FIRM-SIZE HETEROGENEITY IN BANKRUPTCY PREDICTION: EVIDENCE FROM U.S. PUBLIC COMPANIES IN 1999–2018

by

Oleksandra Ovdiienko

A thesis submitted in partial fulfillment of the requirements for the degree of

MA in Business and Financial Economics

Kyiv School of Economics

2025

Thesis Supervisor: _____ Professor_Olesia Verchenko

Approved by _____
Head of the KSE Defense Committee, Professor [Type surname, name]

_____

_____

_____

Date _____

ACKNOWLEDGMENTS

I am large, I contain multitudes…

— Walt Whitman, "Song of Myself"

Firstly, I am deeply thankful to Professor Olesia Verchenko for being my academic supervisor during the master's studies, as well as for all the advice that I have got to improve my thesis and find the solutions in difficult situations. In addition, I am grateful for the courses that deepened my knowledge of the subject and helped me decide on a topic that will continue to interest me in the future.

Secondly, I am grateful to KSE for giving me the opportunity to study at the university with wonderful teachers, and to the grant committee for making my studies easier for me by providing financial support. I am especially appreciative to my family and friends, who have always been there for me, supported me in all situations, and believed in me. Finally, I am grateful to everyone whose contributions helped me during the writing process.

TABLE OF CONTENTS

LIST OF FIGURES

Number *Page*

LIST OF TABLES

LIST OF ABBREVIATIONS

NYSE The New York Stock Exchange

NASDAQ A stock exchange based in the United States

GAAP Generally Accepted Accounting Principles

SMEs Small and medium-sized enterprises

EBITDA Earnings before Interest, Taxes, Depreciation and Amortization

CTA Cash to total assets

WCTA Working capital to total assets

EBITDAIE Ln EBITDA to interest expense

ROA Return on assets

RETA Retained earnings to total assets

TLTA Total liabilities to total assets

RF Random Forest

MDA Mean Decrease Accuracy

MDG Mean Decrease Gini

SMOTE Synthetic Minority Over-Sampling Technique

CI Confidence Interval

ML Machine Learning

DT Decision Tree

GB Gradient Boosting

CHAPTER 1. INTRODUCTION

Bankruptcy prediction has been one of the most explored topics in corporate finance and accounting, as business failures remain a daily reality. Many individuals consider starting their own business, yet the risk of failure serves as a reminder of the uncertainty of future. Furthermore, it is not only business owners who may face negative consequences in the event of bankruptcy, but also other players such as investors, employees, creditors, and regulatory bodies. Predicting bankruptcy is not only essential for firm-level risk assessment but also is necessary for reaching stability on credit markets and for preventing systematic financial crises. As high-quality financial data becomes increasingly more available, the development of robust early warning models keeps drawing attention of many researchers, who try to find better ways to mitigate risks, optimize investment strategies, enhance credit valuation systems etc. Thus, a bankruptcy prediction model is not just a financial tool but a part of larger strategy for economic stability and resilience.

Various models have been already developed over the decades, but many of them have a need to be tested further having more recent and available data, increasing the period range or enhancing the models by emphasizing that companies of different sizes and ages do not respond to risks in the same way.

This thesis aims to answer the question of whether financial indicators can consistently predict corporate bankruptcy across firms of different sizes. Specifically, it investigates whether the determinants of failure vary by small, medium, and large publicly listed companies in the United States. The analysis focuses on the firms that were listed on the exchanges such as NYSE and NASDAQ in the period between 1999 and 2018. While the vast majority of prior research focuses on developing different models on a fixed dataset for bankruptcy prediction, this study highlights the heterogeneity of financial distress

mechanisms and explores how firm size affects both the predictive power and significance of financial factors.

This study's motivation also flows from the availability of an extensive dataset that covers 20 years of financial data from public U.S. firms listed on stock exchanges. The definition of bankruptcy in the U.S. is consistent, and well-documented due to the Bankruptcy Code of Chapter 7 (liquidation) and Chapter 11 (reorganization). This classification gives a clear identification of the default events that is not always available in other countries. Moreover, publicly traded companies in the U.S. are required to report the financial statements in accordance with GAAP, which ensures that they follow the same accounting rules and makes financial reporting comparable. Additionally, the study of company failure in the United States is especially important as the default of large American company can have a wide resonance throughout the world economy. Using a large and well-structured dataset of public firms, this research estimates the model's robustness over the long term, covering both stable years and shocks.

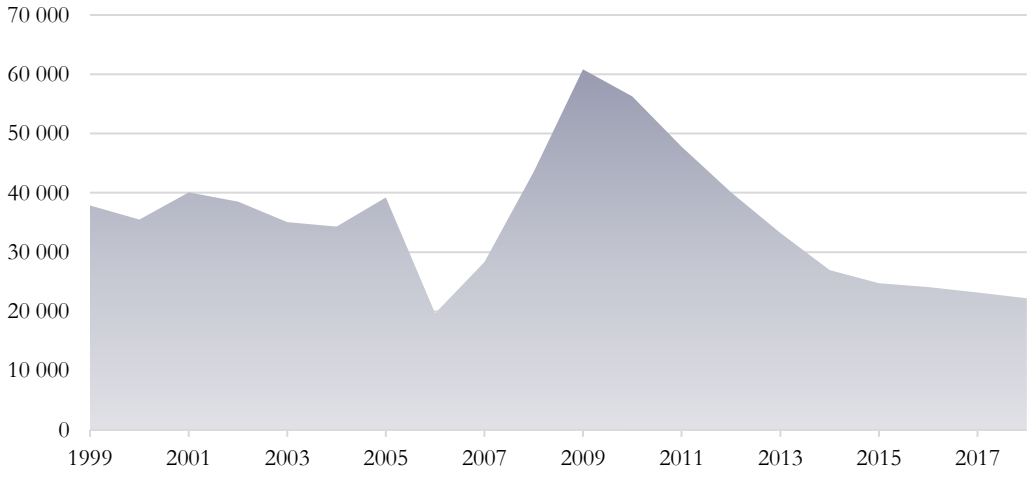
The methodology and findings of this research may be adaptable to other markets having the access to structured financial information at the firm level. In particular, the understanding of how the determinants of bankruptcy differs across firms of various sizes can help stakeholders and policymakers design more accurate and differentiated early-warning systems. For instance, in countries such as Ukraine, where the access to a structured firm-level financial reporting is limited at the moment as many enterprises provide inconsistently reported or not publicly available information, such models could be applied once reliable accounting data become available to identify whether small, medium, and large enterprises exhibit distinct risk patterns. Thus, using the traditional financial indicators when accounting for firm size heterogeneity for predicting bankruptcy, this study contributes to the academic field by revealing that bankruptcy is not driven by uniform financial patterns across all firms, while improving existing models and testing them over long time periods.

CHAPTER 2. U.S. BANKRUPTCY LANDSCAPE AND RELATED STUDIES

The prediction of corporate bankruptcy has been the essential topic of interest in finance for a long time. Creditors, investors, policymakers, employees, and firm managers are the most interested parties in company's well-being. In developed markets like the United States the companies that are listed on the two most famous exchanges called NYSE and NASDAQ are obliged to disclose their financial data under GAAP requirements that gives the opportunity for many researchers to get necessary information for their articles, to develop models and report results to the world.

Between 1999-2018, the number of bankruptcies in the United States fluctuated significantly as shown in Figure 1. During the period under review, the highest number of bankruptcies occurred in 2009-2010, which is probably related to the financial crisis, after which their number began to decline.

Figure 1. The number of bankruptcies in the U.S., the number of companies, 1999-2018



Source: Administrative Office of the U.S. Courts, compiled and published by the American Bankruptcy Institute (ABI), Annual Business and Non-business Filings by Year (1980–2021))

By Administrative Office of the U.S. Courts an all-time highest amount of bankrupt companies from 1980 to 2025 reached the number of 82,446 in year 1987, while the lowest was in 2022 – 12,748 companies.

The U.S. Bankruptcy Code brings transparency to the restructuring and liquidation of the firms. This clarity is crucial for this study, as it ensures that the bankruptcy status of companies is accurately identified and consistently reported across the sample period. Such data reliability allows the models to capture genuine financial distress patterns rather than errors arising from inconsistent reporting. This situation is much less common in the developing markets, where the process of reporting about the bankruptcy is inconsistent and the vast majority of financial data is being difficult to find or unavailable.

The two economists who laid the foundation of quantitative prediction of bankruptcy were Beaver (1966) with his analysis to discover which financial ratios predict the bankruptcy best and Altman (1968), who used multiple discriminant analysis for 66 U.S. manufacturing companies from 1945 to 1965. The latter used such ratios as working capital to total assets, retained earnings to total assets, EBIT to total assets, market value equity to book value equity of total debt and sales to total assets. He significantly advanced the field with his Z-score model and the ratios that were used still serve as a baseline for many studies including mine.

Later, Ohlson (1980) chose logit analysis to avoid the problems connected with MDA (Multivariate discriminant analysis) that was the most popular method with the usage of vectors as predictors. In particular, MDA relies on strong statistical assumptions, such as multivariate normality of the predictors and equal covariance matrices across bankrupt and non-bankrupt firms, assumptions that are rarely satisfied in financial data (Wang, 2024). Ohlson used sample of U.S. firms and incorporated leverage, firm size, and liquidity ratios as predictors.

In the early 2000s, many economists moved their attention to the SMEs and startups that survive through constant financial challenges. The European studies that focused on startup and SMEs survival are:

- Cultrera and Bredart (2016) with the examination of 7,152 Belgian SMEs and identification of current ratio and total equity to total assets ratio as significant predictors. Their findings emphasize the importance of liquidity and capital structure – the two aspects particularly sensitive in smaller firms that often operate with limited reserves;
- EI Kalak and Hudson (2015) used a dataset on U.S. SMEs of year 1980-2013 and applied survival analysis with the result of working capital and retained earnings being the most important risk indicators. This study is important as it connects traditional accounting metrics with indicators that reflect the long-term stability of the company, rather than the short-term setbacks, which closely matches my own use of financial ratios.

These studies are important as they highlight the need to account for firm size and capital flexibility in predictive models, factors that are central to the methodology in this paper. Their approach is going to be expanded by segmenting firms into SMEs and large enterprises in this thesis.

Moreover, the recent studies showed that there is a common usage of hybrid and advanced statistical methods, in particular combinations of traditional classification models with survival techniques and machine-learning algorithms. For example, Fuertes-Callen et al. (2022) applied logistic and survival models to Spanish startups data, while Gallucci et al. (2023) used a merged longitudinal predictive model for Italian data. Both studies included time-dependent financial ratios such as EBITDA to debt and cash flow to turnover as explanatory variables.

Machine learning methods (ML) are another approach adopted by numerous researchers beyond the traditional models. For instance, Severin and Veganzones (2021) used several

different approaches including logistic regression, neural networks, support vector machines (SVMs), and extreme learning machines (a learning algorithm for single-hidden layer feedforward neural networks) on 6,000 different SMEs in France (Yu et al., 2014) . Including time-dependent financial ratios as EBITDA to debt and cash flow to turnover, their findings highlighted the trade-off between forecasting accuracy and model interpretability.

Gallucci (2023) shows that Random Forest outperforms traditional statistical models such as logit or discriminant analysis in bankruptcy prediction. Therefore, following the recent literature I will use both the logit model and a Random Forest approach in my thesis to test whether modern, data-driven models can improve prediction while maintaining some interpretability. Random Forest was chosen among other machine-learning techniques because it balances predictive accuracy and robustness against overfitting, while still allowing for an assessment of variable importance – that is a valuable feature for identifying key financial determinants of bankruptcy. Additionally, among all analyzed codes from different authors who worked with the dataset I have used in this thesis, this model was one of the most accurate and outperformed other models in terms of the overall predictive accuracy, along with the XGBoost model (Singh, 2023).

A synthesis of the reviewed literature indicates that the most frequently used predictors include: CTA (cash to total assets), WCTA (working capital to total assets), EBITDAIE (ln EBITDA to interest expense), ROA using EBITDA and net income, RETA (retained earnings to total assets), TLTA (total liabilities to total assets), size (ln total assets), age (ln current year – year of foundation).

Table 1. Literature review of researches on firms' failure

Author	**Data**	**Country**	**Model or method**
Altman (1968)	1945-1965; 66 companies	US	Multiple discriminant analysis

EI Kalak and Hudson (2015)	1980-2013; 11,117 SMEs	US	Survival analysis
Gupta et al. (2015)	2000-2009	UK	Survival analysis
Cultrera and Bredant (2016)	2002-2012; 7,152 SMEs	Belgium	Logit
Castillo et al. (2018)	2012-2014; (>10,000)	Colombia	Logit
Succurro et al. (2019)	2006-2010; 31,958	Italy	Principal component analysis and Logit
Severin and Veganzones (2021)	2016-2019; 6,000 SMEs	France	Logit, linear discriminant analysis, neural network, ML, support vector machine
Zhang and Xie (2023)	2008-2018; 4,622 firms (hotel industry)	Norway	Survival analysis
Gallucci et al. (2023)	2012-2014; 973 SMEs	Italy	Merged longitudinal predictive model

Source: Wang, 2024

Table 1 summarizes data periods, data geography and empirical approaches of the reviewed studies. In addition, the table shows the progress in the complexity of the models used to predict bankruptcy over time.

Despite the wide range of studies on company bankruptcy across countries including Norway, the U.S., Italy, France, Colombia, Belgium, and the UK, there are several gaps remain unaddressed. Many existing papers focus mostly on relatively short timeframes, use narrow datasets that limit the explanatory power of the models, or rely mostly on a single

predictive method without accounting for firm heterogeneity. Just a few studies combine both traditional and machine learning methods on a dataset that covers a long period of time. Therefore, there is still significant scope for further research.

To address these issues, this thesis compares the performance of Logistic Regression and Random Forest models, while introducing the segmentation of firms by size to discover how predictive performance varies across SMEs and large enterprises. Although the set of data has already been used in 17 Kaggle-based studies, most of them were limited to testing various machine learning models such as Random, Forest, XGBoost, SVM, or Gradient Boosting on the same pooled sample of firms with the aim of maximizing prediction accuracy. None of the reviewed implementations accounted for firm heterogeneity or investigated how bankruptcy determinants may differ across companies of various size. Moreover, many of those codes relied on unbalanced data, redundant variable (for instance, the identical X9 and X16 indicators), and lacked cross-validation or interpretability analysis. In contrast, this study suggests a methodological improvement by dividing companies into small, medium, and large categories and investigating whether financial ratios affect bankruptcy risk differently across these groups. It does also combine two modeling methods of the Logistic regression and Random Forest to balance interpretability and predictive performance.

Thus, the goal of this study is not only to compare classical and modern predictive models, but also to evaluate the stability of financial determinants across firms of different sizes applied over a long period of time in the United States for the potential future application of such models in the emerging markets such as Ukraine.

CHAPTER 3. METHODOLOGY

3.1. Modelling framework and theoretical background

Predicting corporate bankruptcy is a central problem in financial risk modeling, as it combines both theoretical and practical aspects of firm performance assessment. The empirical literature suggests that a list of different financial ratios are being used for the estimation of company bankruptcy. For instance, Altman (1968) used multiple discriminant analysis with such financial ratios as working capital to total assets, retained earnings to total assets, EBIT to total assets, market value equity to book value of total debt, sales to total assets. Many subsequent studies have applied the logit model or survival analysis as the method of prediction. Accordingly, the choice of the logit model as the core one in this thesis is grounded on the literature, its ease of implementation, the high level of interpretability of coefficients, its adequate forecast accuracy, and low computational cost. The logit model allows to directly interpret the marginal effects, which is key to understanding the strength of the impact and the direction of financial indicators on the probability of bankruptcy. Furthermore, the logistic regression model serves as a baseline for comparison with more flexible data-driven models, such as Random Forest, allowing for an estimation of the trade-off between interpretability and predictive accuracy.

The function of the logistic regression model is defined as follows:

$$P(Y_{it} = 1|X_{it-1}) = \frac{1}{1+ e^{-\beta X_{it-1}}} \quad (1)$$

where,

$P(Y_{it} = 1|X_{it-1})$ – is the probability of company i failure at time t;

X_{it-1} – is explanatory financial ratios in a vector form that are observed at the end of the previous period for company i;

β – is the vector of estimated coefficients.

While the logistic regression reflects linear connections between explanatory variables and the odds ratio of failure, it may not fully account for the complex non-linear patterns in financial data. Therefore, the use of other methods of bankruptcy investigation has become widespread.

According to Gurnani et al. (2021) more sophisticated models are used in this topic as well, for example, Random Forest model often achieves higher predictive accuracy, offers average level of interpretation of coefficients, captures non-linear effects, stable to retraining, not very sensitive to noise which makes it particularly suitable for financial datasets characterized by heterogeneity and noise. In addition to the logistic regression model, this study aims to incorporate Random Forest as a second predictive method to evaluate and compare results, assessing non-linear relationships and verifying the stability of the obtained patterns over a long period of time.

Unlike a single decision tree, that tend to overfit the training sample, Random Forest, that is one of the most used methods in machine learning, reduces variance by the method of averaging over a large ensemble of weakly correlated trees. This allows to stabilize the predictions and makes the model resistant to the noise and exclusions (outliers), as well as this model is less sensitive to the multicollinearity, or noise in financial indicators, which is common in real company accounts.

The mathematical description of Random Forest model is presented here:

$$\hat{y} = argmax_y \sum_{t=1}^{T} 1\{f_x(x) = y\}, \quad (2)$$

where,

T – is the number of trees in the model;

$f_x(x)$ – the prediction, received from the t-th tree;

y covers all possible class labels;

$1\{\cdot\}$ – is the indicator function, that takes the value of 1 if the tree classifies the observation as a y class.

The logic of this function is to combine a set of decision trees to increase the overall accuracy and lower the error. Random Forest teaches a lot of weakly correlated trees of solutions on random subsets of data and features. During the prediction the model averages the probabilities returned by individual trees and then classifies the observations in accordance with the selected threshold (Jarvis, 2024). In addition, the choice of this model is justified by the results of research conducted by other scientists, including those who used the same dataset in their analyses to compare which model has the highest accuracy and predictive power in forecasting bankruptcies. A significant number of investigators have concluded that the Random Forest and XGBoost models are among the best in terms of their accuracy metrics for the dataset used in my thesis and moreover they are not extremely heavy machine learning models.

Hypotheses formulated in this thesis:

1. Hypothesis 1. Random Forest will outperform logistic regression in discrimination (AUC/PR-AUC) due to nonlinearity and interactions, while logistic regression remains more interpretable at the coefficient level.
2. Hypothesis 2. After applying SMOTE to the training set (and testing on a clean set), model performance increasing relative to the imbalanced baseline.
3. Hypothesis 3. Predictive performance and key predictors will differ depending on company size; some signals will be stronger in small companies, others in large firms, so size-specific models will outperform the pooled model in terms of interpretability and discrimination quality.

Hence, this research presents a dual-model framework that combines an interpretable, theory-driven statistical model with a non-parametric algorithm, that enables a

comprehensive understanding of bankruptcy determinants. The following sections describe in more detail all the steps that were taken to confirm or disprove the aforementioned hypotheses.

3.2. Replication of existing studies

Methods-to-purpose map:

- Logistic regression provides signs and magnitudes of effects and a transparent baseline;
- Random Forest captures non-linear patterns and interaction that logistic regression is not capable of;
- SMOTE balances rare bankruptcy class in the training data of the dataset;
- Correct labeling (only one failure in the company history) prevents the leakage of information;
- Size distribution checks whether the determining factors and accuracy vary depending on the size of the company;
- AUC, PR-AUC, Brier and VIF assess discrimination, calibration, and multicollinearity;
- RF importance (MDA/MDG) assesses factor stability.

As the initial step of my research, I decided to replicate the code of one of the researchers who worked on the same dataset in order to establish a reliable reference point for further model enhancement. Out of the 17 studies by other authors that I analyzed, I selected the analysis by Utkarsh Singh, who was the author of the dataset as well, because he conducted a brief exploratory data analysis providing summary statistics at the beginning, following with target column distribution, moreover presented a graph of numerical features with outliers, the correlation matrix of features, and showed a comparison between accuracy, ROC-AUC, and F1 score of Random Forest, SVM, Decision Tree, XGBoost and Gradient Boosting models.

The data that contains 18 standardized financial variables was downloaded from Kaggle. In the original dataset, companies that ultimately went bankrupt were labeled as "failed" for all of their historical observations, that is for each year available prior to bankruptcy. To address this issue and bring the dataset in line with the standard event study logic used in the bankruptcy prediction literature (Ohlson, 1980; Shumway, 2001; Giordani et al., 2011), the labeling was changed so that each company is coded as bankrupt only once – in the last fiscal year preceding the bankruptcy filing. All previous observations of the company over the years remain marked as "alive". Although none of the researchers who worked on the same dataset made this modification, this can be considered another improvement in working with these data. To enable further processing, all text labels were converted to numerical values: "alive" = 1, "failed" = 0.

A summary table was generated with the main statistical characteristics of each indicator: data type, proportion of missing values, the number of unique observations, minimum and maximum values, and the first three values of the indicator. This information was visualized in a table format.

Next, the following steps were performed:

- Calculation of the proportions of "alive" and "failed" companies and construction of a pie chart showing a significant imbalance between classes.
- Identification of potential outliers in 18 financial variables using the interquartile range (IQR), and presenting it in a box plot.
- Evaluation of the correlation matrix between all financial ratios and construction of the heatmap.

These steps confirmed the presence of classic problems with the financial data, that are heterogeneity, outliers, and collinearity between indicators. After the preliminary analysis and pre-processing that includes the company data label change into the binary one of 1 and 0, and the selection of features/target columns, a logistic regression was constructed

in the form of L2 regularization (where alpha=0) to prevent model overfitting with correlated variables and to make it as close as possible to the original Python code of the replication, as penalty = "l2" is the default setting in Python's sklearn.linear_model.LogisticRegression(), and my analysis was completely performed in the R statistical tool. The cv.glmnet() function with five-fold cross-validation was used to automatically select the optimal penalty parameter (lambda). The model was trained on the matrix x_mat (all 18 financial variables) and the label vector y_vec. As a result, coefficient estimates were obtained, and their absolute values were used to determine the most important 10 predictors.

To compare the results, a basic Random Forest model was recreated with the same variables, setting such parameters as 1,000 trees, square root of the number of variables in each split (mtry = sqrt(p)), and the seed of 42. To save the full identity to the logit model Random Forest was trained on the full dataset without class balancing. After the training, 10 variables with the biggest impact were defined using the importance() function with the criterion of Mean Decrease Accuracy (MDA).

Accordingly, two separate subsets of variables selected for logistic regressing and Random Forest models were created and saved in a csv file. For each model, an 80/20 stratified division of data was performed into training and test samples. The models were evaluated using the metrics of accuracy, AUC (area under the ROC curve), precision-recall curve, and confusion matrix, which shows the ratio of true and false classifications. As the author of the replication code did not finish his analysis with the logistic regression, I have applied the same system of training to this model and evaluated it similarly like the RF was, to be able to reasonably compare them with each other.

After the final step of replication, having the same dataset and applying the similar methodology the chosen by R factors appeared to be partially different from the variables in the original code written by Utkarsh Singh in Python, and as a result the final models' characteristics and evaluations have a difference as well. It can be explained by data

modification and the difference in internal optimization algorithms between the glmnet library in R and the sklearn library in Python. Other changes applied to the model structures and additional modifications, apart from the use of L2 regularization for logistic regression, did not bring the similarity of the model results closer to the expected ones, so this version of the code remained final.

So, this stage of the thesis provided a baseline for comparison for further model improvements – the introduction of class balancing (SMOTE), duplicate filtering, robustness tests, and company segmentation by asset size.

3.3. Modified replication and methodological extensions

The replication of the code mentioned previously presented the base structure of the models of logistic regression and Random Forest, but the results showed a number of methodological problems. Mainly the existence of the imbalanced classes, where the share of bankruptcies is just 0.8% that lowers a lot the ability of models to learn on the rare observations. Moreover, there were duplicates of the variables (X9=X16), that many other researchers that used the same dataset did not fix. In accordance with this conclusion the other part of my code shows the modified replication with methodological improvements:

- Duplicate feature removal;
- SMOTE on the training set only, to learn the minority class while preserving an uncontaminated test set.

The first change this part of code underwent compared to the previous replication was a check for duplicate variables, which revealed that "net sales" and "total revenue" features are the same, so the "total revenue" column was removed from the set of data.

Another modification was that in order to eliminate class imbalance, the Synthetic Minority Over-Sampling Technique (SMOTE) from the smotefamily library was used, similarly to the original Kaggle example (Espinoza, 2024). This algorithm generates synthetic observations of minority class (bankrupt companies) by interpolating between k-nearest

neighbors (in the study, k=5). After the SMOTE implementation the number of bankruptcies in the training sample reached the share of nearly 50%, that was illustrated by a pie chart diagram. Such balancing allows the model to learn on a rare class and avoid the majority domination more adequately.

After such changes the logit model was rebuild with the same methodology, with the model being trained on a balanced train sample and the testing set on a "clean" test sample without the synthetic observations. Random Forest model was built on the same balanced train data with 500 trees and random subset of predictors at each split.

Then, for both models, a set of robustness tests was conducted. They are the ROC_AUC with 95% confidence interval (low/high), Brier score, PR_AUC, the optimization of F1 threshold. In addition, both models were subjected to visualization of the analysis results with the confusion matrix, ROC curve, and Precision-Recall curve, where RF consistently demonstrates higher resolution and fewer misclassifications compared to the logistic model.

Therefore, as expected after the implementation of the SMOTE balancing, the predictive quality of the models improved, which is also confirmed by studies of various bankruptcy prediction models, concluding that such data class balancing significantly improves classification accuracy with regard to sensitivity and balanced accuracy (Jarvis, 2024).

3.4. Size-oriented approach to modeling

The previous steps of this research were oriented on the development of the basic logistic regression and Random Forest models, their replication, and enhancing in a way of SMOTE balancing and robustness check. Although even after these improvements an important hypothesis is being left – all companies are considered to be homogeneous.

Since bankruptcy risks and financial structures vary much between the different company sizes, this study tests separate logistic and Random Forest models within each segment. This approach addresses a limitation in many prior studies that developed models without

segmenting the companies by size or taking as data just SMEs. By introducing size specific modeling, the analysis aims to deliver more precise conclusions without masking size-related effects to bring actionable insights tailored to different categories of firms. Segmentation by size of the companies adds an original level of analysis, removing the limitations of many previous studies that considered companies as homogeneous. Moreover, none of the existing studies that used the same public Kaggle dataset have applied similar size-oriented approach to overcome the assumption of homogeneity.

Therefore, in this part of the thesis companies were segmented by the total assets (X10) variable, which is used as a proxy variable for firm size. The three groups of companies (mainly, small, medium, and large ones) were formed in the training sample on the base of quantile binning, where the first quantile represents the small company group and so on (Vater, 2023). To prevent data leakage this grouping was performed only on training data, after which the same thresholds were applied to the test sample. This ensures the independence of testing and allows to assess whether the quality of the forecast and the significance of financial indicators differ between segments of different sizes. Separate logistic regression and Random Forest models were then constructed for each of the three groups.

Before evaluating the logistic models for each group, the following steps were performed: removal of duplicates (X16), class balancing using SMOTE in the training sample (k=5). The coefficients of each model were also estimated, and the models were evaluated using ROC-AUC with 95% CI, Brier score, Accuracy, PR-AUC. As well as ROC, Precision-Recall curves were constructed for each segment to visualize the ability of the models to classify bankruptcies in different groups of companies.

To improve predictive quality and verify the stability of results, Random Forest models were also implemented within each group (small, medium, large). The algorithm was realized in a similar way to the logistic regression, where the training data was balanced using SMOTE. A separate RF model was built with 500 trees and importance calculation

activated (importance = TRUE). Then models were tested on a "clean" test sample without the SMOTE observations. Later for each company a set of metrics was calculated (AUC with 95% CI, Brier score, PR-AUC, Accuracy). In addition, ROC, Precision-Recall curves were constructed for each RF model.

Hence, this approach allows to assess whether the patterns of bankruptcy differ between small, medium, and large enterprises. While by simultaneously applying two types of models, it is possible to determine which size groups of bankruptcies are better described by a linear model and which by a more complex non-linear structure.

3.5. Robustness checks and determination of the factors' importance in models

After the establishment of the size-oriented logistic regression and Random Forest model a series of tests were conducted to evaluate the reliability, stability, and interpretability of the results.

This stage of the methodology covers two key areas:

- Testing the robustness of models (among the estimates not mentioned earlier are assessing multicollinearity and the stability of coefficient estimates). Robustness covers discrimination (AUC with 95% CI, PR-AUC), calibration (Brier), and specification (VIF and aliasing checks in logistic regression models), for RF I assess importance stability via MDA and MDG correlation;
- Determination of the variable importance in the models.

In the logistic regression models the main focus was on the problem of multicollinearity between financial characteristics, that can distort the estimates of β coefficients and reduce the interpretability of the results. To check for multicollinearity, the Variance Inflation Factor (VIF) was used, which is defined by Investopedia as:

$$VIF_i = \frac{1}{1 - R_i^2} \tag{3}$$

where,

R_i^2 – is the coefficient of determination obtained by regressing the i-th predictor on all other independent variables.

The indicator was interpreted as follows: if VIF < 5, this is an acceptable level of multicollinearity; if the value is between 5 and 10, this is a moderate collinearity; and finally, if VIF ≥ 10, this is a high level of collinearity. The table for each predictor of the three models was constructed. Before the estimation the variables with zero or near-zero variance and coefficients that appeared to be aliased (NA) because of the collinearity in the subset, were removed.

For an assessment of the reliability of Random Forest results, in addition to indicators mentioned in the section on building size-oriented models, the stability of predictor importance was checked in each of the three groups of companies using two built-in indicators, according to Benard et al. (2021):

- Mean Decrease Accuracy (MDA) – reflects how much the accuracy of the model decreases when a specific variable is randomly shuffled;
- Mean Decrease Gini (MDG) – characterizes the average decrease in the purity index (Gini impurity) when splitting by a given predictor in trees.

For each group of the companies, the correlation between MDA and MDG was calculated. After completing the robustness check, the final part of the thesis concerns the assessment of which factors were the most important in each size-specific model for determining bankruptcy.

For each of the logit models there were built β coefficients graphs, allowing for a visual estimation of the direction of influence (positive/negative effect on the probability of bankruptcy), and the significance of predictors according to the criterion of $p < 0.05$. For each size group of the Random Forest models, the 10 most important financial variables

were determined based on MDA and MDG indicators, which were also presented in the form of graphs.

Therefore, the methodology of this study combines two complementary lines: logistic regression as an interpreted statistical benchmark and Random Forest model as a non-linear algorithm. The study design consists of: reproduction of basic approach; methodological improvements (removing duplicates, balancing SMOTE classes); size-oriented modeling with quantile binning by assets and independent training/testing in each segment; robustness testing of models and identification of key factors influencing the probability of a company's bankruptcy.

CHAPTER 4. DATA

4.1. Description of the data

The dataset used in this study is named as "US Company bankruptcy prediction dataset" and is publicly available on Kaggle. The collection of data was compiled by Utkarsh Singh (2023) from NYSE and NASDAQ stock markets and lastly changed in 2023. The dataset includes:

- 8,971 companies listed on NYSE and NASDAQ
- 78,682 firm-year observations from 1999 to 2018
- 18 standardized firm characteristics labeled as X1-X18
- binary bankruptcy labels as "failed" or "alive".

The dataset contains no missing values or synthetic data and distinguishes bankruptcy events as the year after the last one listed in set. The 18 variables are the financial characteristics, namely:

- X1 – current assets (the sum of all assets that the firm is expected to sell or use).
- X2 – cost of goods sold (direct costs attributable due to the sale of goods).
- X3 – depreciation and amortization (the reduction in value of tangible and intangible assets respectively because of the wear, usage or time).
- X4 – EBITDA (that are earnings before interest, taxes, depreciation, and amortization).
- X5 – inventory (raw materials and goods that are held by a company for production or resale).
- X6 – net income ("clean" profit that is gained by company after the reduction of all expenses, taxed, costs).
- X7 – total accounts receivable (outstanding payments of customers for goods and services delivered).

- X8 – market value (market capitalization to the total value of firm's publicly traded shares).
- X9 – net sales (gross sales after adjusting for returns, allowances, and discounts).
- X10 – total assets (all economic resources that the firm owns).
- X11 – total long-term debt (financial obligations due after one year).
- X12 – EBIT (earnings before interest and taxes).
- X13 – gross profit (revenue after the deduction of the cost of sold goods, excluding other expenses).
- X14 – total current liabilities (short-term financial obligations payable within one year).
- X15 – retained earnings (the profit the company has after paying direct and indirect costs, income taxes and dividends to shareholders).
- X16 – total revenue (all income the company generated before expenses).
- X17 – total liabilities (all debts owed by the firm to external entities).
- X18 – total operating expenses (day-to-day expenses required to run a business).

Each firm in the dataset is anonymized and identified by a code (e.g., C_1, C_2) and each entry corresponds to a specific fiscal year. Table 3 presents more detailed information about the data type, % of missing data, and number of unique values for each variable in the dataset, as well as their minimum and maximum values.

4.2. Analysis of previous studies and change in labeling scheme

Before building the models, I have completed a full analysis of all available studies that used the same open dataset. As of 2025, there are seventeen studies that cover various methodological approaches: from classical models (Logit, Altman Z-score) to machine learning (Random Forest, XGBoost, SVM, Decision Tree, Gradient Boosting, Deep Learning).

After verification, I found that all the works did not change the classification of the status in the dataset, that is the company was marked with a status of "failed" in all years prior to the actual bankruptcy. This approach creates "target leakage" and the model receives a hint about the future state of the company. To prevent this, the current study reformatted the status label by manually changing the status label of the companies as "failed" just for the one year that is prior to the real bankruptcy, while all the previous years retain the status "alive". This eliminates target leakage, so the model learns pre-failure patterns rather than memorizing post-outcome labels. So, this approach brings the task closer to the realistic bankruptcy risk forecasting, where decisions are made based on financial data from the years prior to the actual collapse of the company.

The most common performance metrics for bankruptcy prediction used by authors who utilized this dataset were accuracy and ROC-AUC, although some studies also considered balanced accuracy and F1-score to account for class imbalance. The typical structure of their studies included a basic exploratory analysis (EDA), random division into training and test samples (most often 70/30 or 80/20), and testing of several algorithms for the comparison. A few studies used time-split with a training on earlier years and testing on later ones. The main problem with the studies was the incorrect interpretation of the variable "status_label" mentioned before. In some cases, authors used the year indicator as the sole predictor, while companies were labeled as "failed" in all years prior to filing for bankruptcy, resulting in the model yielding a spurious accuracy near 100% (Table 2).

Table 2. Review of studies using the "US Company bankruptcy prediction dataset"

The name	**Models**	**Method & Conclusion**
A comparison of different methods for bankruptcy prediction	Lasso LR, RF, XGBoost, Z-score	ML approaches outperform traditional methods; SMOTE - an effective way to improve performance
US Bankruptcy Visualize Importance	LightGBM	High overall accuracy but very weak at predicting bankruptcies

US Company Bankruptcy Prediction: 93% Accuracy Bankruptcy prediction via XGBoost: AUC ROC = 0.89	RF, SVM, XGBoost, DT, GB XGBoost	RF and XGBoost – best accuracies Good predictive performance
US Company Bankruptcy Prediction | 100% accuracy	XGBoost	Not acceptable as a valid predictive because of the target leakage
US_company_bankrupt	LR; DT; RF; KNN; Gaussian Naive Bayes	Models mostly predict the majority "alive" class and struggle to identify bankruptcies
Decision Trees 97% accuracy	DT	Heavy oversampling
Bank Thief	Random Forest	70/30 split; 93% accuracy
american_bankruptcy	Random Forest	Over/under sampling; CV ≈ 99%, overfitting likely
US Company Bankruptcy Classification	LR; CART (DT); XGBoost	XGB most stable
EDA for Companies Financial data	Linear, Lasso, DT	Not bankruptcy model
Time Series Classification for Business	MLSTM-FCN	Without class weights or resampling, the model fails to capture rare bankruptcy cases
US Company Bankruptcy Prediction | RF + GB	RF; GB	Gradient Boosting outperforms Random Forest
US Company Bankruptcy Prediction | XGB +SVM	XGBoost; SVM	Higher accuracy in SVM, but recall = 0
US Company Bankruptcy Prediction 87% high recall	KNN, Naive Bayes, SGD	KNN performed best
US Companies Bankruptcy Prediction: Altman vs. ML	Altman Z-Score, RF, DT, GB, XGBoost, SVM	XGBoost & Random Forest performed the best
notebookb76ad15a49	-	-

Source: Kaggle

According to the generalized results, the XGBoost and Random Forest models consistently showed better results, especially after class balancing. As well many authors report high overall accuracy but low accuracy in detecting the rare class. In this regard, this thesis also used a RF model, but with a modification in the dataset to ensure correct labeling, and SMOTE application. In addition, the paper introduces a classification of companies by asset size that was not presented for this dataset previously.

4.3. Exploratory data analysis

A summary overview function confirmed the absence of gaps in the data. The Table 3 shows a strong imbalance between "alive" and "failed" classes, where only 0.8% are reported to be bankrupt.

Table 3. Summary table of the dataset, including the proportions of active and bankrupt companies

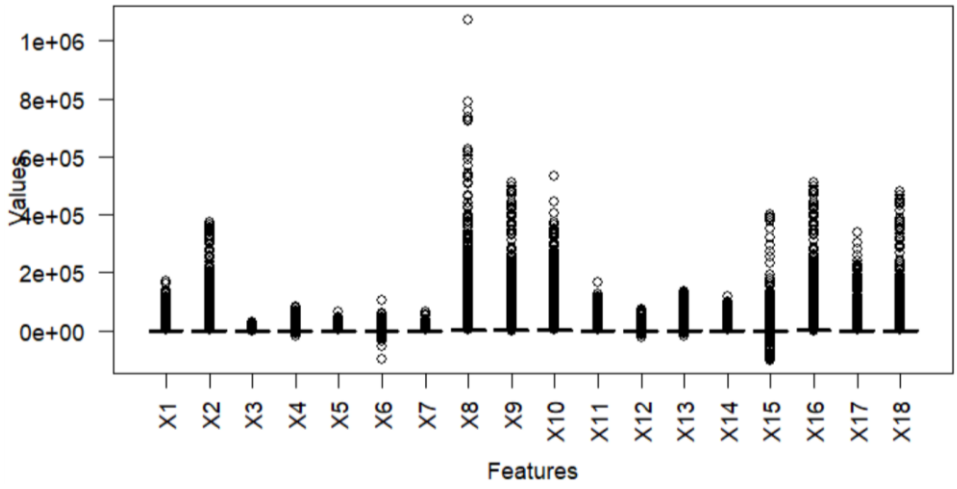
Alive share: 99.2%			Failed share: 0.8%		
	Data type	% missing	# unique	min	max
Company name	Character	0	8971	NA	NA
Status label	Character	0	2	NA	NA
Year	Integer	0	20	1999	2018
X1	Numeric	0	65895	-7.76	169662
X2	Numeric	0	65690	-366.65	374623
X3	Numeric	0	36010	0.00	28430
X4	Numeric	0	59060	-21913.00	81730
X5	Numeric	0	38898	0.00	62567
X6	Numeric	0	55550	-98696.00	104821

X7	Numeric	0	49577	-0.01	65812
X8	Numeric	0	77580	0.00	1073391
X9	Numeric	0	68596	-1965	511729
X10	Numeric	0	71521	0.00	531864
X11	Numeric	0	39741	-0.02	166250
X12	Numeric	0	56949	-25913.00	71230
X13	Numeric	0	64952	-21536.00	137106
X14	Numeric	0	58685	0.00	116866
X15	Numeric	0	72062	-102362.00	402089
X16	Numeric	0	68596	-1965	511729
X17	Numeric	0	64640	0.00	337980
X18	Numeric	0	70840	-317.20	481580

To identify outliers, a box plot of numerical features was constructed for all financial indicators. The visualization showed a pattern, where numerous high outliers present, particularly in variables of market value, net sales, total assets (X8-X10), retained earnings (X15), which indicates significant gaps between company sizes (Figure 2).

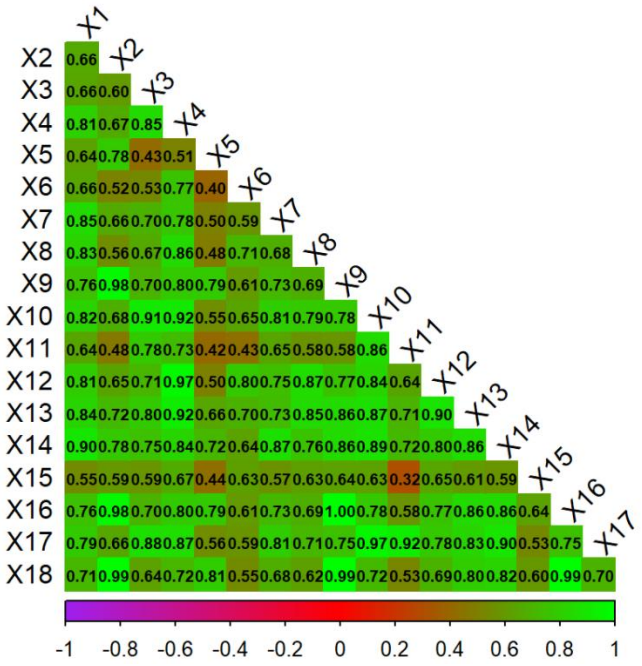
To assess the relationships between numerical variables, a correlation matrix was constructed (Figure 3). As can be seen from the graph, most variables are characterized by a high level of positive correlation (> 0.7), which may be typical for financial data, where indicators are formed on the basis of similar accounting values such as assets, income, equity.

Figure 2. The Box plot of numerical features for outliers detection



The highest coefficients are observed between such variables as total operating expenses and cost of goods sold/net sales/total revenue at 0.99, that is because these variables are economically interdependent.

Figure 3. Correlation matrix of numerical features in the dataset



While the correlation between net sales and total revenue reached 1.00, which led to data verification and revealed that these data columns are identical, so total revenue variable is removed from the data. Therefore, the initial data analysis shower that the dataset is complete and highly unbalanced, financial variables have large number of outliers and a strong positive correlation driven by accounting identities and company size. These findings prompt the elimination of class imbalance with SMOTE and the construction of size-oriented models.

CHAPTER 5. RESULTS

5.1. Results of the replication models and the application of SMOTE

Two baseline models, logistic regression and Random Forest, were tested on the full adjusted dataset, where status "failed" was assigned only for the year immediately preceding a company's bankruptcy.

For each model, the ten most significant predictors were automatically selected based on their contribution to the probability of bankruptcy (using the L2 regularization method for LR and feature importance based on Gini Decrease for RF):

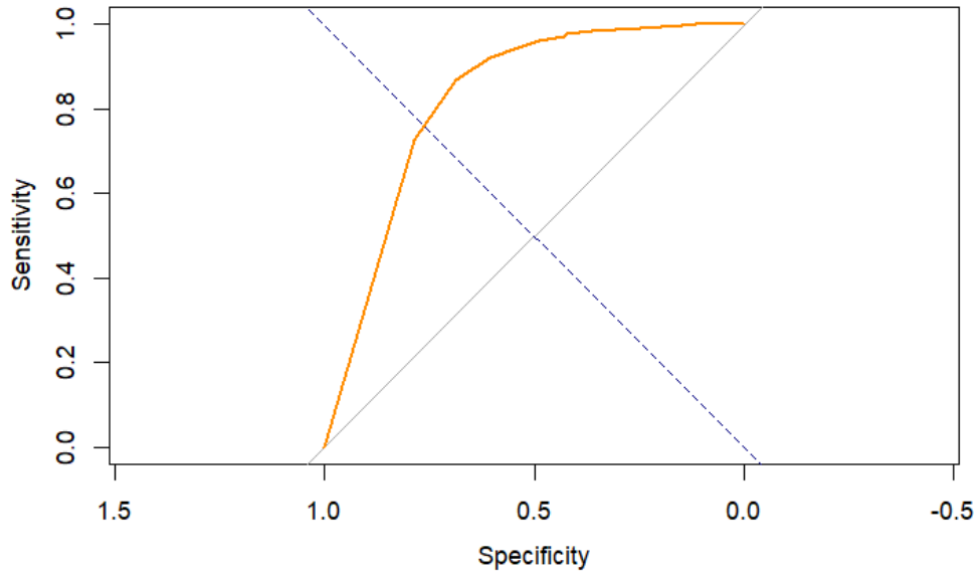
- Logistic regression: year, total receivables, total current liabilities, EBIT, current assets, EBITDA, inventory, depreciation and amortization, total long-term debt, market value.
- Random Forest: year, market value, total current liabilities, inventory, net income, retained earnings, total liabilities, net sales, current assets, total revenue.

This indicates that both algorithms independently recognized the time factor; market value, inventory, current assets (as proxies for size and operating activity), and total current liabilities (the increase in short- and long-term liabilities grows the risk of default when cash flows worsen) among the most significant in bankruptcy prediction. Earnings-related indicators such as net income, retained earnings, EBIT, and EBITDA also play a key role, since a decline in profitability and exhaustion of accumulated profits reduce financial stability. The level and dynamics of revenue (net sales, total revenue) are directly related to the ability to service debt.

The results of the Random Forest model demonstrated high overall predictive performance, although this was largely due to a strong imbalance in the data. The confusion matrix showed that the model correctly classified most non-bankrupt firms (15,613 classified correctly out of 15,614) but incorrectly classified most bankrupt ones. Out of the 121 bankrupt companies in the test set, only five were correctly predicted. This pattern

confirms that Random Forest learns most on the majority classes. Nevertheless, the model achieved a very high overall accuracy of 99.26%, which primarily reflects the overwhelming number of "alive" companies. The ROC curve shows an AUC of 0.82 (calculated with the "alive" class treated as positive here and later), indicating that the classifier can separate bankrupt companies from healthy firms with a probability of about 82% (Figure 4).

Figure 4. The ROC curve of Random Forest model (replication)

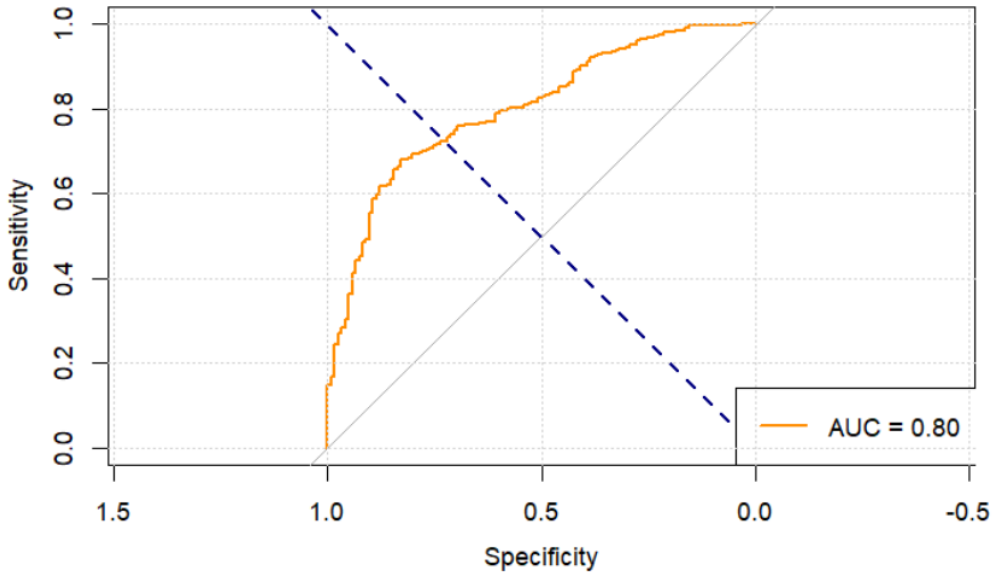


Cross-validation confirmed the stability of the model, with the accuracy values across the five folds ranging from 99.19% to 99.28%. However, high accuracy combined with low level of reproduction of the bankrupt class indicates that the base model is affected by class imbalance, which leads to the underrepresentation of bankrupt companies during training.

In the first stage of replication, the logistic regression achieved an accuracy of 99.21% with the average accuracy on five folds of 99.22%. ROC curve (AUC=0.8) demonstrates a good, though not perfect, model resolution, indicating its ability to identify companies with an increased risk of bankruptcy (Figure 5). At the same time, similarly to the RF model, the high overall accuracy is due to a significant class imbalance – most observations belong to the "alive" class, which makes model oversensitive. The confusion matrix showed only one

correctly identified bankruptcy out of 122, while three cases misclassified in "alive" class (15,611 classified correct).

Figure 5. The ROC curve of logistic regression model (replication)



In summary, the replication results confirm that while both logistic regression and Random Forest achieve very high overall accuracy, they perform poorly in detecting the minority class. This finding is consistent with the earlier research on this dataset, where models often misclassify bankrupt firms due to severe class imbalance and high multicollinearity between financial variables. Therefore, to increase sensitivity to bankruptcies, the next stage of the study applies SMOTE balancing and models' re-evaluation. Moreover, a correlation analysis revealed that two of the financial variables consist of the same data, so one of them (total revenue) is being removed from the future models.

After balancing the sample using the SMOTE method and removing variable X16 (total revenue), which completely duplicated X9 (net sales), the logistic regression and Random Forest models were re-evaluated. The use of SMOTE made it possible to correct the strong imbalance in the training sample: the proportions of companies with "alive" and "failed" status became equal to 50% each (Table 4), whereas previously there were 487 bankruptcies

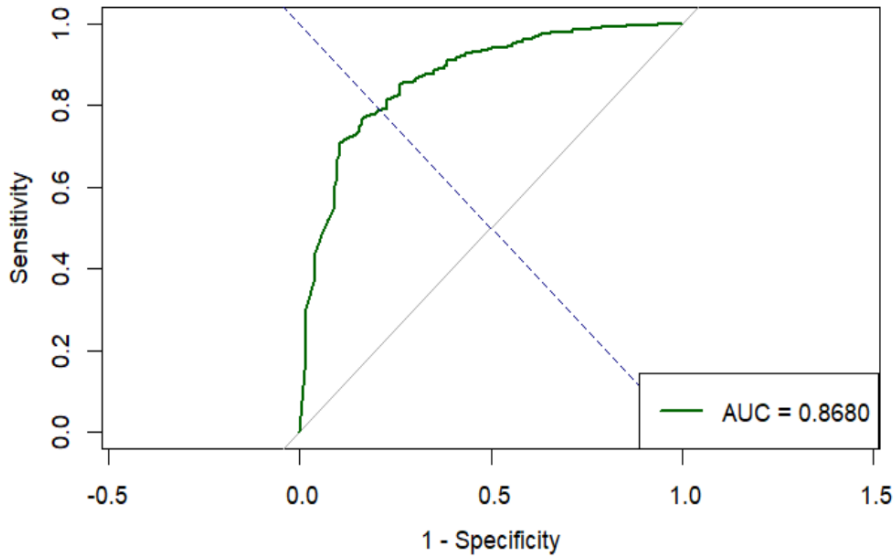
against 62,459 alive companies, and now there are approximately 62,000 in each class. This created conditions for more balanced training.

Table 4. The number of "alive" and "failed" classes in the dataset (before and after SMOTE)

Class distribution before SMOTE (train only):	
"Failed": 487	"Alive": 62459
Class distribution after SMOTE (train only):	
"Failed": 62336	"Alive": 62459
Rows before SMOTE: 62946	Rows after SMOTE: 62946

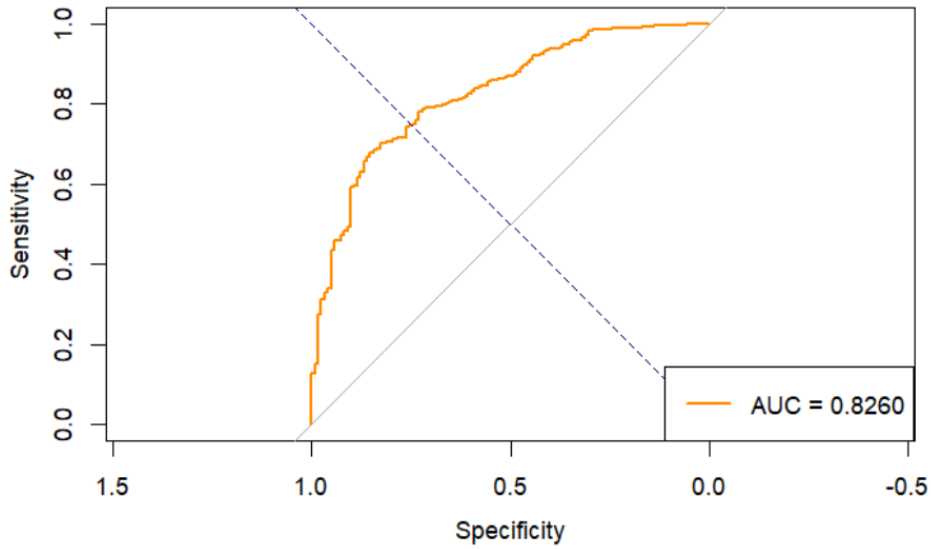
The Random Forest model, trained on a SMOTE sample and tested on a "clean" sample, demonstrates a high general accuracy of 98.32% (previously 99.26% before SMOTE implementation) and good ranking ability with ROC-AUC of 0.87 (AUC before SMOTE: 0.82) (Figure 6). At the same time the confusion matrix shows that, while having a high class imbalance, out of 122 real bankruptcies 27 were classified correctly (5 out of 121 for the model without SMOTE), while 15,444 were classified correctly out of the 15,614 companies in the "alive" class. Five-fold cross-validation yields an average value of 99.28%, which is consistent with the stability of the model during training.

Figure 6. The ROC curve of Random Forest model (after SMOTE)



The logistic regression showed a significant change in the structure of the results. The overall accuracy of the model on the test sample decreased to 68.63%, but this trend is potentially possible, as during training, the model was able to pay the same attention to both classes, and not only the majority one. The most important result is the growth in the ability of the model to recognize the bankrupt companies. If previously without the SMOTE usage just 1 out of 122 bankrupt companies were recognized, after the use, the algorithm identified 103 out of 122 bankruptcy cases. As a result, the sensitivity to the bankrupt class increased from 1% to nearly 84%. But at the same time the sensitivity to the "alive" class decreased as the model started to identify some of the stable companies as a potentially bankrupt one. Thus, 4,918 out of 15,614 cases were mistakenly identified as "failed" class. It explains the decrease in accuracy level as "alive" class observations prevail in the test sample. Nevertheless, the ROC-curve demonstrates a high resolution of the model with AUC=0.83 (0.80 in the previous logit regression), which indicates that the model has a good ability to separate risky firms from the stable ones (Figure 7).

Figure 7. The ROC curve of logistic regression model (after SMOTE)



After SMOTE, ten most significant variables for the logistic regression identified using ridge-regularized logistic regression, were those characterizing profitability, asset structure, and debt burden: year, net income, total receivables, EBIT, total current liabilities, current assets, EBITDA, inventory, depreciation and amortization, and gross profit. These features reflect the most stable predictors rather than those with the largest raw coefficients. So, the result shows an improvement in the overall quality of the classification for the rare class in the logistic regression, even if the accuracy decreases due to class balancing and the average cross-validation accuracy reached 70.25% (compared to 99.2% before SMOTE).

Robustness checks confirm that both models maintain stable discriminatory power, but differ significantly in calibration and remain weak in detecting bankruptcies. For logistic regression, bootstrap estimation (1000 replications) gives AUC of 0.83 with 95% CI, Brier of 0.19, and PR-AUC (failed as positive) of 0.07. For the threshold that maximizes F1 (t=0.75), we obtain precision of 0.14, and recall 0.29. This means that the model ranks risk fairly well overall, but the probabilities are moderately recalibrated and sensitivity remains low to bankruptcies (Table 5).

Table 5. Robustness check of LR and RF models (after SMOTE)

Model	AUC (alive)	Brier (alive)	PR-AUC (failed)	Best F1 threshold	Precision (failed)	Recall (failed)	F1 (failed)
Logistic Regression	0.826: [0.792; 0.860]	0.188	0.073	0.748	0.137	0.287	0.185
Random Forest	0.868: [0.836; 0.900]	0.015	0.081	0.539	0.165	0.213	0.186

Random Forest demonstrates better ranking ability and significantly more accurate probability forecasts of AUC 0.87 with 95% CI, Brier 0.02, and PR-AUC 0.08. At the best F1 (t=0.54), the precision is 0.17, recall 0.21. Thus, RF outperforms logit in terms of AUC and calibration, but both models still miss a proportion of bankruptcies on a "clean" unbalanced dataset.

In summary, SMOTE improved bankruptcy detection, especially for logistic regression, while Random Forest maintained high ranking efficiency. However, both models continued to miss a significant portion of bankruptcies in the unbalanced dataset. This pattern suggests that signals of financial distress are not the same for all companies and are likely to depend on their size. Accordingly, the next section presents the results of size-oriented models for identifying risk factors specific to each company size.

5.2. Results of the size-oriented models

To test whether the determinants of bankruptcy differ across companies of different sizes, separate logistic regression models were estimated for three size groups based on total assets (X10): small, medium, and large companies. Each model was trained on a balanced SMOTE subsample (20,982 records for each bin) and evaluated on a "clean" test set of 5,287 (48 bankrupt and 5,239 alive observations), 5,174 (42 bankrupt and 5,132 alive observations), and 5,275 (32 bankrupt and 5,243 alive observations) records, respectively.

The results show that the discriminatory power of the model improves with increasing company size (Table 6).

Table 6. Metrics of the logistic regression models by asset size

Bin	Train_n	Test_n	AUC (alive as positive)	Brier (alive as positive)	PR_AUC (failed as positive)	Accuracy
Small	41440	5287	0.79: [0.72; 0.86]	0.16	0.05	0.78
Medium	41457	5174	0.87: [0.79; 0.93]	0.10	0.14	0.86
Large	41688	5275	0.92: [0.87; 0.97]	0.13	0.17	0.81

The AUC rises from 0.79 for small companies to 0.92 for large firms, which indicates that financial ratios are more accurate in predicting financial difficulties among larger companies. Similarly, overall accuracy improves from 0.78 (small) to 0.86 (medium), but slightly decreases to 0.81 for large firms, while the Brier score decreases from 0.16 (small) to 0.13 (large), confirming better probability calibration in the large-company segment, but reports even better probability of calibration for medium bin with the value of 0.1.

However, the PR-AUC remains relatively low (0.05-0.17 across all bins), showing that even though the ranking ability improves, the minority "failed" class remains challenging to detect precisely. The ROC and Precision-Recall curves (Appendix B) show that as firm size increases, the model better recognizes the risk of bankruptcy due to more stable financial patterns, but even for large companies the accuracy of default detection remains limited due to the rarity of the bankrupt class.

The robustness of the size-specific logistic models was assessed using multicollinearity diagnostic (VIF). The results show that the level of collinearity differs significantly between

categories of firms. For small companies, most variables have acceptable VIF values below 5, and only a few indicators (three) show elevated values, which may be a consequence of the instability of financial indicators in this group. For medium-sized firms, there are more signs of multicollinearity: ten variables have VIF values above 5, and six have values above 10, indicating partial duplication of information between financial ratios. The highest level of collinearity is observed among large companies, where nine variables exceed the VIF > 5 threshold and seven exceed the VIF > 10, indicating strong correlations. Thus, the robustness of the logistic models is generally preserved, although stability decreases with firm size. For large enterprises, several financial variables are interrelated. Therefore, the coefficient estimates should be interpreted with caution, and the models should be viewed primarily as classification tools rather than instruments for assessing individual effects (Appendix C).

Overall, performance improves with the firm size increase but the minority-class detection remains limited and multicollinearity rises in larger groups. The next subsection presents size-oriented Random Forest models to examine whether nonlinear patterns further enhance prediction within each size segment.

Random Forest models were evaluated separately for three groups of companies based on total assets: small, medium, and large. Each model was trained on a balanced SMOTE sample (41,440 records for small bin, 41,457 for medium, and 41,688 for large) and evaluated on a "clean" test dataset similarly to the logistic-regression models. The test dataset consisted of 5,287 observations for small bin, 5,174 for medium, and 5,275 observations for large bin.

The performance table (Table 7) shows that the AUC values increase from 0.79 for small enterprises to 0.88 for medium-sized and 0.9 for large companies, while accuracy remains consistently high (0.98-0.99) for all bins.

Table 7. Metrics of the Random Forest by asset size

Bin	Train_n	Test_n	AUC (alive as positive)	Brier (alive as positive)	PR_AUC (failed as positive)	Accuracy
Small	41440	5287	0.79: [0.74; 0.84]	0.02	0.06	0.98
Medium	41457	5174	0.88: [0.83; 0.93]	0.01	0.09	0.98
Large	41688	5275	0.90: [0.83; 0.97]	0.01	0.17	0.99

However, this high accuracy largely reflects the correct classification of non-bankrupt ("alive") firms, as they dominate in the test sample. The Brier score decreases slightly with increasing company size (from 0.02 to 0.01), indicating a slight improvement in the calibration of predicted probabilities. The PR-AUC values are low for all models (0.06-0.17), reporting that the classification of the minority "failed" class remains difficult even after balancing. ROC-curves (Appendix D) show an improvement in discriminatory power across different size categories, with the ROC curve approaching the upper left corner, indicating the ability to distinguish between bankrupt and non-bankrupt companies. Overall, the results show that Random Forest models perform consistently well across all groups, with higher AUC scores and lower Brier scores for larger companies. This means that prediction accuracy and ranking ability improve with company size, although bankruptcy detection (as seen in PR-AUC) remains limited.

An additional check of the stability (Table 8) of variable importance in Random Forest shows a fairly high correlation between two independent importance metrics – Mean Decrease Accuracy (MDA) and Mean Decrease Gini (Gini importance), which indicates the consistency of model results in different approaches to assessing the contribution of predictors:

- For small companies, the correlation coefficient (0.71) indicates moderate stability: the model remains relatively reliable, but the importance of individual variables may fluctuate due to higher variation in financial indicators in this segment.
- For medium-sized companies, the correlation reaches the value of 0.89, demonstrating the highest consistency and the most stable estimates of variable importance. This means that the structure of financial factors for these companies is the most predictable.
- For large companies, the correlation is 0.73, which suggests good stability as well, though lower than the medium-sized companies have, possibly due to a more complex and interdependence financial structure of large companies.

Table 8. Correlation between MDA and Gini importance (robustness check) for RF size-oriented models

	Small bin	**Medium bin**	**Large bin**
Correlation between MDA & Gini	0.71	0.89	0.73

In general, the obtained values confirm the robustness of Random Forest size-segmented models, since the high level of correlation between different importance criteria suggests that the models reliably identify the same key determinants of bankruptcy in each dimensional segment.

Therefore, the comparison of logistic regression and Random Forest models by size bins shows higher discriminatory power for larger firms, but their strengths differ. Logistic regression gives higher ROC-AUC than RF in the large bin, while Random Forest consistently has a much lower Brier and very high overall accuracy ("alive" as a positive class) across all bins. At the same time for the task of detecting the rare class "failed" (where PR-AUC is calculated specifically for bankruptcies), both models remain limited. It is also worth noting the robustness in terms of feature importance where RF shows high

consistency between MDA and Gini, while LR shows singularities as noticeable multicollinearity in all bins, so the coefficients should be interpreted with caution.

5.3. Identification of key factors influencing bankruptcy

The graphs (Appendix H-J) that show the predictor, its significance level and beta coefficient were constructed for all three size categories of logistic regression models to identify the key factors influencing the bankruptcy. The visual representation showed that most of the factors in all models are concentrated around zero, and only a few predictors are statistically significant while being not around zero (Appendix E-G). This indicates that individual financial ratios have limited explanatory power when analyzed separately, and the predictive power of the model is mainly due to their combined interaction. For small and medium enterprises, the coefficients show instability and potential separation effects, while for large firms, the estimates become more consistent:

1. For small companies, most indicators are statistically significant ($p < 0.05$) but have a magnitude of about zero, so no specific conclusions can be made regarding their marginal impact. But from the table of results it's visible that current assets, and market value are significant and have positive direction on influence on the company's "alive" status, indicating that as the variable increases, the probability of bankruptcy decreases. While year, total current liabilities, total long-term debt, and total liabilities increase the bankruptcy probability (year and total current liabilities with the biggest amplitude).

2. For medium-sized companies, 11 out of 16 indicators are statistically significant, but all of them are around zero in magnitude as well. While EBIT and depreciation and amortization show the biggest positive effect on the company becoming bankrupt and EBITDA has the biggest negative effect on bankruptcy, all in the value nearly +- 60, they stay not statistically significant with $p > 0.05$. Nevertheless, from the table of coefficient results before each of the predictors, it can be seen that market value, current assets, cost of goods sold, total assets, net income have

a positive impact on the company staying in the "alive" class, while total liabilities, total long-term debt, total current liabilities, and year have a positive impact on the probability of company getting into the "bankrupt" class.

3. For large companies, 12 indicators are statistically significant and the extent and direction of their impact is evident for most of them even out of the graph where the value of beta coefficient is more far away from the zero than in the small- and medium-sized bins. For example, year, total current liabilities, total liabilities, total long-term debt clearly show the negative impact on alive status of the company, that means that they increase the probability of company becoming bankrupt. At the same time total assets, net income, market value, current assets, total receivables, and cost of goods sold (to a lesser extent, since the coefficient is close to 0) present a negative effect on company becoming bankrupt which is consistent with the economic logic of the financial situation of companies.

4. At he same time, some factors have an unexpected effect on the probability of a company going bankrupt. For example, for the medium bin, net sales based on the sign before the coefficient should have a positive effect on bankruptcy, but since all such variables that have a questionable economic impact in this case have a coefficient close to 0, their interpretation should be particularly cautions, as it is not determined by magnitude and may be controversial.

Therefore, logistic regression analysis shows that most financial indicators have limited individual explanatory power, and bankruptcy prediction largely depends on their combined interaction. For small and medium-sized enterprises, the coefficients remain close to zero and unstable, indicating a weaker distinction between bankrupt and operating companies. In contrast, large enterprises show more consistent patterns: total current liabilities, total liabilities, and total long-term debt increase the probability of bankruptcy, while higher total assets, net income, market value, current assets, and total receivables

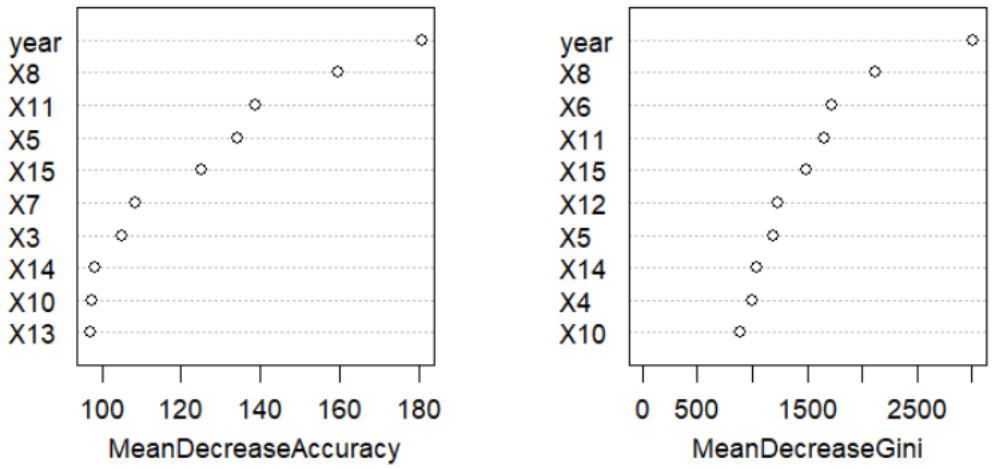
decrease it. These findings confirm that financial structure and profitability play a stabilizing role, while excessive debt increases the risk of bankruptcy, especially in large enterprises.

To identify the key factors that influence bankruptcy prediction in size-oriented Random Forest models Mean Decrease Accuracy and Mean Decrease Gini were used. In RF, importance shows the contribution to the forecast, not the direction, so here is the interpretation of the importance for each firm size.

Small firms (Figure 8):

- Consistently the most important (high in both metrics): year, market value, total long-term debt, retained earnings.
- Method-dependent: net income is high in Gini but lower in MDA; inventory is noticeable in MDA but placed lower in Gini. Total assets, EBITDA, gross profit are secondary.

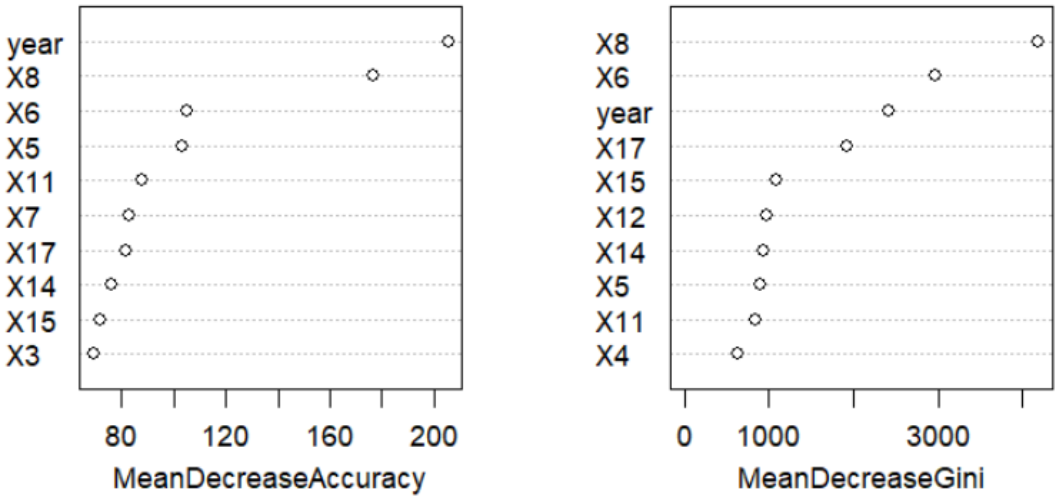
Figure 8. Mean Decrease Accuracy and Mean Decrease Gini for size-oriented Random Forest model (small)



Medium firms (Figure 9):

- Agreed core predictors (high in both metrics): year, market value, net income.
- Additionally: retained earnings, total liabilities, and EBIT rise in Gini. Inventory, total receivables, and total long-term debt are more important in MDA.

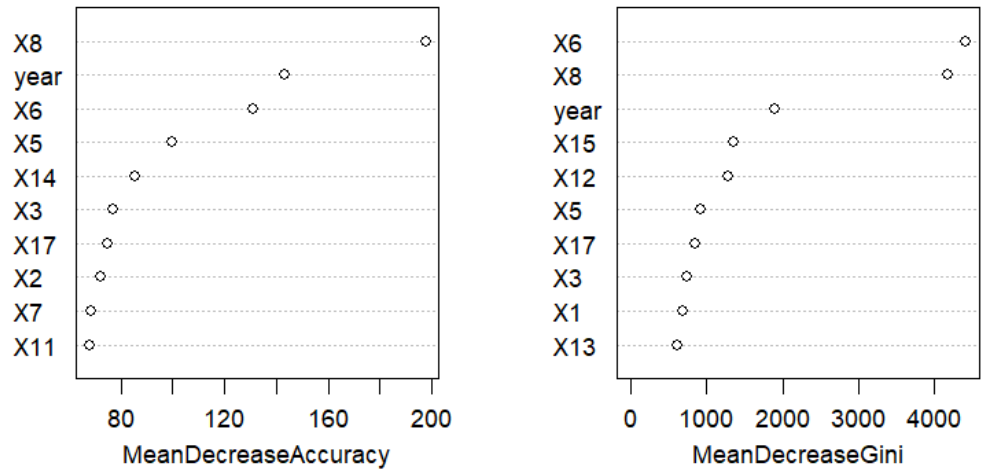
Figure 9. Mean Decrease Accuracy and Mean Decrease Gini for size-oriented Random Forest model (medium)



Large firms (Figure 10):

- Common core (high in both metrics): market value, net income, year.
- Method-dependent: inventory, total long-term debt, total receivables are still noticeable in MDA. Total liabilities, retained earnings, cost of goods sold are among the key ones in Gini.

Figure 10. Mean Decrease Accuracy and Mean Decrease Gini for size-oriented Random Forest model (large)



In summary, in all three groups, the key factors for forecasting are market value, net income (top-3 in MDG for small companies), and year – these are the ones that consistently make the largest contribution, so market valuation, profitability, and time effects are the best ways to distinguish between "failed" and "alive". In small firms, the importance of retained earnings and long-term debt (sensitivity to capital structure) increases. In medium-sized firms, total liabilities and EBIT (debt and margin management risks) are added. While in large firms, the core remains but liabilities/retained earnings are more noticeable in Gini.

CHAPTER 6. CONCLUSIONS AND RECOMMENDATIONS

This thesis aims to investigate whether financial indicators can reliably predict the bankruptcy of companies of different sizes and whether the determinants of bankruptcy differ for small, medium, and large enterprises. Combining classical and machine learning methods, namely logistic regression and Random Forest, on a large and standardized dataset, the study aims to test the reliability, interpretability, and heterogeneity of bankruptcy prediction models.

The study makes three key methodological contributions. First, it corrects a common mislabeling problem in previous Kaggle-based studies by labeling companies as bankrupt only once – in the year immediately preceding bankruptcy, thereby bringing the dataset in line with the proper event study logic. Second, it applies SMOTE balancing to mitigate severe class imbalance and increase sensitivity to the rare class ("failed"). Third, it implements a size segmentation approach based on total assets, allowing small, medium, and large firms to be modeled separately – a new addition to this dataset.

Initial replications confirmed that both logistic regression and Random Forest achieve very high overall accuracy but have low efficiency in identifying bankrupt companies due to class imbalance. After implementing SMOTE balancing, the sensitivity of the logistic model to bankruptcies increased significantly, while Random Forest retained good ranking. This confirmed that class balancing improves the detection of minority classes, even if it slightly reduces overall accuracy. When the models were evaluated separately for small, medium, and large companies, clear patterns emerged. Predictive performance improved with company size in both approaches. Larger companies exhibited more stable and predictable financial relationships, while smaller firms showed higher volatility and weaker signals. Brier estimates confirmed better calibration for larger companies, and Random Forest consistently reported very low Brier values, indicating better calibrated probability estimate. However, PR-AUC values remained low across all models, indicating that precise bankruptcy detection remains challenging even after balancing.

The overall comparison confirmed that Random Forest outperformed logistic regression in ranking ability (AUC) and probability calibration nearly in every model, while the latter remained more interpretable in terms of the size and the direction of influence, especially for large firms. RF demonstrated high internal consistency, with a correlation between MDA and MDG above 0.7. In contrast, logistic regression showed an increase in multicollinearity with company size, which reduced the clarity of coefficient interpretation.

Among the key determinants of bankruptcy both models consistently identified market value, profitability, leverage, and time effects as the most influential determinants of bankruptcy, although their role varied depending on the size of the company:

- Small enterprises: market value, total long-term debt, year, and retained earnings as the variables reflecting sensitivity to capital structure and financial constraints, had the greatest impact on the probability of bankruptcy. As well as net income and inventory important in MDA and MDG, respectively.
- Medium-sized companies: the most predictive indicators were market value, net income, and year, while retained earnings, liabilities and inventory were also important.
- Large firms: the most predictive indicators were market value, net income, and year, inventory (MDA), and total liabilities (MDG). The logistic regression coefficients for large companies were farther from zero than in smaller firms, which made them more interpretable; net income, current assets, and total current liabilities showed the strongest effects.

That is, there is a common "core" that often serves as a key factor influencing the bankruptcy of all companies of different sizes (market value, year, net income), but the importance profile varies depending on the size: small enterprises are more sensitive to capital structure (total long-term debt, retained earnings), medium-sized firms to retained earnings and liabilities, and large enterprises to inventories and total liabilities. This

confirms the relevance of size-oriented modeling. In all groups, the year remained an important variable reflecting systemic cycles and macroeconomic shocks. Importantly, both models confirmed that financial ratios are not universal predictors, their impact and importance depends on the size of the company.

These findings confirm the classic thesis of Altman (1968) and Ohlson (1980) that profitability, leverage, and liquidity remain key factors in predicting financial difficulties for firms. However, they also show that the strength and direction of these effects are not uniform. In practice, results show that the early warning systems and credit risk models need to be adapted to size. A pooled model fails to capture signals of financial distress typical for different companies' size, whereas separate calibration by company size may significantly improve detection accuracy.

Although the thesis provides new insights into size-based bankruptcy modeling, several limitations for future development remain:

- Since the dataset includes only public companies, private SMEs may exhibit different risk patterns.
- The models do not include macroeconomic or sectoral variables that could improve interpretability and accuracy.
- Further work on screening metrics for analysis and testing models can improve the accuracy and quality of identifying bankrupt companies, which will improve overall results and interpretation.
- Applying the same framework to emerging markets such as Ukraine could test the transferability of the results and support the development of size-sensitive early warning systems once structured firm-level reporting becomes available.

In summary, this thesis establishes that bankruptcy risk is not size-neutral: the predictive power and significance of financial factors differ significantly for small, medium, and large firms. Random Forest provides better ranking and calibration, while logistic

regression remains more interpretable, especially for large enterprises. Implementing the proposed enhancements could further improve the detection of the minority class in models without losing interpretability, while also providing important insights into the impact of factors on the stability of homogeneous companies.

REFERENCES

Ahmed Ibrahim. 2023. US Company Bankruptcy Classification. *Kaggle*. https://www.kaggle.com/code/ahmedebrahiim/us-company-bankruptcy-classification.

Alejandro Ignacio Espinoza. 2024. US Company Bankruptcy Prediction | RF + GB. *Kaggle*. https://www.kaggle.com/code/alejandroee/us-company-bankruptcy-prediction-rf-gb.

Alejandro Ignacio Espinoza. 2024. US Company Bankruptcy Prediction | XGB +SVM. *Kaggle*. https://www.kaggle.com/code/alejandroee/us-company-bankruptcy-prediction-xgb-svm.

Altman Edward. 1968. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance* 23(4): 589–609. https://doi.org/10.2307/2978933.

Andrii Yarko. 2023. Desicion Trees 97% accuracy. *Kaggle*. https://www.kaggle.com/code/yarkoandriy/desicion-trees-97-accuracy.

Annual Business and Non-Business Filings by Year (1980-2021). *American Bankruptcy Institute*. https://abi-org.s3.amazonaws.com/Newsroom/Bankruptcy_Statistics/Total-Business-Consumer1980-Present.pdf.

Arcuri, Giuseppe & Succurro, Marianna & Costanzo, Giuseppina. 2019. A combined approach based on Robust PCA to improve bankruptcy forecasting. *Review of Accounting and Finance*. 10.1108/RAF-04-2018-0077.

Ayushmman Saini. 2023. EDA for Companies Financial data. *Kaggle*. https://www.kaggle.com/code/ayushmmansaini/eda-for-companies-financial-data.

Beaver William. 1966. Financial Ratios As Predictors of Failure. *Journal of Accounting Research* 4: 71–111. https://doi.org/10.2307/2490171.

Bénard C., Veiga S., & Scornet E. 2022. MDA for random forests: inconsistency, and a practical solution via the Sobol-MDA. *Cornell University Library*. 1-66. https://arxiv.org/pdf/2102.13347.

Broendsholm. 2023. US Company Bankruptcy Prediction | 100% accuracy. *Kaggle*. https://www.kaggle.com/code/broendsholm/us-company-bankruptcy-prediction-100-accuracy.

Castillo JA, Mora-Valencia A, Perote J. 2018. Moral hazard and default risk of SMEs with collateralized loans. Finance Research Letters, Elsevier, vol. 26(C), pages 95-99.

Cuellar, Beatriz & Serrano-Cinca, Carlos. 2020. Predicting startup survival using first years financial statements. Journal of Small Business *Management*. 60. 1-37. 10.1080/00472778.2020.1750302.

Cultrera, L. and Brédart, X. 2016. Bankruptcy prediction: the case of Belgian SMEs. *Review of Accounting and Finance*, Vol. 15 No. 1, pp. 101-119. https://doi.org/10.1108/RAF-06-2014-0059.

El Kalak, Izidin & Hudson, Robert. 2015. The effect of size on the failure probabilities of SMEs: An empirical study on the US market using discrete hazard model. *International Review of Financial Analysis*. 43. 10.1016/j.irfa.2015.11.009.

Ezeyinwa Prisca. 2023. american_bankruptcy. *Kaggle*. https://www.kaggle.com/code/ezeyinwaprisca/american-bankruptcy.

Gallucci, Carmen & Santullli, Rosalia & Modina, Michele & Formisano, Vincenzo. 2023. Financial ratios, corporate governance and bank-firm information: a Bayesian approach to predict SMEs' default. *Journal of Management and Governance*. 27. 10.1007/s10997-021-09614-5.

Giordani P., Jacobson T., Schedvin E., & Villani M.. 2011. Taking the Twists into Account: Predicting Firm Bankruptcy Risk with Splines of Financial *Ratios. Sveriges Riksbank Working Paper Series*, 256, 1-55. https://www.econstor.eu/bitstream/10419/81877/1/679599274.pdf.

Gupta, J., Gregoriou, A. & Healy, J. 2015. Forecasting bankruptcy for SMEs using hazard function: To what extent does size matter?. *Rev Quant Finan Acc* 45, 845–869. https://doi.org/10.1007/s11156-014-0458-0.

Gurnani, Ishika & Vincent, & Tandian, Febryan & Anggreainy, Maria. 2021. Predicting Company Bankruptcy Using Random Forest Method. *2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS)* 1-5. 10.1109/AiDAS53897.2021.9574384.

HLI111111. 2025. notebookb76ad15a49. *Kaggle*. https://www.kaggle.com/code/hli111111/notebookb76ad15a49.

Jarvis, P. 2024. A comparison of different methods for bankruptcy prediction. *Dissertation*. https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-531744.

Jay Dixit. 2023. US_company_bankrupt. *Kaggle*. https://www.kaggle.com/code/jayrdixit/us-company-bankrupt.

kalashnikov1405. 2024. Bank Thief. *Kaggle*. https://www.kaggle.com/code/kalashnikov1405/bank-thief.

leesstephanie. 2024. Time Series Classification for Business. *Kaggle*. https://www.kaggle.com/code/leesstephanie/time-series-classification-for-business.

Levi Sverdlov. 2023. Bankcuptcy prediction via XGBoost: AUC ROC = 0.89. *Kaggle*. https://www.kaggle.com/code/levisverdlov/bankcuptcy-prediction-via-xgboost-auc-roc-0-89.

Nada H., Newton431, Elgouhary Y. & Rodina A. 2024. US Company Bankruptcy Prediction 87% high recall. *Kaggle*. https://www.kaggle.com/code/naddamuhhamed/us-company-bankruptcy-prediction-87-high-recall.

Ohlson James. 1980. Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research* 18(1): 109–131. https://doi.org/10.2307/2490395.

Séverin E. & Veganzones D. 2021. Can earnings management information improve bankruptcy prediction models?. *Annals of Operations Research*. 306. 10.1007/s10479-021-04183-0.

Shumway, T. 2001. Forecasting Bankruptcy More Accurately: A Simple Hazard Model. *The Journal of Business*, 74(1), 101–124. https://doi.org/10.1086/209665.

Singh U. 2023. US Company Bankruptcy Prediction Dataset. *Kaggle*. https://www.kaggle.com/datasets/utkarshx27/american-companies-bankruptcy-prediction-dataset.

Singh U. 2023. US Company Bankruptcy Prediction: 93% Accuracy. *Kaggle*. https://www.kaggle.com/code/utkarshx27/us-company-bankruptcy-prediction-93-accuracy.

stpete_ishii. 2023. US Bankrupsy Visualize Importance. *Kaggle*. https://www.kaggle.com/code/stpeteishii/us-bankrupsy-visualize-importance.

Team T. I. Variance Inflation Factor (VIF): Definition and Formula. *Investopedia*. https://www.investopedia.com/terms/v/variance-inflation-factor.asp.

United States Bankruptcies. *Trading Economics | 20 million INDICATORS FROM 196 COUNTRIES*. URL: https://tradingeconomics.com/united-states/bankruptcies.

U.S. Code: Title 11 - Bankruptcy. Cornell Law School. *Legal Information Institute*. https://www.law.cornell.edu/uscode/text/11.

Vater T. & Wolf S. 2023. Bankruptcy Prediction via Earnings Distributions. *SSRN*. http://dx.doi.org/10.2139/ssrn.3929521.

Wang, W., Guedes, M.J. 2024. Firm failure prediction for small and medium-sized enterprises and new ventures. *Rev Manag Sci.* https://doi.org/10.1007/s11846-024-00742-4.

Yu Q., Miche Y., Séverin E. & Lendasse A. 2014. Bankruptcy prediction using Extreme Learning Machine and financial expertise. *Neurocomputing*. 128. 296-302. https://doi.org/10.1016/j.neucom.2013.01.063.

Yuval Glasner. 2025. US Companies Bankruptcy Prediction: Altman vs. ML. *Kaggle*. https://www.kaggle.com/code/yuvalglasnermarin/us-companies-bankruptcy-prediction-altman-vs-ml.

Zhang, D., & Xie, J. 2021. Influence of Tourism Seasonality and Financial Ratios on Hotels' Exit Risk. *Journal of Hospitality & Tourism Research*, 47(4), 714-733. https://doi.org/10.1177/10963480211016038 (Original work published 2023).

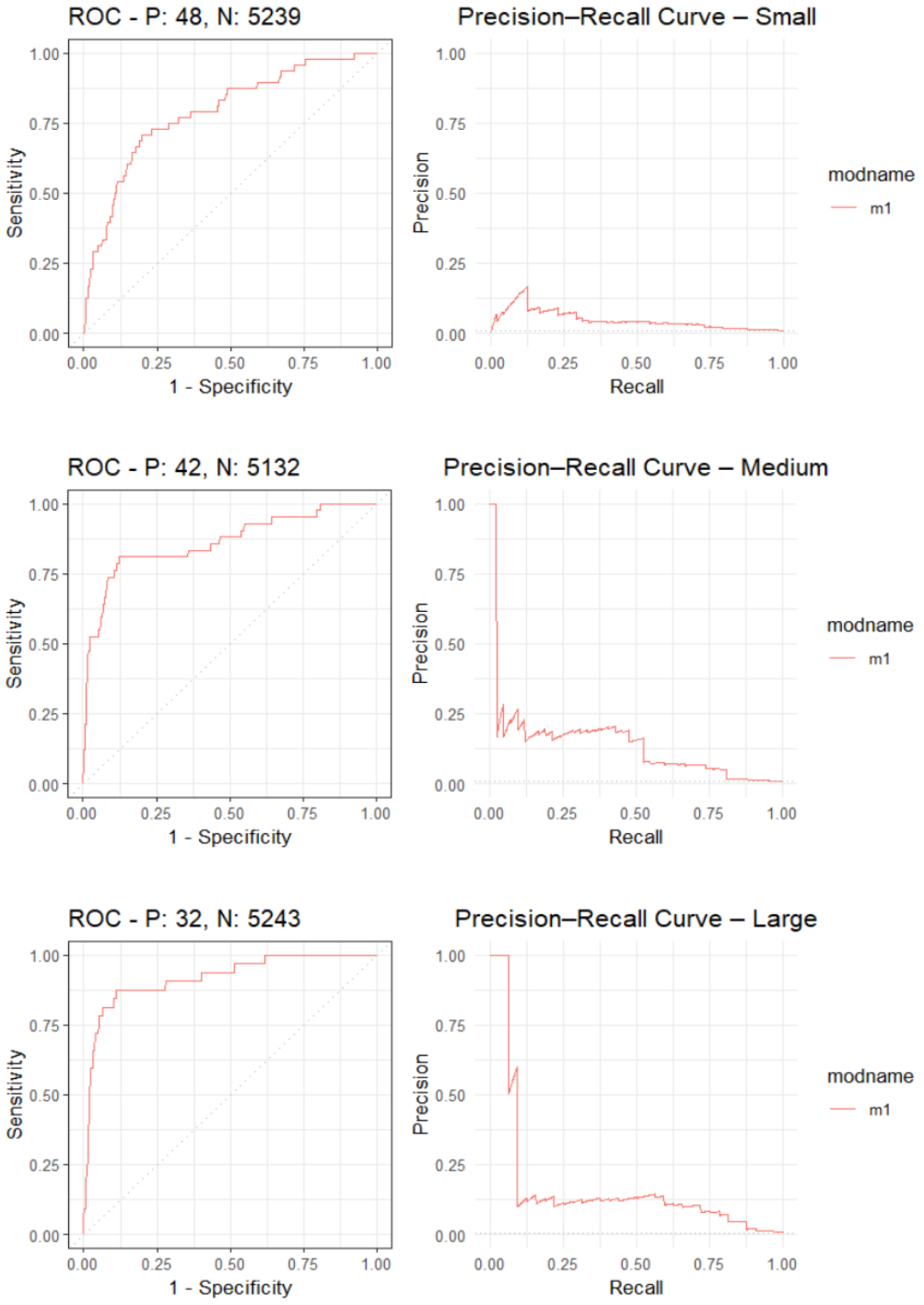
APPENDICES

Appendix A.9. The summary table of logistic regression model (after SMOTE)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.30e+02	2.71e+00	48.09	< 2e-16	***
year	-6.51e-02	1.35e-03	-48.18	< 2e-16	***
X1	3.33e-03	7.71e-05	43.24	< 2e-16	***
X2	1.25e+04	3.61e+05	0.04	0.97	
X3	-7.83e+01	2.94e+03	-0.03	0.98	
X4	1.26e+04	3.61e+05	0.04	0.97	
X5	-2.28e-03	1.09e-04	-20.91	< 2e-16	***
X6	5.33e-04	5.58e-05	9.55	< 2e-16	***
X7	6.18e-04	1.22e-04	5.06	4.27e-07	***
X8	1.13e-03	1.86e-05	60.77	< 2e-16	***
X9	-2.50e+04	7.22e+05	-0.04	0.97	
X10	2.95e-04	2.68e-05	11.01	< 2e-16	***
X11	-6.41e-04	4.07e-05	-15.76	< 2e-16	***
X12	-7.83e+01	2.94e+03	-0.03	0.98	
X13	1.25e+04	3.61e+05	0.04	0.97	
X14	-3.18e-03	6.41e-05	-49.61	< 2e-16	***
X15	5.47e-05	8.34e-06	6.56	5.44e-11	***
X17	-5.00e-04	4.08e-05	-12.26	< 2e-16	***
X18	1.25e+04	3.61e+05	0.04	0.97	

Note: ***p<0.001; **p<0.01; *p<0.05

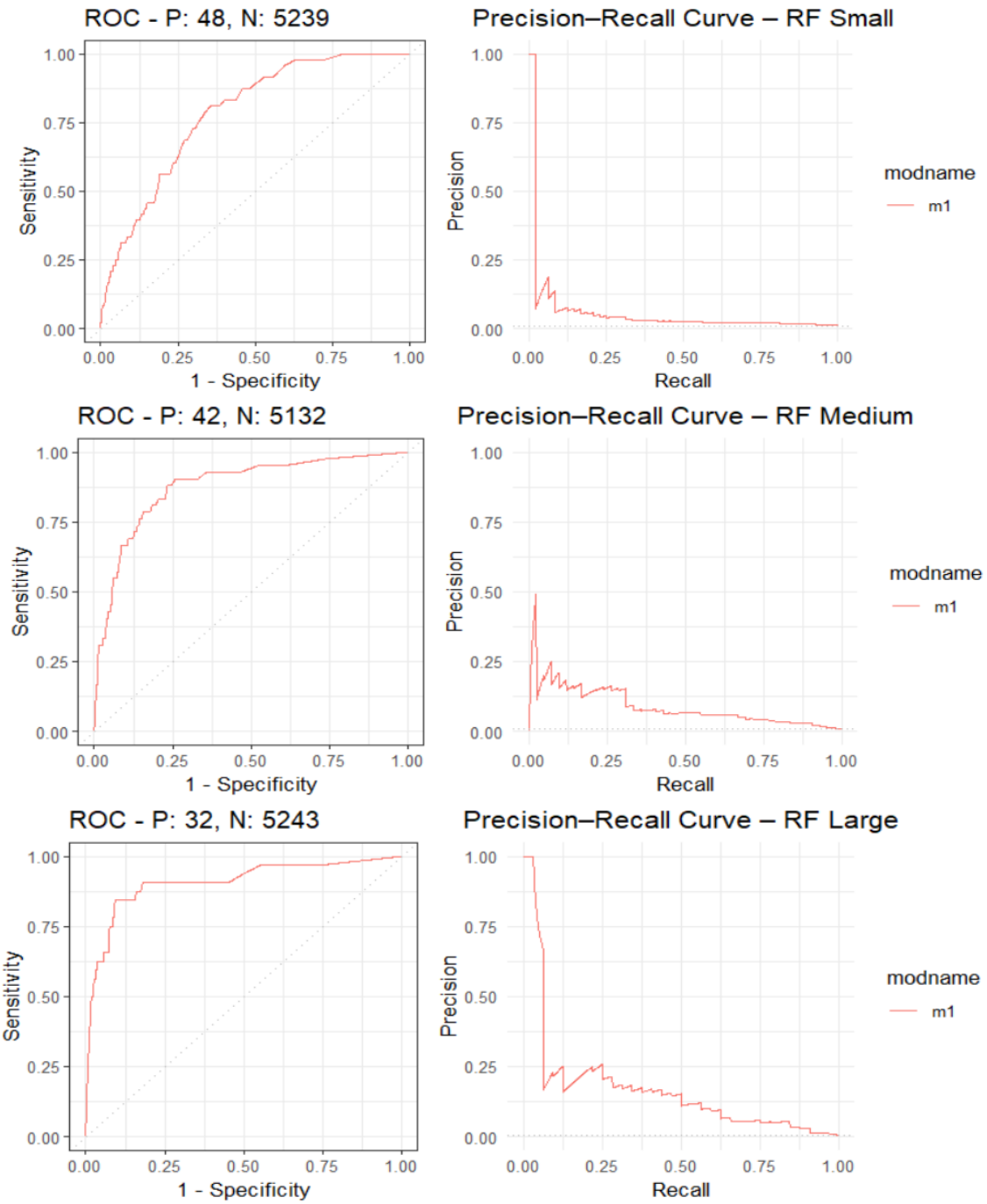
Appendix B.11. ROC and Precision-Recall curves for size-specific logistic regression models



Appendix C.10. The VIFs values for all predictors in size-segmented logistic-regression models (small, medium, and large companies)

Small		**Medium**		**Large**	
Term	*VIF*	*Term*	*VIF*	*Term*	*VIF*
X13	2.35e+16	X4	7.96e+11	X9	71.92
X9	1.49e+15	X12	6.90e+11	X2	56.95
X17	6.44e+00	X3	3.20e+11	X17	47.77
X14	4.15e+00	X9	3.58e+01	X10	29.25
X11	4.04e+00	X2	2.82e+01	X11	13.93
X5	2.84e+00	X17	1.38e+01	X1	11.48
X7	1.85e+00	X1	7.74e+00	X14	10.24
X10	1.31e+00	X11	7.09e+00	X4	6.05
X3	1.28e+00	X10	5.77e+00	X7	5.08
X15	9.60e-01	X14	5.03e+00	X5	4.37
X8	7.20e-01	X7	3.80e+00	X3	3.93
X6	-2.90e-01	X5	3.55e+00	X8	2.99
Year	-5.41e+01	X6	2.10e+00	X6	2.09
X1	-8.91e+01	X8	1.740e+00	X15	1.92
X4	-8.52e+14	X15	1.45e+00	Year	1.05
X2	-1.96e+16	Year	1.11e+00		
X18	-3.12e+16				

Appendix D.12. ROC and Precision-recall curves for size-specific Random Forest models



Appendix E.11. The summary table of logistic regression model (small bin)

Coefficients: (1 not defined because of singularities)					
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	6.12e+01	5.22e+00	11.71	< 2e-16	***
year	-3.04e-02	2.60e-03	-11.67	< 2e-16	***
X1	1.67e-01	3.92e-03	42.55	< 2e-16	***
X2	8.92e+03	1.20e+05	0.08	0.94	
X3	1.44e-01	1.40e-02	10.27	< 2e-16	***
X4	8.89e+03	1.20e+05	0.07	0.94	
X5	-1.65e-01	4.86e-03	-33.92	< 2e-16	***
X6	-2.83e-02	2.36e-03	-11.98	< 2e-16	***
X7	-1.05e-01	5.81e-03	-18.03	< 2e-16	***
X8	2.40e-02	5.97e-04	40.13	< 2e-16	***
X9	-1.78e+04	2.39e+05	-0.08	0.94	
X10	1.94e-03	1.81e-03	1.07	0.28	
X11	-1.15e-02	4.12e-03	-2.79	0.00	**
X12	NA	NA	NA	NA	
X13	8.92e+03	1.20e+05	0.08	0.94	
X14	-4.51e-02	4.21e-03	-10.72	< 2e-16	***
X15	2.55e-03	1.53e-04	16.63	< 2e-16	***
X17	-2.64e-02	3.54e-03	-7.46	8.85e-14	***
X18	8.89e+03	1.20e+05	0.07	0.94	

Note: ***p<0.001; **p<0.01; *p<0.05

Appendix F.12. The summary table of logistic regression model (medium bin)

Coefficients: (2 not defined because of singularities)					
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.16e+02	6.43e+00	17.97	< 2e-16	***
year	-5.76e-02	3.21e-03	-17.97	< 2e-16	***
X1	7.98e-03	5.34e-04	14.94	< 2e-16	***
X2	5.95e-03	3.24e-04	18.38	< 2e-16	***
X3	-5.74e+01	3.98e+02	-0.14	0.89	
X4	5.75e+01	3.98e+02	0.14	0.89	
X5	-8.28e-03	6.65e-04	-12.46	< 2e-16	***
X6	1.80e-03	3.46e-04	5.18	2.18e-07	***
X7	-1.78e-04	7.95e-04	-0.22	0.82	
X8	8.17e-03	1.80e-04	45.30	< 2e-16	***
X9	-6.11e-03	3.05e-04	-20.02	< 2e-16	***
X10	1.88e-03	2.75e-04	6.84	8.15e-12	***
X11	-7.21e-03	4.95e-04	-14.57	< 2e-16	***
X12	-5.74e+01	3.98e+02	-0.14	0.89	
X13	NA	NA	NA	NA	
X14	-1.43e-02	5.83e-04	-24.60	< 2e-16	***
X15	-7.32e-05	4.52e-05	-1.62	0.10	
X17	-9.98e-04	4.65e-04	-2.15	0.03	*
X18	NA	NA	NA	NA	

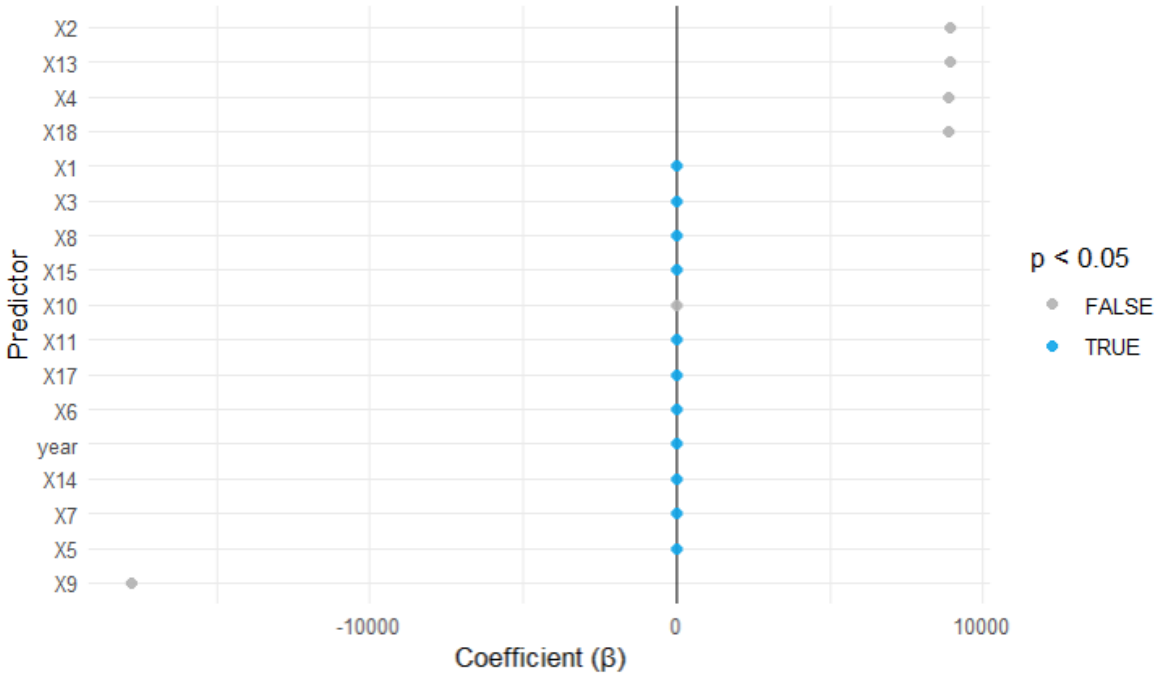
Note: ***p<0.001; **p<0.01; *p<0.05

Appendix G.13. The summary table of logistic regression model (large bin)

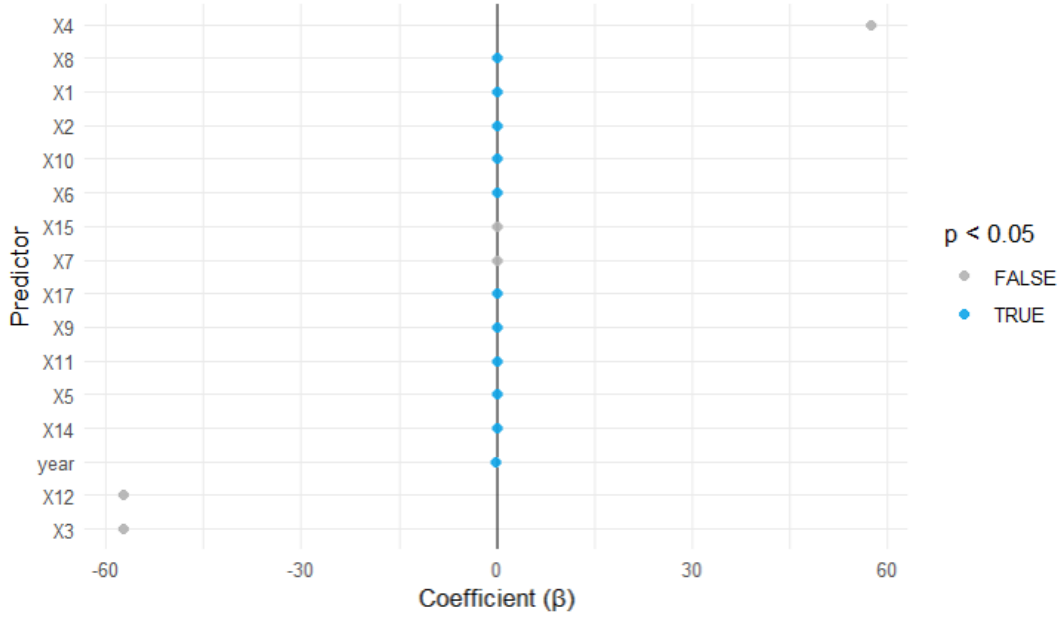
Coefficients: (3 not defined because of singularities)					
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.52e+02	5.34e+00	28.37	< 2e-16	***
year	-7.57e-02	2.66e-03	-28.48	< 2e-16	***
X1	1.23e-03	5.12e-05	24.00	< 2e-16	***
X2	1.54e-04	3.96e-05	3.88	0.00	***
X3	-1.27e-04	1.26e-04	-1.01	0.31	
X4	7.12e-05	6.50e-05	1.10	0.27	
X5	-5.07e-05	8.43e-05	-0.60	0.55	
X6	4.44e-04	3.67e-05	12.13	< 2e-16	***
X7	1.53e-03	9.37e-05	16.38	< 2e-16	***
X8	6.39e-04	1.28e-05	49.87	< 2e-16	***
X9	-1.73e-04	3.53e-05	-4.90	9.73e-07	***
X10	2.82e-04	1.78e-05	15.90	< 2e-16	***
X11	-2.81e-04	2.88e-05	-9.74	< 2e-16	***
X12	NA	NA	NA	NA	
X13	NA	NA	NA	NA	
X14	-1.58e-03	4.59e-05	-34.41	< 2e-16	***
X15	-5.36e-05	9.68e-06	-5.54	3.02e-08	***
X17	-4.23e-04	2.85e-05	-14.88	< 2e-16	***
X18	NA	NA	NA	NA	

Note: ***p<0.001; **p<0.01; *p<0.05

Appendix H.13. The β (beta) coefficients of predictors of logistic regression (small bin)



Appendix I.14. The β (beta) coefficients of predictors of logistic regression (medium bin)



Appendix J.15. The β (beta) coefficients of predictors of logistic regression (large bin)