REGRESSION-BASED CREDIT RISK MODELLING: A CASE STUDY ON THE GERMAN CREDIT DATASET

Lappeenranta–Lahti University of Technology LUT

Bachelor’s Programme in Business Analytics, Bachelor’s thesis

2025

Joonas Sainio

Examiners: M.Sc. Matias Heikkinen

Abstract

Lappeenranta–Lahti University of Technology LUT

LUT Business School

Bachelor's Programme in Business Analytics

Joonas Sainio

Regression-Based Credit Risk Modelling:
A Case Study on the German Credit Dataset

Bachelor's thesis

2025

26 pages, 13 figures, 5 tables and 8 appendices

Examiners: M.Sc. Matias Heikkinen

Keywords: Credit risk, Linear regression, Logistic regression, Tikhonov regularization

This thesis examines which applicant characteristics are most strongly associated with credit risk and how different regression models can be used to estimate the probability of being classified as a bad credit risk. The analysis uses the Statlog German Credit Data, which contains 1000 credit applicants described by their demographic, financial, and credit-related variables and a binary target variable that indicates good or bad credit risk. Ordinary Least Squares (OLS) regression is first used as a baseline model. The reduced OLS is then expanded by Tikhonov regularization to examine multicollinearity and the stability of the model. Finally, a logistic regression model is estimated and evaluated using a stratified train-test split, classification accuracy, confusion matrices, and a ROC curve.

The results show that credit-related variables and the financial characteristics of the applicants are the strongest predictors of credit risk. Longer loan durations, larger credit amounts, and higher installment rates are associated with a higher likelihood of being classified as a bad credit risk. Checking account status and savings categories also show meaningful associations with credit risk classification. Demographic characteristics, such as age, have a milder effect, with age showing a statistically significant negative association in the reduced logistic model, while foreign worker applicants have a strong association with higher estimated credit risk. The reduced OLS model explains about a quarter of the variation in the target variable, and the Tikhonov-regularized model's coefficients are essentially unchanged, indicating the model's stability. The reduced logistic regression model achieves an accuracy of approximately 0.74 at the 0.5 threshold and an area under the ROC curve of 0.78, suggesting that the model is capable of capturing key patterns in credit risk evaluation.

Tiivistelmä

Lappeenrannan–Lahden teknillinen yliopisto LUT

LUT-kauppakorkeakoulu

Kauppatieteiden koulutusohjelma

Joonas Sainio

Regressiopohjainen luottoriskimallinnus:
Case-Tutkimus German Credit -aineistosta

Kandidaatintyö

2025

26 sivua, 13 kuvaa, 5 taulukkoa ja 8 liitettä

Tarkastajat: DI Matias Heikkinen

Avainsanat: Luottoriski, Lineaarinen regressio, Logistinen regressio, Tikhonov regularisaatio

Tässä tutkielmassa tarkastellaan, mitkä lainahakijoiden ominaisuudet ovat vahvimmin yhteydessä luottoriskiin ja miten regressiomalleilla voidaan arvioida huonoksi luottoriskiksi luokittelun todennäköisyyttä. Empiirinen analyysi perustuu Statlog German Credit -aineistoon, joka sisältää 1000 luotonhakijaa; hakijoita kuvataan demografisilla, taloudellisilla ja luottohistoriaa kuvaavilla muuttujilla sekä kohdemuuttujalla, joka ilmaisee, luokitellaanko hakija hyväksi vai huonoksi luottoriskiksi. Pienimmän neliösumman menetelmä (PNS) toimii lähtötason mallina, jota täydennetään Tikhonov-regularisoinnilla vakauden arvioimiseksi, ja logistista regressiota arvioidaan ositetulla (stratifioidulla) 70/30 opetus-testaus jaolla, sekaannusmatriiseilla (confusion matrix) ja ROC-käyrällä.

Tulokset osoittavat, että lainaan ja hakijoiden taloudelliseen asemaan liittyvät muuttujat ovat vahvimpia luottoriskin selittäjiä: pidemmät laina-ajat, suuremmat luottomäärät ja korkeammat maksuerät sekä käyttötilin tila ja säästöluokat ovat yhteydessä suurempaan riskiin. Iän vaikutus on maltillisempi ja negatiivinen pelkistetyssä logistisessa mallissa, kun taas ulkomaalaisen työntekijän indikaattori on vahvasti yhteydessä korkeampaan arvioituun luottoriskiin. OLS-malli selittää noin neljäsosan kohdemuuttujan vaihtelusta, ja Tikhonov-regularisoidun mallin kertoimet pysyvät käytännössä muuttumattomina. Pelkistetyn logistisen mallin luokittelutarkkuus 0,5 määrittelytasolla on noin 0,74 ja ROC AUC noin 0,78.

Declarations

Turnitin

The originality of this thesis has been reviewed with the Turnitin similarity checking service.

AI Usage

The author of the thesis, Joonas Sainio, used the following AI-tools during the preparation of the thesis:

1. Grammarly
   a. Purpose of use: To check spelling, grammar and language issues.
   b. Explanation of the use of the tool: Grammarly was applied to text written by the author to identify grammatical errors and to suggest stylistic improvements. All suggestions were critically reviewed and only the changes by the author were implemented.
2. ChatGPT
   a. Purpose of use: To form ideas to improve the overall structure and clarity of the text and to find related research to support learning.
   b. Explanation of the use of the tool: ChatGPT was used in an interactive manner to discuss the structure of the text to refine it and to clarify text written by the author. The sources were used to learn methodology and to apply it to the study. All suggestions and sources were critically reviewed and only the changes by the author were implemented.

Responsibility

The author, Joonas Sainio, takes full responsibility for the content of this thesis and has reviewed and edited the content generated by the possible use of AI tools.

Table of contents

1 Introduction

Credit risk assessment remains a central challenge in modern finance, as lenders must evaluate the likelihood that borrowers will meet their obligations. As financial institutions provide loans and other forms of credit, they face uncertainty about whether borrowers will repay as agreed. In consumer and retail lending, this uncertainty is managed using credit and behavioural scoring systems, which predict the probability that a borrower will default on their obligations (Thomas, 2000). Accurate estimates of default probabilities affect both the profitability of the lender and the stability of the financial system. If credit risk is underestimated and losses are realised, these costs are passed on to the customers through higher interest margins, much like how car insurance premiums increase when a certain type of vehicle is frequently involved in accidents. This effect was clearly visible in the U.S. mortgage market before the 2007-2008 financial crisis, where supply-driven rise in the amount of risky borrowers was followed by high default rates and large investor losses (Mian & Sufi, 2009). Reliable credit risk assessment is therefore important for both customers and financial institutions.

Over time, the process of evaluating creditworthiness has evolved from subjective judgments to quantitative analysis. Statistical models now play a central role in credit scoring, allowing lenders to base their decisions on measurable financial and personal characteristics rather than intuition or experience. These models help to make lending decisions more consistent and transparent while improving the management of credit risk (Thomas, 2000). Early approaches relied on methods such as linear discriminant analysis and logistic regression. These techniques remain widely used because they are relatively simple and interpretable (Hand & Henley, 1997). Logistic regression, in particular, has become a standard tool in credit scoring because it directly estimates the probability of default and provides interpretable results about the influence of borrower characteristics (Peng, Lee & Ingersoll, 2002). Large benchmarking studies comparing different modern classifiers on multiple credit datasets often find that more flexible methods can improve predictive accuracy, but logistic regression often remains a strong baseline due to its transparency (Lessmann et al., 2015).

The importance of reliable credit risk assessment became especially clear during the global financial crisis of 2008, when poor risk management contributed to widespread instability (Mian & Sufi, 2009). Since then, banks and financial regulators have placed greater emphasis on transparent and data-based risk evaluation. The General Data Protection Regulation (GDPR) introduced by the European Union has increased the importance of using transparent and explainable models in financial decision-making (Goodman & Flaxman, 2017). This requirement for interpretability reinforces the value of traditional statistical methods, which make it possible to explain the relationships between variables and model outcomes clearly.

Although modern research has introduced new tools and datasets, the main objective has remained the same: to identify which factors best explain differences in creditworthiness. In credit scoring applications, scorecards are formed from information collected on the credit application form and from credit bureau records. Thomas (2000) describes how categorical values such as residential status and quantitative variables such as age and income are used as predictors, and that these variables are assigned weights so that the resulting scorecard separates "good" and "bad" borrowers as effectively as possible. Empirical work has examined how demographic characteristics, such as age and marital status, and loan terms, such as loan size and interest rate, affect the borrower's ability to repay on time (Özdemir, 2008). Although Özdemir (2008) finds that financial variables are more important predictors than demographic characteristics in her sample, the study, together with broader credit scoring literature, highlights the value of considering both demographic and financial information when analyzing credit risk.

This thesis addresses this gap by focusing on interpretable regression-based credit risk models. The analysis relies on ordinary least squares (OLS) regression, which is extended to Tikhonov regularization, and logistic regression to examine how applicant characteristics are linked to credit risk classification. In addition, Bayesian probability principles are used to interpret the model parameters in probabilistic terms and to evaluate the uncertainty surrounding the estimates. Together, the chosen methods provide a solid basis for quantifying the impact of different predictors while maintaining transparency in how conclusions are drawn.

1.1 Research Aim and Questions

The purpose of this thesis is to identify and analyze the most significant indicators of credit risk using interpretable statistical models, with an emphasis on statistical interpretation rather than complex computational modelling.

The main research question is:

RQ1: *Which applicant characteristics are most strongly associated with being classified as a bad credit risk?*

To support the main question, the thesis addresses the following sub-questions:

RSQ1: *How are key demographic characteristics, such as age, employment status, and housing, related to credit risk?*

RSQ2: *How can the chosen regression models and their statistical diagnostics be used to evaluate the performance and reliability of risk prediction?*

These questions are investigated based on the Statlog German credit dataset Hofmann (1994), so the findings should be interpreted as evidence from this particular case instead of universal conclusions of all credit markets.

1.2 Structure of the Thesis

This thesis continues as follows: Chapter 2, *Methodology*, presents the theoretical framework and research methods used in the analysis. Methodology introduces OLS regression, Tikhonov regularization, logistic regression, and provides outlines for the Bayesian perspective, which is used to interpret model parameters and uncertainty. Chapter 3, *Experimental Design and Data Exploration*, describes the empirical setup of the study, including the dataset, the variables and how the preprocessing steps applied to them. Chapter 3 also documents the modelling choices before the regression methods are applied. Chapter 4, *Experimental Results*, presents the empirical results of the applied methods. This includes regression model comparisons and a discussion of what the results imply. The final chapter, Chapter 5, *Conclusions*, summarizes the main findings, reflects on the limitations of the study and suggests ideas for future research.

In summary, this thesis examines credit risk using interpretable regression models that combine financial and demographic credit applicant characteristics. By applying the chosen methods to the dataset and comparing the results across different methods, the study aims to contribute to the discussion surrounding credit risk modeling by exploring the factors that affect the likelihood of being classified as a bad credit risk in an interpretable way.

2 Methodology

The Methodology chapter presents the framework used to complete the analysis of this research. The research applies a quantitative approach using regression-based models to explore the relationship between characteristics and credit risk. The selected methods enable both linear and probabilistic modeling of credit risk, and allow robustness of the results to be assessed.

The study employs three analytical methods to form a thorough analysis: Ordinary Least Squares (OLS) regression, Tikhonov regularization, and logistic regression. In addition to the three analytical methods, Bayesian probability principles are used to interpret the parameters in probabilistic terms to assess the model's uncertainty.

2.1 Bayesian Probability

Bayesian inference provides a framework for reasoning about uncertainty. In this method, unknown quantities, such as regression coefficients, are considered random variables. Prior beliefs about these quantities are represented by a prior distribution $P(\beta)$, and these beliefs are updated after observing the data ($\mathbf{y}$).

Bayes' theorem combines the prior distribution $P(\beta)$ with the likelihood $P(\mathbf{y} \mid \beta)$, which quantifies how likely the observed outcomes are under different model parameters. The result is the posterior distribution $P(\beta \mid \mathbf{y})$, which represents updated beliefs about β after observing the data.

$$P(\beta \mid \mathbf{y}) = \frac{P(\mathbf{y} \mid \beta)P(\beta)}{P(\mathbf{y})}, \tag{1}$$

where the denominator $P(\mathbf{y})$ does not depend on β and therefore acts as a normalizing constant that rescales the numerator so that the posterior integrates to one. Therefore, $P(\beta \mid \mathbf{y})$ is a probability distribution over β conditional on the observed data.

Bayes' theorem is often presented in proportional form:

$$P(\beta \mid \mathbf{y}) \propto P(\mathbf{y} \mid \beta)P(\beta). \tag{2}$$

This is justified as the constant $P(\mathbf{y})$ is not dependent on the parameters β, and only rescales the posterior density without changing its shape. Thus, it does not interfere with the location of the maximum. As a result, the maximum a posteriori (MAP) estimate is identical in the hyperplane of β whether computed from the normalized posterior or from its proportional form. In the one-dimensional case $\beta \in \mathbb{R}$, this means the MAP estimate peaks at the same position on the β-axis under both forms.

2.2 Regression and Regularization

Before introducing the Ordinary Least Squares (OLS) model, it is useful to view regression as a function between explanatory variables and outcome predictions. Regression analysis examines how a set of explanatory variables can be used to predict an outcome. In this context, regression can be seen as a function that maps predictor values to estimated responses. Formally, the estimates generated by the regression model define such a function

$$f : \mathscr{X} \to \mathscr{Y}, \tag{3}$$

where $\mathscr{X}$ represents the space of predictor vectors and $\mathscr{Y}$ represents the space of possible outcomes. For observation i, the model maps the predictor vector $\mathbf{x}_i \in \mathscr{X}$ to an the estimate $\hat{y}_i = f(\mathbf{x}_i)$. Collecting all observations into vectors, the vector of observed outcomes $\mathbf{y}$ can be decomposed into an estimate $\hat{\mathbf{y}}$ and residuals ε,

$$\mathbf{y} = \hat{\mathbf{y}} + \varepsilon. \tag{4}$$

The goal of regression analysis is to determine the function f and estimate optimal parameters that best capture the relationship between inputs and outputs.

As a baseline model, OLS regression is applied to examine the linear relationship between the predictors and the target variable. Linear regression can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \tag{5}$$

Here y_i denotes the dependent variable for i:th observation, β_0 is the intercept and β_j are the coefficients associated with the explanatory variables x_{ji}. ε_i is the error term that represents the unexplained part of y_i.

Let $\mathbf{y} = [y_1, y_2, \ldots, y_n]^\top \in \mathbb{R}^n$ denote the vector of observed outcomes, where y_i is the dependent variable for observation i. For each observation i, define the expanded predictor vector

$\mathbf{x}_i = [1, x_{i1}, x_{i2}, \ldots, x_{ik}]^\top \in \mathbb{R}^{k+1}$, where the leading 1 corresponds to the intercept term and x_{ij} denotes the value of predictor j for observation i. Adding the predictors into the design matrix

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix} \in \mathbb{R}^{n\times(k+1)}, \tag{6}$$

and denoting the vector of regression coefficients by $\beta = [\beta_0, \beta_1, \beta_2, \ldots, \beta_k]^\top \in \mathbb{R}^{k+1}$. The estimate is given by

$$\hat{\mathbf{y}} = \mathbf{X}\beta. \tag{7}$$

The Ordinary Least Squares (OLS) estimator is obtained by choosing the parameter values that minimize the sum of squared residuals. This is demonstrated with a linear model in Fig. 1, where the squared errors are plotted between the data points and the fitted line.

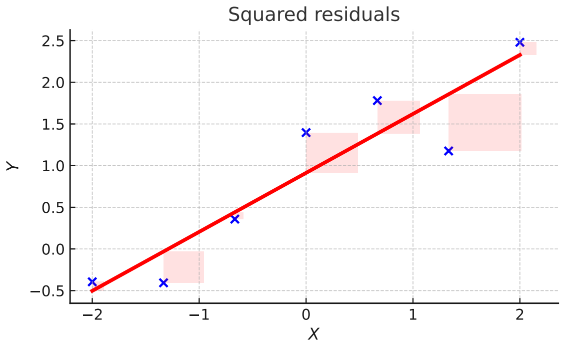


Figure 1: Linear OLS example

The estimator solves the optimization problem

$$\min_{\beta} \sum_{i=1}^{n} (\varepsilon_i(\beta))^2 = \min_{\beta} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \beta)^2 = \min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2. \tag{8}$$

Let

$$L(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2 = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta). \tag{9}$$

Expanding $L(\beta)$ yields

$$L(\beta) = \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{X}\beta - \beta^\top \mathbf{X}^\top \mathbf{y} + \beta^\top \mathbf{X}^\top \mathbf{X}\beta. \tag{10}$$

Because the minima of a differentiable function occur at points where the gradient is zero, the OLS estimator is found by differentiating $L(\beta)$ with respect to β and setting the gradient equal to zero yields

$$-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\beta = 0 \tag{11}$$

which simplifies to

$$\mathbf{X}^\top \mathbf{X}\beta = \mathbf{X}^\top \mathbf{y}, \tag{12}$$

and solving for β gives the closed-form OLS estimator

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \tag{13}$$

Under the standard assumptions of linear regression, including independence, homoscedasticity, normality of residuals, and zero mean error, the OLS estimator produces unbiased and consistent coefficient estimates. These coefficients indicate both the strength and the direction of the relationship between each explanatory variable and the dependent variable.

While Ordinary Least Squares (OLS) regression produces unbiased and consistent parameter estimates, its performance may become worse in the presence of multicollinearity among the explanatory variables. When there are strong correlations between predictors, small changes in the data can result in large variations in the estimated coefficients, reducing the stability and interpretability of the model.

To mitigate this issue, Tikhonov regularization introduces a regularization term (λ) to the OLS estimation problem in Eq. (8), which penalizes large coefficient values. An important part of Tikhonov regularization is to choose a regularization term λ. Larger values of λ produce stronger shrinkage in the coefficients, while smaller values keep the regularized estimates closer to the OLS estimates.

Under perfect multicollinearity, the mapping from coefficients to estimates ($\hat{\beta} \to \hat{\mathbf{y}} = X\hat{\beta}$) is not injective, meaning that if two predictors are perfectly collinear, different coefficient vectors can produce the same predictions. At the same time the mapping is surjective onto the column space of X (set of all estimated value vectors that the model can produce), which means that every prediction vector $\hat{\mathbf{y}}$ that the model can represent can be reached by at least one β. However the mapping generally is not surjective onto $\mathbb{R}^n$, so not every possible outcome vector can be produced by $X\beta$. This motivates regularization. The Tikhonov (ridge) objective adds the penalty term $\lambda\|\beta\|^2$, which selects a unique solution by preferring the coefficient vector with the smallest $\|\beta\|$ among those that fit the data similarly well. In other words, when many β lead to the same estimated values, regularization chooses the smallest coefficients, improving stability and reducing sensitivity to small changes in the data. The concepts of injectivity, surjectivity, and bijectivity are illustrated in Fig. 2.

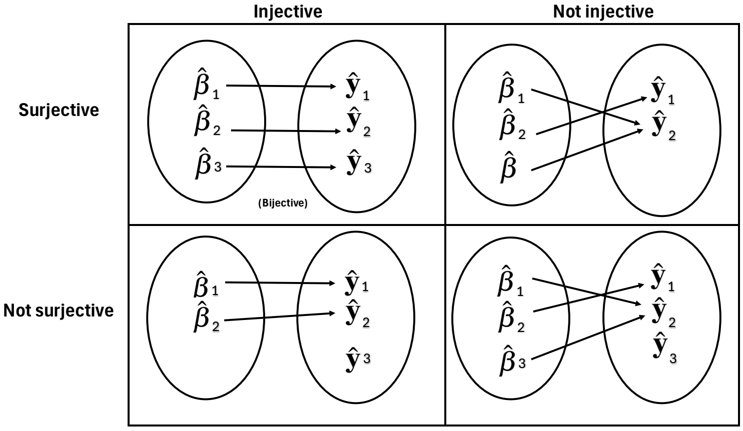


Figure 2: Visualization matrix of injectivity, surjectivity, and bijectivity

This modification reduces variance and enhances the robustness and reliability of the model's estimates by shrinking coefficient magnitudes. The Tikhonov regularization is represented as

$$\min_{\beta}\left(\|\mathbf{y}-\mathbf{X}\beta\|^2+\lambda\|\beta\|^2\right), \tag{14}$$

Differentiating the Tikhonov regularization equation with respect to β gives the gradient

$$\nabla_{\beta}L_R(\beta) = -2\mathbf{X}^{\top}(\mathbf{y}-\mathbf{X}\beta)+2\lambda\beta. \tag{15}$$

Setting the gradient to zero gives the Tikhonov regularization normal equation

$$(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})\beta=\mathbf{X}^\top\mathbf{y}, \tag{16}$$

To demonstrate the effect of the regularization in matrix form, assume that

$$\mathbf{X}^\top\mathbf{X}=\begin{bmatrix}1&2\\2&4\end{bmatrix},$$

which is singular because the second column is twice the first one. This can be seen directly from the determinant $\det(\mathbf{X}^\top\mathbf{X})=1\times4-2\times2$, which equals 0. This leads to a situation where the inverse $(\mathbf{X}^\top\mathbf{X})^{-1}$ is not defined. Adding the Tikhonov penalty then modifies the non-regularized matrix to

$$\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I}=\begin{bmatrix}1+\lambda&2\\2&4+\lambda\end{bmatrix}.$$

This leads to $\det(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})=5\lambda+\lambda^2$, which is positive for any $\lambda>0$. This means that the matrix is now invertible when $\lambda>0$.

Solving the optimization gives the closed-form Tikhonov estimator

$$\hat{\beta}_R=(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}, \tag{17}$$

which minimizes Eq. (14).

Tikhonov regularization can also be interpreted from a Bayesian perspective, under Gaussian likelihood and prior assumptions. Assuming a Gaussian likelihood for the data, we start from the standard regression error assumption $\varepsilon\sim\mathcal{N}(0,\sigma^2\mathbf{I})$ in the model $\mathbf{y}=\mathbf{X}\beta+\varepsilon$. This implies that the conditional distribution of the outcome vector given the coefficients is

$$\mathbf{y}\sim\mathcal{N}(\mathbf{X}\beta,\sigma^2\mathbf{I}). \tag{18}$$

If a normally distributed prior is assumed for the coefficients,

$$\beta \sim \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{I}), \tag{19}$$

The likelihood is proportional to $P(\mathbf{y} \mid \beta) \propto \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\beta\|^2\right)$ under the normal-error assumption. For Tikhonov regularization, the prior is a zero-mean normal prior, which is proportional to $P(\beta) \propto \exp\left(-\frac{1}{2\tau^2}\|\beta\|^2\right)$. By Bayes' theorem, the posterior is proportional to likelihood times prior, $P(\beta \mid \mathbf{y}) \propto P(\mathbf{y} \mid \beta) \times P(\beta)$. Combining the likelihood and prior using Bayes' theorem Eq. (2) leads to a posterior density form

$$P(\beta \mid \mathbf{y}) \propto \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\beta\|^2 - \frac{1}{2\tau^2}\|\beta\|^2\right). \tag{20}$$

Taking the negative logarithm of the posterior distribution converts the maximization of the posterior into a minimization problem. Because the posterior is written up to proportionality, terms that do not depend on β can be left out without changing the value of β that minimizes the objective.

$$-\log P(\beta \mid \mathbf{y}) = \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\beta\|^2 + \frac{1}{2\tau^2}\|\beta\|^2. \tag{21}$$

Minimizing this negative log-posterior with respect to β is therefore equivalent to solving

$$\min_{\beta}\left(\|\mathbf{y} - \mathbf{X}\beta\|^2 + \frac{\sigma^2}{\tau^2}\|\beta\|^2\right). \tag{22}$$

This problem has the same structure as the Tikhonov regularization objective shown in Eq. (14), where $\lambda = \frac{\sigma^2}{\tau^2}$. Consequently, Tikhonov regularization can be viewed as Bayesian ridge regression when assuming Gaussian likelihood and prior distributions.

2.3 Logistic Regression

When the dependent variable is binary, linear regression is not ideal because it can generate estimates outside the range $[0,1]$ and depends on assumptions such as normally distributed errors and homoscedasticity, which are not valid for binary data. Logistic regression solves these problems by modeling the log-odds of the event as a linear function of the predictors and then applying a logistic transformation, which maps the linear predictor $\mathbf{x}_i^\top \beta \in \mathbb{R}$ to a probability between 0 and 1. Let

$$y_i \in \{0, 1\} \tag{23}$$

denote the binary outcome of observations y_i, where $y_i = 1$ indicates that the event happens, and $y_i = 0$ indicates that it does not. Logistic regression models the probability of the event rather than the class label itself.

Let

$$\mathbf{x}_i = [1, x_{i1}, x_{i2}, \ldots, x_{ip}]^\top \tag{24}$$

denote the vector of p predictor values for observation i, and

$$\beta = [\beta_0, \beta_1, \ldots, \beta_p]^\top \tag{25}$$

denote the vector of model parameters, where β_0 is the intercept term. Each coefficient β_j describes how the corresponding predictor contributes to the log-odds of the event.

Logistic regression represents the logarithmic odds (logit) of the event $y_i = 1$ as a linear function of the predictors

$$\log\left(\frac{P(y_i = 1 \mid \beta)}{1 - P(y_i = 1 \mid \beta)}\right) = \mathbf{x}_i^\top \beta, \tag{26}$$

where the left-hand side of the equation is the log-odds of the event happening, and the right-hand side represents these log-odds as a linear combination of the predictors.

Solving for the probability $P(y_i = 1 \mid \beta)$ yields the logistic function

$$P(y_i = 1 \mid \beta) = \frac{1}{1 + \exp(-\mathbf{x}_i^\top \beta)}. \tag{27}$$

Because the negative log-likelihood does not yield a closed-form solution, the coefficients must be computed by numerical optimization, such as the Newton-Raphson algorithm. Each coefficient measures how a one-unit change in a predictor affects the log-odds of the event, holding other variables constant. The logistic function ensures that the predicted probabilities remain within the interval $[0,1]$, which allows the model to capture nonlinear relationships between the predictors and the probability of the event occurring.

Bayesian thinking can also be applied to logistic regression by treating the coefficient vector β as an unknown quantity and combining the likelihood with a prior distribution. Because the outcome is binary, each observation satisfies $y_i \in \{0,1\}$ and follows a Bernoulli distribution (equivalently a binomial model with one trial),

$$y_i \sim \text{Bernoulli}(p_i), \tag{28}$$

where $p_i = P(y_i = 1 \mid \beta)$, which means that each i:th observation has its own probability p_i that can vary with observations through predictors $\mathbf{x}_i$. By this property, logistic regression is estimated by maximizing the joint Bernoulli likelihood under the logit link. Assuming conditional independence across observations, the likelihood function for the full sample $\mathbf{y} = [y_1, y_2, \ldots, y_n]^\top$ is the product of Bernoulli probabilities,

$$L(\beta) = P(\mathbf{y} \mid \beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}. \tag{29}$$

We denote the log-likelihood by $l(\beta) = \log L(\beta)$, which gives

$$l(\beta) = \sum_{i=1}^{n} \left[y_i \log(p_i) + (1-y_i)\log(1-p_i) \right]. \tag{30}$$

The MLE is obtained by choosing β that minimizes the negative log-likelihood $(-l(\beta))$.

3 Experimental Design and Data Exploration

The purpose of this chapter is to describe the setup used to estimate and compare the chosen regression models and methods of the research. The chapter introduces the chosen dataset, summarizes the properties of the variables, outlines the preprocessing steps, and briefly presents how the statistical methods described in the previous chapter are applied in the analysis.

3.1 Data and Variables

The analysis is based on the German Credit Data dataset (Hofmann, 1994), which consists of 1000 credit applicants and 20 predictor variables describing their demographic, financial, and credit-related characteristics. The target variable *Risk* classifies applicants as either good- or bad credit risks.

The distribution of the outcome variable is moderately imbalanced. 700 applicants are labeled as good credit risks and 300 as bad credit risks. A model that would predict every applicant as a good credit risk would achieve an accuracy of 70%, so accuracy alone can be misleading. Therefore, the results are evaluated using confusion matrices.

Tab. 1 summarizes the variables, their types, and basic descriptive statistics.

Table 1: Data description table

Variable	**Type**	**Min**	**Max**	**Mean**	**Median**	**Missing values**
Status	Nominal	1	4	2.6	2.0	0
Duration	Numeric	4	72	20.9	18.0	0
Credit history	Ordinal	0	4	2.5	2.0	0
Purpose	Nominal	0	10	2.8	2.0	0
Credit amount	Numeric	250	18424	3271.3	2319.5	0
Savings	Nominal	1	5	2.1	1.0	0
Employment duration	Ordinal	1	5	3.4	3.0	0
Installment rate	Ordinal	1	4	3.0	3.0	0
Personal status	Nominal	1	4	2.7	3.0	0
Other debtors	Nominal	1	3	1.1	1.0	0
Present residence	Ordinal	1	4	2.8	3.0	0
Property	Nominal	1	4	2.4	2.0	0
Age	Numeric	19	75	35.5	33.0	0
Other installment plans	Nominal	1	3	2.7	3.0	0
Housing	Nominal	1	3	1.9	2.0	0
Number of credits	Ordinal	1	4	1.4	1.0	0
Job	Nominal	1	4	2.9	3.0	0
People liable	Binary	1	2	1.8	2.0	0
Telephone	Binary	1	2	1.4	1.0	0
Foreign worker	Binary	1	2	2.0	2.0	0
Risk	Binary target (0 = good, 1 = bad)	0	1	0.3	0.0	0

The variables in the dataset can be divided into continuous numeric, ordinal, nominal, and binary types. Continuous numeric variables include *Duration*, *Credit amount*, and *Age*. Several variables are measured on an ordered scale (e.g., *Installment rate*, *Present residence*, and *Number of credits*). Some variables, like *Credit history*, *Savings*, and *Employment duration*, are conceptually ordinal in the original coding but consist of discrete categories that do not reflect evenly spaced numerical intervals and are therefore treated as nominal. Nominal categorical variables with more than two categories include *Status*, *Purpose*, *Personal status*, *Other debtors*, *Other installment plans*, *Housing*, *Property*, and *Job*. Finally, the variables *Telephone*, *People liable*, and *Foreign worker* are binary variables indicating the presence or absence of a specific characteristic. For more detailed variable summaries, see Tab. 1.

3.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to gain an initial understanding of the dataset and to identify patterns among the predictor variables before applying regression models. The purpose of the EDA is to summarize distributions, check for imbalances and outliers, and provide context for the modelling choices rather than to draw definitive claims (Morgenthaler, 2009).

The distributional plots for the numeric predictors *Duration*, *Age*, and *Credit amount* are presented in App. C, D, and E. Because the risk groups (good and bad) are normalized separately, the histograms compare the distributional shape rather than counts, making the visual comparison less sensitive to the risk class imbalance. The bad credit risk group is more concentrated at higher credit amounts, while age differences are smaller and less clear to interpret from the histogram.

Discrete predictors (ordinal, nominal, and binary) were analyzed with countplots in App. F and G, which enabled comparison of risk class frequencies between groups within each predictor. For example, the distribution of *Credit history* differs across its categories, with most observations concentrated in a few levels and some categories having significantly fewer observations. Similarly, *Foreign worker* is highly unbalanced with the majority of applicants falling into a single category.

A simple correlation matrix of numeric predictors (*Duration, Credit amount*, and *Age*) is shown in App. H. It indicates a moderate positive correlation between *Duration* and *Credit amount* (0.62), while *Age* is weakly correlated with both.

All of the exploratory plots are provided for reference in the Appendices. The primary purpose of the EDA within the thesis is descriptive and serves as context for the regression models.

3.3 Data Preprocessing

All nominal categorical variables with more than two categories were converted into dummy variables using one-hot encoding, where each category level is represented by a binary indicator (Poslavskaya & Korolev, 2023). For each of these variables, one category was set as the reference level. These variables include *Status*, *Purpose*, *Savings*, *Personal status*, *Other debtors*, *Other installment plans*, *Housing*, *Property* and *Job*. For *Purpose*, categories other than *New car*, *Used car*, and *Education* were grouped into an *Other purpose* class for modelling. The ordinal predictors *Credit history* and *Employment duration* were kept in ordinal form for modelling. The binary variables, *Telephone*, *People liable*, and *Foreign worker*, were recoded from a 1-2 format to a 0-1 format for better compatibility with regression models.

The rest of the variables are categorized as either continuous numeric variables or truly ordinal variables and were kept in their original form. These variables include *Installment rate*, *Present residence*, *Number of credits*, *Duration*, *Credit amount*, and *Age*.

All variables in the original dataset were retained during the preprocessing stage, meaning that no predictors were excluded before model estimation. All excluded variables were selected based on being insignificant in regression modelling, and the exclusion is described in the results section.

3.4 Model Evaluation

The OLS model is first estimated using the entire dataset, and evaluated with standard regression diagnostics such as R^2 and adjusted R^2 measures and the significance of individual coefficients. In addition to point estimates, 95% confidence intervals are reported for the OLS coefficients to measure uncertainty in the estimates. Variables that show no clear significance and lack theoretical justification are removed to obtain a reduced OLS model, which is evaluated using the same diagnostics.

The reduced OLS model is then expanded to a Tikhonov regularization model using the same predictors as the reduced OLS model. Its coefficients and overall fit are compared with the non-regularized model to assess the effect of regularization and the stability of the estimated coefficients in the presence of multicollinearity. These diagnostics are used to evaluate the explanatory power and robustness of the methods.

In the empirical analysis, the Bayesian interpretation of Tikhonov regularization is used to further examine the robustness of the model. Following the Bayesian interpretation, the regression coefficients (β) are viewed as having a zero-mean normal prior with variance (τ^2), and the penalty term (λ) represents the ratio between the error variance (σ^2) and the prior variance (τ^2), where $\lambda = \frac{\sigma^2}{\tau^2}$. Because the prior information about the effect of the individual predictors to the target variable is not strong, the penalty effect is tested using small λ values, which correspond to weak prior knowledge. The purpose of Tikhonov regularization in this study is therefore not to force strong coefficient shrinkage, but to examine if weak regularization affects the coefficients, which would indicate instability due to multicollinearity.

The logistic regression model is first estimated on the entire dataset to analyze the effects of all preprocessed predictors and to obtain general fit measures. Variables that are statistically insignificant and lack substantive justification are then removed, resulting in a simplified model. This reduced model is then re-estimated using a 70/30 stratified train–test split: the model is trained on the training data, and its out-of-sample classification performance is assessed on the test data through confusion matrices and overall accuracy. In addition, a ROC curve and the area under the curve (AUC) are provided for the reduced model as indicators of its ability to distinguish between good and bad credit risks. These evaluation methods are used to get a summary of how effectively the logistic regression model predicts credit risk and to see where the classification performance is limited.

In addition, the Bayesian interpretation is also used when evaluating the logistic regression model. In the logistic model, the coefficient vector β can be assigned a zero-mean Gaussian prior, $\beta \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I})$, which produces a Maximum A Posteriori (MAP) estimate. Because prior information on effect sizes is weak in this research, only mild regularization is considered. The purpose of this is to test whether the logistic regression coefficients and predictions are sensitive to weak shrinkage, which would indicate model instability.

In practice, the MAP estimate is obtained by minimizing the negative log-likelihood derived from the log-likelihood in Eq. (30), with an added penalty term on β. MAP estimation is then implemented by minimizing the logistic negative log-likelihood with an added penalty term on β, which corresponds to the zero-mean Gaussian prior. The baseline logistic regression estimates are compared to weakly regularized MAP estimates by examining coefficient stability and by examining predictive stability on the test set (confusion matrix and ROC AUC).

4 Experimental Results

In this chapter, the empirical findings are presented and interpreted. The analysis applies linear regression and logistic regression to examine which applicant characteristics are associated with credit risk classifications. OLS is first used to establish an interpretable baseline to describe how applicant characteristics relate to risk classifications. Insignificant variables are dropped from the initial full model, and the reduced model's estimates are then examined with Tikhonov regularization to identify multicollinearity and instability. Tikhonov regularization can be interpreted from a Bayesian viewpoint to test how sensitive the estimates are to weak shrinkage.

Logistic regression is used to model the probability of bad credit and its performance is evaluated out-of-sample using confusion matrices and ROC AUC. The reduced logistic model is also re-estimated under Bayesian interpretation, which yields a Maximum A Posteriori (MAP) estimate that provides a robustness check against mild regularization. The results from both linear and logistic models provide stable results.

4.1 Linear Regression

The initial OLS model included all predictor variables. Variables with clearly insignificant coefficients, p-values, and without theoretical justification were sequentially excluded in the regression progress. Categorical variables were evaluated as groups, and if all dummy-encoded levels of the variable group were insignificant, the variable was completely removed from the model. This resulted in a reduced OLS model with more stable estimates.

The dependent variable *Risk* is coded as 1 for bad credit risk, and 0 for good credit risk applicants. The model is estimated by using OLS as defined in Eq. (8) in the Methodology chapter. Because OLS is linear, the model's predicted values are not limited to the $[0,1]$ interval and can take any real value. With the coding used in this study, larger predicted values indicate higher estimated bad credit risk, while lower predicted values indicate higher estimated good credit likelihood. Each estimated coefficient can be interpreted as the one-unit change in the predictors, holding other variables constant.

Before model reduction, the full specification included all continuous, binary, and dummy-encoded categorical variables. The initial model presented several insignificant predictors and signs of multicollinearity in the large condition number ($= 5.68 \times 10^4$). Variables that were clearly insignificant and lacked theoretical justification were removed. Categorical predictor variables were evaluated as groups where if the entire group consisted of only insignificant levels, the entire variable was removed, otherwise the variable was retained in the model.

Predictors *Present residence*, *Number of credits*, *People liable* and *Telephone* as well as dummy-encoded categorical groups *Personal status*, *Other installment plans*, *Housing*, *Property*, *Job* and *Savings* were removed. The reduced model therefore retains only variables that demonstrate statistical relevance or clear theoretical importance in credit-risk modeling.

Even after exclusions, the reduced OLS model still shows limited explanatory power, with an R^2 of 0.24 (Adj. $R^2 = 0.22$) and a lower but still elevated condition number ($= 3.78 \times 10^4$). This outcome is expected because of the binary dependent variable and the linear structure of OLS. Since OLS presumes linearity and constant variance (homoskedasticity), it cannot effectively model the nonlinear link between applicant traits and credit risk. Consequently, residuals and the standard errors of individual coefficients remain substantial, leading to a decrease in statistical significance for some predictors. These limitations highlight why OLS is used primarily as a descriptive tool and a benchmark in the context of this study.

Despite its limitations, the reduced OLS model provides valuable insight. Several predictors remained consistent and statistically significant, especially those related to applicants' financial capacity and credit history, such as *Duration*, *Credit amount*, *Installment rate*, and various levels of *Status* and *Purpose* as key risk indicators. These effects correspond to theoretical expectations, indicating that longer durations and higher credit amounts increase the likelihood of credit risk. These findings confirm that the reduced model serves as an effective descriptive benchmark.

Overall, the OLS analysis serves two main roles, acting as an interpretable baseline model that explains relationships among predictors and as a tool for detecting multicollinearity, helping identify irrelevant predictors and encouraging the use of regularized and probabilistic regression models. OLS also lays a solid foundation for Tikhonov regularization and logistic regression analysis.

Tikhonov regularization was applied to determine if the coefficient values from the reduced OLS model were sensitive to multicollinearity or instability in the design matrix. As noted earlier, the OLS model exhibited a high condition number, indicating multicollinearity in the initial full OLS specification. After the model was reduced, the coefficient set was re-estimated using Tikhonov regularization as defined in Eq. (14).

To examine the sensitivity of the coefficients to regularization, two penalty parameters were tested $\lambda = 0.01$ and $\lambda = 0.001$. Across both values the resulting coefficient estimates remained nearly identical to those of the OLS model. No coefficients changed sign, and the importance of all predictors remained unchanged. This indicates that even though the predictor set has multicollinearity, it does not significantly distort the OLS estimates, and the OLS model is stable. The OLS coefficients with 95% confidence intervals are presented in Tab. 4.

Fig. 3 presents the coefficient estimates from the reduced OLS model with their 95% confidence intervals. Each point represents the estimated effect of a predictor on the target variable, and the vertical lines show how uncertain the estimate is.

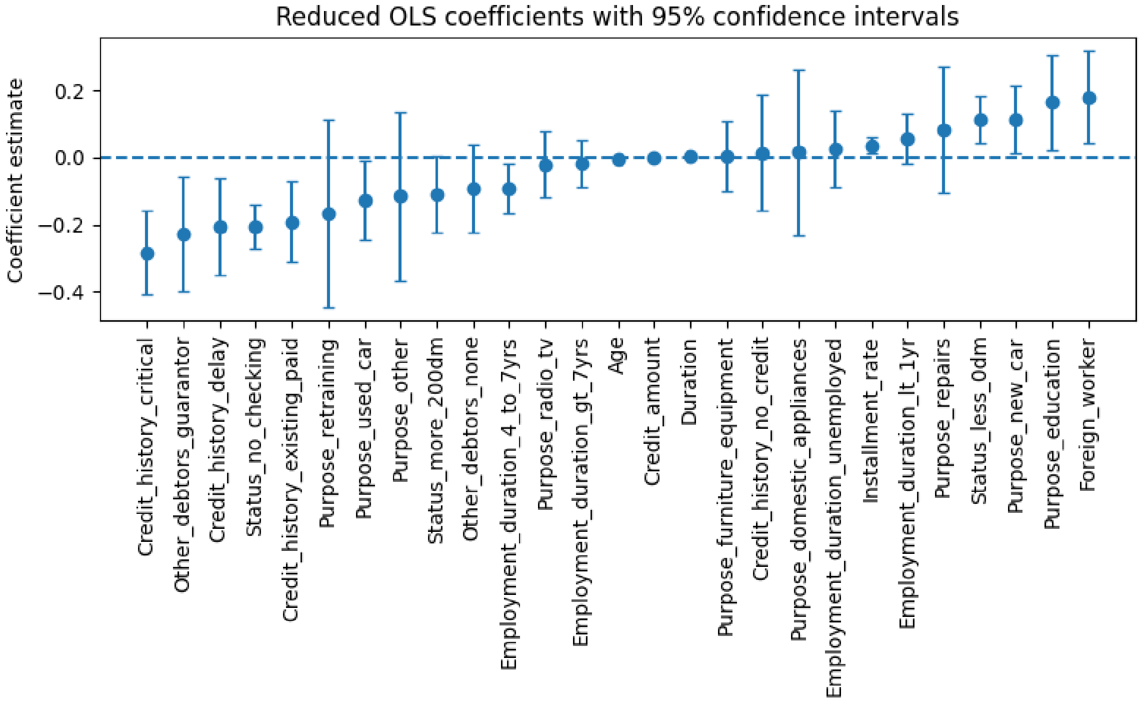


Figure 3: Reduced OLS model Coefficient visualization

From the Bayesian perspective, the small penalty values that were used in the Tikhonov models correspond to weakly informative priors. The fact that the regularized estimates are almost identical to the reduced OLS model's estimates for all predictors shows that despite the presence of multicollinearity, mild regularization does not distort the estimated coefficients. This suggests that the posterior mode is mostly determined by likelihood rather than by substantial prior assumptions. In Bayesian terms, the linear model is therefore not highly sensitive to prior variance choices, which supports the observation that OLS model is stable despite the presence of multicollinearity.

4.2 Logistic Regression

A logistic regression model was used to analyze how the characteristics of applicants influence the likelihood of being classified as a bad credit risk. Since logistic regression is a nonlinear model, unlike previous methods, all predictors were re-entered at this stage. This included variables that were removed from earlier models. The reintroduction was designed to allow variables with no linear explanatory power in the OLS model to potentially show additional explanatory power in the logit model.

To model the probability of credit risk, logistic regression assumes that the log-odds of being classified as a bad credit risk are a linear function of the predictors. The model estimates coefficients β using Eq. (27).

Starting with the complete set of preprocessed predictors, a logistic regression model was first estimated using maximum likelihood. The initial model indicated that several predictors and groups of dummy variables were insignificant both statistically and theoretically. To create a more interpretable model, insignificant variables were removed, resulting in a reduced logistic regression model. Continuous variables with high p-values were dropped, and for categorical variables, entire dummy groups were eliminated if all encoded levels had high p-values. However, if at least one level of a dummy-encoded variable had a significant p-value or a strong theoretical reason, the variable was retained. After this reduction, the final logistic regression model included 13 predictors and an intercept. Fig. 4 visualizes the reduced logistic regression coefficients with 95% confidence intervals, showing only statistically significant predictors for readability. Tab. 5 in the Appendix provides the corresponding numeric estimates for all coefficient estimates in the reduced model.

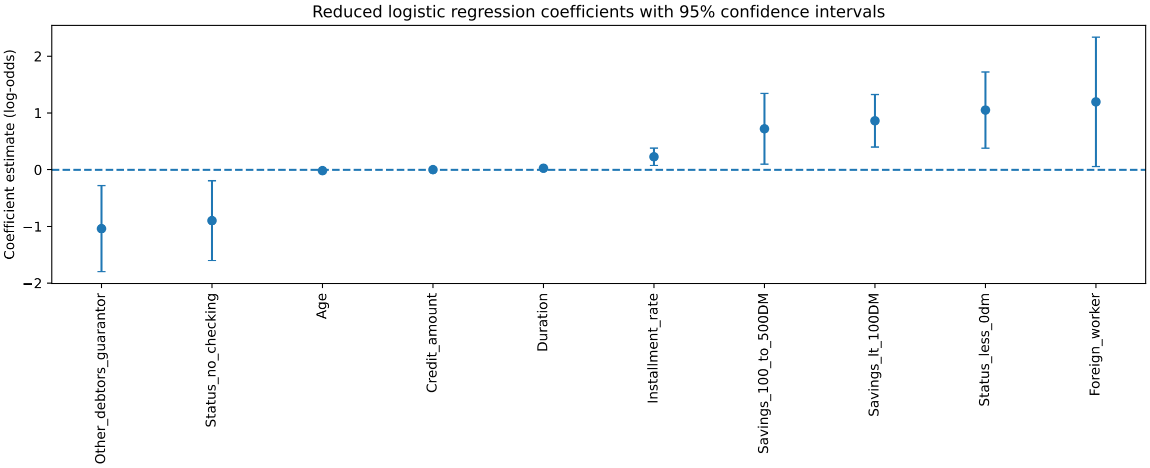


Figure 4: Reduced logistic regression model Coefficient visualization

The reduced logistic regression achieved a pseudo-R^2 of 0.179, and the likelihood ratio test is highly significant ($LLRp = 2.0 \times 10^{-39}$), indicating that the predictors collectively explain a substantial portion of the variation in credit risk.

The estimated coefficients of the reduced model generally align with economic intuition and the estimates obtained from the OLS model. Longer loan *Durations*, higher *Installment rates*, and larger *Credit amounts* increase the likelihood of being classified as a bad credit risk. *Age* has a negative coefficient, indicating that older applicants are, on average, less likely to be considered bad credit risks. The variable for *Foreign workers* has a large positive coefficient, indicating higher estimated log-odds of being classified as bad credit risks compared to the reference group. The *Savings* variable indicates that lower savings amounts are associated with higher estimated log-odds of being classified as a bad credit risk.

The checking account *Status* variable uses applicants with more than 200 DM (Deutsche Mark) as the reference category. Relative to this baseline, applicants with a balance below 0 DM have significantly higher log-odds of being classified as bad credit risks, with an odds ratio of $e^{1.051} \approx 2.86$. The no checking account category has lower estimated log-odds compared to the baseline, with an odds ratio of $e^{-0.898} \approx 0.41$.

To evaluate the performance of the model on unseen data, the sample was divided into a training set (70%) and a test set (30%) using stratified sampling. The predicted probabilities were calculated for the test set and converted into class labels with a threshold of 0.50. Tab. 2 presents the resulting confusion matrix. Out of the 210 good credit applicants, 187 were correctly classified as good, while 23 were misclassified as bad. Out of the 90 bad credit applicants, 36 were correctly classified as bad, and 54 were misclassified as good. This results in an overall accuracy of $(187+36)/300 \approx 0.743$. The model correctly identifies approximately $187/210 \approx 89\%$ of good credit applicants but only $36/90 = 40\%$ of bad credit applicants. Overall, the confusion matrix shows that the model performs significantly better at identifying good credit applicants than bad ones at the 0.50 threshold.

Table 2: Confusion matrix for logistic regression (threshold = 0.5)

	Predicted good (0)	Predicted bad (1)
Actual good (0)	187	23
Actual bad (1)	54	36

The classification performance of the model depends on the chosen threshold. A lower threshold classifies more applicants as bad credit risks, which increases the detection rate of bad applicants (reducing false negatives), but also increases false positives. Since misclassifying a bad applicant as good is a more critical error in credit risk management, performance is reported at a lower threshold (t = 0.2) in addition to the conventional benchmark threshold (t = 0.5) in Tab. 3.

Table 3: Confusion matrix for t = 0.2

	Pred. good (0)	Pred. bad (1)
Actual good (0)	112	98
Actual bad (1)	13	77

At t = 0.2, the model correctly identifies 86% of bad credit applicants (77/90). This reduction in false negatives comes at the cost of flagging more good applicants as bad (98/210).

To complement the confusion matrix results, the reduced logistic regression model's classification performance was also evaluated using the receiver operating characteristic (ROC) curve on the test set. The area under the ROC curve (AUC) is 0.777. This means that if we randomly select a good credit applicant and a bad credit applicant, there is about a 78% chance that the model assigns a higher predicted probability of being a bad credit risk to the bad credit applicant than to the good credit applicant. The model is therefore clearly more accurate than random guessing, but far from perfect at separating good- and bad credit risks. The ROC curve for the reduced logistic regression model is shown in Fig. 5.

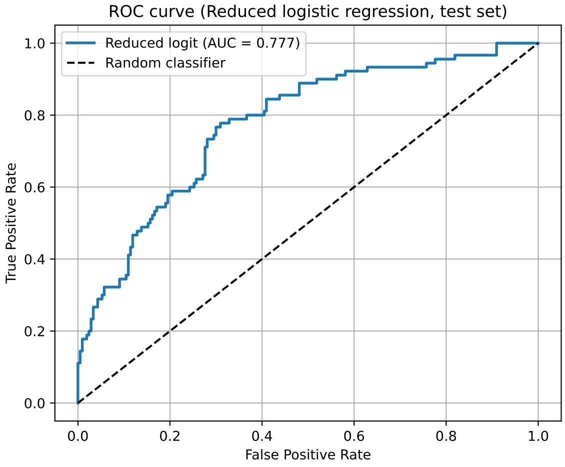


Figure 5: ROC AUC Curve (Reduced logistic regression model)

As a Bayesian robustness check, the reduced logistic regression model was also re-estimated using Maximum A Posteriori (MAP) estimation under weakly informative priors. The resulting coefficient estimates and predictive performance were essentially unchanged compared to the baseline logistic regression model. On the test set, the ROC AUC remained at 0.78, and for both selected thresholds (0.2 and 0.5), the confusion matrix and classification rates were identical. This suggests that the reduced logistic regression results are stable under this Bayesian specification, and that the main conclusions are driven by the data rather than prior assumptions.

4.3 Result Validation to Prior Studies

Comparing the results with previous research on the topic, the findings are consistent with the existing literature. Hand & Henley (1997) and Thomas (2000) emphasize the central role of financial characteristics and repayment history as indicators of credit risk classification. In these articles, demographic variables typically play a supportive role. The significant roles of credit-related and financial characteristics, *Duration*, *Credit amount*, *Checking account status*, and *Credit history* in the models here align with that evidence. Özdemir (2008) finds that financial variables have a stronger influence on risk classification than demographic variables, which is consistent with the findings of this study. The use of logistic regression as the main classification model also follows the pattern of earlier literature. Logistic regression is presented as a standard and widely popular tool for credit risk scoring and related topics (Hand & Henley, 1997; Thomas, 2000; Peng, Lee & Ingersoll, 2002).

5 Conclusions

The purpose of this thesis was to identify the most significant indicators of credit risk using interpretable regression models. The contribution of the study is to demonstrate how classic interpretable regression methods can be combined into a coherent analysis for credit risk using the German Credit Data (Hofmann, 1994). The combination of a linear regression model and the logistic regression model forms a path from descriptive modelling to probabilistic estimation, while the Bayesian interpretation of Tikhonov regularization provides an additional perspective on inspecting parameter uncertainty and model stability.

The analysis examines how demographic, financial, and credit-related characteristics are associated with the likelihood of being classified as a bad credit risk, using statistically interpretable methods. In practice, the results highlight the importance of lenders carefully inspecting the applicant's account status, credit history, and the loan's terms. The study also shows that lenders should pay attention to applicants' demographic characteristics, although in this study their association with the risk classification is generally weaker than that of financial and credit-related variables.

The main research question was which applicant characteristics are most strongly associated with being classified as a bad credit risk. In both linear and logistic models, the most significant predictors were variables that describe the applicant's financial, loan, and credit-related characteristics. Longer *Loan durations*, larger *Credit amounts*, and higher *Installment rates* are associated with a higher likelihood of being classified as a bad credit risk, which aligns with basic economic intuition. In addition, *Checking account* status emerges as an important discriminator in the logistic model. Applicants with a checking account balance below 0 DM (Deutsche Mark) exhibit a substantially higher estimated risk relative to the reference group, while the category of no checking account is associated with a lower estimated risk in this dataset. Savings are also informative. Compared to the reference category, lower savings categories are associated with higher estimated risk, although not all savings levels show statistically significant differences.

The first supporting question was about demographic characteristics. Demographic predictors are less informative than financial and credit-related variables, but *Age* shows a statistically significant, modest negative association with credit risk in both logistic and linear models, indicating that older applicants are less likely to be classified as bad credit risks. The *Foreign worker* indicator is associated with a higher estimated risk in both the logistic and linear regression models, but these results should be interpreted with caution because the foreign worker category is rare in the data. The result may reflect unobserved institutional or socioeconomic differences not captured by the available predictors, rather than a stable causal relationship.

The second supporting research question examined how model performance and robustness can be evaluated using regression diagnostics. In the reduced logistic regression, the overall fit is statistically moderate (pseudo-$R^2 = 0.179$), and the likelihood ratio test is highly significant. In out-of-sample evaluation using a stratified 70/30 train-test split, the model achieves a ROC AUC of 0.78, which indicates that it is more accurate than random guessing with the uneven target variable distribution. At the 0.2 threshold, the model achieves a high recall for bad credit applicants, which is highly important to reduce false negatives. This comes at the cost of more false positives and lower overall accuracy.

The limitations of this study should be considered. The analysis relies on a single dataset from one historical context, consisting of records for 1000 credit applicants with 20 predictors and a binary target variable classifying applicants as good or bad. As a result, the findings may not generalize to broader markets or different time periods. In addition, the binary outcome restricts the analysis to a coarse two-class setting, even though real-world credit risk is often assessed on a continuum or through multiple rating grades. The set of predictors is also limited and omits potentially important information, such as annual income, which could affect both coefficient estimates and predictive performance. Finally, the empirical comparison is restricted to interpretable regression models, so the study does not evaluate whether different machine learning methods would improve predictive accuracy on this dataset.

The limitations open several directions for future research on this topic. An extension for future research would be to compare the regression models used in this thesis with machine learning approaches, such as random forests, to assess the trade-off between interpretability and performance. The research could also focus on finding ways to handle multicollinearity among the predictors. This could make the models more stable and the effects of the variables easier to interpret, for example by studying how different prior strengths affect coefficient shrinkage and stability in the Bayesian interpretations.

References

Goodman, B. & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a “right to explanation”. *AI Magazine* 38(3), pp. 50–57. DOI: 10.1609/aimag.v38i3.2741.

Hand, D. & Henley, W. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 160(3), pp. 523–541. DOI: 10.1111/j.1467-985X.1997.00078.x.

Hofmann, H. (1994). *Statlog (German Credit Data)*. UCI Machine Learning Repository. DOI: 10.24432/C5NC77.

Lessmann, S., Baesens, B., Seow, H.-V. & Thomas, L. (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. DOI: 10.1016/j.ejor.2015.05.030.

Mian, A. & Sufi, A. (2009). The consequences of mortgage credit expansion: Evidence from the U.S. mortgage default crisis. *The Quarterly Journal of Economics* 124(4), pp. 1449–1496. DOI: 10.1162/qjec.2009.124.4.1449.

Morgenthaler, S. (2009). Exploratory data analysis. *WIREs Computational Statistics* 1(1), pp. 33–44. DOI: 10.1002/wics.2.

Özdemir, Ö. (2008). An empirical investigation of payment performance for consumer loans in Turkey. *ODTÜ Gelişme Dergisi (METU Studies in Development)* 35(2), pp. 385–398. ISSN: 1010-9935.

Peng, C.-Y., Lee, K. & Ingersoll, G. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research* 96(1), pp. 3–14. DOI: 10.1080/00220670209598786.

Poslavskaya, E. & Korolev, A. (2023). Encoding categorical data: Is there yet anything “hotter” than one-hot encoding? *arXiv preprint arXiv:2312.16930*. DOI: 10.48550/arXiv.2312.16930.

Thomas, L. (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting* 16(2), pp. 149–172. DOI: 10.1016/S0169-2070(00)00034-0.

A Reduced OLS coefficient table

Table 4: Reduced OLS coefficient estimates with 95% confidence intervals (0.025 and 0.975 quantiles)

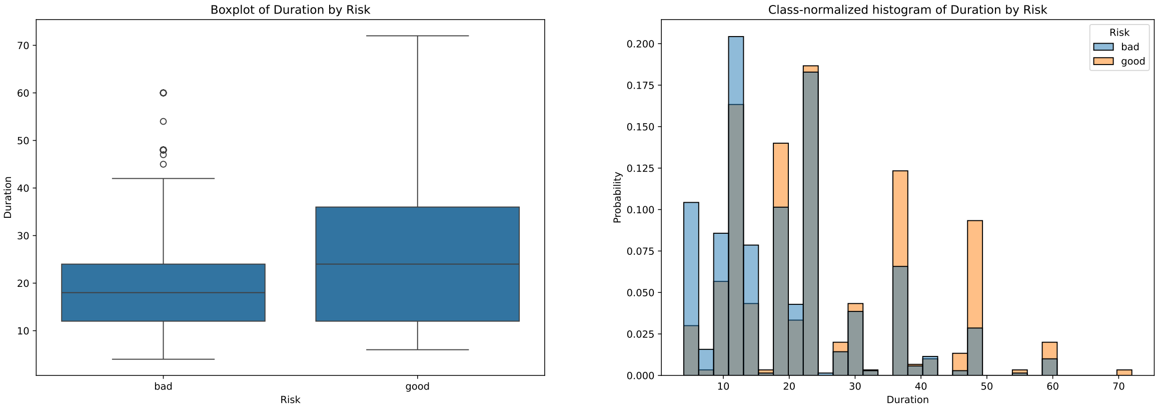
Variable	**β Estimate**	**CI 2.5%**	**CI 97.5%**
const	0.0181	-0.1710	0.2080
Duration	0.0049	0.0020	0.0080
Credit_history	-0.0647	-0.0890	-0.0410
Credit_amount	0.00001536	0.00000242	0.00002830
Employment_duration	-0.0248	-0.0470	-0.0030
Installment_rate	0.0376	0.0130	0.0630
Age	-0.0020	-0.0040	0.0000
Foreign_worker	0.1701	0.0320	0.3080
Status_0_to_200dm	0.2101	0.1450	0.2750
Status_less_0dm	0.3234	0.2590	0.3880
Status_more_200dm	0.1021	-0.0070	0.2110
Purpose_education	0.1776	0.0600	0.2950
Purpose_new_car	0.1222	0.0600	0.1850
Purpose_used_car	-0.1191	-0.2070	-0.0310
Other_debtors_co_applicant	0.0828	-0.0460	0.2110
Other_debtors_guarantor	-0.1469	-0.2630	-0.0310

B Reduced logistic regression coefficient table

Table 5: Reduced logistic regression coefficient estimates with 95% confidence intervals (0.025 and 0.975 quantiles)

Variable	**β Estimate**	**CI 2.5%**	**CI 97.5%**
const	-4.9742	-7.4250	-2.5230
Duration	0.0255	0.0090	0.0420
Credit_amount	0.00007718	0.00000463	0.00015000
Installment_rate	0.2274	0.0740	0.3810
Age	-0.0173	-0.0310	-0.0030
Foreign_worker	1.1956	0.0540	2.3370
Status_0_to_200dm	0.6196	-0.0630	1.3020
Status_less_0dm	1.0512	0.3790	1.7230
Status_no_checking	-0.8984	-1.6010	-0.1960
Other_debtors_co_applicant	0.3589	-0.3690	1.0870
Other_debtors_guarantor	-1.0399	-1.7980	-0.2810
Savings_100_to_500DM	0.7210	0.0980	1.3440
Savings_gt_500DM	0.1197	-0.5750	0.8140
Savings_lt_100DM	0.8621	0.4000	1.3250

C Duration (boxplot and histogram) by Risk

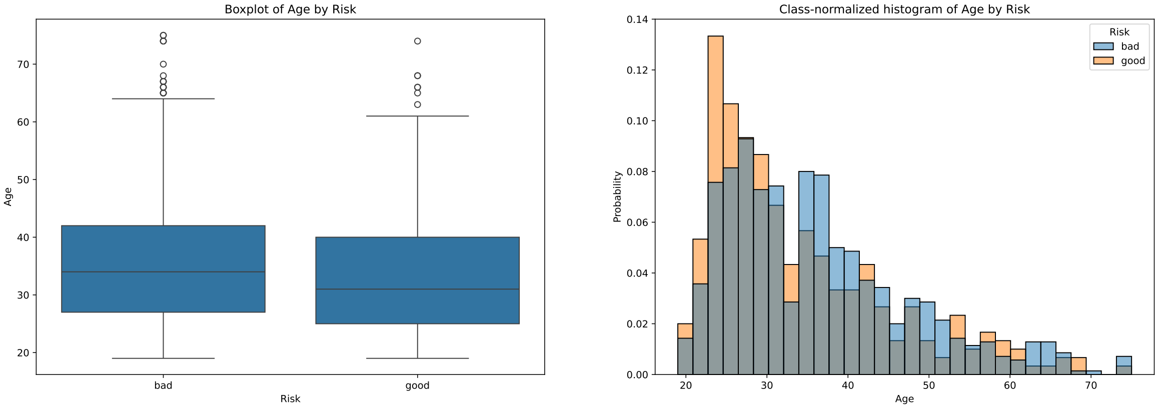


(a) Boxplot of Duration by Risk

(b) Density histogram of Duration by Risk

Figure 6: Duration distributions by Risk class

D Age (boxplot and histogram) by Risk

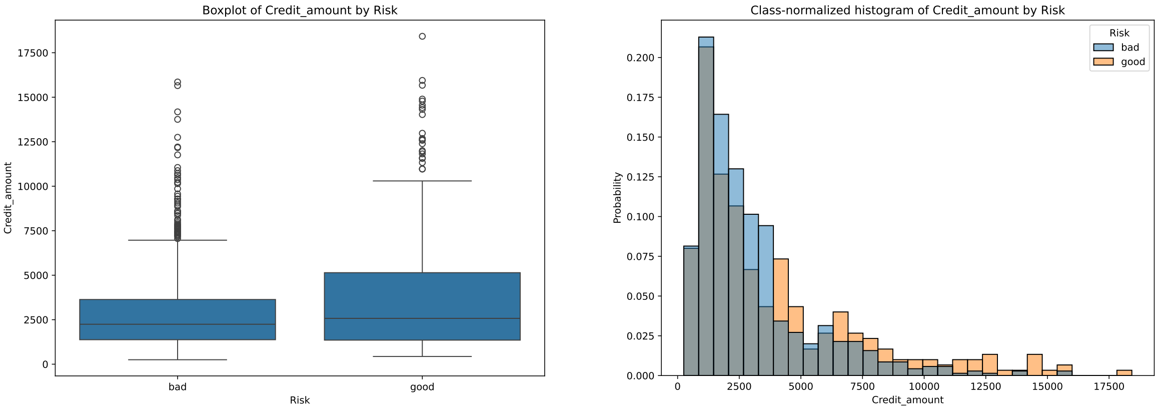


(a) Boxplot of Age by Risk

(b) Density histogram of Age by Risk

Figure 7: Age distributions by Risk class

E Credit amount (boxplot and histogram) by Risk



(a) Boxplot of Credit amount by Risk

(b) Density histogram of Credit amount by Risk

Figure 8: Credit amount distributions by Risk class

F Countplots by Risk (Credit History and Other Debtors)

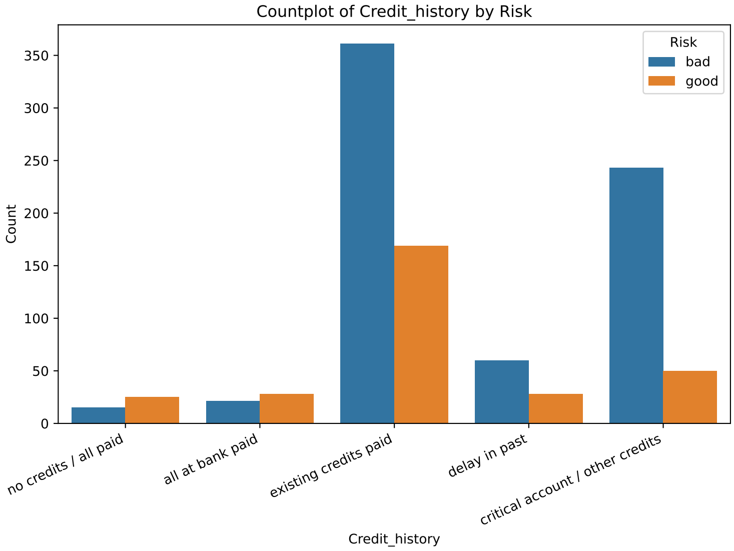


Figure 9: Countplot of Credit History by Risk

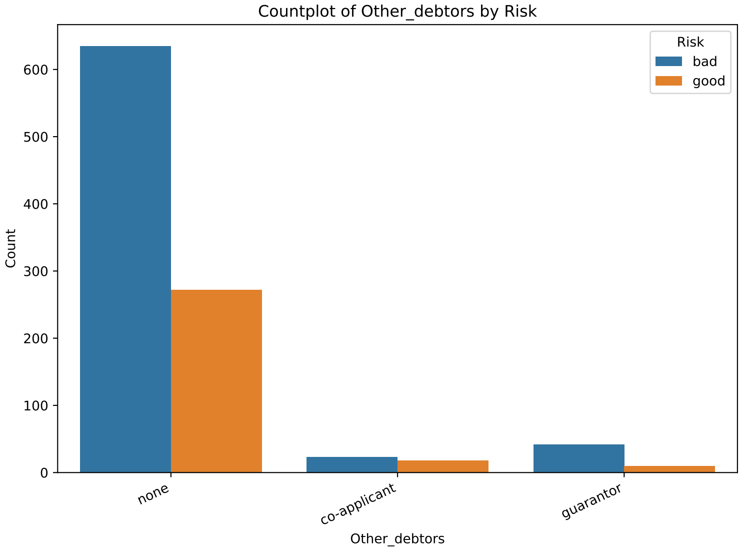


Figure 10: Countplot of Other Debtors by Risk

G Countplots by Risk (Installment Rate and Savings)

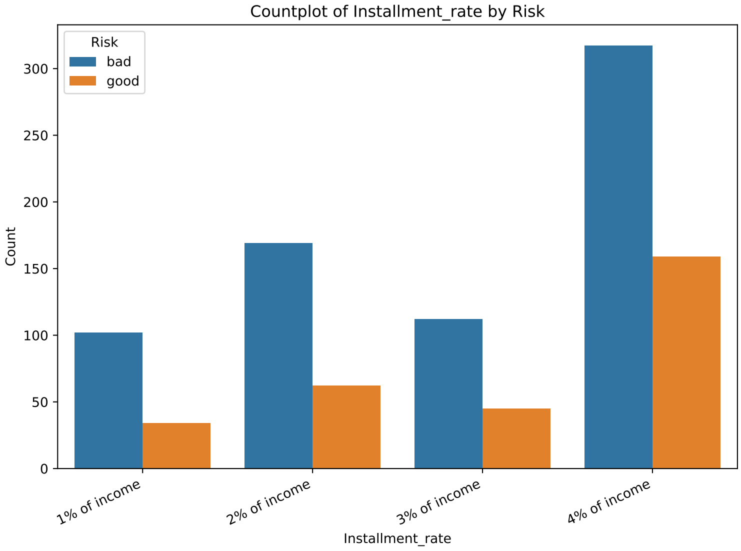


Figure 11: Countplot of Installment Rate (%) by Risk

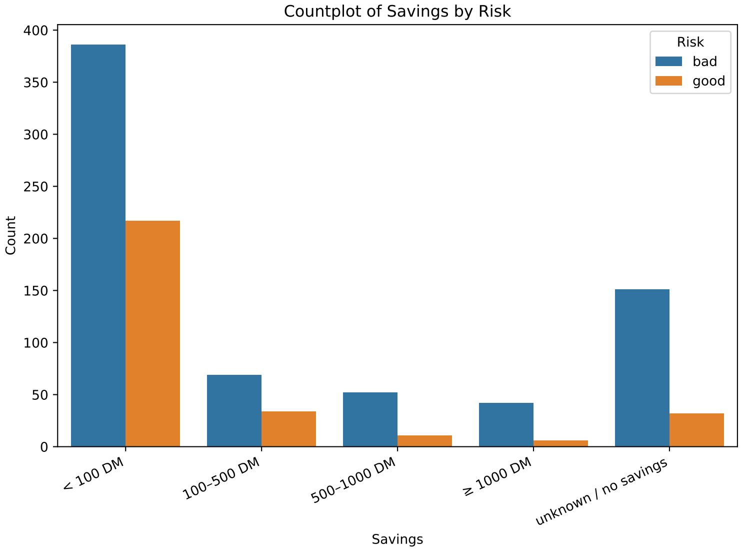


Figure 12: Countplot of Savings (DM) by Risk

H Correlation heatmap of numeric variables

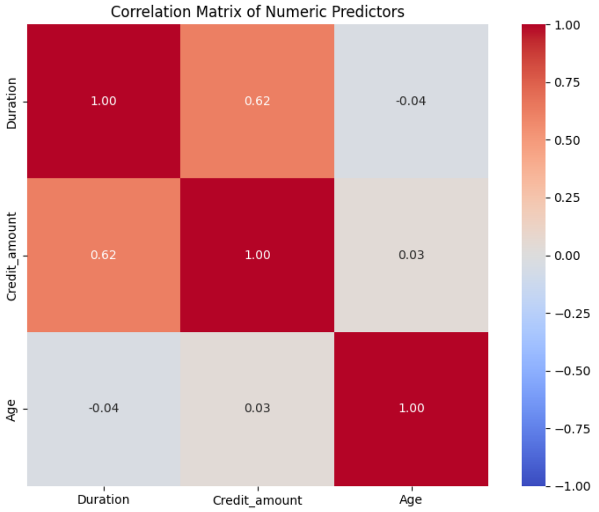


Figure 13: Correlation heatmap of numeric variables