Delft University of Technology
Faculty of EEMCS

Mortgage Default Risk Modeling

Improving Credit Acceptance through Logistic and Multivariate Isotonic Regression

by
Jing Jing Wong

to obtain the degree of Master of Science in Applied Mathematics
at the Delft University of Technology,
to be defended publicly on 25-06-2025.
(Public Version)

Student number:	4555317		
Project duration:	Septemmber 1, 2024 – June 20, 2025		
Thesis committee:	Dr. F. Mies,	TU Delft,	daily supervisor
	Dr. C. Kraaikamp,	TU Delft,	responsible supervisor
	L. van der Poel,	Achmea,	external supervisor

TU Delft

Abstract

The thesis focuses on optimising the mortgage acceptance process at Achmea Bank, based on improved probability of default (PD) modelling. The objective is to improve the existing Advanced Internal Ratings-Based (A-IRB) model, initially designed for calculating capital requirement, and better tailor it to the credit acceptance mechanism. This research identifies areas where adjustments are necessary, including feature selection optimisation, the use of a more suitable target variable, and the exploration of multivariate isotonic regression as a non-parametric model for better estimating the nonlinear interactions among features and the target variable.

Earlier on, the thesis introduces an adapted dataset to mimic the setting of a mortgage acceptance process. This is where only the first observation of each facility in the dataset is used, unlike the original A-IRB model, which worked with more than one observation per facility. Missing values are appropriately addressed, and outliers are addressed through capping and flooring techniques to ensure data quality.

The research employs logistic regression to model the probability of default, with a focus on the feature selection process. Comparison among different models with target defaulting in 12 months and 24 months reveals that changing the target variable and improving the feature selection process results in better model performance.

The second focus of this thesis is isotonic multivariate regression, giving greater flexibility by fitting a nonlinear relationship between the risk drivers and target but with the constraint of monotonicity. The minimum Redundancy Maximum Relevance (mRMR) algorithm is used for the selection of features to reduce computational time, which has much less computational time than traditional methods.

Comparison across models reveals that the isotonic regression model, with greater recall and more accurate default detection, has a higher false-positive rate. Logistic regression with a 24-month target, on the other hand, strikes a better trade-off between precision and recall, leading to fewer false alarms and a lower number of flagged cases.

Overall, the thesis demonstrates that logistic regression and isotonic regression models both provide valuable information to Achmea Bank's mortgage acceptance. While logistic regression with a 24-month goal is an appropriate balance between recall and precision, isotonic regression can provide recall-improvement potential at the cost of precision. Subsequent studies should focus on reducing false positives within the isotonic regression model, exploring reject inference procedures to adjust for potential sample bias, and examining other forms of classification procedures that can more effectively handle class imbalance.

Contents

1 | Introduction

Financial institutions play a crucial role in the contemporary economy by providing financial capital to their customers. Banks normally get most of their revenues from mortgages [15]. Financial institutions' greatest challenge is the possibility that counterparties do not perform in time. Banks should maintain liquid assets to cover such losses. Defaults occurring too frequently can result in substantial financial losses and even bankruptcy.

For example, the large investment bank Lehman Brothers filed for bankruptcy during the 2008 subprime mortgage crisis [1]. The immediate cause of this crisis was the housing bubble in the US. In the early 2000s, interest rates were kept low to stimulate economic growth, making borrowing more affordable. As expected, this resulted in rapid growth in the housing market and a rapid increase in housing prices. As housing prices rise rapidly, lenders are increasingly offering more and more risky loans to borrowers with poor credit. In 2006, house prices reached their peak, and the prices started to decrease. When this happens, many borrowers default on their loans, resulting in a wave of foreclosures and substantial losses for financial institutions.

Credit risk arises when an institution lends money to a counterparty. For Achmea Bank, this occurs when clients are unable to pay their mortgage payments, potentially resulting in a loss for the bank. Since the subprime mortgage crisis of 2008, credit risk modelling has gained increasing importance [21]. To prevent similar events, the Basel Committee on Banking Supervision (BCBS), the primary global standard setter for banking regulation, strengthened the framework that defines minimum capital requirements for banks to cover potential future losses.

1.1 Basel Committee on Banking Supervision (BCBS)

Under Basel regulations, banks are required to maintain a certain amount of capital as a buffer against unexpected losses. The required capital is expressed as a percentage of the bank's risk-weighted assets (RWA), namely 8% [3]. The riskier the asset, the higher the weight assigned. Basel II [2] introduced two approaches for calculating the RWA: the standardised approach and the internal rating-based (IRB) approach. The standardised approach assigns fixed risk weights to assets as provided by regulators, while the IRB approach allows banks to use internal models to estimate risk parameters, such as the probability of default (PD), loss-given-default (LGD), and exposure at default (EAD). A higher accuracy of these models will lead to a lower risk for banks, providing a clear incentive for research in this area.

1.2 A-IRB model

Achmea Bank uses the Advanced Internal Ratings-Based (A-IRB) approach, which offers greater flexibility in modelling credit risk but requires more stringent regulatory oversight. Under the A-IRB framework, Achmea estimates key risk parameters-namely, the probability of default (PD), loss given default (LGD), and exposure at default (EAD)-using its internal data. This enables a more accurate and risk-sensitive calculation of capital requirements.

The capital requirement per unit of exposure is similar to the unexpected loss per unit of exposure and is denoted as K. The total Risk-Weighted Assets (RWA) are then calculated as:

$$\text{RWA} = \frac{1}{0.08} \cdot K \cdot \text{EAD}, \tag{1.1}$$

Where the factor $\frac{1}{0.08}$ is a scaling constant that reflects the regulatory minimum capital requirement of 8%. The capital requirement per unit of exposure, denoted by K, is a function that incorporates the PD and the LGD, as described in Article 31.5 [3]. If the PD and LGD are high, then K will also be high, leading to a higher RWA. This implies that the bank must hold a larger amount of capital to cover potential unexpected losses.

This approach enhances risk sensitivity by reflecting the actual characteristics of the bank's exposures, which can lead to potentially lower capital requirements for low-risk assets.

1.3 Extending the A-IRB Model for Credit Approval

The primary use of the A-IRB model is to model the capital requirements. However, BCBS stated in Article 36.60 [3] that in order to grant permission for the use of the A-IRB approach, banks need to demonstrate that internal ratings used in the calculation of their own capital requirements and associated systems and processes play an essential role in the risk management process and in the credit approval process. Therefore, Achmea has established the acceptance framework, which uses PD to support the decision-making on whether an applicant should be accepted or rejected. Every applicant is given a credit score representing the probability of default. Based on this rating, an applicant will be flagged. If an applicant is flagged, the applicant will be reviewed manually.

The exact accuracy of the A-IRB model within the acceptance framework is currently unknown. However, according to the mortgage desk, the model appears to underperform in practice. A significant number of facilities are being flagged by the model, yet manual reviews often reveal no underlying issues. Furthermore, while the observed probability of default in recent years is approximately 0.05%, the proportion of flagged facilities exceeds 2%. This mismatch between the proportion of defaults and the proportion of flags has led experts at Achmea to conclude that the model's predictive performance can be improved.

The aim of this thesis is to identify potential improvements to the existing A-IRB model when applied to the mortgage acceptance process. The A-IRB model was originally developed for calculating capital requirements, and as such, its design and training phase did not take the mortgage acceptance context into account. To evaluate how well the model functions in this new context, we will begin by assessing the dataset.

The feature set used for A-IRB modelling includes a wide range of attributes that cover data only obtainable after issuing a mortgage. For instance, mortgage payment behaviour features are included in the model. Although these sorts of features yield good predictive power for risk modelling, they are based on post-acceptance data and therefore will not be available during actual acceptance. This creates a mismatch: the model can learn to rely on variables that are not present at use time, rendering it less practical for the acceptance process.

In addition, credit risk modellers dispute the completeness of the dataset in the current acceptance framework and agree that feature selection is where the highest value addition lies. The current approach is very time-consuming, so much so that it is difficult to test different feature combinations effectively. To address this, we look at leveraging the mRMR (Minimum Redundancy - Maximum Relevance) algorithm, which attempts to reduce computation time without influencing the model's performance [26].

Next, we examine the target variable used in the A-IRB model, which is currently defined as default within one year. This target is a common industry standard and is recommended by

experts in credit risk management for capital calculation purposes [4]. However, it is unclear whether this definition is also appropriate for the acceptance process.

Furthermore, to improve the acceptance model, multivariate isotonic regression is introduced. Multivariate isotonic regression is a non-parametric method that models non-linear relationships between multiple independent variables and the target variable while preserving monotonicity. Unlike traditional models that rely on specific functional forms, isotonic regression offers greater flexibility and is particularly well-suited for financial applications, where variables such as income, credit score, and loan size are expected to have monotonic relationships with the probability of default. Compared to other non-parametric techniques, isotonic regression has the advantage of preventing overfitting due to its monotonicity constraint and does not require tuning parameters, making it both robust and easy to implement.

Finally, another key issue in the current framework is the absence of a reliable evaluation metric. Without an appropriate way to assess model performance, it is difficult to determine whether the current A-IRB-based acceptance model is optimal for its intended purpose.

1.4 Research objective and Contribution

Achmea's A-IRB model has been in production for several years, and its performance in the context of capital requirement calculations has been thoroughly analysed and validated. However, similar analysis and validation are lacking in the credit acceptance framework. This thesis contributes to the analysis and development of a more effective credit acceptance process for Achmea Bank.

In recent years, numerous studies have explored the use of advanced machine learning techniques such as random forests, gradient boosting, and neural networks to model the probability of default (PD). These models often demonstrate strong predictive performance. However, due to strict regulatory requirements in the financial sector, particularly those related to explainability and interpretability, such models are typically not permitted in production environments despite their performance. Therefore, a key contribution of this thesis is the development of an interpretable model that aligns with regulatory expectations: **multivariate isotonic regression**.

Univariate isotonic regression has a long-standing history in statistical modelling and has been widely applied in calibration tasks, including in the field of credit risk. For example, it has been used to adjust model outputs so that predicted risk scores correspond more closely to observed default rates [25]. However, despite the monotonic relation between variables such as income and credit rating, to our knowledge, multivariate isotonic regression has not yet been applied to directly model the probability of default in credit scoring in the existing literature.

In summary, the objective of this thesis is to enhance the mortgage acceptance model by improving feature selection, identifying a more suitable target variable, and exploring the potential of multivariate isotonic regression.

The research is guided by the following sub-questions:

1. What modifications should be made to the dataset before applying the model? (Section 2.2)
2. What is the most appropriate target variable for the acceptance model?
3. How should one choose the optimal feature set? (Section 3.2 and 4.4)
4. What evaluation metric should be used to assess model performance? (Section 3.5)
5. How does multivariate isotonic regression perform compared to logistic regression? (Section 4.6.2)

2 | Data and Preprocessing

Before introducing the model, this chapter provides key definitions and background information that will be used throughout the project, along with an explanation of the available dataset. Section 2.1 defines important concepts such as mortgage, facility, and default and section 2.2 presents the available data sources.

2.1 Definitions

This section provides an overview of some definitions used in this report related to credit risk modelling.

2.1.1 Mortgage

A **mortgage** is a type of loan used to purchase (or maintain) property, usually a house. The borrower agrees to repay the lender over a certain period of time, which can range from 10 to 30 years (or even more). The property itself serves as **collateral**, meaning the lender has the right to sell the house if the borrower does not make their payments.

In order to apply for a mortgage, a borrower must ensure that they meet several requirements, for example, a minimum credit score. In the application process, the lender will ask the borrower to provide evidence such as bank statements, proof of current employment and income. If the application is approved, the lender will offer the borrower a loan up to a certain amount, depending on the borrower's income, at a particular interest rate. The ratio between the loan amount and the value of the underlying property is called the **Loan-to-Value (LTV)**.

Mortgages are available in various types, for example, mortgages for self-employed professionals. The various types are also called **product**.

2.1.2 Facility

A **facility** is a credit obligation arising from a contract between an obligor (individual or joint) and, in our case, Achmea Bank or one of its partners. The facility must be economically owned by Achmea Bank, backed by the same collateral. From every facility, Achmea has monthly records about its characteristics. Every record of a facility is called an **observation**. One facility can have multiple observations, but every observation corresponds to only one facility.

2.1.3 Default

The definition of default has been omitted from this publicly available version of the document.

2.2 Available data

This research utilizes the Historical Model Dataset (HMDS) from Achmea Bank, covering a period of 9 years, from [redacted] to [redacted]. The HMDS contains more than 600,000

yearly observations. Table 2.1 provides an overview of the number of observations, number of defaults, and the observed default rate (ODR) in December of each year.

Table 2.1: Overview of the number of observations, number of defaults, and observed default rate (ODR) for each year in the period from [redacted] to [redacted] in the Historical Model Dataset (HMDS) from Achmea Bank.

[REDACTED]

The unique facilities in the HDMS is randomly split into a training and test set based on the 80-20 rule. To assess whether the training and test sets are representative, we compare their characteristics. The various characteristics are presented in Table 2.2. The Loan-to-Value (LTV) is calculated by dividing the loan amount by the appraised value of the property. For instance, if the appraised value is €400,000 and the desired loan is €300,000, the LTV is calculated as $\frac{€300,000}{€400,000} = 0.75$. A lower LTV suggests a less risky facility. The Central Credit Registration Office (in Dutch, "Bureau Krediet Registratie" (BKR)) keeps records of private parties that have taken out a loan and based on the number of loans recorded and payment behaviour, parties receives a score. This score ranges from A to L, with A being the best, and the values being monotonic. The median values for both sets are identical, with other values showing slight differences. The training set has a higher percentage of clients who are or have been in arrears, but a lower default rate and LTV.

Table 2.2: Statistics of the training and test datasets.

[REDACTED]

The dataset contains various risk drivers and the target variable, which can be grouped into three main categories. First, it includes obligor characteristics such as BKR registration, payment arrears, income, and marital status. Second, it provides information on loan characteristics, including the proposition type, outstanding amount, remaining interest period, and the presence of a bridge loan. Finally, the dataset includes collateral characteristics such as the region, loan-to-value (LTV) ratio, and the type of collateral. The target variable is defined as a binary indicator representing whether a default occurs within one year. Formally, it is defined as:

$$y_i = \begin{cases} 1, & \text{if observation i will default within one year,} \\ 0, & \text{otherwise.} \end{cases}$$

2.2.1 Missing data and outliers

The section is subject to a non-disclosure agreement and have been omitted from this publicly available version.

2.3 Datasets

Dataset A refers to the dataset used by Achmea, while **dataset B** is the modified dataset. Specific details about both datasets have been omitted from this publicly available version.

In the following chapters, we will train both the logistic regression model (Chapter 3) and the multivariate isotonic regression model (Chapter 4) using dataset B. The performance of these models will be compared to each other, as well as to the current model trained on dataset A.

3 | Modelling the probability of default using logistic regression

Accurately estimating the probability of default (PD) is a fundamental aspect of credit risk modelling, as it directly informs lending decisions. One commonly used approach for PD modelling is logistic regression - a parametric technique that estimates the likelihood of default based on a set of identified risk drivers.

This chapter explores the application of logistic regression in modelling the probability of default. We begin by introducing the logistic regression model and its role in credit risk assessment (Section 3.1). Subsequently, we present the feature selection process (Section 3.2). Finally, we evaluate the model's performance and interpret the results to assess its predictive accuracy and practical relevance in estimating default risk.

3.1 Logistic Regression

Suppose we are given a training set consisting of independent observations $(X_i, y_i)_{i=1}^n$ with $X_i \in \mathbb{R}^p$ and $y_i \in \{0, 1\}$.

Then the logistic regression model postulates:

$$p(X_i; \beta) = \mathbb{P}(y_i = 1 \mid X_i) = \frac{e^{\boldsymbol{\beta}^T X_i}}{1 + e^{\boldsymbol{\beta}^T X_i}}, \quad \text{for } i = 1, \ldots, n,$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of coefficients associated with the p features.

Logistic regression models are usually fit by maximum likelihood using the conditional likelihood of $y_i = 1$ given X_i. The likelihood function for logistic regression can be written as:

$$\ell(\beta) = \prod_{i=1}^{n} \left[p(X_i; \beta)^{y_i} (1 - p(X_i; \beta))^{1-y_i} \right],$$

The log-likelihood is the natural logarithm of the likelihood:

$$\log(\ell(\beta)) = \sum_{i=1}^{n} \left[y_i \log(p(X_i; \beta)) + (1 - y_i) \log(1 - p(X_i; \beta)) \right].$$

To find the maximum likelihood estimator $\hat{\beta}$, we solve the following equation:

$$\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - p(X_i; \beta)) X_i \tag{3.1}$$

Equation (3.1) can be solved using the Newton-Raphson algorithm (see [16], p. 120).

3.2 Feature Selection

In the previous section, we showed the mathematical framework for logistic regression, where the goal is to model the probability of a binary outcome as a function of the predictors, using a logistic function. The coefficients of the logistic regression model are typically estimated using maximum likelihood estimation (MLE), which requires selecting appropriate features that can effectively explain the variability in the outcome variable. However, in practice, the choice of features can significantly impact the model's performance. When dealing with a large number of predictors, some of them may be irrelevant or highly correlated with others, which can lead to overfitting. Therefore, before fitting the logistic regression model, it is important to carefully select the most relevant features.

Especially, in credit risk assessment, feature selection is an important task due to the large number of features. When datasets contain numerous redundant variables, the computational cost of training models increases significantly. For example, consider a dataset with p features. Given this set of features, we can choose to use all, some, or even none of them to enter into the model. If the goal is to find the best predictive model using these p features, then the total number of unique subsets equals 2^p, where we include the possibility of having an empty set. The number of possible models increases exponentially with the number of features. Therefore, it is not computationally feasible to obtain the best predictive model by analysing all possible subsets of features. Furthermore, reducing the number of features not only simplifies the model but also enhances its interpretability. The accuracy and reliability of these models also rely on this process. Indeed, recent studies on credit risk modelling shows the importance of feature selection, in particular in improving the model's performance [18].

Feature selection has been studied across various financial domains with several approaches to identify the most relevant features for the PD models. The available data to estimate the PD can be divided in three categories:

- Borrower risk characteristics: characteristics related to individual clients such as age, occupation and debt-to-income ratio.
- Transaction risk characteristics: characteristics about the loan. For example, the loan purpose, the house type and the loan-to-value ratio.
- Aggregate risk characteristics: characteristics which are the same for multiple clients. Examples are macroeconomic variables as house price indices, unemployment rates and GDP growth rates.

Depending on the purpose of the loan, the focus can vary between borrower risk characteristics and aggregate risk characteristics. Variables as interest rate, current income and loan-to-value ratio are often highlighted as crucial in PD models [7], [19] [24]. However, studies also show that macro-economic characteristics have a significant impact on the PD models [23]. For example, a rise in the unemployment rate means that more people lose their primary income, affecting their ability to make the payments.

The literature on PD models for mortgages in the Netherlands is very limited. First, Medema [20] published a paper about the validation of a PD model, focussing on the effects of risk drivers. Using data from 2000-2003, the probability of default within a year is modelled using a logit model. The model includes the following risk drivers: loan-to-value (loan-to-value ratio), loan-to-income (loan-to-income ratio), duration (maturity), type of mortgage and a due payment indicator. The model shows that the expired duration has a negative relation with the probability of default, meaning that the closer to the maturity date, the lower the PD. The other variables have a positive relation with the target variable.

In addition to the paper of Medema, Kroot [18] analyses the impact of the financial crisis on the probability of default covering the period from 2001 until 2012. The study develops

a statistical model including both internal and external risk factors. External risk factors are influenced by market conditions and are not directly linked to client characteristics.

The details of the feature selection process have been omitted from this publicly available version.

3.3 Model Performance

The model's performance is evaluated using the GINI index. Further details on the performance is omitted from this publicly available version.

GINI Index

Let us now introduce the notation and procedure used to define and compute the *GINI coefficient*, a measure of discriminatory power commonly used in binary classification tasks, particularly in credit risk modelling.

Let $n \in \mathbb{N}$ denote the total number of observations, and let

$$y = (y_1, \ldots, y_n) \in \{0,1\}^n \quad \text{and} \quad \hat{y} = (\hat{y}_1, \ldots, \hat{y}_n) \in \mathbb{R}^n$$

be the binary outcome vector (e.g., default indicators) and the corresponding predicted probabilities, respectively. We assume that the observations are sorted in descending order of predicted values, so that

$$\hat{y}_1 \geq \hat{y}_2 \geq \cdots \geq \hat{y}_n.$$

For each index $i \in \{1, \ldots, n\}$, we define the cumulative proportion of the population as

$$C_i := \frac{i}{n},$$

and the cumulative proportion of true positives as

$$T_i := \frac{\sum_{j=1}^{i} y_j}{\sum_{j=1}^{n} y_j}.$$

We define $C_0 := 0$ and $T_0 := 0$. The points (C_i, T_i) forms the Lorenz curve as i goes from 0 to n. To approximate the area under this curve, we apply the trapezoidal rule:

$$\text{AUC} := \sum_{i=1}^{n} \frac{T_i + T_{i-1}}{2}(C_i - C_{i-1}).$$

The GINI coefficient is then defined by:

$$\text{GINI} = 2 \cdot \text{AUC} - 1.$$

The training set achieves a GINI of 65.8%, while the test set yields only 30.9%. This significant difference indicates overfitting, implying that the model is overly tailored to the training data. Such a situation may occur when the training set does not include enough examples of defaults or when the model is too complex.

3.4 Comparison across models

Table 4.5 presents a comparison of the risk drivers used in the three models. For convenience, we define the models as follows:

- Model $LR_{A,12}$: Logistic regression model trained on dataset A (see Section 2.3), with the target variable being default within 12 months.
- Model $LR_{B,12}$: Logistic regression model trained on dataset B (see Section 2.3), with the target variable being default within 12 months.
- Model $LR_{B,24}$: Logistic regression model trained on dataset B, with the target variable being default within 24 months.

Table 3.1: List of risk drivers in final model. The first column represents the model trained using logistic regression on Dataset A, with the target variable being default within 12 months. The second and third columns are both trained on Dataset B. In the second column, the target variable is set to default within 12 months, while the third column uses default within 24 months as the target variable.

[REDACTED]

In Table 3.2, the GINI scores of the different models are evaluated on the training and test sets. The GINI score of model $LR_{A,12}$ for the training and test sets is 64.7% and 19.2%, respectively. It is important to note that the model is trained on the training set of dataset A, and its performance is evaluated on the test set of dataset B. This is due to the fact that dataset B only contains the first observation, which mimics the data available during acceptance. The second model $LR_{B,12}$ has a GINI of 65.8% and 30.9%, respectively. And the third model has a GINI of 62.3% and 55.8%, respectively. Based on these GINI scores, $LR_{B,24}$ performs the best in terms of generalisation to the test set.

Table 3.2: Model Performance in terms of GINI on Different Datasets.

Model	Training	Test
$LR_{A,12}$	64.7%	19.2%
$LR_{B,12}$	65.8%	30.9%
$LR_{B,24}$	62.3%	55.8%

3.5 Acceptance Framework

The outcome of the model provides the probability of default for each facility. This probability is then used to determine whether an applicant should be flagged during the acceptance process. Flagging implies that the client requires a manual review.

Initially, the performance of the model was assessed using the GINI score to evaluate its predictive capabilities. However, within the acceptance framework, a more granular approach is necessary. The focus shifts towards analysing the false default (FD), false healthy (FH), true

default (TD), and true healthy (TH) cases. To gain deeper insights into these classifications, the following accuracy metrics are considered:

- Precision: The proportion of true detected defaults, given by:

$$\text{Precision} = \frac{TD}{FD + TD} \tag{3.2}$$

- Recall: The proportion of actual defaults that are correctly identified, given by:

$$\text{Recall} = \frac{TD}{TD + FH} \tag{3.3}$$

- Accuracy Rate: The proportion of correctly predicted outcomes, given by:

$$\text{ACC} = \frac{TD + TH}{TD + TH + FD + FH} \tag{3.4}$$

- F1 Score: The proportion of correctly predicted outcomes, given by:

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.5}$$

3.5.1 Threshold Optimisation and Model Evaluation via F1 Score

The threshold for whether an applicant should be flagged will be chosen by optimising the F1 score. The F1 score is the harmonic mean of precision and recall, and is particularly useful in situations where there is an imbalance between the classes. By maximising the F1 score, we aim to find a threshold that balances the trade-off between false positives and false negatives, ensuring that we identify as many relevant applicants as possible (high recall) while minimising incorrect flags (high precision). The optimal threshold is determined using predictions on the training data and then applied to the test data to evaluate final model performance. The optimal threshold values for models $LR_{B,12}$ and $LR_{B,24}$ are 0.096 and 0.227, respectively, corresponding to F1 scores of 0.069 and 0.160 (see Appendix A.5).

Table 3.3 summarises the key classification metrics: Accuracy, Recall, Precision, and F1 Score for three logistic regression models (see Appendix A.6 for the corresponding confusion matrices). Model $LR_{A,12}$ achieves the highest accuracy (0.9870) and recall (0.1429), though its precision (0.0769) remains relatively low, resulting in a F1 score of 0.100. Model $LR_{B,12}$ results the lowest recall (0.0714) and precision (0.0333), yielding the weakest F1 score (0.045) among the three. Meanwhile, Model $LR_{B,24}$ demonstrates more balanced performance, with a recall of 0.1034 and the highest precision (0.1154), with an F1 score of 0.109. Overall, while $LR_{A,12}$ shows strength in capturing positive cases, $LR_{B,24}$ offers improved precision and slightly better F1 performance.

Table 3.3: The accuracy, recall, precision and F1 score for all three models.

Model	Accuracy	Recall	Precision	F1 Score
Model $LR_{A,12}$	0.9870	0.1429	0.0769	0.100
Model $LR_{B,12}$	0.9848	0.0714	0.0333	0.045
Model $LR_{B,24}$	0.9823	0.1034	0.1154	0.109

In practice, the flagged facilities must be manually reviewed by the acceptance team, who face a high workload. Therefore, we also analyse the false alarm rate (FAR), which is defined as:

$$FAR = \frac{FD}{FD + TH} \tag{3.6}$$

The FAR for models $LR_{A,12}$, $LR_{B,12}$ and $LR_{B,24}$ is equal to 0.87%, 1.06% and 0.84%, respectively. Lastly, the distribution of predicted default probabilities is compared between actual defaults and non-defaults. The predicted probabilities from model $LR_{B,12}$ are visualized in Figure 3.1 (and for model $LR_{A,12}$ and model $LR_{B,24}$ see Appendix A.7). The plot shows that both defaults and non-defaults are concentrated around a predicted probability of approximately 0.01. As a result, lowering the threshold would increase the detection rate (recall), but it would also substantially raise the false positive rate (see Figure 3.2). This illustrates the natural trade-off between sensitivity and specificity in the model's predictions.

[REDACTED]

Figure 3.1: Distribution of the predicted probability of default. The top panel shows the distribution of predicted probabilities that do not default within one year, with the most predictions concentrated near 0. The bottom panel shows the distribution of defaulted loans, with a broader spread of predicted probabilities.

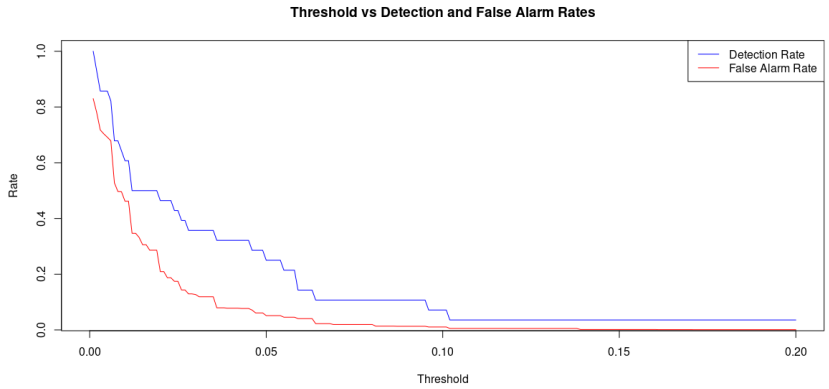

Figure 3.2: Detection and false alarm rates as a function of the classification threshold.

3.5.2 Further Analysis for Credit Policy

The details of this section have been omitted from this publicly available version.

3.6 Conclusion

In this chapter, we explored the application of logistic regression for predicting the probability of default. Through feature selection, we identified the key risk drivers.

The performance of model $LR_{B,12}$ revealed signs of overfitting, as indicated by a substantial gap between the GINI scores on the training set (65.8%) and the test set (30.9%). Despite this,

$LR_{B,12}$ demonstrated improved default probability estimation compared to the current model $LR_{A,12}$, based on GINI score. However, overfitting remains a concern. The last model, $LR_{B,24}$, trained to predict defaults over a 24-month horizon instead of 12 months, yielded a lower GINI score on the training set, but outperformed the other models on the test set, indicating better generalisation. In terms of classification performance, model $LR_{A,12}$ achieved the highest recall, while $LR_{B,24}$ achieved better precision and the lowest false alarm rate.

In the next chapter, we introduce isotonic regression for modelling default probabilities. Unlike logistic regression, isotonic regression makes no parametric assumptions about the relationship between predictors and outcomes, offering greater flexibility to capture complex patterns in the data.

4 | Multivariate Isotonic Regression

Isotonic regression is a non-parametric regression technique that provides greater flexibility compared to parametric methods such as linear regression. For simplicity, multivariate isotonic regression will be referred to as isotonic regression in the following discussion, as we are modelling the probability of default with multiple features. One of its key advantages is the relaxation of restrictive assumptions. While linear regression assumes a strict linear relationship between variables, isotonic regression only requires that the relationship between X and Y be monotonic. This makes it particularly useful in applications where the data is expected to follow a non-decreasing or non-increasing trend, such as credit risk modelling.

Isotonic functions can be used as an underlying assumption in both regression and classification problems. First, consider a nonparametric regression problem of the form: $f^* : [0,1]^d \to \mathbb{R}$, where $d \geq 1$. Let $(\mathbf{x}_i, y_i)_{i=1}^n$ represent a set of independent observations, with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Assume that the relationship between the input and output variables is described by the model

$$y_i = f^*(\mathbf{x}_i) + \varepsilon_i, \quad \text{where } \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n, \tag{4.1}$$

where $\sigma^2 \geq 0$ is unknown, and f^* can be estimated using the isotonic least squares estimator defined in equation (4.2).

The isotonic regression estimator is the least squares estimator over $\mathcal{F}_{EM}^d$, the class of entirely monotone functions on $[0,1]^d$:

$$\hat{f}_{\text{EM}} \in \operatorname*{argmin}_{f \in \mathcal{F}_{\text{EM}}^d} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2. \tag{4.2}$$

In this chapter, we are interested in the performance of multivariate isotonic regression in predicting the probability of default. To this end, we first provide some definitions and properties related to monotonicity in Section 4.1. Then, Section 4.2 describes how the isotonic regression estimator can be found by solving a non-negative least squares problem. We are also interested in the risk behaviour of the estimator $\hat{f}_{\text{EM}}$, which is studied under the standard fixed design squared error loss function. In Section 4.3, we prove that the risk of estimator $\hat{f}_{\text{EM}}$ is bounded by $n^{-2/3}$.

Before applying isotonic regression, preprocessing steps are necessary, for example, to improve computational efficiency. Given the size of the dataset used in this study, selecting an optimal subset of features is essential. To achieve this, we use the Minimum Redundancy Maximum Relevance (MRMR) algorithm, which is described in Section 4.4. The application and evaluation of isotonic regression are presented in Section 4.5. Finally, in Section 4.6, we compare the performance of the different methods.

4.1 Entire Monotonicity

Before defining monotone and entirely monotone functions, we first introduce the concept of domination, which will be used in their definitions. Let $\mathbf{a} = (a_1, \ldots, a_d)$ and $\mathbf{b} = (b_1, \ldots, b_d)$ be

two points in $[0,1]^d$. We say that $\mathbf{a}$ dominates $\mathbf{b}$ if, in every dimension, a_j is greater than or equal to b_j. This is denoted as:

$$\mathbf{a} \succeq \mathbf{b} \iff a_j \geq b_j, \quad \forall j \in \{1, 2, \ldots, d\}.$$

Then, we can define the monotone function.

Definition 4.1 (Monotone Function)**.** *A function $f : [0,1]^d \to \mathbb{R}$ is called monotone if for any $\mathbf{a}, \mathbf{b} \in [0,1]^d$ such that $\mathbf{a} \succeq \mathbf{b}$, it holds that:*

$$f(\mathbf{a}) \geq f(\mathbf{b}).$$

In the univariate case ($d = 1$), the class of monotone functions, $\mathcal{F}_M^1$, coincides with the class of entirely monotone functions, $\mathcal{F}_{EM}^1$. For the bivariate case ($d = 2$), the class $\mathcal{F}_{EM}^2$ consists of all monotone functions that also satisfy the following condition:

$$f(b_1, b_2) - f(a_1, b_2) - f(b_1, a_2) + f(a_1, a_2) \geq 0,$$

for every $0 \leq a_1 \leq b_1 \leq 1$ and $0 \leq a_2 \leq b_2 \leq 1$. This condition can be generalised for $d > 2$ by introducing the quasi-volume.

Definition 4.2 (Quasi-volume)**.** *Let $f : [0,1]^d \to \mathbb{R}$ and let $\mathbf{a}, \mathbf{b} \in [0,1]^d$ satisfy $\mathbf{a} \preceq \mathbf{b}$ (i.e. $a_i \leq b_i$ for every $i = 1, \ldots, d$). The quasi-volume of f over the rectangle $[\mathbf{a}, \mathbf{b}]$ is*

$$\Delta\big(f;\, [\mathbf{a}, \mathbf{b}]\big) \;:=\; \sum_{j_1=0}^{J_1} \cdots \sum_{j_d=0}^{J_d} (-1)^{j_1+\cdots+j_d}\, f\big(b_1 + j_1(a_1 - b_1), \ldots, b_d + j_d(a_d - b_d)\big), \qquad J_i := \mathbb{I}\{a_i \neq b_i\}.$$

Remark 4.1. *The quasi-volume operator is a linear operator: for all $\alpha, \beta \in \mathbb{R}$ and functions $f, g : [0,1]^d \to \mathbb{R}$,*

$$\Delta\big(\alpha f + \beta g;\, [\mathbf{a}, \mathbf{b}]\big) = \alpha\, \Delta\big(f;\, [\mathbf{a}, \mathbf{b}]\big) + \beta\, \Delta\big(g;\, [\mathbf{a}, \mathbf{b}]\big).$$

Definition 4.3 (Entirely Monotone Function)**.** *A function $f : [0,1]^d \to \mathbb{R}$ is called entirely monotone if every quasi-volume $\Delta(f; [\mathbf{a}, \mathbf{b}])$ is non-negative:*

$$\Delta(f; [\mathbf{a}, \mathbf{b}]) \geq 0, \quad \text{for every } \mathbf{a} \neq \mathbf{b} \in [0,1]^d \text{ with } \mathbf{a} \preceq \mathbf{b}. \tag{4.3}$$

Now that we have defined both the set of monotone functions and the set of entirely monotone functions, we establish the relationship between the two sets using the following lemma.

Lemma 4.1. *For $d \geq 2$, we have $\mathcal{F}_{EM}^d \subsetneq \mathcal{F}_M^d$.*

Proof. Let $\mathbf{a}, \mathbf{b} \in [0,1]^d$ and $f \in \mathcal{F}_{\mathrm{EM}}^d$. By the definition of entirely monotone functions, for $\mathbf{a} \preceq \mathbf{b}$, we have:

$$\Delta(f; [\mathbf{a}, \mathbf{b}]) \geq 0, \quad \text{for all } \mathbf{a} \neq \mathbf{b}.$$

Taking $\mathbf{a}$ and $\mathbf{b}$ such that $a_i = b_i$ for all $i \neq k$ and $a_k \neq b_k$, we note that for all $i \neq k$,

$$b_i + j_i(a_i - b_i) = b_i.$$

Since only $a_k \neq b_k$, the summation over all j_i for $i \neq k$ becomes trivial, reducing the quasi-volume expression to a one-dimensional sum over j_k:

$$\begin{aligned} \Delta(f; [\mathbf{a}, \mathbf{b}]) &= \sum_{j_k=0}^{1} (-1)^{j_k} f(b_1, \ldots, b_{k-1}, b_k + j_k(a_k - b_k), b_{k+1}, \ldots, b_d) && (4.4) \\ &= f(b_1, \ldots, b_k, \ldots, b_d) - f(b_1, \ldots, a_k, \ldots, b_d) \geq 0. && (4.5) \end{aligned}$$

Then, for all $k \in \{1, \ldots, d\}$, we have

$$f(b_1, \ldots, b_k, \ldots, b_d) \geq f(b_1, \ldots, a_k, \ldots, b_d).$$

For $\mathbf{a} \preceq \mathbf{b}$, we can apply the equation repeatedly to obtain:

$$f(b_1, \ldots, b_d) \geq f(b_1, \ldots, a_d) \geq \cdots \geq f(b_1, a_2, \ldots, a_d) \geq f(a_1, a_2, \ldots, a_d).$$

Thus, we conclude that $f \in \mathcal{F}_{\mathrm{M}}^d$.

To show that the two sets are not equal, consider the function $f : [0,1]^d \to \mathbb{R}$ defined by:

$$f(\mathbf{u}) := \begin{cases} 0 & \text{if } \max\{u_1, u_2\} < \frac{1}{2}, \\ 3 & \text{if } \min\{u_1, u_2\} \geq \frac{1}{2}, \\ 2 & \text{otherwise.} \end{cases}$$

Note that f is constant in all components except the first two. It is clear that $f \in \mathcal{F}_M^d$. However, for

$$\mathbf{a} = \left(\frac{1}{4}, \frac{1}{4}, 0, \ldots, 0\right), \quad \mathbf{b} = \left(\frac{3}{4}, \frac{3}{4}, 0, \ldots, 0\right),$$

we compute:

$$\Delta(f; [\mathbf{a}, \mathbf{b}]) = 3 - 2 - 2 + 0 = -1 < 0.$$

Since $\Delta(f; [\mathbf{a}, \mathbf{b}]) < 0$, we conclude that $f \notin \mathcal{F}_{\mathrm{EM}}^d$, proving the strict inclusion. □

To conclude, we have defined monotone and entirely monotone functions and highlighted the difference between them. In the following section, we will show how to compute the estimator.

4.2 Solving LSE problem using NNLS

The estimator (4.2) can be computed by solving a non-negative least squares (NNLS) problem. Given a suitable design matrix $\mathbf{A}$, the goal of the NNLS problem is to find

$$\widehat{\boldsymbol{\beta}}_{EM} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^p : \beta_j \geq 0, \forall j \geq 2}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{A}\boldsymbol{\beta}\|_2^2, \tag{4.6}$$

where $\mathbf{y}$ is the $n \times 1$ vector consisting of the output variables $y_1, \ldots, y_n$ from equation (4.1).

In this project, we consider the special setting where the observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ form an equally spaced lattice design. That is, given the positive integers $n_1, \ldots, n_d$ with $n = n_1 \cdots n_d$, $\mathbf{x}_1, \ldots, \mathbf{x}_n$ form an enumeration of the points in

$$\mathbb{L}_{n_1, \ldots, n_d} := \left\{ \left(\frac{i_1}{n_1}, \ldots, \frac{i_d}{n_d} \right) : 0 \leq i_j \leq n_j - 1, j = 1, \ldots, d \right\}. \tag{4.7}$$

We construct a suitable design matrix $\mathbf{A}$, where the (i,j)-th entry of $\mathbf{A}$ is given by:

$$\mathbf{A}(i,j) = \mathbb{I}_{[\mathbf{x}_j, \mathbf{1}]}(\mathbf{x}_i) = \mathbb{I}\{\mathbf{x}_j \preceq \mathbf{x}_i\}.$$

Therefore, in the lattice design, the NNLS problem in equation 4.6 can also be written as

$$\hat{\boldsymbol{\beta}}_{\mathrm{EM}} = \underset{\beta \in \mathbb{R}^p : \beta_j \geq 0, \, \forall j \geq 2}{\arg\min} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{n} \mathbb{I}\{\mathbf{x}_j \preceq \mathbf{x}_i\} \beta_j \right)^2$$

In the lattice design, the matrix $\mathbf{A}$ has dimensions $n \times n$, where n is the number of observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Each column of $\mathbf{A}$ corresponds to a vector $\mathbf{x}_j$, which shows whether each observation $\mathbf{x}_i$ satisfies the inequality $\mathbf{x}_j \preceq \mathbf{x}_i$. An example of the design matrix $\mathbf{A}$ is given in Appendix A.2.

The following theorem shows that the design matrix $\mathbf{A}$ is square and invertible in the lattice design.

Theorem 4.1. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in [0,1]^d$ be the distinct points of a lattice design, relabeled so that*

$$\mathbf{x}_1 \preceq \mathbf{x}_2 \preceq \cdots \preceq \mathbf{x}_n.$$

Define the $n \times n$ matrix $\mathbf{A}$ by

$$\mathbf{A}_{ij} \;=\; \mathbb{I}\{\mathbf{x}_j \preceq \mathbf{x}_i\}, \quad i,j = 1, \ldots, n.$$

Then $\mathbf{A}$ is invertible.

Proof. Under the given ordering, if $j > i$ then $\mathbf{x}_j \not\preceq \mathbf{x}_i$, so

$$A_{ij} = \mathbb{I}\{\mathbf{x}_j \preceq \mathbf{x}_i\} = 0.$$

Moreover, for each i,

$$A_{ii} = \mathbb{I}\{\mathbf{x}_i \preceq \mathbf{x}_i\} = 1.$$

Hence $\mathbf{A}$ is lower-triangular matrix with all diagonal entries equal to 1. It follows that

$$\det(\mathbf{A}) = \prod_{i=1}^{n} A_{ii} = 1,$$

and therefore $\mathbf{A}$ is invertible. □

Solution of the optimisation problem

The solution $\widehat{\boldsymbol{\beta}}_{EM}$ of the NNLS problem in equation (4.6) is not necessarily unique but the projection $\mathbf{A}\widehat{\boldsymbol{\beta}}_{EM}$ of the observation $\mathbf{y}$ onto the closed convex cone is unique. This paragraph shows how to obtain a solution given any $\widehat{\boldsymbol{\beta}}_{EM}$.

Proposition 4.1. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in [0,1]^d$ be the points of a lattice design, and define*

$$A_{ij} \;=\; \mathbb{I}\{\mathbf{x}_j \preceq \mathbf{x}_i\}, \quad i,j = 1, \ldots, n.$$

Then

$$\{\mathbf{A}\boldsymbol{\beta} : \beta_j \geq 0, \forall j \geq 2\} = \left\{(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)) : f \in \mathcal{F}^d_{EM}\right\}.$$

Proof. **(i) $\subseteq$:** Take $\beta_j \geq 0$ and define

$$f(\mathbf{z}) = \sum_{j=1}^{n} \beta_j \, \mathbb{I}\{\mathbf{x}_j \preceq \mathbf{z}\}, \quad \mathbf{z} \in [0,1]^d.$$

Then for each design point,

$$f(\mathbf{x}_i) = \sum_{j=1}^{n} \beta_j \, \mathbb{I}\{\mathbf{x}_j \preceq \mathbf{x}_i\} = (\mathbf{A}\boldsymbol{\beta})_i.$$

By linearity of quasi-volume, we have

$$\Delta\big(f; [\mathbf{a}, \mathbf{b}]\big) = \sum_{j=1}^{n} \beta_j \, \Delta\big(\mathbb{I}\{\mathbf{x}_j \preceq \cdot\}; [\mathbf{a}, \mathbf{b}]\big) = \sum_{j=1}^{n} \beta_j \, \mathbb{I}\{\mathbf{a} \prec \mathbf{x}_j \preceq \mathbf{b}\} \;\geq\; 0.$$

Therefore, $f \in \mathcal{F}^d_{\mathrm{EM}}$,

(ii) $\supseteq$: Conversely, if $f \in \mathcal{F}^d_{\mathrm{EM}}$, then

$$f(\mathbf{x}_i) = \sum_{j=1}^{n} \mathbb{I}\{\mathbf{x}_j \preceq \mathbf{x}_i\} \, \beta_j, \quad i = 1, \ldots, n,$$

has the unique solution $\boldsymbol{\beta} = \mathbf{A}^{-1}(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$. Entire monotonicity of f implies $\Delta(\mathbb{I}\{\mathbf{x}_j \preceq \cdot\}; [\ldots]) \geq 0$, which gives $\beta_j \geq 0$. □

Combining Proposition 4.1 with the fact that $\widehat{\boldsymbol{\beta}}_{EM}$ minimizes $\|\mathbf{y} - \mathbf{A}\beta\|^2$ over $\beta \geq 0$, we conclude

$$\big(\hat{f}_{EM}(\mathbf{x}_1), \ldots, \hat{f}_{EM}(\mathbf{x}_n)\big) = \mathbf{A}\,\widehat{\boldsymbol{\beta}}_{EM},$$

and hence the fitted entirely monotone function has the following form

$$\hat{f}_{EM}(\mathbf{z}) = \sum_{j=1}^{n} (\widehat{\beta}_{EM})_j \, \mathbb{I}\{\mathbf{x}_j \preceq \mathbf{z}\}, \quad \mathbf{z} \in [0,1]^d.$$

4.3 Risk results

In this section, we study risk behaviour under the standard squared loss function. The risk of an estimator $\hat{f}$ is given by:

$$\mathcal{R}(\hat{f}, f^*) := \mathbb{E}\left[L(\hat{f}, f^*)\right],$$

where the loss function $L(f, f^*)$ is:

$$L(\hat{f}, f^*) := \frac{1}{n}\sum_{i=1}^{n}\Big(\hat{f}(x_i) - f^*(x_i)\Big)^2.$$

The following theorem proves that the risk of the estimator $\hat{f}_{EM}$ is bounded in the setting where the observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ form an equally spaced lattice design. Since our goal is to predict probabilities, we additionally assume that the true regression function satisfies

$$f^*(\mathbf{x}), \hat{f}_{EM}(\mathbf{x}) \in [0,1] \quad \text{for all } \mathbf{x} \in [0,1]^d.$$

Theorem 4.2. *Let $f^* \in \mathcal{F}^d_{EM}$ and $\hat{f}_{EM}, f^* \in [0,1]$. For the lattice design (4.7), the estimator $\hat{f}_{EM}$ satisfies*

$$\mathcal{R}(\hat{f}_{EM}, f^*) \leq C_d n^{-2/3}\{\log(2+\sqrt{n})^{\frac{2d-1}{3}} + \log(e\sqrt{n}))^{\frac{2d-1}{2}}\}. \tag{4.8}$$

where C_d is a constant that depends only on the dimension d.

4.3.1 Proof sketch

The proof of the main result (Theorem 4.2) is quite complex, so this section gives a brief sketch and explanation of the proof. The proof of Theorem 4.2 answers the question: How far is our estimator from the true value?

First, we put all fitted values $\hat{f}_{\mathrm{EM}}(\mathbf{x}_i)$ into $\hat{\boldsymbol{\theta}}$ and true values $f^*(\mathbf{x}_i)$ into $\boldsymbol{\theta}^*$. Note that $\hat{\boldsymbol{\theta}} := \Pi_{\mathcal{K}}(\mathbf{y})$ is the projection of the noisy data $\boldsymbol{y} = \boldsymbol{\theta}^* + \boldsymbol{\xi}$ onto the closed convex cone

$$\mathcal{D} := \{(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)) : f \in \mathcal{F}^d_{\mathrm{EM}}\}.$$

Then, we analyse the risk

$$\mathcal{R}(\hat{f}_{\mathrm{EM}}, f^*) = \mathbb{E}\frac{1}{n}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2,$$

where $\boldsymbol{\theta}^*$ is the vector of true function values.

In Theorem 4.3 (which is proven in Section 4.3.3) we have stated that if you can find a radius t_*, so that the expected maximum of all Gaussian inner products within that radius is at most $t_*^2/2$, then the expected risk is of order t_*^2. In order to apply the theorem, we want to find a bound for

$$\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\mathcal{K}:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\leq t_*} \langle \boldsymbol{\xi}, \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle\right].$$

To apply this result, we define the function

$$G(t) := \mathbb{E} \sup_{\boldsymbol{\theta} \in \mathcal{D}_{n_1,\dots,n_d} \cap \mathcal{B}_2(\boldsymbol{\theta}^*, t)} \langle \boldsymbol{\xi}, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle,$$

and derive an upper bound on $G(t)$ using chaining, entropy integrals and covering numbers.

A key step is to show that $G(t) \le t^2/2$ for all $t \ge t_m$, where t_1 is an function of n, $\log n$, and the dimension d. This allows us to conclude that t_* can be taken as t_m, and thus the risk is bounded by Ct_m^2/n for a universal constant C.

4.3.2 Proof of Theorem 4.2

Let

$$\hat{\boldsymbol{\theta}} := \big(\hat{f}_{\mathrm{EM}}(\mathbf{x}_1), \dots, \hat{f}_{\mathrm{EM}}(\mathbf{x}_n)\big), \quad \text{and} \quad \boldsymbol{\theta}^* := \big(f^*(\mathbf{x}_1), \dots, f^*(\mathbf{x}_n)\big),$$

and note that

$$\mathcal{R}(\hat{f}_{\mathrm{EM}}, f^*) = \mathbb{E}\frac{1}{n}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2,$$

where $\|\cdot\|$ denotes the usual Euclidean norm in $\mathbb{R}^n$.

Observe that

$$\hat{\boldsymbol{\theta}} = \mathbf{A}\hat{\boldsymbol{\beta}}_{\mathrm{EM}}$$

is the projection of the data vector $\mathbf{y}$ onto the closed convex cone

$$\mathcal{D}_{n_1,\dots,n_d} := \{\mathbf{A}\boldsymbol{\beta} : \beta_j \ge 0,\ \forall j \ge 2\} = \{(f(\mathbf{x}_1), \dots, f(\mathbf{x}_n)) : f \in \mathcal{F}^d_{\mathrm{EM}}\}.$$

Note that, under the lattice design (4.7), the set $\mathcal{D}_{n_1,\dots,n_d}$ is completely determined by the values of $n_1, \dots, n_d$. We can therefore apply Theorem 4.3 (proved in Section 4.3.3) to bound the risk $\mathbb{E}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2/n$.

Theorem 4.3. *Let $\mathcal{K}$ be a closed convex set in $\mathbb{R}^n$ and let*

$$\hat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta} \in \mathcal{K}} \|\mathbf{y} - \boldsymbol{\theta}\|^2,$$

where $\mathbf{y} \sim \mathcal{N}_n(\boldsymbol{\theta}^*, \mathbf{I}_n)$ for some $\boldsymbol{\theta}^* \in \mathcal{K}$ (the well-specified case). Then there exists a universal positive constant C such that

$$\mathbb{E}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 \le C \max(t_*^2, 1)$$

for every $t_* > 0$ which satisfies

$$\mathbb{E}\left[\sup_{\boldsymbol{\theta} \in \mathcal{K} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \le t_*} \langle \boldsymbol{\xi}, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle\right] \le \frac{t_*^2}{2},$$

where $\boldsymbol{\xi} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$.

To apply this theorem, we need to find a t_* such that

$$\boxed{G(t) := \mathbb{E} \sup_{\boldsymbol{\theta} \in \mathcal{D}_{n_1,\dots,n_d} \cap \mathcal{B}_2(\boldsymbol{\theta}^*, t)} \langle \boldsymbol{\xi}, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle \le \frac{t_*^2}{2},} \tag{4.9}$$

where $\boldsymbol{\xi} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$ and

$$\mathcal{B}_2(\boldsymbol{\theta}^*, t) := \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| < t\}$$

denotes the Euclidean ball of radius t centered at $\boldsymbol{\theta}^*$.

We use the following chaining result, which bounds $G(t)$ in terms of covering numbers. The covering number $N(\epsilon, \mathcal{D}_{n_1,\dots,n_d} \cap \mathcal{B}_2(\boldsymbol{\theta}^*, t))$ is the minimal number of Euclidean balls of radius ϵ needed to cover the set $\mathcal{D}_{n_1,\dots,n_d} \cap \mathcal{B}_2(\boldsymbol{\theta}^*, t)$.

Theorem 4.4 ([9], Theorem 3.2)**.** *For every $\boldsymbol{\theta}^* \in \mathcal{D}_{n_1,\dots,n_d}$ and $t > 0$,*

$$\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\mathcal{D}_{n_1,\dots,n_d}\cap\mathcal{B}_2(\boldsymbol{\theta}^*,t)} \langle \boldsymbol{\epsilon}, \boldsymbol{\theta}-\boldsymbol{\theta}^*\rangle\right] \le \sigma \inf_{0\le\delta\le t}\left\{12\int_\delta^t \sqrt{\log N(\epsilon, \mathcal{D}_{n_1,\dots,n_d}\cap\mathcal{B}_2(\boldsymbol{\theta}^*,t))}\,d\epsilon + 4\delta\sqrt{n}\right\}.$$

Setting $\delta = 0$, we get

$$\boxed{G(t) \le 12\sigma\int_0^t \sqrt{\log N(\epsilon, \mathcal{D}_{n_1,\dots,n_d}\cap\mathcal{B}_2(\boldsymbol{\theta}^*,t))}\,d\epsilon.} \tag{4.10}$$

To bound the covering numbers, we invoke Lemma 4.2. To apply it, we first establish an interval that contains $\boldsymbol{\theta}$. We assumed that $\boldsymbol{\theta}^* \in [0,1]$ and since $\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\| < t$, we have

$$\boldsymbol{\theta} \in [a,b] := [-t, 1+t].$$

Lemma 4.2 ([14], Lemma 8.4)**.** *For $a < b$, we have*

$$\log N_\epsilon(\epsilon, \mathcal{D}_{n_1,\dots,n_d}\cap[a,b]^n) \le C_d\frac{(b-a)\sqrt{n}}{\epsilon}\left(\log\frac{(b-a)\sqrt{n}}{\epsilon}\right)^{d-\frac{1}{2}}\mathbb{I}\{\epsilon \le (b-a)\sqrt{n}\}.$$

Combining this lemma with (4.10) gives

$$\begin{aligned} G(t) &\le 12\sigma\int_0^t \sqrt{\log N(\epsilon, \mathcal{D}_{n_1,\dots,n_d}\cap[a,b]^n)}\,d\epsilon &(4.11)\\ &\le C_d\int_0^t \sqrt{\frac{(b-a)\sqrt{n}}{\epsilon}\left(\log\frac{(b-a)\sqrt{n}}{\epsilon}\right)^{d-\frac{1}{2}}\mathbb{I}\{\epsilon \le (b-a)\sqrt{n}\}}\,d\epsilon &(4.12)\\ &\le C_d\int_0^t \sqrt{\frac{B}{\epsilon}\left(\log\frac{B}{\epsilon}\right)^{d-\frac{1}{2}}}\,d\epsilon, &(4.13)\end{aligned}$$

where $B = (b-a)\sqrt{n} = (1+2t)\sqrt{n}$. The following lemma helps bound this integral.

Lemma 4.3 ([14], Lemma 8.5)**.** *For every $d \ge 1$, there exists a positive constant C_d such that for every $s \in (0, B]$,*

$$\int_0^s \sqrt{\frac{B}{\epsilon}\left(\log\frac{B}{\epsilon}\right)^{d-\frac{1}{2}}}\,d\epsilon \le C_d\sqrt{sB}\left(\log\frac{eB}{s}\right)^{\frac{2d-1}{4}}.$$

Applying Lemma 4.3 to (4.13) with $s := t$, we get

$$\boxed{G(t) \le C_d\sqrt{t}\sqrt{(1+2t)\sqrt{n}}\left(\log\frac{e(1+2t)\sqrt{n}}{t}\right)^{\frac{2d-1}{4}},}$$

We split into two cases:
Case 1: $2t \le 1$.
In this case $B \le 2\sqrt{n}$ and $\mathbb{I}\{2t \le 1\} \le \mathbb{I}\{t \le \sqrt{n}\}$, so

$$\begin{aligned} G(t) &\le C_d\,\sqrt{t}\,n^{1/4}\Big(\log\frac{e\sqrt{n}}{t}\Big)^{(2d-1)/4}\mathbb{I}\{t \le \sqrt{n}\}\\ &= C_d\,\sqrt{t}\,n^{1/4}\Big(\log_+\frac{e\sqrt{n}}{t}\Big)^{(2d-1)/4}.\end{aligned}$$

Case 2: $2t \ge 1$.

Here $B \le 3t\sqrt{n}$ and $\mathbb{I}\{2t \ge 1\} \le 1$, so

$$G(t) \le C_d\, t\, n^{1/4}\Big(\log(e\sqrt{n})\Big)^{(2d-1)/4}.$$

Hence for all $t > 0$,

$$G(t) \;\le\; \underbrace{C_d\,\sqrt{t}\, n^{1/4}\Big(\log_+ \tfrac{e\sqrt{n}}{t}\Big)^{(2d-1)/4}}_{G_1(t)} \;+\; \underbrace{C_d\, t\, n^{1/4}\Big(\log(e\sqrt{n})\Big)^{(2d-1)/4}}_{G_2(t)}.$$

Define

$$t_1 := \max\{1, (4C_d)^{2/3}\}(\sqrt{n})^{1/3}\left[\max\{1, \log_+(e(\sqrt{n})^{2/3})\}\right]^{\frac{2d-1}{6}}.$$

Since $t_1 \ge (\sqrt{n})^{1/3}$, for all $t \ge t_1$,

$$\begin{aligned}\frac{G_1(t)}{t^2} &= C_d \frac{n^{1/4}}{t^{3/2}}\left(\log_+ \frac{e\sqrt{n}}{t}\right)^{\frac{2d-1}{4}} \\ &\le C_d \frac{n^{1/4}}{t_1^{3/2}}\left(\log_+(e(\sqrt{n})^{2/3})\right)^{\frac{2d-1}{4}} \\ &\le \frac{1}{4}.\end{aligned}$$

Similarly, define

$$t_2 := 4C_d(\sqrt{n})^{1/3}(\log(e\sqrt{n}))^{\frac{2d-1}{4}}.$$

Then, for $t \ge t_2$, we have

$$\begin{aligned}\frac{G_2(t)}{t^2} &= C_d\, \frac{n^{1/4}}{t}\left(\log(e\sqrt{n})\right)^{(2d-1)/4} \\ &\le C_d\, \frac{n^{1/4}}{t_2}\left(\log(e\sqrt{n})\right)^{(2d-1)/4} \\ &= \frac{1}{4}.\end{aligned}$$

Hence,

$$G(t) \le \frac{t^2}{2} \quad \text{for all } t \ge \max\{t_1, t_2\}.$$

By Theorem 4.3, the risk satisfies

$$\begin{aligned}\mathcal{R}(\hat{f}_{\mathrm{EM}}, f^*) = \mathbb{E}\frac{1}{n}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 &\le \frac{t_1^2 + t_2^2}{n} \\ &\le \frac{1}{n}\Big\{(4C_d)^{4/3}(\sqrt{n})^{2/3}\max\{1, \log_+(e(\sqrt{n})^{2/3})\}^{\frac{2d-1}{3}} + (4C_d)^2(\sqrt{n})^{2/3}(\log(e\sqrt{n}))^{\frac{2d-1}{2}}\Big\} \\ &\le C_d n^{-2/3}\log(2+\sqrt{n})^{\frac{2d-1}{3}} + C_d n^{-2/3}(\log(e\sqrt{n}))^{\frac{2d-1}{2}},\end{aligned}$$

where the last inequality follows because $\log(2+x)$ dominates $\log_+(ex^{2/3})$ for sufficiently large x. Thus, we have proven that

$$\boxed{\mathcal{R}(\hat{f}_{\mathrm{EM}}, f^*) \le C_d n^{-2/3}(\log(2+\sqrt{n})^{\frac{2d-1}{3}} + \log(e\sqrt{n}))^{\frac{2d-1}{2}}).}$$

□

4.3.3 Proof of Theorem 4.3

Theorem 4.3, which we use in the proof of the risk result, is originally proven in Chatterjee's paper [8]. In his work, Chatterjee provides a general framework for analysing the least-squares estimator under convex constraints, allowing the true parameters $\theta^* \in \mathbb{R}^d$, including the case where $\theta^* \notin K$.

In the proof of the risk result (Theorem 4.2), we only consider the well-specified assumption where $\theta^* \in \mathcal{K}$. Therefore, we extract a variant of the proof of Chatterjee by only considering the well-specified setting $\theta^* \in \mathcal{K}$. This assumption leads to a simplification: rather than relying on the Gaussian supremum functional and identifying the projection error as the maximiser of a concave function (as Chatterjee does in his proof), we can directly use the inequality that characterises projections onto convex sets. In particular, by taking $\theta = \theta^*$ in the inequality

$$\langle \hat{\theta} - \theta^* - \xi, \theta - \hat{\theta} \rangle \leq 0 \quad \text{for all } \theta \in K,$$

we obtain the bound

$$\|\hat{\theta} - \theta^*\|^2 \leq \langle \xi, \hat{\theta} - \theta^* \rangle,$$

which is one of the main steps in the proof of Theorem 4.3

Proof. Since $\hat{\boldsymbol{\theta}}$ is the projection of $\mathbf{y}$ onto the closed convex set $\mathcal{K}$, it satisfies the variational inequality:

$$\langle \mathbf{y} - \hat{\boldsymbol{\theta}}, \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \rangle \leq 0 \quad \text{for all } \boldsymbol{\theta} \in \mathcal{K}.$$

To see this, for any $\boldsymbol{\theta} \in \mathcal{K}$ and $\lambda \in [0, 1]$, define the convex combination

$$\boldsymbol{\theta}_\lambda := \lambda \boldsymbol{\theta} + (1 - \lambda)\hat{\boldsymbol{\theta}} \in \mathcal{K}.$$

By definition of projection,

$$\|\mathbf{y} - \hat{\boldsymbol{\theta}}\|^2 \leq \|\mathbf{y} - \boldsymbol{\theta}_\lambda\|^2.$$

Expanding the right-hand side,

$$\|\mathbf{y} - \boldsymbol{\theta}_\lambda\|^2 = \|\lambda(\mathbf{y} - \boldsymbol{\theta}) + (1 - \lambda)(\mathbf{y} - \hat{\boldsymbol{\theta}})\|^2 = \lambda^2 \|\mathbf{y} - \boldsymbol{\theta}\|^2 + (1 - \lambda)^2 \|\mathbf{y} - \hat{\boldsymbol{\theta}}\|^2 + 2\lambda(1 - \lambda)\langle \mathbf{y} - \boldsymbol{\theta}, \mathbf{y} - \hat{\boldsymbol{\theta}} \rangle.$$

Rearranging terms,

$$\|\mathbf{y} - \hat{\boldsymbol{\theta}}\|^2 - (1 - \lambda)^2 \|\mathbf{y} - \hat{\boldsymbol{\theta}}\|^2 \leq \lambda^2 \|\mathbf{y} - \boldsymbol{\theta}\|^2 + 2\lambda(1 - \lambda)\langle \mathbf{y} - \boldsymbol{\theta}, \mathbf{y} - \hat{\boldsymbol{\theta}} \rangle.$$

Dividing both sides by $\lambda > 0$ and letting $\lambda \to 0^+$ yields

$$2\|\mathbf{y} - \hat{\boldsymbol{\theta}}\|^2 \leq 2\langle \mathbf{y} - \boldsymbol{\theta}, \mathbf{y} - \hat{\boldsymbol{\theta}} \rangle,$$

or equivalently,

$$\|\mathbf{y} - \hat{\boldsymbol{\theta}}\|^2 \leq \langle \mathbf{y} - \boldsymbol{\theta}, \mathbf{y} - \hat{\boldsymbol{\theta}} \rangle.$$

Expanding the right side gives

$$\langle \mathbf{y} - \boldsymbol{\theta}, \mathbf{y} - \hat{\boldsymbol{\theta}} \rangle = \|\mathbf{y} - \hat{\boldsymbol{\theta}}\|^2 + \langle \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}, \mathbf{y} - \hat{\boldsymbol{\theta}} \rangle,$$

which leads to

$$0 \leq \langle \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}, \mathbf{y} - \hat{\boldsymbol{\theta}} \rangle = -\langle \mathbf{y} - \hat{\boldsymbol{\theta}}, \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \rangle,$$

and hence

$$\boxed{\langle \mathbf{y} - \hat{\boldsymbol{\theta}}, \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \rangle \leq 0, \quad \forall \boldsymbol{\theta} \in \mathcal{K}.}$$

Substitute $\mathbf{y} = \boldsymbol{\theta}^* + \boldsymbol{\xi}$, where $\boldsymbol{\xi} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$, to get

$$\langle \boldsymbol{\theta}^* + \boldsymbol{\xi} - \hat{\boldsymbol{\theta}}, \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \rangle \leq 0, \quad \forall \boldsymbol{\theta} \in \mathcal{K}.$$

Taking $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ yields

$$\langle \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* - \boldsymbol{\xi}, \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \rangle \leq 0,$$

which rearranges to

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 \leq \langle \boldsymbol{\xi}, \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \rangle.$$

Now consider any $t_* > 0$. Conditioning on the event $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \leq t_*$, we have

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 \leq \sup_{\boldsymbol{\theta} \in \mathcal{K}: \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq t_*} \langle \boldsymbol{\xi}, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle.$$

Taking expectations gives

$$\boxed{\mathbb{E}\left[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 \mathbf{1}_{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \leq t_*}\right] \leq \mathbb{E} \sup_{\boldsymbol{\theta} \in \mathcal{K}: \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq t_*} \langle \boldsymbol{\xi}, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle \leq \frac{t_*^2}{2}.}$$

Next, consider the complement event $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| > t_*$. Consider the following theorem.

Theorem 4.5 ([6], (Theorem 5.6)). *Let $X = (X_1, \ldots, X_n)$ be a vector of n independent standard normal random variables, and let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-Lipschitz function with respect to the Euclidean norm. Then, for every $t > 0$,*

$$\mathbb{P}\{\, f(X) - \mathbb{E}f(X) \geq t \,\} \;\leq\; \exp\Big(-\tfrac{t^2}{2L^2}\Big).$$

Let

$$\Pi_{\mathcal{K}}(\mathbf{y}) := \arg\min_{\boldsymbol{\theta} \in \mathcal{K}} \|\mathbf{y} - \boldsymbol{\theta}\|_2$$

be the metric projection onto the closed, convex set $\mathcal{K}$. Convexity implies that

$$\|\Pi_{\mathcal{K}}(\mathbf{x}) - \Pi_{\mathcal{K}}(\mathbf{y})\|_2 \leq \|\mathbf{x} - \mathbf{y}\|_2 \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

To verify this, we compare the projections of a fixed point $\boldsymbol{\theta}^* \in \mathcal{K}$ and an arbitrary vector $\mathbf{y}$:

$$\begin{aligned}
\|\Pi_{\mathcal{K}}(\boldsymbol{\theta}^*) - \Pi_{\mathcal{K}}(\mathbf{y})\|_2^2 &= \big\langle \Pi_{\mathcal{K}}(\boldsymbol{\theta}^*) - \Pi_{\mathcal{K}}(\mathbf{y}), \Pi_{\mathcal{K}}(\boldsymbol{\theta}^*) - \Pi_{\mathcal{K}}(\mathbf{y}) \big\rangle \\
&= \big\langle \Pi_{\mathcal{K}}(\boldsymbol{\theta}^*), \Pi_{\mathcal{K}}(\boldsymbol{\theta}^*) - \Pi_{\mathcal{K}}(\mathbf{y}) \big\rangle - \big\langle \Pi_{\mathcal{K}}(\mathbf{y}), \Pi_{\mathcal{K}}(\boldsymbol{\theta}^*) - \Pi_{\mathcal{K}}(\mathbf{y}) \big\rangle \\
&\leq \big\langle \boldsymbol{\theta}^*, \Pi_{\mathcal{K}}(\boldsymbol{\theta}^*) - \Pi_{\mathcal{K}}(\mathbf{y}) \big\rangle - \big\langle \mathbf{y}, \Pi_{\mathcal{K}}(\boldsymbol{\theta}^*) - \Pi_{\mathcal{K}}(\mathbf{y}) \big\rangle \\
&= \big\langle \boldsymbol{\theta}^* - \mathbf{y}, \Pi_{\mathcal{K}}(\boldsymbol{\theta}^*) - \Pi_{\mathcal{K}}(\mathbf{y}) \big\rangle \\
&\leq \|\boldsymbol{\theta}^* - \mathbf{y}\|_2 \, \|\Pi_{\mathcal{K}}(\boldsymbol{\theta}^*) - \Pi_{\mathcal{K}}(\mathbf{y})\|_2.
\end{aligned}$$

Define the Lipschitz function

$$f(\mathbf{y}) := \big\|\Pi_{\mathcal{K}}(\mathbf{y}) - \boldsymbol{\theta}^*\big\|_2.$$

The Gaussian concentration inequality for Lipschitz functions then yields absolute constants $c, C > 0$ such that for all $u \geq 0$,

$$\mathbb{P}\big(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 > t_* + u\big) \leq C \, \exp\big(-c\, u^2\big).$$

Rewriting with $x = t_* + u$ gives the tail bound

$$\boxed{\mathbb{P}\big(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 > x\big) \leq C \, \exp\big(-c\,(x - t_*)^2\big), \qquad x \geq t_*},$$

so the distance $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$ is sub-Gaussian around its mean t_*.

Using this tail bound, estimate

$$\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|^2\mathbf{1}_{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|>t_*}\right]=\int_{t_*^2}^{\infty}\mathbb{P}\left(\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|^2>t\right)dt=\int_{t_*^2}^{\infty}\mathbb{P}\left(\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|>\sqrt{t}\right)dt.$$

Change variables with $t=x^2$, $dt=2xdx$, to write

$$\begin{aligned}\int_{t_*^2}^{\infty}\mathbb{P}\left(\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|>\sqrt{t}\right)dt&=\int_{t_*}^{\infty}2x\,\mathbb{P}\left(\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|>x\right)dx\\&\le\int_{t_*}^{\infty}2x\,C\exp\left(-c(x-t_*)^2\right)dx.\end{aligned}$$

Substitute $u=x-t_*$, so $x=u+t_*$, and

$$\int_{t_*}^{\infty}2xe^{-c(x-t_*)^2}dx=\int_0^{\infty}2(u+t_*)e^{-cu^2}du=\int_0^{\infty}2ue^{-cu^2}du+\int_0^{\infty}2t_*e^{-cu^2}du.$$

Both integrals are finite:

$$\int_0^{\infty}2ue^{-cu^2}du=\frac{1}{c},\quad\text{and}\quad\int_0^{\infty}2t_*e^{-cu^2}du=t_*\sqrt{\frac{\pi}{c}}.$$

Including the constant C, we get

$$\int_{t_*}^{\infty}2xCe^{-c(x-t_*)^2}dx\le C\left(\frac{1}{c}+t_*\sqrt{\frac{\pi}{c}}\right).$$

Therefore, there exists a constant $C'>0$ such that

$$\boxed{\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|^2\mathbf{1}_{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|>t_*}\right]\le C'\max(t_*^2,1).}$$

Combining both parts,

$$\mathbb{E}\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|^2=\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|^2\mathbf{1}_{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|\le t_*}\right]+\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|^2\mathbf{1}_{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|>t_*}\right]\le C\max(t_*^2,1),$$

for some universal constant $C>0$. Thus, we have proven

$$\boxed{\mathbb{E}\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|^2\le C\max(t_*^2,1).}$$

□

4.4 Feature Selection for Multivariate Isotonic Regression

This section discusses the process of selecting risk drivers for the isotonic regression model. Feature selection is an essential step to prevent overfitting and reduce model complexity. In the method outlined in Section 4.2, the total number of observations is $n=\prod_{j=1}^{d}n_j$, where the number of features is denoted by d and each feature has n_j grid points. Therefore, as the number of features increases, the matrix becomes larger, leading to higher computational costs.

Before the approval of the new dataset B, a preliminary feature selection method was applied to assess the existing modelling approach. This was necessary because the current procedure is time-consuming, particularly when applied to large datasets. As mentioned in the introduction, experts have concerns about whether the current set of risk drivers is complete, and whether additional relevant risk drivers possibly absent in both datasets A and B could improve the model's performance within the acceptance framework.

Evaluating the usefulness of new risk drivers in the current model is a time-intensive task, especially when working with large datasets. Therefore, there is a need for a more efficient method to evaluate and refine the set of risk drivers.

Research suggests that combining individually strong features does not always result in optimal performance [10, 17]. This is often due to the overlap in information provided by highly informative features, leading to redundancy. Redundant variables can unnecessarily increase model complexity without improving predictive performance. To address this issue and reduce computational time, we propose the use of the mRMR algorithm, introduced by Ding in 2003 [12], which enhances model performance by selecting features that are not only individually relevant but also minimally redundant, ensuring a more efficient and interpretable model. Since mRMR evaluates relevance and redundancy using simple mutual information criteria, it significantly reduces computational time compared to other search methods [22].

4.4.1 mRMR-algorithm

The **mRMR** (Minimum Redundancy Maximum Relevance) algorithm selects an optimal subset of features by balancing two key aspects: **relevance** and **redundancy**. Relevance is measured using mutual information, a metric that quantifies the dependency between each feature and the target variable. Redundancy, on the other hand, evaluates the overlap of information among the features. Both relevance and redundancy are defined and calculated using the formulas provided by [22].

Algorithm 1 outlines the mRMR algorithm.

Algorithm 1 mRMR Algorithm

Input: Feature set $\mathbf{X} = (X_1, X_2, \ldots, X_p)$, target variable $\mathbf{y}$, and number of features to select k.
Output: Selected subset of features $\hat{\mathbf{X}}$ with $|\hat{\mathbf{X}}| = p$.
Initialize $\hat{\mathbf{X}} \leftarrow \emptyset$ (empty selected feature set).
Step 1: Compute Relevance
for each $X_j \in \mathbf{X}$ **do**
    Compute Relevance$[X_j] = I(X_j; \mathbf{y})$, the mutual information between X_j and $\mathbf{y}$.
end for
Step 2: Select the First Feature
Select $X_{\text{best}} = \arg\max_{X_j \in \mathbf{X}} \text{Relevance}[X_j]$.
Update $\hat{\mathbf{X}} \leftarrow \hat{\mathbf{X}} \cup \{X_{\text{best}}\}$ and $\mathbf{X} \leftarrow \mathbf{X} \setminus \{X_{\text{best}}\}$.
Step 3: Iterative Selection of Features
while $|\hat{\mathbf{X}}| < p$ **do**
    for each $X_j \in \mathbf{X}$ **do**
        Compute Redundancy$[X_j] = \frac{1}{|\hat{\mathbf{X}}|^2} \sum_{X_k \in \hat{\mathbf{X}}} I(X_j; X_k)$,
        Compute mRMR$[X_j]$ = Relevance$[X_j]$ − Redundancy$[X_j]$.
    end for
    Select $X_{\text{best}} = \arg\max_{X_j \in \mathbf{X}} \text{mRMR}[X_j]$.
    Update $\hat{\mathbf{X}} \leftarrow \hat{\mathbf{X}} \cup \{X_{\text{best}}\}$ and $\mathbf{X} \leftarrow \mathbf{X} \setminus \{X_{\text{best}}\}$.
end while
Return $\hat{\mathbf{X}}$.

It begins by initializing an empty set $\hat{\mathbf{X}}$ to hold the selected features. Let x_{ij} denote the i-th observation and the j-th feature. First, the relevance of each feature $X_j = (x_{1j}, \ldots, x_{nj})^T$ in $\mathbf{X}$ to the target variable y is computed. The relevance of feature X_j is given by Relevance$[X_j] = I(X_j, \mathbf{y})$, where $I(X_j, \mathbf{y})$ is the mutual information, defined as:

$$I(X_j, y) = \sum_{i=1}^{n} \sum_{y=1}^{n} p(x_{ij}, y) \log \frac{p(x_{ij}, y)}{p(x_{ij})p(y)},$$

where p denotes the probability function.

At each iteration, for each remaining feature in $\mathbf{X}$, its redundancy with the features already in $\hat{\mathbf{X}}$ is calculated. This redundancy is computed as the average mutual information between the feature and each feature in the selected set. The mRMR score for each feature is then determined by subtracting its redundancy from its relevance. The feature with the highest mRMR score is chosen and added to $\hat{\mathbf{X}}$, while being removed from $\mathbf{X}$.

The process repeats until the selected subset $\hat{\mathbf{X}}$ contains p features. The list produced by the mRMR algorithm, shown in Table 4.1, reflects the trade-off between maximising relevance and minimising redundancy. Features that appear higher on the list contribute both unique and strong predictive value, while lower-ranked features may still be informative but are less critical, as their predictive power overlaps with features already selected. The scores reflect the relative importance of each feature in predicting the target variable.

A higher positive score indicates greater relevance for the prediction task. The most influential feature in the dataset is BKR, followed by LTV and PRODUCT, all of which provide strong predictive value.

Table 4.1: Feature Scores Train Data

[REDACTED]

The remaining issue is how to determine the optimal number of features. The optimal feature set will be selected using the method described in [22]. Start by investigating in the computational complexity.

Computational Complexity

The number of features directly impacts the computational time of the model, with the most computationally expensive step being the generation of matrix A. As mentioned earlier, the size of matrix A is determined by the number of features in the model. For the lattice design, the matrix A has dimensions $n \times n$, where $n = n_1 \cdot n_2 \cdot \cdots \cdot n_d$. The function used to generate matrix A has a time complexity of $O(n^2)$.

The computational time required for different values of n is presented in Table 4.2. As shown, the computational time grows as n increases, and the relationship follows the expected quadratic complexity.

For this project, the acceptable computational time is set to be +-180 seconds, which means $n <= 1620$. This corresponds to selecting the first four risk drivers. Start with the first four features. Then, the first feature set denoted by S_1 contains one feature and the fourth set S_4 contains four features such that $S_1 \subseteq S_2 \subseteq \cdots \subseteq S_4$. Lastly, evaluate the different feature sets

Table 4.2: Computational Time for Various Values of n

n	Computational Time (seconds)
6	0.004
30	0.033
180	1.618
540	17.595
1620	185.113

on the model and choose the one with the highest accuracy. How this can be evaluated will be described in Section 3.5.

4.5 Probability of Default Modelling with Isotonic Regression

In this section, isotonic regression is applied to model the probability of default. The analysis begins with two key risk drivers, providing a visualisation of the relationship between input variables and default risk. Subsequently, additional risk drivers are incrementally introduced to evaluate their impact on model performance. Finally, an appropriate cut-off point is selected, and the model is assessed within the context of a mortgage acceptance framework.

4.5.1 Two risk drivers

We begin by analysing the top two risk drivers identified by the mRMR algorithm: Client BKR Score and LTV Loan to Market Value. The BKR Score is categorised into 5 categories, while LTV is discretised into 5 almost equally sized bins. For every facility, we have the LTV, the BKR and the indicator whether the facility is defaulted within 12 months. We subtract this information from the original dataset. An example of this subset is shown in Table 4.3.

Table 4.3: Sample of Facility Data

Facility	LTV	BKR
1234	0.92	A
5678	5.15	A
91011	0.97	C
1213	4.09	D
1415	3.12	B

Applying isotonic regression to these risk drivers and solving using the aforementioned methods yields the following results. In Figure 4.1 we show the actual default rates and the fitted default rates.

The left heatmap represents the actual default rates across different categories, illustrating the proportion of defaults in various segments of the dataset. Darker shades indicate a higher concentration of defaults, highlighting segments with greater risk. The right heatmap shows the fitted default rates using isotonic regression.

4.5.2 Number of Risk Drivers

Table 4.4 presents the Gini scores for both the training and test data, evaluated across different sets of variables.

Starting with the base variable, BKR, we observe a Gini score of 0.415 for the training data and 0.348 for the test data, indicating limited predictive power. Adding LTV increases the Gini

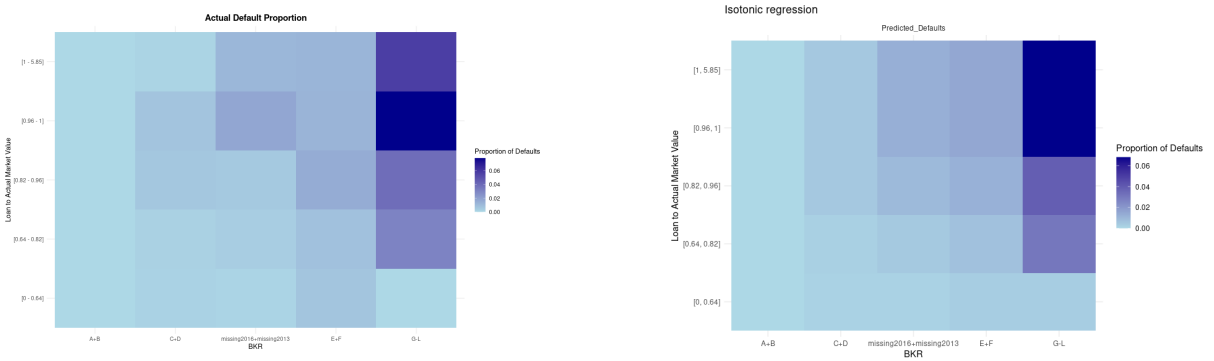

Figure 4.1: Comparison of Actual and Fitted Default Rates

score for the training data to 0.555, suggesting a significant improvement in the model's ability to discriminate between observations. However, the test Gini score of 0.456 remains lower, reflecting a smaller improvement when tested on unseen data.

The inclusion of the PRODUCT variable results in a slight decrease in the training Gini to 0.532 and a corresponding increase in the test Gini to 0.466. This suggests that while PRODUCT adds some value to the model, it does not dramatically improve the model's performance on the test set.

Finally, adding ACTIVE COUNT further improves the training Gini to 0.560, but the test Gini decreases to 0.449. The minor change in the test Gini suggests that ACTIVE COUNT provides some improvement in model discrimination on the training data, but does not offer substantial gains when generalised to unseen data.

In summary, while adding additional variables improves the training Gini score, the test Gini score does not show a corresponding increase, suggesting the potential for overfitting. Among the sets of variables tested, the combination of three variables (BKR, LTV, and PRODUCT) produces the highest Gini score for the test set and the smallest gap between training and test scores. As a result, the set of three variables is selected.

Table 4.4: Gini Scores for Training and Test Data by Number of Variables

Variables	Train Gini	Test Gini
BKR	0.415	0.348
+LTV	0.555	0.456
+PRODUCT	0.532	0.466
+ACTIVE COUNT	0.560	0.449

4.5.3 Cut-off point

The model outputs a default probability for each facility. To convert these probabilities into flags, we select a decision threshold. To determine the optimal cut-off point, we evaluated various performance metrics across a range of threshold values on the training set. Table A.3 presents the performance metrics for the model in different threshold ranges. Performance is evaluated based on several metrics, including accuracy, precision, recall, F1 score, and false alarm rate, along with the components of the confusion matrix: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

At the lowest threshold range, 1e-04 to 1e-04, the model shows low accuracy of 0.036, indicating that the model is biased toward predicting negative outcomes. This is expected since the model has not yet been tuned to identify positive cases effectively. As the threshold increases, the accuracy improves significantly, reaching 0.995 at the highest threshold range, 0.0479 to 0.1,

where the model predicts mostly negative outcomes and achieves the highest accuracy. However, this comes with a substantial decrease in recall, showing a trade-off between accuracy and the ability to capture positive cases.

The default rate is 0.5 per cent, which means that if we set all facilities to non-default, the model would achieve an accuracy of 99.5% . However, this is not a desirable outcome as it indicates that the model is simply predicting the most common outcome (non-defaults) and ignoring potential positives. This would not contribute to the actual goal of the model, which is to correctly identify true defaults (positive cases).

Therefore, to achieve better model performance, we aim to choose a cut-off point where the F1 score is optimised. The F1 score is the harmonic mean of precision and recall and provides a balanced measure between these two metrics. Precision is the proportion of true positive predictions out of all positive predictions made, and recall is the proportion of true positives out of all actual positive cases. The F1 score helps balance the trade-off between precision and recall, especially when both false positives and false negatives carry significant consequences. By optimising the F1 score, we ensure that the model can effectively identify true positive cases (defaults) while minimising false alarms and false negatives, which is especially important in predicting rare events like defaults. Figure 4.2 shows the F1 score of the different thresholds, with an optimum at thresholds equal to 0.0297.

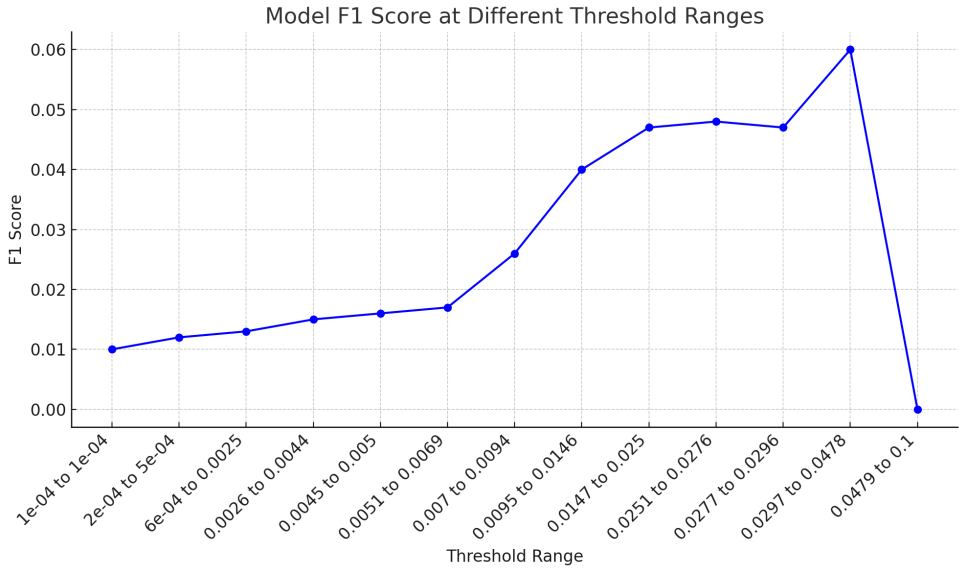

Figure 4.2: Model F1 Score at Different Threshold Ranges

In summary, while the highest accuracy is reached at a cut-off point of 0.0479, the detection rate becomes 0, caused by the low default rate of 0.05%. By optimising the F1 score, we ensure that the model balances precision and recall. The F1 score is optimised for the cut-off point 0.0297. At this point, the accuracy is 99% with a detection rate of 6.5%.

4.5.4 Model Evaluation and Results

The model evaluated on the test set achieved an accuracy of 96.25%, which might seem relatively high. However, this can be misleading, since the precision of the model is relatively low, with a value of 2.6%. This suggests that the model does not effectively identifying positive cases (defaults). Out of all the instances that the model predicted as positive, only a small fraction were true positives. This indicates a significant issue with false positives, where non-defaults are incorrectly predicted as defaults.

Similarly, only 17.8% of the true positive cases were correctly identified by the model. This

low recall further emphasises that the model is biased towards predicting the negative class (non-defaults).

The F1 score, which balances precision and recall, is 4.6%. This is very low, reflecting the poor balance between precision and recall. The model is struggling to effectively identify defaults while maintaining a low rate of false positives.

The false alarm rate is 3.35%, which indicates that 3.35% of the instances predicted as defaults were actually non-defaults. Although not extremely high, this still suggests that the model is making a considerable number of false positive predictions.

Confusion Matrix

The confusion matrix provides further insight into the model's performance:

- True Positives (TP): 5 facilities were correctly identified as defaults.
- True Negatives (TN): 5310 facilities were correctly identified as non-defaults.
- False Positives (FP): 184 facilities were incorrectly predicted as defaults.
- False Negatives (FN): 23 facilities of actual defaults were missed by the model.

Despite a high accuracy, the model's precision and recall indicate that it is not identifying defaults effectively. The low precision suggests that the model is predicting too many non-defaults as defaults, resulting in a high number of false positives. The low recall indicates that the model is missing a large proportion of actual defaults, which is critical in applications like risk assessment for car insurance. The F1-score confirms the imbalance between precision and recall, underscoring the need for model improvements, especially in identifying the minority class (defaults).

These findings suggest that, while the model may appear to perform well at first glance, it needs further refinement. A more effective model should aim to improve recall and precision, potentially by addressing the class imbalance through techniques such as oversampling, undersampling, or by using different classification algorithms that are better suited for imbalanced data.

4.6 Comparative Analysis

In this section, we compare the different feature selection methods and the two models used throughout this project. We will first compare the performance of Forward Stepwise Regression and the mRMR algorithm in combination with the Isotonic Regression model. Following that, we will analyse the performance differences between Logistic Regression and Isotonic Regression.

4.6.1 Forward Stepwise Regression vs mRMR-algorithm

Table 4.5 shows the different feature sets as outcome of the feature selection methods. Features set 1 and 2 consist of features as an outcome of the mRMR algorithm, with target defaults in 12 and 24 months, respectively. Features set 3 and 4 consist of features as outcomes of forward stepwise regression, with target defaults in 12 and 24 months, respectively.

Table 4.5: Feature set 1 and 2 consist of features as outcomes of the mRMR algorithm, with target defaults in 12 and 24 months, respectively. Features set 3 and 4 consist of features as outcomes of forward stepwise regression, with target defaults in 12 and 24 months, respectively.

[REDACTED]

To determine which feature set performs best with the isotonic regression model, we compare the GINI scores of the training and test datasets. The results, shown in Table A.2, reveal that Feature Set 1 achieves the highest GINI score on the test set, with the smallest difference between the training and test GINI scores. This indicates the best performance in terms of generalisation, as it demonstrates a better ability to predict the probability of default on unseen data.

Table 4.6: Gini Scores for Training and Test Data Isotonic Regression

Feature set (number of risk drivers)	Train Gini	Test Gini
F1 (3)	53.2%	46.6%
F2 (4)	53.9%	44.2%
F3 (4)	53.7%	40.1%
F4 (2)	53.2%	43.9%

4.6.2 Logistic vs Multivariate Isotonic Regression

A comparison of values of the GINI score across the test and training data sets is shown in Table 4.7. From the table, both models $LR_{A,12}$ and $LR_{B,12}$ suffer from overfitting. These models perform much better on the training set (64.7% and 65.8%, respectively) than the test set (19.2% and 30.9%, respectively), reflecting that they perform well on the training data but don't generalise to new data well.

On the other hand, the two other models, $LR_{B,24}$ and $IR_{B,12}$, perform more fairly on the training data and test set. Model $LR_{B,24}$ records a comparatively high GINI score of 62.3% for the training data, and it performs equally as well on the test set with a GINI score of 55.8%. This shows that $LR_{B,24}$ generalises very well to new data and still maintains good discriminatory power. Accordingly, the $IR_{B,12}$ model also runs more stably, with a GINI of 53.2% on the training set and 46.6% on the test set.

The results conclude that $LR_{B,24}$ and $IR_{B,12}$ are the most stable models to predict defaults.

The performance results of both the logistic regression models and the isotonic regression model are presented in Table 4.8. The isotonic regression model is denoted as $IR_{B,12}$. Compared to the logistic regression models, $IR_{B,12}$ shows lower accuracy and precision. Specifically, the isotonic regression model has a accuracy of 0.9625 while, the best-performing logistic regression model $LR_{A,12}$ achieves an accuracy of 0.9870. Similarly, the isotonic regression model also has lower precision, with a value of 0.0265 compared to $LR_{B,24}$'s precision of 0.1154.

However, the isotonic regression model achieves the highest recall among all models, with a recall value of 0.1786. This indicates a better ability to correctly identify defaulting facilities

Table 4.7: Model Performance in terms of GINI on Different Datasets.

Model	Training	Test
$LR_{A,12}$	64.7%	19.2%
$LR_{B,12}$	65.8%	30.9%
$LR_{B,24}$	62.3%	55.8%
$IR_{B,12}$	53.2%	46.6%

compared to the logistic regression models, where the recall ranges from 0.0714 to 0.1429. The higher recall of $IR_{B,12}$ means that it is less likely to miss an actual default.

Despite its higher recall, the precision of the isotonic regression model remains relatively low, suggesting a higher amount of false positives, i.e. facilities that were predicted to default but did not. This trade-off between recall and precision is often encountered when the goal is to capture as many defaults as possible, even at the expense of generating more false alarms.

The F1 score, which balances precision and recall, remains similar for both the isotonic regression and logistic regression models, reflecting the trade-off between false positives and false negatives. In this case, the isotonic regression model has an F1 score of 0.045, which is very close to that of the worst-performing logistic regression model, $LR_{B,12}$, with an F1 score of 0.045 as well. The F1 score of all models is below 0.1, indicating a low performance in terms of balancing the precision and recall [13].

Table 4.8: The accuracy, recall, precision and F1 score for three logistic regression models and one isotonic regression model.

Model	Accuracy	Recall	Precision	F1 Score	Flags
Model $LR_{A,12}$	0.9870	0.1429	0.0769	0.100	0.94%
Model $LR_{B,12}$	0.9848	0.0714	0.0333	0.045	1.1%
Model $LR_{B,24}$	0.9823	0.1034	0.1154	0.109	0.94%
Model $IR_{B,12}$	0.9625	0.1786	0.0265	0.045	3.4%

Moreover, the Isotonic Regression model ($IR_{B,12}$) results in the highest number of flagged facilities, with 189 flags, representing 3.4%. Among the other models, $LR_{B,12}$ flags 1.1% of the facilities, while the remaining two models flag 0.94% of the facilities each. In terms of the lowest number of flags, model $LR_{A,12}$ and $LR_{B,24}$ are most desired, where model $LR_{A,12}$ has a higher recall, model $LR_{B,24}$ has a higher precision.

5 | Conclusion

The objective of this thesis was to:

> Enhance the mortgage acceptance model by improving feature selection, identifying a more suitable target variable, and exploring the potential of multivariate isotonic regression.

Models were evaluated on dataset B with targets defaulting within 12 months and defaulting within 24 months. These were compared to the original model trained on dataset A with the 12-month target. Shifting the target from default within 12 months to default within 24 months improved the model's generalisation and results in improved accuracy.

Furthermore, we introduced multivariate isotonic regression as an alternative nonparametric modelling approach. Unlike other nonparametric methods, it does not require tuning parameters and tends to avoid overfitting due to its monotonicity constraint. Theoretical results show that under certain conditions, the risk of the multivariate isotonic regression estimator is bounded by $n^{-2/3}$, implying that its estimation error decreases relatively quickly as the sample size increases.

This theoretical guarantee helps explain the observed stable performance of multivariate isotonic regression compared to logistic regression, particularly in settings with a 12-month default target, where monotonic relationships between risk drivers and the outcome are plausible.

The mRMR (Minimum Redundancy Maximum Relevance) algorithm was used to speed up feature selection, identifying a good feature set in about 20 seconds, much faster than the original approach, which took over 2 minutes.

In terms of performance, the results indicate that logistic regression with a 24-month target offers a balanced trade-off between precision and recall, making it suitable for routine mortgage acceptance decisions, where fewer flagged cases (false positives) are preferable. On the other hand, multivariate isotonic regression showed higher recall, which is beneficial for identifying more defaults, but led to an increased workload due to more false positives. As such, multivariate isotonic regression could be employed in scenarios where higher recall is necessary, such as high-risk cases that warrant more thorough manual review despite the potential for more flagged cases.

To summarise the answers to the research questions:

- **Dataset modification:** Using only the first observation per facility better represents the data available at acceptance.
- **Target variable:** Default within 24 months improves model generalization and reduces overfitting compared to 12 months.
- **Feature selection:** Both forward stepwise regression and mRMR effectively reduce features, with mRMR being faster.
- **Evaluation metric:** The GINI coefficient combined with F1 score and accuracy provides a good balance for model assessment.

- **Model performance:** Logistic regression with the 24-month target performs best in terms of precision and flag management, while multivariate isotonic regression has higher recall but more flagged cases.

Logistic regression requires more preprocessing, such as binning and transformations, to meet model assumptions. multivariate isotonic regression only requires monotonicity, simplifying data preparation.

Choosing a suitable target variable is important. The low default rate within 12 months risks overfitting for logistic regression, which improves by using a 24-month horizon. Multivariate isotonic regression shows less overfitting even with the 12-month target.

In conclusion, both models provide valuable insights for Achmea's mortgage acceptance process. The logistic regression model with a 24-month target offers a good balance between accuracy and manageable flagged cases. The multivariate isotonic regression model offers better recall but at the cost of more flagged applications.

6 | Discussion

This chapter gives the limitations of the results and proposes directions for further research.

6.1 Data

To simulate the mortgage approval process, the model is trained on data of accepted applicants to compute the probability of default. The latter probability score is used to inform applicants of manual checks. However, in practice, the model is applied to both approved and rejected applicants, which could produce a sample selection bias. This is an occurrence known as reject inference because the training information does not include rejected candidates and therefore can limit the model's ability to generalise to the whole candidate population.

The effectiveness of reject inference is debatable in credit risk studies. Past research shows that reject inference will not have a significant impact on model outcomes [11], whereas some show positive effects on prediction performance [5]. The attention for future study should be on analysing the impact of applying rejection inference techniques in the acceptance system of Achmea Bank and on knowing if predictive accuracy can be increased.

6.2 Logistic Regression

The logistic regression models developed within this thesis are overfitting, as evidenced by high variation between training set and test set performance. Overfitting can be attributed to the relatively low default rate and the low number of default observations, which restricts the model from generalising sufficiently.

Furthermore, logistic regression requires heavy preprocessing, i.e., feature binning and transformation. While these transformations improve interpretability and fit, they add complexity to the model.

Future work may explore alternative approaches to beat overfitting, such as regularisation or ensemble methods that maintain interpretability. Further, trying out different target variable specifications, e.g., expanding the default horizon from 12 to 24 months, can improve model stability and performance, as shown here.

6.3 Multivariate Isotonic Regression

Multivariate isotonic regression has been found to be useful in that it is flexible but under monotonicity constraints that prevent overfitting. However, there are some limitations and areas for enhancement.

One of them is computational complexity. The problem of solving the large non-negative least squares is the demand of the method, and this becomes increasingly expensive as it introduces more features. Future work would involve using faster algorithms or approximation methods for scalability.

Another such area of enhancement is handling class imbalance. Because default rates are low, isotonic regression may have problems identifying defaults effectively. Application of techniques such as weighted losses or isotonic regression-specific synthetic sampling would help in such situations.

6.4 Ethical Framework

6.4.1 Transparency and Explainability

In terms of financial regulation, transparency and explainability of models are very important. Logistic regression is easy to interpret from the coefficients since they clearly indicate the effect of each driver of risk on the default probability. Such a transparent explanation aids stakeholders and regulators in understanding and trusting the model.

While multivariate isotonic regression is not as common in financial applications, it remains interpretable due to the fact that it possesses a monotonicity constraint. The monotone assumption among predictors and the response is natural and intuitive.

6.4.2 Bias

Credit risk models can accidentally hold or even enhance biases in the data. For example, the data show that self-employed borrowers default at a higher rate compared to others. This is a pattern in the past data, but not necessarily due to being self-employed leading to greater risk. It could be due to other factors related to self-employment, including income variability or fewer financial documents.

Because of this, special care has to be taken while using such information. Excessive reliance on self-employment status can have a negative effect on some borrowers. To prevent this, the models must undergo constant fairness checks, and steps should be taken to make lending decisions ethical, equitable, and in compliance with the law.

A | Appendix

A.1 Acceptance dataset

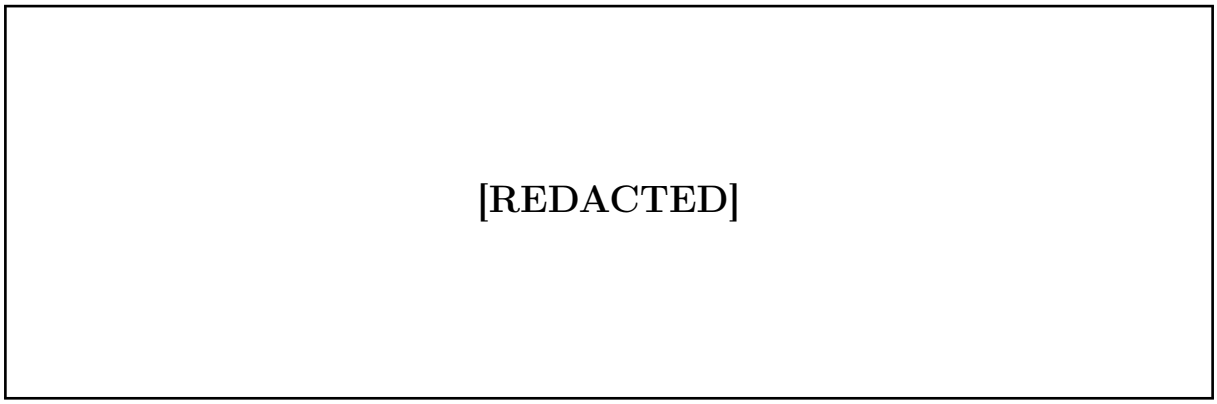

Figure A.1: Comparing the default rates over time of de training set (left) and the test set (right).

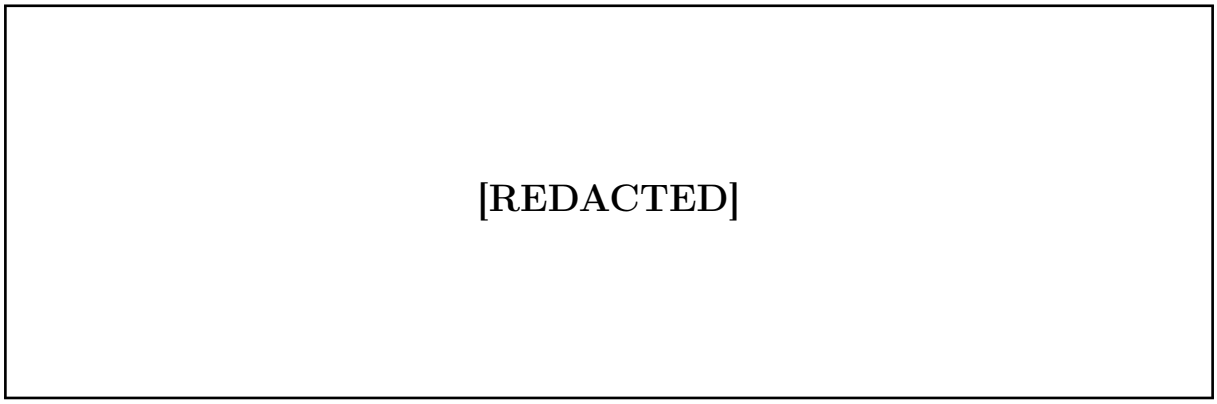

Figure A.2: Comparing the distribution of the BKR scores of de training set (left) and the test set (right).

A.2 Example design matrix A

Suppose we have two risk drivers ($d = 2$) with values ranging between 0 and 1. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_9 \in [0, 1]^2$ denote the nine different observations. These observations lie on an equally spaced 10×10 grid. The design matrix A is defined such that each entry is given by:

$$A(i, j) = \mathbb{I}\{\mathbf{x}_j \preceq \mathbf{x}_i\}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function.

The resulting design matrix A is given by:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

A.3 Performance

[REDACTED]

Table A.1: Performance metrics by number of grouping variables ($IR_{B,12}$).

f_est	Train Gini	Test Gini
f_est_1	0.415	0.348
f_est_2	0.555	0.456
f_est_3	0.532	0.466
f_est_4	0.560	0.449

Table A.2: Gini Scores for Training and Test Data.

Threshold	Acc	Prec	Detec	F1	False Alarm	TP	TN	FP	FN
1e-04 to 1e-04	0.036	0.005	1.000	0.010	0.968	108	680	20828	0
2e-04 to 5e-04	0.157	0.006	0.991	0.012	0.848	107	3276	18232	1
6e-04 to 0.0025	0.257	0.006	0.963	0.013	0.747	104	5443	16065	4
0.0026 to 0.0044	0.401	0.008	0.935	0.015	0.602	101	8562	12946	7
0.0045 to 0.005	0.440	0.008	0.907	0.016	0.562	98	9417	12091	10
0.0051 to 0.0069	0.483	0.009	0.889	0.017	0.519	96	10346	11162	12
0.007 to 0.0094	0.767	0.013	0.611	0.026	0.232	66	16511	4997	42
0.0095 to 0.0146	0.880	0.021	0.500	0.040	0.118	54	18971	2537	54
0.0147 to 0.025	0.905	0.025	0.472	0.047	0.093	51	19504	2004	57
0.0251 to 0.0276	0.912	0.025	0.444	0.048	0.085	48	19671	1837	60
0.0277 to 0.0296	0.914	0.025	0.426	0.047	0.083	46	19714	1794	62
0.0297 to 0.0478	0.990	0.056	0.065	0.060	0.005	7	21390	118	101
0.0479 to 0.1	0.995	0.000	0.000	0.000	0.000	0	21508	0	108

Table A.3: Model Performance Metrics at Different Threshold Ranges for f_est_3.

A.4 Multivariate Results

[REDACTED]

Table A.4: The table presents the multivariate results for various predictors in a statistical model, including the coefficients, standard errors, z-values, and corresponding p-values for each variable.

A.5 F1 Score

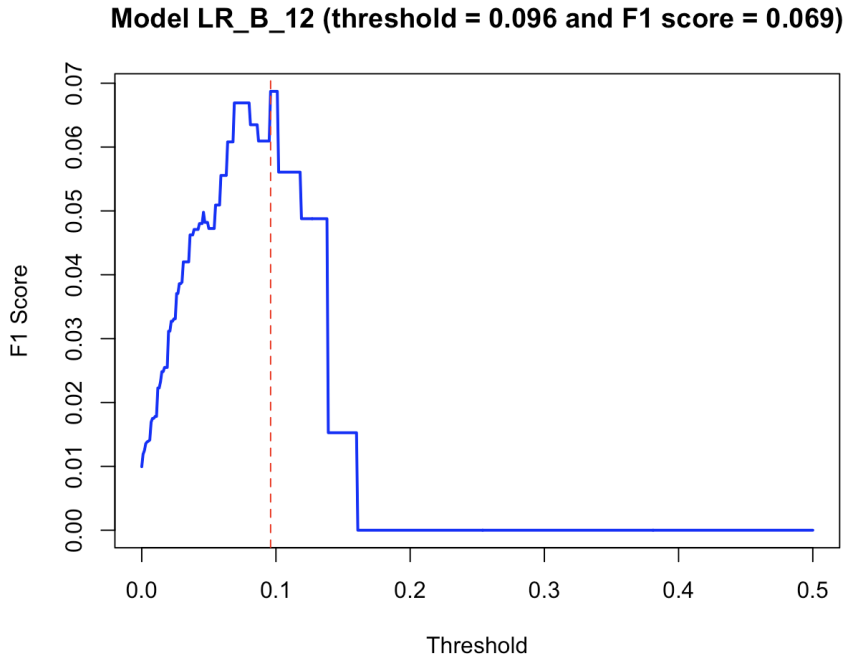

Figure A.3: F1 score as a function of threshold values for Model LR_B_12. The threshold of 0.096 results in an F1 score of 0.069, wich is marked as a vertical dotted red line.

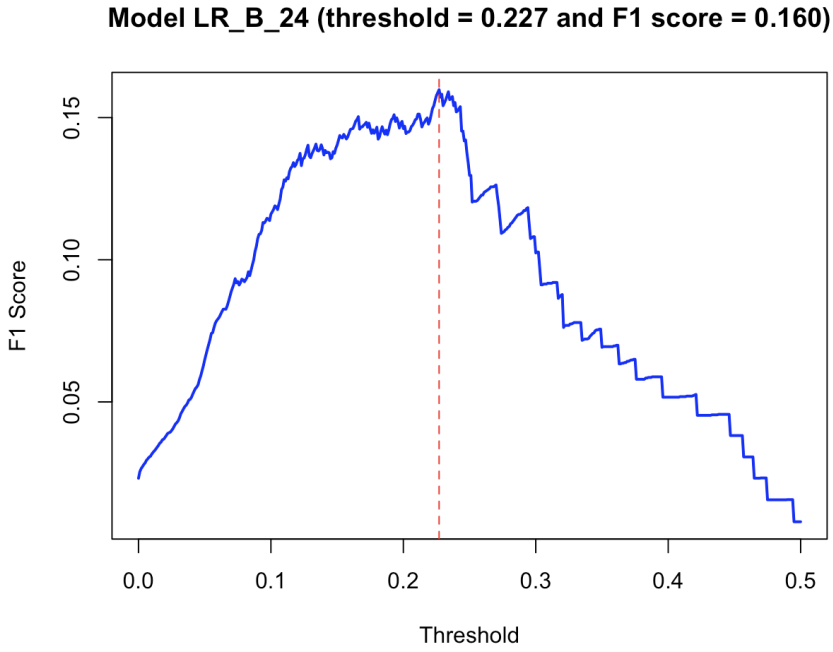

Figure A.4: F1 score as a function of threshold values for Model LR_B_24. The threshold of 0.227 results in an F1 score of 0.160, which is marked as a vertical dotted red line.

A.6 Confusion Matrix

Model	TN	FP	FN	TP	Accuracy	Precision	Recall	F1 Score
$LR_{A,12}$	5446	48	24	4	0.9870	0.0769	0.1429	0.1000
$LR_{B,12}$	5436	58	26	2	0.9848	0.0333	0.0714	0.0455
$LR_{B,24}$	5418	46	52	6	0.9823	0.1154	0.1034	0.1091

Table A.5: Confusion matrix components and classification metrics for three logistic regression models. TN = True Negatives, FP = False Positives, FN = False Negatives, TP = True Positives.

A.7 Trade-off Sensitivity and Specificity

[REDACTED]

Figure A.5: Distribution of the predicted probability of default.

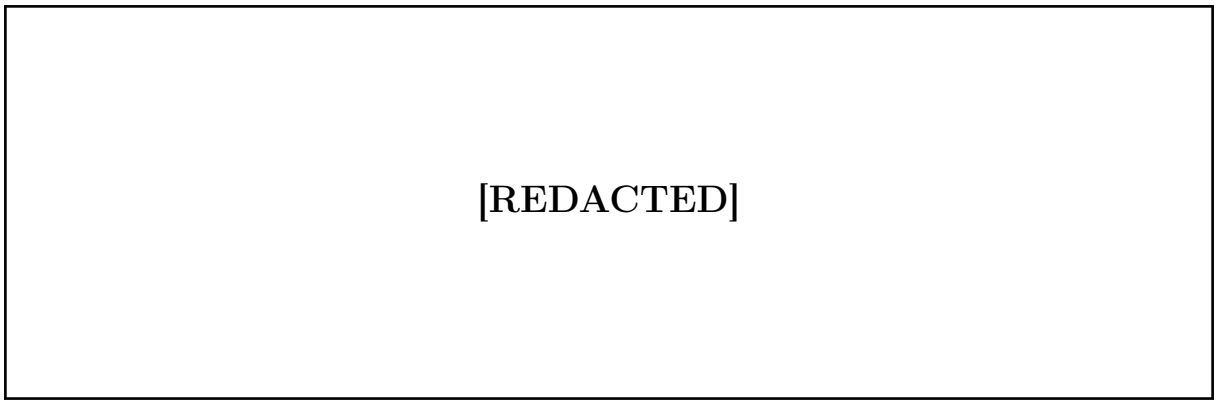

Figure A.6: Distribution of the predicted probability of default.

Bibliography

[1] Amirsaleh Azadinamin. "The Bankruptcy of Lehman Brothers: Causes of Failure & Recommendations Going Forward". In: *SSRN Electronic Journal* (2012). DOI: 10.2139/ssrn.2016892.

[2] Bank for International Settlements. *Calculation of RWA for credit risk*. Accessed: 2025-04-17. 2004. URL: https://www.bis.org/publ/bcbs107.htm.

[3] Bank for International Settlements. *Calculation of RWA for credit risk*. Accessed: 2025-04-17. 2023. URL: https://www.bis.org/basel_framework/chapter/CRE/36.htm.

[4] Basel Committee on Banking Supervision. *Supporting Document to the New Basel Capital Accord*. Tech. rep. Annex 4: The IRB Approach. Bank for International Settlements, Jan. 2001. URL: https://www.bis.org/publ/bcbsca05.pdf.

[5] Marc Baudry, Olivier H. N'Guessan, and Stéphane R. M. L. *Title of the Report*. Tech. rep. Economix - University of Paris Ouest Nanterre La Défense, 2016. URL: https://economix.fr/pdf/dt/2016/WP_EcoX_2016-10.pdf.

[6] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford Series in Probability and Statistics. Oxford: Oxford University Press, 2013. ISBN: 978-0-19-953525-5.

[7] Tim S Campbell and J Kimball Dietrich. "The determinants of default on insured conventional residential mortgage loans". In: *The Journal of Finance* 38.5 (1983), pp. 1569–1581.

[8] S. Chatterjee. "A new perspective on least squares under convex constraint". In: *Annals of Statistics* 42.6 (2014), pp. 2340–2381. DOI: 10.1214/14-AOS1242.

[9] Sabyasachi Chatterjee, Adityanand Guntuboyina, and Bodhisattva Sen. "On matrix estimation under monotonicity constraints". In: *arXiv preprint arXiv:1506.03430* (2015). arXiv: 1506.03430 [math.ST]. URL: https://arxiv.org/abs/1506.03430.

[10] T. M. Cover. "The Best Two Independent Measurements Are Not the Two Best". In: *IEEE Transactions on Systems, Man, and Cybernetics* 4 (1974), pp. 116–117.

[11] Jonathan Crook and John Banasik. "Does reject inference really improve the performance of application scoring models?" In: *Journal of Banking & Finance* 28.4 (2004), pp. 857–874. DOI: 10.1016/j.jbankfin.2003.10.010. URL: https://doi.org/10.1016/j.jbankfin.2003.10.010.

[12] Chris H. Q. Ding and Inna Dubchak. "Multi-class Protein Fold Recognition using Support Vector Machines and Neural Networks". In: *Bioinformatics* 19.4 (2003), pp. 404–411. URL: https://ranger.uta.edu/~chqding/papers/gene_select.pdf.

[13] Encord. *F1 Score in Machine Learning*. Accessed: 2023-05-04. 2023. URL: https://encord.com/blog/f1-score-in-machine-learning/#:~:text=Typically%2C%20an%20F1%20score%20%3E%200.9,to%20have%20a%20poor%20performance..

[14] Billy Fang, Adityanand Guntuboyina, and Bodhisattva Sen. "Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy–Krause variation". In: *Ann. Statist.* 49.2 (2021), pp. 769–792.

[15] Andreas Fuster et al. "The Role of Technology in Mortgage Lending". In: *Review of Financial Studies* 32.5 (2019), pp. 1854–1899.

[16] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. Springer Series in Statistics. Springer, 2009. ISBN: 978-0-387-84857-7.

[17] A. K. Jain, R. P. W. Duin, and J. Mao. "Statistical Pattern Recognition: A Review". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.1 (Jan. 2000), pp. 4–37.

[18] Jan Kroot and Evangelos Giouvris. "Dutch mortgages: Impact of the crisis on probability of default". In: *Finance Research Letters* 18 (2016), pp. 205–217.

[19] Edward C Lawrence, L Douglas Smith, and Malcolm Rhoades. "An analysis of default risk in mobile home credit". In: *Journal of Banking & Finance* 16.2 (1992), pp. 299–312.

[20] Lydian Medema, Ruud H Koning, and Robert Lensink. "A practical approach to validating a PD model". In: *Journal of Banking & Finance* 33.4 (2009), pp. 701–708.

[21] Atif Mian and Amir Sufi. "Household Leverage and the Recession of 2007 to 2009". In: *IMF Economic Review* 58.1 (2009), pp. 74–117.

[22] Hanchuan Peng, Fuhui Long, and Chris Ding. "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy". In: *IEEE Transactions on pattern analysis and machine intelligence* 27.8 (2005), pp. 1226–1238.

[23] Hemlata Sharma et al. "Analysing the Influence of Macroeconomic Factors on Credit Risk in the UK Banking Sector". In: *Analytics* 3.1 (2024), pp. 63–83. DOI: 10.3390/analytics3010005. URL: https://doi.org/10.3390/analytics3010005.

[24] Kerry D Vandell. "Default risk under alternative mortgage instruments". In: *The Journal of Finance* 33.5 (1978), pp. 1279–1296.

[25] Bianca Zadrozny and Charles Elkan. "Transforming classifier scores into accurate multiclass probability estimates". In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2002, pp. 694–699. DOI: 10.1145/775047.775151.

[26] Zhenyu Zhao, Radhika Anand, and Mallory Wang. "Maximum Relevance and Minimum Redundancy Feature Selection Methods for a Marketing Machine Learning Platform". In: *Uber Engineering Blog* (Aug. 2019). Uber AI research; accessed 2025-06-14.