Missing Data techniques for Reject Inference

Master's Thesis submitted

by

Radoslav Evtimov
(570341)

in partial fulfillment of the requirements
for the degree of
Master of Science

Chair of Information Systems
School of Business and Economics
Humboldt-Universität zu Berlin

First Examiner: Prof. Dr. Stefan Lessmann
Second Examiner: Prof. Dr. Benjamin Fabian

Berlin, March 7, 2024

Abstract

Reject Inference in credit risk scoring aims at improving the performance of the models used by labeling a subset of the consumer population that would otherwise be isolated from the training process. As the loan repayment datasets only consist of accepts from previous loan application processes, the scoring models end up not being trained on a suitable set of applicants to be representative of the population. Different methods have been proposed for this problem, but more novel approaches from the missing data literature are still to be explored.

The aim of this Master's Thesis is to discover some of the latest developments in the missing data literature and leverage them for a different context - the Reject Inference (RI) problem in the credit risk scoring domain. Through a literature review, the thesis follows the development of the missing data literature and suggests the potential use of newly developed models suiting the Reject Inference sampling bias problem.

It constructs synthetic credit risk data, simulates the credit approval process, and tests techniques for dealing with both MNAR and MAR missingness. Measuring the potential benefits would give a better answer to the question if the missing data literature is applies to the credit risk setting and if yes, aims at answering the question if and which models show a better performance than the conventional RI techniques.

The repository with all scripts used for the thesis as well as results can be found under `https://github.com/evtimovr/missing_data_techniques_RI`.

List of Figures

Contents

1 Introduction

Reject Inference for credit risk scoring proves to be a growingly interesting topic in the credit risk society. Together with the growth in crediting the efforts to improve the credit risk models is a pivotal task possibly resulting in benefits for the financial institutions employing improved models and the economy on its own.

The issue incepting from not observing the outcome of all applicants remains a challenge for credit risk models. Potential borrowers who are not receiving an approval status in the application process never have the chance to be correctly labeled and thus remain a missing data record. Those are usually completely ignored in the subsequent model optimization, resulting in a significant sampling bias - scoring the overall population of applicants based only on training data from a non-random subset of it.

The Reject Inference techniques are the answer to the question of what could be a mitigation measure. However, they are hard to test based on real-life data as achieving complete data requires financial institutions to finance applicants with a much higher probability of default (PD), an extremely costly path that would usually not be pursued as a strategy. This drives the community towards other solutions that require modeling for the missing values and aiming at improving the accuracy of the credit risk models without employing any costly operations. Employing missing data techniques is one of the usual approaches, but with the advances in the missing data literature, several of the newer techniques are to be explored and validated in different domains, one of which is credit risk modeling.

The scope of this thesis is the overview and analysis of several of the existing methods. To achieve this, this thesis uses simulated data.

2 Background

2.1 Missing data mechanisms

Rubin and Little's work ([40, 39]) introduced a framework for missing data mechanisms in an effort to understand how to approach the missing data problem. They have defined three types of missingness depending on the relationship between the missing and observed data. The *Missing Completely at Random (MCAR)* defines a situation where the missing data is completely independent of the observed variables. As per [9] that looks at the missingness mechanism from the Reject Inference perspective.

In a case where Z is the financing mechanism, Y the repayment status, and X contains the features of an application, in the MCAR case, Z is independent from X and Y:

$$\forall \boldsymbol{x}, y, z, p(z \mid \boldsymbol{x}, y) = p(z)$$

Usually, this type of missingness is perceived as harmless and often ignorable. The *Missing at Random (MAR)* mechanism defines a case where the missingness depends only on the observed data points and not on the repayment Y, whereas the *Missing Not at Random (MNAR)* mechanism is the case where Z depends on X and Y.

$$MAR \quad \forall \boldsymbol{x}, y, z, p(z \mid \boldsymbol{x}, y) = p(z \mid \boldsymbol{x})$$

$$MNAR \quad \exists \boldsymbol{x}, y, z, p(z \mid \boldsymbol{x}, y)! = p(z \mid \boldsymbol{x})$$

MAR and $MNAR$ are the subject of serious attention and many methods deal with them as they require treatment and can't be ignored if one wants to avoid biased results.

2.2 Credit risk scoring

The financial institutions and the risks they take play a crucial role in the modern day's economy. They are often in the focus exactly because of the imbalanced risk they take when allocating capital. In the setting of retail banking, a loan is usually allocated in several steps:

1. A customer requests a loan and submits documentation

2. The customer's documentation and externally collected data form a dataset that is used for evaluation

3. A scoring model evaluates the *Probability of Default (PD)* of the customer

4. Based on the results of the model and a given threshold (symbolising the risk appetite) a loan is allocated or not

The people receiving a loan would be called *accepts* and the ones declined - *rejects.*

Because of the crucial role those models play, they are under close supervision from regulators and state authorities. This process prevents banks from unnecessary risks and ensures profitability. From a societal point of view, the credit risk models ensure the non-risky allocation of resources, and from the bank's view they lower the probability of loss. As financial institutions' (except special institutions like state-owned banks) primary goal is profit and would usually prefer losing a customer rather than a failing loan.

2.3 Reject Inference

Emphasizing the importance of the credit risk models, it is clear it is of utmost importance to maintain the models used to a high level of accuracy as often decisions are taken solely on their results. It is important to note that sometimes there are factors not measurable in a quantitative way that influence the decision of the bank. In a physical bank, this might be a subjective impression or opinion by a credit consultant. This might lead to a difference in the features influencing the financing decision and the repayment.

Lenders train the models used only on the outcome data of already approved applicants. This leads to a sample bias and a deterioration of the model performance as the new applicants are scored on models trained on a non-representative population. This sampling bias leads to performance deterioration over time with some customers remaining underfinanced and the banks potentially declining good customers which could provide additional profit. All techniques dealing with this problem are encompassed by the term *Reject Inference (RI). RI* aims to improve the performance of credit scoring models by obtaining the missing repayment labels for the rejected applicants. The assumption is the more available data, the better the performance of the model trained on it.

Some of the pain points and the criticism on Reject Inference as a set of techniques are well summarised by [24]. Empirical evidence about the value of RI and its efficiency is scarce. Most studies only work with accepted cases ([4]), do not possess an unbiased sample with accepts and rejects ([6]) or use synthetic data. This is mainly due to the unavailability of data of this kind. Apart from that, most studies focus on linear models of SVMs (support vector machines), while there are already techniques showing better performance in the field ([26]).

It is important to mention the two different issues Reject Inference is trying to deal with. One of them is model performance - the models using data with sampling bias are meant to perform worse than they could be with complete data. But also model evaluation - the issue of choosing the best model. The performance of models has usually been evaluated on an ”incomplete” dataset which leads to wrong conclusions when making decisions in this regard.

Another important consideration in RI, but also in credit risk scoring in general is the regulatory environment around the used models. They should per regulation be strictly monitored, but also easy to justify and explain whenever requested. The scrutiny of regulators and society pushes financial institutions into a dilemma of what models exactly to use, but often also takes away some of the potential options the research has been looking at. RI techniques, whatever they are, should be able to also withhold this challenge to be practically successful.

2.3.1 MAR vs. MNAR

The question of whether Reject Inference deals with a MAR or an MNAR problem is one with no clear answer. In the general missing data literature, it is fairly hard to estimate what kind of missing mechanism there exactly is and this is often the judgement of experts. In the case of Reject Inference as per [14] the missingness mechanism in the credit scoring depends on what one believes about the classification performance of the used credit scoring models - if the assumption is they perform well, Reject Inference could be classified as a missing data problem between MAR and MNAR as it depends on both Xs and the missing data itself - the Y outcome.

3 Literature review

The following section will encompass the most important developments in the field of missing data research and will only later focus on its application in the reject inference setting.

Missing data techniques develop constantly as the topic is of utmost importance to better understand and tackle, but a recent review of published machine learning applications suggests that the majority of studies use deletion methods for their missing data handling, despite many software packages offering better alternatives [34].

In his book [10] Enders looks at missing data from the perspective of what he calls three analytical pillars of missing data. Those are maximum likelihood, Bayesian estimation, and multiple imputation. This literature review will therefore follow the structure of those three pillars.

3.1 Maximum Likelihood estimation (MLE)

While the origins of maximum likelihood missing data handling date to the 1950s, the first became practically usable in the 1990s when software advancements started allowing its implementation. The classical MLE aims at optimizing parameters so that a minimum squared standardised difference between real-world data and the predictions is achieved. In the missing data use case, a log-likelihood equation is maximized assuming all the observations share the same model parameters while the contribution of each is limited only to the observed data points[11]. Even though the result is not an actual imputation, the location of the missing data point is inferred from the probability estimator.

An important topic around the MLE remains its robustness to nonnormal data. Simulation work on the matter has proven MLE to be consistent in many applications, but missing data could distort approximate fit indices even with normal data [49]. A topic that has developed in the last decade is the maximum likelihood estimators for factored regression specifications[11]. Flexible approaches that accommodate mixtures of categorical and continuous variables in terms of factorisation are available, but there are still not all combinations of metrics available [38, 28].

3.2 Bayesian Estimation

The Bayesian paradigm considers parameters as random variables and the posterior encompasses our subjective knowledge about the potential realisation while collecting more data. The development in the Bayesian estimation techniques used for handling missing data has outpaced the maximum likelihood methods in the last two decades [11]. Initially, the multivariate normal distribution was the predominantly used Bayesian model and the natural fit of Bayesian estimation with factored regression specification makes it a popular solution for this setting.

Interesting advancements in the factored regression specifications setting include latent variable models. [29, 20, 31], models for missing not at random processes [8], auxiliary variable models [7], multilevel models [15, 12]).

3.3 Multiple Imputation

Multiple imputation [39] is a method that includes an imputation phase, an analysis phase, and a pooling phase. First, the values are imputed iteratively, then the imputations are used for analysis as if the preferred method can leverage a complete dataset and finally, the imputations from all iterations are combined to present a single set of results using standard combining rules introduced by Rubin.

Descriptions of the classic procedure are described in multiple research papers including [41]. There is plenty of literature focusing on ways to deal with the imputation of nonnormal data by transformation prior imputation - methods like Box–Cox, logarithmic, square root, inverse, fourth-root, and logit transformations, among others [11, 43, 44, 25]. [11]

Another set of research papers works on generating nonnormal imputations. As an example, the predictive mean matching uses the classic method, but draws from a pool of observations whose prediction is similar to the missing data sample [43, 22].

When talking about categorical missing variables, initially the research community was using the normal imputation models by rounding the continuous imputation value to a discrete value. Nowadays there are much more suitable ways of dealing with this - the use of a latent response framework to adapt binary, ordinal, and multicategorical nominal variables [3, 37].

An important development in the field of multiple imputation is the introduction of the fully conditional specification. The idea here is to use a sequence of univariate regression models to impute variables one at a time instead of using a multivariate distribution [42, 43]). In this case, each regression is tailored to the incomplete variables metric allowing for a diverse collection of generalized linear models [11]. There have been further developments using the framework of Van Buuren and extending it to incomplete covariates, multilevel data structures, classification and regression trees, and regularised regression.

Potential issues with the fully conditional specification are flagged by [5, 27] and mainly focus on the fact that univariate conditional distribution in certain cases cannot derive from the same multivariate distribution. An example of this is when the model features incomplete interactive or nonlinear effects. In a case like this, even if the missing at random condition is satisfied, approaches based on just-another-variable imputation can produce substantial bias (Kim et al., 2015, 2018).[21]

3.4 Missing Not At Random

The literature around handling Missing Not at Random (MNAR) data encompasses mainly two approaches - selection models and pattern mixture models. They both introduce a model that describes the occurrence of missing data in their effort to fight nonresponse bias. The approach of using an indication of missing data is typical for both approaches, but they are used in completely different ways - the selection model uses the indicator as a dependent variable in a regression while the pattern mixture models leverage it as a predictor.

[11] points out selection models and pattern mixture models are underutilized as software tools are no longer a barrier to implementation and researchers have multiple options as it has never been easier to fit those models from a computational point of view. The pattern mixture models are usually used for the analysis of data measured over time to population-level change and individual differences in change characteristics [47].

Wu & Carroll [46] provide a discussion on using one or more latent variables predicting missingness. In recent years there have been graphical methods developed to study parameter recovery. Examples are [32, 33], their work aims at supporting the evaluation if MNAR data analysis can produce meaningful estimates.

A recent contribution to the MNAR data imputation is the work of [36] that develops a Monte Carlo likelihood approach in correcting the bias in parameter estimation using expectation maximization (EM) to overcome computational issues other approaches previously had.

More will be elaborated on two models - [13] and [2] in the following sections. They are good representatives of both primary approaches in dealing with MNAR missingness - selection models and pattern mixture models.

3.5 Multilevel data

Multilevel data is also known as hierarchical or nested data and describes data structure where observations are nested within higher-level units, which leads to dependencies or correlations within the same group. [47]

All three analytical pillars in the missing data literature have been extended in recent decades to accommodate the multilevel data setting. Papers that give a good summary of those developments are [10, 15, 16].

The multilevel setting is usually tackled by random intercept models and the literature suggests that maximum likelihood is the most suitable approach for this [11]. The popular joint model imputation of Joseph Schafer [35] was extended to multilevel data structures. Newer approaches like [3] use incomplete categorical variables as well as missing data at any level of the data hierarchy. Most implementations of the framework only work with a random intercept and cannot account for relationships between the incomplete variables.

Similar to that, there were also recent developments in Van Buuren's fully conditional specification. They are also reserved for random intercept models as it fails to capture heteroscedastic variation in the conditional distribution of random slope predictors [12]. Fully conditional specification and joint model imputation have both shown to be equally effective when applied to random intercept models. A good example of the MNAR missingness model is [17]. The paper leverages what is already developed by [13] and targets a multilevel data setting.

4 Missing data techniques for Reject Inference

The following section aims at presenting models suitable for the Reject Inference setting that are implemented as part of 5. Those are models that represent the main ideas of more advanced techniques for dealing with MNAR data. *MiceMNAR* representing the ideas of a Hacman's selection models and *PPMM* from the family of the pattern mixture models.

4.1 Heckman imputation models for binary outcomes

4.1.1 Multiple imputation using Heckman's one-step ML estimation (MIHEml)

This missing data technique that is coming from the work of Galimard[13].

In a research paper from 2018[13], Galimard et. al develop an algorithm and corresponding R package aiming at leveraging the MICE (multiple imputation by chained equations) method developed by Van Buuren[42] to impute binary MNAR missing outcomes. This setting exactly replicates the reject inference use case because of the binary character of the credit risk outcome variables (0 - low probability of default, 1 - high probability of default).

Their work aims at introducing a method using a sample selection model to impute missing binary outcomes based on a bivariate probit model associated with a one-step maximum likelihood estimator for the first time. Once defined, this model is leveraged in a MICE procedure.

As mentioned in the previous section, there are two main methods for dealing with MNAR missingness. Galimard et. al [13] focus on the first one - selection models. The idea behind Heckman's selection model is that assuming an MNAR missingness, the conditional expectations of the observed and missing Y are different.

$$\begin{array}{l} R'_{yi} = X_i^s \beta^s + \varepsilon_i^s \\ Y'_i = X_i \beta + \varepsilon_i \end{array}, \text{ with } \begin{pmatrix} \varepsilon^s \\ \varepsilon \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

Where R_{yi} is an indicator of the missingness of Y_i. The model uses the idea of a bivariate probit model, modeling the MNAR missingness based on the error terms between the two equations: the selection equation and the output equation. Those error terms follow a bivariate normal distribution.

The missingness mechanism is controlled by the correlation coefficient ρ. In the case where $\rho = 0$ the two equations have independent error terms and indicate a MAR (Missing At Random) missingness. In the opposite case, the missingness R depends on Y and thus implies an MNAR (Missing Not At Random) mechanism.

An important note about the model is Heckman's model must avoid collinearity between the predictors of the two equations, meaning ideally the variables in the two equations would not be exactly the same and the selection equation would have at least one more covariate. An equivalent of this in the Reject inference would imply the selection and the outcome depend on different features of the applicants and a situation like this would be one where there are factors outside of the usual dataset that influence the decision of credit approval.

Under the MAR mechanism, imputation approaches use the conditional distribution of observed Y given the other covariates to impute the missing Y. In the case of MNAR, however, one would need to model this in another way with the conditional expectations of the observed and missing Y being different. For the binary outcome [45]:

$$P\left(Y_i = 1 \mid X_i, X_i^s, R_{yi} = 0\right) = \frac{\Phi_2\left(X_i \beta, -X_i^s \beta^s, -\rho\right)}{\Phi\left(-X_i^s \beta^s\right)}$$

The model for binary outcome missing variables is described in those 3 steps:

The estimated parameter is denoted as $\hat{\theta}$.

The parameter is estimated to be $\theta \sim \hat{\theta}$.

1. Use the one-step estimator to obtain Heckman's model parameters $(\hat{\theta}, \hat{\Psi})$ where $\hat{\Psi}$ is the variance-covariance matrix of $\hat{\theta}$
2. Draw θ^* from $\mathrm{N}(\hat{\theta}, \hat{\Psi})$
3. Draw Y_i^* from a Bernoulli distribution with parameter p_i^* from:

$$\rho_i^* = \frac{\Phi_2\left(X_i\beta^*, -X_i^s\beta^{s*}, -\rho^*\right)}{\Phi\left(-X_i^s\beta^{s*}\right)}$$

4.1.2 Multiple Imputation by chained equations (MICE)

The final step consists of imputing MNAR outcomes and MAR predictors using the MICE framework [42]. The idea of the framework is to start by imputing a random observed value, followed by obtaining the posterior predictive distribution of the first missing variable, and then on an iterated sequence filling in the following missing variables by the already imputed values. In the case of one missing variable, the ”chained equations” part of the method is skipped.

The final output of the model is usually an M number of complete datasets (M represents the number of iterations). The analysis is then usually performed after the ”pooling phase” of MICE, where the results from any analysis on the imputed datasets are combined using Rubin's rules. More details on the exact implementation can be found in the following section.

4.2 Proxy Pattern-Mixture Model

Another method that will be closely looked at in the current study is the ”Proxy Pattern-Mixture Analysis for a Binary Variable Subject to Nonresponse” [2]. This paper looks at extending the already existing work in the proxy pattern-mixture modeling (PMM) to binary variables [1].

The main idea in the research paper is to create an algorithm that could enable the evaluation of the potential impact of nonresponse on survey estimates assuming they could also be MNAR rather than the usual assumption missing survey responses are usually MAR. The developed method enables three different versions for evaluation - Maximum likelihood, Bayesian, and Multiple imputation.

For the Reject Inference use case, we would leverage the multiple imputation (MI) estimation, basing the MI on the specified proxy pattern mixture model (PPMM) to impute the missing values and measure the model improvement using this method. More details on the exact approach will be given in the following section.

As mentioned before, there are two main approaches for modeling MNAR missing data. The selection models, implemented by [13] and already explained in the previous section. The second approach is the pattern-mixture model (PPM).

While selection models factor the joint distribution of M_i and Y_i as

$$f\left(M_i, Y_i \mid Z_i, \theta, \psi\right) = f\left(Y_i \mid Z_i, \theta\right) f\left(M_i, \mid Z_i, Y_i, \psi\right)$$

the case in the pattern mixture models looks differently:

$$f\left(M_i, Y_i \mid Z_i, \xi, \omega\right) = f\left(Y_i \mid Z_i, M_i, \xi\right) f\left(M_i, \mid Z_i, \omega\right)$$

where the first distribution on the right-hand side models the distribution of Y_i given covariates Z_i in the different pattern of missingness M_i and the second models the probabilities of different pattern M_i.

An important difference between the two is that the Selection model, first defined in [18], requires specifying the missingness model via the density, while the pattern-mixture model, which is the basis for the proposed approach, doesn't require any explicit parametric model for the missingness mechanism. The main interest of the work of [1] is to assess the impact on inference for the proportion of units with Y=1 (target variable).

The framework assumes that Y (target variable with missing values) is related to a normally distributed latent variable U, so that Y=1 when $U > 0$. Similar to previous approaches from the PPM frameworks by the covariates Z are reduced to a proxy X. Denoting the missingness indicator by M (M=1 when Y is missing), the regression of Y on Z is:

$$Pr\left(Y = 1 \mid Z, M = 0\right) = \Phi\left(\alpha_0 + \alpha Z\right)$$

Then after a definition of the pattern-mixture model where we assume the joint distribution of [U,X,M] follows the model discussed in

Of primary interest in the overall model is the marginal mean of Y which can be obtained by averaging over the two patterns (m = 0,1). The estimation looks like this:

$$\begin{aligned}
\mu_y = \Pr(Y = 1) = \Pr(U > 0) & \\
& = \Pr(U > 0 \mid M = 0) \times \Pr(M = 0) + \Pr(U > 0 \mid M = 1) \times \Pr(M = 1) \\
& = \pi\Phi\left(\mu_u^{(0)} / \sqrt{\sigma_{uu}^{(0)}}\right) + (1 - \pi)\Phi\left(\mu_u^{(1)} / \sqrt{\sigma_{uu}^{(1)}}\right)
\end{aligned}$$

As one can't estimate the parameters $\mu_u, \sigma_u u, \rho$ for the nonrespondents, the model remains underidentified. A ρ indicates the correlation between the latent variable U and X. This correlation is defined for the respondents (M=0) and is called "strength of the proxy", indicating a stronger prediction power of the covariates Z of Y in the probit model.

In order to obtain $\mu_u^{(1)}, \sigma_{uu}^{(1)}, \rho^{(1)}$ the parameter restrictions from are used, especially the assumption that the M=1 probability is an unspecified function f of the linear combination of X and U:

$$\operatorname{Pr}(Y=1 \mid U, X)=f\left((1-\phi) X^*+\phi U\right)$$

where ϕ is a sensitivity parameter ranging from 0 to 1 that determines the missingness mechanism, the higher the ϕ, the lower the dependency of Y on the proxy X. For example, a ϕ = 0 indicates a MAR missingness, while a $\phi = 1$ indicates strong MNAR mechanism. A ϕ of 0.5 would indicate an equal weight of X and U in their contribution to Y.

An MNAR case would result in missingness that is a function of U, allowing for "smooth" missingness, which may lie on a continuum instead of only taking two values, as would be the case if missingness depended on Y itself.

The final estimates based on the pattern-mixture model are obtained from three techniques popular in the missing data literature - maximum likelihood, Bayesian inference, and multiple imputation. The choice of technique for the simulation study will be discussed in detail in the next section.

5 Simulation study

5.1 Synthetic Data Generation

Reject Inference (RI) as a process concerns credit risk data and naturally, there are very few completely labeled datasets available as few financial institutions would risk their capital to acquire correct labels. This is one of the reasons why research on the topic is primarily done on synthetically generated datasets. In the case of this simulation study, this will also be the case.

Usually, RI simulated data draws the values for *good* applicants (the ones who will repay the loan) and *bad* applicants (the ones who will not repay the loan) from two different Gaussian distributions. In this case, we will make use of a similar approach that was already used in [30], [9] and [19].

$$\begin{cases} x^g \sim \mathcal{N}\left(\mu^g, \Sigma^g\right) & \text{, where } \mu^g \in R^k \text{ and } \Sigma^g \in R^{k \times k} \\ x^b \sim \mathcal{N}\left(\mu^b, \Sigma^b\right) & \text{, where } \mu^b \in R^k \text{ and } \Sigma^b \in R^{k \times k} \end{cases}$$

where x^b and x^g represent the covariates of the bad and good applicants respectively. Assuming their attributes differ in such a way would provide for a simple and reliable analysis. Additionally to that, a holdout set drawn from the same Gaussian distributions is drawn in order to obtain a dataset for performance evaluation.

To perform also a sensitivity analysis, variations of this approach will be tried out. To make the task more complex there is the option to draw the *good* and *bad* samples from a mixture of Gaussian distributions, where the final value for variable X_1, for example, will be the weighted sum of two different Gaussian distributions.

$$\begin{cases} x^g \sim \sum_{i=c}^{C} \pi_c \cdot \mathcal{N}\left(\mu^g, \Sigma^g\right) \\ x^b \sim \sum_{i=c}^{C} \pi_c \cdot \mathcal{N}\left(\mu^b, \Sigma^b\right) \end{cases}$$

where π represents the weight of each of the Gaussian distribution drawn from. Similar to [**?**] two noisy features are added to make the task more complicated. Both belong to the same distribution $x_n \sim \mathcal{N}(0, 1)$.

5.2 Proxy Pattern-Mixture Model

As explained in the previous section, Aldrigde et. al [2] use three different techniques to get bias estimates from their pattern mixture model. The primary purpose of their work was to give a good estimate of how harmful the missing data is for their survey data. The primary focus of this work is slightly different - to test any techniques that might improve the performance of credit risk models by reducing the sampling bias in the dataset used. Hence, the technique that was leveraged by the simulation study here is the multiple imputation.

Even though only a possible step in the original work of [2] multiple imputation could be used to extract actual imputed values based on the pattern mixture model. In more detail, this simulation study uses the R function documented in the original paper which can also be found at https://github.com/randridge/PPMA.

Leveraging the already existing MI (multiple imputation) function based on the pattern mixture model as defined in [2] one could get each imputation value for each of the variables. Because of the difference of the use case, a reformatting is needed to reach a format similar to the one used in the mice package in R. The result that is aimed for is an object with all the imputed values per imputation cycle. Such an object is of class ”mids” - a class specific to the mice package. The ”mids” object allows one to run an analysis of the actual results of an imputation. The analysis this simulation study is aiming for consists of:

- Convert the output of the ”mi” function into a mids object

- Build logistic regressions over each of the imputed datasets
- Use the created models to predict on a test dataset
- Pool the results from each imputation into a single set of results
- Evaluate the created model on the evaluation measure of choice

5.3 Heckman imputation models for binary outcomes (miceMNAR)

The implementation of the model of Galimard et. al [13] has been done with the help of the miceMNAR package in R. The package is currently not live on CRAN, but could be downloaded from the CRAN archive and installed locally.

The package contains multiple functions for both binary and continuous variables. The function of interest, *mice.impute.heckprob*, provides an imputation model for binary outcomes based on the bivariate probit model.

To obtain a result from [13], the before mentioned function was used within a MICE procedure where it was used to impute repeatedly the missing values in the binary outcome variable. To work with the custom Galimard function there is *MNARargument* object that needs to be created. The MNARargument object contains predefined data on which variables should be part of the selection model and which ones part of the output model.

The result of the *MICE* procedure are M separate datasets from each round of imputations. To be able to compare the performance of the method to usual Reject Inference techniques, the final aim of the implementation was a logistic regression trained on the outputs of the MICE procedure. This has been achieved by creating one logistic regression per imputed dataset, running a test dataset on each of the logistic regression, and obtaining M predictions for each target value.

To obtain final predictions and be able to evaluate the model's performance, the predictions are being pooled using Rubin's rules [40]. Usually, this procedure is done on the estimates from the logistic regressions, but in case the same rules were used on the predictions. This way a final estimation of the method could be done similarly to any other model producing predictions.

5.4 Benchmarks

To evaluate the performance of the novel missing data techniques also some of the more familiar approaches used in the Reject Inference will be used as benchmarks.

5.4.1 Oracle model

An oracle model would be a model that can see all the data. This model operates without the missing data issue and has no sampling bias. It would be the higher bound for the performance metric as any Reject Inference technique is trying to come as close as possible to this setup by inferring the actual label of the rejected applicants.

5.4.2 Complete case analysis (CCA)

Complete case analysis encompasses evaluating a scoring model scored only on the complete cases. It is to be used as a benchmark as the usual expectation would be that it performs worse than the tested other more complex techniques

5.4.3 Augmentation

Augmentation, as implemented, can be found in [9]. It is also documented as a “Re-Weighting method” in [4].

Augmentation is based on the fact that applicants with a certain distribution of features appear in the training data disproportionately due to a non-random sample selection [4]. It refers to the techniques that train an additional model that separates accepts and rejects and predicts the probability of acceptance. The probabilities are then used to compute sampling weights for the scoring model [24].

5.5 Acceptance loop

Following the work of there will be an acceptance loop simulated to most realistically evaluate the already mentioned techniques. The idea of the acceptance loop is to simulate the real-life application scoring process. It starts with applications received and scored. Subsequently, the scoring model the financial institution is using would evaluate the customers based on their features (in our case simply simulated as already explained in 5.1). Once the applicants are scored, only the accepted ones will be given a loan. Any following decision will be based on the data collected from the customers with a low probability of default given a loan (*accepts*).

This creates sampling bias and running many iterations of this process could show us what the impact of the issue would be in the long-term iterative process that credit scoring is. The methods already elaborated in 5.3 and 5.2 would aim at correcting the performance by imputing the missing values with the expectation more complete samples would lead to better model performance. For the aim of simplicity and similar to previous research in Reject Inference, the model of choice will be a logistic regression. This is due to its simplicity, application in the industry, and compatibility with the overall research on the topic.

In our case the mitigating measures used will include the already mentioned methods of [13] and [2], as well as augmentation from [9].

The acceptance loop could be summarised in those steps:

$\boxed{1.}$ Initiating a scoring process by using a powerful feature (feature with the largest difference between *good* and *bad* applicants) to rank applications and declining *1-α* of them (α represents the acceptance rate)

$\boxed{2.}$ Obtaining the actual outcome for the approved applicants (default or repayment) and using those to train a classification model to apply to new applications

$\boxed{3.}$ Using the trained models on new incoming applications. Because of the nature of data simulation, all the outcomes are known and any model could be evaluated on the theoretically existing applicants and their actual outcome label

This setup contains several elements: D^r - dataset for the rejected customers, D^a - dataset for the accepted customers, D - a new set of applicants used for evaluation, α - acceptance rate at each iteration, b - bad rate at data generation, μ^g,μ^b,Σ^g, Σ^b - parameters used in the data generation process from 5.1.

Additionally to that there are special parameters to fit the structure of the novel methods applied. As both *MiceMNAR* and *PPMA* use multiple imputation, we use m - the number of imputations for both methods. For the *PPMA*, different values for the ϕ parameter will be used.

In order to estimate the new techniques proposed, they are implemented as part of the acceptance loop. Here is more on how and on what the different models were estimated:

Method	**Implementation**
MiceMNAR	The model is given the D^r with missing values for the binary outcome and D^a with complete values. Those are used to create multiple imputation using Galimard's R package. Multiple glm's fitted to each imputation are then used to impute the values of the D^r data.
PPMA	The model is given the D^r with missing values for the binary outcome and D^a with complete values. Those are used to create multiple imputation based on the PPMA model from [2]. Multiple glm's fitted to each imputation are then used to impute the values of the D^r data.
Augmentation	The model is given the D^r with missing values for the binary outcome and D_a with complete values. Those are used to create multiple imputation based on the augmentation algorithm used in [9]. The output of the function is a logistic regression that is evaluated on the test dataset D.
CCA	The model is given just the accepted cases (D^a). The goal is to replicate the usual credit risk scoring model.
Oracle model	The model is given all the data with complete labels ($D^a \cup D^r$). The goal is to use the best performance possible as a benchmark.

5.5.1 Parameters

The parameters of interest in the acceptance loop simulation are:

D^r - dataset of the rejected customers D^a - dataset of the accepted customers, D - a new set of applicants used for model evaluation,

Results depend on many different parameter values which are rotated in different combinations where the results are obtained and analysed:

Parameter	Values used
α	0.1; 0.2; 0.3
b	0.3
μ^g	(1,0.5,0.5),(1.2,0.6,0.6), (1.5,0.8,0.8)
μ^b	(2,1,1)
ϕ	0; 0.5; 1
cov^g	0.2
cov^bb	0.2; -0.2

where α - acceptance rate; b - bad rate at data generation; n_{train} - size of training set ; n_{test} - size of test set ; μ^g - means of features g population; μ^b - means of features of b population;

Σ^b and Σ^g representing the covariance matrices in the Multivariate Gaussian distribution. Initial values for those are displayed below with only the covariance between the features in Σ^b being adjusted within the Acceptance loop:

$$\Sigma_1^b = \begin{pmatrix} 1 & -0.2 & -0.2 \\ -0.2 & 1 & -0.2 \\ -0.2 & -0.2 & 1 \end{pmatrix} \quad \Sigma^g = \begin{pmatrix} 1 & 0.2 & 0.2 \\ 0.2 & 1 & 0.2 \\ 0.2 & 0.2 & 1 \end{pmatrix} \quad \Sigma_2^b = \begin{pmatrix} 1 & 0.2 & 0.2 \\ 0.2 & 1 & 0.2 \\ 0.2 & 0.2 & 1 \end{pmatrix}$$

5.6 MNAR Simulation study

An additional simulation study was realised with the idea of creating artificially MNAR missingness and measuring the performance of the models on it.

The MNAR Missingness mechanism was implemented based on [48]. It is essentially just picking a random sample to create missing data with a different probability for the different classes based on the eventually missing outcome variable.

Results were insignificant as because of the MNAR missingness process, there was only a very slight difference between the performance of the oracle model and the accepts-only model. The data pattern was easy to recognise by the models. Only when creating extreme MNAR missingness of 0.9 % and 0.6% in the *goods* and *bads* datasets respectively, there were signs of improvements done by the PPMM and miceMNAR. Additional sensitivity analysis was not conducted in favor of the acceptance loop and is thus not presented in 6

This process will be stored for documentation purposes.

6 Results

6.1 Acceptance loop with five features

The acceptance loop 5.5 was used with several different combinations of the main variables in the algorithms tested. Those results were collected and analysed. The software used was R and the functions come from: PPMM from [2], MiceMNAR from [13], Augmentation from [9] as well as the ideas for the overall structure of the acceptance loop and the data generation process from [23]

The idea of the acceptance loop is to understand if the two novel methods reviewed are applicable to the Reject Inference use case. Some caveats that need attention are:

- Synthetic data is very simplistic and adjustments to it influence the results greatly
- The augmentation function needs available data from both classes (0s and 1s) to run. Because of the setup of the acceptance loop, this is initially not the case, this is why in certain cases the model delivered errors
- In certain cases the models have issues finding an optimum
- Both PPMM and miceMNAR could be computationally intensive depending on the number of iterations and batch sizes

The AUC-ROC was measured for each of the models on every 50th iteration due to computation times - at times one single imputation method (PPMA and MiceMNAR) was taking several minutes.

6.1.1 Overall observations

The changes to the data generation parameters lead to the biggest differences in the results in terms of AUC-ROC. When the simplest setup is used with μ^g far from μ^b with difference covariances between the attributes, no model makes a significant difference as there is simply not much to discover. With a simple pattern, each of the models is able to predict correctly in the majority of cases.

The imputation methods perform better in the cases where the *goods* and *bads* are very similar - mainly controlled by the means of the attributes μ^b and μ^b.

In the acceptance loop with 3 attributes drawn from the Multivariate Gaussian distribution, the number of iterations seems to have little to no influence showing steady performance across the 300 iterations.

6.1.2 Change in the α value

The acceptance rate α proves to be one of the significant drivers of performance in the acceptance loop. A higher acceptance rate would essentially reduce the missing data thus reduces significantly the AUC values for all models. In both settings in 1 the novel methods *PPMM* and *miceMNAR* show minimal improvement over the model trained only on the accepts in the later iterations with performance being extremely similar over the initial iterations.

6.1.3 Change in the ϕ value

As already explained in 5.2 ϕ represents the sensitivity parameter of the PPM model, essentially determining the missingness mechanism. A higher */phi* indicates a higher level of MNAR miss-

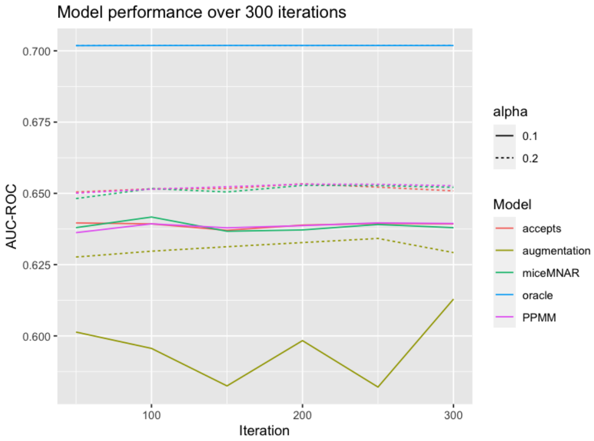

Figure 1: Parameters: μ^g: (1.5,0.8,0.8), μ^b: (2,1,1), α: (0.1,0.2), b:0.3, ϕ: 0.5

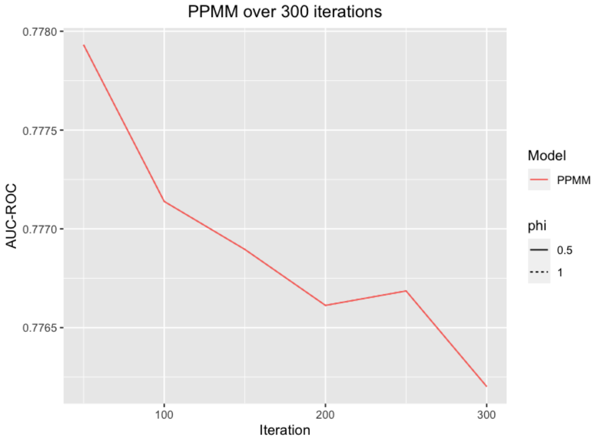

Figure 2: Parameters: μ^g: (1.2,0.6,0.6), μ^b: (2,1,1), α: (0.2), b:0.3, ϕ: (0.5,1)

ingness. 2 shows one of the simulations created with variation of the ϕ value and there is none effect visible. This could be due to the current implementation and the synthetic data used. The analysis of [2] shows that in the case of stronger proxies as defined in the pattern mixture model, ϕ has small to no effect.

6.1.4 Change in mean values (μ)

Adjusting the mean values seems to have a large impact on the model performance. When the difference is very large, especially with 3 variables - which is implemented, so that Galimard works properly. With a diff from (0,0,0) + two noisy features drawn from the same distribution results in almost identical results in all cases. Slight differences are to be observed only compared to the oracle model (Oracle - 98.8 and all other models converge to 98.5 AUC ROC after 300 iterations of the loop)

In the case of smaller differences in the mean values of the Gaussian distribution (μ^g and μ^b), the results show a different behavior. In most of the parameter combinations with those

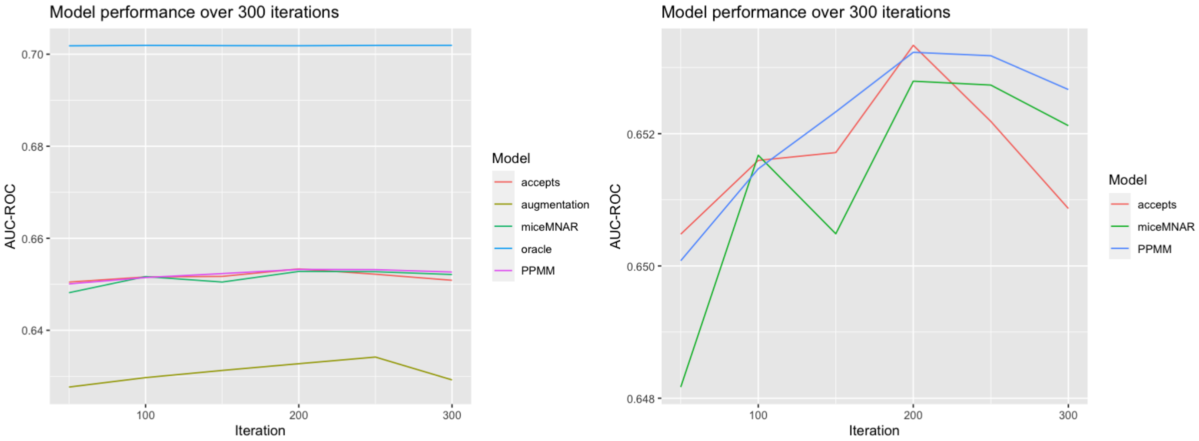

Figure 3: Parameters: μ^g: (1.5,0.8,0.8), μ^b: (2,1,1), α: 0.2, b:0.3, ϕ: 0.5

characteristics *MiceMNAR* and *PPMM* delivered similar results to the model trained only on accepts, while augmentation showed a lower performance. Over 300 iterations the values fluctuate slightly with better AUC-ROC values for the novel techniques of less than 1 percent. 3 shows the development over the 300 iterations (measurement done every 50 iterations) with the graph on the right-hand side showing only 3 of the models with close performance.

6.1.5 Change in covariance matrices

A change in the covariance of the applicant's features is another interesting aspect. The more similar the covariances between the *bads* and *goods* the harder the classification task. This is also visible from 4 where there is a notable decrease in performance when using the same covariances when generating both g and b datasets in the case of the oracle model. Surprisingly the results for the novel methods seem to improve when the similarity of those cases. This is similar throughout different μ and α values.

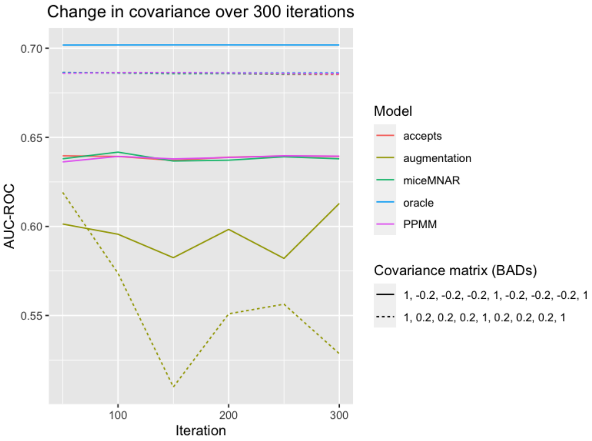

Figure 4: Parameters: μ^g: (1.5,0.8,0.8), μ^b: (2,1,1), α: 0.1, b:0.3, ϕ: 0.5 $cov = (0.2, -0.2)$

The one model standing out is augmentation - augmentation delivers better results when covariances are similar for $\alpha = 0.2$, but worsens its performance when covariances get more similar in the case of $\alpha = 0.1$.

6.2 Acceptance loop with four variables

The same acceptance loop was also implemented with two variables from the multivariate Gaussian distribution and two noisy features. This replicates the initial setup in the implementation of [23]. The results are more meaningful for this setup as because with a simple data generation the models often perform well in recognizing the pattern with few samples and the impact of missing data is limited. Parameters used were similar to the initial version of the loop described in 5.5.1 except the number of variables given to the data generation function - in this case two instead of three in the initial version. There were several interesting observations from this implementation of the acceptance loop.

Interestingly, in 5 Augmentation shows itself as the most successful approach remaining close to the Oracle model even after 300 iterations, while the novel approaches and the model trained on the accepts show a declining performance with the increasing number of iterations. The parameters in this version of the acceptance loop opt in the case of medium mean proximity with $\mu^g = (1.2, 0.6)$ and $\mu^b = (2, 1)$ showing the mean values for X_1 and X_2 respectively.

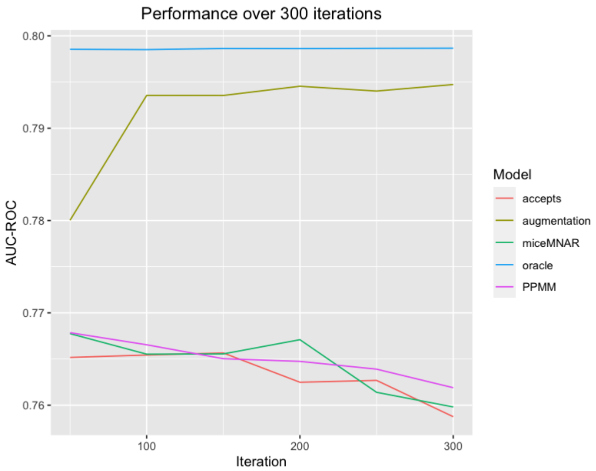

Figure 5: Parameters: μ^g: (1.2,0.6), μ^b: (2,1), α: 0.1, b:0.3, ϕ: 0.5 $cov^b = -0.2$

In another run of the loop visible in 6 *MiceMNAR* and *PPMM* managed to improve the performance over the *accepted* based model while the model corrected by augmentation shows unstable behavior dropping surprisingly in performance in the final measurement (300. iteration).

PPMM and Galimard also performed well on some of the runs with a notable example in

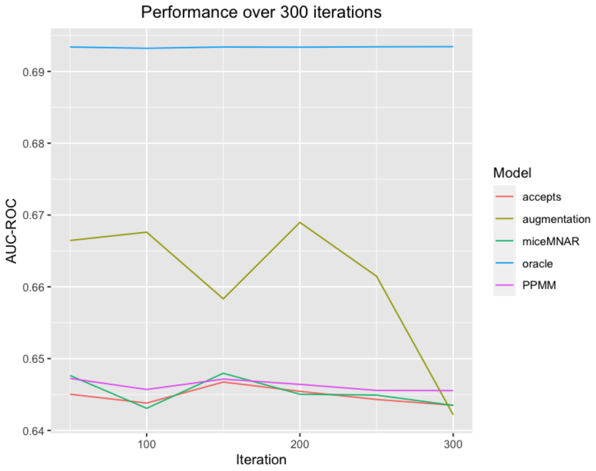

Figure 6: Parameters: μ^g: (1.5,0.8), μ^b: (2,1), α: 0.1, b:0.3, ϕ: 0.5 $cov^b = -0.2$

7. *PPMM* was stable and consistently outperforming the *accepts-trained model*. *MiceMNAR* showed a very unstable performance with serious peaks that are at times outperforming all the other models and coming close to the *Oracle* model. In this particular run the covariance parameter was adjusted to the same values for both g and b populations.

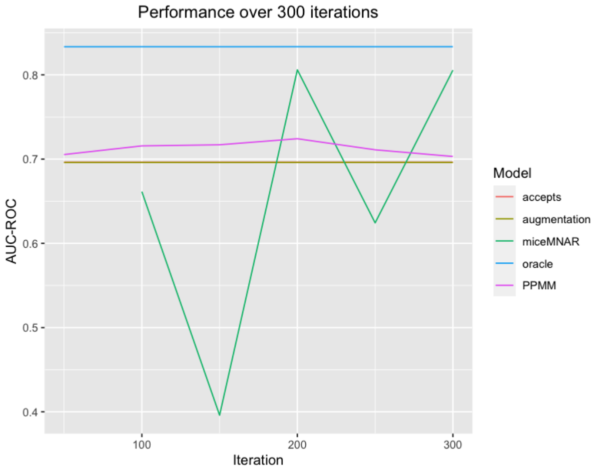

Figure 7: Parameters: μ^g: (1,0.5), μ^b: (2,1), α: 0.1, b:0.3, ϕ: 0.5 $cov^b = 0.2$

7 Conclusion

In this thesis, we conducted an analysis of available novel missing data techniques and aimed at employing them in the context of Reject Inference. Through the usage of synthetic data in a simulation study and a sensitivity analysis, two novel approaches for dealing with MNAR missing data were tested.

The findings uncover the benefits of using more complex missing data techniques in the field of Reject Inference. Some of the used techniques showed promising results in certain very limited settings and their potential application should be a matter of larger research. Even when the results are positive, they often seem unstable and random in the given simulation setting. Often theoretically and practically much simpler techniques that need fewer assumptions reach similar results. Limitations of the study include the limited availability of suitable data and for this reason the use of synthetic data. This may seriously limit the validity of the results. An additional limitation are series of assumptions for the use of those models including a rather theoretical justification for their usage.

Another important note is the fact that those models are to be employed in a regulated environment where stability of results, but also explainability to a rather high degree are required. Future research could look into more of the MNAR techniques and estimate which of them are more realistically usable in the field of Reject Inference.

References

[1] Rebecca Andridge and Roderick Little. Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, 27, 06 2011.

[2] Rebecca R. Andridge and Roderick J.A. Little. Proxy pattern-mixture analysis for a binary variable subject to nonresponse. *Journal of Official Statistics*, 36(3):703–728, September 2020.

[3] Tihomir Asparouhov and Bengt Muthén. Multiple imputation with mplus. *MPlus Web Notes*, 29:238–246, 2010.

[4] J Banasik and J Crook. Reject inference in survival analysis by augmentation. *Journal of the Operational Research Society*, 61(3):473–485, 2010.

[5] Jonathan W Bartlett, Shaun R Seaman, Ian R White, James R Carpenter, and Alzheimer's Disease Neuroimaging Initiative*. Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Statistical methods in medical research*, 24(4):462–487, 2015.

[6] Michael Bücker, Maarten Van Kampen, and Walter Krämer. Reject inference in consumer credit scoring with nonignorable missing data. *Journal of Banking Finance*, 37(3):1040–1045, 2013.

[7] MJ Daniels, Chenguang Wang, and BH Marcus. Fully bayesian inference under ignorable missingness in the presence of auxiliary covariates. *Biometrics*, 70(1):62–72, 2014.

[8] Han Du, Craig Enders, Brian Tinnell Keller, Thomas N Bradbury, and Benjamin R Karney. A bayesian latent variable selection model for nonignorable missingness. *Multivariate behavioral research*, 57(2-3):478–512, 2022.

[9] Adrien Ehrhardt, Christophe Biernacki, Vincent Vandewalle, Philippe Heinrich, and Sébastien Beben. Reject inference methods in credit scoring. *Journal of Applied Statistics*, 48(13-15):2734–2754, November 2021.

[10] Craig K Enders. *Applied missing data analysis*. Guilford Publications, 2022.

[11] Craig K. Enders. Missing data: An update on the state of the art. *Psychological Methods*, March 2023.

[12] Craig K Enders, Han Du, and Brian T Keller. A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and nonlinear terms. *Psychological methods*, 25(1):88, 2020.

[13] Jacques-Emmanuel Galimard, Sylvie Chevret, Emmanuel Curis, and Matthieu Resche-Rigon. Heckman imputation models for binary or continuous MNAR outcomes and MAR predictors. *BMC Medical Research Methodology*, 18(1):90, December 2018.

[14] G Gongyue and Å Thomas. Bound and collapse bayesian reject inference when data are missing not at random.

[15] Simon Grund, Oliver Lüdtke, and Alexander Robitzsch. Multiple imputation of multilevel missing data: an introduction to the r package pan. *Sage Open*, 6(4):2158244016668220, 2016.

[16] Simon Grund, Oliver Lüdtke, and Alexander Robitzsch. On the treatment of missing data in background questionnaires in educational large-scale assessments: An evaluation of different procedures. *Journal of Educational and Behavioral Statistics*, 46(4):430–465, 2021.

[17] Angelina Hammon and Sabine Zinn. Multiple Imputation of Binary Multilevel Missing not at Random Data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 69(3):547–564, June 2020.

[18] James J Heckman. 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. NBER, 1976.

[19] Frank Hoffmann, Bart Baesens, Christophe Mues, Tony Van Gestel, and Jan Vanthienen. Inferring descriptive and approximate fuzzy rules for credit scoring using evolutionary algorithms. *European journal of operational research*, 177(1):540–555, 2007.

[20] Brian T Keller and Craig K Enders. Blimp 3 user's guide (draft 6.29. 2021).

[21] Soeun Kim, Catherine A Sugar, and Thomas R Belin. Evaluating model-based imputation methods for missing covariates in regression models with interactions. *Statistics in medicine*, 34(11):1876–1888, 2015.

[22] Kristian Kleinke. Multiple imputation under violated distributional assumptions: A systematic evaluation of the assumed robustness of predictive mean matching. *Journal of Educational and Behavioral Statistics*, 42(4):371–404, 2017.

[23] Nikita Kozodoi. Fighting the sampling bias: A framework for training and evaluating scoring models. Unpublished preprint 2024.

[24] Nikita Kozodoi, Panagiotis Katsas, Stefan Lessmann, Luis Moreira-Matias, and Konstantinos Papakonstantinou. Shallow self-learning for reject inference in credit scoring. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part III*, pages 516–532. Springer, 2020.

[25] Katherine J Lee and John B Carlin. Multiple imputation in the presence of non-normal data. *Statistics in medicine*, 36(4):606–617, 2017.

[26] Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, and Lyn C. Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136, 2015.

[27] Yu Liu and Craig K Enders. Evaluation of multi-parameter test statistics for multiple imputation. *Multivariate Behavioral Research*, 52(3):371–390, 2017.

[28] Oliver Lüdtke, Alexander Robitzsch, and Stephen G West. Analysis of interactions and nonlinear effects with missing data: a factored regression modeling approach using maximum likelihood estimation. *Multivariate Behavioral Research*, 55(3):361–381, 2020.

[29] Oliver Lüdtke, Alexander Robitzsch, and Stephen G West. Regression models involving nonlinear effects with missing data: A sequential modeling approach using bayesian estimation. *Psychological methods*, 25(2):157, 2020.

[30] Sebastián Maldonado and Gonzalo Paredes. A semi-supervised approach for reject inference in credit scoring using svms. In *Advances in Data Mining. Applications and Theoretical Aspects: 10th Industrial Conference, ICDM 2010, Berlin, Germany, July 12-14, 2010. Proceedings 10*, pages 558–571. Springer, 2010.

[31] Edgar C Merkle and Yves Rosseel. blavaan: Bayesian structural equation models via parameter expansion. *arXiv preprint arXiv:1511.05604*, 2015.

[32] Karthika Mohan and Judea Pearl. Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534):1023–1037, 2021.

[33] Karthika Mohan, Judea Pearl, and Jin Tian. Graphical models for inference with missing data. *Advances in neural information processing systems*, 26, 2013.

[34] Swj Nijman, Am Leeuwenberg, I Beekers, I Verkouter, Jjl Jacobs, Ml Bots, Fw Asselbergs, Kgm Moons, and Tpa Debray. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. *Journal of Clinical Epidemiology*, 142:218–229, 2022.

[35] Maren K Olsen and Joseph L Schafer. A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data. *Journal of the American Statistical Association*, 96(454):730–745, June 2001.

[36] Jiaxu Peng, Jungpil Hahn, and Ke-Wei Huang. Handling Missing Values in Information Systems Research: A Review of Methods and Assumptions. *Information Systems Research*, 34(1):5–26, March 2023.

[37] Matteo Quartagno and James R Carpenter. Multiple imputation for discrete data: evaluation of the joint latent normal model. *Biometrical Journal*, 61(4):1003–1019, 2019.

[38] Sophia Rabe-Hesketh, Anders Skrondal, and Andrew Pickles. Generalized multilevel structural equation modeling. *Psychometrika*, 69(2):167–190, 2004.

[39] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[40] Donald B Rubin. Multiple imputation for survey nonresponse, 1987.

[41] Joseph L. Schafer and John W. Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177, 2002.

[42] Stef Van Buuren. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3):219–242, June 2007.

[43] Stef Van Buuren. *Flexible imputation of missing data*. CRC press, 2018.

[44] Paul T Von Hippel. How to impute interactions, squares, and other transformed variables. *Sociological methodology*, 39(1):265–291, 2009.

[45] Karsten Webel. Greene, w. h., econometric analysis: Prentice hall, new jersey, 2008, 6th edition, xxxvii + 1178 pp., £56.99, isbn 978-0-13-513740-6. *Statistical Papers*, 52(4):983–984, 2011.

[46] Margaret C Wu and Raymond J Carroll. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, pages 175–188, 1988.

[47] Shu Xu and Shelley A. Blozis. Sensitivity Analysis of Mixed Models for Incomplete Longitudinal Data. *Journal of Educational and Behavioral Statistics*, 36(2):237–256, April 2011.

[48] Xijuan Zhang. How to generate missing data for simulation studies. *The Quantitative Methods for Psychology*, 19:100–122, 2023.

[49] Xijuan Zhang and Victoria Savalei. New computations for rmsea and cfi following fiml and ts estimation with missing data. *Psychological Methods*, 28(2):263–283, 2023.

Declaration of Academic Honesty

I, Radoslav Evtimov, hereby declare that I have not previously submitted the present work for other examinations. I wrote this work independently. All sources, including sources from the Internet, that I have reproduced in either an unaltered or modified form (particularly sources for texts, graphs, tables and images), have been acknowledged by me as such. I understand that violations of these principles will result in proceedings regarding deception or attempted deception.

Radoslav Evtimov
Berlin, 07.03.2024