Erasmus School of Economics
Department of econometrics and management science

Sample Selection Inference in Credit Scoring for Microfinance

Master Thesis - MSc Quantitative Finance

Cecile van Riele
581805

University supervisor: Dr. Mikhail Zhelonkin
Company supervisor: John Kamau
Second assessor: Nick Koning

Rotterdam, March 9, 2023

The content of this thesis is the sole responsibility of the author and does not reflect the view of the supervisor, second assessor, Erasmus School of Economics or Erasmus University.

Abstract

Sample selection bias is one of the major problems in credit scoring for microfinance. This paper addresses this problem by applying sample selection inference techniques and evaluating their performance. We compare the performance of logistic regression and three sample selection inference techniques: fuzzy augmentation, extrapolation, and Heckman's two-step bivariate probit model. These methods are combined with three feature selection methods: forward selection, backward elimination, and Lasso. We evaluate the predictive power of the methods and show the accuracy of the probability forecast. Our findings indicate that the bivariate probit model outperforms other methods. Additionally, we show that lasso is a more effective feature selection technique than stepwise regression for all credit scoring methods used. We recommend that MFIs take sample selection bias into account and implement sample selection inference in their credit scoring processes. By doing so, MFIs can make better lending decisions and reduce credit risk.

Contents

1 Introduction

Microfinance institutions (MFIs) are becoming increasingly important in the global financial landscape as they provide access to financial services and capital to those who would be otherwise unable to obtain it. As MFIs continue to grow, they must also implement sound risk management practices (Blanco et al., 2013). Credit scoring has become a key tool for managing credit risk, enabling a fast and efficient assessment of the likelihood of loan default of new loan applicants. Credit scoring is a process that predicts the probability of a credit applicant repaying their loan and determines if they should receive a loan or not. Traditionally, MFIs use methods based on the repayment performance of previously financed clients because the repayment behaviour of the non-financed clients is not available (Van Gool et al., 2012). However, this approach is inherently biased (Hand and Henley, 1997). Van Gool et al. (2012) identify sample selection bias as one of the major problems in credit scoring for microfinance. Sample selection inference refers to the technique of predicting the behaviour of non-financed loan applicants as if they had been financed. This corrects for the bias introduced by only taking into account information of financed clients. Hence, the incorporation of sample selection inference in credit scoring methods can enable better risk management in MFIs by more accurately forecasting the probability of default. Furthermore, sample selection inference might promote fair credit distribution and greater access to credit for individuals with limited credit availability. This outcome can positively impact the lives of those who struggle to access credit, promote financial inclusion, and facilitate economic growth.

In this paper, we investigate whether sample selection bias should be taken into account in credit scoring methods for microfinance. Specifically, we investigate whether sample selection inference outperforms conventional credit scoring models utilized in microfinance. Also, we identify the most effective sample selection inference for MFIs. Additionally, we examine how different methods perform under various missingness mechanisms. Moreover, we aim to gain insights into the effects of selection bias on credit scoring in the context of MFIs, and gain a better understanding of the limitations of credit scoring models currently used in this sector. Finally, this research investigates which feature selection techniques work best in combination with the commonly used logistic regression model and sample selection inference techniques.

It is important to investigate the effect of sample selection bias in microfinance, due to the differences between microloans and traditional loans. Microfinance differs from traditional finance in three ways: the loan amounts are smaller, the default rates are higher, and the reject rates are also higher (Moro-Visconti, 2016; Wright and Hitt, 2017; Wajebo, 2022). The current literature on sample selection inference in microfinance is limited since it only deals with one technique for sample selection inference. However, there are various techniques that can be classified under two different assumptions (Kim and Sohn, 2007). The first assumption is that the distribution pattern of financed applicants can be extended to that of non-financed ones or Missing At Random (MAR) data. The second assumption implies that the probability distribution of financed clients cannot be extended to that of the non-financed ones or Missing Not At Random (MNAR) data. We cannot test whether the data is MAR or MNAR from the observed data when the reason for missingness is unknown or unobservable. We found one paper by Marimo and Chimedza (2022) that demonstrates that sample selection inference improves the performance of a credit scoring model. However, they only introduce one method that deals with the MAR missingness mechanism: fuzzy augmentation. In the traditional credit scoring literature, various other methods are proposed. We contribute to the current literature by implementing some of these methods in a microfinance setting.

Aside from sample selection inference, feature selection is a major problem in credit scoring (Hand and Henley, 1997; Thomas, 2000). The problem of feature selection is especially prevalent for data sets that contain many features and limited data. Current research on sample selection inference often does not deal with the challenge of feature selection. In our case, we have a limited amount of observations and an abundance of features because we use both economic and psychometric features as suggested by Arráiz et al. (2017); Djeundje et al. (2021). The limited observations and a large amount of features result in the need for feature selection to avoid model overfitting. There is no consensus about which feature selection method works best in combination with logistic regression or sample selection inference for credit scoring (Anderson and Hardin, 2013). We differ from the current literature by integrating various feature selection and sample selection inference techniques and demonstrating which of these approaches yield optimal performance for microfinance data.

For this study, we use a dataset comprised of economic and psychometric features based on L-IFT survey data and Finbit app. The survey data consists of 32 surveys with over 400 questions, and the Finbit data consists of weekly financial data from small and medium-sized enterprises (SMEs) collected over one year. The dataset contains information about 547 SMEs across five countries (Ethiopia, Kenya, Nigeria, Indonesia, and Colombia) and includes both formal and informal loan information. Since the loan approval process for the loans in our dataset is unknown we do not know the missingness mechanism in our data. Therefore, we implement methods for both the MAR and MNAR missingness mechanisms.

In this research, we include four credit scoring methods: logistic regression, fuzzy augmentation, extrapolation, and Heckman's two-step bivariate probit. Logistic regression is the most frequently used model for credit scoring in MFIs and performs well (Van Gool et al., 2012). Moreover, logistic regression has been used extensively in credit scoring literature (Hand and Henley, 1997). Fuzzy augmentation deals with MAR data (Crook and Banasik, 2004). This method is implemented in microfinance and improved the accuracy of the predictions (Marimo and Chimedza, 2022). Moreover, Zeng and Zhao (2014) show that this method is the most accurate selection inference method in general credit scoring. Extrapolation also deals with MAR data (Crook and Banasik, 2004). This method has been shown to perform well for data sets with a relatively high default rate (Parnitzke, 2005). Since the default rate in microfinance is high, extrapolation is employed in this study. Heckman's two-step bivariate probit deals with MNAR data (Heckman, 1976, 1979). It is shown that the bivariate probit works well, especially for data sets with high reject rates (Banasik et al., 2003; Banasik and Crook, 2007; Kim and Sohn, 2007). Furthermore, this model is suitable for our application because it has been shown to perform well in business loans (Kim and Sohn, 2007).

Given the large number of features in the dataset, there is a risk of overfitting the models (Guyon et al., 2008). Therefore, three feature selection techniques are applied to each of the four methods such that a total of 12 credit scoring methods are compared. We implement the most commonly used feature selection methods; forward selection and backward elimination (Marshall et al., 2010; Anderson, 2007). These feature selection methods are simple and have been shown to work well for credit scoring in microfinance (Liu and Schumann, 2005; Van Gool et al., 2012). The third method that we implement is feature selection with lasso regularization (Tibshirani, 1996). This is a popular method which has been shown to improve the performance of both the logistic regression model and bivariate probit (Djeundje et al., 2021; Ogundimu, 2022). The performance of the models is evaluated using various measures, including the Gini coefficient, the KS statistic, sensitivity, specificity, calibration curves, the HS statistic and logarithmic score.

Our dataset is relatively small and the missingness mechanism is unknown. Therefore, we conduct a simulation study such that we can evaluate how the different methods perform for the different missingness mechanisms. Here we simulate data that are MAR and MNAR, using this simulated data we evaluate the probability forecast of all 12 credit scoring methods.

This research shows that the bivariate probit model with lasso regularization performs best in terms of predictive power and probability forecast. Fuzzy augmentation performs better than standard logistic regression. None of the extrapolation methods yield better performance compared to logistic regression with feature selection using lasso. This study also shows that feature selection methods impact the performance of the credit scoring methods, with all four methods performing better when combined with lasso regularization. Finally, we show that backward elimination results in a poorly calibrated probability forecast for all methods.

In our simulation study, we evaluate the performance of various methods in estimating probabilities under the MAR and MNAR data. When the missingness mechanism is MAR, we show that logistic regression, fuzzy augmentation, and bivariate probit produce similar results with well-calibrated probability estimates. In contrast, extrapolation gave non-well-calibrated probability estimates for MAR data. For MNAR data we find that logistic regression, fuzzy augmentation, and extrapolation give poorly calibrated probability forecasts. Heckman's two-step bivariate probit model provides reliable probability estimates.

We conclude that sample selection bias should be taken into account when creating credit scoring methods for microfinance. From the simulation study, we found that the bivariate probit methods perform similarly when compared to the other methods for MAR data. The simulation study also shows that bivariate probit methods outperform the other methods for MNAR data. Since bivariate probit with Lasso is found to be the best method for our empirical data, it is likely that our data is MNAR. Therefore, we suggest that MFIs use sample selection inference methods that take into account MNAR data, such as bivariate probit. Furthermore, feature selection techniques impact the performance of credit scoring methods. Lasso regularization outperforms the commonly used stepwise regression methods in terms of all performance measures used. Specifically, Lasso gives a better probability forecast than stepwise regression. Therefore, we suggest using Lasso instead of stepwise regression if feature selection is necessary due to, for example, limited data or a large number of features.

The rest of the paper is structured as follows: in Section 2 the structure of the data and the data transformations are presented. Section 3 presents the credit scoring methods, the feature selection methods, the performance measures and the simulation setup used in this research. In Section 4, the empirical results for the credit scoring methods are presented and Section 5 shows the simulation results for the credit scoring methods. Finally, in Section 6 we present our conclusions, discuss the limitations of this research and present our suggestions for further research.

1.1 Literature

The aim of credit scoring is to classify loan applicants into two categories: clients who are likely to pay back their loans and those who are not likely to pay back their loans. In this paper, we refer to these as default and non-default cases respectively. Over the years, various methods have been developed to classify new loan applicants. Most MFIs rely on judgemental systems or statistical techniques for credit scoring (Van Gool et al., 2012). Judgemental systems involve the use of a credit officer's expertise and knowledge to assess an applicant's creditworthiness. Statistical techniques instead rely on various quantitative methods developed over the years. Judgemental systems provide the benefit of human judgment, while statistical methods are able to process large amounts of data quickly and accurately. However, the most effective statistical method for credit scoring is not identified and depends on factors such as the characteristics used, the extent to which it is possible to separate the classes, and the objectives of the classification (Hand and Henley, 1997). Several credit scoring methods have been applied in a microfinance setting. However, the application of statistical techniques in microfinance has lagged behind the general financial institutions (Van Gool et al., 2012). Research on how credit scoring models such as logistic regression, probit regression, neural networks, and random forest perform in a microfinance setting has been done and there are several papers that compare the performance of these models. However, there is no consensus about which method is most effective and in practice, the logistic regression model is most often used (Blanco et al., 2013; Kiruthika and Dilsha, 2015; Aniceto et al., 2020; Ozgur et al., 2021). One of the major concerns in credit scoring models is sample selection bias, which occurs when a model based on clients who have received a loan in the past is applied to new applicants (Hand and Henley, 1997). Van Gool et al. (2012) identify sample selection inference as one of the major problems in microfinance. However, we found only one paper that implemented sample selection inference techniques in a microfinance setting. Marimo and Chimedza (2022) uses augmentation to correct for the sample selection bias in their credit scoring model and compare the logistic regression model to the inferred model. They demonstrate that taking into account the sample selection bias slightly enhances the accuracy of the credit scoring model. Due to the lack of papers on sample selection inference in microfinance, we turn to the traditional credit scoring literature.

Sample selection bias can be caused by various missingness mechanisms. The different missingness mechanisms are Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR) (Little and Rubin, 2019). Various techniques of sample selection inference, also called reject inference in the credit scoring literature, have been developed to infer the missing status of the non-financed cases and mitigate this bias for different missingness mechanisms. Hsia (1978) proposed augmentation, also called reweighting, which scores the non-financed cases by using a credit scoring model based on the accepted cases. This method yields mixed results. For instance, Crook and Banasik (2004) show that this method impedes useful application and Banasik and Crook (2005) confirm the poor performance of the augmentation method, especially in small samples with high reject rates. Extrapolation is a method that relies upon known performance to infer what might otherwise have happened to non-financed. Similarly to augmentation, the first step is to determine how inferred policy rejects will be treated. Then, a probability of default is derived from this model. After that, the data of the non-financed clients is augmented. This can be done in various ways, the most commonly used are fuzzy and polarised parcelling. The first paper that applies this method is Crook and Banasik (2004). In literature, extrapolation in combination with fuzzy parcelling is often referred to as fuzzy augmentation. Zeng and Zhao (2014) demonstrate the effectiveness of fuzzy augmentation and show that this is the most accurate sample selection inference method. In literature, extrapolation with polarised parcelling is referred to as either extrapolation or reclassification. To stay consistent with Anderson (2007) we will use extrapolation to refer to this method. Crook and Banasik (2004) show that extrapolation yields improvements, although small, compared to the logistic regression model. Parnitzke (2005) shows that the extrapolation majorly improves the performance of their credit scoring methods. The statistical properties and the properties of the estimators of both the augmentation and extrapolation method are shown in Ehrhardt et al. (2021). Another form of extrapolation is iterative reclassification introduced by Joannes (1993). However, Verstraeten and Van den Poel (2005) show that this method neither significantly improved performance nor profitability. Heckman's two-stage bivariate probit model Heckman (1979, 1976) has been used in various studies. Bivariate probit corrects for the correlation between the financed/non-financed model and the default/non-default model. Banasik et al. (2003); Crook et al. (2007); Marshall et al. (2010) show that bivariate probit slightly improves performance and Kim and Sohn (2007) show that the bivariate probit model works well for corporate loans. The main criticism on this model is that it is not able to correct for the sample selection bias when the bias is strong (Wu and Hand, 2007; Chen and Åstebro, 2012). Recently various machine learning techniques have been introduced in the credit scoring literature. Maldonado and Paredes (2010) introduce the idea of extrapolating the distance measurement in support vector machines SVM between the two classes of accepted to the class of rejected. In their approach, they iteratively add rejected applications to retrain the SVM. They show that their approach performs better than other reject inference approaches using SVMs. Building further on the SVM model, Li et al. (2017) introduces a method for solving the sample selection bias involving machine learning. They show that the Semi-supervised Support Vector Machines model improves the performance compared to logistic regression. However, Mancisidor et al. (2020) suggest that SVM models do not scale well in large credit scoring data sets. A further drawback is the black-box nature of SVMs, which makes it very difficult to interpret the resulting model.

One of the other main issues of credit scoring research has been to determine what features significantly influence the probability of default (Thomas, 2000). Models with a large number of features trained on a limited amount of data are prone to overfitting (Guyon et al., 2008). Various methods have been proposed in the literature to address this challenge. In literature, these methods are often combined with logistic regression or machine learning techniques, but not with sample selection inference (Hand and Henley, 1997; Guyon et al., 2008). Liu and Schumann (2005) show that stepwise regression yields good results in a credit scoring context, which is confirmed by Laborda and Ryoo (2021). Similar results are found for credit scoring in microfinance Van Gool et al. (2012). Zhou et al. (2021); Chen and Xiang (2017) show that logistic regression with feature selection outperforms several other feature selection techniques. Furthermore, Djeundje et al. (2021) show that this method also works well in microfinance. Finally, we found two studies that combine feature selection with sample selection inference. Mancisidor et al. (2020) combine fuzzy augmentation and extrapolation with forward selection and Ogundimu (2022) combine bivariate probit with lasso.

2 Data

For this study, survey data of L-IFT and data from the Finbit app developed by L-IFT were used. The survey data includes 32 surveys with more than 400 questions asked to business owners, customers, employees, suppliers, and surveyors. In addition, financial data from the SMEs is available, which was recorded via the Finbit app. This data consists of weekly financial data obtained over a one-year period. Our total data set contains information from 547 SMEs based in five countries (Ethiopia, Kenya, Nigeria, Indonesia, and Colombia). Due to the small sample size we do not consider each country separately but combine the data of the five countries. The dataset contains clients that have and have not taken a loan. In literature, this is normally recorded as accepted versus rejected clients. Instead, in this paper, we will use the terminology financed versus non-financed. We use this terminology because we do not know the reason why someone has not taken a loan. Out of the 547 observations, 267 are non-financed and 280 are financed. This gives us a reject rate of 48.8% which is in line with the average reject rates in microfinance (Wajebo, 2022). Of the financed clients, 35 have defaulted and 245 have not, which results in an overall default rate of 12.4%. In this dataset, a customer is recorded as a default case when the loan was not (fully) repaid over the agreed-upon time period, which may differ for different loans. The loan approval process for the loans is unknown, therefore it is not possible to determine whether our data MAR or MNAR. Despite suspecting that the data is MNAR due to the common use of methods like human judgment or manual overrides by many MFIs, we cannot make any definitive assumptions without further information. Finally, the dataset includes both formal and informal loan information providing a more comprehensive view of microfinance loans as many clients operate in the informal sector.

2.1 Composing of features

For this research, extensive pre-processing of the raw survey responses and financial transaction records. The features constructed from the unprocessed data set are based on commonly used features in credit scoring practices (Hand and Henley, 1997; Abdou and Pointon, 2011). Also, some of the psychometric and business performance features suggested by Klinger et al. (2013), Arráiz et al. (2017), Liberati and Camillo (2018) and Djeundje et al. (2021) are included in the data because it has been shown that these features improve the performance of credit scoring models. These features are constructed using the business diaries data, aspirations surveys, business formalisation surveys, credit scoring surveys, and attitudes towards technology surveys. In total, we have 28 features in our data set. Table 1 shows the features, their variable type, and their description. Appendix A shows the summary statistic of the features. Some of these features serve as proxies for psychometric indicators, meaning they provide indirect measures of certain psychological traits or attitudes. For example, the feature *Word survey feedback* is a proxy of how invested a business owner is in his or her business.

Table 1: Overview of the features, their variable type and description.

Feature	Variable type	Description
Age of Firm	Categorical (8 levels)	How long did the firm exist at the start of study
Country	Categorical (5 levels)	The country that the firm is located in
Formalization	Categorical (5 levels)	Indication of how the firm is registered and via what forms
Industry	Categorical (3 levels)	The industry the firm operates in
Location in the Country	Categorical (17 levels)	The town that the firm is located in
Number of Owners	Categorical (5 levels)	The amount of owners of the firm
Owner/s Gender	Categorical (4 levels)	The gender of the business owners
Sector type	Categorical (120 levels)	The sector that the firm operates in
Age	Numeric	The age of the business owners
Commercial loan	Numeric	Number of distinct loans taken from a vendor over the 52 weeks of study
Distance walked	Numeric	The average distance walked by customers to the firm
Mean expense	Numeric	Average expense of the business over the 52 weeks of study converted to dollars
Mean income	Numeric	Average income of the business over the 52 weeks of study converted to dollars
Mean Loan Repayment	Numeric	Average loan repayments over 52 weeks of the study converted to dollars
Mean Loan Taken	Numeric	Average of the loans taken throughout the 52 weeks of study converted to dollars
Mean Savings	Numeric	Average savings deposited in various accounts over the 52 weeks converted to dollars
Mean Savings Withdrawals	Numeric	Average savings withdrawn in various accounts over the 52 weeks converted to dollars
Number of children	Numeric	The number of children the business owners have
Number of Employees	Numeric	Number of employees that work at the firm at the start of the study
Number of Employees	Numeric	Number of employees that work at the firm at the start of the study
Number of loans	Numeric	The total number of distinct loans taken over the 52 weeks of study
Number of loans repayments	Numeric	The number of loan repayments made over the 52 weeks of study

Feature	Variable type	Description
Q1/Q2	Numeric	The ratio of income in Q1 to Q2, indicator of revenue growth
Q2/Q3	Numeric	The ratio of income in Q2 to Q3, , indicator of revenue growth
Q3/Q4	Numeric	The ratio of income in Q3 to Q4, indicator of revenue growth
Technology	Numeric	Number technological appliances that have been adopted in the business
Total amount loan	Numeric	The total amount of loans taken from MFIs over the 52 weeks of study converted to dollars
Total amount loan repayments	Numeric	The total amount of loans repayments to MFIs over the 52 weeks of study converted to dollars
Transactions count	Numeric	Number of transactions done over the 52 weeks
Words survey feedback	Numeric	Number of words that were given in the feedback sections of the surveys

2.2 Data transformation

An effective and commonly used transformation in the credit scoring literature is the weight of evidence transformation (WoE). It has been shown that the WoE transformation improves the model performance in the credit scoring context (Smith et al., 2002). We use the WoE transformation because it can deal with missing data. The missing data is not removed due to two reasons. First, the data set is small, removing these observations would further reduce the number of observations in the data set which is not desirable. Second, the missing data points may be an indicator of the behaviour of the survey respondent. Another reason for using the WoE transformation is that we need to transform the categorical variables in the dataset because they cannot directly be handled by some of the proposed methods. By transforming the data we turn the categorical variables into numerical variables which can be used by the methods. Most often, a dummy transformation is used. This method creates a dummy variable for each level of the categorical data. However, when the categorical variable has a large number of levels, such as the feature *Sector* in this dataset which has 120 levels, using the dummy transformation is not feasible. Therefore, the WoE transformation is used. Another advantage of the WoE transformation is that it deals with outliers in the numerical variables.

For the categorical data the WoE transformation computes a numerical value for each level of the categorical variable. The value is dependent on a binary target variable, in our case the default indicator. The WoE for each level of feature A is computed as follows:

$$WoE_i^A = \ln\left(\frac{N_i^A/SN}{P_i^A/SP}\right), \tag{1}$$

where, N_i^A is the number of data points that were labelled as negative, which is a non-default case in our application, for the i^{th} attribute of feature A. P_i^A is the number of data points that were labelled as positive, which is a default in our application, for the i^{th} attribute of feature A. SN is the total number of negatives and SP is the total number of positives.

The WoE transformation is not only used for categorical variables, but also for numerical variables, as the numerical variables also contain missing data and this transformation can improve the performance of the credit scoring models (Smith et al., 2002). To apply the WoE transformation to numerical variables, the data must first be divided into ordered bins. In this research, the standard number of 10 bins was used. Then, observations that fall within each bin are treated as one category, and the WoE evidence for this category is computed using (1). If the WoE transformation results in non-monotonic values, the number of bins is reduced until monotonic WoE values are established. Missing data points are placed into a separate bin and assigned a numerical WoE value. The WoE transformation deals with outliers. Due to the binning of the data, the outliers in the data are assigned to the first and last bin and then the average over that bin is taken. Thereby, mitigating the effect of the outliers on the data.

3 Methodology

In this section, we present four credit scoring methods. First, we discuss logistic regression which is a commonly used credit scoring model that does not take into account sample selection bias. We also present three methods that take sample selection bias into account: fuzzy augmentation, extrapolation, and bivariate probit. Additionally, we will discuss three feature selection techniques, namely forward selection, backward elimination, and lasso regularization, and how these can be used in combination with logistic regression and the three sample selection inference techniques. We show how the performance of these methods is evaluated using the Gini coefficient, the Kolmogorov-Smirnonov statistic, sensitivity, specificity, the Hosmer-Lemeshow statistic and calibration curves and the logarithmic score. Finally, we present our simulation setup.

3.1 Logistic Regression

Logistic regression is one of the most used credit scoring models in MFIs because the model is easy to estimate and interpret (Van Gool et al., 2012). Furthermore, the statistical properties and the performance of this model in credit scoring have been extensively studied (Hand and Henley, 1997). Logistic regression estimates the probability of default, based on a set of explanatory variables. From the probability of default, a binary outcome variable can be deducted. In a credit scoring context, the binary outcome variable is $y_i = 1$ when the customer defaulted on their loan or $y_i = 0$ when the customer did not default on their loan. The set of explanatory variables is $x_{i,j}$, where $i = 1, ..., n$ are the observations in the financed group and $j = 1, ..., M$ are the features. The probability of default given the set of explanatory variables is defined by:

$$p(x_i) = P(y_i = 1|x_i) = \frac{1}{1 + \exp\{-(\beta_0 + \beta^T x_i)\}}, \tag{2}$$

where, β_0 is the intercept and β is a vector of the regression coefficients with size M. The log-odds ratio of the logistic regression model is deducted from (2) and is given by:

$$\frac{p(x_i)}{1 - p(x_i)} = \exp(\beta_0 + \beta^T x_i). \tag{3}$$

The logistic regression model is linear in the log-odds which makes it easy to interpret the model. As shown in (3), the odds multiply by $\exp(\beta_j)$ as x_j increases and all other predictors stay constant. This means that if β_j is positive, increasing x_i will result in a higher probability of default p_i. However, when β_j is negative, an increase in x_i results in a decline of the probability of default $p(x_i)$.

The regression coefficients of the logistic regression model are estimated with Maximum Likelihood Estimation (MLE). We do this by maximizing the log-likelihood for a set of n

observations, which is given by:

$$\ell(\beta_0, \beta) = \sum_{i=1}^{n} \Big[y_i \log\{p(x_i)\} + (1 - y_i) \log\{1 - p(x_i)\} \Big], \tag{4}$$

where $p(x_i) = P(y_i = 1|x_i; \beta_0, \beta)$. Logistic regression is biased for the MNAR missingness mechanisms.

3.2 Fuzzy augmentation

Fuzzy augmentation method takes the sample selection bias into account by dealing with the MAR missingness mechanism (Hand and Henley, 1993). The paper of Crook and Banasik (2004) first shows how this method performs in practice. Furthermore, this is the only method that is implemented in microfinance where it is shown to improve the performance of the credit scoring model (Marimo and Chimedza, 2022). Furthermore, Zeng and Zhao (2014) show that this method, combined with the WoE transformation performs well and gives the most accurate results. With fuzzy augmentation, we estimate a posterior probability model using the data of the financed clients. We augment the data of the non-financed cases with the probability of default from the model based on the financed clients. Then, a new credit scoring model is estimated based on the extrapolated dataset containing both the financed and the augmented non-financed groups.

The dataset with both financed and non-financed clients can be described as follows. Assume we have a total of $n + m$ observations. Where $i = 1, ..., n$ observations are financed and $i = n + 1, ..., m$ observations are non-financed. The outcome variable y_i is known for the first n observations and is denoted as y_i^f (financed). This variable $y_i = 1$ in case of a defaulting client and $y_i = 0$ in case of a non-defaulting client. The outcome variable for the second m observations is unknown and is denoted as y_i^{nf} (non-financed). Let x_{ij} be a set of explanatory variables, where $i = 1, ..., n + m$ are the observations in the financed and non-financed group and $j = 1, ..., M$ are the features. The fuzzy augmentation algorithm is then executed as follows:

Step 1: Estimate a logistic regression model for the financed clients (y_i^f) as shown in Section 3.1.

Step 2: Use the logistic regression model to augment the data with the estimated probability of default for the non-financed clients $P(y_i^{nf}|x_i)$. Do so by adding a record for default $y_i^{nf} = 1$ with weight $P(y_i^{nf}|x_i)$, and adding a record for non-default $y_i^{nf} = 0$ with weight $1 - P(y_i^{nf}|x_i)$. Such that the sum of the weights equals 1.

Step 3: Estimate a new logistic regression model (Section 3.1) on the extrapolated dataset with financed observations and augmented non-financed observations.

Fuzzy augmentation is formalized as follows. Let $\hat{\beta}$ be the solution obtained from the logistic regression model estimated using y_i^f with a corresponding maximum likelihood estimate of $\hat{p_i} = P(y_i|x_i)$. Then the weighted log-likelihood for the fuzzy augmentation method is defined by:

$$\ell(\beta_0, \beta) = \sum_{i=1}^{n} \Big[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \Big] + \sum_{i=n+1}^{n+m} \Big[\hat{p}_i \log(p_i) + (1 - \hat{p}_i) \log(1 - p_i) \Big], \quad (5)$$

where $p_i = P(y_i = 1|x_i; \beta_0, \beta)$ and β_0, β are the regression coefficients that are estimated. Ehrhardt et al. (2021); Zeng and Zhao (2014); Hand and Henley (1993) show that MLE gives estimators that are consistent and unbiased for the MAR missingness mechanism. Fuzzy augmentation is implemented using the R-package scoringTools (Ehrhardt, 2020).

3.3 Extrapolation

The first paper that uses extrapolation as selection inference is Crook and Banasik (2004). Research has demonstrated that extrapolation can lead to significant performance improvements, particularly when default rates are high (Parnitzke, 2005). The literature proposes that this method deals with the MAR missingness mechanism. Extrapolation uses a method that is similar to fuzzy augmentation. However, instead of augmenting the data with two weights, we add a single observation to the dataset. This observation is the predicted outcome of the logistic regression model, indicating whether we predict a non-financed client to be a default case or a non-default case.

Assume the same setting as used in the fuzzy augmentation method where we have n financed clients and m non-financed clients. Extrapolation is executed as follows:

Step 1: Estimate a logistic regression model for the financed clients y_i^f as shown in Section 3.1.

Step 2: Use the logistic regression model to augment the data with the estimated $\hat{y}_i^{nf}$ where $\hat{y}_i^{nf} = 0$ indicates a non-default case and $\hat{y}_i^{nf} = 1$ indicates a default case.

Step 3: Estimate a new logistic regression model (Section 3.1) on the new dataset with financed observations and augmented non-financed observations.

Again, let $\hat{\beta}$ be the solution obtained from the logistic regression model. From this $\hat{\beta}$ we can obtain the prediction of the logistic regression model $\hat{y}_i \in 0, 1$. The weighted log-likelihood for the extrapolation method is the following:

$$\ell(\beta_0, \beta) = \sum_{i=1}^{n} \Big[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \Big] + \sum_{i=n+1}^{n+m} \Big[\hat{y}_i \log(p_i) + (1 - \hat{y}_i) \log(1 - p_i) \Big], \quad (6)$$

where $p_i = P(y_i = 1|x_i; \beta_0, \beta)$ and β_0, β are the regression coefficients that are estimated. Ehrhardt et al. (2021) shows that the MLE estimates of the extrapolation method are asymptotically biased for the MAR missingness mechanism. However, they show that extrapolation does produce a sharper decision boundary. This means that the predicted probabilities are closer to 0 and 1 than their true values. Even though extrapolation has biased estimates, it has been shown that extrapolation performs well empirically (Parnitzke, 2005; Ehrhardt et al., 2021). Extrapolation is implemented using the R-package scoringTools (Ehrhardt, 2020).

3.4 Bivariate probit

Heckman's bivariate two-stage model Heckman (1979, 1976) has been used for sample selection inference problems and was applied in a credit scoring context by Banasik et al. (2003); Banasik and Crook (2007); Kim and Sohn (2007). These papers show that the bivariate probit model works well for data sets with high reject rates and in business loans. The bivariate probit model assumes that the distribution of financed clients differs from that of non-financed clients. Mathematically this imposes that $P(\text{default}|x_{ij}^{f}, \text{financed}) \neq P(\text{default}|x_{ij}^{nf}, \text{non-financed})$, where x_{ij}^{f} is the set of explanatory variables of the financed clients and x_{ij}^{nf} is the set of explanatory variables of the non-financed clients. The bivariate probit model is constructed as follows: y_{1i}^* and y_{2i}^* are unobserved continuous random variables defined by:

$$y_{1i}^* = x_{1i}'\beta_1 + \epsilon_{1i}, \tag{7}$$

$$y_{2i}^* = x_{2i}'\beta_2 + \epsilon_{2i}, \tag{8}$$

where (7) is the selection equation and (8) is the default equation. Here, β_1 and β_2 are unknown parameters of x_{1i} and x_{2i}. We observe the binary variables y_{1i} and y_{2i}, where y_{1i} takes a value of 1 if it is financed and 0 if it is not financed and y_{2i} takes a value of 1 if the loan defaults and 0 if it does not default.

$$y_{1i} = \begin{cases} 1, & \text{if } y_{1i} > 0 \text{ (financed)} \\ 0, & \text{if } y_{1i} \leq 0 \text{ (non-financed)}, \end{cases}$$

$$y_{2i} = \begin{cases} 1, & \text{if } y_{2i} > 0 \text{ (default)} \\ 0, & \text{if } y_{2i} \leq 0 \text{ (non-default)}. \end{cases}$$

The error terms are assumed to be bivariate normally distributed, such that

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma), \quad \mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Whether someone is financed (y_{1i}) is always observed but whether someone defaults (y_{2i}) is only observed when someone is financed ($y_{1i} = 1$). Thus, we have three types of observations, non-financed loans, defaulting financed loans and non-defaulting financed loans. These different types of loans have corresponding probabilities:

$$P(y_{1i} = 0) = \Phi_U(x'_{1i}\beta_1),$$

$$P(y_{1i} = 1, y_{2i} = 1) = \Phi_B(x'_{1i}\beta_1, x'_{2i}\beta_2, \rho),$$

$$P(y_{1i} = 1, y_{2i} = 0) = \Phi_B(x'_{1i}\beta_1, -x'_{2i}\beta_2, -\rho),$$

where Φ_B is the cumulative distribution function (CDF) of the bivariate normal distribution and Φ_U is the CDF of the univariate normal distribution. The selection equation (7) is always estimated separately since it is fully observed. The likelihood function of the bivariate probit model is given by:

$$L(\beta_1, \beta_2, \rho) = \prod_{i=1}^{n_1} \Phi_B(x'_{1i}\beta_1, x'_{2i}\beta_2, \rho) \prod_{i=n_1+1}^{n} \Phi_B(x'_{1i}\beta_1, -x'_{2i}\beta_2, -\rho) \prod_{i=n+1}^{m} \Phi_U(x'_{1i}\beta_1), \quad (9)$$

where the cases $i = 1$ to n_1 are the financed default cases. The cases n_1 up to n are the financed non-default cases and the cases n up to m are the non-financed cases.

3.5 Feature selection

The dataset used for this study has a large set of features. To avoid overfitting and improve the performance of the methods, we implement three feature selection methods on the logistic regression model. The model with the best fit is used as the benchmark model. The three methods include forward selection, backward elimination, and least absolute shrinkage and selection operator (lasso). Forward selection and backward elimination are two stepwise regression methods that are designed to identify the most important features that influence a given outcome. Through an iterative process, these methods add or subtract features from the model until no further improvement to the Akaike Information Criterion (AIC) is achieved. Lasso regression is a regularisation method that ensures that the absolute sum of the coefficients is less than a certain fixed value. The details of these selection methods are further discussed in the next two paragraphs.

3.5.1 Stepwise regression

Stepwise regression iteratively evaluates the statistical significance of the independent variables and is often used as a feature selection technique in credit scoring (Hand and Henley, 1997). These methods are simple but yield good results when applied in a credit scoring context (Liu and Schumann, 2005). Laborda and Ryoo (2021) show that stepwise regression in combination with logistic regression yield better results than other feature selection methods. Furthermore, Van Gool et al. (2012) use forward selection in a microfinance setting and yields good results. The advantage of stepwise regression is that it is relatively easy. However, a disadvantage is that it can lead to a local optimal solution and is sensitive to data type (Fahrmeir et al., 1994). There is no research on how well forward selection and backward elimination work in combination with sample selection inference. Mancisidor et al. (2020) combine forward selection with fuzzy augmentation and extrapolation. However, they do not compare them to other feature selection techniques.

Forward selection and backward elimination are both wrapper methods and can easily be combined with the various proposed methods. Forward selection starts with no selected features, and add new features incrementally, prioritizing the feature that improves the AIC the most. In the backward elimination method, we start with a model that uses all features and then we remove one feature in each iteration. We stop adding or subtracting features when doing so does not further improve the overall AIC of the model.

3.5.2 Lasso

Lasso is a regularisation technique introduced by Tibshirani (1996). This regularisation technique takes into account all features, but only a subset is selected as a predictor in the final model. Lasso has previously been combined with both logistic regression and bivariate probit. It has not been combined with fuzzy augmentation and extrapolation. Zhou et al. (2021) show that logistic regression with lasso feature selection improves the credit scoring models the most compared to various other feature selection techniques such as multivariate adaptive regression splines. Furthermore, Chen and Xiang (2017) shows that lasso yields better results than backward elimination. Finally, Djeundje et al. (2021) shows that enhancing the logistic regression model with lasso also improves the performance of the credit scoring in a microfinance setting where both economic and psychometric features are used. Ogundimu (2022) shows that bivariate probit with lasso yields better results than bivariate probit without lasso.

Lasso adds regularisation and selects features by adding an additional restriction to the negative log-likelihood. This restriction ensures that the sum of the absolute value of the regression coefficients is less than a fixed value t. Lasso in the logistic regression setting can be formalized as follows:

$$\begin{aligned}(\hat{\beta}_0^{lasso}, \hat{\beta}^{lasso}) = \min_{\beta_0, \beta} \quad & -\ell(\beta_0, \beta) \\ \text{s.t.} \quad & \textstyle\sum_{j=1}^{p} |\beta_j| \leq t,\end{aligned} \tag{10}$$

where β_0 is the intercept and β is the vector of the regression coefficients of size M. The log-likelihood $\ell(\beta_0, \beta)$ is the log-likelihood of logistic regression shown in (4), fuzzy augmentation shown in (5) or extrapolation shown in (6). From (10) we can derive the Lagrangian:

$$(\hat{\beta}_0^{lasso}, \hat{\beta}^{lasso}) = \min_{\beta_0, \beta}\{-\ell(\beta_0, \beta) + \lambda(\sum_{j=1}^{M} |\beta_j|)\},$$

where λ is a non-negative penalty term. A sufficiently large penalty term λ will shrink the coefficients of the least important features to zero, thus effectively removing them from the model. When $\lambda = 0$, the optimization problem simplifies to the well-known maximum likelihood estimation This allows lasso to estimate the parameters and choose the important features simultaneously. Unfortunately, lasso fails to distinguish irrelevant predictors from the true ones when predictors are highly correlated and drops them arbitrarily (Zhao and Yu, 2006). Furthermore, lasso has the issue of introducing an additional bias by shrinking non-zero coefficients towards zero (Bühlmann and Van De Geer, 2011). Thus, this needs to be investigated before using lasso as a feature selection technique.

We determine λ by using the AIC to estimate the optimal tuning parameter. The amount of regularisation is amplified as λ becomes larger. The tuning parameter is determined by performing a 10-fold cross-validated search over a wide range of values. The tuning parameter that results in the lowest AIC value is then used. We implement lasso using the glmnet package in R.

Combining lasso with bivariate probit is less straightforward than combining it with logistic regression, fuzzy augmentation or extrapolation. Ogundimu (2022) introduce Lasso penalized Heckman-type bivariate probit model and assess its performance in identifying predictive features in credit scoring. They suggest that lasso methods should be preferred for optimal predictions. We consider the likelihood function $L(\beta_1, \beta_2, \rho)$ shown in (9). The first elements of β_1, β_2 are $\beta_{1,0}, \beta_{2,0}$ and the last elements are β_{1i}, β_{2i} where $i = 1, ..., n + m$. This is constructed such that $\beta_{1,0}, \beta_{2,0}, \rho$ are not penalised. The lasso estimator is given by:

$$(\hat{\beta}_1^{lasso}, \hat{\beta}_2^{lasso}, \hat{\rho}^{lasso}) = \min_{\beta_1, \beta_2, \rho}[-\log\{L(\beta_1, \beta_2, \rho)\} + \lambda\{\sum_{j=1}^{n+m}(|\beta_{1j}| + |\beta_{2j}|)\}],$$

where $\lambda > 0$ and the second term in (3.5.2) is the penalty term.

3.6 Performance measures

Performance measures are used to compare the performance of the methods based on various aspects. Anderson (2007) suggests measuring the performance based on two main aspects: predictive power and accuracy of the probability forecast. Predictive power is the ability of the method to separate default versus non-default. In this study, the performance measures to assess the predictive power are the sensitivity, the specificity, the Gini coefficient and the Kolmogorov-Smirnov (KS) statistic. The calibration is the accuracy of the estimated probability of default relative to actual default rates. In this study, the performance of the method in terms of the accuracy of the probability forecast is measured by the calibration curve and Hosmer-Lemeshow (HL) statistic. Finally, the quality of the probability forecast is evaluated using the logarithmic score (LS) which maximises the sharpness of the predictive distributions, subject to calibration. These performance measures are further elaborated in the paragraphs below.

3.6.1 Predictive power

The goal of a credit scoring model is to classify the loans into default and non-default cases. To do so, a classification threshold is needed. This threshold divides the loans into default cases and non-default cases and is chosen based on the estimated probabilities. There is no universal way to determine the threshold. However, ideally, the threshold would be determined by optimising a cost function that reflects the actual cost of an MFI. This cost function should take into account that the cost of granting credit to someone who will eventually default is higher than the cost of not granting credit to someone who will not default. However, the cost function is unknown nevertheless, a threshold needs to be determined in order to obtain classification results. Thus, we obtain a threshold by minimizing the sum of the error frequencies which is the same as maximising the sum of sensitivity and specificity. The sensitivity and specificity are calculated based on the confusion matrix as shown in Table 2. The confusion matrix provides a visual representation of the performance of a classification model. In the context of credit scoring, this matrix and the measures derived from it are valuable due to the fact that the economic consequences of a false positive (FP) are not the same as those of a false negative (FN).

Table 2: Confusion matrix.

		Predicted	
		Default	Non-default
Actual	Default	True positive (TP)	False negative (FN)
	Non-Default	False positive (FP)	True negative (TN)

Sensitivity is the ratio of correctly classified default cases (TP) to the total number of actual default cases and is computed as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

A high sensitivity means that the model correctly classifies the default cases (TP), while a low sensitivity indicates that many default cases are incorrectly classified as non-default cases (FN). Thus, a low sensitivity leads to a too-liberal model and grants credit to many risky customers. This is an important consideration for credit scoring, as a great number of false negatives can lead to a higher amount of defaulting loans.

Specificity measures the ratio of correctly classified non-default cases to the total number of actual non-default cases and is computed as follows:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}.$$

A high specificity means that the model correctly classifies the non-default cases, while a low specificity could lead to a high false positive rate. Thus, a low specificity results in a model that falsely classifies too many non-default cases as a default case, making the model too conservative and denying credit to many potential customers. It is important to carefully consider the trade-off between sensitivity and specificity in order to find a balance that maximizes the performance of the model.

The Gini coefficient is an important measure of discriminatory power used to evaluate the performance of a classification model (Anderson, 2007). To calculate the Gini coefficient, the Area Under the Curve (AUC) is used which is derived from the Receiver Operating Characteristic (ROC) curve. The ROC curve plots the true positive rate (sensitivity) of the model against the false positive rate (1 - specificity) for various thresholds and is the cumulative distribution function of the model's prediction scores. The Gini coefficient is a simplified representation of the AUC and is calculated by $\text{Gini} = 2 \cdot \text{AUC} - 1$. The values of the Gini coefficient are between 0 and 1. Here a value of 0 corresponds to random classification and 1 is an asymptotic value that is not reached in practice. A higher Gini coefficient indicates that the model is better able to distinguish between the two groups (e.g. defaulting and non-defaulting clients).

The Kolmogorov-Smirnov (KS) statistic is a widely used measure for quantifying the predictive power of a model (Anderson, 2007). The statistic can be used to test the null hypothesis that the two samples are drawn from the same distribution. If the KS-statistic is high, the null hypothesis is rejected, indicating a significant difference between the Cumulative Distribution Functions (CDFs) of the two samples. In the context of credit scoring, the KS-statistic measures the maximum difference between the CDFs of the default group and the non-default group. A KS-statistic value of 0 suggests that the model is not able to differentiate between the default group and non-default group, while a value greater than 0 implies that the model is able to distinguish between the two groups. The larger the KS statistic, the more successful the model is at distinguishing the default group and non-default group. The

KS-statistic is computed as follows:

$$D_{KS} = \sup_{x}\{|F_d(x) - F_{nd}(x)|\},$$

where $F_d(x)$ and $F_{nd}(x)$ are the cumulative probability distribution functions of the default group and non-default group respectively.

3.6.2 Accuracy

The quality of a credit risk model cannot solely be determined by its ability to accurately identify default cases and non-default cases, but also by the quality of its probabilistic forecast. To evaluate the probabilistic forecast both its calibration performance and sharpness should be investigated. The calibration performance focuses on the appropriateness of the probability of default (PD) estimate. In other words, the calibration performance provides an indication of how likely a borrower classified as a default case will actually default. On the other hand, sharpness refers to the concentration of the predictive distributions and is a property of the forecasts only. In this research, we evaluate the calibration of the models using the calibration curve and HL statistic, followed by an evaluation of the calibration and sharpness simultaneously using the logarithmic score, a proper scoring rule.

Gneiting and Ranjan (2013) showed that probabilistic and conditional calibration are equivalent and can be evaluated with a calibration curve. A calibration curve is constructed by plotting the conditional event frequencies against the binned forecast probabilities. A perfectly calibrated model will give a calibration curve that is equal to the diagonal between $(0, 0)$ and $(1, 1)$. Deviations from the diagonal of the graph indicate that the model is not well calibrated.

The Hosmer-Lemeshow (HL) or chi-square test is one of the most commonly used methods to assess the goodness of fit or calibration of credit risk models (Anderson, 2007). The HL-test assesses the null hypothesis which is that the observed and expected probabilities are not the same. It is computed by first ordering the predicted probabilities and putting them into bins. Then the observed and expected default rates within each bin are compared. The HL statistic is computed as follows:

$$HL = \sum_{k=1}^{b} \left[n_k \frac{(p_k - \hat{p}_k)^2}{\hat{p}_k(1 - \hat{p}_k)} \right],$$

where k is an index for each bin and b is the total number of bins. The number of observations in bin k is n_k, the average observed probability in bin k is p_k and $\hat{p}_k$ is the estimated average probability of bin k. The conventional number of bins used for this statistic is $b = 10$. This is also used in this study. The values of the HL statistic are chi-square distributed with 2 degrees of freedom. A small p-value or a significant HL statistic means that there is evidence that the model is not well-calibrated. Thus, a large p-value indicates that there is no evidence that the model is not correctly specified.

The logarithmic Score, also known as the log score, is a proper scoring rule (Gneiting and Katzfuss, 2014). This scoring rule indicates how close the predicted probability of default is to the corresponding true value. It is a local proper scoring rule, meaning assesses both calibration and sharpness simultaneously and it only depends on the value the model attains at a specific observation. The logarithmic score is calculated by taking the natural logarithm of the ratio between the predicted probability of an event occurring and the observed frequency of the event. The average logarithmic score is obtained by:

$$\text{Logarithmic score} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right],$$

where, N is the number of observations, p_i is the probability of default and y_i is the binary outcome observation of observation i. The values of the logarithmic score range from 0 to infinity, with lower values indicating better performance. The logarithmic score metric is sensitive to extreme predictions and assigns higher values to predictions in the correct direction. For instance, if a default occurs ($y_i = 1$) and the predicted probability of default is 0, the logarithmic score would be equal to infinity. This means that even if the other predictions are accurate, the average logarithmic score will be very poor. However, the logarithmic score is closely related to the loss function of the proposed methods and is a suitable metric for model evaluation of proposed credit scoring methods, as these extreme values are not likely to occur. Therefore, logarithmic score is suitable for the model evaluation of the proposed credit scoring methods.

3.7 Simulation setup

To study the behaviour of the methods under different missingness mechanisms we conduct a simulation study. For this simulation study, we generate data that incorporates the MAR or MNAR missingness mechanisms. Similarly to our empirical setting, we generate thirty features. For this setup we assume that these simulated features are normally independently distributed with feature x_j, where $j = 1, ..., 30$ such that $x_j \sim \mathcal{N}(0, \sigma^2)$, where we fix σ^2 to 0.5. The two error terms ϵ_1 and ϵ_2 are generated from the multivariate normal distribution with mean 0, variance 1 and covariance ρ. We set $\rho = 0$ for the MAR simulation and $\rho = -1$ for the MNAR simulation. The binary outcomes are generated based on the selection equation and outcome equation. The selection equation is $y_1 = \beta_{1,0} + \sum_{j=1}^{30} \beta_{1,j} x_j + \epsilon_1$, if $y_1 > 0$, then $Y_1 = 1$ (financed); otherwise $Y_1 = 0$ (non-financed). The outcome equation is $y_2 = \beta_{2,0} + \sum_{j=1}^{30} \beta_{2,j} x_j + \epsilon_2$, if $y_2 > 0$, then $Y_2 = 1$ (default); otherwise $Y_2 = 0$ (non-default). In both the MAR and MNAR settings, β_1 and β_2 are sparse, where we set 50% of their entries to non-zero, where five out of the thirty coefficients are non-zero for both β_1 and β_2. For simplicity, we set all non-zero entries to 0.5. Furthermore, we set $\beta_{1,0} = -0.25$ and $\beta_{2,0} = 0$. This simulation setup results in a reject rate of approximately 50% and a default rate in the financed sample of approximately 13% which is similar to our actual dataset. For each scenario, 500 sets of data, each consisting of 2000 independent observations, are generated.

4 Empirical results

This section shows the results of logistic regression, fuzzy augmentation, extrapolation and bivariate probit, which are introduced in Sections 3.1 to 3.4. Each method is combined with forward selection, backward elimination and lasso, as presented in Section 3.5. In total 12 credit scoring methods are compared. The method's forecasting ability is evaluated using a 70%/30% train-test split. The training set contains 25 default and 382 non-default cases and the test set contains 10 default cases and 73 non-default cases. The methods are evaluated using the Gini coefficient, KS statistic, sensitivity, specificity, HL statistic, calibration curves and logarithmic score presented in Section 3.6.

As described in Section 3.5.2, the assumption that the correlation between the predictors is limited needs to hold if we want to use lasso for feature selection. The assumption is checked and holds as shown in Appendix B.

4.1 Predictive power

Table 3 shows the Gini coefficient, KS statistic, sensitivity and specificity for all methods used in this study. First, we evaluate the discriminatory power of the methods using the Gini coefficient and KS statistic. Next, we evaluate how well each method performs in classifying default and non-default cases using sensitivity and specificity. The results of this evaluation will provide insight into the predictive power of each method.

Table 3: Performance measures of logistic regression (LR), fuzzy augmentation (FA), extrapolation (EX) and bivariate probit (BP) with forward selection, backward elimination and lasso. Boldface indicates the best model for that metric.

	Gini	KS D	Sensitivity	Specificity
LR forward	0.844	0.789	0.9	0.889
LR backward	0.836	0.792	**1.0**	0.792
LR lasso	0.883	0.831	0.9	0.931
FA forward	0.853	0.761	0.9	0.861
FA backward	0.886	0.831	0.9	0.931
FA lasso	0.903	0.844	0.9	0.944
EX forward	0.817	0.758	0.8	**0.958**
EX backward	0.836	0.819	**1.0**	0.819
EX lasso	0.875	0.758	0.8	0.944
BP forward	0.833	0.719	0.9	0.819
BP backward	0.908	**0.861**	**1.0**	0.861
BP lasso	**0.919**	0.847	**1.0**	0.847

Table 3 shows that the bivariate probit using lasso has the highest Gini coefficient overall. However, both the bivariate probit with lasso and backward elimination outperform the other methods. Fuzzy augmentation with lasso has the third-highest Gini coefficient, followed by fuzzy augmentation with backward elimination. Logistic regression with lasso performs similarly to fuzzy augmentation with backward elimination, as well as logistic regression with forward and backward elimination. Finally, logistic regression and extrapolation with backward elimination, as well as extrapolation and bivariate probit with forward selection, perform the least well in terms of the Gini coefficient.

The bivariate probit with backward elimination has the highest KS statistic. Bivariate probit with lasso is a close second followed by fuzzy augmentation with lasso. Fuzzy augmentation with backward and logistic regression with lasso are both the fourth best and equal in terms of KS statistics. Fuzzy augmentation with forward selection and extrapolation with forward selection and lasso also have similar KS statistics and perform less well than the other methods. Bivariate probit with forward selection has the lowest KS statistic.

The sensitivity and specificity are based on a threshold for each method. The thresholds minimize the sum of the error frequencies and are shown in Appendix D. Since the sum of the error frequencies is minimized, we can only evaluate the performance by taking both sensitivity and specificity simultaneously into account. The methods presented in Table 3 all have a high sensitivity, with all methods classifying more than 80% of the default cases correctly. The highest sensitivities are achieved by logistic regression with backward elimination, extrapolation with backward elimination and bivariate probit with backward elimination and lasso, all of which correctly classify all default cases. Logistic regression with forward selection and lasso, fuzzy augmentation with forward selection, backward elimination and lasso and bivariate probit with forward selection have a sensitivity of 90%. The bivariate probit model with backward elimination performs the best when considering both sensitivity and specificity. However, bivariate probit with lasso is a close second. These methods have the highest specificities of the methods with a sensitivity of one. The third-highest sum of sensitivity and specificity is achieved by fuzzy augmentation with lasso, followed by fuzzy augmentation with backward elimination and logistic regression with lasso which have the same performance. However, these methods do not have a sensitivity of one and thus perform less well at classifying default cases. This means that, for these thresholds, these methods would perform less well in practice due to the high cost of classifying default cases as non-default cases. Even though extrapolation with forward selection has the highest specificity it is one of the worst performers in terms of both sensitivity and specificity simultaneously.

In short, these results provide insights into the performance of various methods in predicting credit defaults. The bivariate probit model with backward elimination demonstrated the highest discriminatory power and classification accuracy among all the methods evaluated in this study. However, the bivariate probit model with lasso is also highly effective and shows slightly better performance in terms of the Gini coefficient. Fuzzy augmentation with lasso demonstrates promising results in terms of discriminatory power and performs reasonably well in terms of classification accuracy, although it is not as good as some of the other methods in terms of sensitivity. Notably, logistic regression with lasso stands out as an effective method, showing good performance in terms of the Gini coefficient, KS statistic, sensitivity, and specificity. There is insufficient evidence to suggest that all sample selection inference outperform logistic regression because, methods such as fuzzy augmentation with forward selection, extrapolation methods, and bivariate probit with forward selection, do not perform well in terms of predictive power. Overall, our findings suggest that bivariate probit with backward elimination, lasso, and fuzzy augmentation with backward elimination and lasso are promising methods for modelling credit scores.

4.2 Accuracy

The HL statistic, calibration curves and logarithmic score are used to evaluate the accuracy of the various methods. Table 4 shows the HL p-value and logarithmic score for each method and Figure 1 shows the calibration curve for each method. The HL statistic gives an indication of the method's calibration, and the calibration curves provide insight into how well the model performs in different probability areas. The logarithmic score is a proper scoring rule and measures the accuracy of probabilistic forecasts.

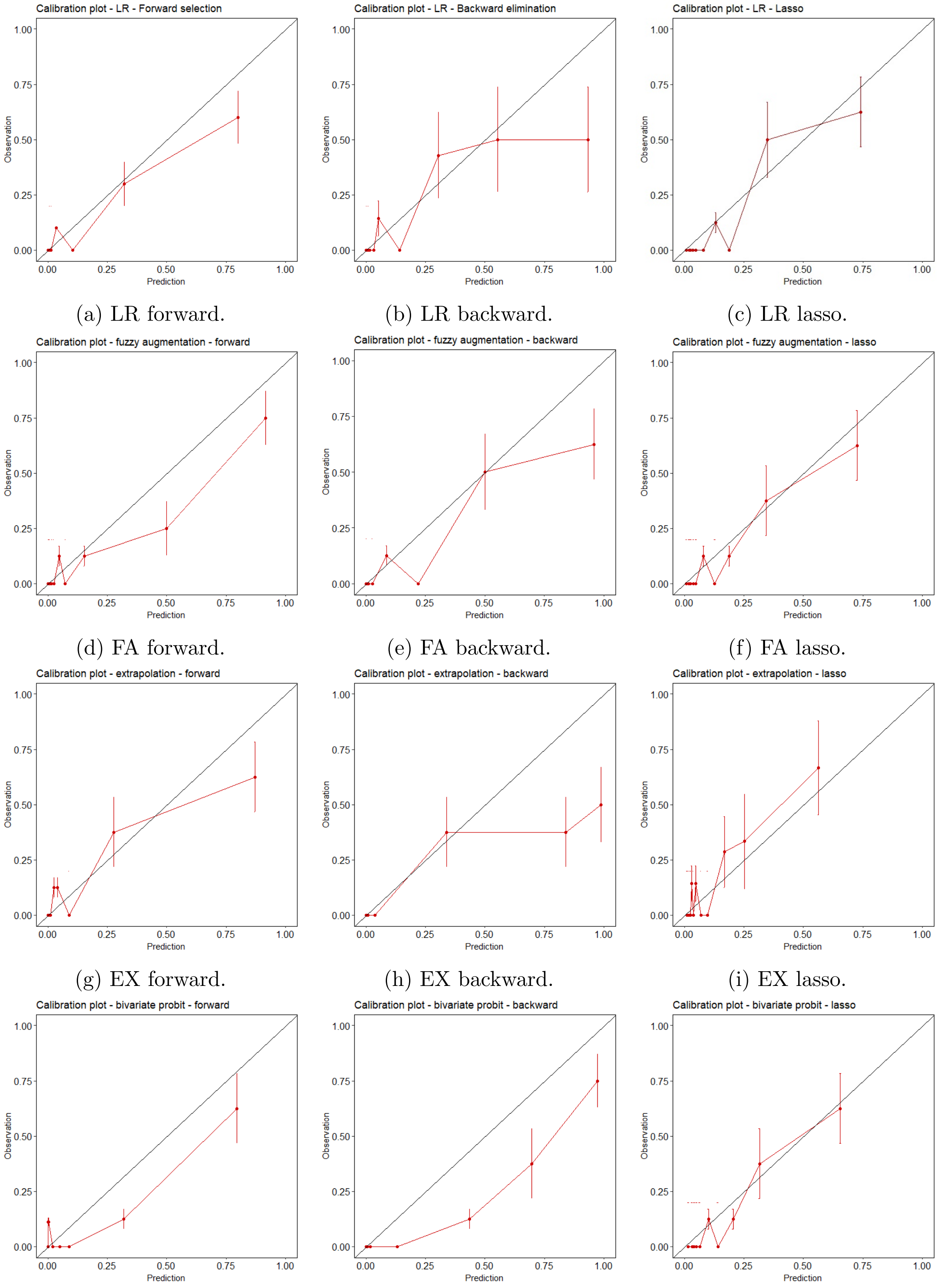
Table 4: Perfromance measures of logistic regression (LR), fuzzy augmentation (FA), extrapolation (EX) and bivariate probit (BP) with forward selection, backward elimination and lasso. A star means that the p-value is significant (99% confidence level) and boldface indicates the best model for that metric.

	HL p-value	Log score
LR forward	0.408	0.259
LR backward	0.000*	0.299
LR lasso	0.885	0.217
FA forward	0.335	0.258
FA backward	0.071	0.257
FA lasso	0.840	0.196
EX forward	0.500	0.276
EX backward	0.000*	0.464
EX lasso	0.783	0.246
BP forward	0.174	0.278
BP backward	0.003*	0.298
BP lasso	**0.928**	**0.188**

First, the HL-statistic and calibration curves are evaluated. The calibration curves presented in Figure 1 are not smooth which is likely caused by the small number of observations in the test set. However, some methods are better calibrated than others, and most calibration curves are in line with the HL p-values presented in Table 4. A significant HL p-value means that the null hypothesis of observed and expected probabilities being the same, is not rejected. Thus, when we have a p-value of 0.01 or higher we reject the null hypothesis and have no evidence that the model is not well-calibrated. Logistic regression with backward elimination, extrapolation with backward elimination and bivariate probit with backward elimination have significant p-values which means that these methods are not well-calibrated. The calibration curves shown in Figures 1b, 1h and 1k support these findings since the calibration curves of these methods lie far from the diagonal. The other methods have non-significant p-values, which indicates that there is no evidence that these methods are not well-calibrated. This is confirmed by most calibration plots, as most methods with significant p-values have calibration curves that lie relatively close to the diagonal. Although not significant, fuzzy augmentation with backward elimination has a low p-value. Figure 1e shows that the calibration curve lies quite far from the diagonal for two probability bins. Which points towards calibration issues in the probability areas of 0.25 and 0.9. The calibration curve of the bivariate probit with forward selection also deviates relatively much from the diagonal in the two highest probability bins, as shown in Figure 1j. These probability bins lie under the diagonal thus, this method overpredicts the probability of default. The HL p-value of this method is non-significant but relatively low compared to the other methods. The calibration curve of bivariate probit with lasso shown in Figure 1l, is the probability curve that lies closest to the diagonal. Moreover, the HL p-value is close to one. Thus, both the calibration curve and HL p-value indicate that bivariate probit with lasso is the best-calibrated method out of the 12 methods presented here.

The methods that are well-calibrated give accurate probability estimates which can be interpreted as probabilities of default. However, these estimated probabilities should be evaluated further by also considering the sharpness of the probability forecast, which refers to how concentrated or spread out the forecast probabilities are around the event's true probability. Sharpness should only be evaluated if the method is well-calibrated. The logarithmic score takes into account both sharpness and calibration. The lower the logarithmic score the better the method performs. Table 4 shows that the bivariate probit with lasso has the lowest logarithmic score. This indicates that this method gives the best probability estimates. The second lowest logarithmic score is logistic regression with lasso and the third lowest is fuzzy augmentation with lasso. However, these are already somewhat higher than that of bivariate probit with lasso. This indicates that the lasso feature selection method actually results in better overall probability estimates. The methods that are not well-calibrated also give high logarithmic scores. This further indicates that those methods do not perform well in predicting the probability of default.

In short, we find that bivariate probit with lasso was found to be the best model in terms of explanatory power, with a well-calibrated calibration curve and the lowest logarithmic score. Furthermore, we find that logistic regression with backward elimination, extrapolation with backward elimination, and bivariate probit with backward elimination are not well-calibrated.



(a) LR forward.

(b) LR backward.

(c) LR lasso.

(d) FA forward.

(e) FA backward.

(f) FA lasso.

(g) EX forward.

(h) EX backward.

(i) EX lasso.

(j) BP forward.

(k) BP backward.

(l) BP lasso.

Figure 1: Calibration curves for the classification methods. Where the error bars are calculated by bootstrapping parcels in each probability bin.

The lasso feature selection method was also found to improve the overall probability estimates. Overall, the calibration plots do not give nice calibration curves which is mainly caused by the lack of observations in our dataset. This makes it hard to draw definitive conclusions from these plots. Therefore, we conduct a simulation study in the next section, where we simulate a larger dataset and show how these methods perform for different missingness mechanisms.

After evaluating various methods based on their predictive power and accuracy of the probability forecast, the bivariate probit model with lasso stands out as the best-performing method. While the bivariate probit model with backward elimination performs well in predictive power, its probability estimates are not well-calibrated and cannot be relied upon for accurate default probabilities. Fuzzy augmentation with lasso also demonstrates strong performance in both predictive power and accuracy. Furthermore, logistic regression with lasso is a reliable method that consistently performs well, outperforming many of the other methods evaluated. The extrapolation methods, in contrast, perform considerably worse than the other methods in terms of both predictive power and accuracy. All the methods with backward elimination exhibit issues with calibration. These results confirm the results of Marimo and Chimedza (2022) and show that fuzzy augmentation indeed improves the performance compared to logistic regression. However, here we show that bivariate probit outperforms fuzzy augmentation. We confirm that bivariate probit works well when the reject rates are high (Banasik et al., 2003; Banasik and Crook, 2007; Kim and Sohn, 2007). We show that bivariate probit in combination with lasso performs well and we show that this method outperforms other methods Ogundimu (2022). We show that logistic regression with lasso performs better than both forward and backward elimination. This is in line with the paper of Chen and Xiang (2017) which shows that logistic regression with lasso is better than backward elimination. We show that backward elimination in combination with all algorithms gives bad calibration performance, this may be caused by some of the disadvantages of the backward elimination method such as, biased regression coefficients. We show that lasso in combination with fuzzy augmentation and extrapolation is better than stepwise regression. These methods in combination with fuzzy augmentation and extrapolation are used in Mancisidor et al. (2020). Our results suggest that it is better to use lasso instead.

4.3 Feature selection and interpretation

Table 5 shows which features were selected for the credit scoring method and the corresponding parameter estimates. We do not show the significance of the parameter estimates because we implement feature selection methods. This means that the parameter estimates become conditional on the selection process and are therefore biased. Therefore, we can not say anything about the feature's importance based on the significance of the features. However, we can look at how many times the features are selected and the trace plots of lasso to give some indication of feature importance. The table shows that the features *Mean income* and *Q1/Q2*, *Q3/Q4* and *Sector type* are selected in all methods. This might imply that these features have better predictive power than the other features. Moreover, it is shown that some of the features are never or rarely selected by the feature selection methods. This implies that these features are not as predictive as the other features. Appendix E shows the trace plots of the different methods with lasso. These plots also show that *Mean income*, *Q2/Q3* and *Sector type* are important features. Furthermore, Table 5 shows that lasso selects as many as or more features than the other methods. The reason for this is that the optimal turning parameter of lasso is relatively low, such that there is not a large penalty for selecting many features.

Table 5: Features selected by Logistic regression (LR), fuzzy augmentation (FA), extrapolation (EX) and bivariate probit (BP) with forward selection (Fs), backward elimination (Be) and lasso (La). The coefficient of the feature is shown if the feature was included in the regression such that it was selected by the stepwise regression method or had a coefficient that was larger than zero for lasso.

	LR			FA			EX			BP		
	Fs	Be	La	Fs	Be	La	Fs	Be	La	Fs	Be	La
Intercept	-5.15	-4.77	-3.91	-4.80	-5.32	-3.50	-5.98	-8.46	-3.78	-2.34	-2.91	-2.34
Mean income	1.38	1.18	0.71	1.53	1.01	0.74	1.90	2.19	1.16	0.68	0.64	0.59
Q1 / Q2	1.73	2.17	0.86	1.31	1.98	0.73	1.57	2.90	1.09	0.84	0.94	0.76
Q3 / Q4	1.62	1.87	0.31	1.92	2.46	0.34	1.95	3.08	0.36	1.09	1.02	0.19
Sector type	2.10	2.90	1.08	1.77	1.71	1.11	2.03	2.84	1.25	1.24	1.70	0.98
Q2 / Q3	1.92	2.20	0.25			0.29	1.14	2.98	0.38	0.75	1.10	0.22
Words survey feedback	1.72	1.66		1.40	1.28	0.05	2.28	2.89		0.75	0.71	
Mean saving withdraw			0.45		0.75	0.27			0.95	0.49	0.52	0.34
Nr of Employees	1.47	1.42				0.12		3.06	0.28	0.97	1.18	
total amount loan	1.58	2.04	0.32		1.23	0.14	1.09	0.23		1.15	0.72	0.13
Nr of loans repay	1.72		0.68		1.33	0.29	2.23		0.92	1.26		0.45
Location in Country	1.55		1.27			1.32	1.82	2.11	1.43	0.73		1.19
Mean loan taken		2.28	0.62		2.08	0.46	1.12	3.37	1.12		1.06	0.46
Nr of children			0.69		2.37	1.12		2.28	0.83			0.68
Country		2.01	0.05		2.09	0.09		3.28	0.36		0.98	
Mean expense			0.69		1.57	0.77		3.51	1.11		1.00	0.58
Technology	1.27	1.45	0.45			0.36		1.88			0.81	0.28
Owners gender		2.44	0.56		2.13	0.43		4.26	0.87		1.44	0.40
Mean savings					1.39		1.49	2.14				
Age of Firm						0.10			0.50			
Nr of loans	2.76		0.15			0.46				1.33		
Age					1.18			1.73				
Transactions count						0.14		1.32				
Commercial loan			0.89			1.14			1.54			0.71
Total loan repayment			0.03			0.74						
Mean loan repay												
Distance walked												
Industry												
Nr of owners												
Total features selected	12	12	18	10	15	22	11	17	17	12	14	15

Table 5 shows that the parameter estimates are all positive. So, we have to look at the transformed data to see what the direction of the relationship is between the feature and the probability of default. The WoE transformation gives us monotonic values. However, these values can be monotonically increasing or decreasing, as shown in Appendix C. In the case of monotonically decreasing WoE values, the probability of default will decrease as the actual feature increases. There are four features that have increasing WoE values, which are; *Mean expense, Age, Number of children* and *Transaction count*. This means that an increase in these features will result in an increase of the probability of default. This means that if an SME has more expenses, the probability of default increases, which is in line with results presented in Djeundje et al. (2021); Blanco et al. (2013). Literature shows that the direction of the relationship between loan default and age or transaction amounts is not consistently positive or negative (Blanco et al., 2013). So far we have discussed the direction of the relationship between the features and the probability of default. However, we also have the strength of the relationship between the two. Table 5 shows that the strength of the relationship between the features and the probability of default is not consistent between the different methods. However, it is shown that lasso often has smaller coefficients, this is likely caused by the regularisation imposed on the model.

In summary, *Mean income, Q1/Q2, Q2/Q3 and Q3/Q4*, and *Sector type* are consistently selected by different feature selection techniques, while some other features are rarely selected by the feature selection methods. Four features have a negative relation with the probability of default and all other features have a positive relationship with the probability of default. The coefficients vary much from method to method and we do not have reliable p-values, therefore we can not draw any conclusions about the strength of the relation between the features and the probability of default.

5 Simulation results

In this section, we examine how well the methods discussed in Sections 3.1 to 3.5 perform in terms of probabilistic forecasting for different missingness mechanisms. To achieve this, we conduct a controlled study in which we simulate data under different missingness mechanisms. The data is generated using the simulation setup presented in Section 3.7. In this section, we only focus on the performance of the probability forecast because this can give us better insights into the bias of the different methods for the different missingness mechanisms.

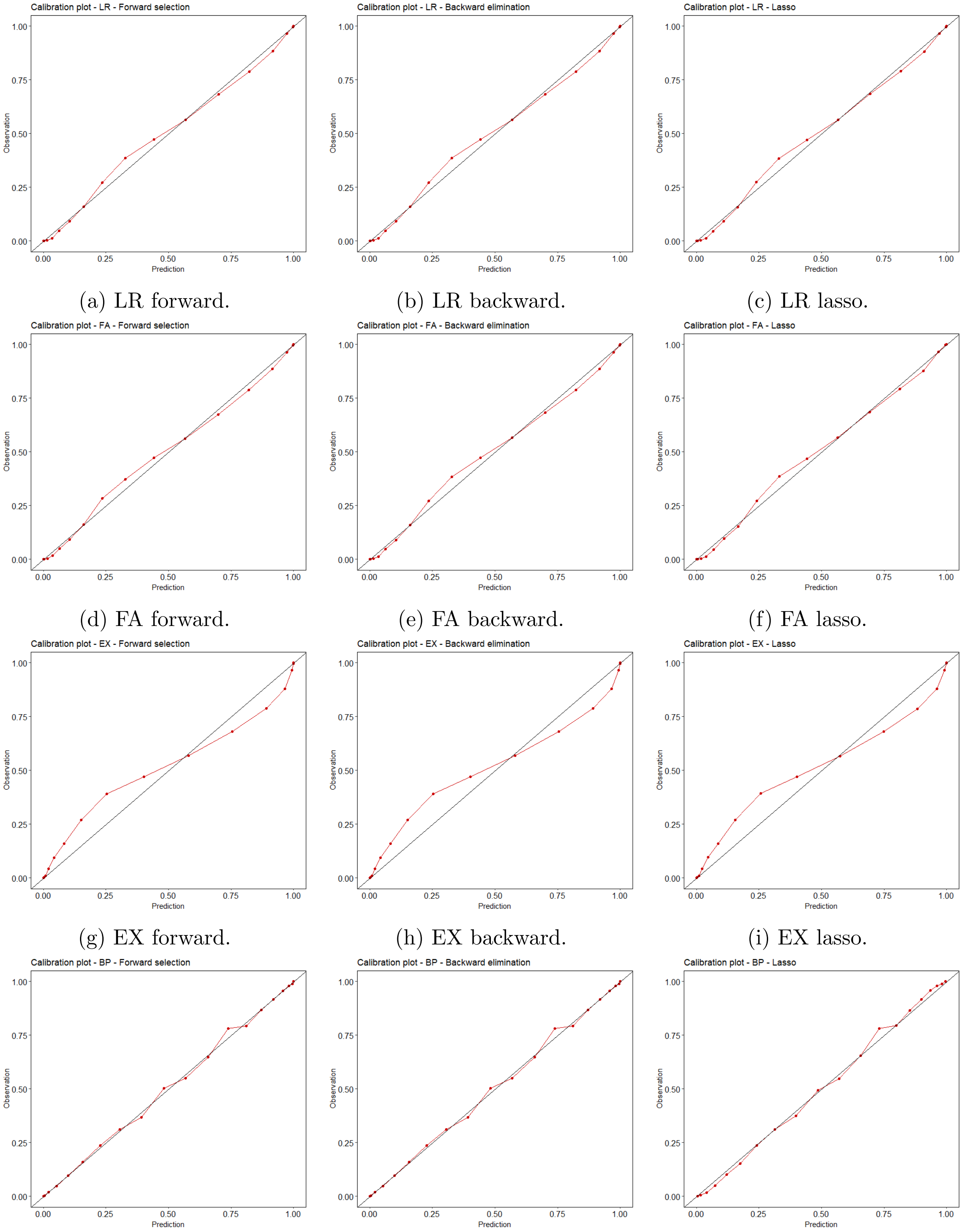
Figure 2 shows the calibration plots for the 12 different methods under the MNAR missingness mechanism. The calibration curves of bivariate probit do not deviate much from the diagonal which indicates that this method is well calibrated. The calibration curves of logistic regression, fuzzy augmentation and extrapolation deviate more from the diagonal than the calibration curves of bivariate probit. However, the differences between the methods are relatively small. Furthermore, the different feature selection techniques do not show large differences in their results. This is likely caused by the larger number of observations included in the data set.

Logistic regression mainly shows deviations in the lower probability regions. Here, the calibration curve lies above the diagonal. This means that the average predicted probability of default is smaller than the actual default rate of that group. Thus, we underpredict the probability of default for the cases with a low probability of default. The calibration curves of logistic regression also slightly deviate from the optimal curve in the higher probability areas. Here the calibration curve lies slightly under the optimal curve which indicates that we over-predict the probability of default in these areas. Although, the deviations are less strong than in the lower probability areas they are still present.

Fuzzy augmentation gives a slightly better-calibrated probability forecast than logistic regression. The deviations are similar and in the same direction as for logistic regression. So this method also under-predicts the probability for the lower probabilities of default and over-predicts for the higher probabilities of default.

The calibration curves of the extrapolation method are the least well-calibrated. The effects of over-predicting the high probabilities of default and under-predicting the low probabilities of default are stronger in these calibration plots. This effect can be explained by the bias that is introduced in this method. Ehrhardt et al. (2021) show that the extrapolation method has a sharper decision boundary than logistic regression and fuzzy augmentation. Thus, the predicted probabilities of this method are more extreme, which explains the effects we see in the calibration plots.

We show that Heckman's two-step bivariate probit method corrects the sample selection bias introduced due to the MNAR missingness mechanisms. The method does not give perfectly calibrated results. Especially in the higher probability regions, the calibration plots slightly deviate from the optimal curve. However, the calibration curves of all variations of the bivariate probit model lie closer to the true calibration curve than the calibration



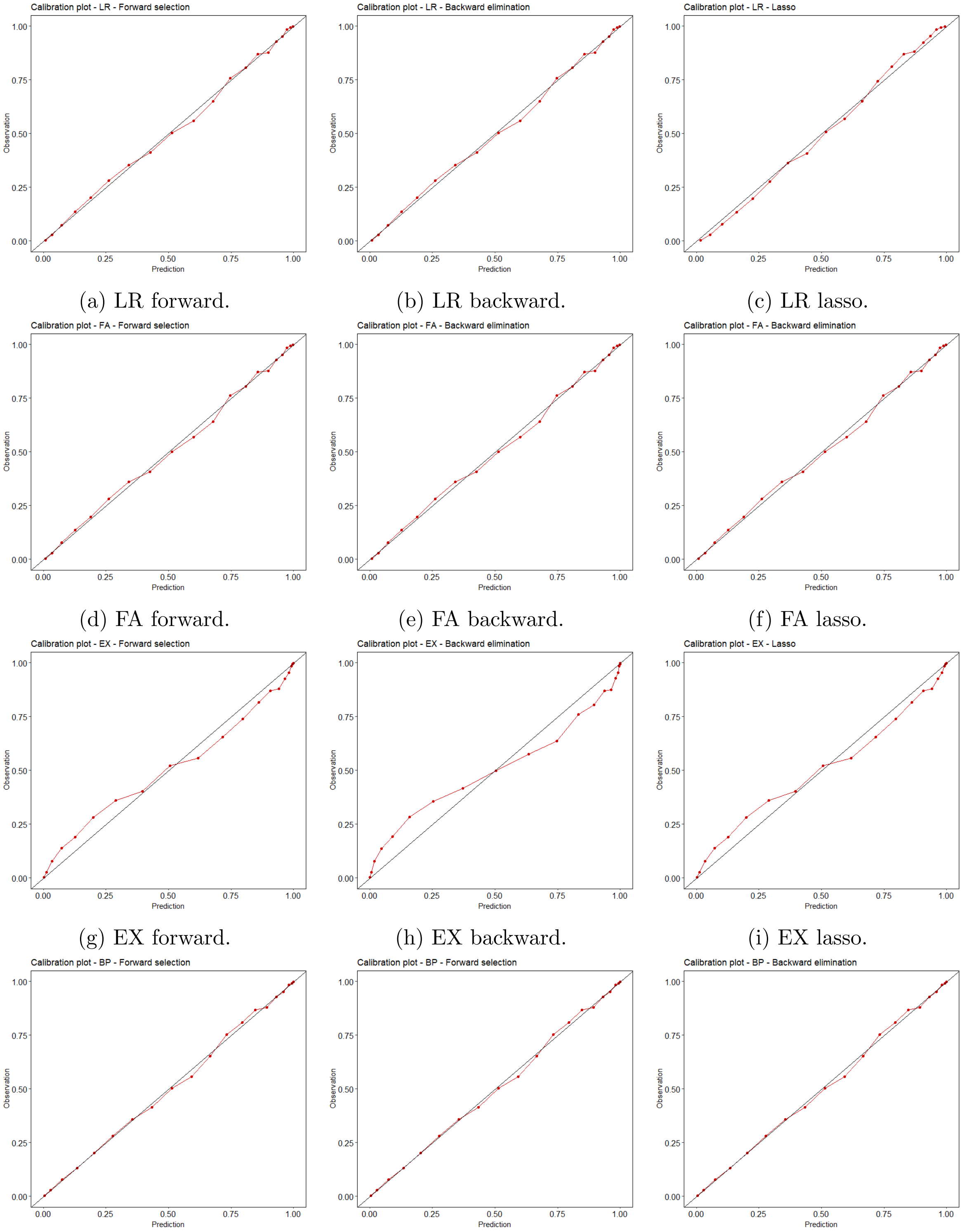
(a) LR forward. (b) LR backward. (c) LR lasso.

(d) FA forward. (e) FA backward. (f) FA lasso.

(g) EX forward. (h) EX backward. (i) EX lasso.

(j) BP forward. (k) BP backward. (l) BP lasso.

Figure 2: Calibration curves for the classification methods for the simulated data with the MNAR missingness mechanism.



(a) LR forward.

(b) LR backward.

(c) LR lasso.

(d) FA forward.

(e) FA backward.

(f) FA lasso.

(g) EX forward.

(h) EX backward.

(i) EX lasso.

(j) BP forward.

(k) BP backward.

(l) BP lasso.

Figure 3: Calibration curves for the classification methods for the simulated data with the MAR missingness mechanism.

curves of logistic regression, fuzzy augmentation and extrapolation. Especially for the lower probability regions where the calibration curve almost exactly follows the diagonal. Even though the difference between the calibration of logistic regression, fuzzy augmentation and bivariate probit are relatively small in credit scoring small differences in performance can make big differences in economic results. Our simulation results demonstrate that bivariate probit provides the most accurate calibration curve among all tested methods. Therefore, we show that Heckman's two-step bivariate probit model indeed infers MNAR data (Heckman, 1976).

Figure 4 shows a boxplot of the logarithmic scores based on MNAR simulated data. Here we can see that the difference between logistic regression, fuzzy augmentation and bivariate probit are small. However, the logarithmic score of bivariate probit is consistently smaller

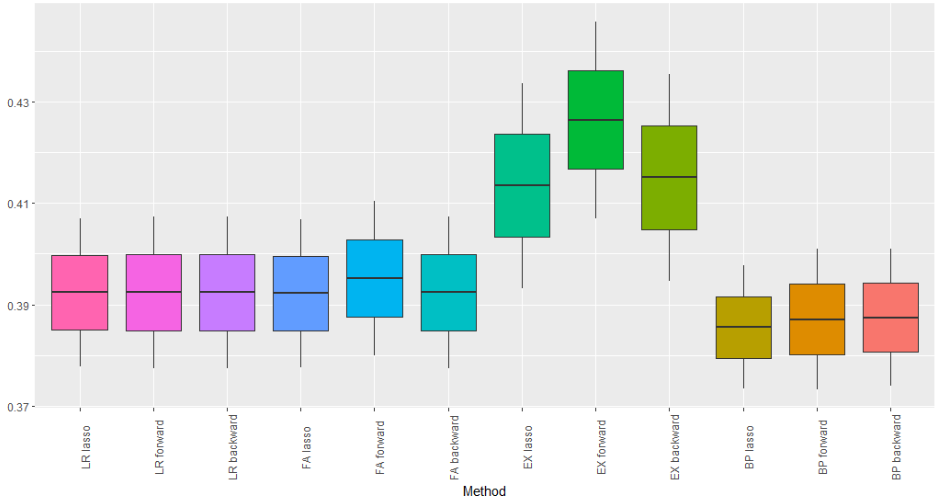


Figure 4: Logarithmic scores under MNAR data.

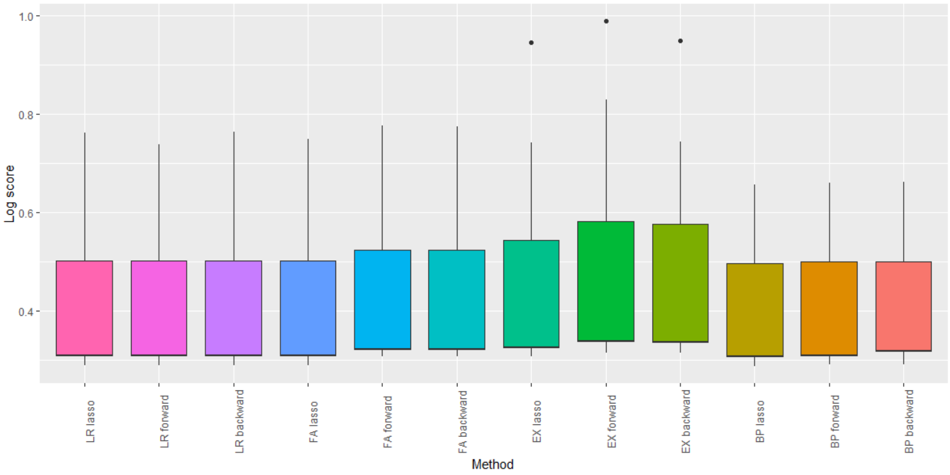


Figure 5: Logarithmic scores under MAR data.

than that of logistic regression and fuzzy augmentation. This indicates that bivariate probit gives a better probability forecast than the other methods. Furthermore, the figure shows that the logarithmic score for all extrapolation methods is higher that that of the other methods. This confirms that the probability forecast of extrapolation is not accurate.

Figure 3 shows the calibration plots for the 12 different methods under the MAR missingness mechanism. The calibration curves show that the calibration under the MAR missingness mechanism is similar for logistic regression, fuzzy augmentation and bivariate probit. Zeng and Zhao (2014) and Ehrhardt et al. (2021) show that for MAR data fuzzy augmentation and logistic regression give the same estimates for the coefficients if the features are not re-selected in the third step of the algorithm presented in Section 3.2. This is likely the reason that these two methods give similar results in our simulation study. The calibration curve of extrapolation deviates from the optimal curve. Here again, we see that the higher probabilities are over-predicted and the lower probabilities are under-predicted. Ehrhardt et al. (2021) show that the extrapolation is biased even for MAR data, this is confirmed by the results presented in Figure 3. Therefore, the probability forecast of this method is not reliable.

Figure 5 also shows that logistic regression, fuzzy augmentation and bivariate probit give similar logarithmic scores. This indicates that these methods give a similar probability forecast. Again the logarithmic scores of extrapolation are somewhat higher which indicates that these methods perform less well for this performance measure.

In summary, the bivariate probit model corrects for sample selection bias due to MNAR missing data. It gives better-calibrated probability estimates compared to logistic regression, fuzzy augmentation and extrapolation. Logistic regression, fuzzy augmentation and bivariate probit give a similar well-calibrated probability forecast for MAR data. On the contrary, extrapolation does not give a good probability forecast for MAR data.

6 Conclusion and Discussion

Our research compares logistic regression, fuzzy augmentation, extrapolation, and Heckman's two-step bivariate probit model for credit scoring in microfinance. Each of these methods is combined with forward selection, backward elimination and lasso. We evaluate the predictive power and probability forecast using multiple metrics. Also, we conduct a simulation study for MAR and MNAR data.

Our analysis reveals that bivariate probit with lasso outperforms all other methods used in this study. This is evident from its superior predictive performance and accuracy of the probability forecast. Additionally, our findings indicate that the probability forecast of bivariate probit is comparable to that of the other methods for MAR data. Our study also shows that bivariate probit gives a better probability forecast compared to the other methods for the MNAR assumption. Although the differences under this assumption are small they are relevant.

Our empirical results show that fuzzy augmentation yields better results in terms of predictive power than logistic regression, which is in line with earlier studies. Additionally, we show that the probability forecast of fuzzy augmentation does not consistently outperform logistic regression. In our simulation study, we find no conclusive evidence to suggest that fuzzy augmentation outperforms logistic regression for either MAR or MNAR data. When compared to logistic regression, our analysis indicates that extrapolation does not improve performance for the empirical data. In our simulation results, we demonstrate that extrapolation fails to provide a well-calibrated probability forecast. Our empirical results highlight that the choice of feature selection method has a noticeable impact on the performance of credit scoring methods. Specifically, all four methods demonstrate their optimal performance when combined with lasso regularization. Our findings suggest that the bivariate probit model with backward elimination or lasso, exhibits comparable predictive power. However, the probability forecast of the bivariate probit with backward elimination is not well-calibrated.

MFIs often overlook sample selection bias in their credit scoring process, resulting in biased credit scoring models. Sample selection bias arises due to missing data, which may occur due to various missingness mechanisms. The limited literature on sample selection inference in microfinance focuses on the MAR missingness mechanism. However, we doubt if this is correct since many MFIs use methods like human judgement in their loan approval process which often causes the data to be MNAR. Since the missingness mechanism in our data is unknown we use methods that assume MAR or MNAR data. We show that for MAR data the different methods yield similar results. When using MNAR data, bivariate probit gives a better probability forecast. Since, bivariate probit outperforms our other methods in our empirical study and bivariate probit assumes MNAR data, we think it is likely that our data is MNAR. Therefore, we recommend that MFIs consider sample selection bias in their credit scoring process and use methods such as bivariate probit, which infer MNAR data.

In our empirical results, we find that lasso provides better results in terms of both predictive power and accuracy compared to stepwise regression for all four methods used in this study. The superior performance of lasso over stepwise regression is likely caused by the ability of lasso to find a global optimum instead of getting stuck in a local optimum like stepwise regression. Our research, where data is limited and there are many features to consider, suggests that lasso is a better choice for feature selection than stepwise regression.

We encountered several limitations while conducting this research. One of these limitations is the limited amount of observations in our dataset. We conclude that our data is likely MNAR. There are several other methods that deal with this type of missing data however, these methods need larger data sets. To enhance the credit scoring research in microfinance, we suggest that future research uses larger data sets in combination with bivariate probit and other methods that deal with MNAR data.

Another limitation of our research is that a limited amount of feature selection methods are employed. Since this study shows that feature selection has an impact on the performance of the credit scoring method, we suggest that the methods used in this research are also combined with other feature selection methods. Furthermore, we only show the performance of the four credit scoring methods in combination with feature selection to prevent overfitting. When more data is available we can use more features because the risk of overfitting becomes less prevalent. Hence, we suggest that the currently used methods are also implemented without feature selection in further research when more data is available.

Our dataset only contains information about SMEs which may limit the generalisability of the findings. The results may only apply to business loans, and it is uncertain whether they also apply to consumer loans. Moreover, our dataset contains data from five countries which are used together because of limited data availability. Investigating multiple countries together may overlook country-specific differences that may affect the results. It is also unclear whether these findings are applicable to other countries that are not included in the dataset, as loan approval mechanisms may vary from country to country. Future research could investigate whether the results presented in our research hold for consumer loans, explore if there are country-specific differences, and assess the generalizability of these findings to other countries.

Another limitation is that we only show how these methods perform for one reject rate. The traditional credit scoring literature shows that the effects of sample selection bias and the effectiveness of different variations of sample selection inference vary for different reject rates. Further research could investigate whether these findings also apply to microfinance.

Finally, future studies could build on our findings by investigating the impact of not taking sample selection bias into account on the inclusion or exclusion of certain groups. Specifically, they could explore whether not accounting for sample selection bias may result in the exclusion of individuals who do not fit the typical borrower profile, further exacerbating existing disparities in credit access. This could inform the development of more inclusive credit scoring models that expand access to credit for individuals who are currently underserved by traditional credit scoring methods.

In conclusion, it is important to consider sample selection bias in credit scoring for microfinance since sample selection inference offers a more accurate way for MFIs to assess credit risk compared to traditional methods. The most effective sample selection inference technique may depend on the specific missingness mechanism present. However, we show that bivariate probit performs well in various settings. Specifically, our analysis demonstrates that bivariate probit with lasso regularization performs well when applied to real-world microfinance data. Therefore, MFIs should consider implementing sample selection inference in their credit scoring process.

Bibliography

Abdou, H. A. and Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting, Finance and Management*, 18(2-3):59–88. 7

Anderson, B. and Hardin, J. M. (2013). Modified logistic regression using the em algorithm for reject inference. *International Journal of Data Analysis Techniques and Strategies*, 5(4):359–373. 2

Anderson, R. (2007). *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation.* Oxford University Press. 3, 5, 17, 19, 20

Aniceto, M. C., Barboza, F., and Kimura, H. (2020). Machine learning predictivity applied to consumer creditworthiness. *Future Business Journal*, 6(1):1–14. 5

Arráiz, I., Bruhn, M., and Stucchi, R. (2017). Psychometrics as a tool to improve credit information. *The World Bank Economic Review*, 30:S67–S76. 2, 7

Banasik, J. and Crook, J. (2005). Credit scoring, augmentation and lean models. *Journal of the Operational Research Society*, 56(9):1072–1081. 5

Banasik, J. and Crook, J. (2007). Reject inference, augmentation, and sample selection. *European Journal of Operational Research*, 183(3):1582–1594. 3, 14, 28

Banasik, J., Crook, J., and Thomas, L. (2003). Sample selection bias in credit scoring models. *Journal of the Operational Research Society*, 54(8):822–832. 3, 6, 14, 28

Blanco, A., Pino-Mejías, R., Lara, J., and Rayo, S. (2013). Credit scoring models for the microfinance industry using neural networks: Evidence from peru. *Expert Systems with Applications*, 40(1):356–364. 1, 5, 30

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media. 17

Chen, G. G. and Åstebro, T. (2012). Bound and collapse bayesian reject inference for credit scoring. *Journal of the Operational Research Society*, 63(10):1374–1387. 6

Chen, H. and Xiang, Y. (2017). The study of credit scoring model based on group lasso. *Procedia computer science*, 122:677–684. 6, 16, 28

Crook, J. and Banasik, J. (2004). Does reject inference really improve the performance of application scoring models? *Journal of Banking & Finance*, 28(4):857–874. 2, 5, 12, 13

Crook, J. N., Edelman, D. B., and Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 183(3):1447–1465. 6

Djeundje, V. B., Crook, J., Calabrese, R., and Hamid, M. (2021). Enhancing credit scoring with alternative data. *Expert Systems with Applications*, 163:113766. 2, 3, 6, 7, 16, 30

Ehrhardt, A. (2020). *Credit Scoring Tools: the 'scoringTools' package*. `https://adimajo.github.io/scoringTools`. 13, 14

Ehrhardt, A., Biernacki, C., Vandewalle, V., Heinrich, P., and Beben, S. (2021). Reject inference methods in credit scoring. *Journal of Applied Statistics*, 48(13-15):2734–2754. 5, 13, 14, 31, 35

Fahrmeir, L., Tutz, G., Hennevogl, W., and Salem, E. (1994). *Multivariate statistical modelling based on generalized linear models*, volume 425. Springer New York. 16

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1(1):125–151. 21

Gneiting, T. and Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, 7:1747–1782. 20

Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. A. (2008). *Feature extraction: foundations and applications*, volume 207. Springer. 3, 6

Hand, D. and Henley, W. (1993). Can reject inference ever work? *IMA Journal of Management Mathematics*, 5(1):45–55. 12, 13

Hand, D. J. and Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society Series A*, 160(3):523–541. 1, 2, 4, 5, 6, 7, 11, 15

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of economic and social measurement, volume 5, number 4*, pages 475–492. 3, 6, 14, 34

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161. 3, 6, 14

Hsia, D. C. (1978). Credit scoring and the equal credit opportunity act. *The Hastings Law Journal*, 30:371±448. 5

Joannes, D. N. (1993). Reject inference applied to logistic regression for credit scoring. *IMA Journal of Management Mathematics*, 5(1):35–43. 6

Kim, Y. and Sohn, S. (2007). Technology scoring model considering rejected applicants and effect of reject inference. *Journal of the Operational Research Society*, 58(10):1341–1347. 1, 3, 6, 14, 28

Kiruthika and Dilsha, M. (2015). A neural network approach for microfinance credit scoring. *Journal of Statistics and Management Systems*, 18(1-2):121–138. 5

Klinger, B., Khwaja, A. I., and Del Carpio, C. (2013). *Enterprising psychometrics and poverty reduction*. Springer. 7

Laborda, J. and Ryoo, S. (2021). Feature selection in a credit scoring model. *Mathematics*, 9(7):746. 6, 16

Li, Z., Tian, Y., Li, K., Zhou, F., and Yang, W. (2017). Reject inference in credit scoring using semi-supervised support vector machines. *Expert Systems with Applications*, 74:105–114. 6

Liberati, C. and Camillo, F. (2018). Personal values and credit scoring: new insights in the financial prediction. *Journal of the Operational Research Society*, 69(12):1994–2005. 7

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons. 5

Liu, Y. and Schumann, M. (2005). Data mining feature selection for credit scoring models. *Journal of the Operational Research Society*, 56(9):1099–1108. 3, 6, 16

Maldonado, S. and Paredes, G. (2010). A semi-supervised approach for reject inference in credit scoring using svms. In Perner, P., editor, *Advances in Data Mining. Applications and Theoretical Aspects*, pages 558–571. Springer. 6

Mancisidor, R. A., Kampffmeyer, M., Aas, K., and Jenssen, R. (2020). Deep generative models for reject inference in credit scoring. *Knowledge-Based Systems*, 196:105758. 6, 16, 28

Marimo, M. and Chimedza, C. (2022). Sme credit risk modelling in south africa: A case study. *Acta Economica*, 20(36):9–30. 2, 5, 12, 28

Marshall, A., Tang, L., and Milne, A. (2010). Variable reduction, sample selection bias and bank retail credit scoring. *Journal of Empirical Finance*, 17(3):501–512. 3, 6

Moro-Visconti, R. (2016). Microfinance vs. traditional banking in developing countries. *International Journal of Financial Innovation in Banking*, 1:43–61. 1

Ogundimu, E. O. (2022). On lasso and adaptive lasso for non-random sample in credit scoring. *Statistical Modelling*. 3, 6, 16, 17, 28

Ozgur, O., Karagol, E. T., and Ozbugday, F. C. (2021). Machine learning approach to drivers of bank lending: evidence from an emerging economy. *Financial Innovation*, 7(1):1–29. 5

Parnitzke, T. (2005). Credit scoring and the sample selection bias. *Institute of Insurance Economics, University of St. Gallen (mimeo)*. 2, 5, 13, 14

Smith, E. P., Lipkovich, I., and Ye, K. (2002). Weight-of-evidence (woe): quantitative estimation of probability of impairment for individual and multiple lines of evidence. *Human and Ecological Risk Assessment*, 8(7):1585–1596. 9, 10

Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International journal of forecasting*, 16(2):149–172. 2, 6

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288. 3, 16

Van Gool, J., Verbeke, W., Sercu, P., and Baesens, B. (2012). Credit scoring for microfinance: is it worth it? *International Journal of Finance & Economics*, 17(2):103–123. 1, 2, 3, 4, 5, 6, 11, 16

Verstraeten, G. and Van den Poel, D. (2005). The impact of sample bias on consumer credit scoring performance and profitability. *Journal of the Operational Research Society*, 56:981–992. 6

Wajebo, T. W. (2022). Micro, small and medium entreprises acces to finance constraints in ethiopia: Demand side analysis. *International Journal of Small and Medium Enterprises*, 5(1):32–39. 1, 7

Wright, M. and Hitt, M. A. (2017). Strategic entrepreneurship and sej: Development and current progress. *Strategic Entrepreneurship Journal*, 11(3):200–210. 1

Wu, I.-D. and Hand, D. J. (2007). Handling selection bias when choosing actions in retail credit applications. *European Journal of Operational Research*, 183(3):1560–1568. 6

Zeng, G. and Zhao, Q. (2014). A rule of thumb for reject inference in credit scoring. *Mathematical Finance Letters*, pages 1–13. 2, 5, 12, 13, 35

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563. 17

Zhou, Y., Uddin, M. S., Habib, T., Chi, G., and Yuan, K. (2021). Feature selection in credit risk modeling: an international evidence. *Economic Research-Ekonomska Istraživanja*, 34(1):3064–3091. 6, 16

A Summary statistics

The summary statistics of numeric features are shown in Table 6.

Table 6: Summary statistics of numeric features.

Feature	Mean	Standard deviation	Min	Max
Age	40.55	11.15	2	80
closeassociates_loan	0.09	0.29	0	1
closefamily_loan	0.10	0.30	0	1
commercial_loan	0.11	0.31	0	1
distance_walked	3.74	18.34	0	360
Mean_expense	387.60	882.18	2.80	10864.70
Mean_income	669.10	1737.18	9.10	27409.60
Mean_LoansTakenRepayment	190.60	1582.75	0	34200
Mean_LoanTaken	522.90	2448.33	0	26600
Mean_Savings	788.60	2105.00	0	29479.20
Mean_SavingsWithdrawals	444.60	1001.59	0	11350.20
Number of children	2.38	2.06	0	23
Number of Employees	6.01	4.56	0	50
number_of_loans	0.76	2.06	0	18
number_of_loans_repayments	1.20	3.71	0	40
pre_pay	0.07	0.25	0	1
Q1/Q2	0.44	3.52	-1	71.55
Q2/Q3	0.14	1.20	-1	12.94
Q3/Q4	0.20	2.46	-1	31.10
Technology	1.64	1.42	0	9
total_amount_loan	492.30	2304.73	0	26649
total_amount_loan_repayments	369.80	2068.98	0	34295
Transactions_count	826.80	697.65	3	5102
words_survey_feedback	52.40	59.64	0	410

Table 7: Summary statistics of categorical features.

Feature	Levels	Nr of observations
Owner/s Gender	All men	294
	All women	166
	Both men and women	86
	Undefined	1
Sector type	F	48
	A	37
	AD	32
	AQ	24
	D	23
	S	21
	Other levels	362

Feature	Levels	Nr of observations
Age of Firm	1 year	28
	2 years	46
	3 - 4 years	69
	5 - 6 years	75
	7 - 10 years	95
	Longer than 10 years	216
	Other	18
Country	Colombia	101
	Ethiopia	134
	Indonesia	77
	Kenya	76
	Nigeria	159
Industry	Agri	113
	Light Manu	290
	Other	144
Location in the Country	Adama/Mojo	38
	Addis Ababa	66
	Bandung	13
	Barranquilla	36
	Bogotá	41
	Cali	24
	Diredawa	11
	Enugu	40
	Harar	19
	Kaduna	46
	Kisumu	17
	Kwale	26
	Lagos	73
	Makassar	14
	Medan	21
	Nairobi	33
	Yogyakarta	29
Marital Status	Divorced	12
	Married	425
	Single	99
	Widowed	11
Number of Owners	1	422
	2	75
	3	20
	4 or more	29
	Prefer not to answer	1

B Correlation of features

Figure 1 shows a heat map of the correlations between the predictors. Given these correlations the neighborhood stability condition is not broken.

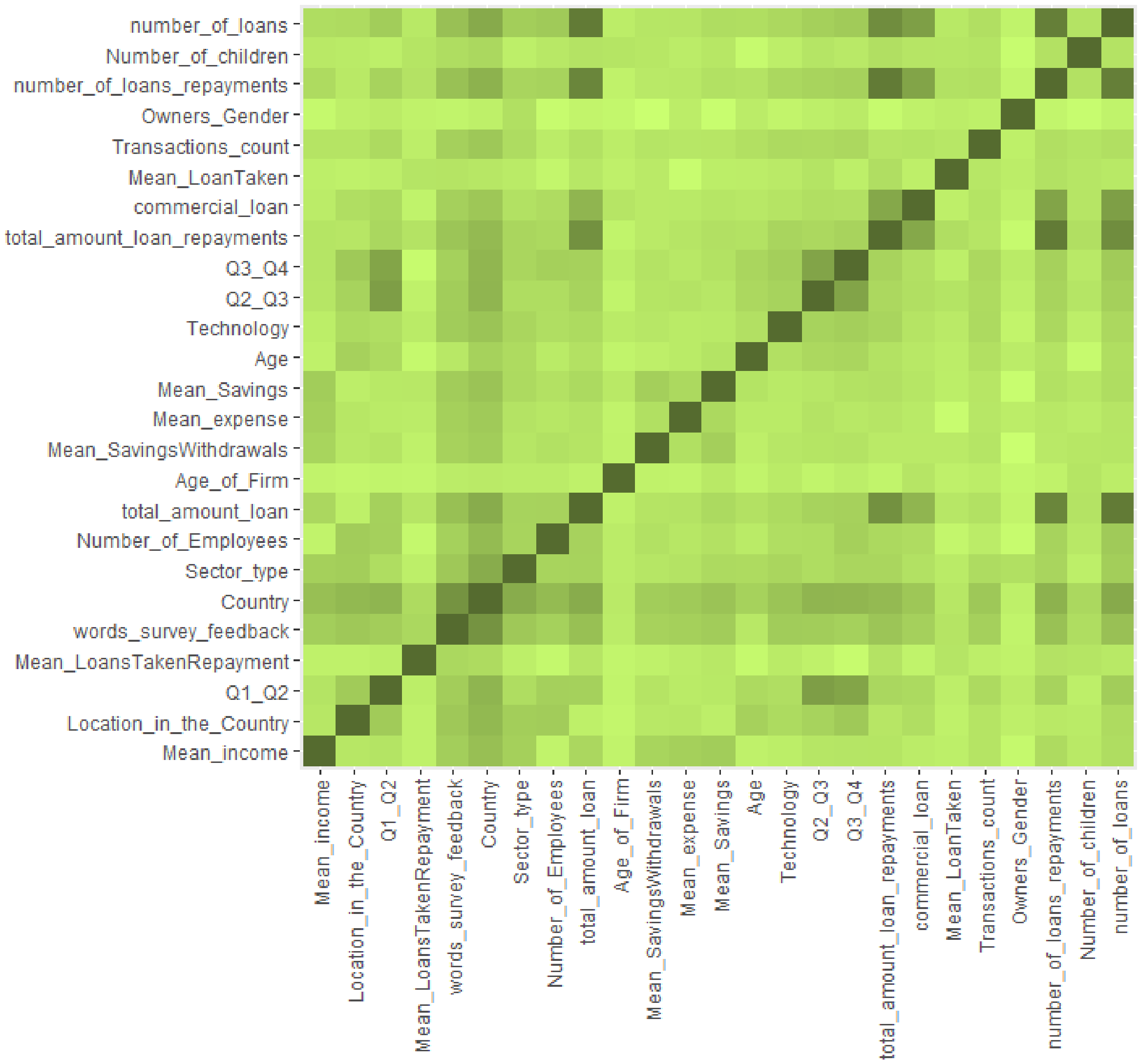


Figure 1: Correlation plot.

C Weight of Evidence transformation

Owners Gender	WOE	IV
All men	0.078	0.003
All women	0.197	0.016
Both men & women	-1.050	0.127

Table 8: WoE Ownwers Gender.

Industry	WOE	IV
Agri	0.241	0.015
Light Manu	-0.259	0.045
Other	0.201	0.057

Table 9: WoE Industry.

Distance walked	WOE	IV
[0:0]	0.058	0.003
[2:60]	-0.657	0.038
NA	0.395	0.012

Table 10: WoE Distance walked.

Commercial loan	WOE	IV
[0:0]	0,165	0.023
[1:1]	-0.963	0.1564

Table 11: WoE Commercial loan.

Nr loans	WOE	IV
[0:1]	-0.162	0.019
[2:3]	0.052	0.001
[4:18]	2.182	0.232

Table 12: WoE Number of loans.

Mean loan repay	WOE	IV
[0:28]	0.280	0.061
[29:99]	-0.619	0.091
[100:245]	-1.350	0.201
[247:34200]	-1.386	0.320
NA	-1.193	0.011

Table 13: WoE Mean loan repayments.

Location Country	WOE	IV
Adama/Mojo	0.000	0.000
Addis Ababa	-0.675	0.074
Diredawa	0.442	0.083
Enugu	1.212	0.375
Harar	-0.134	0.376
Kaduna	0.388	0.405
Lagos	-0.298	0.426

Table 14: WoE Location in the country.

Age Firm	WOE	IV
< 1 year	0.000	0.231
1 year	0.480	0.016
2 years	0.154	0.018
3 - 4 years	0.847	0.142
5 - 6 years	-0.074	0.251
7 - 10 years	-0.251	0.231
> 10 years	-0.916	0.233

Table 15: WoE Age of Firm.

Nr Employees	WOE	IV
[0:3]	-0.494	0.094
[4:5]	-0.368	0.096
[6:6]	0.394	0.119
[7:9]	0.406	0.162
[9:26]	0.666	0.164

Table 16: WoE Nr of employees.

Nr children	WOE	IV
[0:2]	-0.015	0.022
[3:23]	0.016	0.032
NA	0.067	0.014

Table 17: WoE Mean income.

Mean income	WOE	IV
[10.5:44.5]	1.723	0.485
[44.7:64.2]	0.420	0.506
[65.2:111.8]	0.154	0.508
[112.4:144]	0.000	0.508
[145.4:181.6]	-0.174	0.539
[181.9:271.9]	-0.696	0.541
[283.4:400.8]	-0.782	0.544
[401.8:539.6]	-1.350	0.654
[550.5:845]	-1.429	0.657
[858.6:27409.6]	-1.489	0.657

Table 18: WoE Number of children.

Nr loans repay	WOE	IV
[0:1]	0.560	0.007
[2:3]	0.100	0.068
[4:6]	0.000	0.068
[5:7]	-0.693	0.182
[9:40]	-0.728	0.113

Table 19: WoE Number of loan repayments.

Mean expense	WOE	IV
[2.8:20.3]	-0.134	0.002
[21.1:32.2]	-1.350	0.112
[33:48.4]	0.154	0.114
[49.2:76.2]	0.154	0.117
[76.4:100.1]	-0.174	0.120
[102.1:153.7]	0.420	0.140
[154.6:228.9]	0.647	0.193
[232.5:390.3]	0.154	0.196
[399.8:603.8]	0.000	0.196
[608.8:8785.2]	0.377	0.213

Table 20: WoE Mean expense.

Mean Savings	WOE	IV
[10.4:51.2]	0.134	0.002
[52.5:80.8]	0.154	0.004
[84.4:145.7]	0.420	0.025
[146.2:214.3]	0.420	0.045
[214.4:285.8]	0.420	0.066
[286.7:381.8]	0.000	0.066
[386.1:506.6]	-0.154	0.068
[509.6:710.5]	-0.420	0.089
[726.8:1136.7]	-1.350	0.199
[1178:29479.2]	-1.214	0.203

Table 21: WoE Mean savings.

Total loan repay	WOE	IV
[0:209.7]	0.160	0.019
[209.91:371.7]	0.154	0.021
[399.7:1415.78]	-0.619	0.052
[1498.75:34295]	-1.386	0.171
[5:9]	-1.693	0.182

Table 22: WoE Total loan repayments.

Total loan	WOE	IV
[0:114]	0.907	0.000
[141:275.73]	0.847	0.097
[278:459]	0.154	0.099
[478.06:1689]	-0.619	0.130
[1779.5:26649]	-1.386	0.249

Table 23: WoE Total loan.

Sector type	WOE	IV
A	0.036	0.000
AD	-0.496	0.018
AQ	-0.251	0.022
D	0.154	0.024
E	1.946	0.072
G	0.693	0.092
K	1.946	0.187
N	-0.619	0.202
Printing	1.946	0.250
S	0.211	0.253

Table 24: WoE Sector type, this feature contains 110 other levels.

Age	WOE	IV
[18:26]	-0.251	0.004
[27:29]	-0.211	0.008
[30:33]	-0.201	0.016
[34:35]	-0.112	0.118
[36:38]	0.000	0.165
[39:39]	0.634	0.165
[40:42]	0.781	0.185
[43:45]	0.789	0.185
[46:51]	0.853	0.192
[52:80]	0.860	0.200

Table 25: WoE Age.

Mean savings withdraw	WOE	IV
[0:31.8]	0.896	0.106
[33.5:54.2]	0.154	0.109
[54.8:79.9]	0.154	0.111
[80.9:111.4]	-0.124	0.114
[111.8:144.9]	-0.154	0.116
[145.1:191.5]	-0.000	0.116
[192.5:281.7]	-0.647	0.169
[283:400.3]	-0.660	0.190
[401.9:644.3]	-0.662	0.190
[652.1:11350.2]	-0.667	0.255

Table 26: WoE Mean savings withdrawals.

Transactions count	WOE	IV
[7:232]	-0.580	0.026
[237:355]	0.154	0.029
[358:475]	-0.174	0.031
[479:572]	0.197	0.035
[574:699]	0.377	0.052
[700:808]	-0.619	0.083
[809:953]	-0.619	0.113
[958:1208]	0.420	0.134
[1217:1477]	0.154	0.136
[1490:3005]	0.113	0.137

Table 27: WoE Transaction count.

Nr Owners	WOE	IV
1	0.100	0.008
2	0.074	0.009
3	0.000	0.009
4 or more	-0.762	0.034
Prefer not to answer	0.000	0.034

Table 28: WoE Number of owners.

Technology	WOE	IV
[0:0]	0.430	0.006
[1:1]	0.241	0.083
[2:2]	-0.270	0.123
[3:4]	-0.305	0.142
[5:9]	-0.693	0.182

Table 29: WoE Technology.

Q1/ Q2	WOE	IV
[-1:-0.54]	0.511	0.025
[-0.55:-0.39]	0.465	0.054
[-0.38:-0.2]	0.421	0.181
[-0.19:-0.05]	0.197	0.185
[-0.03:0.09]	0.120	0.190
[0.11:0.26]	0.114	0.210
[0.27:0.44]	0.064	0.235
[0.45:0.8]	-0.174	0.354
[0.82:1.43]	-0.214	0.357
[1.44:29.74]	-1.375	0.361

Table 30: WoE Q1/Q2.

Q2/Q3	WOE	IV
[-1:-0.6]	1.134	0.002
[-0.59:-0.35]	0.464	0.026
[-0.34:-0.25]	0.312	0.128
[-0.21:-0.15]	0.214	0.131
[-0.14:-0.03]	0.174	0.135
[-0.02:0.09]	0.134	0.137
[0.1:0.28]	0.116	0.151
[0.29:0.5]	0.054	0.153
[0.51:1.08]	-0.420	0.174
[1.09:12.94]	-0.614	0.178

Table 31: WoE Q2/Q3.

Q3/Q4	WOE	IV
[-1:-0.82]	0.464	0.025
[-0.81:-0.55]	0.420	0.045
[-0.53:-0.3]	0.380	0.071
[-0.29:-0.16]	0.320	0.092
[-0.15:-0.05]	0.241	0.098
[-0.04:0.05]	0.036	0.098
[0.06:0.17]	-0.619	0.128
[0.18:0.39]	-0.691	0.129
[0.4:0.73]	-0.693	0.169
[0.76:29]	-0.814	0.173

Table 32: WoE Q3/Q4.

Words survey feedback	WOE	IV
[0:18]	0.580	0.026
[19:29]	0.452	0.041
[30:40]	0.260	0.049
[41:50]	0.236	0.060
[52:66]	-0.193	0.100
[68:76]	-0.251	0.106
[77:95]	-1.134	0.208
[96:116]	-1.154	0.210
[117:165]	-1.377	0.257
[167:410]	-1.501	0.315

Table 33: WoE Words in survey feedback.

D Optimal thresholds

Table 34: Thresholds that minimize the sum of the error frequencies of the credit scoring methods.

	Threshold
LR forward	0.116
LR backward	0.149
LR lasso	0.295
FA forward	0.139
FA backward	0.350
FA lasso	0.175
EX forward	0.283
EX backward	0.424
EX lasso	0.273
BP forward	0.128
BP backward	0.457
BP lasso	0.326

E Trace plots

Figures 2 to 5 show the trace plots of logistic regression, fuzzy augmentation, extrapolation and bivariate probit with lasso regularisation respectively. A trace plot of lasso is a graphical representation of the regularization path of the coefficients and plotted for different penalty terms λ. As lambda increases, more coefficients are shrunk to zero, leading to a simpler model with fewer variables. We can interpret a trace plot by identifying the value of lambda where each coefficient is first set to zero. This value of lambda corresponds to the point where the model has ”selected” a subset of the features that are most important for predicting the response variable. Additionally, we can use the trace plot to identify which coefficients remain non-zero even as lambda increases, indicating that these features are the most important for the model to include. These figures show that variables such as; *Mean income, Sector type, Q2/Q3* and *Number of employees* are essential features.

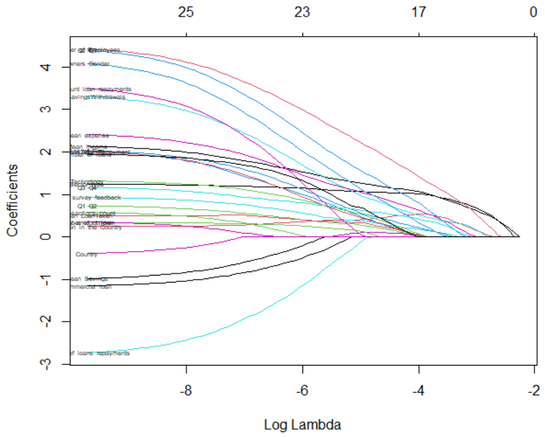


Figure 2: Trace plot of LR with lasso.

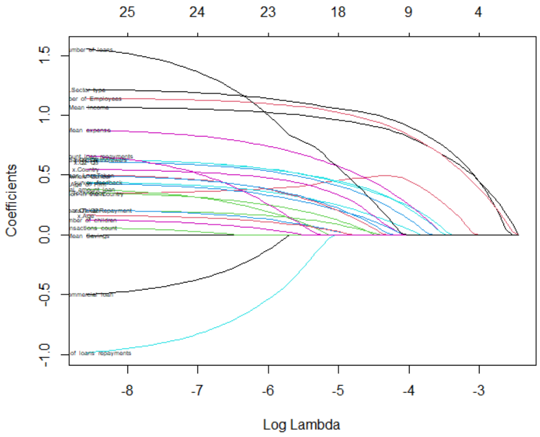


Figure 3: Trace plot of FA with lasso.

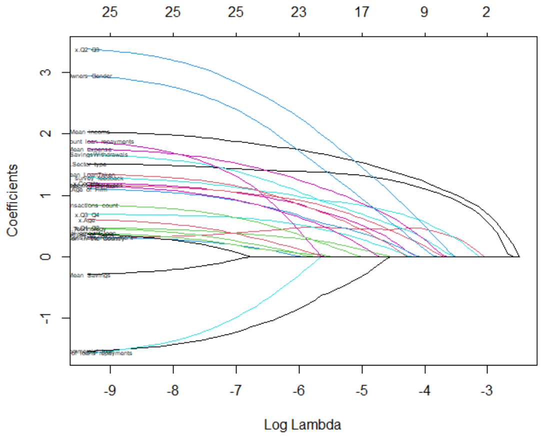


Figure 4: Trace plot of EX with lasso.

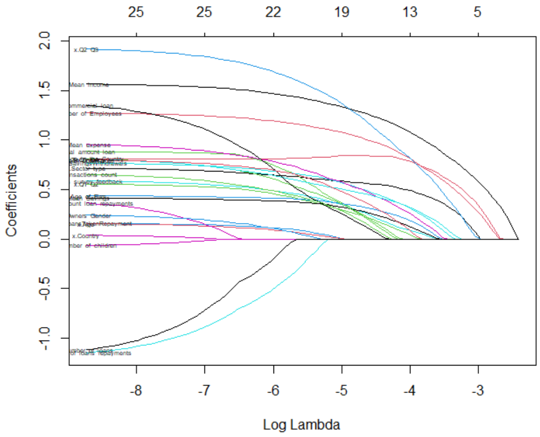


Figure 5: Trace plot of BP with lasso.