Charles University

Faculty of Social Sciences
Institute of Economic Studies

MASTER'S THESIS

Performance Ranking of Czech Credit Scoring Models

Author: **Bc. Peter Smolár**

Supervisor: **doc. PhDr. Tomáš Havránek, Ph.D.**

Academic Year: **2019/2020**

Declaration of Authorship

The author hereby declares that he compiled this thesis independently; using only the listed resources and literature, and the thesis has not been used to obtain a different or the same degree. This thesis consists of 107 086 characters, excluding its abstract, bibliography and addenda.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, July 27, 2020

Peter Smolár

Signature

Acknowledgments

First and foremost, I would like to thank my supervisor doc. PhDr. Tomáš Havránek, Ph.D. for his patience, thoughtful insights and professional guidance. Furthermore, I would like to thank my family for all the support they have been more than willing to give. Last but not least, I would also like to thank my great friend Miroslav Jakab.

Abstract

This thesis provides a comprehensive ranking of 11 Czech statistical and 4 foreign credit scoring models. The ranking is based on the predictive performance of individual models, as measured by the area under curve, evaluated on a randomly sampled set of 250 training and validation samples. After establishing a baseline comparison, 3 avenues of estimation setup optimization are explored, namely missing value treatment, estimation method and the use of additional non-financial variables. After being optimized, the models are once again ranked based on their predictive performance. Statistical inference is drawn using ANOVA and the Friedman test, along with the corresponding Tukey and Nemeyi pos-hoc tests. In their baseline form, the Czech credit scoring models are found to be outperformed by the foreign benchmark model. Treating the missing values by OLS imputation and estimating the models by probit, significantly is found to significantly improve their predictive performance. In their optimized form, the difference in predictive performance between Czech and foreign credit scoring model is found to be only marginal.

JEL Classification	G28, G32, G33, G38
Keywords	credit scoring, multiple discriminant analysis, logit analysis, probit analysis

Author's e-mail	71247263@fsv.cuni.cz
Supervisor's e-mail	tomas.havranek@ies-prague.org

Abstrakt

Tato práce srovnává 11 českých a 4 zahraničních kredit skóringových modelů. Toto srovnání je založeno na schopnosti jednotlivých modelů předpovídat úpadky firem na měřené plochou pod ROC křivkou. Srovnání se zakládá na sadě 250 trénovacích dat a testovacích dat. Na základě tohoto základního srovnání, tato práce zkoumá 3 potencionální způsoby zlepšení predikčních schopností skóringových modelů, a to konkrétně metody nahrazení chybějících hodnot, různé statistické metody uplatněné při odhadu modelu a možnost přidání nefinančních proměnných. Po aplikaci těchto způsobů, predikční schopnost modelů byla opět vyčíslena a modely srovnány. Tato práce používá ANOVA a Friedman test, jakož i jim odpovídající post-hoc Tukey a Nememyi testy pro testování hypotéz. Ve své základní podobě jsou zahraniční modely

lepší než jejich české protějšky v predikování úpadku společností. Nahrazení chybějících hodnot pomocí OLS a odhad modelů za pomoci probit signifikantně zlepšuje predikční schopnosti srovnaných modelů. Po aplikaci těchto zlepšení je rozdíl v predikčních schopnostech českých a zahraničních modelů marginální.

Klasifikace	G28, G32, G33, G38
Klíčová slova	kredit skóring, diskriminanční analýza více proměnných, logit analýza, probit analýza

E-mail autora	71247263@fsv.cuni.cz
E-mail vedoucího práce	tomas.havranek@ies-prague.org

Contents

List of Tables

List of Figures

Acronyms

AIES	Artificial intelligence expert system
ANOVA	Analysis of variance
APCP	Average probability of correct prediction
AUC	Area under curve
CNB	Czech National Bank
CSM	Credit Scoring Model or Credit Scoring Models
CSO	Czech Statistical Office
ICO	Identification number of firms registered in the Czech Republic
MDA	Multivariate discriminant analysis
ROC	Receiver operating characteristic curve
SME	Small and medium-sized enterprises

Master's Thesis Proposal

Author:	Bc. Peter Smolár
Supervisor:	doc. PhDr. Tomáš Havránek, Ph.D.
Defense Planned:	September 2020

Proposed Topic:

Performance Ranking of Czech Credit Scoring Models

Motivation:

Although Czech credit scoring literature is particularly rich, it is in dire need of a comprehensive empirical overview. It does not take much effort to create a new, original statistical credit scoring model. One only needs a fresh data sample and few as of yet unidentified significant variables. As a consequence, a recent overview of Czech papers on the topic by Prusak (2018) lists at least 19 distinct mixed industry models. Authors of these models often boast about their superior performance compared with a chosen foreign credit scoring benchmark model such as Altman's Z-score. The few papers which actually do pit relevant Czech credit scoring models against each other (see for example Němec and Pavlík (2016), or Machek (2014)) do so unsystematically. They mostly select only a small number of such models, use various cut-off values and their datasets usually span just a few years. Moreover, these papers often present a new model of their own which could make one doubt the honest intentions of the authors.

The primary goal of this paper is therefore to provide an unbiased and comprehensive long-term performance ranking of Czech credit scoring models. The models should thus be simultaneously pitted against Czech and foreign credit scoring models, the latter justifying the usage of a particular Czech model. The performance of the models will be evaluated using established methods for credit scoring models ranking.

The scope of the performance ranking will be limited by several criteria. As for the data only models based on financial statements will be taken into consideration. As for the techniques, this paper will only be concerned with statistical techniques. Excluding theoretical models can be justified by their absence in the Czech literature. Exclusion of artificially intelligent expert system (AIES) models is imposed by the technical skills required to set up such models. As for the industry scope, the paper will be limited to mixed industries credit scoring models with the aim of making the results comparable. Using credit scoring models specific to individual industries would greatly reduce their comparability, due to a low number of Czech models existing for one given industry.

All the secondary goals are set out conditional on the overall extent of the paper. The first secondary goal is to optimally calibrate the explored scoring models given the observed data. The extent of this calibration could stretch from optimal model cut-off values and optimal model parameters' re-estimation frequency to optimal timespan for the estimation dataset. The second secondary goal is to extend the

paper to include not only statistical but also artificially intelligent expert systems techniques.

Hypotheses:

1. **Czech models outperform foreign benchmark models:**
   The crux of this paper lies in ranking the Czech credit scoring models. However, as Machek (2014) shows in his article, out of 4 Czech credit scoring models included in his study, only 1 beats the general benchmark Altman's Z-score model. A fundamental case-by-case question which this hypothesis aims to answer is therefore whether there even is a need for Czech Republic specific models.
2. **Calibration increases the performance of the models:**
   The models predict bankruptcy based on a comparison of score function value with a cut-off value. Taking the familiar example, in his original article Altman (1968), determines that the lowest number of bankruptcy misclassifications are obtained for a Z-score value of 2.675. For practical purposes such cut-off values are often taken at face value and copied from the original study. The problem with this is that these cut-off values are inherently data dependent. One could therefore make a case for periodic model calibration. Apart from affirming this hypothesis and in order to achieve a practical implementation, the study will be concerned with the magnitude of the increase in performance as well as the magnitude of calibration changes.
3. **Hypothesis 3: AIES techniques outperform statistical techniques**
   AIES models are harder to set up than statistical models. Apart from personal experience of the author, this fact is well illustrated by their proportional representation the overview study of Aziz and Dar (2006). Czech credit scoring literature mirrors this trend and AIES models form only a small minority of credit scoring models. For a practitioner, the motivation to set up an AIES model, can only stem from its superior performance. This superior performance was confirmed by Aziz and Dar (2006). It remains to be seen whether a different methodology and a selection of a single country will lead to identical results.

Methodology:

When it comes to evaluating the performance of different models, we will follow the framework developed by Jackson and Wood (2013). The authors are able to compare 25 different credit scoring methods. The compared models are based on statistical, theoretical and artificially intelligent expert systems techniques. Instead of minimizing type I and type II errors, the authors use receiver operating characteristic curves to score the models. Doing so allows them to eliminate the impact of arbitrary cut-off choices, which need to be made in real life application of the models, by maximizing the performance over all potential cut-off choices. For optimal calibration, the same methodology will be used in an iterative fashion.

Expected Contribution:

This paper aims to enrich both academia as well as finance practitioners. For academic purposes, its aims to evaluate the added value of developing new Czech credit scoring models. If Czech credit scoring models are not able to consistently beat foreign benchmark models, there is very little justification for their existence. As for the practitioners, the paper aims to give an answer as to which model to select with regards to overall performance, stability of results and ease of use, thus making sense of the mess which is credit scoring literature in the Czech Republic.

Outline:

1. Introduction: The chapter will introduce the topic of credit scoring as well as clarify the motivation behind this paper.
2. Literature review: The chapter will be introduced by a brief review of international credit scoring literature. The aim is to select benchmark models for comparison and also to create a framework for cataloguing the Czech models. After this, a structured catalogue of Czech credit scoring models will then delimit the overall scope of the empirical part of this paper.
3. Methodology and data: As described before, this paper will use the framework developed by Jackson and Wood (2013) to rank the performance of individual models. In order to be well replicable, the data should be either publicly or at least widely accessible. One such source is the Magnus Web database, used in several empirical studies on the topic.
4. Results: The chapter will contain a discussion over results of the hypothesis testing.
5. Concluding remarks: The chapter will present a summary of the main findings with their practical use in mind.

Core Bibliography:

1. Jackson, R. and Wood, A. (2013). *The performance of insolvency prediction and credit risk models in the UK: A comparative study*. The British Accounting Review, 45(3), pp.183-202.
2. Machek, O. (2014). *Long-term Predictive Ability of Bankruptcy Models in the Czech Republic: Evidence from 2007-2012*. Central European Business Review, [online] 3(2), pp.14–17.
3. Jakubik, P. and Teplý, P. (2008). *The Prediction of Corporate Bankruptcy and Czech Economy's Financial Stability through Logit Analysis*. Working Papers IES 2008/19, Prague: Institute of Economic Studies
4. Prusak, B. (2018). *Review of Research into Enterprise Bankruptcy Prediction in Selected Central and Eastern European Countries*. International Journal of Financial Studies, 6(3), p.60.
5. Němec, D. and Pavlík, M. (2016) *Predicting insolvency risk of the Czech companies*. Proceedings of the International Scientific Conference Quantitative Methods in Economics, Bratislava: University of Economics
6. Adnan Aziz, M. and Dar, H.A. (2006). *Predicting corporate bankruptcy: where we stand?* Corporate Governance: The international journal of business in society, 6(1), pp.18–33.

1 Introduction

For a country of its size and age, credit scoring literature in the Czech Republic is particularly rich. A recent overview of Czech papers on the topic by Prusak (2018) lists at least 19 distinct papers, introducing new original credit scoring models (CSM). Authors of these models often boast about their CSM having superior performance over a chosen foreign credit scoring benchmark model such as Altman's Z-score. The few papers that pit relevant Czech credit scoring models against each other, like (Machek, 2014) or (Němec and Pavlík, 2016), do so unsystematically. Although badly needed, as of writing of this thesis, no comprehensive predictive performance ranking exists. As explained in further detail, CSM often suffer from being overfitted and data sensitive. For these reasons, a comparison based purely on the results reported in original papers is not sufficient. Any reasonable comparison must be derived using a common, representative dataset. This thesis aims to fill this hole in the Czech CSM literature. Providing a comprehensive prediction performance ranking of both Czech and foreign benchmark CSM helps to assess the justification for developing and using CSM specific to the Czech Republic. Moreover, the thesis explores avenues, in which the estimation setup can be optimized in order to increase the overall predictive performance of the compared models. Doing so allows to judge the performance of the CSM more objectively, without needing to repeat the arbitrary choices made by the model author. The impact of these choices on the prediction performance is then judged by comparing the baseline models with their optimized counterparts.

This thesis compares the predictive performance of 11 Czech and 4 foreign statistical general industry non-financial CSM. The compared models were re-estimated using a set of 250 training samples and their predictive performance was then evaluated on a set 250 validation samples. The compared models were then ranked based on their respective distributions of the prediction performance indicator. At its crux, the comparison of CSM aimed to be as general and as objective as possible. Any source of data sensitivity of the results must have been minimised. Any factors introducing arbitrariness into prediction performance must have been avoided. In order to reduce the impact any individual firm would have on the results, a dataset 60 times larger than that of any other Czech CSM paper was used. Since macroeconomic conditions (Mensah, 1984), industry distribution (Platt and Platt, 2002) and bankruptcy prevalence (Zmijewski, 1984) all affect the predictive performance of CSM, the dataset was constructed to span long enough, to be representative enough and balanced enough to

provide the most truthful image of the Czech non-financial market. To avoid overfitting of the models, as well as reliance on a single result rather than their distribution, the procedure described in (Jackson and Wood, 2013) was followed. Accordingly, in the process of random sampling, the dataset was repeatedly split into disjoint training and validation samples. To decrease the data sensitivity of the results even further, a third exclusion sample was introduced into the process. Also, rather than randomly discarding multiple observations per firm, the thesis kept all the available information. Area under curve, the prediction performance indicator employed by (Jackson and Wood, 2013) and this thesis, was also considered to be a superior indicator when compared with the more popular alternatives, which all require that a researcher selects an arbitrary "cut-off" value.

In the baseline comparison, Ohlson logit model contained in (Ohlson, 1980) was found to provide the best predictive performance, having a median AUC of 78.9 %. On average, foreign benchmark models outperformed the Czech CSM by 4.2 % AUC. In general, using Czech CSM therefore seems hardly. In order to improve the predictive performance of the compared CSM, 3 avenues of estimation setup were explored. The imputation of missing values using OLS was found to increase the predictive performance of the compared CSM on average by 3.2 % AUC. Similarly, using probit estimation instead of multivariate discriminant analysis (MDA) improved the predictive performance of CSM by 2.2 % AUC. Addition of non-financial variables was also analysed, but none of the compared variables increased the predictive performance over the baseline. With the best performing optimization methods at hand, the compared models were all re-estimated using OLS imputation and probit estimation. The resulting ranking saw the JT index model, developed in (Jakubik and Teplý, 2011), rise to the top with the median AUC of 86.6 %. Since a Czech model became the performing CSM and the gap between Czech and foreign CSM narrowed to only 0.6 %, the usage of Czech CSM in their optimized form could finally be justified.

The thesis delivers on its promise of being a contribution to academic research as well as of being a contribution to financial practitioners. From the academic standpoint, the thesis affirms that the theoretical superiority of conditional probability models over MDA, as derived by (Ohlson, 1980), translates into higher predictive performance. Moreover, it points out that in the context of Czech Republic, the issue of missing value treatment, which up until now received little attention, has serious negative impact on the predictive performance. On top of this high-level observation, the study also provides an exploratory analysis of different missing value treatments. The study may also prove useful to practitioners, trying to the select the highest performing Czech

CSM or simply trying to increase the predictive performance of their CSM by estimation setup optimization.

The remainder of this study is organised thusly. Chapter 2 presents the Czech and foreign CSM literature overview. The ultimate goal of this chapter was to delimit the scope of the comparison, with respect to the Czech and foreign CSM. Chapter 3 the discusses the dataset employed in this study. Apart from familiarising the reader with the dataset construction, dataset composition, employed data sources and explanatory variables, it also presented the many ways in which the representativeness of the dataset was verified. Chapter 4 then presents most important methodological aspects of the thesis. Special attention is given to the random sampling, predictive performance indicators and the statistical inference used in hypothesis testing, since all of these aspect differed in some way from the usual CSM literature. Chapter 5 then contains the results of the CSM ranking, discusses the potential performance optimizing treatments and ranks the CSM in their optimized form. Finally, Chapter 6 summarizes the aim of the thesis, the results as well as its contribution in a few concluding remarks.

2 Literature overview

For the purposes of this thesis, literature overview serves 3 main goals and is therefore split into 3 subsections. First, a common framework for classifying credit scoring models (CSM) needs to be developed. Second, for each class of models, representative foreign benchmark CSM need to be selected. Including foreign benchmark allows to debate whether the existence of CSM specific for the Czech Republic is justified. Third, a structured overview of Czech CSM is provided. Models are divided depending on methods applied and based on industry specificity. After exclusion of irrelevant articles, the overall scope of this thesis can be delimited.

2.1 Model classification framework

A common CSM classification divides the models into 3 classes – statistical methods, artificial intelligence expert systems (AIES) and theoretical methods (Aziz and Dar, 2006). A higher degree of detail is applied to statistical methods, also called parametric or classical class of methods, as these form the crux of the thesis. In developing the classification framework, it is useful to follow the evolution of individual methods as doing so provides the broader motivation for their introduction as well as their critique.

Early research into corporate bankruptcy began in the first half of the 20th century. A comprehensive overview of this stage which spanned from 1930 to 1965 can be found in (Bellovary et al., 2007). The early studies concentrated on identifying indicators of corporate bankruptcy in the form of financial ratios. Although, some of these studies introduced industry benchmarks, with which the individual values of financial indicators could be compared, the usefulness of these ratios in a predictive setup was limited.

The first generally recognised CSM was coined in (Beaver, 1966). Unlike his predecessors, Beaver actually tested the predictive abilities of financial ratios in differentiating between non-bankrupt firms and bankrupt firms. Since the prediction was made on a single financial ratio, this statistical class of CSM was dubbed univariate or single variable models. In his original paper Beaver identifies inclusion of more financial ratios as an avenue for further research. As a response, two ways to approach this problem and therefore two new statistical methods arose.

On one hand, the lesser known risk index models were introduced in (Tamari, 1966). Individual companies are assigned a score on a scale between 0 and 100 based on weights assigned to different financial ratios. Higher score is then equivalent to better financial health. One of the critiques of this approach was that these weights were subjective (Balcaen and Ooghe, 2006). On the other hand, multivariate discriminant analysis (MDA), popularised in a landmark article by (Altman, 1968), became the most employed CSM method in the empirical literature (Jackson and Wood, 2013). The ease of estimation of MDA is a likely the reason for its popularity to this day. Instead of subjective weights, MDA uses the coefficients extracted from a bankruptcy model estimated using least squares. In the predictive context, the fitted values are simply compared with a predetermined "cut-off" value in order to classify a firm as healthy or bankrupt. Although widely used, MDA suffers from several shortcomings. It is based on a set of restrictive model assumptions, namely normal distribution of the independent variables and identical variance-covariance matrices of bankrupt and healthy firms (Balcaen and Ooghe, 2006).

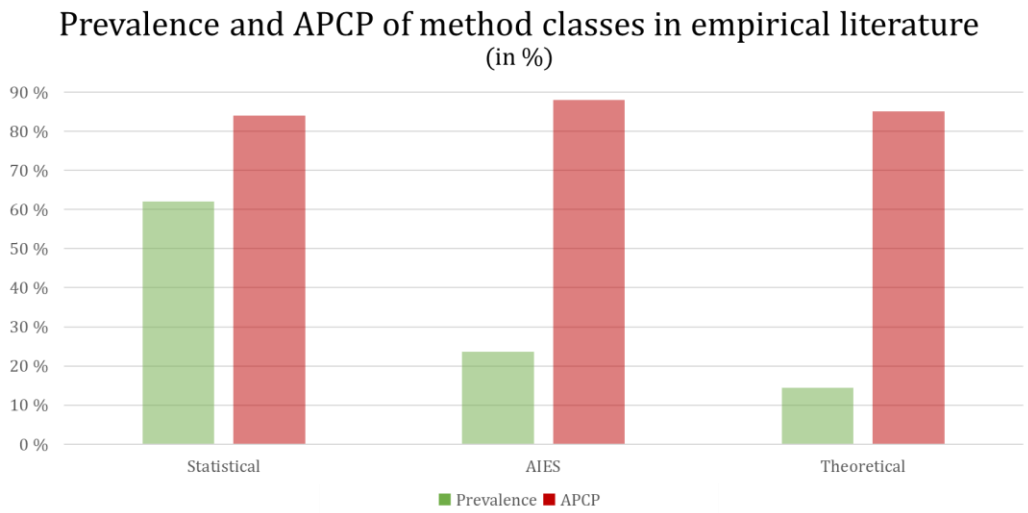
Aforementioned shortcomings of the MDA were addressed by conditional probability models. These models employ a cumulative distribution function to projection the dependent variable of the latent regression into a 0 to 1 space. Even though the cumulative distribution function may take many functional forms, it was the logistic cumulative distribution function and hence logit model which became the most common of conditional probability models. Currently, logit models are also the second most commonly used CSM overall (Jackson and Wood, 2013). Unlike MDA, logit CSM, which owes its fame to (Ohlson, 1980), uses maximum likelihood estimation to derive its results. Therefore, logit models manage to avoid the restrictive assumptions about the distributional properties of independent variables required by ordinary least squares (Balcaen and Ooghe, 2006). Popularisation of probit can be traced to (Zmijewski, 1984). Even though this model did not become as prevalent as logit, the main goal of the paper was to point out the serious flaws connected with construction of non-random datasets. Conditional probability models are not the last milestone in the development of the statistical class of methods. The research continued with e cumulative sums procedures, partial adjustment processes or hazard models (Jackson and Wood, 2013). Nevertheless, popularity of these models has not yet matched that of the MDA and logit.

AIES methods, a new class of CSM, was brought about by advances in computational power in the 1990s. A comprehensive review of this class can be found in (Kirkos, 2015). On top of single classifiers, AIES allow for hybridization and ensemble methods

which combine multiple algorithms and hence allow for a wider variety of CSM methods than the statistical class (Sun et al., 2013).

The most recently developed class of theoretical models helps to address the overfitting caused by variable selection. Credit scoring as a field of research had been lacks a common theoretical basis. Before theoretical models, variable selection was based on empirical procedures, such as stepwise regression, which can potentially lead to overfitting (Balcaen and Ooghe, 2006). A common theoretical basis helps to address this problem by reducing the number of regressors which are taken into consideration in construction of the model.

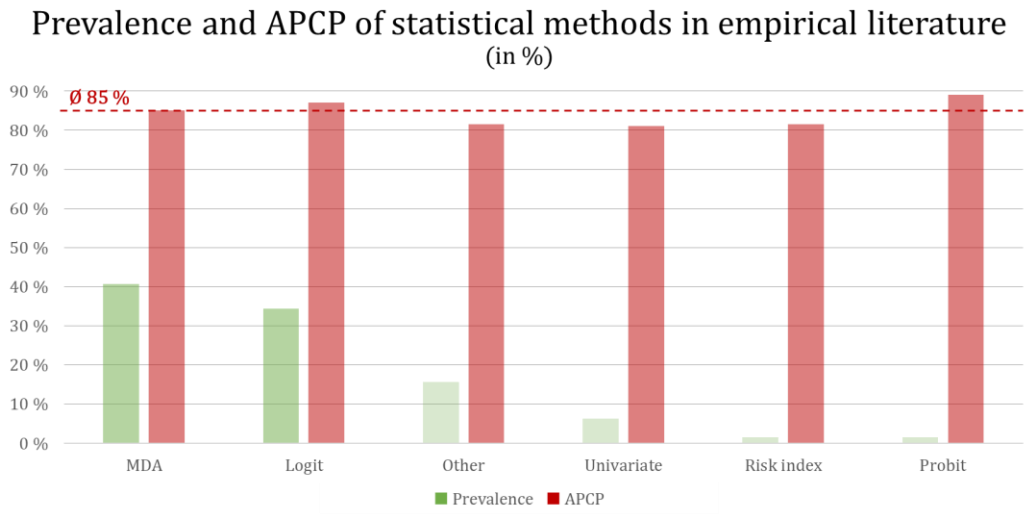
Figure 1:



Source: (Jackson and Wood, 2013)

When it comes to predictive performance, an often cited article by (Aziz and Dar, 2006) compares the results reported in 89 unique CSM. Even though there are small differences across the various method classes, their average probability of correct prediction (APCP) seems to converge at around 86 %. AIES methods only slightly outperform theoretical and statistical methods.

Figure 2:



Source: (Aziz and Dar, 2006), (Jackson and Wood, 2013)

In a similar way, (Aziz and Dar, 2006) provide a ranking of individual statistical methods. Using APCP to measure predictive performance, they find all the methods converging at around 85 %. Univariate method is on average outperformed by MDA, in turn outperformed by conditional probability models. The authors conclude their article by stating:

"Despite a dedicated effort of more than 35 years, there is apparently still no academic consensus as to the most useful method for predicting corporate bankruptcy. The major finding of this study, that the various approaches are broadly comparable, may indicate that consensus is not necessarily important."

However, it is the opinion of the author of this thesis that a simple comparison of reported results is not sufficient to rank neither individual methods, nor their classes. As discussed later in great detail, CSM tend to be data sensitive as a different country, industry composition, bankruptcy prevalence or market conditions all affect the predictive performance of CSM. Due to this fact, the reported results cannot simply be generalised. On the contrary, any sensible CSM comparison must be based on a common dataset, constructed in a way to minimise the data sensitivity of the results. To be more specific, any such dataset should be representative from both cross-sectional and temporal point of view. Moreover, such dataset should be separated into a training sample, used to derive model coefficients and a "hold-out" validation sample, serving to establish its predictive performance. Although, to the knowledge of the author of this thesis, no large-scale comprehensive performance ranking of this sort exists, the best example of this approach for select 13 models can be found in (Jackson and Wood, 2013).

2.2 Benchmark models

For each of the methods applied a foreign benchmark model was selected. For each statistical method, the first landmark article was selected. Ideally, these first exploratory models, developed more than 30 years ago, should play the role of the lowest bar to pass for all the subsequent models. In foreign as well as in the Czech CSM literature, Altman and Ohlson are commonly included as benchmark models, justifying the development of a new CSM on the basis of predictive power. Also, the inclusion of the best performing univariate model contained in (Beaver, 1966) is justified as rather surprisingly, according to results provided in (Jackson and Wood, 2013), in some cases, univariate CSM can be expected to outperform MDA. The model by (Zmijewski, 1984), even though used rather scarcely, may provide interesting results. Unlike other landmark CSM and similarly to this thesis, the model was estimated on a sample with overall bankruptcy rates close to the true population rates. Finally, one of the landmark models discussed in the previous section will be excluded. For objectivity concerns, risk index models will not be considered. The selected benchmark models were numbered for easier referencing and are contained in the following table.

Table 1: Overview of compared foreign papers and CSM

Paper	Estimation method	Model n.
(Beaver, 1966)	Univariate	12
(Altman, 1968)	MDA	13
(Ohlson, 1980)	Logit	14
(Zmijewski, 1984)	Probit	15

Source: Mentioned papers

2.3 Overview of Czech credit scoring models

All the Czech CSM contained in this thesis were sourced from articles discussed in a recent literature review by (Prusak, 2018). The literature review identified articles, which introduced new CSM. After reading though these articles, all the Czech CSM contained therein were extracted. This included additional Czech CSM which were discussed in these articles but not included in (Prusak, 2018). From all the identified articles a list was created and demonstrated by the following table. For a few Czech specific CSM the original article could not be accessed but these models were well specified in some other article. This was the case of Kralicek's Quick test which was discussed in (Machek, 2014) and then IN99 and IN01 discussed in (Neumaierová, 2002).

Table 2: Method class and industry focus of Czech papers and CSM

Paper	Method class	Industry sector
(Vochozka and Rowland, 2015)	AIES	Construction
(Dvořáček et al., 2012b)	AIES, Statistical	Unspecified
(Karas and Režňáková, 2014)	AIES, Statistical	Manufacturing
(Koráb, 2001)	Statistical	Unspecified
(Neumaierová, 2002)	Statistical	Unspecified
(Neumaierová and Neumaier, 2005)	Statistical	Industrial
(Dvořáček et al., 2008)	Statistical	Industrial
(Jakubik and Teplý, 2011)	Statistical	Non-financial
(Dvořáček et al., 2012a)	Statistical	Mixed industry
(Hampel et al., 2012)	Statistical	Agriculture
(Valecký and Slivková, 2012)	Statistical	Unspecified
(Kalouda and Vaníček, 2013)	Statistical	Mixed industry
(Karas and Režňáková, 2013)	Statistical	Industrial
(Kocmanová et al., 2014)	Statistical	Manufacturing
(Machek, 2014)	Statistical	Unspecified
(Bemš et al., 2015)	Statistical	Mixed industry
(Machek et al., 2015)	Statistical	Culture
(Vochozka et al., 2015)	Statistical	Transportation
(Němec and Pavlík, 2016)	Statistical	Non-financial

Source: (Prusak, 2018) and references contained in the mentioned articles

As far as the prevalence of methods employed, Czech statistical CSM copy the general trend observed in (Jackson and Wood, 2013) for foreign models. To be more precise, MDA is the most popular, followed by logit. Together these two account for the majority of CSM. The remaining two methods are quite interesting. Historically, the oldest Czech CSM, Králiček's Quick test discussed in (Machek, 2014), is a risk index, a rare and relatively subjective multifactor credit scoring method. The last article on the list (Bemš et al., 2015) is interesting in since it actually develops and tests a brand new statistical method, so called "Magic squares". An apparent advantage of using "Magic squares" is that the results can easily be visualised.

Next, the relevance of CSM contained in the articles was judged based on two requirements. First, its method needed to be of the statistical class. This requirement was raised by the scope of this thesis. Second, the models could not be industry specific. Applying this criteria meant that only non-financial, mixed industry, industrial firms or firms for which the industry was unspecified were selected. Industry specific models provide higher predictive performance for the cost of inefficiency, when applied to different sector (Sun et al., 2013). Comparing the general models hence puts them on an equal footing and makes the comparison more sensible.

A few additional CSM were excluded after further individualised considerations. The CSM contained in (Koráb, 2001) was excluded as this article could not be accessed. IN05 model introduced in (Neumaierová and Neumaier, 2005) is a re-estimated version

of IN01. Since the original article for IN01 could not be accessed, the newer version was kept, which allowed to assess the model in bigger detail. A somewhat similar issue occurred with the models contained in (Dvořáček et al., 2012a) and (Dvořáček et al., 2012b). The latter contains the same MDA model, which is then re-estimated using logit and AIES. As far as can be discerned, the model was re-estimated using a different dataset and so even though only one MDA model was included, both of the source articles were kept for later analysis. IN95, a model described in (Neumaierová and Neumaier, 2005), contains industry specific explanatory variable weights. In discussing how the industry weights were derived, the source article is not specific enough and so these weights could not have been re-estimated. The model described in (Bemš et al., 2015), although truly original, would be difficult to setup and would make comparisons in the empirical part of this thesis cumbersome. The latter part of the same argument also applies to the "Quick test" risk index model described in (Machek, 2014). After these considerations the final scope of this thesis could be delimited. The Czech specific CSM contained in this thesis were numbered for easier referencing and are summarised by the following table.

Table 3: Overview of compared Czech papers and CSM

Paper	Estimation method	Name	Model n.
(Karas and Režňáková, 2013)	MDA		1
(Dvořáček et al., 2008)	MDA		2
(Neumaierová and Neumaier, 2005)	MDA	IN01/IN05	3
(Kalouda and Vaníček, 2013)	MDA	CZ2	4
(Kalouda and Vaníček, 2013)	MDA	CZ2	5
(Neumaierová, 2002)	MDA	IN99	6
(Dvořáček et al., 2012a)	MDA		7
(Dvořáček et al., 2012b)	MDA		7
(Jakubik and Teplý, 2011)	Logit	JT index	8
(Němec and Pavlík, 2016)	Logit		9
(Valecký and Slivková, 2012)	Logit		10
(Dvořáček et al., 2012b)	Logit		11

Source: Mentioned papers

3 Dataset

The pooled dataset employed by this thesis spanned from 2004 to 2018 and contained information on 125 297 Czech based firms. Out of this total, 1 367 firms were classified as bankrupt. The dataset itself contains 832 074 observations. Each observation is uniquely identified by a combination of NACE, year and firm identification number (ICO). For 86 % of the firms, the dataset contains multiple observations. Apart from observation identifiers and a dummy variable for bankruptcy, observations are complete with 54 financial indicators constructed using financial statements and 4 non-financial indicators. Firm specific data was obtained exclusively from the Magnus Web database. Some of the non-financial indicators were sourced from the Czech Statistical Office (CSO) and Czech National Bank (CNB) public database. Finally, Eurostat database and Crefoport press releases were used to check the distributional properties of the dataset. The 15 compared models employ a grand total of 47 unique explanatory variables. Although a detailed overview of individual explanatory variables as well as model composition is provided in the appendix, a more general discussion contrasting general trends in explanatory variable selection and composition with insights gained from the compared models is warranted. If the dataset employed in this thesis was to be considered representative of the Czech non-financial market, it not only needed to have a sufficiently large cross-sectional and temporal dimensions, but also needed to satisfy several distributional properties. The dataset industry distribution, proxied by NACE classification, of both healthy and bankrupt firms needed to correspond with the population industry distribution. The yearly bankruptcy rate should reflect the different historical bankruptcy rates and the overall bankruptcy rate should be close to the population bankruptcy prevalence.

3.1 Dataset construction

Overall, the dataset contained information on some 125 297 Czech based firms. The criteria imposed onto the selection of individual firms were carefully chosen in order to exclude as few potential firms as possible. Rather than comparing a small number of firms disposing with pristine data, this thesis allowed to evaluate the usefulness of compared CSM as mixed industry models. With this objective in mind, this thesis imposed the following selection criteria.

First, this thesis aimed to compare credit scoring models for firms operating in the non-financial sector. Consequently, all the companies included in the financial sector with the NACE denomination K were excluded. Second, without a loss of generality only joint-stock and limited liability companies were included. According to CSO, from the total number of business companies and partnerships in the Czech Republic in 2018, these two legal forms of commercial enterprises accounted for 98 %. Of the two, limited liability companies alone accounted for 93 % of total. Whereas both joint-stock and limited liability companies share many similarities from the legal and accounting standpoint, they are quite dissimilar when compared with the remaining legal forms. Therefore, limiting the thesis in this regard made sense as the various financial ratio provided information of better quality. This qualitative criterion was imposed solely because the number of firms included in the dataset was not significantly affected.

Third, a set of 3 further selection criteria were imposed based on the financial statements of individual firms. For a firm to be included, it needed to have financial statements available for at least 2 subsequent year. This was necessary since some of the models included in this thesis employ so called “indexes”, i.e. explanatory variables derived from financial statement data for two consecutive years. The remaining two conditions require that both the sales and assets for a given year should be greater than 0. These two conditions are put in place in order to filter out economically inactive firms for which CSM are of little use.

Lastly, for the subset of bankrupt firms, a bankruptcy event and prediction horizon needed to be selected. It should be reiterated that in line with other selection criteria, both were chosen in a way to include as many firms as possible. The Magnus Web database contained mention of 34 813 firms listed as an insolvent party in an insolvency proceedings. If a match based on ICO could be established, the firm was considered bankrupt. Such definition of bankruptcy event is rather wide and must also include firms, which were listed as insolvent without ever becoming insolvent. Indeed, before being excluded, the dataset contained multiple firms which went bankrupt multiple times. However, since the insolvency proceedings tend to be lengthy, including only firms which were pronounced as bankrupt in an insolvency proceedings would have excluded a great number of firms that went bankrupt in recent years from the dataset.

To be included in the final dataset, firm’s financial statements needed to be available in the prediction horizon, else the explanatory variables could not have been constructed. The prediction horizon corresponds with the maximum time difference between the end of the last period for which financial statements were available and the bankruptcy event. The Czech credit scoring literature focuses on short-term

bankruptcy prediction, as can be seen from the following table. Accommodating the above mention general selection criteria greatly reduced the number of potential bankrupt firms. To offset this reduction and thus guarantee a sufficiently large number of bankrupt firms, the wider two-year prediction horizon was adopted.

Table 4: Prediction horizon of compared Czech CSM

Paper	Prediction horizon
(Dvořáček et al., 2008)	2 Years
(Jakubik and Teplý, 2011)	1 Year
(Dvořáček et al., 2012a)	17 Months
(Dvořáček et al., 2012b)	1 Year
(Valecký and Slivková, 2012)	1 Year
(Kalouda and Vaníček, 2013)	2 Years
(Karas and Režňáková, 2013)	1 Year
(Němec and Pavlík, 2016)	1 Year

Source: Mentioned papers

Each of the 125 297 firms which fulfilled the selection criteria was then on average represented by 6.6 observations in the final dataset. An observation represented a unique NACE, year and ICO combination. Statistical credit scoring methods derive their results from a pooled dataset in which each firm is represented by a single observation. Since papers on the subject usually report estimation results derived from a single dataset, a single observation must be selected at random from all the available NACE, year and ICO combinations. Otherwise some of the firms would be overrepresented in the dataset. Through this random selection much of the wealth of available information on individual firms is lost as illustrated by the following table, summarizing dataset composition of this thesis. Instead of eliminating the additional 5.6 observations per firm upfront, this thesis kept all the available information. The description of the necessary process through which multiple observations per firm were reduced to one observation per firm can be found in the chapter dedicated to random sampling.

Table 5: Breakdown of firms included in the dataset by observation availability

	Single year	Multiple years
Single NACE	14 %	68 %
Multiple NACE	3 %	16 %

3.2 Data sources

The data employed in this thesis was one of two kinds. Firstly, the firm specific data which was obtained from the Magnus Web database. This data was both of financial and non-financial nature and could be characterised as being linked with individual

companies. Secondly, the remaining firm non-specific data came from a multitude of sources. On one hand, this data was used to construct a firm non-specific explanatory variable. On the other hand, and more importantly, it allowed to construct the dataset which was representative of the overall Czech non-financial market by providing information about its distributional properties.

Magnus Web is a closed access database operated by Bisnode Czech Republic. Bisnode products, namely Magnus Web or Albertina, provide a user-friendly environment, through which information about a large pool of Czech based firms can be filtered and later downloaded. As such, it is commonly employed in the Czech CSM literature. Taking into account that ČEKIA is an older title of the Bisnode company, out of the 6 papers reviewed in this thesis for which the source of data is provided, 4 employ Bisnode databases as a source of information.

Table 6: Data sources of compared Czech papers

Paper	Source
(Jakubik and Teplý, 2011)	ČEKIA
(Dvořáček et al., 2012a)	Online Press Releases
(Dvořáček et al., 2012b)	Official Business
(Valecký and Slivková, 2012)	Magnus Web
(Kalouda and Vaníček, 2013)	Albertina
(Němec and Pavlík, 2016)	Albertina

Source: Mentioned papers

The Bisnode databases aggregate data from a number of mostly publicly accessible sources, chief among them the official business register and ARES database. Importantly, Bisnode databases also aggregate data from the official insolvency register. As a consequence, the database contains a large pool of both financial and non-financial firm-specific information. For the purposes of this thesis, the term financial information encompassed the information contained in financial statements, namely in balance sheet and in profit and loss statements. Non-financial firm-specific data included information about insolvency proceedings, NACE classification, date of incorporation and legal form.

The remaining firm non-specific data was obtained from CSO and CNB official sites, Eurostat database and Crefoport press releases. Data used for the construction of the non-financial explanatory variables, namely unemployment rate and GDP growth was obtained from official sites of CSO whereas the data about 2W repo rate were sourced using the official site of CNB. Finally, the information about overall firm NACE distribution and bankruptcy prevalence was primarily obtained from the Eurostat public database and Crefoport press releases respectively. Both were chosen amongst other potential sources for having the longest respective timespan, in turn guarantying

their methodological consistency. Crefoport is a Czech based company specialising in market research and accounts receivable management.

3.3 Explanatory variables

As far as explanatory variable selection, a detailed critique is presented in (Balcaen and Ooghe, 2006). Variables are most often selected empirically by choosing a subset from a list of potential financial indicators. In this regard, stepwise regression or t-test are commonly employed tools (Sun et al., 2013). The empirical approach to variable selection has multiple negative implications (Balcaen and Ooghe, 2006). Among the most significant implications, from a scientific point of view, is that the results cannot be generalised. Even today, there is no consensus on superior predictors. Across 165 papers reviewed in (Bellovary et al., 2007), 90 % of all explanatory variable employed in CSM literature appear only once or twice. No consensus on best performing explanatory variables and a lack of theoretical basis means that CSM of the statistical class run a high risk of being overfitted. Although a few underlying theories for variable selection exist, many papers still rely on empirics rather than theory. Authors of (Balcaen and Ooghe, 2006) name common sense as the ultimate filter for variable selection:

"[A general viewpoint is that] a good failure prediction model should include some carefully chosen variables from the whole spectrum of financial analysis—liquidity, indebtedness, profitability, and activity…and that it should use these variables in the intuitively right sense."

Out of the 47 unique variables, 32 appear in only one of the models. The lack of consensus is underlined by the fact, that even though many of the explanatory variables use identical financial items as inputs, they differ in the specification of functional forms. To cite an example, both variables n. 1 and 4 compare assets to debt in a simple ratio. The only difference is that the former applies the debt in the nominator, whereas the latter uses debt in the denominator. At first glance, such functional differences may seem trivial. However, in the process of generating the explanatory variables the latter measure also generated 0.9 % of missing values whereas the former generated none. In this regard, variables n. 8, 30 and 32 respectively are the worst perpetrators, generating 50.7 % of missing values. Were these explanatory variables specified with the interest expense in the nominator, the proportion of missing values would drop to 10.0 %, 9.0 % and 9.1 % respectively.

As far as variable composition, the explanatory variables first needed to be classified according to their type. This categorisation split the explanatory variables into 4 bins

titled solvency, profitability, liquidity and activity explanatory variables. A clarification of the difference between solvency and liquidity variables is warranted. Solvency is the capacity to repay long-term debt whereas liquidity encapsulates the capacity to repay current debt. The assigned type of individual explanatory variables can be found in the appendix to this thesis. As expected, the compared models rely most heavily on solvency and liquidity variables. Every model includes at least one solvency or liquidity measure. Of the 11 compared Czech CSM, only 4 models numbered 3, 4, 8 and 6 employ explanatory variables of all types.

Table 7: Breakdown of explanatory variables by type and model

Model	Solvency	Profitability	Liquidity	Activity
1	0	0	1	2
2	2	1	0	5
3	1	1	2	1
4	1	1	2	3
5	0	0	3	0
6	1	1	1	1
7	2	0	3	3
8	3	2	1	1
9	4	0	1	1
10	2	1	2	0
11	2	0	3	3
12	1	0	0	0
13	2	1	1	1
14	3	3	2	1
15	1	1	1	0
Total	25	12	23	22

3.4 Dataset representativeness

3.4.1 Dataset size and timespan

For the purposes of this thesis, as far as datasets are concerned, the bigger the better. On one hand, as discussed later in this thesis, sufficiently large dataset helped in minimising the issue of data sensitivity. On the other hand, large size was required so that both the overall dataset bankruptcy rate and the industry distribution of bankrupt firms matched the population parameters. Consequently, this thesis aimed to include the largest pool of data possible. In total 125 297 firms or estimated 63 % of the Magnus Web was included in the dataset.

Table 8: Samples size of compared Czech papers

Paper	# of firms
(Neumaierová and Neumaier, 2005)	1526
(Dvořáček et al., 2008)	73
(Jakubik and Teplý, 2011)	757
(Dvořáček et al., 2012a)	170
(Dvořáček et al., 2012b)	144
(Valecký and Slivková, 2012)	242
(Karas and Režňáková, 2013)	207
(Němec and Pavlík, 2016)	2061

Source: Mentioned papers

Compared with the previous widest study of Czech non-financial market, the observations come from approximately 60 times more firms. In fact, obtaining more data was not possible as the limits of the Magnus Web were reached. More data could not be included as the dataset already contained all the firms operating in the agricultural sector.

To a certain degree, sufficiently long timespan can also help in reducing the problem of data sensitivity. According to (Mensah, 1984), the predictive performance and structure of models differs with regards to "economic environment", proxied by inflation, interest rates and business cycle. This means that a model developed on a longer timespan can be expected to trade some of its predictive accuracy for more temporal stability. In order to be representative of the Czech non-financial market, the timespan needed to be long enough to represent diverse market conditions. The dataset employed in this thesis spanned from 2004 to 2018. In this period, the average yearly GDP growth slowed from 7 % p.a. between 2004-2008, to 0 % p.a. during the period of 2009-2013. just to rise again to 5 % p.a. in the following 5-year window. Hence the data covered an economic cycle from the boom to the bust and back.

Once again, data availability was the limiting factor with respect to the timespan of the dataset. Since the empirical part of this thesis was carried out in the first half of 2020, the majority of financial statement data for the year 2019 were not yet published. As for the lower bound of the timespan, including the year 2003 and below would significantly reduce the number of available observations per year, which would in turn make safeguarding all the discussed distributional properties impossible.

3.4.2 Dataset industry distribution

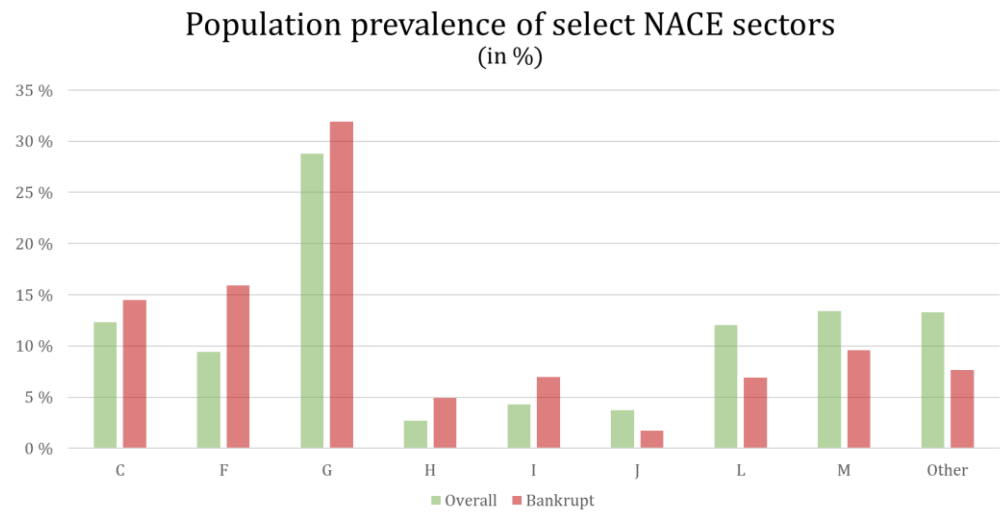
Industry is generally recognised as an important factor in CSM empirical literature. The models themselves are commonly divided into industry specific models and general models. This categorisation is justified. Even the early research into financial

ratio analysis showed that different industries manifest “clustering of ratios” (Chudson, 1945). In the article (Platt and Platt, 2002) we learn that a model trained on one industry is often ineffective for another industry. To evaluate the predictive performance of general models, the industry distribution of the dataset must match the population industry distribution.

For the purposes of this thesis, the widely used NACE industry classification was employed. The population NACE distribution was proxied by the Eurostat database. Since sector composition of the Czech non-financial market remained relatively stable throughout the years 2008 to 2017, the prevalence of population NACE sectors was proxied by their long-term arithmetic average. The Eurostat database does not include information on agricultural firms. The prevalence of agricultural sector was therefore extrapolated from Crefoport press releases as accounting for 4.5 % of the total Czech non-financial market. In some Czech papers, for example in (Němec and Pavlík, 2016), the distribution of firms contained in the dataset is compared with the distribution of firms available in the Bisnode databases. If the data reported by Eurostat is to be trusted, this is not an advisable practice. Some sectors, such as real estate, display differences of almost 5 %.

As for the subset of bankrupt firms, their NACE distribution was matched with the information contained in the 2010 to 2017 Crefoport press releases. Once again, the reason for preferring this source was its overall timespan. With Crefoport press releases the argument was even stronger since it also contained the yearly bankruptcy rates. In a single sector, the prevalence of bankruptcy exhibits high levels of year to year variation. The variation can be explained by the fact that bankruptcy of a firm is not an isolated effect. In the very least it affects its suppliers and distributors. The so called domino effect for the UK firms accounts approximatively for 27 % of bankruptcies (Jackson and Wood, 2013). Since none of the compared models tries to explain this effect, a long-term average for the years 2010 to 2017 was taken as the measure of bankruptcy prevalence.

Figure 3:



Source: Eurostat, Crefoport

Finally, the following table presents the cross-sectional NACE distribution comparison between the Czech non-financial market and the dataset employed in this thesis. Across the 149 865 unique NACE and ICO combinations, the overall firm's distribution falls within 0.01 % of the population distribution. The match between the population and dataset for the subset of 1 486 unique NACE and ICO combinations representing bankrupt firms, is less stellar. However, all the differences are within a tolerance of 1 %. The higher discrepancies are at least partly due to discrepancies introduced by the rounding error.

Table 9: Comparison of population and dataset NACE distribution

NACE	Firms overall			Bankrupt firms		
	Eurostat	Dataset	Difference	Crefoport	Dataset	Difference
A	4.3%	4.3%	0.00%	2.0%	2.5%	-0.56%
B	0.1%	0.1%	0.00%	0.1%	0.1%	0.03%
C	12.3%	12.3%	0.01%	14.7%	14.5%	0.20%
D	0.5%	0.5%	0.01%	0.3%	0.2%	0.14%
E	0.9%	0.9%	0.00%	0.7%	0.6%	0.07%
F	9.4%	9.4%	0.01%	16.1%	15.9%	0.17%
G	28.8%	28.8%	0.00%	32.2%	31.9%	0.26%
H	2.7%	2.7%	0.00%	4.4%	4.9%	-0.47%
I	4.3%	4.3%	0.01%	7.0%	6.9%	0.05%
J	3.7%	3.7%	0.00%	2.3%	1.7%	0.58%
L	12.1%	12.1%	-0.01%	7.0%	6.9%	0.07%
M	13.4%	13.4%	-0.01%	9.6%	9.6%	-0.01%
Other	7.5%	7.5%	0.00%	3.7%	4.3%	-0.55%

Source: Eurostat, Crefoport

3.4.3 Dataset bankruptcy prevalence

A common practice in CSM literature is to employ so-called paired samples. In the formation of a sample, each bankrupt company is paired with a company of a similar size, age or industry. Such procedure was employed by Altman in his landmark article and has been replicated many times since. Creating such non-random samples was challenged in (Zmijewski, 1984). With the true rate of bankruptcy well below 50 %, pairing basically represents oversampling of bankrupt firms. The author remarked that among the reviewed empirical papers, the lowest proportion of bankrupt firms was twice as high as the level of the true population. Zmijewski then showed a negative relationship between proportion of distressed firms and type I error as well as a positive relationship between the bankruptcy rate and type II error (Sun et al., 2013).

Even if the landmark article by Zmijewski dates back to 1984, 4 out of 7 Czech CSM continue to employ sample pairing, as is evident from the following table.

Table 10: Overall sample bankruptcy rate of compared Czech papers

Paper	Pairing	% of bankrupt
(Dvořáček et al., 2008)	No	45%
(Jakubik and Teplý, 2011)	No	20%
(Dvořáček et al., 2012a)	Yes	50%
(Dvořáček et al., 2012b)	Yes	50%
(Valecký and Slivková, 2012)	Yes	41%
(Karas and Režňáková, 2013)	No	15%
(Němec and Pavlík, 2016)	Yes	50%

Source: Mentioned papers

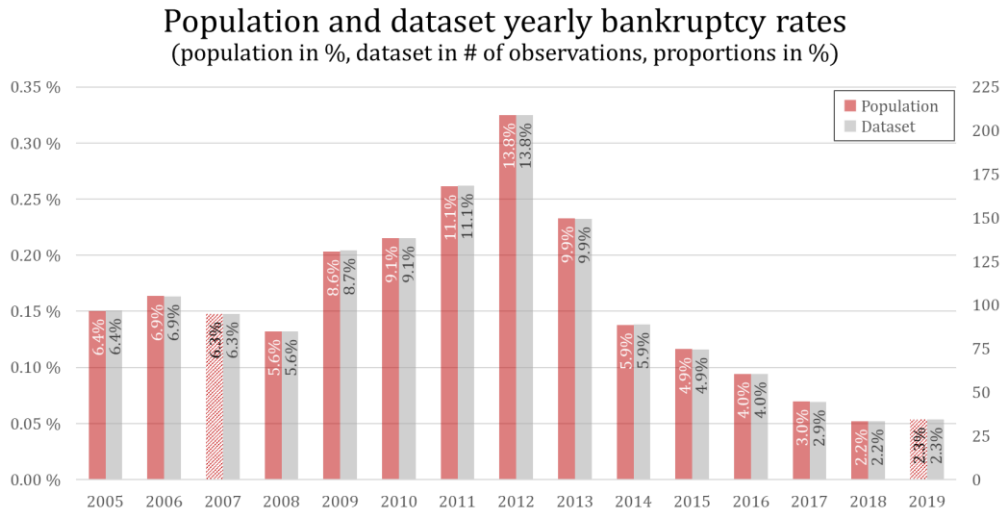
At the same time, the observed average prevalence of bankruptcy among Czech firms sits at around 0.16 %. This means that the paper by (Karas and Režňáková, 2013), the one with the lowest overall bankruptcy, still manages to overrepresent this figure approximately 94 times. Consequently, the reported predictive performance of Czech CSM must be taken with more than a pinch of salt.

Though this thesis does not employ pairing, the applied overall bankruptcy rate of 1 % is still about 6.5 times higher than the true population rate. The reason behind employing the higher rate was purely pragmatic. As mentioned before, the overall size of the sample was limited to 149 865 unique NACE and ICO combinations. Applying the true population rate would mean that bankrupt firms from sectors such as mining, which make up only a relatively small share of non-financial firm but which have an above average bankruptcy rate, would not be represented in the dataset. In this regard, satisfying the NACE distributional properties was preferred which required using the higher average bankruptcy rate.

On top of the overall bankruptcy rate, the yearly bankruptcy prevalence also needed to be verified. Apart from safeguarding the representatives of the dataset, corresponding yearly bankruptcy rates were needed to draw conclusions about macroeconomic non-financial explanatory variables. If unchecked, the links which exist between these macroeconomic variables and the yearly bankruptcy rates would potentially be destroyed due to non-corresponding dataset distribution. The yearly bankruptcy rates were obtained from the Crefoport press releases for the years 2004 to 2018. Since the press releases for the years 2007 and 2019 could not be found, the bankruptcy rates were extrapolated. Bankruptcy prevalence for the year 2007 was computed as the arithmetic average of the years 2006 and 2008. The bankruptcy rate for 2019 was in turn computed based on the 2018 Crefoport rate and the 3 % yearly increase reported by CRIF, a competing market research firm. Yearly bankruptcy rate displays a high degree of variability. At its high in 2012 the bankruptcy rate was approximately 6 times higher than at its low in 2018. As demonstrated by the following figure, when

compared with the population, the dataset values are within 0.1 % tolerance for the entire timespan.

Figure 4:



Source: Crefoport, CRIF, CSO

4 Methodology

In order to reduce the data sensitivity of the results, predictive performance was evaluated on a set of samples generated through random sampling. The broad dataset was repeatedly and at random split into 3 disjoint parts. Two equal parts were used for training and validation, whereas the last part was omitted. In this process, the multiple observations per firm were also reduced to a single observation per firm. To quantify the predictive performance of compared models, area under curve (AUC) was employed. This measure, related to commonly used "Gini coefficients", was selected since it allows for direct comparison of different estimation methods without the need of arbitrary selection of cut-off values. In many ways, the design of this thesis diverges from the typical economics research. The unusual setup required the usage of somewhat unusual statistical methods to draw inference. On one hand, in case of parametric estimation, ANOVA accompanied by the Tukey's test was employed. On the other, non-parametric estimation was performed with Friedman test and the Nemeyi's post-hoc test.

4.1 Random sampling

As discussed in the preceding chapters, data sensitivity is a serious flaw plaguing the credit scoring literature. To minimise the issue, this thesis built on the approach contained in a similar comparative study by (Jackson and Wood, 2013).

First, similarly to 50 % of Czech papers, the dataset was split into a training and a validation subsets. The training subset was used to obtain the estimates of coefficients assigned to independent variables. The validation subset was then used to evaluate the predictive performance of the model. Doing so simulated the real-life use of CSM in the predictive context. More importantly and disregarding this practical aspect, the separation limited the extent to which a an overfitted model affected the results. Each result was derived not from two samples, reducing the data sensitivity.

Second, similarly to (Jackson and Wood, 2013), the splitting of the dataset into training and validation samples was done through a process of random sampling with replacement. Therefore, rather than a single resulting performance indicator per sample, an entire distribution could be obtained. Doing so diluted the impact of individual sample on the results. It also allowed to observe not just the level of predictive performance, but also its stability.

Third, on top of these precautions, supplementary steps other than those contained in (Jackson and Wood, 2013) were taken in order to reduce data sensitivity. These included the introduction of an exclusion sample and the random reduction of multiple observations per firm into one.

In the process of random sampling, the original dataset was repeatedly split into 3 sample. Besides the training and validation samples, a third exclusion sample was designated. This sample was selected in a way to be disjoint from both the training and validation. As a consequence, for the given split generated by random sampling, this sample was simply excluded from training and validation. Through the introduction of the exclusion sample, further sample variation was introduced at the cost of reducing the number of available observations. s

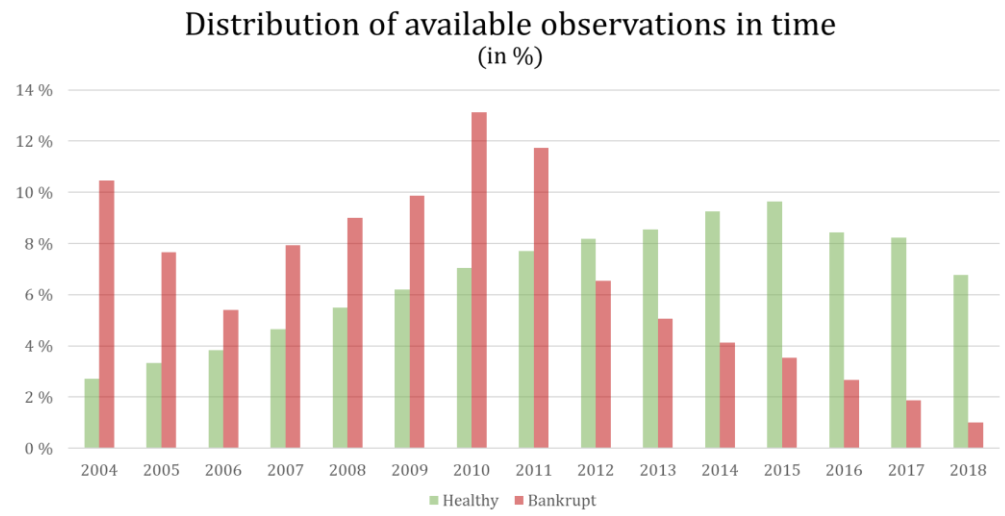
The random sampling was done on the basis of ICO. As mentioned before, the original broad sample often contained multiple observations per firm, corresponding with unique ICO, NACE and year combinations. For each ICO and year combination, the values contained in financial statements and therefore the values of explanatory variables derived from these statements differed. To estimate the models, each firm needed to be represented by a single observation. As discussed, 84 % of all firms contained in the dataset were represented in the dataset by more than one ICO and year combination. The process of randomised reduction of multiple observations therefore represented an important potential source of sample variation. Rather than randomly selecting an observation only once, the process of randomised reduction of observation was repeated with each random sampling. Since the selected bankruptcy prediction horizon was 2 years, for each bankrupt firm there were only 2 observations available at most. Healthy firms therefore accounted for a greater source of variation.

The process through which multiple NACE and ICO combinations were reduced to one was identical. Since the financial statements for ICO and year combinations did not change with NACE, the classification did not affect the explanatory variables and hence the results directly. As NACE served as an important grouping factor, its variation effected the overall results indirectly. For example, in the treatment of missing values, the imputed values of mean and median were computed per NACE group.

In total, for the purposes of this thesis 250 sets of training, validation and exclusion samples were created. Both training and validation samples were created equal in the number of observations included and contained observations from the entire timespan of the broad dataset. Of the 50000 observations in the non-excluded samples, 500 represented bankrupt firms, meaning that the exclusion sample accounted for 25297

firms. Even though the random sampling helped to address the issue of data sensitivity, it also potentially reduced the representativeness of each sample. To mitigate the problem, the most important characteristics on which the representativeness of the sample relied were safeguarded through a set of limitations imposed on the random sampling process.

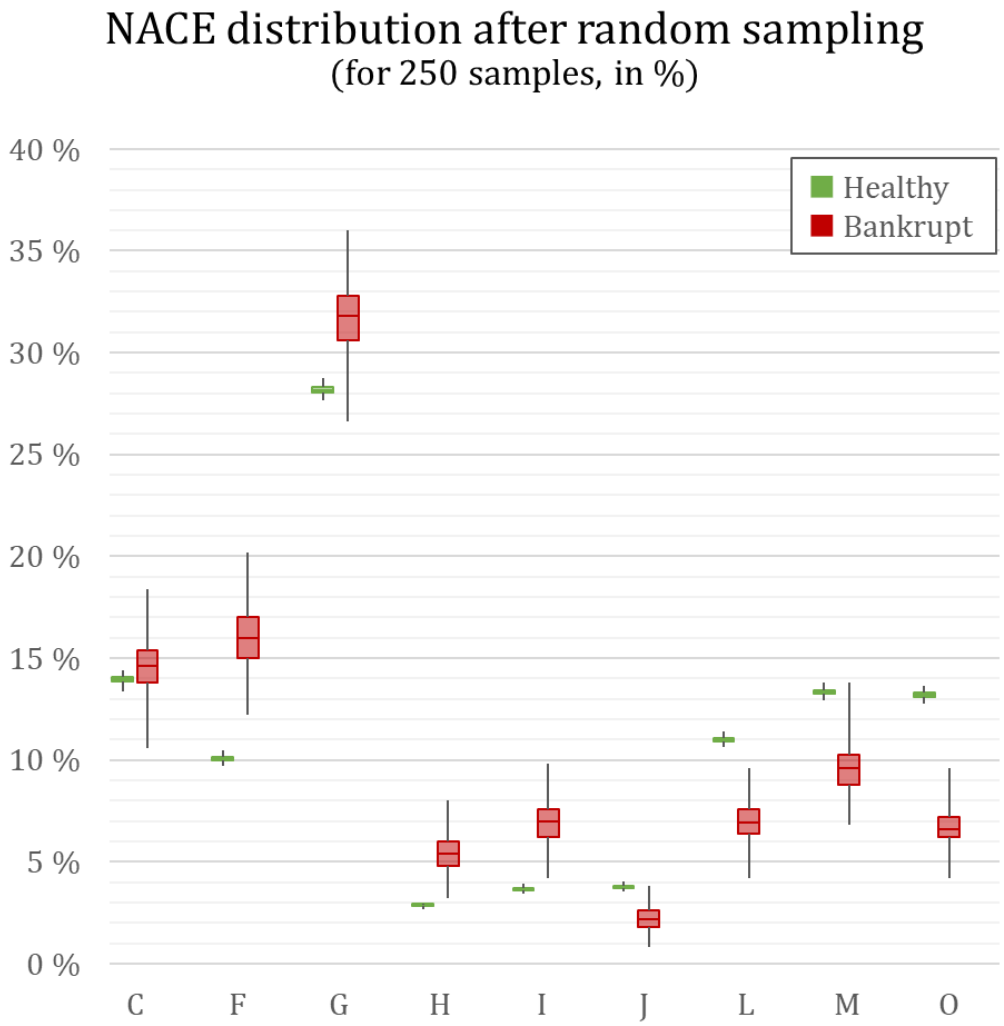
Figure 5:



A limitation needed to be imposed on the number of observations of healthy firms per year. The preceding figure shows that significantly more observations were available in the latter half of the dataset timespan. In fact, even though the years 2013 to 2018 stretched roughly a third of the timespan, they accounted for the majority of observations. Limiting the number of observations of healthy firms per year should conserve the desired yearly bankruptcy rates for the cost of higher variation attributable to the later years. An equivalent limitation was not imposed on bankrupt firms since they were chosen so as to match the population yearly bankruptcy rate variations.

Other than temporal distribution, the overall NACE distribution was verified visually. Naturally, the subset of bankrupt firms being smaller displayed a higher degree of dispersion. However, as can be seen on the accompanying boxplot, even in the case of this subset half of the observations fell within 1 % of the mean. Therefore, it can be concluded that the distributional properties of both bankrupt and healthy firms in these samples were at large satisfied.

Figure 6:



4.2 Prediction performance indicator

After the CSM were estimated on the set of training samples, the model coefficient could be extracted. These coefficients were then used to obtain a set of fitted values using a validation sample. With the predicted values at hand, the next step was to evaluate the performance of individual models. Credit scoring literature uses a plethora of prediction performance indicators. Of those listed in (Balcaen and Ooghe, 2006), Czech literature employs the measures of average probability of correctly classified, type I and type II errors and "Gini coefficient".

By the frequency of use, the first three measures are the most popular. Of the 9 Czech articles compared, 7 employ APCP and 6 employ Type I and II errors. The main benefit of these measures is their intuitiveness. Their main drawback is that they introduce arbitrariness into the comparison of model performance. To establish whether a firm should be categorised as bankrupt in the predictive context, the fitted values are

compared with the so called "cut-off" value (Balcaen and Ooghe, 2006). In the case of conditional probability models, the selection of such "cut-off" value may be implicit. Since their fitted values of probit and logit ultimately fall between 0 and 1, the firms can be categorised simply by rounding the fitted values to the closest integer. As for the MDA models, the selection of "cut-off" value can be done based on a reasonable criterion such as APCP minimisation (Altman, 1968). However sophisticated, the selection of a unique "cut-off" value inevitably introduces arbitrariness. The results of any cross-model comparison are dependent not only on the predictive performance of compared models but also on the selection of the "cut-off" value.

Table 11: Prediction performance indicators of compared Czech papers

Paper	Type I/II error	APCP	Gini
(Karas and Režňáková, 2013)	Yes	Yes	No
(Dvořáček et al., 2008)	Yes	Yes	No
(Neumaierová and Neumaier, 2005)	No	Yes	No
(Kalouda and Vaníček, 2013)	No	No	Yes
(Dvořáček et al., 2012a)	Yes	Yes	No
(Dvořáček et al., 2012b)	Yes	Yes	No
(Jakubik and Teplý, 2011)	No	No	Yes
(Němec and Pavlík, 2016)	Yes	Yes	No
(Valecký and Slivková, 2012)	Yes	Yes	No

Instead of employing APCP, Type I or Type II measures, this thesis follows the approach discussed by (Jackson and Wood, 2013) by employing average area under curve or AUC as the prediction performance indicator. The main benefit of this measure is that evaluates the predictive performance of models on the whole spectrum of potential "cut-off" values.

AUC is derived using receiver operating characteristic curve or ROC. First a set of true positive rates is obtained by assessing the predictive performance of a model across the potential "cut-off" values. The procedure is repeated in order to obtain a set of false positive rates. ROC are then created by plotting the true positive rate against the false positive rate. (Schechtman and Schechtman, 2019). By calculating the area under a ROC, one obtains the AUC measure.

At first sight, it may seem that the article introduces yet another type of prediction performance indicator. Nevertheless, there exists a simple arithmetic relationship between AUC and the aforementioned "Gini coefficient". Although (Schechtman and Schechtman, 2019) explain that the terminology used by the authors comparing the two measures leaves a lot to be desired, in laymen terms the $Gini = 2 * AUC - 1$, where $Gini$ stands for the so-called "Gini coefficient". The overall interpretation is similar. A model with $AUC = 1$ classifies the variables into the two categories perfectly. A model

with $AUC = 0.5$ has the same predictive power as a coin toss and a model with $AUC < 0.5$ is worse than random chance in categorising the two variables.

4.3 Statistical inference

In the most condensed way possible, the aim of this thesis was to evaluate how the predictive power of CSM differs with individual models and model characteristics. To fulfil this goal, median AUC and AUC standard deviations could serve as a measure of the expected level of predictive power and its variability. A subsequent question was whether the AUC means obtained by employing different models and methods were statistically significantly different from each other. To answer this question, one must resort to statistical inference.

Usually in the field of economics, researchers observe a process over which they have little control. A person studying the causality between GDP growth and income inequality cannot easily manipulate the former to observe the effects on the latter. Rather than manipulating the causes themselves, the links between cause and effect are commonly extrapolated after controlling for other relevant factors. Within the framework of this thesis however, the application of different models or model characteristics can be introduced at will. In this light, the setup allows for experiments to be conducted on the given set of samples. Naturally, inference in experimental setup demands the usage of statistical methods typical for natural sciences or psychology that are somewhat atypical for classic economics.

When selecting the appropriate statistical method, the first thing to consider is the experiment design. Experiment design is important for a great number of scientific fields, each imparting its own terminology on the issue. For the sake of consistency, this thesis applied the terminology contained in (Dean et al., 2017). The response variable was represented by the average AUC. Each variation in model or model characteristics could be viewed as one of v treatments. Each sample-treatment combination could be seen as an experimental unit. For reasons explained below, all treatments were applied to all samples. Each sample hence represents a block of identical experimental units numbering k, where $k = v$. The experimental design thus described is called randomized complete block design.

As discussed before, CSM are known to be data sensitive. If the results of this thesis are to be generalized, it was paramount to limit the extent to which the variation in the specific firm composition of a sample could have an impact on the results. In a completely randomized design, a sample would be assigned to a single treatment at random. Each sample would present a unique experimental unit. To reduce the weight

sample variability had on the results, one could increase the number of samples employed and hope that the individual differences average out. A more effective way achieving the same goal was to tweak the experiment design. Rather than being exposed to a single treatment, each sample would be subjected to all treatments. Individual samples then needed to be considered as blocks of identical experimental units rather than unique experimental units. The main advantage of such a design is that for a given block, individual differences across the methods are effectively eliminated (Maxwell et al., 2004).

To test the hypothesis that all responses are identical for different treatments, the ANOVA framework can be used. In this regard, the step by step procedure laid down in (Dean et al., 2017) was once again referred to. ANOVA relies on a set of distributional properties. In case these assumptions were not met or where their verification proved to be inconvenient, a non-parametric Friedman test, as described in (Hollander et al., 2013) was used.

Following the notation in (Dean et al., 2017), ANOVA model for randomized complete block design could be defined as

$$Y_{hi} = \mu + \theta_h + \tau_i + \varepsilon_{hi}$$

where

$$\varepsilon_{hi} \sim N(0, \sigma^2)$$

$$\varepsilon_{hi}\ are\ mutually\ independent$$

$$h = 1,2, \dots, b$$

$$i = 1,2, \dots, v$$

and Y_{hi} and ε_{hi} are respectively the response and error variables connected with block h and treatment i, θ_h and τ_i respectively encapsulate the effect of the block h and the treatment i on the response Y_{hi}, b and v respectively are the total number of blocks and treatments and finally μ is the overall mean. The hypothesis to be tested can be stated thusly

$$H_0: \{\tau_1 = \tau_2 = \cdots = \tau_v\}$$

against the alternative that

$$H_0: \{\tau_1, \tau_2, .., \tau_v \text{ not all equal}\}$$

The basic idea behind the inference as outlined by (Dean et al., 2017) relies on sum of squares as a measure of explanatory power. To test the H_0, it compares the sum of squares generated by the full model ssE_f with ssE_r, a sum of squares of a hypothetical reduced model, which assumes that H_0 holds. It can be shown that

$$\frac{ssE_f}{\sigma^2} \sim \chi^2_{n-v};\ \frac{ssE_r - ssE_f}{\sigma^2} \sim \chi^2_{v-1};\ ssE_r - ssE_f \text{ and } ssE_f \text{ independent}$$

and therefore

$$\frac{\frac{ssE_r - ssE_f}{\sigma^2(v-1)}}{\frac{ssE_f}{\sigma^2(n-v)}} = \frac{\frac{ssE_r - ssE_f}{(v-1)}}{\frac{ssE_f}{(n-v)}} \sim F_{v-1,n-v}$$

This hypothetical distribution than can be used to test the H_0 at a specified level of significance.

The same hypothesis can be tested using the Friedman test. The model relaxes the constraining assumption of normal distribution of errors (Hollander et al., 2013). Modifying the notation to more closely match the one connected with the ANOVA model, Friedman model can be summarised thusly

$$Y_{ijt} = \mu + \theta_h + \tau_i + \varepsilon_{hi}$$

where

$$\varepsilon_{hi}\ are\ iid$$

$$\varepsilon_{hi}\ are\ independent\ of\ \tau_i$$

$$h = 1,2,\dots,b$$

$$i = 1,2,\dots,v$$

and Y_{hi} and ε_{hi} are respectively the response and error variables connected with block h and treatment i, θ_h and τ_i respectively encapsulate the effect of the block h and the treatment i on the response Y_{hi}, b and v respectively are the total number of blocks and treatments and finally μ is the overall median. To obtain the decision rule, Friedman test statistic S is utilized

$$S = \frac{12v}{b(b+1)} \sum_{h=1}^{b} \left(R_{.j} - \frac{n+1}{2} \right)^2$$

where

$$R_{.j} = \frac{R_{ij}}{v}$$

and R_{ij} denotes the rank of Y_{hi} in the block h. Friedman test statistic S has an asymptotic χ^2_{b-1} distribution is used, which can be finally used to test the H_0.

Even though ANOVA and Friedman test can establish that at least two treatments are statistically different, it does not examine the differences amongst individual treatments. As is common, these tests were therefore followed by a post-hoc test. From the 4 potential post-hoc tests proposed for ANOVA by (Dean et al., 2017), the Tukey Method was chosen as it delivers the tightest confidence intervals amongst the compared methods, allowing for all pairwise comparisons. As for the Friedman post-hoc test, the source material for this thesis (Hollander et al., 2013) only provides a description of Wilcoxon, Nemenyi, McDonald-Thompson also shortened to Nemeyi test, making the choice rather simple.

To draw inference, a decision whether to apply the parametric ANOVA or non-parametric Friedman test had to be taken. The decision ultimately hinged on the core distributional assumptions being met. For ANOVA, one primarily needs to check for normality of population errors, represented by the sample residuals. Other than normality, the assumptions of equal variance of errors is also important in selecting the appropriate estimation method. Unlike violations of normality, ANOVA can accommodate the violations of equal constant variance using the Satterthwaite's method (Dean et al., 2017). Even if the parametric approach was selected, the results of the non-parametric Friedman and the corresponding post-hoc Nemeyi test were still computed and presented for comparison.

Other than the normality and homoskedasticity, the assumptions of error independence were verified. Following the reasoning and the procedure in (Dean et al., 2017), all the assumptions were checked visually. Verifying the assumptions and selecting an appropriate method with all relevant modifications could prove impractical. Therefore, when individual models were concerned, inference was drawn using the non-parametric tests exclusively.

5 Results

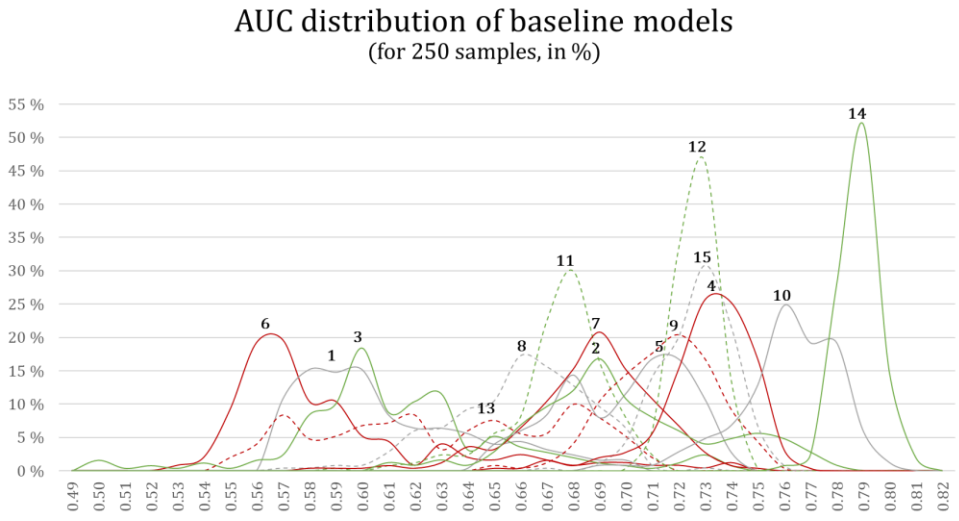
The goal of the first subchapter is to evaluate the predictive performance of the compared models, as described in the Czech and foreign CSM literature. Each model is estimated and evaluated on a set of 250 pairs of training and validation samples, which are identical for all the models compared. The obtain sets of 250 prediction performance indicators per model were then used to create a baseline CSM ranking. The second subchapter explores 3 different ways, in which the CSM estimation setup could be optimized in order to deliver higher predictive performance. Motivated by the findings of the empirical CSM literature and stylized facts presented in the first subchapter, it explores the impact of employing different missing value treatments, estimation methods and additional non-financial variables on the predictive performance. The third and last subchapter then applies the optimal estimation setup, that is to say the estimation setup which provides the highest economically and statistically significant improvement of predictive performance over the baseline estimate, to the compared models. Thus, a new optimized model ranking is established and contrasted with the baseline. Finally, Czech and foreign benchmark models are grouped together and compared, both in their baseline and in their optimized form, in order to evaluate which, present a better suited alternative for bankruptcy prediction in the Czech Republic.

5.1 Baseline model performance comparison

As a first step a baseline model comparison needed to be established. To achieve this, the 15 compared models were first estimated on a set of 250 training samples, after which it was evaluated on another set of 250 validation samples. The set of training and validation samples were identical for all the models compares. For the baseline comparison, the missing values were omitted.

For illustration purposes, the resulting AUC were rounded to 2 decimal points and ploted against their relative occurrence. The results are captured by the following figure. Based on this representation, one can conclude that the best performing model was model n. 14, the worst model n. 6. Between these two extremes, the remaining models are spread a uniform fashion. Importantly, judging by the plot, the individual distributions seem roughly bilaterally symmetric. This means that both mean and median are equally suitable as a statistic for model performance evaluation.

Figure 7:



Next the model specific summary statistics were computed, and the models were ranked using their median. The graphical intuition was confirmed, as the highest recorded median predictive performance of around 78.9 % AUC was achieved by using model n. 14. The median of the lowest scoring model, model n. 6 was just 57.4 % AUC. The resulting median spread from 57.4 % AUC to 78.9% AUC is similar with the results obtained for statistical methods by (Jackson and Wood, 2013), that is to say from 58.3 % AUC to 80.5 % AUC.

Table 12: Summary statistics for the baseline CSM comparison

Model	Median	Mean	Min	Max	Std.	Kurt.	Skew.
14	78.9%	78.8%	76.7%	80.8%	0.0077	3.49	-0.12
10	76.4%	76.1%	69.2%	79.8%	0.0195	4.02	-0.87
4	73.4%	73.1%	64.8%	76.6%	0.0184	6.30	-1.46
15	72.8%	72.7%	66.7%	76.1%	0.0140	3.96	-0.58
13	72.7%	63.3%	70.3%	74.7%	0.0075	2.91	-0.22
9	71.5%	71.4%	65.1%	75.6%	0.0194	3.10	-0.33
5	70.2%	69.8%	64.3%	74.3%	0.0238	2.10	-0.33
2	69.4%	69.9%	61.0%	77.7%	0.0349	2.79	0.18
7	68.9%	68.6%	58.1%	74.5%	0.0262	4.83	-0.85
11	67.7%	67.4%	58.7%	71.4%	0.0189	5.56	-1.19
8	66.2%	66.0%	57.4%	73.1%	0.0261	3.22	-0.47
12	63.5%	72.7%	54.7%	71.3%	0.0453	1.77	-0.09
3	60.9%	61.6%	49.7%	73.8%	0.0427	4.37	0.42
1	60.0%	61.0%	56.6%	70.1%	0.0333	2.85	0.85
6	57.4%	59.0%	52.6%	74.5%	0.0432	5.34	1.63

Note: Models ordered based on median AUC rank

Note that choosing mean instead of median would only affect the respective rank of models n. 12 and 15. With the exception of models n. 1 and 6, the difference between these central measures is within 1 % tolerance. Both the lowest AUC of 49.7 % and consequently the biggest maximum-minimum spread was obtained using model n. 3. The highest score of 80.8% and the smallest lowest maximum-minimum spread was

obtained using model n. 14. The lowest standard deviation was linked with model n. 13, the highest with model n. 12. Out of 15 compared distributions, kurtosis of 9 lay between 2 and 4. All except 2 of the models had skewness between -1 and 1. Compared with a normal distribution, the obtained distributions tended to have positive excess kurtosis and to be negatively skewed.

Table 13: Breakdown of baseline performance

Model	Median	Est. method	NA values	Origin
14	0.7885	Logit	3%	Foreign
10	0.7635	Logit	8%	CZ
4	0.73375	MDA	62%	CZ
15	0.72815	Probit	8%	Foreign
12	0.72675	Univariate	1%	Foreign
9	0.71535	Logit	8%	CZ
5	0.7021	MDA	8%	CZ
2	0.69365	MDA	40%	CZ
7	0.68885	MDA	2%	CZ
11	0.6774	Logit	2%	CZ
8	0.66215	Logit	58%	CZ
13	0.63455	MDA	1%	Foreign
3	0.6087	MDA	62%	CZ
1	0.6001	MDA	0%	CZ
6	0.57435	MDA	8%	CZ

With the individual model ranking out of the way, the discussion about a few stylized facts is due. Hand in hand with the empirical CSM literature, these simple observations serve as a starting point for analysis carried out in the following subchapters. As depicted in the figure, among the top 5 highest ranking models, 3 derive their results using conditional probability models. On the other side of the ranking, 4 of the bottom 5 models employ MDA. Grouping the probit and logit models under the header of conditional probability models on one side and the univariate and MDA models as models derived using OLS on the other allows to establish a comparison between the estimation methods. The former group of models has an average AUC of 72.3 %, outperforming the latter by 6.0 % AUC. This comparison is in line with the theoretical justification behind conditional probability models and also the empirical literature on the subject. However, properly answering the question which estimation method is superior requires an experimental design, where each set of explanatory variables characteristic for different models is estimated using all of the compared statistical methods. Simply concluding that the conditional probability models provide better predictive power would be a mistake.

As demonstrated by their slightly negative correlation coefficient of -11.6 %, the median AUC and the proportion of missing values seem to be locked in a negative

relationship. On one hand such result could have been expected. A model derived from a larger dataset should theoretically better reflect the coefficients of a model derived from the entire population. One would therefore expect missing values to negatively affect the predictive performance and the stability of the results. With the correlation coefficient of 22.9 % between the model specific standard deviation and the proportion of missing values, the expectation is reaffirmed. On the other hand, there are reasons to believe that the relationship between proportion of missing values and the predictive performance is in fact positive. If the observations are omitted in a way that they exclude observation, for which the predictive performance is lower, the omission of observations may in fact inflate the overall results. In this regard, for the bottom 10 % of smallest firms contained in the dataset, there are 6 times more missing values than for the top 10 %. As many of the compared models were not originally estimated using micro firms, which form the bulk of the Czech non-financial market and the dataset of this thesis, the inclusion of these companies may in fact prove to have a negative effect on the predictive performance. The true impact of missing values should be evaluated based on an experimental design, where the missing values are imputed. To judge the impact of missing values on the predictive performance it then suffices to compare the model specific predictive performance before and after the missing values are dealt with. Although an improvement due to missing value imputation is expected, with the third best performing model only employing around 38 % of the dataset the results promise to be interesting.

Lastly, out of the 4 foreign models included in the thesis, 3 placed in the top 5. It is not surprising then that averaging the median AUC across foreign and Czech models, gives 71.9% AUC and 67.5 % AUC respectively. Since the models differ only in estimation method and explanatory variable composition, the difference can be partially explained in light of the preceding discussion. Czech models rely more heavily MDA estimation method. Only 4 out of 11 Czech models employ a conditional probability models, compared with 2 out of 4 for their foreign counterparts. Moreover, the Czech CSM composition generates on average 23.5 % missing values, compared with only 3.3 % generated by the foreign benchmark models. If the intuition about the impact of estimation method and the missing value treatment on the predictive performance is correct, optimizing the estimation setup should lead to an overall improvement of predictive performance of Czech models.

5.2 Estimation setup optimization

5.2.1 Missing values treatment

The treatment of missing values is a topic which receives little attention in the foreign CSM literature. Although the CSM overview by (Balcaen and Ooghe, 2006) recognizes missing values as a problem connected with accounting-based models, they only briefly mention potential solutions. Even its source, (Tucker, 1996) only mentions potential solutions, such as mean, OLS and random imputation, rather than providing an inquiry into their effectiveness. If the issue receives little attention, it may be because it is not that common with bigger public firms, which are commonly used in foreign CSM literature. Indeed, even in the dataset employed by this thesis, for top 10 % of biggest firms there are on average only 0.9 missing values per firm across the 15 compared models. The corresponding figure for the bottom 10 % is 5.2, almost 6 times higher. Unlike foreign research, Czech CSM can tap into the rich resource of publicly available accounts of smaller firms. The issue therefore needed to be analyzed and dealt with.

The analysis focused on 4 potential treatments, namely imputation by OLS, by mean, by median and lastly by assets and sales. For the purposes of the evaluation, they were labeled *OLS*, *Mean*, *Med* and *A&S* respectively. On top of the 4 treatments, *Omit* was included as control, representing the baseline estimate where the missing values were simply omitted. With the exception of OLS imputation, the 3 imputation methods were made NACE sector specific, to help reflect the variation across industries. This meant that instead of a single sample mean for example, the missing values were imputed according to their NACE sector mean. The imputation by assets and sales needs further explanation. Since the firms were selected on the basis of having non-zero assets and sales, these financial statement metrics could be used to impute other missing metrics. In case of imputation by assets, first the average ratios of individual balance sheet items over assets were calculated for all NACE sectors. To impute a missing balance sheet item for a given firm, the NACE corresponding ratio was multiplied by the firm specific assets. The explanatory variables were then computed based on the available balance sheet items and the imputed values. The process was identical for imputation by sales with the exception of imputing profit and loss statement items.

The share of missing values in the sample ultimately depends on the explanatory variable selection. The distribution of missing values is far from being uniform across the compared models. The baseline estimation of the model n. 3, the gravest perpetrator, omits 62 % of the observations whereas model n. 1 employs 100 % of the

dataset. This uneven distribution of missing values must be taken into account, when interpreting the cross-model comparisons.

Figure 8:

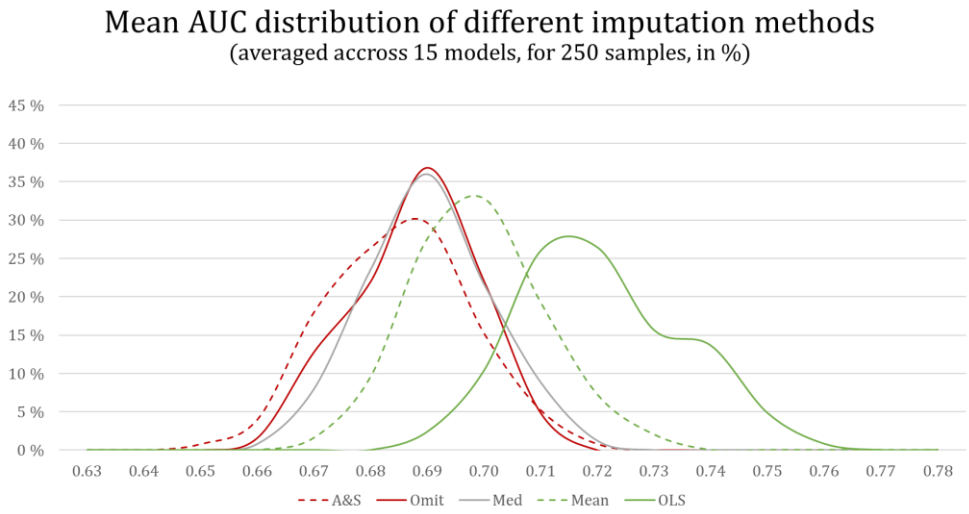


Table 14: AUC summary statistics for different missing value treatments

Treatment	Median	Mean	Min	Max	Std.	Kurt.	Skew.
OLS	72.0%	72.1%	68.7%	75.8%	0.0143	2.52	0.25
Mean	69.8%	69.9%	67.3%	73.3%	0.0117	2.94	0.27
Med	69.0%	69.0%	66.3%	72.5%	0.0113	2.96	0.19
Omit	68.8%	68.8%	65.8%	71.4%	0.0107	2.89	-0.12
A&S	68.5%	68.5%	64.7%	72.4%	0.0129	3.10	0.10

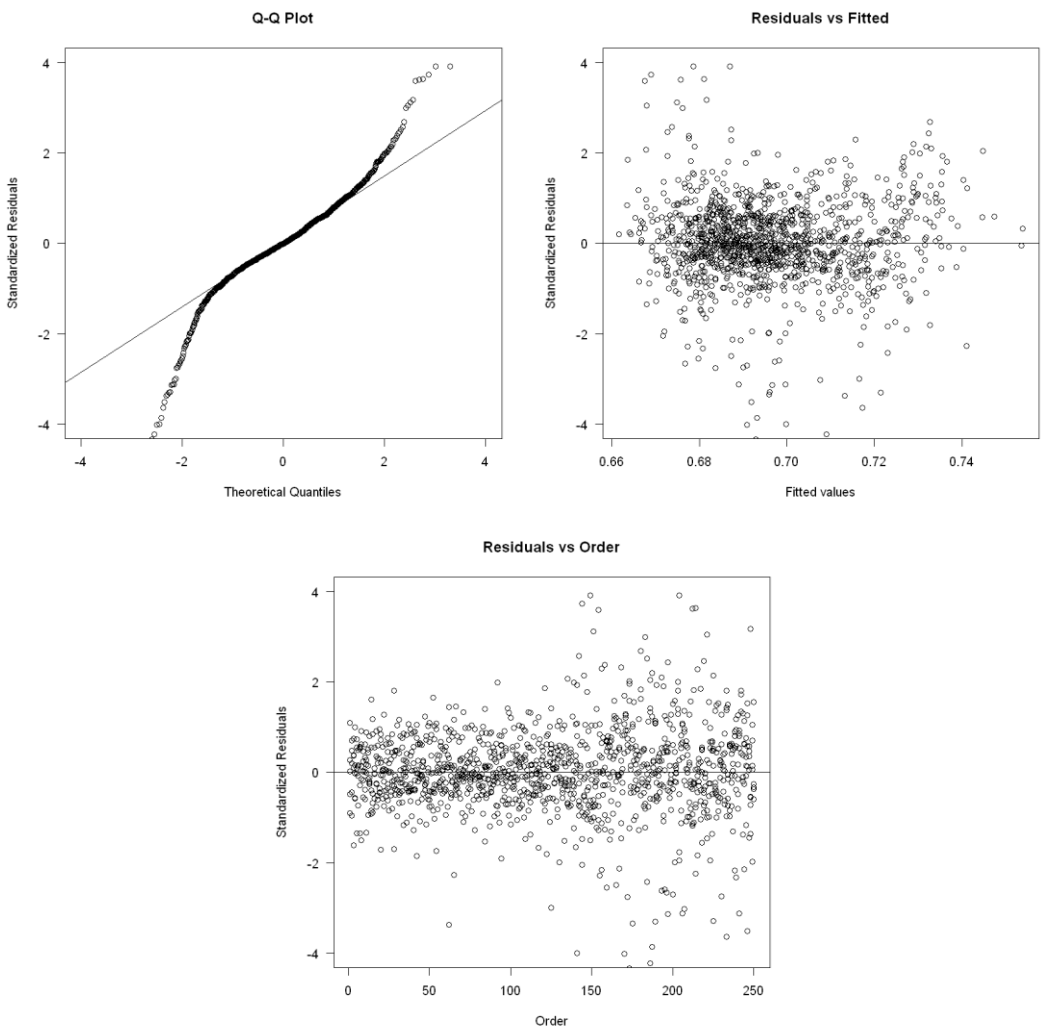
A clear winner emerges in comparing the different imputation techniques on the basis of predictive power. On average, the imputation by OLS improves AUC by 3.2 %. Replacement missing values by means, the second closest contender provides an improvement of around 1.0 %. Median imputation provides only a marginal AUC increase of around 0.2 %. The results for assets and sales imputation are somewhat surprising. This sophisticated technique has the lowest predictive power, underperforming the baseline by 0.2 %.

The graphical representation and low skewness scores hint at a bilaterally symmetrical distribution. Consequently, the above-mentioned conclusions hold for mean just as well as for the median with a 0.1 % tolerance. Compared with the control, the missing value treatment methods come with a slightly higher standard deviation, pointing to uneven improvement across models. Since the proportion of missing value varies widely across the models, this result is in line with the expectations.

After this initial discussion about the overall results, what remains to be seen is whether the changes in median predictive power due to different treatment methods are

statistically significant. Such conclusions can be drawn using either the parametric ANOVA or the non-parametric Friedman test. ANOVA is preferred, if the assumptions upon which it relies are satisfied. The following paragraphs therefore discuss the assumptions of error normality, equal variance and independence using the diagnostic plots as a basis.

Figure 9: Diagnostic plots for different imputation methods



First, the assumption that errors are normally distributed needs to be checked. In this regard, the Q-Q plot is useful in detecting deviations from normality. A normally distributed random variable should be distributed close to the 45 ° line representing the equality between observed and theoretical quantiles. In the Q-Q plot at hand, the residuals at both extremes deviate from the diagonal line. Since the absolute value of the standardized residual surpasses that of the theoretical quantiles, the distribution is most likely heavy-tailed. At the same time, the distribution seems concave rather than convex, hinting at a negatively skewed distribution. This intuition is supported by the kurtosis statistic sitting at 8.57 and the skewness statistic at -0.83. In an effort to

mitigate the violation of normality, logarithmic, square and cubic root, as well as 2nd and 3rd power transformations were applied. None of these transformations reduced the kurtosis below 4. In case of a leptokurtic distribution, (Dean et al., 2017) suggest employing non-parametric tests and so in the end the inference was drawn using the Friedman and Nemeyi tests.

Another ANOVA assumption, that of constant error variance can be visually assessed using the residuals vs fitted plot. In essence, the higher the variance, the wider the spread of standardized residuals for a given fitted value. The assumption of constant variance then requires that the variance does not change across the range of fitted values. Due to violation of normality, the non-parametric Friedman test was preferred. Whereas homoskedasticity is essential for ANOVA, the assumption is not required for the Friedman test and so the assumption was not checked.

The assumption of error independence is critical for both ANOVA and Friedman test. In experiment design, the temporal or cross-sectional differences in sampling may affect the results. The last plot, which display standardized residuals plotted against individual block allows to check for apparent error dependencies. A priori, no such dependencies are expected from the cross-sectional point of view, as each block contains a set of completely identical sample. Any temporal dependency would show if the process of random sampling by which the samples were obtained was somehow contaminated. A tell-tale sign would be a linear relationship between the order in which the blocks were created and the resulting standardized residuals. Looking at the plot, the residuals are spread equally around the zero line, except for a few outliers. In conclusion, the random sampling seems to have worked correctly and no apparent violation of error dependency occurred.

Table 15: Friedman and Nemeyi tests for missing value treatments

Friedman test

χ^2	df	p-value
756.3	4	0.0000

Nemeyi test

Treatments compared	Median diff.	p-value
OLS-Omit	0.0321	0.0000
Mean-Omit	0.0104	0.0000
Med-Omit	0.0024	0.0003
A&S-Omit	-0.0024	0.9972
OLS-Mean	0.0217	0.0000

The results of Friedman test lead to the rejection of H_0 of equal treatment effects at 1 %, meaning that amongst the different treatments, at least two effect the predictive

performance differently. To analyse the significance of individual pairs of treatments, the post-hoc Nemeyi test was carried out. At 1 % significance level, the only insignificant difference is the one between assets and sales imputation and omission of the missing value. As a consequence, OLS imputation was found to be the best performing optimization method.

5.2.2 Statistical methods

When constructing a CSM, every researcher is faced with the question of selecting a statistical method. Whilst exploring this question, we are going to limit ourselves to the comparison of MDA, logit and probit models. The comparison is warranted not only due to the popularity of these statistical methods, but also due to their interchangeability. Indeed, modern statistical software allows to switch between these methods with ease. As mentioned before, the introduction of conditional probability models was motivated by theoretical shortcomings of MDA models. Remains to be seen whether this theoretical superiority translates into a significantly higher predictive performance in practice.

On top of the papers mention if literature overview, namely (Aziz and Dar, 2006) and (Jackson and Wood, 2013), probably the most relevant answer to the question in the context of Czech Republic can be found in (Dvořáček et al., 2012b). The authors find that logit with 95.24 % APCP outperforms the MDA with 90.48 % APCP. The overall sample size of 186 firms and the fact that the study only compares the estimation methods for one model leaves a lot to be desired. By estimating multiple models using all three statistical methods on a common set of large samples, this thesis eliminates shortcomings of all the comparisons mentioned so far. In the estimation, all the missing values were omitted. The following graph and table summarise results of this part of analysis.

Figure 10:

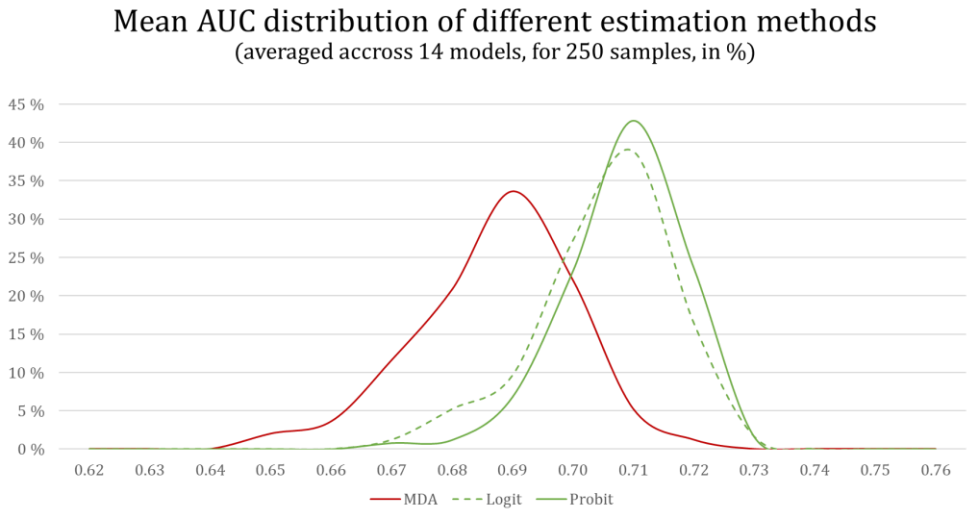
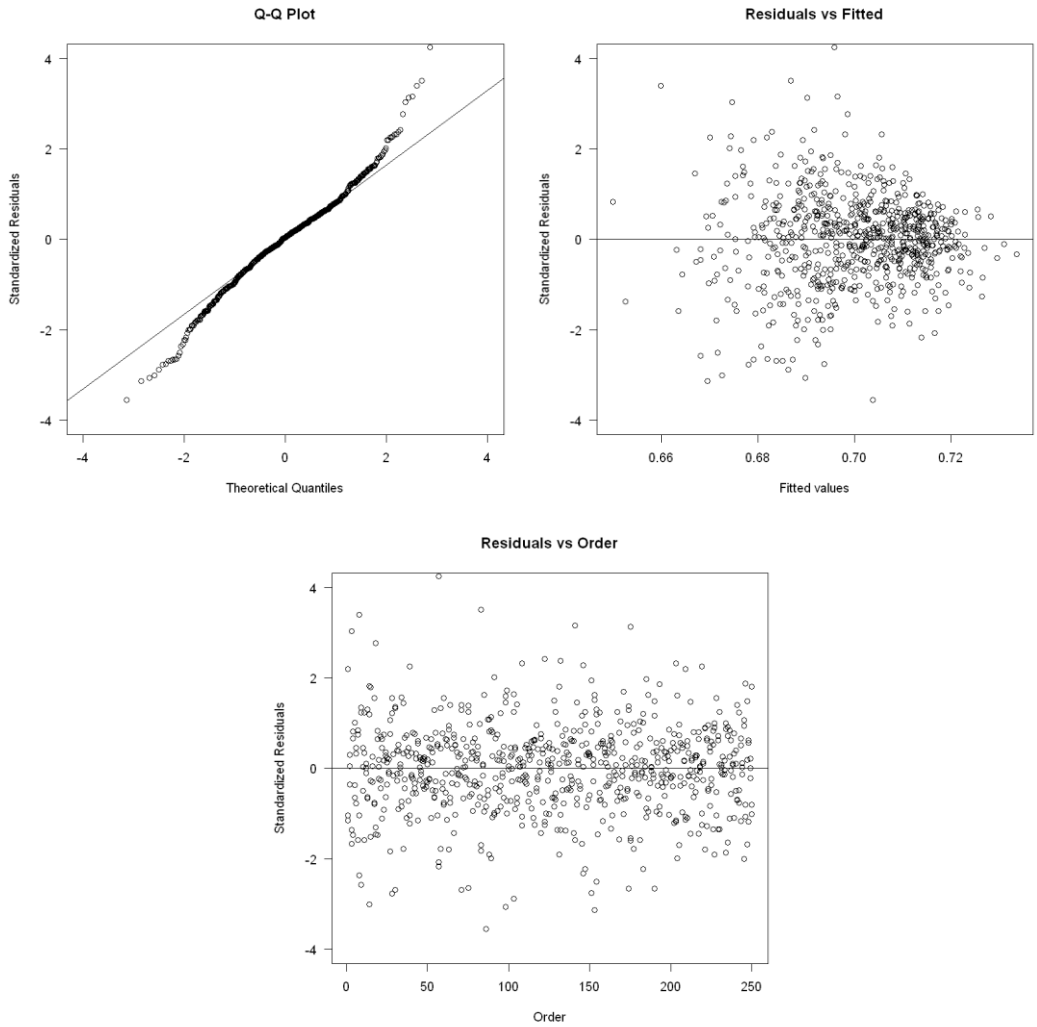


Table 16: AUC summary statistics for different estimation methods

Treatment	Median	Mean	Min	Max	Std.	Kurt.	Skew.
Probit	70.9%	70.8%	66.8%	73.2%	0.0102	4.31	-0.79
Logit	70.6%	70.5%	66.6%	73.0%	0.0109	3.56	-0.71
MDA	68.8%	68.7%	64.5%	71.9%	0.0128	3.35	-0.35

Once again, the following conclusions are similar even using arithmetic average, the alternative central measure. Based on the preliminary results, general conditional probability models seem to present a better suited alternative for CSM when compared with MDA. The former entails an approximately 2 % higher average AUC. In a mutual comparison, conditional probability models are neck and neck, with probit providing a 0.27 % increase in predictive power. These findings are in line with previous empirical research by (Aziz and Dar, 2006) and does justice to the theoretical justification of their usage. A question is whether these economically significant difference also prove to be statistically significant

Figure 11: Diagnostic plots for different estimation methods



Looking at the normal Q-Q plot, the extreme values diverge from the diagonal. Once again, the absolute value of these extreme values is higher than the value of corresponding theoretical quantiles, pointing to a leptokurtic distribution. The distribution does not seem particularly concave or convex and so if present, skewness is negligible. These insights are confirmed by the higher than normal kurtosis of 4.67 and the low skewness of 0.05. Applying the aforementioned battery of transformations does not address the problem and so in the end the distribution of residuals is deemed not normal. Just as before, verifying the equal variance assumption is therefore of little interest. Suffices to say that residuals vs fitted plot displays a textbook case of a heteroskedastic “megaphone”. Before applying the non-parametric Friedman test, the assumption of error independence needed to be checked. The residual vs order does not display any apparent dependency as save for a few outliers, the standardized residuals are spread uniformly around the 0.

Table 17: Friedman and Nemeyi tests for estimation methods

Friedman test		
χ^2	df	p-value
414.1	2	0.0000

Nemeyi test		
Treatments compared	Median diff.	p-value
Probit-MDA	0.0218	0.0000
Logit-MDA	0.0185	0.0000
Probit-Logit	0.0033	0.0000

The Friedman H_0 of equal treatment effect could be rejected at 1 % significance level. The post-hoc analysis comparing individual means bared similar results. All the previously mentioned comparisons are statistically significant at 1 %. Overall, MDA is outperformed by both conditional probability estimation methods. Probit outperforms logit slightly but statistically significantly and can therefore be considered the best performing estimation method.

5.2.3 Non-financial variables

Whether based on accounting or the market data, CSM rely on financial information, most commonly in the form of financial ratios. However, using financial ratios and indicators exclusively could be justified only if they contained all the information relevant to bankruptcy prediction. Otherwise their omission causes lower predictive performance (Sun et al., 2013). In fact, a body of literature suggests that the predictive performance of CSM can be improved by including non-financial information.

A fairly extensive list of papers employing non-financial variables can be found in (Balcaen and Ooghe, 2006). The authors split these into firm-specific and external factors. Among the former, the authors list such factors as interest rates, industry growth rate or business cycle stage. The latter are said to include company age, size or industry classification. Yet other scholars revoke biases in managerial decisions explored by behavioural economics and, without being too specific, call for inclusion of "psychological phenomena" (Constand and Yazdipour, 2011).

The research into inclusion of these non-financial factors was at least partly brought about by the lack of publicly available financial statements for small and medium enterprises (Balcaen and Ooghe, 2006). This is not the case for the Czech Republic where anybody can easily obtain access to financial statements of thousands of Czech firms. Moreover, the above mentioned non-financial variables can easily be constructed from public sources or from the very same databases the authors use to

construct datasets for their models. However, all the Czech CSM taken into consideration fail to take notice of non-financial information and hence potentially hinder their predictive performance.

Amongst the large battery of potential non-financial variables, this thesis compared 4, easily accessible measures. First, *Age* captures the age of the company in years at the moment when the financial statement was elaborated. Second, *JS* is a dummy variable reflecting the legal form of a company. Limited liability companies are flagged with 0, joint-stock companies take the value of 1. Third, *Avg_b* variable is connected with historical NACE specific bankruptcy prevalence. It takes value of 1 if a firm operates under the B, C, F, G, H or I sector classifications, which have an above average historical bankruptcy prevalence. Fourth, *Unem* is the percentage unemployment year, as reported by CSO. In the evaluation, these 4 variables were accompanied by *None*, the control variable, standing for no non-financial variable added to the baseline.

On top of the ultimately included non-financial variables, others were also taken into consideration. First, such variable was the firm size category. The generally recognised categorisation relies on the number of employees. Since Magnus Web contained this information only for 44 % of the firms, the non-financial variable was not included. Second, the GDP growth and 2W repo were considered. Ultimately, neither of these variables were included in the comparison due to their low predictive performance. Averaged across models it equalled only 59.8 % AUC and 60.8 % AUC, underperforming the baseline by over 8 % AUC. Since the aim of comparison was to establish the best performing methods, there was little interest in including these variables into the comparison.

Figure 12:

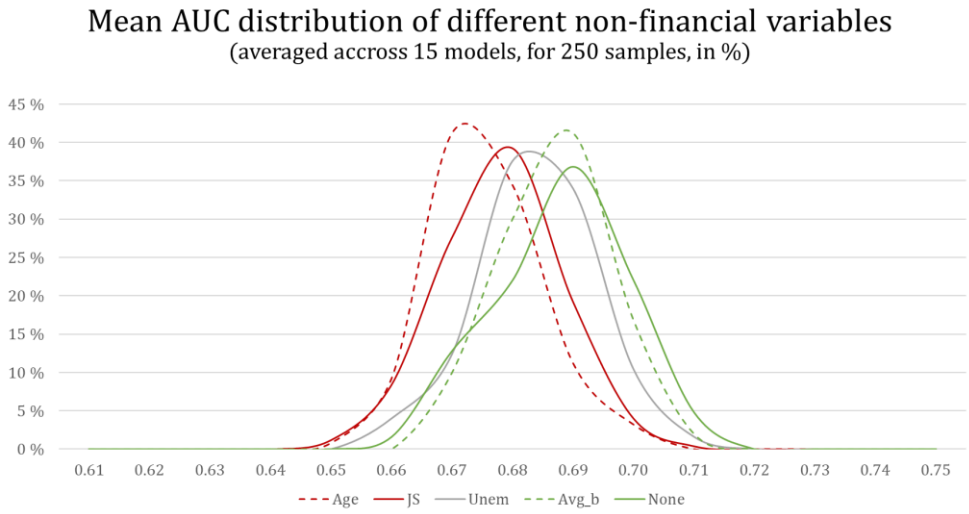
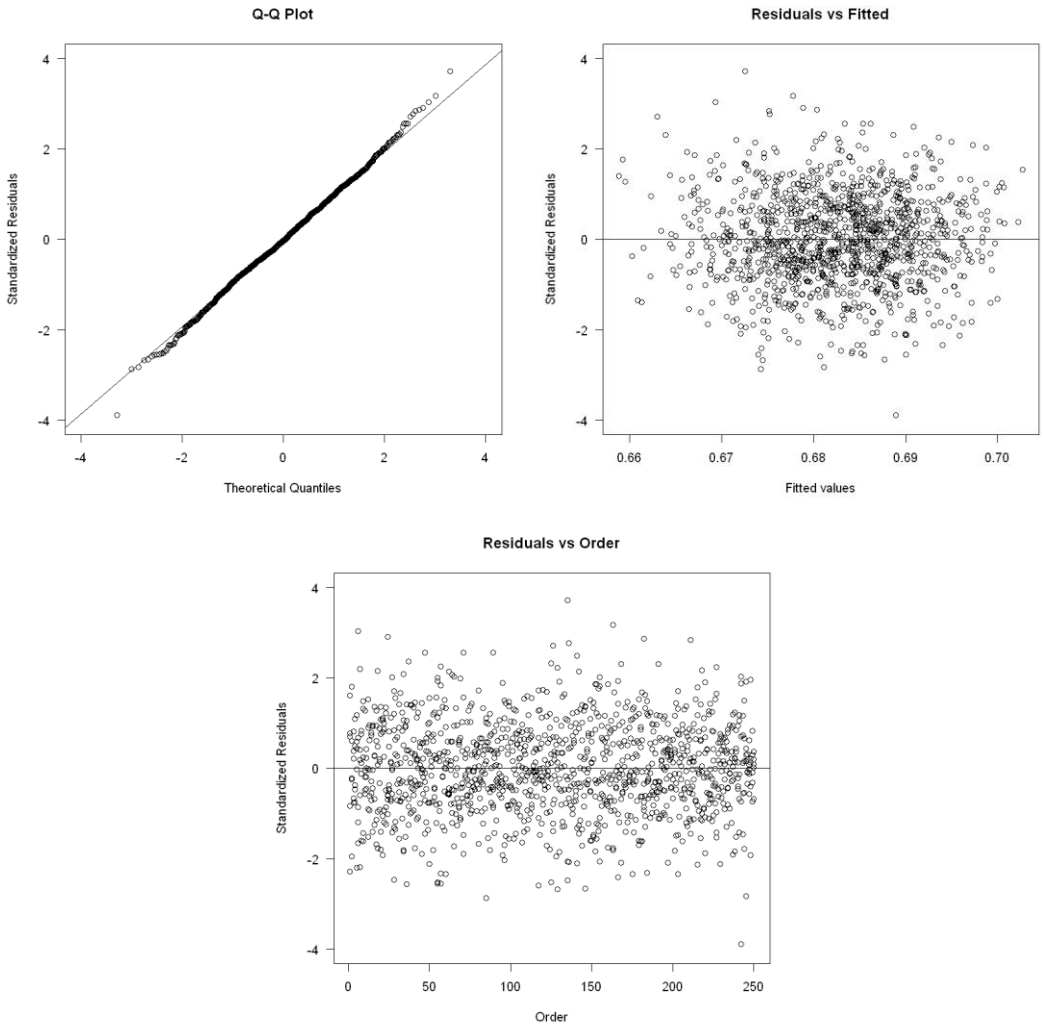


Table 18: AUC summary statistics for different non-financial variables

Treatment	Median	Mean	Min	Max	Std.	Kurt.	Skew.
None	68.8%	68.8%	65.8%	71.4%	0.0107	2.89	-0.12
Avg_b	68.8%	68.7%	66.6%	70.9%	0.0089	2.58	-0.07
Unem	68.4%	68.4%	65.7%	70.7%	0.0095	3.34	-0.21
JS	67.8%	67.8%	65.1%	70.9%	0.0100	2.96	-0.06
Age	67.5%	67.6%	65.3%	70.1%	0.0090	2.94	0.22

With the median AUC of 68.8 %, the control ever so slightly outperforms the closest contender Avg_b, the difference being just 0.02 %. Since the remaining treatments underperform the control by a larger margin, the results indicate that none of 4 additional non-financial variables increase the predictive performance. These conclusions hold irrespective of whether mean or median is used.

Figure 13: Diagnostic plots for different non-financial variables



Judging by the Q-Q plot, residuals seem to be normally distributed. The kurtosis and skewness statistics equal 3.26 and 0.01 respectively. With a p-value of p-value 0.07, the residuals even pass the Shapiro-Wilk normality test at 5 % significance level. Since the residuals are equally spread across the fitted values, one can conclude that the assumption of equal variances is met. The same can be said for the assumptions of error independence. The average residuals sit at 0 for all the compared treatment methods. Although unnecessary at this point, it can also be concluded that the random sampling did not introduce any apparent error dependency.

Table 19: ANOVA, Friedman and post-hoc tests for non-financial variables

ANOVA			Friedman test		
F	Df	p-value	χ^2	Df	p-value
104.6	4	0.0000	320.9	4	0.0000

Tukey test		
Treatments compared	Mean diff.	p-value
Avg_b-None	-0.0005	0.9613
Unem-None	-0.0035	0.0000
Age-None	-0.0096	0.0000
JS-None	-0.0117	0.0000
Avg_b-Unem	0.0030	0.0004

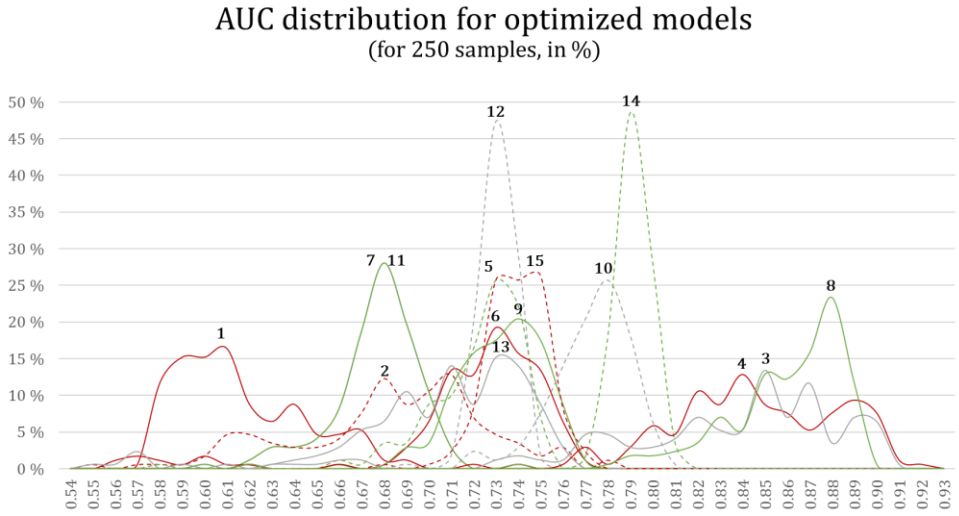
Nemeyi test		
Treatments compared	Median diff.	p-value
Avg_b-None	-0.0000	0.5261
Unem-None	-0.0034	0.0001
Age-None	-0.0096	0.0000
JS-None	-0.0127	0.0000
Avg_b-Unem	0.0034	0.0425

Both the parametric ANOVA and non-parametric Friedman test lead to the rejection of their H_0 at 1 % significance level. Accordingly, the interpretation of the results for both post-hoc tests at 5 % significance level is identical. Only the difference between the Avg_b and the control seem to be insignificant. Since in both of the comparisons, the difference between Avg_b and control was negative, no the financial variables was added to the optimized models.

5.3 Comparison of optimized models

After identifying avenues of estimation setup optimization, the highest performing treatments were put to the test. All of the 15 baseline models were re-estimated using probit and OLS imputation. The following figure visualises the resulting AUC distribution. Judging by the graphical representation, model n. 8 attained the highest predictive performance whilst model n. 1 attained the lowest predictive performance. Note that since both model n. 7 and 11 were estimated using probit, their respective AUC distributions are identical. Compared with the baseline estimation, the former best performing baseline model number 14 seems outperformed by 3 different models. Even after optimization, most of the distributions still look centrally symmetrical.

Figure 14:



As before, the model ranking was done on the basis of median AUC, once the summary statistics were computed. Both median and mean confirm the intuition that model n. 14 was outperformed by models n. 3, 4 and 8. The highest scoring model n. 8 reached a median AUC of 86.6 %, improving on the baseline model n. 14 estimate by almost 8 %. The worst performing amongst compared models was the model n. 1 with the median AUC of 61.1%. As a consequence, the spread between the best and worst median AUC increased from 25.5 % to 21.5 %.

Table 20: Summary statistics for the optimized CSM comparison

Model	Median	Mean	Min	Max	Std.	Kurt.	Skew.
8	86.6%	85.8%	62.1%	89.8%	0.0325	19.46	-2.92
3	84.4%	82.0%	55.4%	90.9%	0.0755	6.39	-1.81
4	84.2%	84.3%	61.9%	91.5%	0.0434	7.87	-1.40
14	79.2%	79.1%	76.7%	81.2%	0.0075	3.16	0.05
10	77.5%	77.2%	69.2%	80.9%	0.0182	5.17	-1.06
15	74.0%	73.9%	68.4%	77.4%	0.0134	4.03	-0.48
9	73.3%	73.2%	65.1%	76.8%	0.0190	3.55	-0.58
13	73.1%	71.2%	71.2%	75.1%	0.0075	2.80	-0.12
6	72.8%	71.8%	55.7%	76.8%	0.0430	8.56	-2.37
5	72.7%	72.1%	58.1%	75.1%	0.0231	11.71	-2.17
12	71.5%	73.1%	59.0%	76.0%	0.0303	3.70	-0.82
2	69.0%	68.4%	55.1%	77.6%	0.0429	3.01	-0.48
7	68.0%	67.7%	60.2%	71.3%	0.0189	4.70	-1.03
11	68.0%	67.7%	60.2%	71.3%	0.0189	4.70	-1.03
1	61.1%	61.6%	57.3%	69.3%	0.0279	2.61	0.68

Note: Models ordered based on median AUC rank

Once again, the biggest maximum-minimum spread of 35.5 % was achieved by model n. 3 which also accounted for the overall highest AUC of 90.9 %. The smallest maximum-minimum spread could be observed with the model n. 13 at just 3.9 % AUC. Although the average standard deviation increased by 0.03, 10 model saw their standard deviation decrease. As for the remaining 5 models, models n. 3 and 4

accounted for a combined 80 % of the increase in standard deviation. Unsurprisingly, both the kurtosis and skewness statistics increased significantly. Whereas before the majority of compared distributions had kurtosis between 2 and 4, only 6 fulfulied this criteria after optimization. More importantly, the average skewness increased approximately by factor of 4. This thesis relies on median as the main metric by which the results are interpreted. In presence of non-zero skewness mode is generally superior to median which is in turn superior to mean as a measure of the central tendency. To see whether the resulting ranking was robust to different central measures, the median ranking was compared with the alternative mode and mean rankings.

Table 21: Comparison of model ranking based on different central measures

Mode	Median	Mean
8	8	8
3	3	4
4	4	3
14	14	14
10	10	10
15	15	15
9	9	9
5	13	12
6	6	5
13	5	6
12	12	13
2	2	2
7	7	7
11	11	11
1	1	1

Overall, the ranking derived using median can be relied upon. The only difference in rank occurred with the models n. 13 and 5 which is no cause for alarm. The difference in medians of these two models is just 0.4 %. With 250 sample sets, the mode ranking needed to be derived by rounding up the results to 2 decimal points. Consequently, the resulting difference can potentially be attributed to a rounding error.

Finally, the performance of Czech and foreign benchmark models could be contrasted. First, contrasting the baseline results allowed for a comparison of Czech and foreign CSM as described by the CSM literature. Second, the contrast between baseline results and optimized results for either Czech or foreign models allowed to judge relative importance of using the optimization treatments in a predictive context. Third, the contrast between the optimized results allowed for a more objective comparison between the Czech and foreign models by controlling for the contamination of the baseline estimation due to different estimation methods and proportions of missing values.

Figure 15:

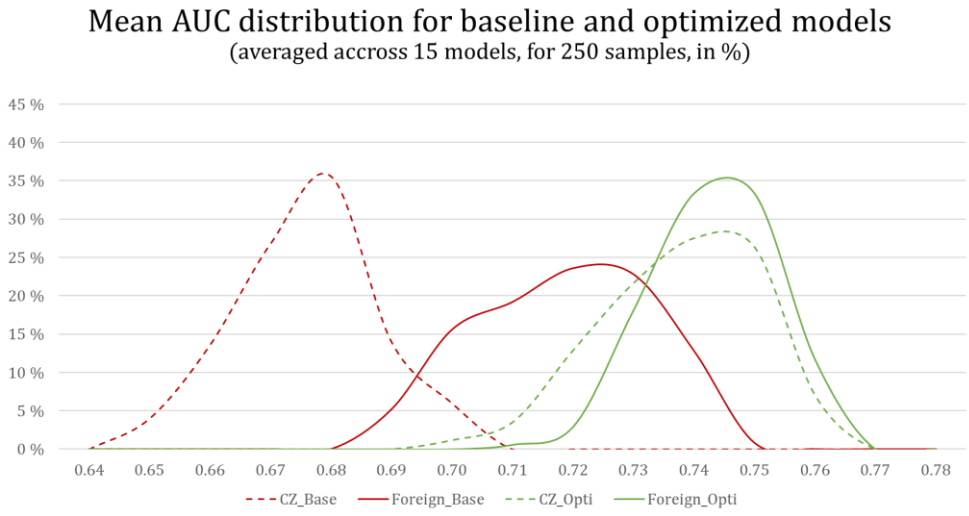
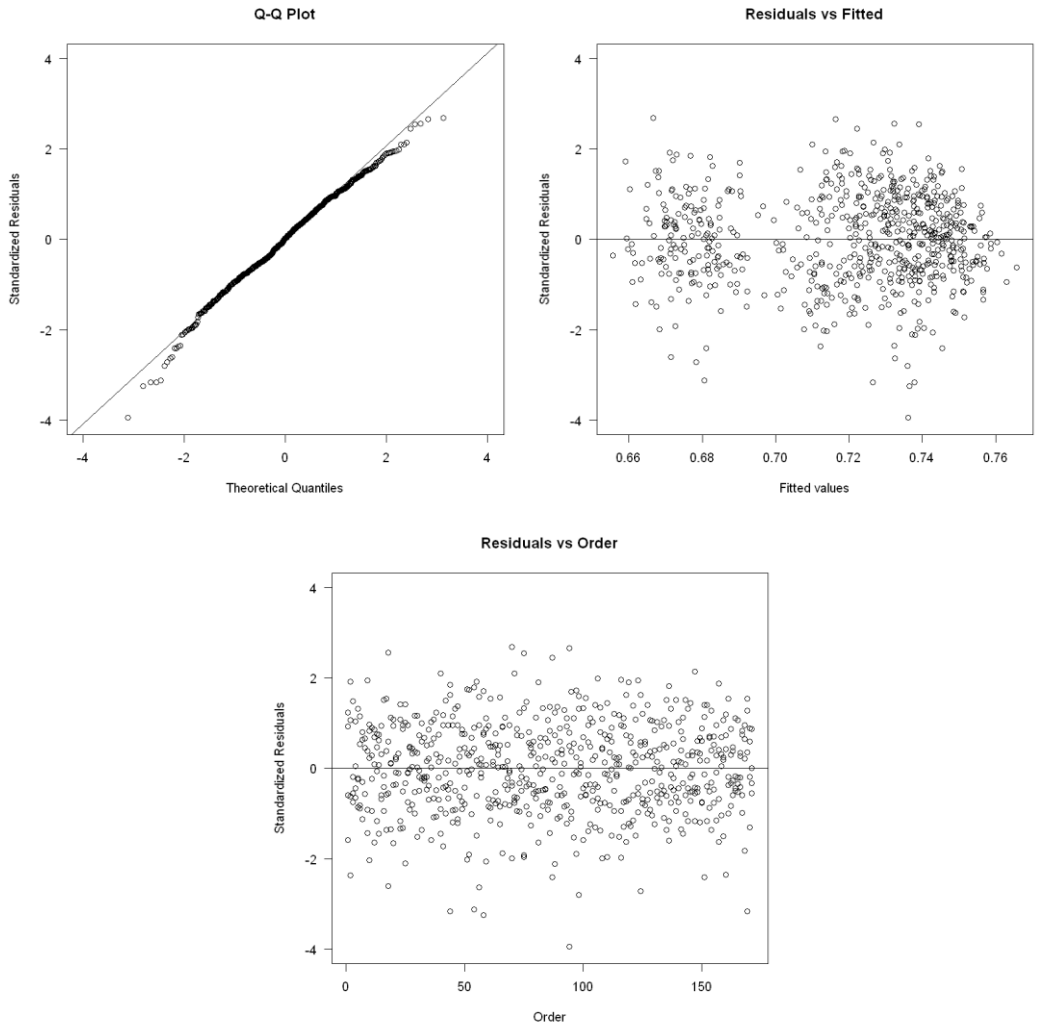


Table 22: AUC summary statistics for baseline and optimized models

Treatment	Median	Mean	Min	Max	Std.	Kurt.	Skew.
Foreign_Opti	74.5%	74.3%	70.9%	76.4%	0.0100	2.90	-0.37
CZ_Opti	73.9%	73.8%	70.0%	76.1%	0.0126	2.84	-0.52
Foreign_Base	72.0%	71.9%	68.6%	74.8%	0.0141	2.15	-0.16
CZ_Base	67.7%	67.6%	64.5%	70.5%	0.0114	3.11	-0.09

In the baseline comparison, Czech models are outperformed by their foreign counterparts by about 4.1 %. Optimization the Czech models improves their median predictive performance by about 6.2 %. The corresponding improvement of foreign models equals only 2.5 %. After optimization, the difference between the two groups of models shrinks down from 4.1 % to less than 0.6 %. Same as before, the differences between median AUC of different treatments do not differ by more than 0.1 % from the differences between means. All the excess kurtosis and skewness measures deviate form 0 by no more than 1, speaking in favour of the normality.

Figure 16: Diagnostic plots for baseline and optimized models



Before estimating the statistical significance of individual contrasts, the usual diagnostic plots need to be discussed. All the residuals lie on or close to the diagonal of the Q-Q plot meaning that their distribution can be considered approximately normal. A curve going through the residuals would be concave and so the distribution is most likely negatively skewness. Kurtosis equals 3.32 and skewness equals -0.27, further strengthening the argument in favour of normality of the errors. Moving on to the plot displaying residuals vs fitted, save for a few outliers the residuals are distributed uniformly around 0 for the whole range of fitted values. Consequently, the assumption of constant error variance can also be considered to hold. Since both of its assumptions about the error distribution are satisfied, the inference was drawn using ANOVA. The results of Friedman test were also provided for more robust conclusions. Finally, the visual inspection of the residuals plotted against the order does not reveal any apparent relationship. Once again one can conclude that the independence of errors is satisfied.

Table 23: ANOVA, Friedman and post-hoc tests for baseline and optimized models

ANOVA			Friedman test		
F	Df	p-value	χ^2	Df	p-value
1579.0	3	0.0000	429.9	3	0.0000

Tukey test

Treatments compared	Mean diff.	p-value
Foreign_Base-CZ_Base	0.0423	0.0000
Foreign_Opti-Foreign_Base	0.0247	0.0000
CZ_Opti-CZ_Base	0.0617	0.0000
Foreign_Opti-CZ_Opti	0.0053	0.0000

Nemeyi test

Treatments compared	Median diff.	p-value
Foreign_Base-CZ_Base	0.0431	0.0000
Foreign_Opti-Foreign_Base	0.0248	0.0000
CZ_Opti-CZ_Base	0.0622	0.0000
Foreign_Opti-CZ_Opti	0.0057	0.0200

Using both the parametric ANOVA and non-parametric Friedman test, the H_0 of equal treatment effects can be rejected at 1 % significance level. After carrying out the post-hoc Tukey and Nemeyi tests, H_0 of equal treatment effects for the displayed differences can be rejected at 5 % significance level. These tests confirm the conclusion, that in their baseline form, foreign CSM outperformed Czech CSM, that the optimization improves the performance of Czech CSM more and that in their optimized form, the predictive performance of the two groups is only marginally different.

Table 24: Rank comparison of baseline and optimized models

Model	Origin	Median		Optimized-Baseline		Rank	
		Baseline	Optimized	Median diff.	p-value	Baseline	Change
8	CZ	66.2%	86.6%	0.2043	0.0000	11	↑10
3	CZ	60.9%	84.4%	0.2356	0.0000	13	↑11
4	CZ	73.4%	84.2%	0.1084	0.0000	3	0
14	Foreign	78.9%	79.2%	0.0033	0.0000	1	↓3
10	CZ	76.3%	77.5%	0.0119	0.0000	2	↓3
15	Foreign	72.8%	74.0%	0.0117	0.0000	4	↓2
9	CZ	71.5%	73.3%	0.0176	0.0000	6	↓1
12	Foreign	72.7%	73.1%	0.0041	0.0000	5	↓3
6	CZ	57.4%	72.8%	0.1537	0.0000	15	↑6
5	CZ	70.2%	72.7%	0.0254	0.0000	7	↓3
13	Foreign	63.2%	71.5%	0.0824	0.0000	12	↑1
2	CZ	69.3%	69.0%	-0.0038	0.0390	8	↓4
7	CZ	68.9%	68.0%	-0.0084	0.0000	9	↓4
11	CZ	67.7%	68.0%	0.0031	0.0000	10	↓4
1	CZ	60.0%	61.1%	0.0108	0.0002	14	↓1

To understand which of the compared models were most affected by the optimization, a rank comparison contrasting the baseline and optimized models was prepared. Apart

from summarizing the already discussed information, 15 individual Friedman tests were carried out to see whether the results obtained after optimization are statistically significantly different from the baseline mode. Except for model n. 2, the difference between baseline median AUC and optimized median AUC was statistically significant at 1 % significance level. Among the statistically significant results, 13 models present an increase in median AUC.

However, the increase in performance was not shared uniformly across the models. Models 3, 4, 6, 8 and 13 account for 91.2 % of the overall improvement. The performance of all but one of the remaining models increased by less than 2 % AUC. If these 5 models were omitted from the ranking, with the exception of model n. 12, no models would have changed their place. The unequal effect of optimization can be partly explained by already obtained results. Comparing OLS imputation and probit estimation with their respective controls, the former improved the baseline on average model by 3.2 % whereas the latter only by 2.2 %. If we take into account that MDA, the control to the estimation method treatments, was applied only in 9 of the 15 baseline models, the expected effect of using probit should be even less. To illustrate the fact that the missing value treatment is more important of the two optimization treatments, it suffices to realizes, that the top 3 optimized models are the same models that had the highest proportion of missing values in the baseline models.

6 Conclusion

In its essence, this thesis provided a structured comparison of Czech credit scoring models (CSM) based on their predictive performance. For the subset of general industry statistical models, it established how Czech CSM fare in a comparison with other Czech CSM and with foreign CSM benchmarks, how their predictive performance can be optimized and finally how the models compare after being optimized. The importance of these questions stemmed from the fact that Czech credit scoring literature introduces a fair number of CSM whose predictive performance cannot be compared simply by contrasting the reported results. At the time of writing, there were 19 different papers introducing new and original Czech CSM. As was shown, Czech CSM are derived using small and often wildly unrepresentative samples of Czech market. Due to variable selection procedures based on empirics rather than theory and the lack of validation, the reported results run a risk of being overfitted. Finally, the results obtained from the prediction performance indicators used by the overwhelming majority of papers are affected by choices made by the researcher. All of these factors combine in making a simple comparison of reported results without much interest. Instead, with the goal of making the results generalizable, this thesis evaluated the predictive performance of the compared CSM by using an arbitrariness excluding indicator, obtained by techniques aimed at minimizing the data sensitivity from a large, long-term, representative dataset.

As far as the actual results, the baseline ranking was established first, comparing Czech and foreign CSM as defined by the original papers. The Ohlson logit model contained in (Ohlson, 1980) with the median AUC of 78.9 % was found to provide the best predictive performance. The worst median AUC of 57.4 % was achieved by the IN99 model as described in (Neumaierová, 2002). In their aggregate, Czech CSM underperformed the foreign benchmark models by 4.2 % AUC. Based on the baseline estimate, the usage of foreign CSM to predict the bankruptcy of Czech firms is just as justified, if not more, than the usage of Czech CSM.

Next, three potential avenues of the estimation setup optimization, namely missing value imputation, estimation method selection, and non-financial variable addition, were explored. Out of 4 compared missing value treatments, OLS imputation improved the baseline results by the highest margin of 3.2 % AUC. The second closest contender - mean imputation - provided only an increase of 1.0 % AUC. Next, in the comparison

of 3 estimation methods, probit significantly outperformed both logit and MDA. Compared with MDA, it improved the results by almost 2.2 %. In the comparison with logit, probit improved the baseline results only marginally but statistically significantly. Finally, 6 non-financial variables were compared in order to determine, which of them improved the results of the baseline estimation the most. It was found that the addition of none of the compared non-financial variables presented an economically and statistically significant improvement over the baseline CSM. Consequently, only OLS imputation and probit estimation were identified as treatments that in general increase the predictive performance of CSM.

Finally, the 15 baseline CSM were ranked after being optimized, that is to say after imputing the missing values by OLS and after being estimated using probit. In the optimized model ranking, JT index model, as introduced in (Jakubik and Teplý, 2011), replaced the Ohlson logit model as the best performing CSM with the median AUC of 86.6 %. The worst performing CSM was the one introduced in (Karas and Režňáková, 2013) with the median AUC of 61.1 %. Although the optimization improved the results of 13 out of 15 compared models, the improvement was not shared uniformly. After the optimization, the gap in predictive performance of Czech and foreign CSM shrank significantly to just 0.6 % AUC. This means that whereas in the baseline form one is better off by using a foreign CSM to predict the bankruptcy of Czech firms, after the optimization the difference in predictive performance is marginal. It is also important to note that only 5 CSM accounted for over 90 % of the improvement due to optimization. Given that the top 3 optimized models generated over 50 % of the missing data in their baseline form, the overall ranking has changed mainly due to OLS imputation.

The overall contribution of the thesis depends on whether the reader is a practitioner or a member of the academia. When looking for an existing CSM with the highest predictive performance, the ranking contained in this study presents a useful guide to any practitioner. The findings that using probit and OLS imputation should generally increases the predictive performance of CSM is then useful to practitioners in construction and estimation of such models. As for the academic contribution, this thesis validates already existing findings on one hand and as well as introducing new observations on the other. The findings of this thesis are just one in a long line studies providing empirical evidence in favour of the theoretically higher predictive conditional probability models over MDA, as proposed by (Ohlson, 1980). By suggesting that the omission of missing values has an economically and statistically significant negative impact on the predictive performance of CSM, the thesis could

also bring the attention of the academia to the so far overlooked issue of missing value imputation.

Bibliography

Altman, E.I., 1968. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance 23, 589–609. https://doi.org/10.2307/2978933

Aziz, M., Dar, H., 2006. Predicting corporate bankruptcy: Where we stand? Corporate Governance 6, 18–33. https://doi.org/10.1108/14720700610649436

Balcaen, S., Ooghe, H., 2006. 35 Years of Studies on Business Failure: An Overview of the Classic Statistical Methodologies and Their Related Problems. The British Accounting Review 38, 63–93. https://doi.org/10.1016/j.bar.2005.09.001

Beaver, W.H., 1966. Financial Ratios as Predictors of Failure. Journal of Accounting Research (Wiley-Blackwell) 4, 71.

Bellovary, J.L., Giacomino, D.E., Akers, M.D., 2007. A Review of Bankruptcy Prediction Studies: 1930 to Present. Journal of Financial Education 33, 1–42.

Bemš, J., Starý, O., Macaš, M., Žegklitz, J., Pošík, P., 2015. Innovative default prediction approach. Expert Systems With Applications 42, 6277–6285. https://doi.org/10.1016/j.eswa.2015.04.053

Chudson, W.A., 1945. The Pattern of Corporate Financial Structure: A Cross-section View of Manufacturing, Mining, Trade, and Construction, 1937. National Bureau of Economic Research.

Constand, R., Yazdipour, R., 2011. Firm Failure Prediction Models: A Critique and a Review of Recent Developments, in: Advances in Entrepreneurial Finance: With Applications from Behavioral Finance and Economics. pp. 185–204.

Dean, A., Voss, D., Draguljić, D., 2017. Design and Analysis of Experiments. Springer.

Dvořáček, J., Sousedíková, R., Barták, P., Štěrba, J., Novák, K., 2012a. Forecasting Companies' Future Economic Development. Acta Montanistica Slovaca 111.

Dvořáček, J., Sousedíková, R., Domaracká, L., 2008. Industrial enterprises bankruptcy forecasting. Metalurgija 33.

Dvořáček, J., Sousedíková, R., Řepka, M., Domaracká, L., Barták, P., Bartošíková, M., 2012b. Choosing a method for predicting economic performance of companies. Metalurgija 525.

Hampel, D., Vavřina, J., Janová, J., 2012. Predicting bankruptcy of companies based on the production function parameters. Presented at the 30th International Conference Mathematical Methods in Economics, Karviná, Czech Republic.

Hollander, M., Wolfe, D.A., Chicken, E., 2013. Nonparametric Statistical Methods. John Wiley & Sons.

Jackson, R.H.G., Wood, A., 2013. The performance of insolvency prediction and credit risk models in the UK: A comparative study. The British Accounting Review 45, 183–202. https://doi.org/10.1016/j.bar.2013.06.009

Jakubik, P., Teplý, P., 2011. The JT Index as an Indicator of Financial Stability of Corporate Sector. Prague Economic Papers 20, 157–176. https://doi.org/10.18267/j.pep.394

Kalouda, F., Vaníček, R., 2013. Alternativní bankrotní modely - první výsledky. Masaryk University.

Karas, M., Režňáková, M., 2014. A parametric or nonparametric approach for creating a new bankruptcy prediction model: The Evidence from the Czech Republic. International Journal of Mathematical Models and Methods in Applied Sciences 8.

Karas, M., Režňáková, M., 2013. Bankruptcy prediction model of industrial enterprises in the Czech Republic. International Journal of Mathematical Models and Methods in Applied Sciences 7, 519–531.

Kirkos, E., 2015. Assessing methodologies for intelligent bankruptcy prediction. Artif Intell Rev 43, 83–123. https://doi.org/10.1007/s10462-012-9367-6

Kocmanová, A., Dočekalová, M.P., Němeček, P., 2014. Sustainable Corporate Performance Index for Manufacturing Industry, in: 18th World Multi-Conference on Systemics, Cybernetics and Informatic, Orlando, FL, USA, July. pp. 15–18.

Koráb, V., 2001. ONE APPROACH TO SMALL BUSINESS BANKRUPTCY PREDICTION: THE CASE OF THE CZECH REPUBLIC. Presented at the 8th SIGEF CONGRESS, Naples, Italy.

Machek, O., 2014. Long-term Predictive Ability of Bankruptcy Models in the Czech Republic: Evidence from 2007-2012. Central European Business Review 3, 14–17. https://doi.org/10.18267/j.cebr.80

Machek, O., Smrcka, L., Strouhal, J., 2015. How to predict potential default of cultural organizations, in: 7th International Scientific Conference Finance and Performance of Firms in Science, Education and Practice, Zlín, Czech Republic, April. pp. 23–24.

Maxwell, S.E., Maxwell, S.E., Delaney, H.D., Kelley, K., 2004. Designing Experiments and Analyzing Data: A Model Comparison Perspective. Psychology Press.

Němec, D., Pavlík, M., 2016. Predicting insolvency risk of the Czech companies.

Neumaierová, I., 2002. Výkonnost a tržní hodnota firmy, 1. vyd. ed, Finance: [edice, Grada]. Grada.

Neumaierová, I., Neumaier, I., 2005. Index IN05. Presented at the Evropské finanční systémy, Brno, Czech Republic, pp. 143–148.

Ohlson, J., 1980. Financial Ratios And The Probabilistic Prediction Of Bankruptcy. Journal of Accounting Research 109.

Platt, H.D., Platt, M.B., 2002. Predicting corporate financial distress: Reflections on choice-based sample bias. J Econ Finan 26, 184–199. https://doi.org/10.1007/BF02755985

Prusak, B., 2018. Review of Research into Enterprise Bankruptcy Prediction in Selected Central and Eastern European Countries. International Journal of Financial Studies 6, 60. https://doi.org/10.3390/ijfs6030060

Schechtman, E., Schechtman, G., 2019. The relationship between Gini terminology and the ROC curve. Metron 77, 171–178.

Sun, J., Li, H., Huang, Q.-H., He, K.-Y., 2013. Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. Knowledge-Based Systems 57. https://doi.org/10.1016/j.knosys.2013.12.006

Tamari, M., 1966. Financial Ratios as a Means of Forecasting Bankruptcy. Management International Review 6, 15–21.

Tucker, J., 1996. Neural Networks Versus Logistic Regression In Financial Modelling: A Methodological Comparison, in: In Proceedings of the 1996 World First Online Workshop on Soft Computing (WSC1.

Valecký, J., Slivková, E., 2012. Mikroekonomický scoringový model úpadku českých podniků. https://doi.org/10.7327/cerei.2012.03.02

Vochozka, M., Rowland, Z., 2015. Prediction of the future development of construction companies by means of artificial neural networks on the basis of data from the Czech Republic. Математичне моделювання в економіці 62–76.

Vochozka, M., Straková, J., Váchal, J., 2015. Model to predict survival of transportation and shipping companies. Nase More.

Zmijewski, M.E., 1984. Methodological Issues Related to the Estimation of Financial Distress Prediction Models. Journal of Accounting Research 22, 59–82. https://doi.org/10.2307/2490859

Appendix A: Explanatory variables

The following table presents a summary of the explanatory variables, used by the compared models. Individual variables were numbered for easier referencing, based on type of variable and the number of models that employed them. Apart from these two characteristics, the table contains formulas by which the variables were defined and the proportion of observations, for which the variable could not be calculated.

Variable	Formula	# of uses	Not available	Variable type
1	$\frac{\text{Debt}_t}{\text{Assets}_t}$	6	0.0 %	Solvency
2	$\frac{\text{Equity}_t}{\text{Debt}_t}$	2	0.9 %	Solvency
3	$\frac{\text{Debt}_t}{\text{Assets}_t} * \frac{\text{Assets}_{t-1}}{\text{Debt}_{t-1}}$	2	0.9 %	Solvency
4	$\frac{\text{Assets}_t}{\text{Debt}_t}$	2	0.9 %	Solvency
5	$\frac{\text{Retained earnings}_t}{\text{Assets}_t}$	2	0.0 %	Solvency
6	$\frac{\text{Debt}_t}{\text{Equity}_t}$	1	0.1 %	Solvency
7	$\frac{\text{Long term debt}_t}{\text{Equity}_t}$	1	0.1 %	Solvency
8	$\frac{\text{Operating profit}_t}{\text{Interest expense}_t}$	1	50.7 %	Solvency
9	$\frac{\text{Assets}_t}{\text{Equity}_t}$	1	0.1 %	Solvency
10	$\frac{\text{EBITDA}_t}{\text{Debt}_t}$	1	0.9 %	Solvency
11	$\frac{\text{Equity}_t}{\text{Assets}_t}$	1	0.0 %	Solvency
12	$\frac{\text{Bank loans}_t}{\text{Assets}_t}$	1	0.0 %	Solvency
13	1 for $\text{Debt}_t > \text{Assets}_t$ 0 otherwise	1	0.0 %	Solvency
14	$\frac{\text{Equity}_t}{\text{Equity}_{t-1}}$	1	0.1 %	Solvency
15	$\frac{\text{Profit}_t + \text{Depreciation}_t}{\text{Debt}_t}$	1	0.9 %	Solvency
16	$\frac{\text{Operating CF}_t}{\text{Debt}_t}$	1	0.9 %	Solvency
17	$\frac{\text{Profit}_t}{\text{Assets}_t}$	4	0.0 %	Profitability

Variable	Formula	# of uses	Not available	Variable type
18	$\frac{\text{EBIT}_t}{\text{Assets}_t}$	3	0.0 %	Profitability
19	$\frac{\text{Operating profit}_t}{\text{Sales}_t}$	1	0.0 %	Profitability
20	$\frac{\text{Profit}_t}{\text{Profit}_{t-1}}$	1	2.7 %	Profitability
21	1 for $\text{Profit}_t > 0$ & $\text{Profit}_{t-1} > 0$ 0 otherwise0	1	0.0 %	Profitability
22	$\frac{\text{Profit}_t - \text{Profit}_{t-1}}{\text{abs}(\text{Profit}_t) + \text{abs}(\text{Profit}_{t-1})}$	1	1.4 %	Profitability
23	$\frac{\text{Profit}_t}{\text{Equity}_t}$	1	0.1 %	Profitability
24	$\frac{\text{Current assets}_t}{\text{Current debt}_t}$	5	6.8 %	Liquidity
25	$\frac{\text{Working capital}_t}{\text{Assets}_t}$	3	0.0 %	Liquidity
26	$\frac{\text{Financial assets}_t}{\text{Current debt}_t}$	3	6.8 %	Liquidity
27	$\frac{\text{Current debt}_t}{\text{Debt}_t}$	3	0.0 %	Liquidity
28	$\frac{\text{Current assets}_t}{\text{Assets}_t}$	2	0.0 %	Liquidity
29	$\frac{\text{Financial assets}_t}{\text{Current assets}_t}$	2	0.1 %	Liquidity
30	$\frac{\text{EBITDA}_t}{\text{Interest expense}_t}$	1	50.7 %	Liquidity
31	$\frac{\text{Inventory}_t}{\text{Current debt}_t}$	1	6.8 %	Liquidity
32	$\frac{\text{EBIT}_t}{\text{Interest expense}_t}$	1	50.7 %	Liquidity
33	$\frac{\text{Current debt}_t}{\text{Current assets}_t}$	1	0.1 %	Liquidity
34	$\frac{\text{Current assets}_t}{\text{Sales}_t}$	1	0.0 %	Liquidity
35	$\frac{\text{Sales}_t}{\text{Assets}_t}$	6	0.0 %	Activity
36	$\frac{\text{Current assets}_t}{\text{Current assets}_{t-1}}$	3	0.1 %	Activity
37	$\frac{\text{Inventory}_t * 365}{\text{Sales}_t}$	1	0.0 %	Activity
38	$\frac{\text{Financial assets}_t * 365}{\text{Sales}_t}$	1	0.0 %	Activity
39	$\frac{\text{Recievables}_t}{\text{Current assets}_t}$	1	0.1 %	Activity

Variable	Formula	# of uses	Not available	Variable type
40	$\frac{\text{Inventory}_t}{\text{Current assets}_t}$	1	0.1 %	Activity
41	$\frac{\text{Accounts payable}_t}{\text{Sales}_t}$	1	0.0 %	Activity
42	$\frac{\text{Accounts receivable}_t}{\text{Sales}_t}$	1	0.0 %	Activity
43	$\frac{\text{Inventory}_t}{\text{Sales}_t}$	1	0.0 %	Activity
44	Assets_t	1	0.0 %	Activity
45	$\frac{\text{Fixed assets}_t}{\text{Fixed assets}_{t-1}}$	1	21.2 %	Activity
46	$\frac{\text{Recievables}_t}{\text{Recievables}_{t-1}}$	1	9.5 %	Activity
47	$\log\left(\frac{\text{Assets}_t}{\text{GDP deflator}_t}\right)$	1	0.1 %	Activity

The following table captures model composition with respect to individual variables.

Explanatory variable	Model														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1							●		●	●	●			●	●
2									●				●		
3							●				●				
4			●			●									
5				●									●		
6								●							
7								●							
8								●							
9									●						
10									●						
11		●													
12										●					
13														●	
14		●													
15												●			
16														●	
17				●						●				●	●
18			●			●							●		
19								●							
20		●													
21														●	
22														●	
23								●							
24			●		●	●			●						●
25								●					●	●	
26				●	●					●					
27							●			●	●				
28							●				●				
29							●				●				
30				●											
31					●										
32			●												
33														●	
34	●														
35	●		●			●	●				●		●		
36		●					●				●				
37								●							
38									●						
39		●													
40		●													
41				●											
42				●											
43				●											
44	●														
45		●													
46		●													
47														●	