Bankruptcy Prediction: A Comparison of Data Mining Models on Unbalanced Data and Effects of Sampling

Gunin Ruthwik Javvadi

A Thesis

In The Department

of

Supply Chain and Business Technology Management

Presented in Partial Fulfillment of the Requirements

For the Degree of

Master of Science (Business Analytics and Technology Management)

at Concordia University

Montréal, Québec, Canada

November 2023

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Gunin Ruthwik Javvadi

Entitled: Bankruptcy Prediction: A Comparison of Data Mining Models on Unbalanced Data and Effects of Sampling

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Business Analytics and Technology Management)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

_____ Chair

Dr. Rustom Vahidov

_____ Examiner

Dr. *Suchit Ahuja*

_____ Examiner

Dr. *Danielle Morin*

_____ Supervisor

Dr. *Salim lahmiri*

Approved by _____

Dr. Suchit Ahuja, Graduate Program Director

November 28th 2023 _____

Dr. Anne-Marie Croteau, Dean of Faculty

ABSTRACT

Bankruptcy Prediction: A Comparison of Data Mining Models on Unbalanced Data and Effects of Sampling

Gunin Ruthwik Javvadi

With the very unbalanced data found in financial risk prediction, this study hopes to aid in anticipating the financial risk that corporations may encounter. We can improve performance by employing oversampling and under-sampling algorithms. We were able to better understand how the performance of each classifier changed in each dataset by using a variety of classifiers across three distinctively sampled datasets. In addition, we analyzed our dataset using three different evaluation metrics: accuracy, sensitivity, and specificity, rather than being limited to just one. The results indicate that the accuracy on three separate datasets with various sampling methods differs greatly. The sensitivity and specificity of the under-sampled dataset differ from those of the original dataset and the oversampled dataset, which are fairly comparable to one another. It was discovered that gradient boosting trees produce better outcomes than other algorithms. When using oversampled data and measuring accuracy, logistic regression was found to be the most effective. However, when using under-sampled data, LightGBM Classifier had the best performance. Additionally, when considering sensitivity and specificity, CatBoost Classifier was the best choice.

Keywords: Bankruptcy, unbalanced data, classification

Acknowledgments

I would like to express my deepest appreciation to my supervisor, Dr. Salim Lahmiri. I'm grateful to Dr. Mohsen Farhadloo for their crucial support and guidance in my research. I could not have undertaken this journey without their help. I am also thankful to my advisor, Dr. Suchit Ahuja, for their support and for being an internal examiner. Sincere thanks to Dr. Danielle Morin for being an internal examiner for my thesis. Finally, I am extremely grateful to my parents for taking care of me and supporting me all the time. Thank you all.

Table of Contents

List of Figures VII
List of Tables VIII
Chapter 1 1
Introduction 1
Chapter 2 5
Literature Review 5
2.1 General overview of Bankruptcy 5
2.2 Bankruptcy with imbalance data 6
Chapter 3 11
Methodology 11
Figure 1. Flowchart of the experiments 12
3.1 Machine Learning Models 12
3.1.1 Logistic Regression 13
3.1.2 XGBoost 14
3.1.3 Decision Trees 14
3.1.4 Random forest 15
3.1.5 LightGBM Classifier 16
3.1.6 CatBoost Classifier 17
3.2 Evaluation Metrics 17
Chapter 4 19
Dataset 19
Table 1. Description of the Polish dataset 20
Table 2. Description of each attribute in the dataset 20
Chapter 5 25
Results 25
Table 3. Summary of obtained accuracy 25
Table 4. Summary of obtained sensitivity 27
Table 5. Summary of obtained specificity 29
Table 6. Overall averages of performance measures across years. 31
Table 7. Average time taken by each model to run 33
Table 8. Comparison with previous studies validated on the same database 36
Chapter 6 38
Discussions and Implications 38
6.1 Theoretical Implications 39
6.2 Practical & Academic Implications 39
Chapter 7 40
Limitations 40
Chapter 8 41

Conclusions 41
Chapter 9 44
Future Work 44
References 46

List of Figures

Figure 1. Flowchart of the experiments 12

List of Tables

Table 1. Description of the Polish dataset 20

Table 2. Description of each attribute in the dataset 20

Table 3. Summary of obtained accuracy 25

Table 4. Summary of obtained sensitivity 27

Table 5. Summary of obtained specificity 29

Table 6. Overall averages of performance measures across years. 31

Table 7. Average time taken by each model to run 33

Table 8. Comparison with previous studies validated on the same database 35

Chapter 1

Introduction

One of the most crucial jobs in managing financial risk is predicting bankruptcy or default. The ability to assess a company's solvency, its capacity to meet financial obligations, is a critical responsibility in the realm of financial decision-making (Taffler, 1983). The prediction of corporate bankruptcy serves as a linchpin in this process, holding the key to navigating the complex terrain of fiscal stability. The evaluation of a company's financial stability and its ability to meet financial commitments is a pivotal aspect of making sound financial decisions.

Foreseeing corporate bankruptcy holds significant importance in this domain, acting as a cornerstone in navigating the intricate landscape of financial security. Mitigating the risk of bankruptcy not only fosters economic growth but also amplifies profits for financial entities and generates higher revenues for the government. Hence, it becomes imperative for businesses to possess precise and dependable models for assessing financial risks, enabling them to make well-informed choices. This line of work holds relevance across diverse sectors, encompassing banks, insurance firms, investment enterprises, and governmental bodies. It also affects the sustainable growth of business organizations. Corporate sustainability, as told by Artiach et al. (2010), is considered to be a business and investment strategy that seeks to use the best business practices to meet and balance the needs of current and future stakeholders (Artiach et al., 2010).

Financial ratios have historically been used to assess bank performance as presented in Board et Al. (2003), but developments in artificial intelligence (A.I.) and operational research (O.R.) have led to a move toward these quantitative methods (Board et al., 2003). Naturally, this is not shocking given that O.R. has been widely employed in other financial applications over the past 50 years. By employing data mining models to forecast financial risks, these entities can enhance investment strategies, optimize fund allocation, and steer clear of economic downturns. For the economy as a whole, reducing bankruptcy risk is critical for promoting economic growth and stability. When companies go bankrupt, they often lay off workers, which can lead to higher unemployment rates and reduced consumer spending. This, in turn, can lead to a decline in economic growth and stability. By accurately predicting bankruptcy risk, organizations can take steps to

prevent financial crises and promote economic growth. Furthermore, this research bears implications for policymaking.

Forecasting bankruptcy holds substantial implications for market stability. Employing predictive techniques in bankruptcy analysis contributes to averting potential financial crises and fostering sustainability within markets. This drive towards bankruptcy prediction stems from a sense of duty among auditors, creditors, and stakeholders vested in securing their enterprises' future. High business failure rates can have disastrous effects on partners, society, the nation's economy as a whole, and business owners themselves (Li et al., 2009). It is therefore reasonable to conduct a great deal of research into creating bankruptcy prediction models (BPM) for businesses. The choice of tool used in its construction has a significant impact on the model's performance, among other things (Alaka et al., 2016).

The previous several decades have seen a rise in financial market globalization, rivalry between businesses, financial institutions, and organizations, as well as quick changes in the economy, society, and technology. All of these factors have contributed to a more uncertain and unstable business and financial climate. As a result, there is now a greater need than ever for the deployment of efficient techniques for gauging the financial health of enterprises and for a thorough approach to problem-solving. In this "new" world, scientists and financial professionals alike understand that the challenges of forecasting a company's financial health must be addressed with comprehensive, practical solutions built upon advanced quantitative analysis methodologies. The relationship that develops between mathematical programming and financial theory is highly significant (Horváthová et al., 2023).

The dataset employed in this study spans 14 years, assessing bankruptcy likelihood in Polish companies. It consists of 64 attributes and a class column indicating bankruptcy status. Over time, instances of bankruptcy gradually increased, showing a rise in the number of bankrupt companies across five years. Government entities and investment managers overseeing financial institutions, banking sectors, manufacturing industries, etc., benefit from informed decision-making to safeguard their enterprises against failure and economic downturns. Consequently, the relevance of bankruptcy prediction continues to grow, leading researchers to propose diverse financial indicators like net revenue, net profit, and liabilities. Despite its advantages, misclassifying bankruptcy predictions can incur costs surpassing actual bankruptcy expenses.

Although filing for bankruptcy allows companies to reboot, it can severely hamper their future prospects, hindering access to future loans. Thus, researchers persistently strive to minimize error probabilities in bankruptcy prediction. Governments can leverage these models to oversee financial institutions, ensuring their operations adhere to acceptable risk thresholds. This proactive approach aids in averting financial crises and fostering sustained economic development. However, it is not merely a routine task; rather, it constitutes one of the most formidable challenges in the field of financial risk management. The dynamics of bankruptcy prediction present a classic example of an imbalanced classification problem. The crux of the challenge lies in the disproportionate distribution between instances of defaults or bankruptcies and those of non-defaults or non-bankruptcies.

This imbalance adds a layer of complexity to predictive analytics, requiring a nuanced approach to ensure accuracy and reliability. Numerous studies have delved into the intricate web of bankruptcy prediction, recognizing its significance in safeguarding financial health. The uneven nature of the classification problem underscores the need for sophisticated methodologies and models. As financial markets continue to evolve, the exploration of diverse avenues in bankruptcy prediction remains a focal point, with researchers and practitioners striving to enhance the efficacy of predictive tools and methodologies. In essence, the quest to foresee financial distress and insolvency is an ongoing journey marked by continuous refinement and adaptation to the ever-changing landscape of financial dynamics (Edum-Fotwe et al., 1996).

Our research methodology focuses on employing oversampling and undersampling approaches to address uneven financial data in our study methodology. Using a 10-fold cross-validation approach, we assess six models that were selected based on their simplicity and efficacy: logistic regression, XGBoost, decision trees, random forest, LightGBM, and CatBoost. This strategy guarantees thorough testing under a range of data settings and offers insightful information about the intricacies of financial risk assessment. By means of this methodical process, we want to make a significant and thorough contribution to the field of financial risk management.

The thesis appeals to a diverse audience, including finance professionals, policymakers, academics, and data scientists, offering insights into financial risk management, predictive models, and their practical applications in various sectors such as banking, investments, and regulatory frameworks.

The remainder of the paper is organized as follows: Section 2 presents a comprehensive Literature Review. Section 3 technically describes selected machine learning models for forecasting and performance measures of accuracy, specificity, and sensitivity. Section 4 introduces our data and provides forecasting results, as well as comparisons of different models' performance and their strengths or weaknesses. Section 5 concludes our main findings and discusses future research directions.

Chapter 2

Literature Review

Since 1968, research in the area of bankruptcy prediction has been conducted. Various classifications of data have been made based on sampling techniques and the overall process of bankruptcy. This study presents the literature review between general bankruptcy concerns and those that specifically address the imbalance data in bankruptcy cases in the two sections that are presented below.

2.1 General overview of Bankruptcy

One of the first approaches used discriminant analysis mode Springate (1978) which was a powerful tool for predicting the possibility of failure in Canadian and U.S. firms (Springate, 1978). The model was based on a sample of 120 firms, of which 60 failed and 60 did not fail. The model was able to correctly classify 85% of the firms in the sample. A comparative study between various models Boritz et Al (2007) including Springate (1978) (Boritz et al., 2007)

In Chen (2011), the authors used principal component analysis for features dimension reduction and decision trees and logistic regression for bankruptcy prediction in Taiwan. They found that the decision trees achieve a better prediction accuracy than the logistic regression (M.-Y. Chen, 2011). In Fedorova (2013), the authors predicted the bankruptcy of Russian companies by making use of various classifiers and designed an improved AdaBoost, named AsymBoost that introduced a cost-sensitive learning conception into the boosting framework to transform the optimization objective function from maximizing the predictive accuracy to minimizing the total misclassification cost. The model achieved an impressive accuracy rate of 86.1%, correctly classifying 87% of the bankrupt companies and 84% of the healthy companies (Fedorova et al., 2013).

Another notable study by Acharjya and Rathi (2021b) involved feature selection through principal component analysis (PCA) to reduce the dimensionality of the dataset. The authors then conducted a comparative analysis, pitting various methodologies such as statistical, rough computing, and mixed computing approaches against each other (Acharjya & Rathi, 2021). They utilized rough set (RS), rough set hybridization with neural network (RSNN), rough set hybridization with binary coded genetic algorithm

(RSBCGA), rough set hybridization with real-coded genetic algorithm (RSRCGA), and fuzzy-rough hybridization with real-coded genetic algorithm (FRSRCGA) to process the data. The results showed that RSBCGA outperformed other methods in terms of accuracy.

2.2 Bankruptcy with imbalance data

The problem of class inequality is difficult to undertake in the context of financial risk evaluation and management. In this regard, few studies have examined how well-unbalanced models predict financial risk and examined the performance of various methods under imbalanced datasets. For instance, the authors in Altman et Al (1968) proposed a methodology based on the resampling technique of the credit score datasets in accordance with their imbalance ratio and a predetermined threshold(Altman, 1968). They extended the balance cascade approach Shultz et Al (1991) to generate adjustable balanced subsets based on the imbalance ratios of the training data(Shultz & Schmidt, 1991). The proposed model uses an ensemble of classifiers, each trained on a subset of the balanced data, and then combined to make the final prediction. The results showed that the proposed model outperformed the existing methods in terms of accuracy and F1 score.

The experimental results showed that their proposed model achieved an accuracy rate of 81.3% and an area under the receiver operating curve (ROC) curve (AUC) of 0.847. Besides, eight different sample techniques were used in Van et Al (2007) (Van Hulse et al., 2007). It was demonstrated that the random sampling method outperforms intelligent sampling methods like SMOTE. The impact of the imbalance ratio on classifier outcomes on various resampling techniques was examined by Garca et Al (2009) which showed that evolutionary undersampling outperforms the nonevolutionary models when the degree of imbalance is increased (García & Herrera, 2009). To balance the initial unbalanced credit datasets, Crone and Finlay (2012) used both over- and under-sampling techniques (Crone & Finlay, 2012). Based on unbalanced credit scoring data sets, Brown and Mues (2012) developed experimental comparisons with a number of methodologies and found that random forest and gradient-boosting classifiers perform very well in a credit scoring context (Brown & Mues, 2012). As a result of the findings, it was concluded that random forest and gradient-boosting classifiers function well in a credit-scoring scenario with observable class imbalances.

To deal with unbalanced data in bankruptcy prediction, the authors in Sun et Al (2014) used a synthetic minority over-sampling technique (SMOTE) to balance the

distribution of the company failure dataset. In another study, to create minority class samples for Chinese tourism business failure alerts, the authors Li et al. (2014b) (Li et al., 2014) presented a k-nearest neighbors (kNN) algorithm based on an oversampling technique Chen et Al (2019) (Cheng et al., 2019).

A cost-sensitive SVM was proposed in Kim (2014) work which assigned a higher weight to the misclassified majority class samples, thus modifying the decision boundary and the optimal classification hyperplane (Kim & Upneja, 2014). On the imbalanced datasets used in the experiments, the proposed cost-sensitive SVM method achieved an average accuracy of 91.52%, which was significantly higher than the accuracies achieved by the other methods.

Zięba et al. (2016) classified firms using an extension of Extreme Gradient Boosting and said that the findings were applicable to all data in this area of study (Zięba et al., 2016). Additionally, they created a brand-new technique known as synthetic features to guarantee that data appropriately depicts higher-order statistics. The authors in Sun et Al (2018) developed an ensemble model for imbalanced credit evaluation based on the SMOTE algorithm and the bagging technique with various sample rates (Sun et al., 2018). The SMOTE algorithm is used to oversample the minority class, while the bagging technique is used to reduce the overfitting of the model. The proposed model achieved better results than other methods like DT, over-sampling DT, over-under-sampling DT, SMOTE DT, etc. in terms of accuracy, precision, recall, and F1 score.

In Veganzones et Al (2018), the authors examined the degree of imbalance, loss of performance, and sampling techniques under Spanish unbalanced bankruptcy data (Veganzones & Séverin, 2018). It was found that the prediction loss rises with the imbalanced proportion. In addition, the support vector machine method is less influenced by imbalanced datasets than linear discriminant analysis, logistic regression, random forests, and neural networks. Furthermore, the SMOTE performs better than other sampling techniques (for instance, random oversampling, random undersampling, and easy ensemble) for all types of prediction models and different training set sizes.

In Garcia et al. (2019), six different bankruptcy prediction models were tested on a dataset of Brazilian firms with different class distribution ratios. The models included logistic regression, support vector machines, random forests, neural networks, gradient boosting machines, and k-nearest neighbors (V. García et al., 2019). The evaluation was performed using various performance metrics, including accuracy, precision, recall, F1-

score, and the area under the ROC curve. The results showed that the performance of the models varied depending on the class distribution ratio in the dataset. In particular, the random forest and gradient boosting models performed better when the dataset was more balanced, while the logistic regression and k-nearest neighbors models performed better when the dataset was more unbalanced.

In recent research, various techniques have been employed to improve the performance of classification models on imbalanced datasets. One such approach, introduced by Quynh et al. (2020), is the synthetic minority over-sampling technique (SMOTE). This technique has been used to balance class distribution in datasets, and its successful application led to an impressive accuracy score of 99.52% (Quynh & Phuong, 2020). Smiti and Soui (2020) proposed a Borderline Synthetic Minority oversampling technique with a Stacked auto-encoder (BSM-SAES) approach to balance the Polish data sets and decrease the dimensionality of variables for predicting bankruptcy(Smiti & Soui, 2020).

Three separate financial data sets with various types of attributes were used by Lahmiri et al. (2020). These included a quantitative data set that was comparable to the data from Polish companies in their first year, a qualitative data set, and a credit scoring data set that combined quantitative and qualitative data (Lahmiri et al., 2020). On such data sets, ensemble classifiers such as AdaBoost, LogitBoost, RUSBoost, subspace, and Bagging were used. The outcomes demonstrated that AdaBoost outperformed other ensemble financial classification techniques for this Polish data set, with the lowest error of 0.0532; RUSBoost, LogitBoost, Bagging, and Subspace had the lowest errors in that order.

In a similar vein, Keya et al. (2021) adopted a combination of feature extraction and recursive feature elimination (RFE) to select the most informative features from a pool of 64. They also leveraged SMOTE to address the data imbalance issue (Keya et al., 2021). The models they employed, including AdaBoost, decision tree, random forest, J48, and bagging, achieved an accuracy of 97%. Aljawazneh et al. (2021) assessed the effectiveness of five ensemble techniques and a variety of DL models using three distinct data sets, including polish data. Following the use of the oversampling approach, the authors employed the XGBoost, RF, SVM, K_NN, Deep belief network (DBN), Long-short term memory (LSTM), and Multilayer Perceptron With 6 Layers (MLP 6L)(Aljawazneh et al., 2021).

Zahiri (2022) used SMOTE in order to deal with the imbalance in the dataset. Apart from trying out various preprocessing and balancing techniques, they used complex neural networks like convolutional neural networks and artificial neural networks in order to deal with such dataset and performed exceptionally well especially in terms of accuracy and sensitivity (Zahiri, 2022).

In this study, we aim to conduct a comparative analysis of the performance of several classifiers on the original dataset and various resampled datasets. Both random oversampling and random undersampling are considered in our analysis. On all three different datasets, we compare the accuracy, specificity, and sensitivity of six different classifiers, including logistic regression, random forest, decision trees, XGBoost, LightGBM, and CatBoost. Logistic regression, random forest, and decision trees are commonly used in previous studies. However, examining and comparing the effectiveness of Gradient Boosting models like XGBoost, LightGBM, and CatBoost is missing. Hence, our contributions follow. First, we examine the performance of various ensemble learning models in the context of bankruptcy prediction under various sampling schemes. Second, we compare their effectiveness to that of standard classification models, namely, the logistic regression and decision trees.

Our study fills a significant vacuum in the literature by evaluating machine learning models for bankruptcy prediction using a variety of sampling techniques. In particular, we go further into previous research by looking at a wider range of models and comparing their performance over various years and sampling techniques, all the while utilizing the original unbalanced data as a point of reference. Furthermore, we advance this field by introducing undersampling approaches, an approach that was not as thoroughly investigated in earlier studies that mostly concentrated on oversampling. With the use of this thorough method, we can investigate and contrast the effectiveness of undersampling and oversampling techniques for reducing data imbalance and improving the accuracy of bankruptcy predictions.

Chapter 3

Methodology

As the dataset is highly imbalanced, we use random oversampling and random under-sampling. Our approach incorporates a robust 10-fold cross-validation protocol, ensuring the reliability of our results by systematically rotating through subsets of the data during training and testing.

The overall experimental architecture is illustrated in Figure 1, outlining a comprehensive framework for our investigations. Within this framework, six distinct machine learning models are harnessed to discern between bankrupt and non-bankrupt companies. These models encompass logistic regression, XGBoost, decision trees, random forest, LightGBM, and CatBoost. Their selection is rooted in their simplicity and efficiency, requiring minimal parameter tuning, which aligns with common practices in the literature.

The utilization of these classifiers is underlined by their widespread applicability and efficacy in handling imbalanced datasets. In the subsequent subsections, we provide concise details on each of these classifiers. The systematic integration of resampling techniques, cross-validation, and a diverse set of classifiers encapsulates a holistic strategy aimed at robustly addressing the intricacies of imbalanced datasets and enhancing the generalizability of our findings.

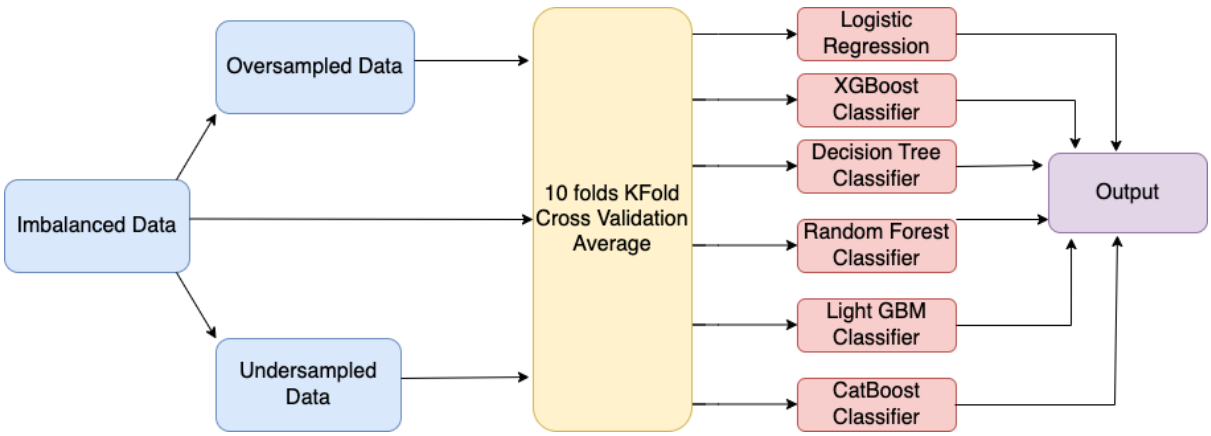

Figure 1. Flowchart of the experiments

Our research methodology unfolds in a systematic fashion, as visualized in Figure 1. The essence of our approach lies in addressing the intricacies of highly imbalanced financial data and meticulously assessing the performance of six chosen models.

The first step involves the collection of data spanning multiple years, with each year's dataset exhibiting significant imbalance due to the rarity of default and bankruptcy events. We drop all the null values from our dataset. To mitigate this imbalance, we apply both oversampling and undersampling techniques to each year's data, resulting in three distinct sets: the original dataset, the undersampled dataset, and the oversampled dataset. This step is pivotal in ensuring that our models are tested under varying data conditions, capturing the nuances of financial risk patterns.

To gauge the reliability and consistency of our results, we employ a 10-fold cross-validation strategy. Each of the three dataset types for each year is further divided into ten subsets, ensuring that our models are rigorously tested and evaluated across a diverse array of data samples.

The crux of our evaluation lies in the performance of the six selected models: logistic regression, XGBoost classifier, decision tree classifier, Random Forest classifier, LightGBM classifier, and CatBoost classifier. Each model, trained independently, is assessed based on three critical evaluation metrics: accuracy, sensitivity, and specificity. This multifaceted evaluation approach enables us to comprehensively understand how each model performs in the intricate task of identifying financial risk within imbalanced datasets.

By meticulously adhering to this systematic workflow, we aim to provide valuable insights into the comparative performance of these models, shedding light on the complexities of financial risk assessment in the face of highly imbalanced data. The robustness of our approach ensures that our findings are both reliable and comprehensive, contributing to the ongoing discourse surrounding financial risk management.

3.1 Machine Learning Models

There were various machine learning algorithms and models that were used in our research namely, Logistic Regression, XGBoost, Decision Trees, random forest,

LightGBM, and Catboost. Here is a brief discussion and the mathematical approach behind each algorithm in a bit of detail:

3.1.1 Logistic Regression

Logistic Regression Stoltzfus et Al (2011) is a classification method used in machine learning to model the dependent variable, and a logistic (sigmoid) function is used(Stoltzfus, 2011). There are only two viable classes because the dependent variable is dichotomous in nature. In logistic regression, the main modeling assumption is that the function $a(x)$ is linear in x. In particular, in logistic regression, we have :

$$y(x) = p(C1\,|x) = \sigma(w\,\top x) \tag{2}$$

where w is the weight vector for the linear model and $\top$ denotes the transpose of a matrix (vector).

The connection is nonlinear overall because of the activation function $\sigma(.)$, despite the linear assumption first appearing restricted. Moreover, logistic regression works quite robustly, as seen by its application in several disciplines. Since the feature vector in this model is M dimensional, there will be M parameters corresponding to w = ($w1$, $z2$, …, xM). In other words, the number of parameters in this model matches the number of features. It goes without saying that as features grow, so will the number of model parameters. Using basis functions or other reduction strategies is one way to lower the number of parameters.

Logistic Regression offers interpretability, facilitating the understanding of how predictor variables impact the likelihood of a specific outcome. It is particularly efficient with smaller datasets and is well-suited for constrained data availability. However, it does have limitations, including the assumption of linear relationships between predictors and the log odds of the outcome, making it less suitable for capturing complex nonlinear associations. Additionally, it can be susceptible to overfitting in cases with numerous predictors and has an assumption of predictor independence, which may not hold in practical contexts. Outliers in the data can also exert a notable influence on its performance and parameter estimates.

3.1.2 XGBoost

Extreme gradient boosting Chen et Al (2016), often known as XGBoost classifier, is a gradient tree-boosting technique. It is an ensemble additive model that chooses the function that minimizes the loss from among numerous base learners or functions(T. Chen & Guestrin, 2016). A tree is eagerly constructed using XGBoost, and the split that minimizes loss the most is selected. Here's how the XGBoost algorithm decides the best loss function:

$$L_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \quad (2)$$

Here, Lsplit refers to the loss function of the current split, which is the sum of the losses of the left and right splits, ILis the set of instances of the left split, IRis the set of instances of the right split, gi is the gradient of the loss function with respect to the predicted value of the *i*th instance, hi is the second derivative of the loss function with respect to the predicted value of the *i*th instance, is the regularization parameter and is the minimum loss reduction required to make a further partition on a leaf node of the tree.

Advantages of XGBoost include its high predictive accuracy, efficient handling of missing data, incorporation of regularization techniques to mitigate overfitting, and support for parallel processing, leading to faster computation times. However, it is associated with complexity, particularly for beginners, can be computationally intensive, demands thorough hyperparameter tuning, and may pose challenges in terms of model interpretability due to its advanced features and complexity.

3.1.3 Decision Trees

The decision tree classifier Breiman et Al (2017) creates the classification model by building a decision tree(Breiman, 2017). A test on an attribute is specified by each node in the tree, and each branch descending from that node represents one of the possible values for that attribute. A decision tree is mainly made using the Gini Index and Entropy. The Gini Index is calculated as follows:

$$GiniIndex = 1 - \sum_{i=1}^{n} p_i^2 \quad (3)$$

where p refers to the probabilities of each class. The Gini Index measures the probability of misclassifying a randomly chosen sample from the dataset, and it is minimized when all samples in a node belong to the same class.

Decision trees approach is particularly lauded for its interpretability, allowing users to comprehend the decision-making process. Decision trees are versatile in capturing both linear and nonlinear relationships in the data and are computationally efficient for prediction, especially with smaller datasets. They also serve as useful feature selection tools by virtue of the tree structure. However, they are susceptible to overfitting, especially when the trees become complex, necessitating pruning to achieve optimal generalization. Decision trees may exhibit instability in response to minor variations in the data, and in imbalanced datasets, they can display a bias toward dominant classes. Their limited ability to represent intricate decision boundaries is another constraint (García & Herrera, 2009).

3.1.4 Random forest

Each decision tree in the ensemble that makes up the random forest classifier Breiman et Al (2001) is composed of a data sample taken from a training set with a replacement known as the bootstrap sample(Breiman, 2001). It is a meta-estimator that applies multiple decision tree classifiers to different dataset sub-samples and averages the outcomes. Calculating the relevance of each attribute for each tree, then dividing that total by the number of trees yields:

$$RFfi_i = \frac{\sum_{j \in alltrees} normfi_{ij}}{T} \quad (4)$$

$RFfi_i$= the importance of feature i calculated from all trees in the random forest model, nor $normfi_{ij}$= the normalized feature importance for i in tree j and T = Total number of trees.

Random Forest offers several merits, including high predictive accuracy, robustness to outliers and noise in the data, efficient handling of large datasets, and the ability to identify feature importance, aiding in feature selection and interpretability. Despite its numerous strengths, random forest has some limitations, such as potential overfitting, computational resource requirements, and reduced transparency compared to individual decision trees. Interpretability can be compromised when dealing with a

large number of trees in the ensemble. In summary, random forest is a formidable ensemble method that excels in predictive performance and robustness.

3.1.5 LightGBM Classifier

The LightGBM classifier Ke et Al (2017) is a quickly distributed high-performance gradient-boosting framework used for many different machine learning applications, including classification and ranking(Ke et al., 2017). While other tree-based learning algorithms grow trees horizontally, LightGBM develops trees vertically. In contrast to other algorithms, LightGBM grows like a tree from the leaves up. The leaf with the greatest delta loss will be chosen to grow. When expanding the same leaf, the leaf-wise algorithm can reduce loss more than a level-wise algorithm.

$$\underline{V_j}(d) = \frac{1}{n}\left(\frac{\left(\sum_{x_i \in A_l} g_i + \frac{1-a}{b}\sum_{x_i \in B_l} g_i\right)^2}{n_l^j(d)} + \frac{\left(\sum_{x_i \in A_r} g_i + \frac{1-a}{b}\sum_{x_i \in B_r} g_i\right)^2}{n_r^j(d)}\right) \quad (5)$$

where $\underline{V_j}(d)$ refers to the score of node j when splitting on feature d, x_i refers to the number of samples in the dataset that are assigned to node i, A_l refers to the sum of the target values for the samples assigned to node i, B_l refers to the sum of the squared target values for the samples assigned to node i, A_r refers to the sum of the gradient values for the samples assigned to node i, B_r refers to the sum of the squared gradient values for the samples assigned to node i, and n refers to the total number of instances. LightGBM natively accommodates categorical features, reducing the need for extensive pre-processing. It also excels at feature selection, quantifying the significance of predictors. It is sensitive to outliers in the data and it is necessary for precise hyperparameter tuning for optimal performance.

3.1.6 CatBoost Classifier

The CatBoost classifier Prokhorenkova et Al. (2018) is based on gradient-boosted decision trees. A series of decision trees are built sequentially during training(Prokhorenkova et al., 2018). In comparison to the preceding trees, each subsequent tree is constructed with less loss.

$$h(x) = \sum_{j=1}^{J} b_j 1_{\{x \in R_j\}} \quad (6)$$

where R_j are the disjoint regions corresponding to the leaves of the tree, x is the input variable, which can be a scalar or a vector, b_j is the constant value assigned to x in the jth interval. This is also known as the weight or coefficient of the interval.

CatBoost's ability to efficiently handle categorical features without the need for one-hot encoding offers a significant advantage, simplifying the preprocessing phase and reducing the risk of dimensionality explosion. It further distinguishes itself by its innate capacity to handle missing data, enhancing the algorithm's resilience when confronted with real-world, noisy datasets. It is essential to consider that CatBoost's inherently deep model structure may compromise the interpretability of the model in complex ensemble scenarios. Also, its computational demands should not be underestimated.

3.2 Evaluation Metrics

We evaluate all classification models based on accuracy (correct classification rate), sensitivity, and specificity which are widely used in classification problems. They are expressed as follows:

$$Accuracy = \frac{Classified\ Samples}{Total\ Number\ of\ Samples} \tag{7}$$

$$Sensitivity = \frac{Correctly\ Classified\ Positive\ Samples}{True\ Positive\ Samples} \tag{8}$$

$$Specificity = \frac{Correctly\ Classified\ Negative\ Samples}{True\ Negative\ Samples} \tag{9}$$

We implement the 10-fold cross-validation protocol to ensure the robustness of our results. This involves dividing the dataset into ten subsets, training the model on nine of them, and testing on the remaining one in a systematic rotation. This process is repeated ten times, each time using a different subset as the test set.

To gauge the impact of resampling techniques on the distribution of scores, we calculate the average and standard deviation of each performance metric across these ten folds. This statistical analysis provides a comprehensive view of how our model performs under various conditions, considering the fluctuations introduced by different subsets. The average gives us a central tendency measure, offering insight into the typical

performance, while the standard deviation quantifies the degree of variation or consistency in our results.

By examining both the average and standard deviation, we gain a nuanced understanding of the stability and reliability of our model. A lower standard deviation indicates less variability in the performance metrics, suggesting a more consistent and robust model across different subsets. This meticulous evaluation process ensures that our results are not skewed by the particularities of a single subset, enhancing the generalizability and validity of our findings.

Chapter 4

Dataset

The dataset, sourced from the Emerging Markets Information Service database, examines the likelihood of Polish businesses declaring bankruptcy (Tomczak, 2016). The data covers the period from 2000 to 2013, focusing on companies that filed for bankruptcy between 2000 and 2012 and those that remained operational from 2007 to 2013. There are 64 attributes labeled X1 to X64 a, with a class column indicating bankruptcy status: '0' signifies a company that did not file for bankruptcy, while '1' indicates one that did. Table 1 summarizes the dataset, while Table 2 summarizes all the columns used in the dataset and their information.

Five categorization cases are defined by the dataset according to various forecast periods. Using financial rates from the first forecasting period, 7,027 cases were examined in the first year; 271 of those enterprises declared bankruptcy, while 6,756 remained solvent. This makes up 271 out of 7027, or roughly 3.85% of all the companies mentioned.

A total of 10173 instances were analyzed in the second year, which focused on financial rates from the second year of the projection period. Of these, 400 companies filed for bankruptcy, and 9,773 companies remained in existence. This amounts to around 400 out of 10173, or 3.93%, of all the companies that filed for bankruptcy during this time frame. Taking into account the financial rates from the third year of the forecasted period, 10503 cases were noted in the third year; 495 of these companies declared bankruptcy, leaving 10008 businesses still in business. In this time period, 495 out of 10503 enterprises, or around 4.71% of all corporations, declared bankruptcy.

A total of 9,792 cases were evaluated in the fourth year of the forecasted period based on an examination of financial rates; 515 firms were declared bankrupt, and 9,277 did not file for bankruptcy. This means that 515 out of 9792 firms, or around 5.26% of all enterprises, will file for bankruptcy during this time. Using financial rates from the fifth year of the forecasted period, 5,910 instances were assessed in the fifth year; 410 of these enterprises declared bankruptcy, while 5,500 continued to operate. This accounts for

approximately 6.94% of the total companies (410 out of 5910) that filed for bankruptcy during this period.

Table 1. Description of the Polish dataset

Year	Total Companies	Bankrupted	Non-Bankrupted
1	7027	271	6756
2	10173	400	9773
3	10503	495	10008
4	9792	515	9277
5	5910	410	5500

Table 2. Description of each attribute in the dataset

Attribute	Description
X1	Net profit / Total assets
X2	Total liabilities / Total assets
X3	Working capital / Total assets
X4	Current assets / Short-term liabilities
X5	[(Cash + Short-term securities + Receivables - Short-term liabilities) / (Operating expenses - Depreciation)] * 365
X6	Retained earnings / Total assets

X7	EBIT / Total assets
X8	Book value of equity / Total liabilities
X9	Sales / Total assets
X10	Equity / Total assets
X11	(Gross profit + Extraordinary items + Financial expenses) / Total assets
X12	Gross profit / Short-term liabilities
X13	(Gross profit + Depreciation) / Sales
X14	(Gross profit + Interest) / Total assets
X15	(Total liabilities * 365) / (Gross profit + Depreciation)
X16	(Gross profit + Depreciation) / Total liabilities
X17	Total assets / Total liabilities
X18	Gross profit / Total assets
X19	Gross profit / Sales
X20	(Inventory * 365) / Sales
X21	Sales (n) / Sales (n-1)
X22	Profit on operating activities / Total assets
X23	Net profit / Sales
X24	Gross profit (in 3 years) / Total assets
X25	(Equity - Share capital) / Total assets

X26	(Net profit + Depreciation) / Total liabilities
X27	Profit on operating activities / Financial expenses
X28	Working capital / Fixed assets
X29	Logarithm of total assets
X30	(Total liabilities - Cash) / Sales
X31	(Gross profit + Interest) / Sales
X32	(Current liabilities * 365) / Cost of products sold
X33	Operating expenses / Short-term liabilities
X34	Operating expenses / Total liabilities
X35	Profit on sales / Total assets
X36	Total sales / Total assets
X37	(Current assets - Inventories) / Long-term liabilities
X38	Constant capital / Total assets
X39	Profit on sales / Sales
X40	(Current assets - Inventory - Receivables) / Short-term liabilities
X41	Total liabilities / ((Profit on operating activities + Depreciation) * (12/365))
X42	Profit on operating activities / Sales
X43	Rotation receivables + Inventory turnover in days

X44	(Receivables * 365) / Sales
X45	Net profit / Inventory
X46	(Current assets - Inventory) / Short-term liabilities
X47	(Inventory * 365) / Cost of products sold
X48	EBITDA (Profit on operating activities - Depreciation) / Total assets
X49	EBITDA (Profit on operating activities - Depreciation) / Sales
X50	Current assets / Total liabilities
X51	Short-term liabilities / Total assets
X52	(Short-term liabilities * 365) / Cost of products sold
X53	Equity / Fixed assets
X54	Constant capital / Fixed assets
X55	Working capital
X56	(Sales - Cost of products sold) / Sales
X57	(Current assets - Inventory - Short-term liabilities) / (Sales - Gross profit - Depreciation)
X58	Total costs / Total sales
X59	Long-term liabilities / Equity
X60	Sales / Inventory

X61	Sales / Receivables
X62	(Short-term liabilities * 365) / Sales
X63	Sales / Short-term liabilities
X64	Sales / Fixed assets

From our above observations, one can derive that there needs to be some solution to this massive difference between counts of companies that go bankrupt and those not which may heavily affect how the model reacts to it. Then, we can either make use of over-sampling or under-sampling in order to equally balance the dataset and make use of machine learning models on it.

This study explores the use of six different classifiers consisting of logistic regression, XGBoost classifier, decision tree classifier, random forest classifier, LightGBM classifier, and CatBoost classifier over three differently sampled datasets over a period of five years. The results were calculated over 10 folds and here as result, the average of all folds has been shown.

Chapter 5

Results

As we can see from Tables 3, 4, 5 and 6 the accuracy has been the highest for the original dataset, then for the undersampled dataset for the first year.. The standard deviation for accuracy is relatively less than the oversampled and undersampled datasets. The sensitivity score is quite high for the undersampled dataset which means that the undersampled data set correctly classifies the companies that went bankrupt as compared to the other two datasets for each of the classifiers. As for the classifiers, the XBGClassifier performed the best for the original dataset. As for the oversampled dataset, logistic regression performed well for accuracy and sensitivity but lacked predicting specificity. In the undersampled dataset, CatBoost seemed to perform the best. Thus, gradient boosting methods seem to outperform other traditional machine learning algorithms.

Table 3. Summary of obtained accuracy

Models	**Year 1**	**Year 2**	**Year 3**	**Year 4**	**Year 5**
Original dataset					
Logistic	0.9897±0.0064	0.9805±0.1039	0.9772±0.0098	0.9737±0.0006	0.9641±0.0084
XGB	0.9906±0.0063	0.9810±0.0908	0.9788±0.0091	0.9739±0.0062	0.9673±0.0109
Decision tree	0.9819±0.0095	0.9634±0.0938	0.9605±0.0088	0.9505±0.0091	0.9502±0.0075
Random forest	0.9888±0.0077	0.9794±0.1042	0.9753±0.0086	0.9723±0.0081	0.9674±0.0091
LGBM	0.9891±0.	**0.9835±0**	**0.9809±**	**0.9765±0.**	**0.9706±0**

	0072	**.0719**	**0.0081**	**0074**	**.0106**
CatBoost	**0.9897±0 .0066**	0.9827±0. 0650	0.9799±0. 0088	0.9754±0.0 080	0.9700±0. 0104
Over-sampled dataset					
Logistic	**0.6314±0 .1502**	**0.6282±0 .1870**	0.5699±0. 0855	**0.6351±0. 0721**	**0.7064±0 .0753**
XGB	0.5627±0. 2271	0.5694±0. 1387	**0.6410± 0.1168**	0.5874±0.1 141	0.6493±0. 1424
Decision tree	0.5750±0. 2163	0.5723±0. 1179	0.6121±0. 1174	0.5494±0.1 040	0.5982±0. 0678
Random forest	0.5633±0. 2413	0.5422±0. 0999	0.5359±0. 1360	0.5211±0.1 008	0.5958±0. 0912
LGBM	0.5614±0. 2700	0.5676±0. 9046	0.6410±0. 1217	0.5580±0.1 175	0.6430±0. 0135
CatBoost	0.5631±0. 2415	0.5726±0. 1033	0.6349±0. 1218	0.5620±0.1 146	0.6533±0. `1266
Under-sampled dataset					
Logistic	0.6359±0. 1727	0.5769±0. 1870	0.5695±0. 0860	0.6436±0.1 131	0.6874±0. 1069
XGB	0.8653±0. 2052	0.720±0.1 387	0.7302±0. 0910	0.7007±0.0 757	0.8022±0. 0719
Decision tree	0.7971±0. 2141	**0.7245±0 .1179**	0.6932±0. 1083	0.662±0.09 76	0.8057±0. 1350
Random forest	0.8389±0. 2224	0.7241±0. 0999	0.7482±0. 08884	0.6940±0.0 623	0.7860±0. 1100

LGBM	0.8365±0.2072	0.7175±0.0946	0.7387±0.1041	**0.7366±0.0689**	**0.8201±0.0077**
CatBoost	**0.9097±0.1674**	0.7225±0.1033	**0.7625±0.0784**	0.7236±0.0751	0.8076±0.0644

As we can see from Tables 3, 4, 5 and 6 that the accuracy has been the highest for the original dataset, then for the undersampled dataset for the second. The standard deviation for accuracy is relatively less than the oversampled and undersampled datasets. The sensitivity score is quite high for the undersampled dataset which means that the undersampled data set correctly classifies the companies that got bankrupt as compared to the other two datasets for each of the classifiers. As for the classifiers, XBG Classifier performed the best for the original dataset. As for an oversampled dataset logistic regression performed well for accuracy and sensitivity but lacked predicting specificity. In the undersampled dataset, CatBoost seemed to perform the best. Thus, gradient boosting methods seem to outperform other traditional machine learning algorithms.

Table 4. Summary of obtained sensitivity

Models	**Year 1**	**Year 2**	**Year 3**	**Year 4**	**Year 5**
Original dataset					
Logistic	0	0±0.2421	0	0	0.0503±0.0539
XGB	0	0.01±0.1517	0.0611±0.0849	0.0629±0.0968	0.1919±0.1069
Decision tree	**0.05±0.1067**	0.0683±0.1562	**0.2231±0.1297**	0.1613±0.1018	**0.2940±0.1423**
Random forest	0	0±0.0791	0.01180±0.0236	0.0615±0.1078	0.168±0.1259

LGBM	0	**0.0916±0.1254**	0.2057±0.1349	**0.1437±0.0916**	0.2882±0.1514
CatBoost	0	0.0308±0.1253	0.1011±0.1060	0.08893±0.1051	0.2369±0.1366
Over-sampled dataset					
Logistic	**0.5250±0.3758**	**0.4762±0.1870**	**0.2910±0.2240**	**0.6091±0.1693**	**0.6933±0.1662**
XGB	0 ±0.0480	0.0953±0.1287	0.2466±0.1425	0.1684±0.1584	0.2954±0.2640
Decision tree	0.0283±0.1441	0.1094±0.1179	0.2095±0.1157	0.1008±0.1133	0.1975±0.0683
Random forest	0	0.0395±0.0999	0.02724±0.0923	0.0293±0.0942	0.1811±0.1607
LGBM	0	0.0898±0.0946	0.2576±0.1367	0.1081±0.1545	0.2867±0.2388
CatBoost	0	0.10143± 0.1033	0.2420±0.1750	0.1246±0.1311	0.2940±0.2259
Under-sampled dataset					
Logistic	0.4583±0.4108	0.6758±0.3164	0.6012±0.1031	0.6885±0.1614	0.7996±0.1686
XGB	0.6750±0.4163	0.7383±0.14941	0.7228±0.1545	0.7171±0.0926	0.8035±0.1244
Decision tree	0.5583± 0.4349	0.6516±0.1379	0.6711±0.2321	0.6709±0.0992	0.8190±0.1993
Random forest	0.7166±0.4203	0.7025±0.2731	0.7453±0.1522	0.7084±0.1217	0.8111±0.2205

LGBM	0.7083±0.43588	**0.7716±0.1553**	0.7170±0.1628	0.7568±0.1146	0.7989±0.1072
CatBoost	**0.8000±0.4163**	0.7350±0.1498	**0.7794±0.1610**	**0.7793±0.0811**	**0.8503±0.1308**

As we can see from Tables 3, 4, 5 and 6 that the accuracy has been the highest for the original dataset, then for the undersampled dataset of the third year. The standard deviation for accuracy is relatively less than the oversampled and undersampled datasets. The sensitivity score is quite high for the undersampled dataset which means that the undersampled data set correctly classifies the companies that went bankrupt as compared to the other two datasets for each of the classifiers. As for the classifiers, the XBG classifier performed the best for the original dataset. As for the oversampled dataset, logistic regression performed well for accuracy and sensitivity but lacked predicting specificity. In the undersampled dataset, CatBoost seemed to perform the best. Thus, gradient boosting methods seem to outperform other traditional machine learning algorithms.

Table 5. Summary of obtained specificity

Models	**Year 1**	**Year 2**	**Year 3**	**Year 4**	**Year 5**
Original dataset					
Logistic	0.9990+0.0014	0.9985±0.05754	0.9991±0.00103	**0.9989±0.0014**	0.9959±0.0044
XGB	**1**	0.9987±0.0029	0.9993±0.0009	0.9978±0.0009	0.9948±0.0044
Decision tree	0.9905±.00729	0.9792±0.00878	0.9771±0.0051	0.9714±0.0072	0.9761±0.0078
Random forest	0.9981±0.0207	0.9973±0.0021	0.9968±0.0017	0.9963±0.0033	0.9950±0.0031

LGBM	0.9984±0.0051	0.9992±0.0025	0.9983±0.0015	0.9982±0.0021	0.9948±0.0058
CatBoost	0.9990±0.0014	**1**	**0.9995±0.0008**	**0.9989±0.0014**	**0.9965±0.0030**
Over-sampled dataset					
Logistic	0.6788±0.8814	0.7695±0.18702	0.8246±0.0636	0.6673±0.0354	0.7227±0.0708
XGB	0.9976±0.0017	0.9984±0.1387	**0.9976±0.0026**	0.9946±0.0021	0.9884±0.0060
Decision tree	0.9881±0.0074	0.9860±0.1179	0.9791±0.0073	0.9802±0.0045	0.9783±0.0098
Random forest	**0.9988±0.0030**	0.9975±0.0999	0.9979±0.0017	**0.9967±0.0032**	**0.9927±0.0075**
LGBM	0.9956±0.0055	**0.9991±0.0946**	**0.9976±0.0032**	0.9940±0.0026	0.9901±0.0063
CatBoost	0.9984±0.0026	0.9988±0.1033	0.9966±0.0029	0.9951±0.0029	0.9880±0.0078
Under-sampled dataset					
Logistic	0.6666±0.2362	0.5527±0.2953	0.5541±0.1918	0.5992±0.1983	0.6103±0.1219
XGB	**0.9166±0.2291**	0.6933±0.1795	0.7228±0.1242	0.6958±0.0949	0.8066±0.1507
Decision tree	0.8833±0.2291	**0.8155±0.24835**	0.6872±0.1324	0.6702±0.1658	0.8039±0.1805
Random forest	0.8333±0.2891	0.7661±0.1719	0.7339±0.1486	0.7009±0.1301	0.7836±0.0978

LGBM	0.8333±0.3167	0.6733±0.18786	**0.7526±0.1750**	**0.7221±0.1243**	**0.8637±0.1621**
CatBoost	0.8833±0.2260	0.7377±0.1757	0.7259±0.1148	0.6986±0.1275	0.7799±0.1108

As we can see from Tables 3, 4, 5 and 6 that the accuracy has been the highest for the original dataset, then for the undersampled dataset for the fourth. The standard deviation for accuracy is relatively less as compared to the oversampled and undersampled datasets. The sensitivity score is quite high for the undersampled dataset which means that the undersampled data set correctly classifies the companies that got bankrupt as compared to the other two datasets for each of the classifiers. As for the classifiers, XBG classifier performed the best for the original dataset. As for oversampled dataset logistic regression performed well for accuracy and sensitivity but lacked predicting specificity. In the undersampled dataset, CatBoost seemed to perform the best. Thus, gradient boosting methods seem to outperform other traditional machine learning algorithms.

Table 6. Overall averages of performance measures across years.

Models	**Average accuracy across years**	**Average sensitivity across years**	**Average specificity across years**
Original dataset			
Logistic	0.97704	0.01006	0.99828
XGB	0.97832	0.06518	0.99812
Decision tree	0.9613	0.15934	0.97886
Random forest	0.97664	0.04826	0.9967
LGBM	0.98012	0.14584	0.99778

CatBoost	0.97954	0.091546	0.99878
Over-sampled dataset			
Logistic	0.6342	0.51892	0.73258
XGB	0.60196	0.16114	0.99532
Decision tree	0.5814	0.1291	0.98234
Random forest	0.55166	0.055428	0.99672
LGBM	0.5942	0.14844	0.99528
CatBoost	0.59718	0.152406	0.99538
Under-sampled dataset			
Logistic	0.62266	0.64468	0.59658
XGB	0.76368	0.73134	0.76702
Decision tree	0.7365	0.67418	0.77202
Random forest	0.75824	0.59346	0.76356
LGBM	0.76988	0.60886	0.769
CatBoost	0.78518	0.62874	0.76508

As we can see from Tables 3, 4, 5 and 6 that the accuracy has been the highest for the original dataset, then for the undersampled dataset, and is least for the oversampled dataset for the fifth year. The standard deviation for accuracy is relatively less than the oversampled and undersampled datasets. The Sensitivity score is quite high for the undersampled dataset which means that the undersampled dataset correctly classifies the companies that went bankrupt as compared to the other two datasets for each of the classifiers. As for the classifiers, the LGBM classifier performed the best for the original dataset. As for the oversampled dataset, logistic regression performed well for accuracy and sensitivity but lacked predicting specificity. In the undersampled dataset, the LGBM

Classifier seemed to perform the best. Thus, gradient boosting methods seem to outperform other traditional machine learning algorithms.

The analysis of five years of data suggests that the LightGBM classifier exhibited superior performance across metrics of accuracy, sensitivity, and specificity when neither oversampling nor undersampling techniques were employed. However, when utilizing oversampled data and evaluating the accuracy, the logistic regression algorithm demonstrated exceptional performance because logistic regression can handle imbalanced data better than tree-based models, which can be sensitive to class imbalance. Conversely, when utilizing under-sampled data, the LightGBM classifier emerged as the top performer in terms of accuracy. Additionally, when assessing sensitivity and specificity, the catBoost classifier emerged as the optimal choice.

Table 7. Average time taken by each model to run

Models	**Time Taken(in seconds)**
Original dataset	
Logistic	1.98
XGB	8.25
Decision tree	2.05
Random forest	5.45
LGBM	6.24
CatBoost	123
Over-sampled dataset	
Logistic	2.17
XGB	12.1
Decision tree	1.28
Random forest	13.1

LGBM	6.91
CatBoost	125
Under-sampled dataset	
Logistic	0.382
XGB	3.51
Decision tree	0.084
Random forest	2.28
LGBM	0.155
CatBoost	25

The above table summarizes the average time taken by each model compiled by averaging the time to run each model over 10 folds on average on a Python notebook. Across the original dataset, the Decision Tree model emerged as the most efficient, executing in a modest 2.05 seconds. Its capability to rapidly process data while maintaining competitive performance positions it favorably. In the oversampled dataset scenario, despite increased computational demands, the Decision Tree model continued to outperform others, clocking in at 1.28 seconds. Its ability to swiftly handle the increased dataset size underscores its efficiency. Meanwhile, in the undersampled dataset context, the Decision Tree model remained exceptional, executing in a mere 0.084 seconds. Its remarkable speed in handling reduced data volumes signifies its efficiency in resource optimization. The Decision Tree model's consistent swift performance across varied datasets highlights its suitability for efficient processing in both balanced and imbalanced dataset contexts.

The performance and efficiency of Logistic Regression were notably prominent in the oversampled dataset, demonstrating both swift computation and remarkable accuracy. This model attributed considerable importance to specific attributes, notably Rotation Receivables + Inventory Turnover in Days, Operating Expenses / Short-term

Liabilities, Sales / Short-term Liabilities, Sales / Receivables, and Sales / Inventory, assigning them significant coefficients within the logistic regression results.

Contrarily, in tree-based models, attributes such as (Receivables * 365) / Sales, Profit on Operating Activities / Financial Expenses, Total Costs / Total Sales, (Current Assets - Inventory) / Short-term Liabilities, and Operating Expenses / Total Liabilities held more significance. It's evident that these features significantly contributed to the performance of tree-based models, contrasting with their lesser impact on the logistic regression model's outcomes. This observation suggests that while certain attributes were pivotal for tree-based models, their importance was relatively diminished within the logistic regression framework.

In comparison in Table 8, the current study employed oversampling and undersampling techniques and tested multiple models: Logistic Regression, XGBoost, Decision Tree, Random Forest, LightGBM, and CatBoost, achieving a notably higher accuracy of 98.97% with a sensitivity of 0.8503 and a specificity of 0.9995. The current study's superiority over past research lies in its significantly higher accuracy rate, reaching 98.97%, which surpasses all other studies listed in the table. Moreover, it showcases a superior specificity score of 0.9995, indicating its exceptional capability to identify true negative instances, which is crucial in bankruptcy prediction to avoid false alarms about companies facing financial distress. While the sensitivity score might not be the highest among all studies, the combined high accuracy and specificity reflect the effectiveness of the models used in this current research for accurate bankruptcy prediction while minimizing false positives.

However, when we compare our own models, particularly the LightGBM classifier, to the approaches mentioned above, we find that our gradient-boosting models consistently outperform them in various scenarios. Our research demonstrates the efficacy of the LightGBM classifier and highlights its potential as a promising solution for handling imbalanced datasets and achieving higher accuracy in classification tasks, along with being extremely fast and reliable.

Table 8. Comparison with previous studies validated on the same database

Study Details	Techniqu-es Used	Models	Best Accuracy Score	Best Sensitivity Score	Best Specificity Score
Zięba et al (2016)	Synthetic features	Extreme Gradient Boosting	95.9%	Not Reported	Not Reported
Lahmiri et al. (2020)	SMOTE	AdaBoost	94.6%	Not Reported	Not Reported
Keya et al. (2021)	Feature extraction, RFE, SMOTE	AdaBoost, Decision Tree, Random Forest, J48, Bagging	97%	0.94	Not Reported
Smiti and Soui (2020)	BSM-SAES	AutoEncoder	97.4%	Not Reported	Not Reported
Zahiri et Al. (2022)	SMOTE	CNN_2D	91.9%	0.45	0.938
This Study	Oversampling and Undersampling	Logistic Regression, XgBoost, Decision Tree, Random Forest, LightGBM, CatBoost	98.97%	0.8503	0.9995

Chapter 6
Discussions and Implications

This research highlights the importance of predicting bankruptcy risk for financial institutions, governments, and the economy as a whole. By using data mining models and oversampling and under-sampling algorithms, organizations can make informed decisions, reduce financial risk, and increase profits and revenues.

Foreseeing corporate bankruptcy stands as a crucial aspect in navigating the complex realm of financial security. Predicting and averting bankruptcy risk not only fosters economic stability but also amplifies profitability for financial institutions while bolstering government revenues. Consequently, businesses necessitate precise and reliable models to gauge financial risks, empowering them to make informed decisions. This pursuit holds relevance across various sectors, spanning banks, insurers, investment firms, and governmental entities (Fulmer et al., 1984). Leveraging data mining models for financial risk forecasting enables entities to refine investment strategies, optimize fund allocation, and steer clear of potential economic pitfalls. The broader economic implications of mitigating bankruptcy risk are profound. When companies face bankruptcy, there's often a domino effect leading to job layoffs, higher unemployment rates, and decreased consumer spending, ultimately hampering economic growth and stability. Accurate bankruptcy predictions empower organizations to take proactive measures, averting financial crises and nurturing economic progress. Additionally, this research bears significance in the realm of policymaking, offering insights that can shape robust policies aiming to prevent financial turmoil and bolster overall economic growth.

This thesis is intended for a wide range of readers, including academics, data science enthusiasts, policymakers, and financial professionals. It offers an in-depth understanding of prediction models, financial risk management, and its applications. The comprehensive examination of bankruptcy prediction models by financial and credit analysts may be used to assess the financial standing of businesses, and investors and investment managers can use this information to improve their approach to making investments. It provides vital insights into efficient bankruptcy prediction models to banking organizations, assisting with risk management and loan approval procedures. Furthermore, by comprehending the macroeconomic ramifications of company bankruptcy, economists and policymakers may promote economic stability. This

extensive thesis contains useful instructional and technical materials for academics, students, data scientists, and other industry practitioners.

6.1 Theoretical Implications

This study makes a substantial theoretical contribution by highlighting the critical function that predictive modeling plays in predicting bankruptcy risks. With the addition of sophisticated oversampling and undersampling algorithms and state-of-the-art data mining models, scholars have a comprehensive framework to investigate and improve prediction approaches. The integration of finance, data science, and economics in an interdisciplinary manner fosters cooperative research and idea exchange. Moreover, the examination of macroeconomic consequences expands the theoretical domain, providing researchers with a chance to investigate the cascading consequences of financial instability across a wider economic spectrum.

6.2 Practical & Academic Implications

In practice, the Decision Tree model's effectiveness has obvious management ramifications, particularly when it comes to situations where datasets are over- or undersampled. The model is a viable option for real-time decision-making because of its capacity to analyze data quickly without sacrificing accuracy, especially in scenarios when computational resources are scarce. This efficiency may be helpful to financial institutions and other businesses that work with massive datasets for rapid risk assessments and strategic decision-making.

On the managerial front, the attribute significance variations across Logistic Regression and tree-based models offer actionable insights. Depending on the particular features each model emphasizes, decision-makers can modify their approach accordingly. For example, in oversampled cases where Logistic Regression performed exceptionally well, it is critical to pay attention to variables like Sales / Short-term Liabilities and Rotation Receivables + Inventory Turnover in Days. As opposed to this, characteristics such as (Receivables * 365) / Sales and Operating Expenses / Total Liabilities are more important for tree-based models. Based on the selected modeling approach, managers may align their focus on key financial indicators with this practical assistance.

Chapter 7

Limitations

Our current study on bankruptcy forecasting, although exhibiting a sturdy technique and yielding significant findings, has some constraints that require meticulous examination. The study's reliance on oversampling and undersampling techniques to resolve class inequalities is one of its main limitations. Although these methods reduce data imbalances, there is a chance that they could oversimplify minority or majority class representations in the model, which could hinder the model's capacity to generalize to real-world situations. This restriction implies that one should exercise caution when interpreting the model's forecasts since they might not fully capture the intricacies of the wider economic environment.

Moreover, the study largely ignores the complex interplay of variables that lead to financial distress in favor of concentrating on certain economic indicators to forecast bankruptcy. This limited breadth might hinder a thorough knowledge of bankruptcy triggers and, as a result, impair the model's prediction accuracy in identifying high-risk enterprises. Furthermore, the study's assessment criteria prioritize accuracy, sensitivity, and specificity over other important metrics like precision, F1-score, and area under the ROC curve. By including these extra measures, the model's performance may be examined in a more detailed and nuanced manner, giving rise to a better knowledge of its advantages and disadvantages.

One major issue concerns the study's retrospective design, which limits its forecasting power to past data and may compromise its relevance to future economic situations. When used to dynamic and changing financial conditions, the models produced under this framework may not sufficiently account for shifting economic landscapes, making their predictions less dependable.

Chapter 8
Conclusions

In the intricate domain of financial risk assessment, the infrequent nature of default and bankruptcy events, when contrasted with the routine functioning of businesses and regular account activities, poses a distinctive challenge. This challenge stems from the inherent imbalance in financial datasets, where instances of financial distress are notably outweighed by non-distress cases. To effectively address this intricacy and extract meaningful insights, various methodologies within the realm of machine learning have been devised.

The pivotal concern lies in the judicious selection of an algorithm capable of both efficiency and accuracy in classifying instances of financial risk. This study meticulously probes this concern, subjecting six distinct classifiers—logistic regression, XGBoost Classifier, decision tree classifier, random forest classifier, LightGBM classifier, and CatBoost Classifier—to comprehensive evaluation. Selected for their simplicity and effectiveness, these classifiers undergo scrutiny across three different datasets: the original dataset, an oversampled dataset, and an undersampled dataset. This methodological diversity enables a nuanced examination of classifier performance under varying conditions.

The recognition of models based on gradient boosting trees as superior performers in the evaluation of financial risk is consistent with the advancements in machine learning methodologies. In managing unbalanced data, it highlights the value of ensemble approaches and the possibilities of gradient boosting algorithms, offering theoretical support for upcoming developments in algorithmic development within financial risk assessment. Credit analysts and investors may use reliable bankruptcy prediction models built from these insights to make educated investment decisions, lowering financial risk and increasing profitability. Furthermore, governments may use such models to properly allocate funds, avert economic downturns, and encourage economic stability by monitoring firm financial health and preventing crises (He et al., 2018).

Over a comprehensive five-year timeframe, the research intricately evaluates the efficacy of these classifiers using three distinct evaluation metrics. These metrics, encompassing accuracy, sensitivity, and specificity, serve as robust benchmarks for

gauging the nuanced dynamics of each classifier's performance in the context of imbalanced financial datasets.

Significantly, the study unveils a noteworthy trend—the ascendancy of gradient boosting trees-based models, exemplified by XGBoost, LightGBM, and CatBoost. This discovery underscores the heightened effectiveness of ensemble methods, particularly those grounded in gradient boosting techniques, in discerning intricate patterns indicative of financial risk. In an era of evolving financial landscapes, this research not only provides crucial insights into the relative performance of classifiers but also contributes substantively to the ongoing discourse surrounding the optimal integration of machine learning tools in financial risk assessment. From this study, we concluded that features like (Receivables * 365) / Sales, Profit on Operating Activities / Financial Expenses, Total Costs / Total Sales, (Current Assets - Inventory) / Short-term Liabilities, etc. help a lot, especially in the tree-based models but are not of much use when it comes to logistic regression. Logistic Regression depends a lot, especially in utilizing the normal columns like Rotation Receivables + Inventory Turnover in Days, Operating Expenses / Short-term Liabilities, Sales / Short-term Liabilities, Sales / Receivables, Sales / Inventory, etc.

Utilizing data mining models to predict financial risk, credit analysts and investors can make informed decisions about which companies to invest in and which to avoid. This, in turn, can reduce financial risk and increase profits. Accurate bankruptcy prediction models is useful for the government to allocate funds and investment, avoid economic downturns, and promote economic growth. By monitoring the financial health of companies, governments can prevent financial crises and promote economic stability. This, in turn, can lead to increased investment and economic growth.

Future research directions in this domain could involve exploring the application of deep learning techniques like ANNs, LSTMs, CNNs, etc. Alongside this, the optimization of the parameters of all the models could have helped us to enhance further the accuracy and robustness of financial risk assessment models. In evaluations one can make use of many statistics and machine learning techniques, the use of statistics and machine learning tactics (Devi, 2018) addresses the bankruptcy predicting problem (Devi & Radhika, 2018). In order to enhance predictions even further, optimization techniques like Particle Swarm Optimization (PSO) and Genetic Algorithms (GA) can be combined. Additionally, investigating the integration of alternative data sources, such as social

media sentiment analysis and macroeconomic indicators, into the modeling process could yield valuable insights for even more precise risk evaluation.

Chapter 9
Future Work

Future research initiatives could investigate different approaches to get over these restrictions and improve the effectiveness of bankruptcy prediction models. One promising direction involves delving into the intricate web of financial data by incorporating more sophisticated deep-learning models. Convolutional Neural Networks (CNNs) and Artificial Neural Networks (ANNs) present themselves as potent tools with enhanced pattern identification capabilities, capable of capturing complex interdependencies within financial datasets that traditional models may overlook. This exploration into more intricate modeling approaches could unveil a new frontier in the quest for heightened predictive accuracy.

Moreover, the amalgamation of conventional statistical and machine learning methods with optimization techniques such as Particle Swarm Optimization (PSO) and Genetic Algorithms (GA) emerges as a strategic pathway. This fusion aims to not only enhance the performance of models but also to refine the predictive accuracy. By marrying the strengths of different methodologies, researchers can potentially overcome existing constraints and push the boundaries of predictive capabilities.

Expanding the scope of data sources utilized in the modeling process is another avenue for future exploration. Integrating sentiment analysis from social media or incorporating macroeconomic indicators into the predictive framework could inject fresh information, providing a more holistic understanding of the factors influencing bankruptcy risk. This expansion beyond traditional financial metrics may offer novel insights and contribute to an improved accuracy of risk assessment.

Furthermore, the fine-tuning of model parameters represents a crucial facet in advancing predictive robustness. Employing advanced hyperparameter tuning techniques like grid search or Bayesian optimization adds a layer of sophistication to the model optimization process. This meticulous adjustment of parameters holds the potential to enhance the overall robustness and accuracy of predictions, thus addressing one of the noted drawbacks in current models.

In culmination, the confluence of these diverse approaches – from intricate, deep learning models to innovative data source integration and advanced parameter tuning techniques – holds the promise of overcoming existing limitations. The integration of

these methodologies in subsequent studies may pave the way for more comprehensive and trustworthy bankruptcy prediction models, offering a nuanced perspective on financial risk assessment.

References

1. Acharjya, D. P., & Rathi, R. (2021). An extensive study of statistical, rough, and hybridized rough computing in bankruptcy prediction. *Multimedia Tools and Applications*, *80*(28–29), 35387–35413. https://doi.org/10.1007/s11042-020-10167-2
2. Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Ajayi, S. O., Bilal, M., & Akinade, O. O. (2016). Methodological approach of construction business failure prediction studies: A review. *Construction Management and Economics*, *34*(11), 808–842. https://doi.org/10.1080/01446193.2016.1219037
3. Aljawazneh, H., Mora, A. M., García-Sánchez, P., & Castillo-Valdivieso, P. A. (2021). Comparing the performance of deep learning methods to predict companies' financial failure. *IEEE Access*, *9*, 97010–97038.
4. Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(4), 589–609.
5. Artiach, T., Lee, D., Nelson, D., & Walker, J. (2010). The determinants of corporate sustainability performance. *Accounting & Finance*, *50*(1), 31–51. https://doi.org/10.1111/j.1467-629X.2009.00315.x
6. Board, J., Sutcliffe, C., & Ziemba, W. T. (2003). Applying Operations Research Techniques to Financial Markets. *Interfaces*, *33*(2), 12–24. https://doi.org/10.1287/inte.33.2.12.14465
7. Boritz, J. E., Kennedy, D. B., & Sun, J. Y. (2007). Predicting Business Failures in Canada*. *Accounting Perspectives*, *6*(2), 141–165. https://doi.org/10.1506/G8T2-K05V-1850-52U4
8. Breiman, L. (2001). [No title found]. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
9. Breiman, L. (2017). *Classification and regression trees*. Routledge.

https://www.taylorfrancis.com/books/mono/10.1201/9781315139470/classification-regression-trees-leo-breiman

10. Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446–3453.
11. Chen, M.-Y. (2011). Predicting corporate financial distress based on integration of decision tree classification and logistic regression. *Expert Systems with Applications*, *38*(9), 11261–11272.
12. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785
13. Cheng, C.-H., Chan, C.-P., & Sheu, Y.-J. (2019). A novel purity-based k nearest neighbors imputation method and its application in financial distress prediction. *Engineering Applications of Artificial Intelligence*, *81*, 283–299.
14. Crone, S. F., & Finlay, S. (2012). Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting*, *28*(1), 224–238.
15. Devi, S. S., & Radhika, Y. (2018). A survey on machine learning and statistical techniques in bankruptcy prediction. *International Journal of Machine Learning and Computing*, *8*(2), 133–139.
16. Edum-Fotwe, F., Price, A., & Thorpe, A. (1996). A review of financial ratio tools for predicting contractor insolvency. *Construction Management and Economics*, *14*(3), 189–198. https://doi.org/10.1080/014461996373458
17. Fedorova, E., Gilenko, E., & Dovzhenko, S. (2013). Bankruptcy prediction for Russian companies: Application of combined classifiers. *Expert Systems with*

Applications, *40*(18), 7285–7293.

18. Fulmer, J. G., Moon, J. E., Gavin, T. A., & Erwin, M. (1984). A bankruptcy classification model for small firms. *Journal of Commercial Bank Lending*, *66*(11), 25–37.
19. García, S., & Herrera, F. (2009). Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. *Evolutionary Computation*, *17*(3), 275–306.
20. He, H., Zhang, W., & Zhang, S. (2018). A novel ensemble method for credit scoring: Adaption of different imbalance ratios. *Expert Systems with Applications*, *98*, 105–117.
21. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html
22. Keya, M. S., Akter, H., Rahman, M. A., Rahman, M. M., Emon, M. U., & Zulfiker, M. S. (2021). Comparison of different machine learning algorithms for detecting bankruptcy. *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, 705–712. https://ieeexplore.ieee.org/abstract/document/9358587/
23. Kim, S. Y., & Upneja, A. (2014). Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models. *Economic Modelling*, *36*, 354–362.
24. Lahmiri, S., Bekiros, S., Giakoumelou, A., & Bezzina, F. (2020). Performance assessment of ensemble learning systems in financial data classification. *Intelligent*

Systems in Accounting, Finance and Management, *27*(1), 3–9. https://doi.org/10.1002/isaf.1460

25. Li, H., Li, C.-J., Wu, X.-J., & Sun, J. (2014). Statistics-based wrapper for feature selection: An implementation on financial distress identification with support vector machine. *Applied Soft Computing*, *19*, 57–67.
26. Li, H., Sun, J., & Sun, B.-L. (2009). Financial distress prediction based on OR-CBR in the principle of k-nearest neighbors. *Expert Systems with Applications*, *36*(1), 643–659.
27. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, *31*. https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html
28. Quynh, T. D., & Phuong, T. T. L. (2020). Improving the bankruptcy prediction by combining some classification models. *2020 12th International Conference on Knowledge and Systems Engineering (KSE)*, 263–268. https://ieeexplore.ieee.org/abstract/document/9287707/
29. Shultz, T. R., & Schmidt, W. C. (1991). A cascade-correlation model of balance scale phenomena. *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, 635–640. https://www.psych.mcgill.ca/perpg/fac/shultz/personal/Recent_Publications_files/balance91.pdf
30. Smiti, S., & Soui, M. (2020). Bankruptcy Prediction Using Deep Learning Approach Based on Borderline SMOTE. *Information Systems Frontiers*, *22*(5), 1067–1083. https://doi.org/10.1007/s10796-020-10031-6

31. Springate, G. L. (1978). *Predicting the possibility of failure in a Canadian firm: A discriminant analysis* [PhD Thesis]. Simon Fraser University.

32. Stoltzfus, J. C. (2011). Logistic Regression: A Brief Primer. *Academic Emergency Medicine*, *18*(10), 1099–1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x

33. Sun, J., Lang, J., Fujita, H., & Li, H. (2018). Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. *Information Sciences*, *425*, 76–91.

34. Taffler, R. J. (1983). The Assessment of Company Solvency and Performance Using a Statistical Model. *Accounting and Business Research*, *13*(52), 295–308. https://doi.org/10.1080/00014788.1983.9729767

35. Tomczak, S. (2016). Polish Companies Bankruptcy Data, Data Set. *UCI–Machine Learning Repository*.

36. Van Hulse, J., Khoshgoftaar, T. M., & Napolitano, A. (2007). Experimental perspectives on learning from imbalanced data. *Proceedings of the 24th International Conference on Machine Learning*, 935–942. https://doi.org/10.1145/1273496.1273614

37. Veganzones, D., & Séverin, E. (2018). An investigation of bankruptcy prediction in imbalanced datasets. *Decision Support Systems*, *112*, 111–124.

38. Zahiri, P. (2022). *Bankruptcy Prediction by Deep Learning and Machine Learning Methods* [PhD Thesis, Concordia University]. https://spectrum.library.concordia.ca/id/eprint/991305/

39. Zięba, M., Tomczak, S. K., & Tomczak, J. M. (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications*, *58*, 93–101.