is the same as that of the training set. Using prior default probability is expected to alleviate further any potential bias caused by class imbalance.

All the metrics introduced above will be calculated using the testing set. Data partitioning is a common practice in machine learning and for predictive models in general. Rather than investigating the performance of a classifier on the data it was trained, the performance on unseen data is evaluated. Therefore, all the results presented in Section 5 are obtained using the testing set. Each data set is split into a training set comprising 80% of the observations and a testing set containing the remaining 20%.

In addition to partitioning the data into a training and testing set, a 3-fold cross-validation during training is employed to find the optimal parameters for each model. As discussed above, all the utilized estimation methods apart from the Naïve Bayes classifier require a parameter setting before estimation. Therefore, since the model using raw variables may require a different param-

[9] Accuracy is defined as the percentage of correctly classified observations.
[10] Also sometimes referred to as recall or sensitivity.
[11] $precision = \frac{TP}{TP+FP}$
[12] Type II error occurs when an ineligible customer is predicted to be a good customer.

eter setting for optimal performance than the one using binned variables, a grid search is performed. For each model, the training set is split into three folds. The model is estimated on two folds and evaluated on the third one. Therefore, the estimation and evaluation take place three times for each parameter setting, and the mean average of the three performances is calculated. The optimal binning algorithm is applied repeatedly during each iteration to prevent data leakage during cross-validation. The common practice is to utilize the AUC metric for hyperparameter boosting. The best parameter setting according to the average AUC is then used to estimate the final model on the entire training set and subsequently evaluated on the testing set. Table B.5 shows the list of parameters entering the search for each model.

The last part of the evaluation process is the statistical test of the difference in performance between the model utilizing binned variables and the model estimated on raw features. While a comparison of the metrics calculated on the testing set provides a certain indication of the differences, it does not represent a rigorous test. Therefore, the null hypothesis that the model with raw variables performs better or worse than model with binned variables will be tested to obtain some significance level of the result. With this approach, rejecting the null hypothesis in favor of the alternative results in the conclusion that the binned model performs better.

Even though five data sets are employed, the total number of five values of each metric for each model is not sufficient to perform a reliable statistical test. One feasible approach would be the paired bootstrap test.[13] However, as Good (2004) argues, bootstrap testing is more appropriate to estimate the confidence intervals of a metric. When it comes to statistical hypotheses testing, the value returned by the paired bootstrap test cannot be interpreted as the p-value since the distribution of the test statistic was obtained under the true distribution and not under the null hypothesis.

An alternative and more appropriate approach constitutes the permutation test.[14] The intuition behind the test is that if the models predict equally well, it should not be of importance from which one we obtain the predictions. As a result, if all possible alternatives of prediction origins could be considered, then the distribution of performances under the null hypothesis would have been obtained. However, enumerating every alternative is infeasible, and thus a Monte

[13] Konietschke & Pauly (2014).
[14] See, for example, Collingridge (2013).

Carlo simulation is often performed (Collingridge 2013). The permutation test is then employed in the following manner. Given two models whose performances are to be compared, for each observation from the test set, one model is randomly selected and used for prediction. Subsequently, the performance of the resulting set of predictions is evaluated. This process is repeated R times, and each time the resulting value of the metric is stored, forming the distribution under the null hypothesis. To obtain the p-value, the ratio of performances greater or equal to the performance of the model in question is calculated. As Collingridge (2013) indicates, the resulting p-value is only an approximation of the true p-value since not all alternatives were considered. However, since the goal of the test is to reject the null hypothesis at some significance level α, then if the estimated p-value is smaller than α, the probability that the true p-value is below α as well converges to zero with the number of repetitions.

4.4 Supplementary analyses

As outlined in Section 4.1, the missing values are disregarded for the primary analysis. However, with this approach, the main strength of the binning algorithm is not exploited. And since credit risk data sets often suffer from missing values (Leung *et al.* 2008), their appropriate treatment may significantly affect estimation outcomes. As a result, for the purpose of the supplementary analysis, missing values will be replaced by the means, medians, or modes of the respective columns. This approach follows Lessmann *et al.* (2015). For categorical and numerical variables with a small number of unique values, mode imputation will be used. For continuous variables with a skewed distribution, missing values will be replaced with median to avoid biasing the results (Jadhav *et al.* 2019). For a non-skewed variable, the value of its skewness should be between -2 and 2 (Hair *et al.* 2022).[15] These variables will be imputed using the mean. The resulting imputed data set will be used for the estimation of the model with raw variables.[16] The results will be compared to the model utilizing binning and assigning missing observations to a separate bin. The resulting bin has its own WoE value, which is allocated to all missing observations for a given variable during the encoding process. Only three of the available data

[15]Skewness for a variable X is calculated as $\frac{E[(X-\mu)^3]}{\sigma^3}$, where μ and σ are the mean and standard deviation of X, respectively (Kokoska & Zwillinger 2000).

[16]The imputation is performed after data partitioning to avoid leakage into the test set.

sets will be employed for the current analysis since the remaining two do not suffer from missing observations.

An additional widely recognized benefit of the binning transformation is its effective treatment of outlying observations. Since, for example, the performance of logistic regression may be hindered by the presence of extreme outliers (Jennings 1986), the robustness of the results after the exclusion of irregular observations will be investigated. As a result, for estimation purposes, all values above the 99th percentile and below the 1st percentile of each variable within each data set will be disregarded.

A third supplementary analysis will be performed to inspect the effect of including categorical variables in the estimation process. For this purpose, the same steps will be taken as in the main analysis, but only the numerical variables will be utilized. Considering that the GMSC data set does not contain categorical features, the results from the primary analysis are identical to the results of the supplementary analysis.

The fourth auxiliary analysis concerns the transformation of the binned variables. As mentioned above, the literature suggests that using dummy encoding appears inferior to other advanced forms of encoding. However, inspecting whether this approach still surpasses the usage of raw non-binned numerical variables might be interesting. Since the dummy variable transformation results in the formation of many features, the analysis is feasible only for sizeable data sets. In addition, given the binary nature of the explanatory variables, the Bernoulli Naïve Bayes becomes feasible and will be used instead of the Gaussian.

Lastly, the current thesis employs a quantitative approach to evaluate the effect of binning on the performance of credit risk models. However, as Leung *et al.* (2008) indicates, the process of building a scorecard is often of a more qualitative nature. In addition, the aim of this thesis thus far was not to find a model with optimal performance, which is one of the main goals of credit risk analyses besides interpretability. As a result, the last supplementary analysis will employ a more qualitative approach to model estimation, including variable transformations and sophisticated variable selection. The preprocessing of the raw data will include the logarithmic transformation of non-normal variables and the trimming of outliers. Since such an analysis for all estimation methods and data sets would be infeasible, only the logistic regression with

the GMSC data set will be considered. The logistic regression represents the industry standard, and the data set has a sufficiently small number of variables while bearing a reasonable number of observations. Two models will be developed, one using raw (possibly transformed) variables and one with binned features. Subsequently, their performance will be evaluated. Missing values will be disregarded since they have a devoted separate analysis.

Chapter 5

Empirical analysis

The current chapter provides a description of the empirical analysis. The initial subsection is concerned with the results of data preprocessing and variable selection. The second part illustrates the utilization of the binning algorithm. The following five subsections discuss the results of the main analysis for each considered estimation method. The subsequent subsection summarizes the results of the previous analyses. Finally, the last four subsections address the robustness checks.

5.1 Variable selection

As outlined in Section 4.1, several data preprocessing steps were performed before estimation. Firstly, the elimination of variables with insufficient fill rate affected only the HCDR data set, where 50 variables had to be disregarded since they contained less than 80% observations. Secondly, categorical features with more than two categories were removed. This particular step impacted three out of the five data sets and eliminated 29 variables in total.

Thirdly, in an attempt to alleviate potential multicollinearity, the available variables were filtered using the Pearson correlation coefficient. For demonstration purposes, Table B.6 presents the correlation matrix of the GMSC data set. As indicated in Section 3.1, the features capturing customers' delinquency are very severely correlated with each other.[1] Therefore, only the one with the greatest correlation with the dependent variable can be retained. Expectedly, these variables also have the strongest pairwise relationship with the default

[1] The correlation is as high as 0.99.

indicator. It stands to reason that more delinquent customers are more likely to default. An additional variable with a high correlation with the target is *Age*. The sign of the coefficient is negative, suggesting that older people are expected to be less prone to default. A moderately high correlation is also between the features recording the number of loans and the number of mortgages. Nevertheless, the degree of correlation is not large enough to cause any concern. With respect to the remaining features, none of the pairwise correlations exceed the defined threshold. Moreover, the VIF analysis did not result in any further disqualification. The highest calculated value is 4.71 for the variable capturing the number of loans.

For the sake of brevity, the correlation matrices for the remaining data sets are not disclosed.[2] Table 5.1 shows the final statistics for all data sets after data preprocessing and variable selection. As can be seen, the correlation analysis eliminated eight features from the HCDR data set. Furthermore, five additional variables were eliminated due to the high values of the VIF. Given the large number of features, the VIF likely captured some degree of interdependence among multiple variables, which cannot be discovered by mere pairwise correlation. The DCCCT data set contains two groups of features that share large similarities, and as a result, eight variables in total had to be disregarded. The correlation and VIF analyses did not affect the remaining two data sets.

The number of observations in Table 5.1 also reflects the elimination of missing values. As can be seen, while three data sets are unaffected, the HCDR and GMSC data sets suffer from a moderate loss of observations. Nevertheless, for the purpose of the main analysis, the data is assumed to be missing at random. In addition, any potential bias caused by the removal of missing values is expected to affect both models; therefore, the results should remain comparable. Furthermore, the effect of missing values is studied in Section 5.9.

Table 5.1: Data sets summary (final form)

Data set	n	# of numerical features	# of categorical features	% of defaults
Give Me Some Credit	120 268	8	0	6.95
Home Credit	245 155	18	29	7.78
Credit Approval	654	6	4	54.74
Default of Credit Card Clients in Taiwan	30 000	12	1	22.12
South German Credit Card	1 000	3	3	30.00

Source: Author's computations

[2]They are available upon request.

5.2 Binning

Considering that one of the main reasons for utilizing the binning algorithm is to increase the interpretability of complex models, the current subsection demonstrates the outcomes of the binning transformation and the resulting relationship of the transformed features with the dependent variable. Since going through all variables in all data sets is infeasible, only several examples will be discussed.[3] Given its moderate number of variables, the GMSC data set will be utilized for the demonstration.

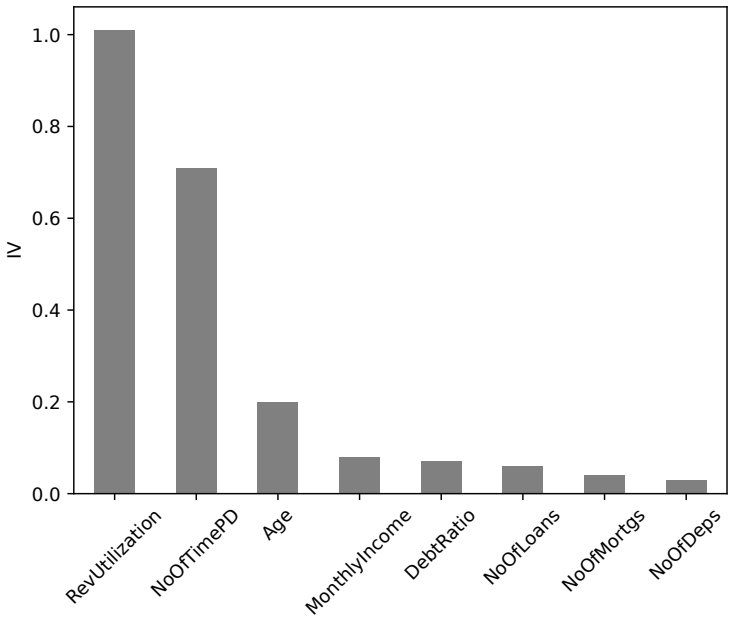
As mentioned in Section 2, the binning transformation can also serve as a variable selection method. Figure 5.1 shows the values of Information Value for the variables in GMSC.[4] The highest value of IV, and thus the highest expected predictive power, is attained by the variable capturing the balance vs. credit limit ratio (*RevUtilization*). The runner-up is the feature recording the number of times a customer was 30-59 days past due (*NoOfTimePD*). As discussed in Section 3.1, variables indicating customer delinquency tend to constitute strong predictors. The remaining variables attain substantially lower values of IV, signaling possibly weak predictive power.

Table 5.2 shows the optimal bin decomposition of *RevUtilization*. The first column outlines the intervals for which the values of the variable are grouped together. As can be seen, all observations below 7% utilization belong to the first category, which also attains the lowest Event Rate. As a result of the imposed monotonicity constraint, with each subsequent group, the ER increases. This outcome appears intuitive since customers who draw only a small portion of their available credit are expected to be more likely to repay it successfully. On the other hand, the last bin, which realizes the highest ER, suggests that almost every fourth customer with utilization higher than 90% defaulted. Moreover, the seventh column of Table 5.2 shows the values of WoE assigned to all observations falling into the respective bins. Since the WoE values decrease with increasing utilization, the coefficient in the logistic regression for this variable would be expected to have a negative sign such that higher utilization results in an increased probability of default. As can be seen in Table B.7, this is indeed the case. Finally, the last column of the table records the values of IV for individual bins. The total IV of the variable is the sum of all individ-

[3] The results for any given variable are available upon request.

[4] Two variables were eliminated during correlation analysis and thus did not enter the binning process.

Figure 5.1: Information Values of binned variables in GMSC



Source: Author's computations

ual IVs. As can be seen, the combined IV of the first and last bin constitutes most of the IV for *RevUtilization*. These two bins also contain a large portion of total observations. As a result, a very low and very high utilization appears to be informative when predicting the probability that customer defaults.

Table 5.2: Binning of *RevUtilization*

Bin	Count	Count (%)	Non-event	Event	ER*	WoE**	IV***
(-inf, 0.07)	32859	34.15	32220	639	1.94	1.33	0.35
[0.07, 0.11)	8174	8.50	7985	189	2.31	1.15	0.07
[0.11, 0.22)	10564	10.98	10225	339	3.21	0.81	0.05
[0.22, 0.30)	6107	6.35	5863	244	4.00	0.59	0.02
[0.30, 0.39)	5791	6.02	5487	304	5.25	0.30	0.00
[0.39, 0.49)	5036	5.23	4723	313	6.22	0.12	0.00
[0.49, 0.70)	8205	8.53	7406	799	9.74	-0.37	0.01
[0.70, 0.90)	6905	7.18	5887	1018	14.74	-0.84	0.07
[0.90, inf)	12573	13.07	9729	2844	22.62	-1.36	0.43
Totals	96214	100.00	89525	6689	6.95		1.01

*Event Rate
**Weight of Evidence
***Information Value

Source: Author's computations

As an additional example, Table 5.3 presents the result of optimal binning for

NoOfTimePD, which has the second highest IV. As can be seen, three bins were found. The first bin covers all observations with a value of zero, which signals no recorded delinquency. Consequently, the first bin contains the majority of the observations.[5] The second bin captures customers, which were precisely once 30-59 days past due. It can be argued that the second bin represents moderate but not severe delinquency. The remaining customers are assigned to the last bin, which comprises the least observations but the highest Event Rate. Again, the ER is monotonic across bins, and given the decreasing values of WoE, the coefficient in the logistic regression would be expected to have a negative sign.[6] As for IV, the last bin constitutes more than 50% of the total IV of the variable.

Table 5.3: Binning of *NoOfTimePD*

Bin	Count	Count (%)	Non-event	Event	ER*	WoE**	IV***
(-inf, 0.50)	80002	83.15	76666	3336	4.17	0.54	0.19
[0.50, 1.50)	10768	11.19	9152	1616	15.01	-0.86	0.12
[1.50, inf)	5444	5.66	3707	1737	31.91	-1.84	0.40
Totals	96214	100.00	89525	6689	6.95		0.71

*Event Rate
**Weight of Evidence
***Information Value

Source: Author's computations

5.3 Logistic regression

Moving away from data preprocessing, the current subsection discusses the results of the analysis of the industry standard in credit scoring, the logistic regression. Table 5.4 presents the optimal parameters for each data set obtained via hyperparameter boosting. As can be seen, in all cases, the highest average performance was obtained with models using some form of regularization. The prevalent regularization method is ℓ_1, which, as discussed in Section 4.2, allows for the shrinkage of coefficient values to zero and, as a result, effectively works as a feature selection method. For example, for the relatively high-dimensional

[5] As discussed in Section 3.1, the *NoOfTimePD* variable attains the value of zero for the majority of observations.

[6] As can be seen in Table B.7, the estimated coefficient for *NoOfTimePD* is negative.

HCDR dataset, this characteristic of the regularization method resulted in the elimination of seven variables.[7]

The λ coefficient representing the inverse of regularization strength appears to be quite similar across the data set. An exception is the GMSC data set, where for the model with raw variables, the coefficient is relatively high, signifying low regularization strength. Overall, the raw models seem less regularized than those utilizing binned variables. On the other hand, the average AUCs seem slightly in favor of the binned models.

Table 5.4: Logistic regression - hyperparameters

Data set	Model	λ^*	Penalty	Average AUC
CA	binned	0.126	ℓ_1	0.938
CA	raw	0.184	ℓ_1	0.931
DCCCT	binned	0.126	ℓ_2	0.765
DCCCT	raw	0.028	ℓ_1	0.720
GMSC	binned	1.758	ℓ_2	0.818
GMSC	raw	6866.488	ℓ_1	0.660
HCDR	binned	0.829	ℓ_1	0.740
HCDR	raw	0.829	ℓ_1	0.737
SGCC	binned	0.184	ℓ_2	0.635
SGCC	raw	1.758	ℓ_1	0.649

*Inverse of the regularization strentgh
The table presents the optimal parameters found through 3-fold cross-validation of the training set. The last column shows the average out-of-sample AUC over the three iterations.

Source: Author's computations

The final models for each data set were re-estimated on the entire training set using the optimal parameters. Table 5.5 presents the resulting performance evaluation on the test set. As can be seen, on average, the model utilizing the binning algorithm outperforms the raw model for all evaluation metrics. However, the unweighted average may be skewed by large differences for one particular data set. Therefore, the third row from the bottom shows the number of data sets for which the binned model outperforms the raw model for each evaluation metric. For all metrics, the binned model outperforms the raw model on most data sets. Nevertheless, a closer inspection reveals that the raw model surpasses the binned model for the Credit Approval data set according to four evaluation metrics. The CA data set does not suffer from class imbalance and

[7]The coefficients for all models and all data sets are not presented in this text due to space constraints. They are available upon request.

contains the fewest observations. As a result, out of the considered data sets, it represents the actual credit scoring environment the least. Apart from the CA, the raw model outperforms the binned model only for the SGCC data set according to the Partial GINI Index. Similarly, the SGCC data set comprises only a few observations. Consequently, the evidence favoring the binned model appears stronger for larger data sets. As a result, the binning transformation seems to prove beneficial to performance with increasing sample size when it comes to logistic regression. It can be argued that more sizeable data sets contain more noise, which is eliminated by the binning transformation.

Considering the AUC metric, the performances vary across data sets. For the GMSC data set, the difference in performance between the raw and binned model is substantial. Even though the difference is not so large for the remaining data sets, even a slight performance improvement may have a considerable positive impact on the financial institution,[8] which encourages the usage of the binning transformation. The absolute performance of the models is acceptable in a majority of the cases,[9] however, the aim of the analysis was not to find the best model. A closely related to AUC is the H-measure, which seems to indicate very similar results. The last member of the group of metrics evaluating the model's ability to differentiate between creditworthy and unreliable customers is the Partial GINI index. As mentioned above, the raw model seems to achieve better performance in terms of PGI for the SGCC and CA data sets. In the former case, the difference is minimal; however, in the latter case, the raw model performs substantially better. As a result, the binning transformation seems to slightly hinder the model's ability to assign proper probabilities to well-behaved clients in small data sets.

The second group of metrics, evaluating the correctness of categorical predictions, includes the F2-score. As can be seen, apart from the CA data set, the absolute performances in terms of this metric are quite low. However, the utilized threshold was not selected such that the performance is optimal but such that the predicted percentage of defaults matches the training set. Moreover, the aim of the analysis is to compare relative rather than absolute performance. In this regard, the binned model seems superior for all data sets. The second

[8] See Section 2.3.

[9] Mandrekar (2010) suggests that values above 0.7 are considered acceptable and values above 0.8 excellent.

metric from the current group is the KS statistic. Apart from the CA data set, the binned model attains a higher value of the statistic than the raw model.

Lastly, the Brier score measures the accuracy of probabilistic predictions. Unlike the remaining metrics, the lower the value of this measure, the better. Even though the differences appear to be minor, the binned model outperforms the raw model for all data sets.

Table 5.5: Logistic regression - results

Data set	Type	AUC	F2-score	Brier score	KS statistic	Partial GINI Index	H-measure
CA	binned	0.892	**0.847**	**0.122**	0.679	0.573	0.564
CA	raw	**0.895**	0.833	0.125	**0.701**	**0.661**	**0.601**
DCCCT	binned	**0.762**	**0.531**	**0.137**	**0.405**	**0.322**	**0.268**
DCCCT	raw	0.723	0.510	0.145	0.378	0.215	0.237
GMSC	binned	**0.816**	**0.340**	**0.056**	**0.491**	**0.587**	**0.319**
GMSC	raw	0.670	0.167	0.063	0.249	0.335	0.090
HCDR	binned	**0.741**	**0.264**	**0.06709**	**0.365**	**0.472**	**0.183**
HCDR	raw	0.739	0.262	0.06714	0.357	0.469	0.182
SGCC	binned	**0.625**	**0.456**	**0.200**	**0.249**	0.081	**0.138**
SGCC	raw	0.607	0.421	0.205	0.223	**0.083**	0.115
binned > raw	-	4	5	5	4	3	4
Average	binned	**0.767**	**0.488**	**0.117**	**0.438**	**0.407**	**0.295**
Average	raw	0.727	0.439	0.121	0.381	0.353	0.245

The table presents the results of calculating the six evaluation metrics (columns (3)-(8)) on the test set comprising 20% observations of each data set. Values in **bold** signal a better performance of the given model type for the given data set. Note that except for the Brier score, the higher the value of the metric, the better. The third row from the bottom shows the number of data sets for which the binned model outperformed the raw model for a given evaluation metric. The last two rows show the average value for each metric for each model type across all data sets.

Source: Author's computations

The discussion so far only concerned the comparison of a single value for each metric and each data set. However, it is desirable to obtain the significance level of the differences. Therefore, Table 5.6 presents the results of the permutation test for each metric and each data set. For two data sets (DCCCT and GMSC), the null hypothesis that the raw model performs better or equally to the binned model is rejected for all evaluation metrics. For the remaining data sets, the evidence is not as strong. As discussed above, the CA and SGCC data sets do not seem to benefit from the binning transformation, which is confirmed by the inability to reject the null hypothesis. With respect to the HCDR data set, the null is rejected only for two metrics at the 5% significance level, namely the AUC and the KS statistic. The p-value for the PGI and H-measure appears to be quite close to the 10% significance level, suggesting that some weak evidence in favor of the alternative hypothesis does exist. As a result, the binning transformation seems to positively affect the model's ability to

differentiate between eligible and ineligible customers according to most data sets. For the CA data set, the p-values are quite high. On the other hand, for the SGCC data, the null is rejected at the 10% significance level for three metrics, while for one metric at the 5% significance level. As a result, with more data, the null could possibly be rejected even for most of the metrics for the SGCC data set.

Table 5.6: Logistic regression - permutation tests

Data set	AUC	F2-score	Brier score	KS statistic	Partial GINI Index	H-measure
CA	0.562	0.553	0.381	0.813	0.766	0.907
DCCCT	**0.000**	**0.001**	**0.000**	**0.000**	**0.000**	**0.000**
GMSC	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
HCDR	**0.035**	0.407	0.210	**0.004**	0.161	0.116
SGCC	0.149	0.096	**0.036**	0.090	0.532	0.070
p <= 0.05	3	2	3	3	2	2

The values in the table represent the p-values of the null hypothesis that the model with raw variables performs better or equally to the binned model. As a result, p-values below 0.05 (in **bold**) signal the rejection of the null hypothesis in favor of the alternative that the binned model performs better at the 5% significance level. The last row shows the number of data sets for which the null is rejected for a given evaluation metric. The permutation test was performed for 5000 repetitions.

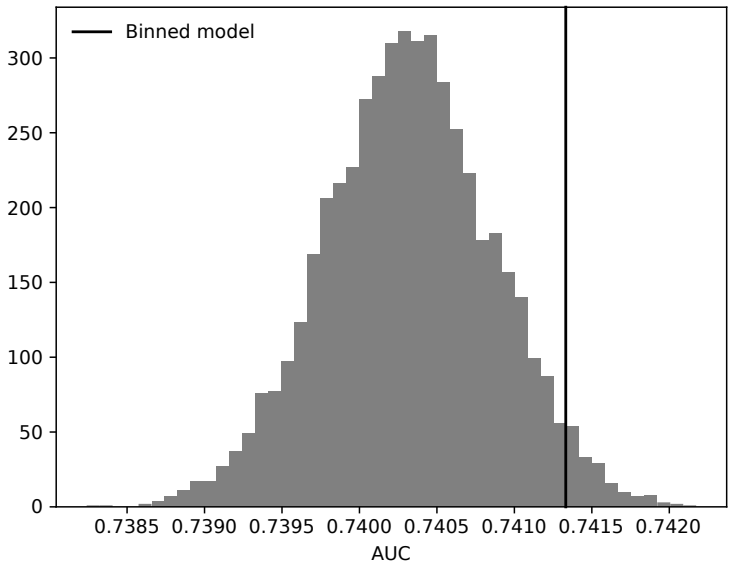
Source: Author's computations

For illustration purposes, Figure 5.2 shows the results of the permutation test for the HCDR data set and the AUC metric. As can be seen, the majority of the performances for randomly selected predictions from both models are worse than the performance of the binned model. Consequently, the p-value of the null hypothesis is approximately 0.035.[10]

To conclude, the logistic regression appears to benefit from the binning transformation, especially for larger data sets, which serves as evidence for Hypothesis 1. The most significant impact appears to be on the model's ability to differentiate between good and bad customers. As a result, the findings of this thesis advocate the usage of the binning transformation and subsequent WoE encoding for credit risk models. The results seem to align with Sharma (2011) and Leung *et al.* (2008).

[10]See Table 5.5 row 4, column 2.

Figure 5.2: Logistic regression - Permutation test (HCDR, AUC)



The figure shows the distribution of the results of the permutation test for the AUC metric. The predictions were obtained using logistic regression estimated on the HCDR data set. The simulation was run for 5000 iterations.

Source: Author's computations

5.4 Decision tree

The next inspected estimation method is the CART decision tree. Table 5.7 presents the results of the grid search for optimal parameters. As can be seen, for all models, the preferred criterion is entropy. Concerning regularization, in most cases, the algorithm restricting the number of total leaves of the tree appears superior. Nevertheless, the CA and SGCC data sets seem to favor the maximum depth parameter, which could be connected to the lower number of observations. The average AUCs are quite similar for the model types for all data sets.

As in the previous subsection, Table 5.8 presents the absolute performances of the models with optimal parameters for each data set and model type. As can be seen, the results are less convincing than in the case of the logistic regression. While, on average, the binning transformation appears to be mostly superior, the differences within individual data sets are minimal. In addition, the average performance according to PGI of the binned model seems to be substantially worse than that of its raw counterpart. In fact, the value of the metric is higher for the binned model only for the GMSC data set. In addition, the binned model appears to perform very poorly for low-probability clients on the

Table 5.7: Decision tree - hyperparameters

Data set	Model	Criterion	Max depth	Max leaves	Average AUC
CA	binned	entropy	5	-	0.896
CA	raw	entropy	5	-	0.909
DCCCT	binned	entropy	-	25	0.761
DCCCT	raw	entropy	-	25	0.765
GMSC	binned	entropy	-	50	0.814
GMSC	raw	entropy	-	50	0.814
HCDR	binned	entropy	-	100	0.719
HCDR	raw	entropy	-	50	0.718
SGCC	binned	entropy	-	25	0.624
SGCC	raw	entropy	5	-	0.617

The table presents the optimal parameters found through 3-fold cross-validation of the training set. The last column shows the average out-of-sample AUC over the three iterations.

Source: Author's computations

CA data set where the PGI attains a negative value.

Table 5.8: Decision tree - results

Data set	Type	AUC	F2-score	Brier score	KS statistic	Partial GINI Index	H-measure
CA	binned	**0.877**	**0.897**	**0.131**	**0.710**	-0.057	**0.572**
CA	raw	0.873	0.879	0.135	0.702	**0.229**	0.548
DCCCT	binned	0.763	0.536	0.1361	0.3986	0.333	0.2631
DCCCT	raw	**0.766**	**0.553**	**0.1362**	**0.3987**	**0.350**	**0.2632**
GMSC	binned	**0.812**	**0.3644**	0.0562	0.481	**0.593**	**0.3143**
GMSC	raw	0.811	0.3643	**0.0559**	**0.482**	0.581	0.3142
HCDR	binned	0.721	0.253	0.0680	0.337	0.432	0.159
HCDR	raw	**0.723**	**0.271**	**0.0678**	**0.340**	**0.434**	**0.160**
SGCC	binned	**0.679**	**0.574**	**0.193**	**0.272**	0.150	**0.149**
SGCC	raw	0.656	0.466	0.210	0.268	**0.184**	0.102
binned > raw	-	3	3	3	2	1	3
Average	binned	**0.771**	**0.525**	**0.117**	**0.440**	0.290	**0.292**
Average	raw	0.766	0.507	0.121	0.438	**0.356**	0.277

The table presents the results of calculating the six evaluation metrics (columns (3)-(8)) on the test set comprising 20% observations of each data set. Values in **bold** signal a better performance of the given model type for the given data set. Note that except for the Brier score, the higher the value of the metric, the better. The third row from the bottom shows the number of data sets for which the binned model outperformed the raw model for a given evaluation metric. The last two rows show the average value for each metric for each model type across all data sets.

Source: Author's computations

As a verification of the results, Table 5.8 shows the p-values of the permutation tests. Notably, the null hypothesis is rejected at the 5% significance level only in three cases. Moreover, in the majority of the instances, the p-values are very high, which suggests that the binned model does not seem to perform significantly better when it comes to the decision tree classifier.

Table 5.9: Decision tree - permutation tests

Data set	AUC	F2-score	Brier score	KS statistic	Partial GINI Index	H-measure
CA	0.434	**0.032**	0.412	0.486	0.933	0.424
DCCCT	0.761	1.000	0.365	0.694	0.908	0.721
GMSC	0.358	0.386	0.935	0.519	**0.025**	0.501
HCDR	0.843	1.000	0.939	0.805	0.751	0.726
SGCC	0.312	**0.015**	0.102	0.689	0.576	0.399
p <= 0.05	0	2	0	0	1	0

The values in the table represent the p-values of the null hypothesis that the model with raw variables performs better or equally to the binned model. As a result, p-values below 0.05 (in **bold**) signal the rejection of the null hypothesis in favor of the alternative that the binned model performs better at the 5% significance level. The last row shows the number of data sets for which the null is rejected for a given evaluation metric. The permutation test was performed for 5000 repetitions.

Source: Author's computations

As discussed in Section 2.1, the literature seems to be divided with respect to the effect of binning on the performance of the decision tree classifier. The results of the current analysis support the view of Ventura & Martinez (1995), who found that the ID3 decision tree algorithm does not seem to benefit from binning numerical variables. It is important to note that the CART algorithm was not used in any of the considered studies, and therefore the current thesis inspects a new aspect of the issue. As Table 5.8 suggests, the performances of the models are very similar. This could be attributed to the fact that the CART algorithm is utilized during pre-binning, and therefore, the resulting tree structures of the raw and binned models are very similar.[11] Consequently, the available evidence does not seem to support Hypothesis 2.

5.5 Random Forest

The second tree-based algorithm utilized in this thesis is the Random Forest model. Once again, Table 5.10 enumerates the optimal parameters found through cross-validation of the training test. Even though for the decision tree the selected criterion was entropy for all models, the Random Forest seems to attain higher performance with the gini criterion for three instances. In addition, unlike for the decision tree, the prevalent regularization parameter appears to be the tree depth since the maximum number of leaves achieved better average performance only for four models. Since the selected maximum depth is either 5 or 10, the algorithm seems to prefer a rather shallow tree structure. On

[11] See Figures A.1 and A.2. As can be seen, the initial splits are essentially the same.

the other hand, the number of estimators is relatively high since, apart from two cases, the maximum considered number of trees was selected as best performing. The average AUCs seem to be similar, with the raw model achieving slightly higher performance for all data sets.

Table 5.10: Random Forest - hyperparameters

Data Set	Model	Criterion	Max depth	# of trees	Max leaves	Average AUC
CA	binned	gini	5	50	-	0.938
CA	raw	entropy	5	100	-	0.945
DCCCT	binned	entropy	-	500	100	0.772
DCCCT	raw	entropy	-	500	500	0.780
GMSC	binned	entropy	-	500	100	0.821
GMSC	raw	gini	10	500	-	0.827
HCDR	binned	entropy	10	500	-	0.736
HCDR	raw	entropy	-	500	500	0.739
SGCC	binned	gini	5	500	-	0.639
SGCC	raw	entropy	5	500	-	0.661

The table presents the optimal parameters found through 3-fold cross-validation of the training set. The last column shows the average out-of-sample AUC over the three iterations.

Source: Author's computations

The results for the RF classifier are available in Table 5.11. As can be seen, the evidence is even less favorable than for the decision tree. In a majority of cases, the raw model performs better, and for the exceptions where it does not, the differences in performance are very small. Moreover, on average, the raw model achieves higher performance according to all evaluation metrics.

Table 5.11: Random Forest - results

Data set	Type	AUC	F2-score	Brier score	KS statistic	Partial GINI Index	H-measure
CreditApproval	binned	0.890	0.847	0.123	0.715	0.399	0.591
CreditApproval	raw	**0.900**	0.847	**0.119**	**0.719**	**0.704**	**0.606**
CreditCardTaiwan	binned	0.771	**0.535**	0.136	0.416	0.343	0.277
CreditCardTaiwan	raw	**0.778**	0.528	**0.134**	**0.419**	**0.363**	**0.286**
GiveMeSomeCredit	binned	0.818	0.345	0.056	0.499	**0.623**	0.324
GiveMeSomeCredit	raw	**0.826**	**0.361**	**0.055**	**0.508**	0.621	**0.340**
HomeCredit	binned	0.736	0.260	0.0678	0.354	0.473	0.178
HomeCredit	raw	**0.740**	**0.267**	**0.0676**	**0.357**	**0.479**	**0.183**
SouthGermanCredit	binned	0.686	0.476	0.1923	0.2800	0.247	**0.194**
SouthGermanCredit	raw	**0.695**	**0.502**	**0.1915**	**0.306**	**0.251**	0.186
binned > raw	-	0	1	0	0	1	1
Average	binned	0.780	0.493	0.115	0.453	0.417	0.313
Average	raw	**0.788**	**0.501**	**0.114**	**0.462**	**0.484**	**0.320**

The table presents the results of calculating the six evaluation metrics (columns (3)-(8)) on the test set comprising 20% observations of each data set. Values in **bold** signal a better performance of the given model type for the given data set. Note that except for the Brier score, the higher the value of the metric, the better. The third row from the bottom shows the number of data sets for which the binned model outperformed the raw model for a given evaluation metric. The last two rows show the average value for each metric for each model type across all data sets.

The previous findings are confirmed in Table 5.12 since the null hypothesis is not rejected at the 5% significance level in any of the considered scenarios. In addition, the p-values for all of the tests are quite high. Consequently, it appears that the obtained results are aligned with the extant literature (Sharma 2011) since binning numerical variables does not seem to improve the performance of the Random Forest classifier. The inability to reject the null hypothesis serves as evidence for Hypothesis 3 of this thesis.

Table 5.12: Random Forest - permutation tests

Data set	AUC	F2-score	Brier score	KS statistic	Partial GINI Index	H-measure
CreditApproval	0.876	0.814	0.881	0.730	0.958	0.835
CreditCardTaiwan	0.999	0.208	0.999	0.779	0.969	0.993
GiveMeSomeCredit	1.000	0.991	1.000	0.968	0.369	1.000
HomeCredit	1.000	0.995	1.000	0.872	1.000	1.000
SouthGermanCredit	0.630	0.592	0.577	0.808	0.491	0.289
p <= 0.05	0	0	0	0	0	0

The values in the table represent the p-values of the null hypothesis that the model with raw variables performs better or equally to the binned model. As a result, p-values below 0.05 (in **bold**) signal the rejection of the null hypothesis in favor of the alternative that the binned model performs better at the 5% significance level. The last row shows the number of data sets for which the null is rejected for a given evaluation metric. The permutation test was performed for 5000 repetitions.

Source: Author's computations

5.6 Neural network

Abandoning the tree-based algorithms, the next machine learning method under inspection is the neural network. As shown in Section 2, the effect of binning transformation on the NN's performance was not thoroughly evaluated. As a result, this thesis will attempt to shine some light on this issue. For this purpose, Table 5.13 shows the optimal hyperparameters for each data set. As can be seen, the prevalent activation function is the ReLU, which is expected for reasons outlined in Section 4.3. Nevertheless, for some of the data sets, the tanh and sigmoid activation functions seem to achieve competitive performance. As for the number of nodes, it seems to depend on the data set size. For data sets with fewer variables, the number of nodes tends to be closer to the number of features. On the other hand, for data sets such as HCDR, the optimal number of nodes appears to be equal to two-thirds or even half of the number of variables. With respect to the average AUCs, the binned model seems to achieve superior performance across all data sets. In addition, the differences seem to be substantial.

Table 5.13: Neural network - hyperparameters

Data set	Model	Activation function	α^*	# of nodes	Average AUC
CA	binned	tanh	0.0100	5	0.932
CA	raw	sigmoid	0.0100	10	0.880
DCCCT	binned	tanh	0.0001	9	0.769
DCCCT	raw	ReLU	1.0000	7	0.669
GMSC	binned	ReLU	0.0100	8	0.820
GMSC	raw	ReLU	0.0100	8	0.792
HCDR	binned	tanh	0.0100	32	0.741
HCDR	raw	ReLU	0.0100	24	0.555
SGCC	binned	ReLU	0.0010	6	0.642
SGCC	raw	ReLU	10	6	0.574

*Regularization strength
The table presents the optimal parameters found through 3-fold cross-validation of the training set. The last column shows the average out-of-sample AUC over the three iterations. The maximum number of epochs for the neural network was set to 500.

Source: Author's computations

The pattern spotted in the cross-validated AUCs seems to be confirmed by the results in Table 5.14. Notably, the binned model outperforms its raw counterpart for almost all metrics across all data sets, with the differences being economically important. The only instances for which the binned model slightly lacks behind are the F2-score for the HCDR data set and the PGI for the CA data set. Consequently, the binning transformation appears to have a strong positive effect on the model's ability to differentiate between bad and good customers (AUC, PGI, H-measure) as well as on the accuracy of probability predictions. Since the KS statistic represents a more rigorous approach to evaluating the correctness of categorical predictions than the F2-score, the third model characteristic also seems to be improved.

As further evidence in support of Hypothesis 4, Table 5.15 presents the results of the permutation tests. As can be seen, the null hypothesis of the raw model performing better or equal to the binned model is rejected for the majority of the cases. Similarly to the logistic regression, the evidence seems to be stronger for large data sets. In fact, the null hypothesis is rejected at the 5% significance level according to all metrics for all of the three largest data sets. In addition, for the SGCC data set, the null hypothesis is rejected for two metrics at the 5% significance level and for three metrics at the 10% significance level. The evidence is insufficient to reject the null hypothesis for any of the considered metrics only for the CA data set. Nevertheless, the analysis provides substantial

Table 5.14: Neural network - results

Data set	Type	AUC	F2-score	Brier score	KS statistic	Partial GINI Index	H-measure
CA	binned	**0.899**	**0.847**	**0.121**	**0.719**	0.615	**0.608**
CA	raw	0.875	0.820	0.134	0.652	**0.831**	0.557
DCCCT	binned	**0.765**	**0.541**	**0.136**	**0.413**	**0.333**	**0.273**
DCCCT	raw	0.643	0.358	0.215	0.219	0.247	0.082
GMSC	binned	**0.819**	**0.353**	**0.056**	**0.502**	**0.596**	**0.326**
GMSC	raw	0.771	0.306	0.060	0.426	0.542	0.252
HCDR	binned	**0.742**	0.265	**0.067**	**0.366**	**0.473**	**0.185**
HCDR	raw	0.498	**0.298**	0.078	0.004	-0.004	0.000
SGCC	binned	**0.633**	**0.429**	**0.202**	**0.237**	**0.170**	**0.128**
SGCC	raw	0.539	0.388	0.217	0.175	-0.140	0.090
binned > raw	-	5	4	5	5	4	5
Average	binned	**0.772**	**0.487**	**0.116**	**0.447**	**0.437**	**0.304**
Average	raw	0.665	0.434	0.141	0.295	0.295	0.196

The table presents the results of calculating the six evaluation metrics (columns (3)-(8)) on the test set comprising 20% observations of each data set. Values in **bold** signal a better performance of the given model type for the given data set. Note that except for the Brier score, the higher the value of the metric, the better. The third row from the bottom shows the number of data sets for which the binned model outperformed the raw model for a given evaluation metric. The last two rows show the average value for each metric for each model type across all data sets.

Source: Author's computations

evidence in favor of the binning transformation for the neural network.

Table 5.15: Neural network - permutation tests

Data set	AUC	F2-score	Brier score	KS statistic	Partial GINI Index	H-measure
CA	0.01	0.323	0.160	0.163	0.620	0.237
DCCCT	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
GMSC	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
HCDR	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
SGCC	**0.044**	0.304	0.087	0.068	**0.043**	0.075
p <= 0.05	4	3	3	3	4	3

The values in the table represent the p-values of the null hypothesis that the model with raw variables performs better or equally to the binned model. As a result, p-values below 0.05 (in **bold**) signal the rejection of the null hypothesis in favor of the alternative that the binned model performs better at the 5% significance level. The last row shows the number of data sets for which the null is rejected for a given evaluation metric. The permutation test was performed for 5000 repetitions.

Source: Author's computations

5.7 Gaussian Naïve Bayes

The last machine learning algorithm considered in this thesis is the Naïve Bayes classifier. As demonstrated in Section 2, the literature seems to support the performance enhancement provided by the binning transformation when it comes to this estimation method.

Since the Naïve Bayes does not have any trainable parameters, the hyperparameter optimization did not have to be performed for this method. Consequently, Table 5.16 presents the results of the performance evaluation. As can be seen, the binned model appears to be superior in most cases, but the results are not completely unambiguous. For example, in terms of the Brier score, the raw model seems to be performing better, especially for the two largest data sets. Overall, the binning transformation appears to boost the model's ability to differentiate between eligible and ineligible customers as measured by the AUC and the H-measure. However, the evidence supporting the correctness of categorical predictions and the precision of estimated probabilities is slightly weaker. Nevertheless, on average, the binned model outperforms the raw model for all metrics apart from the Brier score. Moreover, the binned model also achieves better performance on a majority of the data sets for all metrics but the Brier score.

Table 5.16: Gaussian Naïve Bayes - results

Data set	Type	AUC	F2-score	Brier score	KS statistic	Partial GINI Index	H-measure
CA	binned	**0.886**	0.820	**0.139**	**0.683**	**0.713**	**0.579**
CA	raw	0.855	0.820	0.172	0.670	0.238	0.523
DCCCT	binned	**0.760**	**0.520**	**0.187**	**0.406**	**0.237**	**0.260**
DCCCT	raw	0.670	0.395	0.417	0.263	0.148	0.119
GMSC	binned	**0.805**	**0.328**	0.088	**0.486**	**0.458**	**0.306**
GMSC	raw	0.691	0.219	**0.067**	0.271	0.353	0.127
HCDR	binned	**0.688**	**0.200**	0.891	**0.289**	0.130	**0.115**
HCDR	raw	0.605	0.135	**0.073**	0.159	**0.211**	0.036
SGCC	binned	**0.635**	0.442	0.2181	0.225	0.094	**0.117**
SGCC	raw	0.629	**0.453**	**0.2180**	**0.230**	**0.153**	0.104
binned > raw	-	5	3	2	4	3	5
Average	binned	**0.755**	**0.462**	0.305	**0.418**	**0.326**	**0.275**
Average	raw	0.690	0.404	**0.189**	0.319	0.220	0.182

The table presents the results of calculating the six evaluation metrics (columns (3)-(8)) on the test set comprising 20% observations of each data set. Values in **bold** signal a better performance of the given model type for the given data set. Note that except for the Brier score, the higher the value of the metric, the better. The third row from the bottom shows the number of data sets for which the binned model outperformed the raw model for a given evaluation metric. The last two rows show the average value for each metric for each model type across all data sets.

Source: Author's computations

The results of the permutation tests in Table 5.17 support the conclusions derived so far. For large data sets, the null hypothesis of the raw model performing better or equal to the binned model is rejected consistently across most metrics. The exceptions are the Brier score and the PGI. For the smaller data sets, the results seem to be more ambiguous since while for CA the null is rejected for two metrics at the 5% significance level, for SGCC the null hypothesis fails to be rejected for any of the considered metrics.

Table 5.17: Gaussian Naïve Bayes - permutation tests

Data set	AUC	F2-score	Brier score	KS statistic	Partial GINI Index	H-measure
CA	**0.047**	0.589	0.092	0.204	**0.034**	0.140
DCCCT	**0.000**	**0.000**	**0.000**	**0.000**	0.100	**0.000**
GMSC	**0.000**	**0.000**	1.000	**0.000**	**0.000**	**0.000**
HCDR	**0.000**	**0.000**	1.000	**0.000**	1.000	**0.000**
SGCC	0.327	0.370	0.503	0.471	0.727	0.321
p <= 0.05	4	3	1	3	2	3

The values in the table represent the p-values of the null hypothesis that the model with raw variables performs better or equally to the binned model. As a result, p-values below 0.05 (in **bold**) signal the rejection of the null hypothesis in favor of the alternative that the binned model performs better at the 5% significance level. The last row shows the number of data sets for which the null is rejected for a given evaluation metric. The permutation test was performed for 5000 repetitions.

Source: Author's computations

To conclude, while the Naïve Bayes seems to benefit from the binning of numerical variables in certain aspects of model performance, in other areas, such as the accuracy of the estimated probabilities, the binning transformation appears to have a hindering effect. Nevertheless, the provided evidence is mostly aligned with the extant literature (Abraham *et al.* 2006; Lustgarten *et al.* 2008) and serves in favor of Hypothesis 5.

5.8 Summary

To provide a comprehensive overview of the results of the primary analysis and also as a means of comparison for the subsequent robustness checks, Table 5.18 presents the summarized results across all estimation methods. A clear pattern arises from the summary since the estimation methods appear to be divided into two groups. Firstly, the tree-based algorithms do not seem to benefit from the binning transformation since the null hypothesis is rarely rejected for any estimation metric within any data set. On the other hand, for the logistic regression, neural network, and the Naïve Bayes classifier, strong evidence in favor of the binning transformation was found. In addition, overall, the binning transformation appears superior for larger data sets.

As discussed in Section 4.3, multiple evaluation metrics were utilized to capture different aspects of model performance. To demonstrate the relationships between the utilized metrics, Table 5.19 shows the Pearson correlation coefficient for the metrics across all estimation methods and data sets. The metrics are divided into three groups. The first group, evaluating the model's ability

Table 5.18: Results - summary

Method	Data set	Type	AUC	F2-score	Brier score	KS statistic	Partial GINI Index	H-measure
LogReg	binned > raw	-	4	5	5	4	3	4
LogReg	Average	binned	**0.767**	**0.488**	**0.117**	**0.438**	**0.407**	**0.295**
LogReg	Average	raw	0.727	0.439	0.121	0.381	0.353	0.245
LogReg	p <= 0.05	-	3	2	3	3	2	2
DecTree	binned > raw	-	3	3	3	2	1	3
DecTree	Average	binned	**0.771**	**0.525**	**0.117**	**0.44**	0.29	**0.292**
DecTree	Average	raw	0.766	0.507	0.121	0.438	**0.356**	0.277
DecTree	p <= 0.05	-	0	2	0	0	1	0
RandForest	binned > raw	-	0	1	0	0	1	1
RandForest	Average	binned	0.78	0.493	0.115	0.453	0.417	0.313
RandForest	Average	raw	**0.788**	**0.501**	**0.114**	**0.462**	**0.484**	**0.32**
RandForest	p <= 0.05	-	0	0	0	0	0	0
NN	binned > raw	-	5	4	5	5	4	5
NN	Average	binned	**0.772**	**0.487**	**0.116**	**0.447**	**0.437**	**0.304**
NN	Average	raw	0.665	0.434	0.141	0.295	0.295	0.196
NN	p <= 0.05	-	4	3	3	3	4	3
GaussNB	binned > raw	-	5	3	2	4	3	5
GaussNB	Average	binned	**0.755**	**0.462**	0.305	**0.418**	**0.326**	**0.275**
GaussNB	Average	raw	0.69	0.404	**0.189**	0.319	0.22	0.182
GaussNB	p <= 0.05	-	4	3	1	3	2	3

The table presents the summarized results of calculating the six evaluation metrics (columns (3)-(8)) on the test set comprising 20% observations of each data set. Values in **bold** signal a better average performance of the given model type for the evaluation metric. Note that except for the Brier score, the higher the value of the metric, the better. The first row for each method shows the number of data sets for which the binned model outperformed the raw model for a given evaluation metric. The second and third rows for each estimation method show the average value for each metric for each model type across all data sets. The last row shows the number of data sets for which the null hypothesis of the raw model performing better or equal to the binned model was rejected.

Source: Author's computations

to differentiate between bad and good customers, contains the AUC, PGI, and H-measure. As can be seen, these three measures appear to be highly positively correlated. However, the PGI is slightly less related to the other two metrics and thus may capture an additional aspect of the model's performance. Secondly, the accuracy of the estimated probabilities is measured by the Brier score, which, as expected, has a negative relationship with the remaining metrics. In addition, the correlation does not appear to be particularly strong, and therefore its employment secures the inspection of an additional element of the model's performance. Lastly, the KS statistic and the F2-score capture the correctness of categorical predictions and are quite highly correlated. In addition, the KS statistic seems to be substantially related to the AUC statistic and the H-measure.

Table 5.19: Correlation matrix - evaluation metrics

	AUC	F2-score	BS	KS	PGI	H-measure
AUC	1	0.64	-0.26	0.97	0.72	0.93
F2-score		1	-0.01	0.76	0.15	0.84
BS			1	-0.21	-0.43	-0.20
KS				1	0.61	0.98
PGI					1	0.55
H-measure						1

The table presents the Pearson correlation coefficient between the metrics in the main analysis across all data sets and estimation methods.

Source: Author's computations

5.9 Missing values

For the purposes of the main analysis, missing values were disregarded from all affected data sets in order to maximize comparability between the models. However, as discussed in Section 2.1, one of the main strengths of the binning transformation is its ability to handle missing values. As a result, an additional round of estimations was performed with missing values included. Only three data sets were considered for the robustness analysis: CA, GMSC, and HCDR since the remaining data sets do not contain missing values. Since, for the raw estimation, missing values were replaced by mean, median, or mode, the current analysis ultimately compares the binning algorithm's ability to handle missing values with simple imputation. It is possible that the results may differ

for other more advanced forms of imputation. However, such an investigation is behind the scope of this text.

To demonstrate the management of missing values, Table 5.20 presents the results of optimal binning for the *MonthlyIncome* variable from the GMSC data set. As can be seen, missing values comprise nearly 20% of all observations. In addition, since the Event Rate appears to be decreasing, customers with higher income seem to be less prone to default. The bin consisting of missing values attains only the fourth-highest ER of all the bins. Considering that the resulting WoE assigned to the missing bin is quite close to the bin representing the second largest income group, customers for which the income information is unavailable appear to be moderately creditworthy. Even though the total Information Value of the variable is relatively low, the information derived from missing observations may be useful for estimation.

Table 5.20: Binning of *MonthlyIncome* (with missing values)

Bin	Count	Count (%)	Non-event	Event	ER*	WoE**	IV***
(-inf, 3332.50)	23098	19.25	20947	2151	9.31	-0.35	0.03
[3332.50, 4833.50)	18455	15.38	16904	1551	8.40	-0.24	0.01
[4833.50, 6642.50)	18878	15.73	17595	1283	6.80	-0.01	0.00
[6642.50, 9950.50)	19905	16.59	18843	1062	5.34	0.25	0.01
[9950.50, inf)	15943	13.29	15258	685	4.30	0.47	0.02
Missing	23720	19.77	22385	1335	5.63	0.19	0.01
Totals	119999	100.00	111932	8067	6.72		0.08

*Event Rate
**Weight of Evidence
***Information Value

Source: Author's computations

Table 5.21 presents the summarized results of the robustness analysis with missing values.[12] As can be seen, the results do not seem to differ substantially from the previous analysis. However, the evidence in favor of the logistic regression is slightly weakened by the introduction of missing values. Even though the binned model outperforms the raw model, on average, for all metrics, the null hypothesis is rejected only for a single data set in most cases. In addition, the p-values for the data set with the greatest portion of missing values (HCDR) are very high, suggesting that handling unavailable observations using the binning transformation is not superior to simple imputation when it comes to logistic regression.

[12]For complete results see Tables B.8 and B.9.

In the previous analysis, no convincing evidence was obtained in favor of the binning transformation regarding tree-based algorithms. The inclusion of missing values further weakens the evidence since the null hypothesis of the raw model performing better or equally to the binned model can be rejected at the 5% significance only for a single instance, which is the F2-score for the decision tree. Since standard credit risk data sets usually contain a nonnegligible share of missing values, applying the binning transformation for a tree-based algorithm likely does not provide a performance enhancement based on the available evidence.

On the other hand, the results in Table 5.21 confirm the findings of the previous analysis since the binning transformation appears to improve the performance of the NN and Naïve Bayes classifiers even after the inclusion of missing values. Moreover, the evidence is even stronger since the null hypothesis is rejected for most evaluation metrics in most data sets for both methods. The only larger drawback of the binning transformation seems to be the sub-optimal Brier score for the Naïve Bayes, which, as in the previous analysis, seems to be surpassed by the raw model. In addition, in line with the previous investigation, the neural network seems to provide slightly weaker evidence for the CA data set, which contains few observations.

Table 5.21: Results - estimation with missing values

Method	Data set	Type	AUC	F2-score	Brier score	KS statistic	Partial GINI Index	H-measure
LogReg	binned > raw	-	2	1	2	2	2	2
LogReg	Average	binned	**0.816**	**0.493**	**0.083**	**0.516**	**0.557**	**0.359**
LogReg	Average	raw	0.753	0.427	0.088	0.417	0.429	0.263
LogReg	p <= 0.05	-	2	1	2	1	1	1
DecTree	binned > raw	-	3	0	2	3	1	2
DecTree	Average	binned	**0.803**	0.499	**0.086**	**0.487**	0.562	**0.341**
DecTree	Average	raw	0.797	**0.501**	0.093	0.47	**0.564**	0.323
DecTree	p <= 0.05	-	0	1	0	0	0	0
RandForest	binned > raw	-	0	0	0	1	1	1
RandForest	Average	binned	0.817	0.5	0.081	**0.529**	0.586	0.375
RandForest	Average	raw	**0.825**	**0.513**	**0.08**	0.525	**0.601**	**0.381**
RandForest	p <= 0.05	-	0	0	0	0	0	0
NN	binned > raw	-	3	2	3	3	2	3
NN	Average	binned	**0.813**	**0.492**	**0.083**	**0.521**	**0.519**	**0.361**
NN	Average	raw	0.623	0.472	0.097	0.216	0.212	0.172
NN	p <= 0.05	-	2	2	3	3	2	2
GaussNB	binned > raw	-	3	3	1	3	2	3
GaussNB	Average	binned	**0.784**	**0.45**	0.372	**0.485**	**0.417**	**0.32**
GaussNB	Average	raw	0.705	0.381	**0.128**	0.356	0.286	0.21
GaussNB	p <= 0.05	-	3	3	1	3	1	3

The table presents the summarized results of calculating the six evaluation metrics (columns (3)-(8)) on the test set comprising 20% observations of each data set. Only three data sets containing missing values were considered: CA, GMSC, and HCDR. For the raw model estimation, unavailable observations were replaced by mean, median, or mode. Values in **bold** signal a better average performance of the given model type for the evaluation metric. Note that except for the Brier score, the higher the value of the metric, the better. The first row for each method shows the number of data sets for which the binned model outperformed the raw model for a given evaluation metric. The second and third rows for each estimation method show the average value for each metric for each model type across all data sets. The last row shows the number of data sets for which the null hypothesis of the raw model performing better or equal to the binned model was rejected. For detailed results, see Tables B.8 and B.9.

Source: Author's computations

5.10 Outliers

The second supplementary analysis investigates the effect of removing outliers on the findings obtained in the primary analysis. Table 5.22 presents the results. In general, the findings appear to be robust to the exclusion of outlying observations, with the exception of the logistic regression. As indicated in Section 4.4, logistic regression may be sensitive to extreme observations. While in the main analysis, the binning transformation handles outliers by grouping them together with regular observations, the raw estimation does not have any inherent outlier management. Consequently, the performance of the logistic regression might be deteriorated. This appears to be the case since the null hypothesis of the raw model performing better or equally to the binned model can be consistently rejected for most of the metrics only for a single data set. As a result, the treatment of outliers seems to have a substantial effect on performance when it comes to logistic regression.

On the other hand, the evidence in favor of the binned model for the neural network is still quite strong, even after the elimination of outliers. For all metrics, the null hypothesis is rejected for the majority of data sets. Similarly, significant results are obtained for the Naïve Bayes classifier for four of the six considered evaluation metrics. The exceptions are the Brier score and the F2-score, which is in line with the primary analysis. Lastly, the tree-based algorithms still do not seem to benefit from the binning transformation.

Table 5.22: Results - estimation without outliers

Method	Data set	Type	AUC	F2-score	Brier score	KS statistic	Partial GINI Index	H-measure
LogReg	binned > raw	-	3	3	3	5	4	3
LogReg	Average	binned	**0.767**	**0.483**	**0.11**	**0.444**	**0.441**	**0.301**
LogReg	Average	raw	0.752	0.473	0.113	0.435	0.375	0.286
LogReg	p <= 0.05	-	1	0	1	1	2	1
DecTree	binned > raw	-	5	2	2	5	1	3
DecTree	Average	binned	**0.762**	**0.504**	**0.112**	**0.428**	0.268	**0.281**
DecTree	Average	raw	0.749	0.49	0.119	0.404	**0.284**	0.263
DecTree	p <= 0.05	-	0	1	1	0	0	0
RandForest	binned > raw	-	1	2	1	1	2	1
RandForest	Average	binned	**0.778**	**0.488**	0.107	0.45	**0.419**	0.307
RandForest	Average	raw	0.777	0.485	**0.107**	**0.457**	0.391	**0.319**
RandForest	p <= 0.05	-	0	0	0	0	0	0
NN	binned > raw	-	5	4	5	5	4	5
NN	Average	binned	**0.775**	**0.484**	**0.108**	**0.444**	**0.427**	**0.306**
NN	Average	raw	0.677	0.427	0.143	0.314	0.313	0.211
NN	p <= 0.05	-	4	3	4	3	3	3
GaussNB	binned > raw	-	4	4	2	3	3	3
GaussNB	Average	binned	**0.756**	**0.463**	**0.148**	**0.436**	**0.356**	**0.286**
GaussNB	Average	raw	0.715	0.41	0.155	0.355	0.248	0.224
GaussNB	p <= 0.05	-	3	2	1	3	3	3

The table presents the summarized results of calculating the six evaluation metrics (columns (3)-(8)) on the test set comprising 20% observations of each data set. Outliers above 99% and below 1% were trimmed from all data sets. Values in **bold** signal a better average performance of the given model type for the evaluation metric. Note that except for the Brier score, the higher the value of the metric, the better. The first row for each method shows the number of data sets for which the binned model outperformed the raw model for a given evaluation metric. The second and third rows for each estimation method show the average value for each metric for each model type across all data sets. The last row shows the number of data sets for which the null hypothesis of the raw model performing better or equal to the binned model was rejected. For detailed results, see Tables B.10 and B.11.

Source: Author's computations

5.11 Omitting categorical variables

An additional robustness investigation regards the inclusion of categorical variables. While in the main analysis, categorical features were retained since their interaction with the numerical variables may affect performance, in the supplementary examination, they will be disregarded, and the impact on the results will be inspected. Four data sets were utilized since the GMSC data set does not contain categorical variables. The unavailable observations are disregarded to avoid any effects caused by a different treatment of missing values, as in the primary analysis.

The results are presented in Table 5.23. The overall picture seems to be very similar to the previous analyses. However, the logistic regression does not seem to benefit significantly from the binning transformation. While the absolute values of the metrics tend to be, on average, better for the binning transformation, the differences are statistically significant only for a single data set. Nevertheless, even though the positive results for GMSC are not presented, they are relevant since the data set contains only numerical variables. Consequently, it appears that the presence of categorical variables does not have to be necessary for the binning transformation to induce a performance improvement. However, their omission seems to deteriorate the effect.

The evidence regarding the tree-based algorithms seems consistent with the previous analyses, given the inability to reject the null hypothesis in nearly all cases. On the other hand, the NN and Naïve Bayes estimators appear to benefit from the binning transformation even after the exclusion of categorical variables.

Table 5.23: Results - estimation without categorical variables

Method	Data set	Type	AUC	F2-score	Brier score	KS statistic	Partial GINI Index	H-measure
LogReg	binned > raw	-	4	3	2	3	3	3
LogReg	Average	binned	**0.744**	**0.518**	**0.142**	**0.401**	**0.412**	**0.259**
LogReg	Average	raw	0.724	0.493	0.143	0.391	0.362	0.244
LogReg	p <= 0.05	-	1	1	1	1	1	1
DecTree	binned > raw	-	1	1	2	1	2	2
DecTree	Average	binned	0.731	0.514	**0.144**	0.373	**0.261**	0.227
DecTree	Average	raw	**0.735**	**0.545**	0.145	**0.399**	0.24	**0.239**
DecTree	p <= 0.05	-	0	1	0	0	0	0
RandForest	binned > raw	-	0	0	0	0	0	0
RandForest	Average	binned	0.746	0.512	0.138	0.398	0.381	0.262
RandForest	Average	raw	**0.765**	**0.531**	**0.136**	**0.435**	**0.454**	**0.285**
RandForest	p <= 0.05	-	0	0	0	0	0	0
NN	binned > raw	-	4	3	3	3	3	3
NN	Average	binned	**0.743**	**0.521**	**0.157**	**0.4**	**0.357**	**0.257**
NN	Average	raw	0.664	0.42	0.175	0.296	0.196	0.163
NN	p <= 0.05	-	3	2	3	2	4	2
GaussNB	binned > raw	-	3	4	3	3	2	3
GaussNB	Average	binned	**0.736**	**0.521**	**0.162**	**0.392**	**0.226**	**0.248**
GaussNB	Average	raw	0.688	0.453	0.233	0.321	0.214	0.178
GaussNB	p <= 0.05	-	3	3	2	2	1	2

The table presents the summarized results of calculating the six evaluation metrics (columns (3)-(8)) on the test set comprising 20% observations of each data set. Only four data sets containing categorical variables were considered: CA, DCCCT, HCDR, and SGCC. Categorical variables were removed from the data sets prior to estimation. Values in **bold** signal a better average performance of the given model type for the evaluation metric. Note that except for the Brier score, the higher the value of the metric, the better. The first row for each method shows the number of data sets for which the binned model outperformed the raw model for a given evaluation metric. The second and third rows for each estimation method show the average value for each metric for each model type across all data sets. The last row shows the number of data sets for which the null hypothesis of the raw model performing better or equal to the binned model was rejected. For detailed results, see Tables B.12 and B.13.

Source: Author's computations

5.12 One-hot encoding

The last quantitative supplementary analysis performed in this thesis is the utilization of one-hot encoding of the binned variables. In all the previous analyses, the resulting discretized features were encoded using the Weight of Evidence. While this is the industry standard, one-hot encoding is one of the most popular forms of categorical variable transformation. In addition, the question stands whether one-hot encoding of the binned variables still surpasses the usage of raw features.

Based on the results in Table 5.24, the binning transformation seems to positively affect performance even when a different form of encoding is utilized. For logistic regression, the null hypothesis of the raw model performing better or equal to the binned model is rejected only for the larger data sets. However, the HCDR does not seem to benefit from the transformation. It can be argued that the large number of variables inflated by the usage of one-hot encoding may result in very few degrees of freedom, and thus worse performance.

However, the neural network and the Bernoulli Naïve Bayes appear to benefit from the binning transformation for all large data sets, including the HCDR. As a result, this thesis provides robust evidence in favor of binning transformation for these two estimation methods. On the other hand, the findings for the decision tree and Random Forest algorithms seem to follow the trend of the previous analyses.

Table 5.24: Results - estimation with one-hot encoding

Method	Data set	Type	AUC	F2-score	Brier score	KS statistic	Partial GINI Index	H-measure
LogReg	binned > raw	-	5	4	5	4	2	4
LogReg	Average	binned	**0.759**	**0.477**	**0.122**	**0.415**	**0.397**	**0.276**
LogReg	Average	raw	0.713	0.428	0.127	0.362	0.357	0.213
LogReg	p <= 0.05	-	2	2	2	2	2	2
DecTree	binned > raw	-	1	1	1	1	1	1
DecTree	Average	binned	0.742	0.492	**0.127**	0.386	**0.379**	0.238
DecTree	Average	raw	**0.75**	**0.509**	0.127	**0.416**	0.308	**0.254**
DecTree	p <= 0.05	-	0	1	0	0	0	0
RandForest	binned > raw	-	0	0	0	0	0	0
RandForest	Average	binned	0.759	0.482	0.123	0.415	0.418	0.263
RandForest	Average	raw	**0.777**	**0.497**	**0.12**	**0.449**	**0.487**	**0.296**
RandForest	p <= 0.05	-	0	0	0	0	0	0
NN	binned > raw	-	5	4	5	5	4	5
NN	Average	binned	**0.757**	**0.48**	**0.122**	**0.428**	**0.408**	**0.281**
NN	Average	raw	0.685	0.397	0.152	0.322	0.265	0.181
NN	p <= 0.05	-	3	3	5	4	4	4
BernoulliNB	binned > raw	-	4	4	4	4	4	4
BernoulliNB	Average	binned	**0.747**	**0.447**	**0.132**	**0.399**	**0.394**	**0.239**
BernoulliNB	Average	raw	0.689	0.406	0.2	0.311	0.242	0.168
BernoulliNB	p <= 0.05	-	4	2	3	3	3	4

The table presents the summarized results of calculating the six evaluation metrics (columns (3)-(8)) on the test set comprising 20% observations of each data set. For the binned model, the discretized variables were one-hot encoded. Values in **bold** signal a better average performance of the given model type for the evaluation metric. Note that except for the Brier score, the higher the value of the metric, the better. The first row for each method shows the number of data sets for which the binned model outperformed the raw model for a given evaluation metric. The second and third rows for each estimation method show the average value for each metric for each model type across all data sets. The last row shows the number of data sets for which the null hypothesis of the raw model performing better or equal to the binned model was rejected. For detailed results, see Tables B.14 and B.15.

Source: Author's computations

5.13 Qualitative case study

All the previous analyses employed a quantitative approach. To at least partially imitate the actual development of a credit-scoring model, a more qualitative approach will be applied using the logistic regression on the GMSC data set.

For the estimation of the model with raw variables, outliers from both tails will be trimmed to alleviate the distortion of the results.[13] In addition, upon the inspection of the histograms of the variables, the logarithmic transformation was applied to two of them to attempt to satisfy the assumption of normality.[14] As for the second model, the binning transformation is expected to handle outliers well, and thus the data was unaltered.

The second step of the development process is the forward sequential feature selection based on the average value of AUC. For each candidate, 5-fold cross-validation is performed, and the average AUC over the five iterations is considered for comparison. The final models were re-estimated on the entire training set and evaluated on the testing set utilizing the same procedures as in the previous analyses. Table 5.25 presents the results.

As can be seen, the raw model's performance improved substantially compared to the main analysis. Consequently, the appropriate treatment of outliers appears to greatly impact the logistic regression. The performance of the binned model was enhanced only slightly, and the overall results for both models seem to be very similar. An exception is the AUC and Brier score metrics for which the null hypothesis is rejected.

To conclude, taking a qualitative approach seems to reduce the performance differences between the models for the logistic regression. However, with the increasing number of variables, the feasibility of the qualitative approach rapidly decreases. The advantage of the binning transformation is its ability to handle conveniently raw unaltered data, including outliers.

[13] The 1% and 99% quantiles were used for trimming.
[14] *DebtRatio* and *MonthlyIncome*

Table 5.25: Results - qualitative analysis

	AUC	F2-score	Brier score	KS statistic	Partial GINI Index	H-measure
raw	0.813	0.328	0.058	0.488	0.582	**0.316**
binned	**0.815**	**0.337**	**0.056**	**0.492**	**0.585**	0.315
p	**0.046**	0.619	**0.000**	0.141	0.195	0.530

The table presents the results of calculating the six evaluation metrics for the logistic regression on the test set of the GMSC data set. Values in **bold** signal a better performance of the given model type for the given metric. Note that except for the Brier score, the higher the value of the metric, the better. The last row shows the p-value of the null hypothesis that the raw model performs better or equally to the binned model for the given evaluation metric.

Source: Author's computations

Chapter 6

Conclusion

The development of credit risk models is a thoroughly scrutinized topic in the extant literature. However, while many studies focus on improving performance using advanced methods, the impact of the data preprocessing phase appears to be fairly neglected. In addition, the existing regulatory framework requires credit risk models to be interpretable, hindering the usage of complicated estimation methods. As a result, the potential performance enhancements generated by methods capitalizing on rapidly increasing computational power cannot be exploited in practice. One way to make the estimation more interpretable is to discretize numerical features and a subsequent Weight of Evidence transformation. However, the effect of this transformation on performance has not yet been thoroughly investigated. Consequently, this thesis attempted to inspect the impact of the binning transformation on the performance of five widely known estimation methods.

To address the methodological deficiencies of the existing articles, six evaluation metrics along with a statistical test were employed to compare the performance of the questioned methods. The evaluation metrics are divided into three groups, each inspecting a different aspect of model performance. In addition, five publicly available credit risk data sets of different sizes and characteristics were utilized to secure the robustness of the results.

Three of the five considered estimation methods seem to significantly benefit from utilizing the binning transformation of numerical features. Namely, the logistic regression, the feedforward artificial neural network, and the Naïve Bayes classifier all seem to achieve higher performance when the binning transforma-

tion is utilized prior to estimation. The most substantial impact appears to be on the model's ability to differentiate between eligible and ineligible customers. The results are particularly strong for moderately-sized data sets which represent the industry standard in credit risk modeling. The findings for the neural network and the Naïve Bayes classifier are robust to the inclusion of missing values, the elimination of outliers, and the exclusion of categorical features. Furthermore, the results hold even for utilizing one-hot encoding of the binned variables. In the case of logistic regression, the effect appears to be weakened by the presence of unavailable and extreme observations. In addition, employing a qualitative approach seems to diminish the differences in performance between the raw and binned model. As a result, the main strength of the binning algorithm appears to be its ability to handle outlying observations conveniently without individual investigation of each variable in the data set.

On the other hand, no statistically significant effect of the binning transformation on performance was detected for two tree-based methods, namely, the CART decision tree and the Random Forest model.

The findings of this thesis implicate several recommendations for the development of credit scoring models. While the application of the binning transformation and subsequent WoE encoding is usually justified by its ability to handle missing values, reduce non-linearity, increase interpretability, and alleviate the impact of outliers, the performed analyses suggest that it also positively affects performance. However, since the results do not seem to be entirely robust for small data sets, the data size should be taken into account when applying the transformation. In addition, even though a qualitative approach is usually employed to develop credit risk models, the need for a qualitative procedure grows with increasing dimensions of the data. The binning transformation provides a convenient way of handling missing values and outliers without extensive user input.

The selection of an appropriate algorithm is vital since the effects do not seem to be homogeneous across methods. The recommended frameworks based on the results of this thesis are the Naïve Bayes classifier and the neural network. While the industry standard logistic regression appears to benefit from the binning transformation, the data size, the presence of missing values, and the presence of outliers, all need to be considered.

The analysis performed within the current thesis is not without limitations.

Firstly, the findings relate only to the utilized binning algorithm. While they may be generalized to most supervised statistics-based algorithms, the results may differ for other types. Secondly, only the five most common machine learning estimation methods were considered. The impact of the binning transformation on the performance of more complex algorithms such as heterogenous ensembles is an opportunity for future research. Thirdly, while the utilized data sets are expected to cover the main aspects of actual credit scoring data sets, unseen characteristics may affect the results. Lastly, the stringency of the existing regulatory framework may hinder the utilization of advanced machine learning methods even after applying the binning transformation since the improvement in interpretability may not be sufficient. Therefore, more sophisticated ways of introducing interpretability into "black-box" models may be required.

Bibliography

Abraham, R., J. B. Simha, & S. S. Iyengar (2006): "A comparative analysis of discretization methods for medical datamining with naive bayesian classifier." In "9th International Conference on Information Technology (ICIT'06)," pp. 235–236.

Addo, P. M., D. Guegan, & B. Hassani (2018): "Credit risk analysis using machine and deep learning models." *Risks* **6(2)**: pp. 1–20.

Agre, G. & S. Peev (2002): "On supervised and unsupervised discretization." *Cybernetics and information technologies* **2(2)**: pp. 43–57.

Ali, J., R. Khan, N. Ahmad, & I. Maqsood (2012): "Random forests and decision trees." *International Journal of Computer Science Issues (IJCSI)* **9(3)**: pp. 272–278.

Alin, A. (2010): "Multicollinearity." *WIREs Computational Statistics* **2(3)**: pp. 370–374.

Anna Montoya, Kirill Odintsov, M. K. (2018): "Home credit default risk." Accessed: 2022-10-03, `https://kaggle.com/competitions/home-credit-default-risk`.

Augasta, M. G. & T. Kathirvalavakumar (2013): "An empirical comparison of discretization methods for neural classifier." In R. Prasath & T. Kathirvalavakumar (editors), "Mining Intelligence and Knowledge Exploration," pp. 38–49. Cham: Springer International Publishing.

Baesens, B., T. Van Gestel, S. Viaene, M. STEPANOVA, J. Suykens, & J. Vanthienen (2003): "Benchmarking state-of-the-art classification algorithms for credit scoring." *Journal of the Operational Research Society* **54(6)**: pp. 627–635.

Bekkar, M., H. Djema, & T. Alitouche (2013): "Evaluation measures for models assessment over imbalanced data sets." *Journal of Information Engineering and Applications* **3**: pp. 27–38.

Bhatore, S., L. Mohan, & Y. R. Reddy (2020): "Machine learning techniques for credit risk evaluation: a systematic literature review." *Journal of Banking and Financial Technology* **4(1)**: pp. 111–138.

Bircanoğlu, C. & N. Arica (2018): "A comparison of activation functions in artificial neural networks." In "2018 26th Signal Processing and Communications Applications Conference (SIU)," pp. 1–4.

BIS (2011): *Basel III: A Global Regulatory Framework for More Resilient Banks and Banking Systems*. Basel, Switzerland: Bank for International Settlements.

Branco, P., L. Torgo, & R. P. Ribeiro (2016): "A survey of predictive modeling on imbalanced domains." *ACM Comput. Surv.* **49(2)**.

Breeden, J. (2021): "A survey of machine learning in credit risk." *Journal of Credit Risk* **17(3)**: pp. 1–62.

Breiman, L. (2001): "Random forests." *Machine Learning* **45(1)**: pp. 5–32.

Breiman, L., J. Friedman, C. Stone, & R. Olshen (1984): *Classification and Regression Trees*. Taylor & Francis.

Brier, G. W. (1950): "Verification of forecasts expressed in terms of probability." *Monthly Weather Review* **78(1)**: pp. 1–3.

Bühlmann, Peterand van de Geer, S. (2011): "Generalized linear models and the lasso." In "Statistics for High-Dimensional Data: Methods, Theory and Applications," pp. 45–53. Berlin, Heidelberg: Springer Berlin Heidelberg.

Chen, T. & C. Guestrin (2016): "Xgboost: A scalable tree boosting system." In "Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining," ACM.

Collingridge, D. S. (2013): "A primer on quantitized data analysis and permutation testing." *Journal of Mixed Methods Research* **7(1)**: pp. 81–97.

Cukierski, W. (2011): "Give me some credit." Credit Fusion. Accessed: 2022-12-27, `https://kaggle.com/competitions/GiveMeSomeCredit`.

Dash, R., R. L. Paramguru, & R. Dash (2011): "Comparative analysis of supervised and unsupervised discretization techniques." *International Journal of Advances in Science and Technology* **2(3)**: pp. 29–37.

DCCCT (2016): "Default of credit card clients in taiwan." UCI Machine Learning Repository. Accessed: 2022-12-27, `http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients`.

De La Bourdonnaye, F. & F. Daniel (2021): "Evaluating categorical encoding methods on a real credit card fraud detection database." *CoRR* **abs/2112.12024**.

Dougherty, J., R. Kohavi, & M. Sahami (1995): "Supervised and unsupervised discretization of continuous features." In A. Prieditis & S. Russell (editors), "Machine Learning Proceedings 1995," pp. 194–202. San Francisco (CA): Morgan Kaufmann.

Esmaily, H., M. Tayefi, H. Doosti, M. Ghayour-Mobarhan, H. Nezami, & A. Amirabadizadeh (2018): "A comparison between decision tree and random forest in determining the risk factors associated with type 2 diabetes." *Journal of Research in Health Sciences* **18(2)**: p. 412.

Flynn, C. (2023): "Pypi stats." Accessed: 2023-07-10, `https://pypistats.org/packages/optbinning`.

Friedman, J. H., T. Hastie, & R. Tibshirani (2010): "Regularization paths for generalized linear models via coordinate descent." *Journal of Statistical Software* **33(1)**: pp. 1–22.

Good, P. I. (2004): *Permutation, Parametric, and Bootstrap Tests of Hypotheses (Springer Series in Statistics)*. Berlin, Heidelberg: Springer-Verlag.

Grömping, U. (2019): "South german credit data: Correcting a widely used data set." *Technical Report Report 4/2019*, Beuth University of Applied Sciences Berlin.

Gunnarsson, B. R., S. vanden Broucke, B. Baesens, M. Ă"skarsdĂłttir, & W. Lemahieu (2021): "Deep learning for credit scoring: Do or don't?" *European Journal of Operational Research* **295(1)**: pp. 292–305.

Hair, J., G. T. M. Hult, C. Ringle, & M. Sarstedt (2022): *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM).* Sage Publishing.

Hand, D. J. (2009): "Measuring classifier performance: a coherent alternative to the area under the roc curve." *Machine Learning* **77(1)**: pp. 103–123.

Hawkins, D. M. (2004): "The problem of overfitting." *Journal of Chemical Information and Computer Sciences* **44(1)**: pp. 1–12.

Heaton, J. (2008): *Introduction to Neural Networks for Java, 2nd Edition.* Heaton Research, Inc., 2nd edition.

Hernandez-Orallo, J., P. Flach, & C. Ferri (2011): "Brier curves: a new cost-based visualisation of classifier performance." In "Proceedings of the 28th International Conference on Machine Learning, ICML 2011," pp. 585–592.

Higham, N. J. (2002): *Accuracy and Stability of Numerical Algorithms.* Society for Industrial and Applied Mathematics, second edition.

Jadhav, A., D. Pramod, & K. Ramanathan (2019): "Comparison of performance of data imputation methods for numeric dataset." *Applied Artificial Intelligence* **33(10)**: pp. 913–933.

Jennings, D. E. (1986): "Outliers and residual distributions in logistic regression." *Journal of the American Statistical Association* **81(396)**: pp. 987–990.

Kerber, R. (1992): "Chimerge: Discretization of numeric attributes." In "Proceedings of the Tenth National Conference on Artificial Intelligence," AAAI'92, pp. 123–128. AAAI Press.

Khandani, A. E., A. J. Kim, & A. W. Lo (2010): "Consumer credit-risk models via machine-learning algorithms." *Journal of Banking & Finance* **34(11)**: pp. 2767–2787.

Kohavi, R. & M. Sahami (1996): "Error-based and entropy-based discretization of continuous features." In "Proceedings of the Second International Conference on Knowledge Discovery and Data Mining," KDD'96, pp. 114–119. AAAI Press.

Kokoska, S. & D. Zwillinger (2000): *CRC Standard Probability and Statistics Tables and Formulae, Student Edition.* Chapman & Hall.

Konietschke, F. & M. Pauly (2014): "Bootstrapping and permuting paired t-test type statistics." *Statistics and Computing* **24(3)**: pp. 283–296.

Kornbrot, D. (2005): "Point biserial correlation." In "Encyclopedia of Statistics in Behavioral Science," John Wiley & Sons, Ltd.

Kotsiantis, S. & D. Kanellopoulos (2005): "Discretization techniques: A recent survey." *GESTS International Transactions on Computer Science and Engineering* **32**: pp. 47–58.

Krizhevsky, A., I. Sutskever, & G. E. Hinton (2012): "Imagenet classification with deep convolutional neural networks." In F. Pereira, C. Burges, L. Bottou, & K. Weinberger (editors), "Advances in Neural Information Processing Systems," volume 25. Curran Associates, Inc.

Lavangnananda, K. & S. Chattanachot (2017): "Study of discretization methods in classification." In "2017 9th International Conference on Knowledge and Smart Technology (KST)," pp. 50–55.

LeCun, Y., Y. Bengio, & G. Hinton (2015): "Deep learning." *Nature* **521(7553)**: pp. 436–444.

Lessmann, S., B. Baesens, H.-V. Seow, & L. C. Thomas (2015): "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research." *European Journal of Operational Research* **247(1)**: pp. 124–136.

Leung, K., F. Cheong, C. Cheong, S. O'Farrell, & R. Tissington (2008): "Building a scorecard in practice." In "The 7th International Conference on Computational Intelligence in Economics and Finance," pp. 1–7.

Liu, H. & R. Setiono (1997): "Feature selection via discretization." *IEEE Transactions on Knowledge and Data Engineering* **9(4)**: pp. 642 – 645.

Lustgarten, J. L., V. Gopalakrishnan, H. Grover, & S. Visweswaran (2008): "Improving classification performance with discretization on biomedical datasets." *AMIA Annu Symp Proc* **2008**: pp. 445–449.

Mandrekar, J. N. (2010): "Receiver operating characteristic curve in diagnostic test assessment." *Journal of Thoracic Oncology* **5(9)**: pp. 1315–1316.

Marcoulides, K. & T. Raykov (2018): "Evaluation of variance inflation factors in regression models using latent variable modeling methods." *Educational and Psychological Measurement* **79(5)**: pp. 874–882.

Navas-Palencia, G. (2020): "Optimal binning: mathematical programming formulation." *CoRR* **abs/2001.08025**.

Navas-Palencia, G. (2023): "Optbinning: The python optimal binning library." Accessed: 2023-07-10, `http://gnpalencia.org/optbinning/`.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, & E. Duchesnay (2011): "Scikit-learn: Machine learning in Python." *Journal of Machine Learning Research* **12**: pp. 2825–2830.

Peng, Y. & G. Kou (2008): "A comparative study of classification methods in financial risk detection." In "2008 Fourth International Conference on Networked Computing and Advanced Information Management," volume 2, pp. 9–12.

Petch, J., S. Di, & W. Nelson (2022): "Opening the black box: The promise and limitations of explainable machine learning in cardiology." *Canadian Journal of Cardiology* **38(2)**: pp. 204–213. Focus Issue: New Digital Technologies in Cardiology.

Potdar, K., T. Pardawala, & C. Pai (2017): "A comparative study of categorical variable encoding techniques for neural network classifiers." *International Journal of Computer Applications* **175(4)**: pp. 7–9.

Powers, D. M. W. (2012): "The problem of area under the curve." In "2012 IEEE International Conference on Information Science and Technology," pp. 567–573.

Prajwala, T. (2015): "A comparative study on decision tree and random forest using r tool." *International journal of advanced research in computer and communication engineering* **4(1)**: pp. 196–199.

Pundir, S. & R. Seshadri (2012): "A novel concept of partial lorenz curve and partial gini index." *International Journal of Engineering, Science and Innovative Technology* **1(2)**: pp. 296–301.

Putri, N. H., M. Fatekurohman, & I. M. Tirta (2021): "Credit risk analysis using support vector machines algorithm." *Journal of Physics: Conference Series* **1836(1)**: p. 010239.

Quinlan, Q. (2017): "Credit approval." UCI Machine Learning Repository. Accessed: 2022-12-27, `https://archive.ics.uci.edu/ml/datasets/Credit+Approval`.

Raymaekers, J., W. Verbeke, & T. Verdonck (2022): "Weight-of-evidence through shrinkage and spline binning for interpretable nonlinear classification." *Applied Soft Computing* **115(4)**: p. 108160.

SGCC (2019): "South german credit card." UCI Machine Learning Repository. Accessed: 2022-12-27, `https://archive.ics.uci.edu/ml/datasets/South+German+Credit+%28UPDATE%29`.

Sharma, D. (2011): "Evidence in favor of weight of evidence and binning transformations for predictive modeling." *Econometrics: Econometric & Statistical Methods - General eJournal* .

Siddiqi, N. (2012): *Scorecard Development Process, Stage 4: Scorecard Development*, chapter 6, pp. 73–130. John Wiley & Sons, Ltd.

Smirnov, N. V. (1939): "Estimate of deviation between empirical distribution functions in two independent samples." *Bulletin Moscow University* **2(2)**: pp. 3–16.

Soria, D., J. M. Garibaldi, F. Ambrogi, E. M. Biganzoli, & I. O. Ellis (2011): "A "non-parametric" version of the naive bayes classifier." *Knowledge-Based Systems* **24(6)**: pp. 775–784.

Statnikov, A., C. F. Aliferis, I. Tsamardinos, D. Hardin, & S. Levy (2005): "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis." *Bioinformatics* **21(5)**: pp. 631–643.

Ventura, D. & T. R. Martinez (1995): "An empirical comparison of discretization methods." In "Proceedings of the Tenth International Symposium on Computer and Information Sciences," pp. 443–450.

Verster, T. (2018): "Autobin: a predictive approach towards automatic binning using data splitting." *South African Statistical Journal* **52(2)**: pp. 139–155.

Wójciak, M. & A. Ćupińska Dubicka (2018): "Empirical comparison of methods of data discretization in learning probabilistic models." *Advances in Computer Science Research* **14**: pp. 177–192.

Wooldridge, J. M. (2013): *Introductory econometrics a modern approach*. South-Western, Cengage Learning, 5 edition.

Wu, Q., D. Bell, T. Mcginnity, G. Prasad, G. Qi, & X. Huang (2006): "Improvement of decision accuracy using discretization of continuous attributes." In "Fuzzy Systems and Knowledge Discovery: Second International Conference," pp. 674–683. Springer Berlin Heidelberg.

Yeh, I.-C. & C. hui Lien (2009): "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients." *Expert Systems with Applications* **36(2)**: pp. 2473–2480.

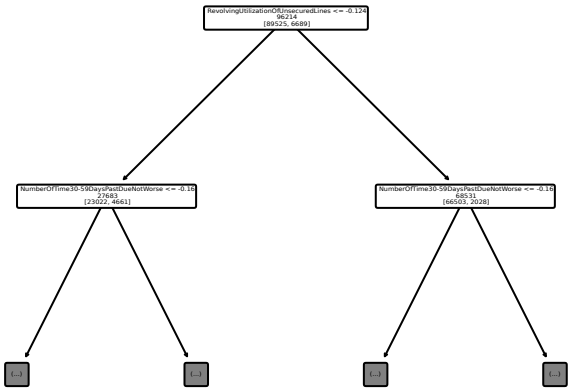
Zeng, G. (2013): "Metric divergence measures and information value in credit scoring." *Journal of Mathematics* **2013(1)**: pp. 1–10.

Zeng, G. (2014): "A necessary condition for a good binning algorithm in credit scoring." *Applied Mathematical Sciences* **8(65)**: pp. 3229–3242.

Appendix A

Figures

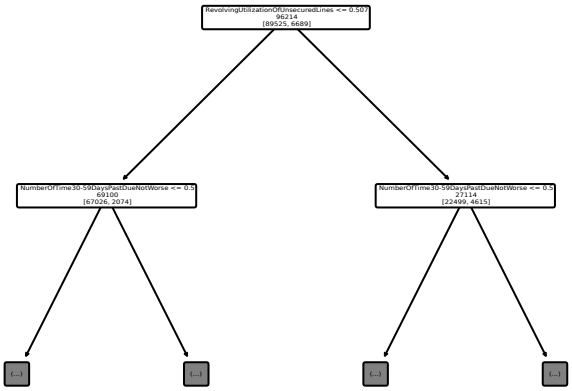
Figure A.1: Decision tree structure (GMSC, binned)



The figure shows the top of the structure of a decision tree estimated on binned features of the GMSC data set. The whole structure is not presented due to space constraints.

Source: Author's computations. Generated using the Python package "scikit-learn" (Pedregosa *et al.* 2011).

Figure A.2: Decision tree structure (GMSC, raw)



The figure shows the top of the structure of a decision tree estimated on raw features of the GMSC data set. The whole structure is not presented due to space constraints.

Source: Author's computations. Generated using the Python package "scikit-learn" (Pedregosa *et al.* 2011).

Appendix B

Tables

Table B.1: Home Credit Default Risk data set - summary

Statistic	n	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
TARGET	307,511	0.081	0.272	0	0	0	0	1
CNT_CHILDREN	307,511	0.417	0.722	0	0	0	1	19
AMT_INCOME_TOTAL	307,511	168,797.900	237,123.100	25,650.000	112,500.000	147,150.000	202,500.000	117,000,000.000
AMT_CREDIT	307,511	599,026.000	402,490.800	45,000.000	270,000.000	513,531.000	808,650.000	4,050,000.000
AMT_ANNUITY	307,499	27,108.570	14,493.740	1,615.500	16,524.000	24,903.000	34,596.000	258,025.500
AMT_GOODS_PRICE	307,233	538,396.200	369,446.500	40,500.000	238,500.000	450,000.000	679,500.000	4,050,000.000
REGION_POPULATION_RELATIVE	307,511	0.021	0.014	0.0003	0.010	0.019	0.029	0.073
DAYS_BIRTH	307,511	−16,037.000	4,363.989	−25,229	−19,682	−15,750	−12,413	−7,489
DAYS_EMPLOYED	307,511	63,815.050	141,275.800	−17,912	−2,760	−1,213	−289	365,243
DAYS_REGISTRATION	307,511	−4,986.120	3,522.886	−24,672.000	−7,479.500	−4,504.000	−2,010.000	0.000
DAYS_ID_PUBLISH	307,511	−2,994.202	1,509.450	−7,197	−4,299	−3,254	−1,720	0
OWN_CAR_AGE	104,582	12.061	11.945	0	5	9	15	91
CNT_FAM_MEMBERS	307,509	2.153	0.911	1	2	2	3	20
EXT_SOURCE_1	134,133	0.502	0.211	0.015	0.334	0.506	0.675	0.963
EXT_SOURCE_2	306,851	0.514	0.191	0.00000	0.392	0.566	0.664	0.855
EXT_SOURCE_3	246,546	0.511	0.195	0.001	0.371	0.535	0.669	0.896
APARTMENTS_AVG	151,450	0.117	0.108	0.000	0.058	0.088	0.148	1.000
BASEMENTAREA_AVG	127,568	0.088	0.082	0.000	0.044	0.076	0.112	1.000
YEARS_BEGINEXPLUATATION_AVG	157,504	0.978	0.059	0.000	0.977	0.982	0.987	1.000
YEARS_BUILD_AVG	103,023	0.752	0.113	0.000	0.687	0.755	0.823	1.000
COMMONAREA_AVG	92,646	0.045	0.076	0.000	0.008	0.021	0.052	1.000
ELEVATORS_AVG	143,620	0.079	0.135	0.000	0.000	0.000	0.120	1.000
ENTRANCES_AVG	152,683	0.150	0.100	0.000	0.069	0.138	0.207	1.000
FLOORSMAX_AVG	154,491	0.226	0.145	0.000	0.167	0.167	0.333	1.000
FLOORSMIN_AVG	98,869	0.232	0.161	0.000	0.083	0.208	0.375	1.000
LANDAREA_AVG	124,921	0.066	0.081	0.000	0.019	0.048	0.086	1.000
LIVINGAPARTMENTS_AVG	97,312	0.101	0.093	0.000	0.050	0.076	0.121	1.000
LIVINGAREA_AVG	153,161	0.107	0.111	0.000	0.045	0.074	0.130	1.000
NONLIVINGAPARTMENTS_AVG	93,997	0.009	0.048	0.000	0.000	0.000	0.004	1.000
NONLIVINGAREA_AVG	137,829	0.028	0.070	0.000	0.000	0.004	0.028	1.000
OBS_30_CNT_SOCIAL_CIRCLE	306,490	1.422	2.401	0	0	0	2	348
DEF_30_CNT_SOCIAL_CIRCLE	306,490	0.143	0.447	0	0	0	0	34
OBS_60_CNT_SOCIAL_CIRCLE	306,490	1.405	2.380	0	0	0	2	344
DEF_60_CNT_SOCIAL_CIRCLE	306,490	0.100	0.362	0	0	0	0	24
DAYS_LAST_PHONE_CHANGE	307,510	−962.859	826.808	−4,292	−1,570	−757	−274	0
AMT_REQ_CREDIT_BUREAU_HOUR	265,992	0.006	0.084	0	0	0	0	4
AMT_REQ_CREDIT_BUREAU_DAY	265,992	0.007	0.111	0	0	0	0	9
AMT_REQ_CREDIT_BUREAU_WEEK	265,992	0.034	0.205	0	0	0	0	8
AMT_REQ_CREDIT_BUREAU_MON	265,992	0.267	0.916	0	0	0	0	27
AMT_REQ_CREDIT_BUREAU_QRT	265,992	0.265	0.794	0	0	0	0	261
AMT_REQ_CREDIT_BUREAU_YEAR	265,992	1.900	1.869	0	0	1	3	25

Source: Anna Montoya (2018)

Table B.2: Credit Approval data set - summary

Statistic	n	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
V2	678	31.568	11.958	13.750	22.602	28.460	38.230	80.250
V3	690	4.759	4.978	0.000	1.000	2.750	7.207	28.000
V8	690	2.223	3.347	0.000	0.165	1.000	2.625	28.500
V11	690	2.400	4.863	0	0	0	3	67
V14	677	184.015	173.807	0	75	160	276	2,000
V15	690	1,017.386	5,210.103	0	0	5	395.5	100,000
Target	690	0.445	0.497	0	0	0	1	1

Source: Quinlan (2017)

Table B.3: Default of Credit Card Clients in Taiwan data set - summary

Statistic	n	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
LIMIT_BAL	30,000	167,484.300	129,747.700	10,000	50,000	140,000	240,000	1,000,000
SEX_FEMALE	30,000	0.604	0.489	0	0	1	1	1
EDUCATION	30,000	1.853	0.790	0	1	2	2	6
MARRIAGE	30,000	0.455	0.498	0	0	0	1	1
AGE	30,000	35.486	9.218	21	28	34	41	79
PAY_0	30,000	−0.017	1.124	−2	−1	0	0	8
PAY_2	30,000	−0.134	1.197	−2	−1	0	0	8
PAY_3	30,000	−0.166	1.197	−2	−1	0	0	8
PAY_4	30,000	−0.221	1.169	−2	−1	0	0	8
PAY_5	30,000	−0.266	1.133	−2	−1	0	0	8
PAY_6	30,000	−0.291	1.150	−2	−1	0	0	8
BILL_AMT1	30,000	51,223.330	73,635.860	−165,580	3,558.8	22,381.5	67,091	964,511
BILL_AMT2	30,000	49,179.070	71,173.770	−69,777	2,984.8	21,200	64,006.2	983,931
BILL_AMT3	30,000	47,013.150	69,349.390	−157,264	2,666.2	20,088.5	60,164.8	1,664,089
BILL_AMT4	30,000	43,262.950	64,332.860	−170,000	2,326.8	19,052	54,506	891,586
BILL_AMT5	30,000	40,311.400	60,797.160	−81,334	1,763	18,104.5	50,190.5	927,171
BILL_AMT6	30,000	38,871.760	59,554.110	−339,603	1,256	17,071	49,198.2	961,664
PAY_AMT1	30,000	5,663.580	16,563.280	0	1,000	2,100	5,006	873,552
PAY_AMT2	30,000	5,921.164	23,040.870	0	833	2,009	5,000	1,684,259
PAY_AMT3	30,000	5,225.682	17,606.960	0	390	1,800	4,505	896,040
PAY_AMT4	30,000	4,826.077	15,666.160	0	296	1,500	4,013.2	621,000
PAY_AMT5	30,000	4,799.388	15,278.310	0	252.5	1,500	4,031.5	426,529
PAY_AMT6	30,000	5,215.503	17,777.470	0	117.8	1,500	4,000	528,666
Target	30,000	0.221	0.415	0	0	0	0	1

Source: DCCCT (2016)

Table B.4: South German Credit Card data set - summary

Statistic	n	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
duration	1,000	20.903	12.059	4	12	18	24	72
amount	1,000	3,271.248	2,822.752	250	1,365.5	2,319.5	3,972.2	18,424
age	1,000	35.542	11.353	19	27	33	42	75
Target	1,000	0.300	0.458	0	0	0	1	1

Source: SGCC (2019)

Table B.5: Hyperparameter grid

Classifier	Hyperparameter	Values	# of models[1]
LogReg	Regularization	ℓ_1, ℓ_2, None	101
	λ[2]	$10^{-4}, 10^{-4+\frac{1}{49}}, 10^{-4+\frac{2}{49}}, \ldots, 10^{4}$	
DecTree	Criterion	gini, entropy	16
	Max depth	5, 10, 20, 50	
	Max leaves	25, 50, 100, 500	
RandForest	# of trees	10, 50, 100, 500	32
	Criterion	gini, entropy	
	Max depth	5, 10, 20, 50	
	Max leaves	25, 50, 100, 500	
NN	# of nodes	$\frac{K}{2}$, $\frac{2K}{3}$, K[3]	54
	α	0.0001, 0.001, 0.01, 0.1, 1, 10	
	activation	ReLU, sigmoid, tanh	

[1] The number of models corresponds to a single data set. In addition, for each model, 3-fold cross-validation is performed. Moreover, each procedure is performed for the binned and non-binned cases totaling (30 * # of models) estimations
[2] λ is the inverse of regularization strength for ℓ_1 and ℓ_2. Not applied when regularization is None
[3] K represents the number of independent variables

Source: Author's selection

Table B.6: Give Me Some Credit data set - Correlation matrix

	Age	NoD	MI	DR	NoOCLaL	NoRELoL	RUoUL	NoT30-59DPDNW	NoT60-89DPDNW	NoT90DL
Age	1									
NumberOfDependents	-0.21*	1								
MonthlyIncome	0.038*	0.063*	1							
DebtRatio	0.024*	-0.041*	-0.029*	1						
NumberOfOpenCreditLinesAndLoans	0.15*	0.065*	0.091*	0.05*	1					
NumberRealEstateLoansOrLines	0.033*	0.12*	0.12*	0.12*	0.43*	1				
RevolvingUtilizationOfUnsecuredLines	-0.0059*	0.0016	0.0071*	0.004	-0.011*	0.0062*	1			
NumberOfTime30-59DaysPastDueNotWorse	-0.063*	-0.0027	-0.01*	-0.0065*	-0.055*	-0.031*	-0.0013	1		
NumberOfTime60-89DaysPastDueNotWorse	-0.057*	-0.011*	-0.011*	-0.0075*	-0.071*	-0.04*	-0.001	0.99*	1	
NumberOfTimes90DaysLate	-0.061*	-0.01*	-0.013*	-0.0083*	-0.08*	-0.045*	-0.0011	0.98*	0.99*	1
SeriousDlqin2yrs	-0.12*	0.046*	-0.02*	-0.0076*	-0.03*	-0.007*	-0.0018	0.13*	0.1*	0.12*

*Significant at the 5% significance level

Source: Author's computations

Table B.7: Model coefficients for logistic regression (GMSC, binned)

Variable	Coefficient
Intercept	-2.596
NumberOfTime30-59DaysPastDueNotWorse	-0.756
DebtRatio	-0.773
NumberRealEstateLoansOrLines	-0.634
Age	-0.395
NumberOfOpenCreditLinesAndLoans	-0.205
RevolvingUtilizationOfUnsecuredLines	-0.749
NumberOfDependents	-0.416
MonthlyIncome	-0.299

The table contains the estimated coefficients of the logistic regression on the GMSC data set. The variables were binned prior to estimation.

Source: Author's computations

Table B.8: Complete results - estimation with missing values

Method	Data set	Type	AUC	F2-score	Brier score	KS statistic	Partial GINI Index	H-measure
LogReg	CA	binned	**0.899**	0.881	**0.126**	**0.712**	**0.633**	**0.591**
LogReg	CA	raw	0.872	0.881	0.136	0.666	0.531	0.545
LogReg	GMSC	binned	**0.825**	**0.351**	**0.053**	**0.5**	**0.608**	**0.331**
LogReg	GMSC	raw	0.658	0.145	0.06	0.244	0.308	0.077
LogReg	HCDR	binned	0.724	0.247	0.07	0.335	0.429	0.156
LogReg	HCDR	raw	**0.73**	**0.255**	**0.069**	**0.34**	**0.448**	**0.166**
LogReg	binned > raw	-	2	1	2	2	2	2
LogReg	Average	binned	**0.816**	**0.493**	**0.083**	**0.516**	**0.557**	**0.359**
LogReg	Average	raw	0.753	0.427	0.088	0.417	0.429	0.263
DecTree	CA	binned	**0.879**	0.881	**0.135**	**0.655**	0.667	**0.554**
DecTree	CA	raw	0.864	**0.884**	0.156	0.606	**0.674**	0.496
DecTree	GMSC	binned	**0.82**	0.362	0.053	**0.496**	**0.611**	0.326
DecTree	GMSC	raw	0.818	**0.362**	**0.053**	0.494	0.607	**0.329**
DecTree	HCDR	binned	**0.711**	0.256	**0.07**	**0.312**	0.41	**0.145**
DecTree	HCDR	raw	0.709	**0.256**	0.07	0.311	**0.412**	0.143
DecTree	binned > raw	-	3	0	2	3	1	2
DecTree	Average	binned	**0.803**	0.499	**0.086**	**0.487**	0.562	**0.341**
DecTree	Average	raw	0.797	**0.501**	0.093	0.47	**0.564**	0.323
RandForest	CA	binned	0.897	0.895	0.119	**0.74**	0.662	**0.622**
RandForest	CA	raw	**0.909**	**0.909**	**0.119**	0.713	**0.7**	0.618
RandForest	GMSC	binned	0.826	0.345	0.053	0.507	**0.639**	0.334
RandForest	GMSC	raw	**0.833**	**0.363**	**0.052**	**0.514**	0.637	**0.352**
RandForest	HCDR	binned	0.729	0.26	0.07	0.34	0.458	0.168
RandForest	HCDR	raw	**0.734**	**0.268**	**0.07**	**0.347**	**0.468**	**0.174**
RandForest	binned > raw	-	0	0	0	1	1	1
RandForest	Average	binned	0.817	0.5	0.081	**0.529**	0.586	0.375
RandForest	Average	raw	**0.825**	**0.513**	**0.08**	0.525	**0.601**	**0.381**
NN	CA	binned	**0.891**	**0.881**	**0.126**	**0.726**	0.515	**0.595**
NN	CA	raw	0.865	0.852	0.148	0.639	**0.63**	0.514
NN	GMSC	binned	**0.824**	**0.35**	**0.053**	**0.501**	**0.606**	**0.331**
NN	GMSC	raw	0.503	0.259	0.061	0.006	0.006	0.001
NN	HCDR	binned	**0.725**	0.244	**0.07**	**0.335**	**0.435**	**0.155**
NN	HCDR	raw	0.5	**0.304**	0.08	0.002	0.0	0.0
NN	binned > raw	-	3	2	3	3	2	3
NN	Average	binned	**0.813**	**0.492**	**0.083**	**0.521**	**0.519**	**0.361**
NN	Average	raw	0.623	0.472	0.097	0.216	0.212	0.172
GaussNB	CA	binned	**0.882**	**0.881**	**0.139**	**0.697**	**0.558**	**0.57**
GaussNB	CA	raw	0.83	0.824	0.242	0.638	0.303	0.49
GaussNB	GMSC	binned	**0.812**	**0.315**	0.088	**0.484**	**0.496**	**0.309**
GaussNB	GMSC	raw	0.677	0.179	**0.066**	0.265	0.338	0.104
GaussNB	HCDR	binned	**0.656**	**0.155**	0.891	**0.275**	0.197	**0.082**
GaussNB	HCDR	raw	0.609	0.139	**0.075**	0.166	**0.217**	0.037
GaussNB	binned > raw	-	3	3	1	3	2	3
GaussNB	Average	binned	**0.784**	**0.45**	0.372	**0.485**	**0.417**	**0.32**
GaussNB	Average	raw	0.705	0.381	**0.128**	0.356	0.286	0.21
All	binned > raw	-	11	6	8	12	8	11
All	Average	binned	**0.807**	**0.487**	0.141	**0.508**	**0.528**	**0.351**
All	Average	raw	0.741	0.459	**0.097**	0.397	0.419	0.27

The table presents the results of calculating the six evaluation metrics (columns (3)-(8)) on the test set comprising 20% observations of each data set. The missing values were removed from the data set prior to estimation. For the raw model estimation, unavailable observations were replaced by mean, median, or mode. Values in **bold** signal a better performance of the given model type for the given data set. Note that except for the Brier score, the higher the value of the metric, the better. The third row from the bottom for each method shows the number of data sets for which the binned model outperformed the raw model for a given evaluation metric. The last two rows for each estimation method show the average value for each metric for each model type across all data sets. The third row from the bottom of the table shows the total number of cases where the binned model outperformed the raw model across all methods and data sets. The very last two rows of the table present total averages across all methods and data sets.

Source: Author's computations

Table B.9: Permutation tests - estimation with missing values

Method	Data set	AUC	F2-score	Brier score	KS statistic	Partial GINI Index	H-measure
LogReg	CA	**0.004**	0.666	**0.015**	0.126	0.255	0.131
LogReg	GMSC	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
LogReg	HCDR	0.982	0.906	1.000	0.794	1.000	0.985
LogReg	p <= 0.05	2	1	2	1	1	1
DecTree	CA	0.179	0.423	0.07	0.192	0.508	0.104
DecTree	GMSC	0.251	**0.001**	0.861	0.374	0.279	0.805
DecTree	HCDR	0.133	0.529	0.19	0.439	0.785	0.261
DecTree	p <= 0.05	0	1	0	0	0	0
RandForest	CA	0.832	0.953	0.532	0.442	0.56	0.458
RandForest	GMSC	1.000	1.000	1.000	0.967	0.159	1.000
RandForest	HCDR	1.000	1.000	1.000	0.975	1.000	1.000
RandForest	p <= 0.05	0	0	0	0	0	0
NN	CA	0.114	0.634	**0.012**	**0.033**	0.572	0.058
NN	GMSC	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
NN	HCDR	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
NN	p <= 0.05	2	2	3	3	2	2
GaussNB	CA	**0.000**	**0.002**	**0.000**	**0.002**	0.121	**0.001**
GaussNB	GMSC	**0.000**	**0.000**	1.000	**0.000**	**0.000**	**0.000**
GaussNB	HCDR	**0.000**	**0.000**	1.000	**0.000**	0.997	**0.000**
GaussNB	p <= 0.05	3	3	1	3	1	3

The values in the table represent the p-values of the null hypothesis that the model with raw variables performs better or equal to the binned model. As a result, p-values below 0.05 (in **bold**) signal the rejection of the null hypothesis in favor of the alternative that the binned model performs better at the 5% significance level. The last row for each estimation method shows the number of data sets for which the null is rejected for a given evaluation metric. The permutation tests were performed for 5000 repetitions.

Source: Author's computations

Table B.10: Complete results - estimation without outliers

Method	Data set	Type	AUC	F2-score	Brier score	KS statistic	Partial GINI Index	H-measure
LogReg	CA	binned	**0.922**	**0.849**	**0.113**	**0.733**	**0.765**	**0.665**
LogReg	CA	raw	0.907	0.822	0.121	0.708	0.578	0.618
LogReg	DCCCT	binned	**0.757**	**0.518**	**0.144**	**0.391**	**0.302**	**0.258**
LogReg	DCCCT	raw	0.714	0.51	0.151	0.373	0.173	0.23
LogReg	GMSC	binned	0.803	0.312	0.049	**0.483**	**0.595**	0.293
LogReg	GMSC	raw	**0.804**	**0.324**	**0.049**	0.48	0.583	**0.297**
LogReg	HCDR	binned	0.722	0.238	0.066	**0.342**	0.44	0.158
LogReg	HCDR	raw	**0.723**	**0.241**	**0.066**	0.341	**0.441**	**0.158**
LogReg	SGCC	binned	**0.63**	**0.496**	**0.178**	**0.274**	**0.102**	**0.133**
LogReg	SGCC	raw	0.613	0.467	0.179	0.273	0.101	0.126
LogReg	binned > raw	-	3	3	3	5	4	3
LogReg	Average	binned	**0.767**	**0.483**	**0.11**	**0.444**	**0.441**	**0.301**
LogReg	Average	raw	0.752	0.473	0.113	0.435	0.375	0.286
DecTree	CA	binned	**0.909**	0.889	**0.121**	**0.704**	0.039	**0.59**
DecTree	CA	raw	0.903	**0.891**	0.126	0.682	**0.065**	0.573
DecTree	DCCCT	binned	**0.756**	**0.531**	0.143	**0.394**	0.293	**0.259**
DecTree	DCCCT	raw	0.755	0.531	**0.142**	0.385	**0.306**	0.258
DecTree	GMSC	binned	**0.797**	0.305	0.049	**0.469**	0.574	0.279
DecTree	GMSC	raw	0.797	**0.329**	**0.049**	0.466	**0.579**	**0.284**
DecTree	HCDR	binned	**0.703**	0.23	0.067	**0.305**	**0.406**	0.131
DecTree	HCDR	raw	0.701	**0.233**	**0.067**	0.301	0.402	**0.131**
DecTree	SGCC	binned	**0.648**	**0.564**	**0.181**	**0.266**	0.028	**0.147**
DecTree	SGCC	raw	0.588	0.466	0.213	0.184	**0.066**	0.069
DecTree	binned > raw	-	5	2	2	5	1	3
DecTree	Average	binned	**0.762**	**0.504**	**0.112**	**0.428**	0.268	**0.281**
DecTree	Average	raw	0.749	0.49	0.119	0.404	**0.284**	0.263
RandForest	CA	binned	0.929	0.861	0.108	0.713	**0.543**	0.631
RandForest	CA	raw	**0.934**	**0.877**	**0.105**	**0.768**	0.437	**0.669**
RandForest	DCCCT	binned	0.765	**0.532**	0.142	0.411	0.324	0.271
RandForest	DCCCT	raw	**0.769**	0.529	**0.141**	**0.421**	**0.343**	**0.276**
RandForest	GMSC	binned	0.807	0.314	0.049	0.483	0.614	0.296
RandForest	GMSC	raw	**0.813**	**0.328**	**0.048**	**0.492**	**0.623**	**0.311**
RandForest	HCDR	binned	0.718	0.237	0.067	0.336	0.436	0.153
RandForest	HCDR	raw	**0.722**	**0.241**	**0.067**	**0.337**	**0.445**	**0.16**
RandForest	SGCC	binned	**0.672**	**0.498**	**0.169**	**0.305**	**0.179**	**0.183**
RandForest	SGCC	raw	0.646	0.447	0.173	0.269	0.107	0.18
RandForest	binned > raw	-	1	2	1	1	2	1
RandForest	Average	binned	**0.778**	**0.488**	0.107	0.45	**0.419**	0.307
RandForest	Average	raw	0.777	0.485	**0.107**	**0.457**	0.391	**0.319**
NN	CA	binned	**0.924**	**0.849**	**0.111**	**0.712**	0.661	**0.645**
NN	CA	raw	0.909	0.822	0.127	0.684	**0.669**	0.614
NN	DCCCT	binned	**0.761**	**0.525**	**0.142**	**0.404**	**0.318**	**0.266**
NN	DCCCT	raw	0.627	0.344	0.232	0.191	0.25	0.072
NN	GMSC	binned	**0.804**	**0.31**	**0.049**	**0.489**	**0.602**	**0.294**
NN	GMSC	raw	0.796	0.307	0.049	0.468	0.592	0.283
NN	HCDR	binned	**0.722**	**0.237**	**0.066**	**0.336**	**0.44**	**0.158**
NN	HCDR	raw	0.469	0.053	0.133	0.066	-0.046	0.0
NN	SGCC	binned	**0.665**	0.498	**0.173**	**0.278**	**0.115**	**0.164**
NN	SGCC	raw	0.583	**0.61**	0.175	0.161	0.102	0.088
NN	binned > raw	-	5	4	5	5	4	5
NN	Average	binned	**0.775**	**0.484**	**0.108**	**0.444**	**0.427**	**0.306**
NN	Average	raw	0.677	0.427	0.143	0.314	0.313	0.211
GaussNB	CA	binned	**0.923**	**0.849**	**0.12**	**0.76**	**0.747**	**0.673**
GaussNB	CA	raw	0.885	0.808	0.153	0.661	0.294	0.558
GaussNB	DCCCT	binned	**0.753**	**0.507**	**0.202**	**0.391**	**0.236**	**0.246**
GaussNB	DCCCT	raw	0.644	0.375	0.274	0.214	0.093	0.099
GaussNB	GMSC	binned	**0.792**	0.286	0.074	0.473	0.469	0.276
GaussNB	GMSC	raw	0.791	**0.297**	**0.072**	**0.474**	**0.472**	**0.287**
GaussNB	HCDR	binned	**0.677**	**0.183**	0.114	**0.278**	**0.277**	**0.1**
GaussNB	HCDR	raw	0.592	0.123	**0.07**	0.135	0.185	0.028
GaussNB	SGCC	binned	0.638	**0.49**	0.233	0.28	0.051	0.135
GaussNB	SGCC	raw	**0.661**	0.447	**0.205**	**0.29**	**0.194**	**0.148**
GaussNB	binned > raw	-	4	4	2	3	3	3
GaussNB	Average	binned	**0.756**	**0.463**	**0.148**	**0.436**	**0.356**	**0.286**
GaussNB	Average	raw	0.715	0.41	0.155	0.355	0.248	0.224
All	binned > raw	-	18	15	13	19	14	15
All	Average	binned	**0.768**	**0.484**	**0.117**	**0.44**	**0.382**	**0.296**
All	Average	raw	0.734	0.457	0.128	0.393	0.322	0.261

*

Source: Author's computations

The table presents the results of calculating the six evaluation metrics (columns (3)-(8)) on the test set comprising 20% observations of each data set. Outliers above 99% and below 1% were trimmed from all data sets. Values in **bold** signal a better performance of the given model type for the given data set. Note that except for the Brier score, the higher the value of the metric, the better. The third row from the bottom for each method shows the number of data sets for which the binned model outperformed the raw model for a given evaluation metric. The last two rows for each estimation method show the average value for each metric for each model type across all data sets. The third row from the bottom of the table shows the total number of cases where the binned model outperformed the raw model across all methods and data sets. The very last two rows of the table present total averages across all methods and data sets.

Table B.11: Permutation tests - estimation without outliers

Method	Data set	AUC	F2-score	Brier score	KS statistic	Partial GINI Index	H-measure
LogReg	CA	0.111	0.319	0.081	0.178	0.061	0.1
LogReg	GMSC	0.634	0.981	0.72	0.344	**0.016**	0.795
LogReg	DCCCT	**0.0**	0.089	**0.0**	**0.001**	**0.0**	**0.0**
LogReg	HCDR	0.743	0.754	0.825	0.459	0.689	0.493
LogReg	SGCC	0.203	0.15	0.453	0.351	0.477	0.358
LogReg	p <= 0.05	1	0	1	1	2	1
DecTree	CA	0.443	0.47	0.384	0.453	0.517	0.461
DecTree	GMSC	0.491	0.833	0.736	0.158	0.819	0.774
DecTree	DCCCT	0.492	0.478	0.847	0.135	0.848	0.575
DecTree	HCDR	0.22	0.715	0.781	0.147	0.22	0.723
DecTree	SGCC	0.134	**0.016**	**0.043**	0.256	0.559	0.095
DecTree	p <= 0.05	0	1	1	0	0	0
RandForest	CA	0.605	0.855	0.693	0.996	0.184	0.925
RandForest	GMSC	1.0	0.995	1.0	0.96	0.999	1.0
RandForest	DCCCT	0.968	0.286	0.993	0.963	0.989	0.985
RandForest	HCDR	1.0	0.921	1.0	0.319	1.0	1.0
RandForest	SGCC	0.181	0.211	0.204	0.535	0.234	0.64
RandForest	p <= 0.05	0	0	0	0	0	0
NN	CA	0.114	0.354	**0.028**	0.494	0.291	0.172
NN	GMSC	**0.001**	0.387	**0.001**	**0.005**	**0.032**	**0.012**
NN	DCCCT	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
NN	HCDR	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
NN	SGCC	**0.021**	**0.025**	0.404	0.08	0.202	0.095
NN	p <= 0.05	4	3	4	3	3	3
GaussNB	CA	**0.018**	0.113	0.062	**0.013**	**0.018**	**0.008**
GaussNB	GMSC	0.139	0.447	0.947	0.083	0.31	0.794
GaussNB	DCCCT	**0.0**	**0.0**	**0.0**	**0.0**	**0.001**	**0.0**
GaussNB	HCDR	**0.0**	**0.004**	1.0	**0.0**	**0.0**	**0.0**
GaussNB	SGCC	0.585	0.178	1.0	0.323	0.75	0.441
GaussNB	p <= 0.05	3	2	1	3	3	3

The values in the table represent the p-values of the null hypothesis that the model with raw variables performs better or equal to the binned model. As a result, p-values below 0.05 (in **bold**) signal the rejection of the null hypothesis in favor of the alternative that the binned model performs better at the 5% significance level. The last row for each estimation method shows the number of data sets for which the null is rejected for a given evaluation metric. The permutation tests were performed for 5000 repetitions.

Source: Author's computations

Table B.12: Complete results - estimation without categorical variables

Method	Data set	Type	AUC	F2-score	Brier score	KS statistic	Partial GINI Index*	H-measure
LogReg	CA	binned	**0.841**	**0.793**	**0.151**	**0.63**	**0.6**	**0.479**
LogReg	CA	raw	0.821	0.766	0.154	0.59	0.586	0.45
LogReg	DCCCT	binned	**0.762**	**0.527**	**0.137**	**0.403**	**0.317**	**0.268**
LogReg	DCCCT	raw	0.723	0.515	0.145	0.381	0.221	0.238
LogReg	HCDR	binned	**0.729**	0.255	0.068	**0.344**	0.448	0.168
LogReg	HCDR	raw	0.728	**0.256**	**0.068**	0.34	**0.45**	**0.168**
LogReg	SGCC	binned	**0.642**	**0.497**	0.214	0.228	**0.285**	**0.122**
LogReg	SGCC	raw	0.622	0.437	**0.204**	**0.252**	0.193	0.121
LogReg	binned > raw	-	4	3	2	3	3	3
LogReg	Average	binned	**0.744**	**0.518**	**0.142**	**0.401**	**0.412**	**0.259**
LogReg	Average	raw	0.724	0.493	0.143	0.391	0.362	0.244
DecTree	CA	binned	0.808	0.764	0.166	0.541	**0.076**	0.401
DecTree	CA	raw	**0.824**	**0.849**	**0.153**	**0.621**	0.013	**0.452**
DecTree	DCCCT	binned	0.765	0.536	**0.136**	0.399	0.337	**0.264**
DecTree	DCCCT	raw	**0.766**	**0.553**	0.136	**0.399**	**0.35**	0.263
DecTree	HCDR	binned	0.719	**0.269**	0.068	**0.333**	0.427	0.156
DecTree	HCDR	raw	**0.721**	0.269	**0.068**	0.332	**0.431**	**0.157**
DecTree	SGCC	binned	**0.631**	0.487	**0.208**	0.217	**0.204**	**0.087**
DecTree	SGCC	raw	0.628	**0.509**	0.223	**0.246**	0.167	0.084
DecTree	binned > raw	-	1	1	2	1	2	2
DecTree	Average	binned	0.731	0.514	**0.144**	0.373	**0.261**	0.227
DecTree	Average	raw	**0.735**	**0.545**	0.145	**0.399**	0.24	**0.239**
RandForest	CA	binned	0.832	0.793	0.153	0.608	0.562	0.46
RandForest	CA	raw	**0.844**	**0.806**	**0.15**	**0.612**	**0.712**	**0.475**
RandForest	DCCCT	binned	0.771	0.536	0.135	0.413	0.347	0.277
RandForest	DCCCT	raw	**0.777**	**0.538**	**0.135**	**0.414**	**0.366**	**0.281**
RandForest	HCDR	binned	0.728	0.254	0.068	0.343	0.456	0.167
RandForest	HCDR	raw	**0.734**	**0.263**	**0.067**	**0.352**	**0.466**	**0.177**
RandForest	SGCC	binned	0.654	0.463	0.196	0.23	0.16	0.145
RandForest	SGCC	raw	**0.704**	**0.518**	**0.191**	**0.361**	**0.272**	**0.205**
RandForest	binned > raw	-	0	0	0	0	0	0
RandForest	Average	binned	0.746	0.512	0.138	0.398	0.381	0.262
RandForest	Average	raw	**0.765**	**0.531**	**0.136**	**0.435**	**0.454**	**0.285**
NN	CA	binned	**0.836**	0.793	0.216	0.612	-	0.463
NN	CA	raw	0.834	**0.806**	**0.168**	**0.626**	0.246	**0.464**
NN	DCCCT	binned	**0.764**	**0.543**	**0.136**	**0.414**	**0.331**	**0.275**
NN	DCCCT	raw	0.66	0.385	0.24	0.241	0.219	0.1
NN	HCDR	binned	**0.73**	**0.25**	**0.068**	**0.345**	**0.451**	**0.169**
NN	HCDR	raw	0.571	0.084	0.078	0.136	0.142	0.02
NN	SGCC	binned	**0.643**	**0.497**	**0.207**	**0.228**	**0.287**	**0.121**
NN	SGCC	raw	0.589	0.405	0.214	0.181	0.178	0.069
NN	binned > raw	-	4	3	3	3	3	3
NN	Average	binned	**0.743**	**0.521**	**0.157**	**0.4**	**0.357**	**0.257**
NN	Average	raw	0.664	0.42	0.175	0.296	0.196	0.163
GaussNB	CA	binned	**0.827**	**0.806**	**0.169**	**0.608**	0.145	**0.454**
GaussNB	CA	raw	0.802	0.78	0.227	0.564	**0.217**	0.402
GaussNB	DCCCT	binned	**0.759**	**0.523**	**0.187**	**0.402**	**0.237**	**0.258**
GaussNB	DCCCT	raw	0.67	0.396	0.417	0.263	0.148	0.119
GaussNB	HCDR	binned	**0.696**	**0.242**	0.081	**0.299**	**0.302**	**0.136**
GaussNB	HCDR	raw	0.605	0.134	**0.073**	0.159	0.211	0.036
GaussNB	SGCC	binned	0.662	**0.514**	**0.212**	0.26	0.221	0.144
GaussNB	SGCC	raw	**0.675**	0.502	0.217	**0.299**	**0.28**	**0.154**
GaussNB	binned > raw	-	3	4	3	3	2	3
GaussNB	Average	binned	**0.736**	**0.521**	**0.162**	**0.392**	**0.226**	**0.248**
GaussNB	Average	raw	0.688	0.453	0.233	0.321	0.214	0.178
All	binned > raw	-	12	11	10	10	10	11
All	Average	binned	**0.74**	**0.517**	**0.149**	**0.393**	**0.326**	**0.251**
All	Average	raw	0.715	0.489	0.166	0.368	0.293	0.222

The table presents the results of calculating the six evaluation metrics (columns (3)-(8)) on the test set comprising 20% observations of each data set. Categorical variables were removed from the data sets prior to estimation. Values in **bold** signal a better performance of the given model type for the given data set. Note that except for the Brier score, the higher the value of the metric, the better. The third row from the bottom for each method shows the number of data sets for which the binned model outperformed the raw model for a given evaluation metric. The last two rows for each estimation method show the average value for each metric for each model type across all data sets. The third row from the bottom of the table shows the total number of cases where the binned model outperformed the raw model across all methods and data sets. The very last two rows of the table present total averages across all methods and data sets. *A missing value for PGI can occur when the selected probability threshold causes all predictions to belong to the same class, since then the metric is not defined.

Source: Author's computations

Table B.13: Permutation tests - estimation without categorical variables

Method	Data set	AUC	F2-score	Brier score	KS statistic	Partial GINI Index	H-measure
LogReg	CA	0.128	0.282	0.363	0.229	0.383	0.323
LogReg	DCCCT	**0.0**	**0.03**	**0.0**	**0.0**	**0.0**	**0.0**
LogReg	HCDR	0.231	0.273	0.52	0.125	0.717	0.549
LogReg	SGCC	0.088	0.057	0.932	0.376	0.051	0.107
LogReg	p <= 0.05	1	1	1	1	1	1
DecTree	CA	0.75	0.96	0.72	0.914	0.583	0.819
DecTree	DCCCT	0.62	1.0	0.333	0.694	0.823	0.609
DecTree	HCDR	0.943	**0.0**	0.865	0.367	0.926	0.718
DecTree	SGCC	0.526	0.603	0.152	0.676	0.486	0.61
DecTree	p <= 0.05	0	1	0	0	0	0
RandForest	CA	0.873	0.872	0.711	0.688	0.932	0.765
RandForest	DCCCT	1.0	0.754	1.0	0.774	0.998	0.999
RandForest	HCDR	1.0	0.997	1.0	1.0	1.0	1.0
RandForest	SGCC	0.965	0.816	0.834	0.996	0.904	0.97
RandForest	p <= 0.05	0	0	0	0	0	0
NN	CA	0.217	0.546	1.0	0.333	**0.0**	0.292
NN	DCCCT	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
NN	HCDR	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
NN	SGCC	**0.029**	0.2	**0.003**	0.172	**0.03**	0.057
NN	p <= 0.05	3	2	3	2	4	2
GaussNB	CA	**0.044**	**0.022**	**0.003**	0.093	0.419	0.057
GaussNB	DCCCT	**0.0**	**0.0**	**0.0**	**0.0**	0.101	**0.0**
GaussNB	HCDR	**0.0**	**0.0**	1.0	**0.0**	**0.0**	**0.0**
GaussNB	SGCC	0.551	0.055	0.132	0.729	0.643	0.392
GaussNB	p <= 0.05	3	3	2	2	1	2

The values in the table represent the p-values of the null hypothesis that the model with raw variables performs better or equal to the binned model. As a result, p-values below 0.05 (in **bold**) signal the rejection of the null hypothesis in favor of the alternative that the binned model performs better at the 5% significance level. The last row for each estimation method shows the number of data sets for which the null is rejected for a given evaluation metric. The permutation tests were performed for 5000 repetitions.

Source: Author's computations

Table B.14: Complete results - estimation with one hot encoding

Method	Data set	Type	AUC	F2-score	Brier score	KS statistic	Partial GINI Index	H-measure
LogReg	CA	binned	**0.847**	**0.806**	**0.147**	**0.621**	0.495	**0.485**
LogReg	CA	raw	0.821	0.766	0.154	0.59	**0.586**	0.45
LogReg	DCCCT	binned	**0.767**	**0.539**	**0.137**	**0.411**	**0.349**	**0.274**
LogReg	DCCCT	raw	0.723	0.515	0.145	0.381	0.221	0.238
LogReg	GMSC	binned	**0.816**	**0.34**	**0.056**	**0.494**	**0.588**	**0.32**
LogReg	GMSC	raw	0.67	0.167	0.063	0.249	0.335	0.09
LogReg	HCDR	binned	**0.729**	0.254	**0.068**	**0.342**	0.448	0.167
LogReg	HCDR	raw	0.728	**0.256**	0.068	0.34	**0.45**	**0.168**
LogReg	SGCC	binned	**0.635**	**0.447**	**0.2**	0.207	0.104	**0.132**
LogReg	SGCC	raw	0.622	0.437	0.204	**0.252**	**0.193**	0.121
LogReg	binned > raw	-	5	4	5	4	2	4
LogReg	Average	binned	**0.759**	**0.477**	**0.122**	**0.415**	**0.397**	**0.276**
LogReg	Average	raw	0.713	0.428	0.127	0.362	0.357	0.213
DecTree	CA	binned	0.781	0.785	0.176	0.487	**0.461**	0.338
DecTree	CA	raw	**0.824**	**0.849**	**0.153**	**0.621**	0.013	**0.452**
DecTree	DCCCT	binned	0.755	0.544	0.137	0.396	0.306	0.259
DecTree	DCCCT	raw	**0.766**	**0.553**	**0.136**	**0.399**	**0.35**	**0.263**
DecTree	GMSC	binned	0.809	**0.368**	0.057	0.474	0.58	0.305
DecTree	GMSC	raw	**0.811**	0.364	**0.056**	**0.482**	**0.581**	**0.314**
DecTree	HCDR	binned	0.715	0.256	0.068	0.323	0.419	0.153
DecTree	HCDR	raw	**0.721**	**0.269**	**0.068**	**0.332**	**0.431**	**0.157**
DecTree	SGCC	binned	**0.65**	0.509	**0.197**	**0.251**	0.131	**0.134**
DecTree	SGCC	raw	0.628	**0.509**	0.223	0.246	**0.167**	0.084
DecTree	binned > raw	-	1	1	1	1	1	1
DecTree	Average	binned	0.742	0.492	**0.127**	0.386	**0.379**	0.238
DecTree	Average	raw	**0.75**	**0.509**	0.127	**0.416**	0.308	**0.254**
RandForest	CA	binned	0.836	0.793	0.155	0.603	0.516	0.457
RandForest	CA	raw	**0.844**	**0.806**	**0.15**	**0.612**	**0.712**	**0.475**
RandForest	DCCCT	binned	0.767	0.532	0.137	0.406	0.324	0.27
RandForest	DCCCT	raw	**0.777**	**0.538**	**0.135**	**0.414**	**0.366**	**0.281**
RandForest	GMSC	binned	0.812	0.34	0.057	0.494	0.61	0.306
RandForest	GMSC	raw	**0.826**	**0.361**	**0.055**	**0.508**	**0.621**	**0.34**
RandForest	HCDR	binned	0.711	0.23	0.069	0.321	0.423	0.141
RandForest	HCDR	raw	**0.734**	**0.263**	**0.067**	**0.352**	**0.466**	**0.177**
RandForest	SGCC	binned	0.666	0.512	0.198	0.251	0.219	0.14
RandForest	SGCC	raw	**0.704**	**0.518**	**0.191**	**0.361**	**0.272**	**0.205**
RandForest	binned > raw	-	0	0	0	0	0	0
RandForest	Average	binned	0.759	0.482	0.123	0.415	0.418	0.263
RandForest	Average	raw	**0.777**	**0.497**	**0.12**	**0.449**	**0.487**	**0.296**
NN	CA	binned	**0.853**	0.806	**0.146**	**0.648**	**0.589**	**0.506**
NN	CA	raw	0.834	0.806	0.168	0.626	0.246	0.464
NN	DCCCT	binned	**0.768**	**0.535**	**0.137**	**0.409**	**0.358**	**0.277**
NN	DCCCT	raw	0.66	0.385	0.24	0.241	0.219	0.1
NN	GMSC	binned	**0.819**	**0.348**	**0.056**	**0.505**	**0.601**	**0.325**
NN	GMSC	raw	0.771	0.306	0.06	0.426	0.542	0.252
NN	HCDR	binned	**0.728**	**0.251**	**0.068**	**0.34**	**0.445**	**0.167**
NN	HCDR	raw	0.571	0.084	0.078	0.136	0.142	0.02
NN	SGCC	binned	**0.614**	**0.457**	**0.203**	**0.237**	0.049	**0.128**
NN	SGCC	raw	0.589	0.405	0.214	0.181	**0.178**	0.069
NN	binned > raw	-	5	4	5	5	4	5
NN	Average	binned	**0.757**	**0.48**	**0.122**	**0.428**	**0.408**	**0.281**
NN	Average	raw	0.685	0.397	0.152	0.322	0.265	0.181
BernoulliNB	CA	binned	**0.863**	**0.793**	**0.141**	**0.612**	**0.589**	**0.507**
BernoulliNB	CA	raw	0.802	0.78	0.227	0.564	0.217	0.402
BernoulliNB	DCCCT	binned	**0.736**	**0.504**	**0.17**	**0.373**	**0.217**	**0.218**
BernoulliNB	DCCCT	raw	0.67	0.396	0.417	0.263	0.148	0.119
BernoulliNB	GMSC	binned	**0.782**	**0.24**	0.07	**0.457**	**0.531**	**0.24**
BernoulliNB	GMSC	raw	0.691	0.219	**0.067**	0.271	0.353	0.127
BernoulliNB	HCDR	binned	**0.697**	**0.204**	**0.07**	**0.3**	**0.385**	**0.117**
BernoulliNB	HCDR	raw	0.605	0.134	0.073	0.159	0.211	0.036
BernoulliNB	SGCC	binned	0.656	0.492	**0.206**	0.255	0.246	0.114
BernoulliNB	SGCC	raw	**0.675**	**0.502**	0.217	**0.299**	**0.28**	**0.154**
BernoulliNB	binned > raw	-	4	4	4	4	4	4
BernoulliNB	Average	binned	**0.747**	**0.447**	**0.132**	**0.399**	**0.394**	**0.239**
BernoulliNB	Average	raw	0.689	0.406	0.2	0.311	0.242	0.168
All	binned > raw	-	15	13	15	14	11	14
All	Average	binned	**0.753**	**0.475**	**0.125**	**0.409**	**0.399**	**0.259**
All	Average	raw	0.723	0.447	0.145	0.372	0.332	0.222

*

Source: Author's computations

*The table presents the results of calculating the six evaluation metrics (columns (3)-(8)) on the test set comprising 20% observations of each data set. For the binned model, the discretized features were one hot encoded. Values in **bold** signal a better performance of the given model type for the given data set. Note that except for the Brier score, the higher the value of the metric, the better. The third row from the bottom for each method shows the number of data sets for which the binned model outperformed the raw model for a given evaluation metric. The last two rows for each estimation method show the average value for each metric for each model type across all data sets. The third row from the bottom of the table shows the total number of cases where the binned model outperformed the raw model across all methods and data sets. The very last two rows of the table present total averages across all methods and data sets.

Table B.15: Permutation tests - estimation with one hot encoding

Method	Data set	AUC	F2-score	Brier score	KS statistic	Partial GINI Index	H-measure
LogReg	CA	0.125	0.184	0.265	0.408	0.492	0.322
LogReg	GMSC	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
LogReg	DCCCT	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
LogReg	HCDR	0.286	0.48	0.476	0.451	0.772	0.69
LogReg	SGCC	0.236	0.192	0.084	0.957	0.896	0.201
LogReg	p <= 0.05	2	2	2	2	2	2
DecTree	CA	0.942	0.615	0.832	0.985	0.109	0.971
DecTree	GMSC	0.844	0.065	1.0	0.95	0.565	0.987
DecTree	DCCCT	0.987	**0.014**	0.87	0.847	0.997	0.915
DecTree	HCDR	0.998	0.881	0.889	0.964	0.998	0.969
DecTree	SGCC	0.362	0.123	0.053	0.516	0.645	0.274
DecTree	p <= 0.05	0	1	0	0	0	0
RandForest	CA	0.815	1.0	0.848	0.753	0.946	0.852
RandForest	GMSC	1.0	0.999	1.0	0.977	0.97	1.0
RandForest	DCCCT	0.999	0.793	1.0	0.901	0.999	1.0
RandForest	HCDR	1.0	1.0	1.0	1.0	1.0	1.0
RandForest	SGCC	0.918	0.338	0.943	0.982	0.73	0.979
RandForest	p <= 0.05	0	0	0	0	0	0
NN	CA	0.132	0.637	**0.01**	0.109	**0.037**	0.069
NN	GMSC	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
NN	DCCCT	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
NN	HCDR	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
NN	SGCC	0.082	0.254	**0.018**	**0.037**	0.609	**0.036**
NN	p <= 0.05	3	3	5	4	4	4
BernoulliNB	CA	**0.003**	0.248	**0.0**	0.058	**0.011**	**0.004**
BernoulliNB	GMSC	**0.0**	0.606	1.0	**0.0**	**0.0**	**0.0**
BernoulliNB	DCCCT	**0.0**	**0.0**	**0.0**	**0.0**	0.189	**0.0**
BernoulliNB	HCDR	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
BernoulliNB	SGCC	0.553	0.144	0.226	0.803	0.457	0.836
BernoulliNB	p <= 0.05	4	2	3	3	3	4

The values in the table represent the p-values of the null hypothesis that the model with raw variables performs better or equal to the binned model. As a result, p-values below 0.05 (in **bold**) signal the rejection of the null hypothesis in favor of the alternative that the binned model performs better at the 5% significance level. The last row for each estimation method shows the number of data sets for which the null is rejected for a given evaluation metric. The permutation tests were performed for 5000 repetitions.

Source: Author's computations

Appendix C

Software implementation

All data preprocessing steps and analyses in this thesis were performed using the Python programming language. The main two utilized packages are the "OptBinning" module containing the implementation of the optimal binning algorithm (Navas-Palencia 2023) and the "scikit-learn" package providing various functionalities for predictive modeling, including estimation methods, data partitioning, and evaluation metrics (Pedregosa *et al.* 2011).

The code necessary for the replication of the results of this thesis is publicly available in the following GitHub repository: `https://github.com/Matyas-Mattanelli/Master-Thesis`