Predicting Corporate Bankruptcy with Ensemble Machine Learning: Modeling the Temporal Dynamics of Financial Distress in Finnish Companies

University of Turku
Department of Computing
Master's Thesis
June 2025
Mikael Nordqvist

Supervisors:
Dr. Luca Zelioli
Msc. Adrian Borzyszkowski

UNIVERSITY OF TURKU
Department of Computing

Mikael Nordqvist: Predicting Corporate Bankruptcy with Ensemble Machine Learning: Modeling the Temporal Dynamics of Financial Distress in Finnish Companies

Master's Thesis, 53 p.

June 2025

The accurate prediction of corporate bankruptcy is a cornerstone of financial risk management, crucial for investors, creditors, and the broader economy. This thesis investigates the efficacy of ensemble machine learning models in forecasting the insolvency of Finnish limited liability companies, with a specific focus on modeling the temporal dynamics of financial distress.

This study employs XGBoost, LightGBM, and Random Forest models on time series data composed of three and five consecutive financial statements. A key contribution is the use of realistic prediction horizons that account for the inherent lag in the public availability of financial reports, a factor often overlooked in prior research. Model performance and the importance of features are evaluated on time series sequences with prediction horizons of 1-4 years and sequence lengths of 3 and 5 financial statements.

The findings reveal that while the models, particularly XGBoost, demonstrate strong predictive power, their accuracy degrades with longer horizons, with a notable decline in performance for predictions made more than three years prior to bankruptcy. The analysis of feature importance consistently highlights profitability metrics, such as gross profit and net financial expenses relative to revenue, as the most significant predictors of distress.

Future research on this topic should investigate the integration of either non-financial or real-time data, such as public debt judgments, to improve forecasting accuracy and further address the challenges posed by financial reporting lags.

Keywords: Bankruptcy prediction, Ensemble model, Prediction horizon, Feature importance, Time series

UNIVERSITY OF TURKU
Department of Computing

Mikael Nordqvist: Predicting Corporate Bankruptcy with Ensemble Machine Learning: Modeling the Temporal Dynamics of Financial Distress in Finnish Companies

Master's Thesis, 53 p.

June 2025

Yrityskonkurssien tarkka ennakointi on taloudellisen riskinhallinnan perusta ja tärkeää niin sijoittajille, velkojille kuin koko kansantaloudellekin. Tässä tutkielmassa arvioidaan ensemble-koneoppimismallien soveltuvuutta suomalaisten osakeyhtiöiden maksukyvyttömyyden ennustamiseen. Erityisenä painopisteenä on taloudellisen ahdingon kehittymisen mallintaminen ajan funktiona.

Tutkimuksessa sovelletaan XGBoost-, LightGBM- ja Random Forest -malleja kolmesta tai viidestä peräkkäisestä tilinpäätöksestä koostettuun aikasarja-aineistoon. Tutkimuksen keskeinen tutkimuksellinen arvo on realististen ennustehorisonttien määrittäminen. Horisontit huomioivat tilinpäätöstietojen julkistamisviiven, joka on aiheen kirjallisuudessa jäänyt laajalti huomioimatta. Mallien ennustuskykyä ja muuttujien merkittävyyttä arvioidaan 1–4 vuoden ennustehorisonteilla ja sekä kolmen, että viiden tilinpäätöksen mittaisilla aikasarjoilla.

Tulokset osoittavat, että vaikka malleilla, erityisesti XGBoostilla, on vahva ennustuskyky, niiden tarkkuus heikkenee ennustehorisontin pidentyessä. Suorituskyky laskee huomattavasti, kun ennusteet ulottuvat yli kolmen vuoden päähän konkurssista. Muuttujien merkittävyysanalyysissä keskeisimmiksi ennustetekijöiksi korostuvat kannattavuusmittarit, kuten myyntikate ja liikevaihtoon suhteutetut nettorahoituskulut. Tulevaisuudessa olisi arvokasta selvittää, voiko ei-taloudellista tai muuta reaaliaikaista tietoa, kuten julkisia velkomustuomioita, hyödyntää ennustetarkkuuden parantamiseen ja raportointiviiveiden aiheuttamien ongelmien ehkäisyyn.

Keywords: Konkurssi, Ennustus, Ensemble-malli, Aikasarja

Contents

1 Introduction **1**

1.1 Research Motivation 2

1.2 Research Questions and Structure 3

2 Background **4**

2.1 Corporate Insolvency 4

2.2 Literature Review 5

2.2.1 Classical Statistical Models 6

2.2.2 Machine Learning and Ensemble Methods 6

2.2.3 Prediction Horizon Challenges 8

2.2.4 Methodologies for Handling Class Imbalance 9

2.2.5 Bankruptcy Prediction in the Nordic Context 10

2.3 Class Imbalance 10

2.4 Ensemble Machine Learning Models 13

2.5 Model Performance Metrics 15

3 Methodology **19**

3.1 Research Design 19

3.2 Dataset Description 21

3.3 Data Considerations and Validity 22

3.4 Data Preprocessing 24

3.5 Model Implementation and Training 28
3.5.1 XGBoost 29
3.5.2 LightGBM 31
3.5.3 Random Forest 33

4 Results **35**
4.1 Basic Metrics 35
4.2 Sequence Length and Prediction Horizon 39
4.3 Feature Importance 41
4.4 Further Analysis 43

5 Discussion **45**
5.1 Research Question 1 45
5.2 Research Question 2 46
5.3 Research Question 3 47
5.4 Limitations of the Study 48
5.5 Future Research 50

6 Conclusion **52**

References **54**

Appendix **64**

List of Figures

3.1 KDE Plot of Final FS Operating Profit IQR Grouped by End Date . 24

3.2 Data Processing and Model Evaluation Pipeline 25

4.1 Receiver Operating Characteristic Curves 38

4.2 XGBoost Logarithmic Loss Grouped by Prediction Horizon for Sequence Length 3 40

4.3 XGBoost Recall Grouped by Prediction Horizon for Sequence Length 3 40

4.4 XGBoost Feature Importance 42

4.5 LightGBM Feature Importance 42

4.6 Random Forest Feature Importance 43

4.7 XGBoost Logarithmic Loss Grouped by Revenue 44

List of Tables

2.1 Confusion Matrix for Binary Classification 16

3.1 XGBoost Hyperparameter Search Space 29
3.2 LightGBM Hyperparameter Search Space 32
3.3 Random Forest Hyperparameter Search Space 33

4.1 Model Performance Comparison for Sequence Length 3 Across Prediction Horizons 36
4.2 Model Performance Comparison for Sequence Length 5 Across Prediction Horizons 37

A.1 Income Sheet Variables 65
A.2 Balance Sheet Variables 66
A.3 Profitability and Solvency Indicators 67
A.4 Liquidity, Size, and Efficiency Indicators 68

Acronyms

ANN Artificial Neural Network.

ARMA Autoregressive Moving-Average.

AUC Area Under Curve.

DNN Deep Neural Network.

DT Decision Tree.

FS Financial Statement.

GBDT Gradient Boosted Decision Trees.

GOSS Gradient-Based One-Side Sampling.

IQR Interquartile Range.

KDE Kernel Density Estimate.

KNN K-Nearest Neighbors.

LR Logistic Regression.

ML Machine Learning.

MLP Multi-Layer Perceptron.

OCR Optical Character Recognition.

ORK Oikeusrekisterikeskus (Finnish Legal Register Centre).

PRH Patentti- ja rekisterihallitus (Finnish Patent and Registration Office).

RF Random Forest.

RNN Recurrent Neural Network.

ROC Receiver Operating Characteristic.

SGB Stochastic Gradient Boosting.

SMOTE Synthetic Minority Over-sampling Technique.

SVM Support Vector Machine.

WPD Warp Path Distance.

1 Introduction

Predicting corporate bankruptcy is a cornerstone of financial risk management, with significant implications for creditors, investors, and the broader economy. Predictive identification of companies at risk of insolvency allows for proactive measures, potentially mitigating financial losses for both stakeholders, creditors, and the broader public, who may be affected through pension funds, employment disruptions and general market uncertainty. While traditional statistical methods have long been employed in bankruptcy prediction, machine learning models, particularly ensemble techniques, have gained popularity in recent years, offering a more sophisticated and potentially more accurate method for assessing potential insolvency.

This thesis investigates the application of ensemble machine learning models, specifically XGBoost [1], LightGBM [2], and Random forest [3], to predict the bankruptcy of Finnish limited liability companies using a time series approach. The dataset consists of financial variables and indicators extracted from companies' financial statements and is used to create sequences of three and five financial statements to evaluate the influence of sequence length on prediction performance. Additionally, to handle the imbalance inherent to datasets in this domain, oversampling methods are applied. A notable aspect of this study is the careful consideration given to the often overlooked temporal misalignment between the public availability of financial statements and the actual date of bankruptcy filings, ensuring that the prediction horizons realistically reflect practical scenarios.

The practical utility of predictive models hinges on understanding the reliability of predictions as a function of how long into the future the prediction is being made. The selected models are applied to prediction horizons of 1 to 4 financial periods to assess the reliability of predictions. Furthermore, the importance of features is evaluated to identify the financial variables with the highest predictive power.

1.1 Research Motivation

Accurate bankruptcy prediction models offer significant practical applications for a wide range of economic actors. Creditors are equipped to make smarter lending criteria and minimize losses from defaulting debtors. Investors and shareholders benefit from reliable early warnings of financial distress, enabling them to divest from risky companies. Furthermore, government agencies can also leverage these models for economic monitoring and intervene to support vulnerable sectors. Finally, individual companies can use these predictions externally to make more informed partnership decisions, as well as internally to identify and rectify financial weaknesses before they lead to insolvency.

Despite a wealth of research in bankruptcy prediction, gaps persist in the literature, which this study aims to address. Firstly, the vast majority of studies overlook the temporal misalignment between the end of a financial reporting period and the public availability of the financial statement, which leads to overoptimistic results. This study confronts this issue by implementing a realistic prediction horizon that accounts for the up-to-8-month delay in the filing and availability of financial statements for Finnish companies. The models' performance on these realistic prediction horizons is further informed by examining the effect of data sequence length.

Secondly, while ensemble models are popular in the literature, there is a lack of comparative studies focusing specifically on Finnish companies. This study contributes by applying and comparing three of the most popular and promising

ensemble models on a comprehensive and up-to-date dataset of Finnish limited liability companies.

1.2 Research Questions and Structure

To address the identified gaps, this study is guided by the following research questions:

RQ1: To what extent do specific ensemble learning models (XGBoost, LightGBM, and Random Forest) differ in their predictive performance when forecasting bankruptcy for Finnish limited liability companies using financial time series data?

RQ2: How do variations in input financial data sequence length and prediction horizon influence the predictive performance of ensemble models in the context of corporate bankruptcy prediction for Finnish limited liability companies?

RQ3: Which financial features are identified as most important for predicting bankruptcy in Finnish limited liability companies when using ensemble models?

To answer these questions, the thesis is structured as follows. Chapter 2 provides the theoretical background, defining corporate insolvency and reviewing the relevant literature on bankruptcy prediction. Chapter 3 details the research methodology, including the dataset description, data preprocessing steps, model implementation, and the performance metrics. Chapter 4 presents the empirical results of the experiments. Chapter 5 discusses these results in the context of the research questions and explores the limitations of the study. Finally, Chapter 6 concludes the thesis by summarizing the key findings and suggesting directions for future research.

2 Background

The chapter details the existing works in the context of corporate bankruptcy prediction and provides general descriptions of the methods used in this study. First, a description of corporate insolvency is given. Then, existing works in the domain are reviewed and models, methods and metrics used in the study are defined.

2.1 Corporate Insolvency

Corporate insolvency refers to a financial state in which a company is unable to pay its debts as they become due. When a company has enough assets to cover its debt, but is unable to liquidate in time, it is referred to as a cash-flow insolvency. A balance-sheet insolvency refers to a situation where the total due debts of a company exceed the value of its current assets. [4] A balance-sheet insolvency is more likely to lead to an official bankruptcy proceeding, as a cash-flow insolvency can more easily be negotiated around.

In Finland, a bankruptcy can be initiated by a debtor, in this case a company, or a creditor, via a district court. The court appoints a bankruptcy trustee, who will be responsible for managing the debtor's assets and will oversee their liquidation. For the bankruptcy application to be valid, the debtor must be insolvent. The district courts also allow corporate debt restructuring to avoid liquidation. It is an alternative to bankruptcy, which allows the company to continue operations after the procedure. However, the company can still be put up for bankruptcy later, if the

debt restructuring fails to resolve the debt situation. [5]

In the case of bankruptcy, the final accounting period of a company ends when the company files for bankruptcy and the company is expected to file a financial statement for that period after the initiation of the bankruptcy. In the case of debt restructuring, the company can continue operations under the same name and the same business ID. Debt restructuring proceedings do not end the accounting period like bankruptcy does, and the reporting of financial statements can continue as normal. [5]

2.2 Literature Review

In the literature, the prediction of financial distress is approached through different study objectives, ranging from predicting formal bankruptcy to predicting more abstract metrics such as the ratings of official credit rating agencies, such as in [6], where ratings from the Korean Credit Information Service are used and in [7], where the ratings from Fitch Ratings are used. Credit ratings, while distinct from formal bankruptcy, are fundamentally still attempting to assess the underlying financial well-being of a company. However, the interpretation of the results can differ depending on the point of view of the study. Studies from the point of view of a potential creditor focus on modeling the likelihood a company will honor its debts, and studies from the point of view of a potential investor might focus on modeling aspects like the impact of official credit ratings on the company's stock price, such as [8].

Models in the literature employ both cross-sectional and time series financial data. While cross-sectional data represents a snapshot of the companies' finances at a single point in time, time series data captures the historical trends and volatility, which should provide superior predictive performance by being able to distinguish between companies that reach a given cross-sectional snapshot through different means. Despite their limitations, cross-sectional studies still provide valuable insights

into feature and model selection for the time series context. For some markets, the availability of financial statements is limited due to differences in regulations. For example, in the US, the Securities and Exchange Commission does not require small private companies to report their financial data, whereas in Finland, the reporting of financial data is mandatory for all companies that meet a set of requirements [9][10].

2.2.1 Classical Statistical Models

The early research on predicting financial distress used primarily statistical methods. A foundational study [11] introduced a single variable discriminant model using financial ratios was introduced. This work was expanded upon in [12] to create a multivariate discriminant model based on five input variables depicting the main aspects of a company's financial profile: profitability, leverage, liquidity, solvency and activity. A logistic regression model was proposed in [13]. However, a subsequent analysis [14] argued that these seminal works have a fundamental issue in their sample selection procedures that led to biased results. In [15], it was suggested that already in the 1980s the economic environment had shifted, such that the coefficients from the seminal studies were now unsuitable.

2.2.2 Machine Learning and Ensemble Methods

In recent times, studies have largely focused on the application of machine learning models to financial distress predictions. The findings in [16] suggest that machine learning models demonstrate a substantial improvement in accuracy over traditional statistical techniques. The authors suggest that machine learning models are generally more robust and rely on less assumptions about the distribution of the features and financial indicators. However, a review of the literature in [17] argues that it is inconclusive which types of models or methods perform best, as results are often highly context-dependent. The impact of sequential data was explored using

Recurrent Neural Network models in [18]. The findings indicated that sequential data increased prediction accuracy, especially in machine learning methods. In [19], shallow neural networks were applied to a dataset of Belgian companies. A subsequent study [20] found that no significant improvement could be made by using a deeper neural network and XGBoost. In [21], decision trees, SVM and MLP models were used. Though the review of the literature done as part of the study indicated that SVM and MLP models would produce the best results, it was decision trees that did so in the experiments. A trend towards ensemble models is noted in [22], where the authors also emphasize the importance of incorporating non-financial features like market reputation.

The experimental section of our study specifically involves applying ensemble models to financial time series data for predicting bankruptcy. Ensemble models show great promise in this particular area, but also in the field of time series forecasting or analysis in general. In the most recent Makridakis Competitions, also known as the M Competitions, ensemble models have been both the most popular and the best performing models overall [23][24][25]. The M Competitions are series of open competitions aiming to evaluate and compare methods for time series forecasting. LightGBM [2] has been particularly dominant among the ML models, best exemplified by M5, where nearly all of the top 50 competitors employed LightGBM in some form in their solution [24]. In [26], ensemble models, including XGBoost [1], AdaBoost [27] and GBDT [28] were used in combination with sampling methods SMOTE [29] and EasyEnsemble [30] as well as feature selection and importance methods to produce better results than Support Vector Machines [31], traditional statistical models and deep neural networks. After testing numerous base learners for ensemble methods, the research in [32] found that ensemble techniques produce better results than single classifiers. A study of US companies [33] used machine learning models to predict bankruptcy risk, finding that neural network models produced better results than

ensemble models. XGBoost was applied specifically to sequential data in [34]. It produced more accurate credit rating predictions than the baseline ARMA model [35]. In [36], an ensemble method with MLP hyperparameter optimization based on Altman Z-Score was proposed. The results indicated that bagging approaches were able to improve models in all metrics and also produced a solution that can handle class imbalances, which are often found in bankruptcy datasets. In [37], incorporating textual features from company reports is studied. The findings suggest that textual features improve prediction performance and that ensemble methods, XGBoost, Random Forest, and GBDT, were particularly effective on the numerical financial data.

2.2.3 Prediction Horizon Challenges

A common finding in studies where the length of the prediction is considered is that longer prediction horizons correspond to worse prediction results. However, prediction horizons are often stated in a way that is too abstract. The horizon is often phrased along the lines of "one year prior to the bankruptcy filing", which is ambiguous, or it is not stated at all. Data of companies' finances are typically gathered in either yearly or quarterly increments and are actually filed and available only after the financial period has already ended, whereas bankruptcies can be filed at any date and information about them is publicly available almost immediately. This timing discrepancy means that shorter prediction horizons can inadvertently include data that became publicly available only after the bankruptcy filing, effectively predicting bankruptcies that have already occurred. This distinction is frequently overlooked in the literature, rendering comparisons of prediction horizons across different studies highly unreliable. In [37], predictions are made with data from 3, 4 and 5 years prior to bankruptcy, with the prediction horizon of 4 years surprisingly producing the best results. Similarly, a study using prediction horizons from 1 to

5 years [38] found that each successive year produced worse prediction results. In [39], ensemble methods like XGBoost, LightGBM and GBDT were used to predict financial distress with prediction horizons of 1 to 5 years. The conclusions indicate a similar decrease in prediction performance but also that the importance of features differs between shorter and longer horizons. In [40], prediction horizons of 1 to 5 years were also run with similar results. The results also demonstrate that the selected ensemble models, RF and XGBoost outperform other models like neural networks. A novel undersampling method to deal with class imbalance is also used, which produced better results than oversampling the minority class, contradicting findings from other studies.

2.2.4 Methodologies for Handling Class Imbalance

Many studies bring attention to this imbalanced nature of most datasets used in bankruptcy prediction, where the number of non-bankrupt companies typically far outweighs the number of bankrupt ones. A common solution is to oversample the bankrupt class using methods like SMOTE or to undersample the non-bankrupt class. A review in [22] found that variations of SMOTE are both the most common and the best performing methods. Different preprocessing methods were applied to data from Polish companies in [38]. Among the models tested, RF produced the best results, particularly on data that was balanced using SMOTE. In [41], the authors focused primarily on the handling of imbalance in financial datasets. The key findings were that the choice of balancing method impacted ML models performance and that the performance of the balancing method was also impacted by the model selection. No conclusion about the generally most optimal balancing methods could be made. In [42], numerous SMOTE variations were experimented with on an RF model, out of which ADASYN [43], an implementation of SMOTE that emphasizes harder to learn samples, produced the best results. Out of all methods in the paper

an XGBoost model with an imbalance-resilient loss function produced the strongest result, indicating that boosting models like XGBoost are innately robust against imbalanced datasets.

2.2.5 Bankruptcy Prediction in the Nordic Context

There are also a few studies that have applied machine learning models to a dataset of Finnish or Nordic companies, similar to ours. One study focusing on Nordic listed companies [44] applied LightGBM, neural network, and logistic regression models to a dataset of quarterly financial key figures. LightGBM produced the best results with precision, recall and ROC-AUC scores of 0.52, 0.78 and 0.93, respectively. The meaning of each performance metric is explained in section 2.5. A similar time series approach was taken with Danish companies in [45]. The models applied were logistic regression, random forest and a neural network model, with the RF model demonstrating the strongest performance with precision, recall and ROC-AUC scores of 0.9473, 0.9775 and 0.9614, respectively. Research focused specifically on Finnish companies applied multiple types of decision trees to companies listed on the Helsinki Nasdaq, achieving a ROC-AUC score of 0.847 using a two-class decision tree [46]. In [47], a comprehensive dataset of financial statement numbers from all privately held Norwegian small to medium-sized enterprises is used. The data is used to predict one-year-ahead bankruptcy with a ROC-AUC score of about 0.87.

2.3 Class Imbalance

To handle imbalance in the target variable sampling methods can be applied. One of the most popular oversampling techniques in the literature is SMOTE (Synthetic Minority Over-sampling Technique) and its different derivations. In [48], a comparison of techniques for handling class imbalance in credit scoring was conducted. It was

found that SMOTE and its variant outperform random oversampling, but also that ensemble methods like random forest and XGBoost produced good results even on highly imbalanced credit scoring data.

SMOTE, which was first proposed in [29], generates synthetic examples from the minority class by interpolating from the existing instances within that class. Equation 1 defines the creation of a synthetic example in SMOTE:

$$S = A + \alpha \cdot (B - A) \tag{1}$$

where S is the generated example, A is a minority class instance, B one of its k-nearest neighbors (KNN) [49] and α a random number between 0 and 1 ($\alpha \sim \text{Uniform}(0, 1)$). [29] More than one neighbor can also be used in the generation of the sample and on multivariate datasets, SMOTE can be applied on each feature individually. However, this baseline SMOTE is not directly applicable to data consisting of many time series sequences instead of individual data points. It can be applied to generate time series sequences by applying the sampling of A and its k-nearest neighbors on a per timestep basis. However, this still retains the possibility of generating sequences that are nonsensical with respect to the temporal dependency in the values for a feature.

To account for this, a slightly modified version is applied, where KNN is used to compare differences in entire time series sequences, instead of individual values. The bankrupt sequences are flattened to a vector of size timesteps $\times$ features. The k-nearest neighbors of a sequence are found by taking the Euclidean distance between the full sequences. This captures the similarity in both the feature values and their temporal patterns. The synthetic samples are then generated by interpolating between a sequence and one of its five nearest neighbors on a per-timestep basis. The value of a feature at a given timestep in the generated sequence is interpolated from the values of that feature at the same timestep in the genuine sequences.

M-SMOTE is a variation of SMOTE designed specifically for handling imbalance in time series datasets, originally proposed in [50]. M-SMOTE aims to generate samples that are more representative of the true distribution of the minority class, by generating samples around the centroid. The application of M-SMOTE consists of four steps: (1) normalizing all features to the range $[-1, 1]$ using a min-max scaler, (2) calculating the centroid of minority class samples, which is the arithmetic mean of all normalized minority class samples and will serve as the reference point in the generation of synthetic samples, (3) partitioning the minority class samples into safe points, samples close to the centroid, and noise points, samples further away from the centroid, and (4) generating the synthetic samples by interpolating feature values between the safe points and the centroid. Equation 2 defines the creation of a synthetic example in M-SMOTE:

$$X_{\text{new}} = X_{\text{safe}} + \alpha \cdot (X_{\text{center}} - X_{\text{safe}}) \tag{2}$$

where X_{new} is the generated sample, X_{safe} a safe point from the minority class, α a random number between 0 and 1 ($\alpha \sim \text{Uniform}(0, 1)$) and X_{center} the centroid of the minority class. The partitioning of the minority class samples into safe and noise samples is done using the warp path distance (WPD) (Eq. 3) between all minority class samples and the centroid, X_{center}. The WPD between two time series is defined as:

$$\text{WPD}(X, Y) = \min_{P} \sum_{k=1}^{K} d(x_{i_k}, y_{j_k}) \tag{3}$$

where $P = (p_1, p_2, \ldots, p_K)$ is the optimal warping path, and $d(x_{i_k}, y_{j_k})$ is the distance (e.g., Euclidean distance) between x_{i_k} and y_{j_k}. The optimal warping path is a sequence of monotonically increasing indices from $p_1 = (1, 1)$ to $p_K = (m, n)$, where m and n are the lengths of time series X and Y, across which the cumulative distance between the points is minimized. [50]

2.4 Ensemble Machine Learning Models

Ensemble learning, a concept initially introduced in [51], has proven effective in addressing the limitations of individual machine learning models. The ensemble approach is split into two different strategies: Boosting, introduced in [52], and Bootstrap Aggregation (Bagging) introduced in [53]. In boosting methods, many weak learners are trained to create a sequence of predictors. Each individual base-learner provides a prediction that is only marginally better than a random guess. Learners are trained in sequence on a weighted version of the original dataset, with previously misclassified points having a higher weight in future iterations. The final prediction of the ensemble can be produced in different ways like a weighted sum for regression and majority voting for classification. [54]

In bagging methods learners are trained instead in parallel. Each learner is trained on a subset of the dataset generated randomly through bootstrapping. Therefore, each learner is independent, unlike in boosting, where the input of each learner depends on the output of earlier learners. In bagging methods for regression, final outputs can be produced through averaging over the outputs of all learners. For classification problems, majority voting can be used. [55]

Random forest models were first proposed in [3]. The model is an ensemble of sequential decision trees. For each tree, a bootstrap sample of the dataset is created that includes a feature or features selected using a splitting rule that maximizes the impurity reduction introduced by the split. In classification tasks, the Gini impurity (Eq. 4) criterion is the most common choice:

$$C(t) = \sum_{j=1}^{J} \hat{p}_j(t)\,(1 - \hat{p}_j(t)) \tag{4}$$

where $\hat{p}_j(t)$ is the class frequency for class j in the node t. [56] Maximizing the Gini impurity reduction through a split in classification tasks means choosing the feature

and split point that best separate the data into groups where each group contains as many instances of the same target variable value as possible. Equation 5 defines the output of a random forest model, which is the average output of all individual trees

$$M_{\text{rf}}(x) = \frac{1}{K}\sum_{k=1}^{K} t_k(x) \tag{5}$$

where M_{rf} is the RF model's calculation result, K is the number of decision trees, and t_k is a single decision tree model [3].

GBDT is a boosting model that is the foundation of modern boosting ensemble methods like XGBoost and LightGBM. It was first introduced in [28]. GBDT builds an ensemble of decision trees sequentially, where each new tree corrects the error in residuals of the previous trees. GBDT, like the random forest bagging model, uses the CART decision tree algorithm proposed in [57], which creates binary decision trees. The optimization process in gradient boosting models can be represented by the update rule at each iteration t presented in equation 6:

$$F_t(x) = F_{t-1}(x) + \eta \cdot h_t(x) \tag{6}$$

where $F_t(x)$ is the model's prediction at iteration t, $F_{t-1}(x)$ is the model's prediction at the previous iteration, η is the learning rate, and $h_t(x)$ is the decision tree trained at iteration t to approximate the negative gradient of the loss function [28].

XGBoost is an optimized implementation of GBDT introduced in [1]. XGBoost adds regularization terms to the loss function which aim to prevent overfitting. Equation 7 defines the regularized objective function:

$$\mathcal{L}^{(t)}_{XGB} = \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t) \tag{7}$$

where g_i and h_i are the first and second derivatives of the loss function with respect to the previous prediction, $f_t(x_i)$ represents the prediction of the new tree for sample i, and $\Omega(f_t)$ is the regularization term penalizing complex tree structures. XGBoost

includes additional improvements such as a histogram-based splitting method, which compares binned values instead of comparing all possible splitting points, reducing computation time, and a tree pruning strategy where the tree is first constructed up to its maximum depth and then pruned backwards by removing splits that do not improve the objective function. [1]

LightGBM is another further optimized implementation of GBDT proposed in [2]. LightGBM includes the additional innovations of Gradient-Based One-Side Sampling (GOSS), which prioritizes training examples with larger gradients and randomly samples less informative examples, and Exclusive Feature Bundling (EFB), which bundles together features that rarely hold non-zero values simultaneously, reducing dimensionality without losing information. Equation 8 defines The regularized objective function with GOSS:

$$\mathcal{L}_{LGBM}^{(t)} \approx \sum_{i \in A} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \frac{1-a}{b} \sum_{i \in B} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad (8)$$

where g_i and h_i are the first and second derivatives of the loss function with respect to the previous prediction, $f_t(x_i)$ represents the prediction of the new tree for sample i, $\Omega(f_t)$ is the regularization term, A is the set of instances with large gradients and a their sampling ratio, B is a random subset of instances with small gradients and b their sampling ratio.

2.5 Model Performance Metrics

In binary classification tasks, such as predicting whether a company will go bankrupt, each data point is predicted to be either positive or negative. In this case, a positive classification means that the company is predicted to go bankrupt within the prediction horizon. Therefore, each predicted label has one of four different designations: a true positive (TP) is a correctly predicted positive label, a false

positive (FP) is a negative label predicted as positive, a true negative (TN) is a correctly predicted negative label and a false negative (FN) is an incorrectly predicted negative label. [58]

The values of these four categories can be visualized in a confusion matrix. It provides a summary of how well the model is able to discriminate between and in the context of binary classification models, it typically takes this form:

Table 2.1: Confusion Matrix for Binary Classification

		Actual Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Further performance metrics that provide more insight into the discriminative performance of the model can be calculated using the four categories.

Precision (Eq. 9) measures the accuracy of positive predictions and answers the question: "Of all instances predicted as positive, how many were actually positive?"

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \in [0, 1] \tag{9}$$

Recall (Eq. 10), also called the True Positive Rate (**TPR**) measures the ability of the model to identify actual positive instances and answers the question: "Of all actual positive instances, how many were correctly predicted as positive?"

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \in [0, 1] \tag{10}$$

Specificity (Eq. 11), also called the True Negative Rate (**TNR**) measures the ability of the model to identify actual negative instances and answers the question: "Of all actual negative instances, how many were correctly predicted as negative?"

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \in [0, 1] \tag{11}$$

Accuracy (Eq. 12) calculates the number of correctly classified positive and negative instances against the total number of instances evaluated. It answers the question: "Out of all predictions made, what fraction were correct?" [58]

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \in [0, 1] \tag{12}$$

The aforementioned evaluation metrics are useful, but can be misleading, especially on an imbalanced dataset. For example, if 95% of sequences in the test set are non-bankrupt, the model will achieve a 95% accuracy simply by predicting every sequence as non-bankrupt. To address this problem, more evaluation metrics must be included.

F1-score (Eq. 13) is the harmonic mean of **Precision** and **Recall**. It represents the balance between high precision, minimizing false positives, and high recall, minimizing false negatives, at a given threshold value. [58] For example, a threshold of 0.5 means a predicted probability of 0.5 represents a bankruptcy prediction and anything less represents a non-bankrupt prediction.

$$\text{F1-score} = \frac{2 * \text{TP}}{2 * \text{TP} + \text{FP} + \text{FN}} \in [0, 1] \tag{13}$$

AUC-ROC (Eq. 14), an acronym for Area Under the Receiver Operating Characteristic Curve, is a metric that measures the ability of the model to discriminate between the two classes across all threshold values. The Receiver Operating Characteristic (ROC) curve is a graphical representation of **Recall** against the False Positive Rate (**FPR**), which is the ratio between false positive predictions and the total number of actually negative instances. The area under the ROC curve equals the probability that a randomly selected, actually positive instance has a higher predicted probability of belonging to the positive class than a randomly selected negative instance. [59]

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}) \in [0, 1] \tag{14}$$

Binary Cross-Entropy (Eq. 15), also known as logarithmic loss, is a metric that quantifies how close the predicted probability values were to the actual binary outcomes of 0 and 1. [60] It heavily penalizes any predictions that assign a high probability to the incorrect class and can be used to evaluate predicted probabilities, not just binary predictions, like the metrics mentioned above.

$$\text{Log loss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \in [0, \infty) \tag{15}$$

Where N is the total number of instances, y_i is the actual binary label (0 or 1) for instances i, p_i is the predicted probability that instance i belongs to class 1.

Brier Score (Eq. 16) is another metric that measures the accuracy of predicted probabilities. It calculates the mean squared difference between the predicted probability and the actual binary outcome. [61] It is less sensitive to extremely large prediction errors from outliers and is commonly used as the numeric measure of a model's calibration.

$$\text{Brier Score} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \in [0, 1] \tag{16}$$

3 Methodology

This chapter details the methodology employed in this study. It begins by outlining the research design and objectives, followed by a comprehensive description of the dataset, including its source and considerations regarding data validity and external factors. Subsequently, the chapter describes the data preprocessing steps undertaken. Finally, it details the implementation, hyperparameter optimization, and specific configurations of the ensemble models used for bankruptcy prediction.

3.1 Research Design

This study aims to predict insolvency of Finnish limited liability companies with ensemble machine learning models using time series of companies' financial statements. Model performance is assessed using time series of 3 and 5 consecutive financial statements. We define the prediction horizon as the length of the future time frame for which the models make insolvency predictions. While not all judicial corporate insolvency proceedings necessarily end in bankruptcy, insolvent companies will be referred to as bankrupt in this study for simplicity and consistency.

The general deadline for the filing of financial statements is 6 months [62]. However, for limited liability companies, this deadline is extended to 8 months. The Annual General Meeting, where financial statements are to be approved, must be held within 6 months of the end of the accounting period. [63] The approved financial statement must be filed with the Finnish Patent and Registration Office

(PRH) within two months of the approval. [64] Therefore, limited liability companies should be expected to have their financial statements available through PRH within 8 months of the end of the accounting period. Overdue filings are punishable by fines or de-registration of the company if the delay is longer than 1 year.

Therefore, we define the prediction horizon of 0 as financial statement sequences, whose final accounting period ends within 8 months of the official bankruptcy filing date. Although the financial statements in prediction horizon 0 end before the bankruptcy filing date, they were not necessarily publicly available before the filing. Models using such statements are therefore predicting an event that might have already happened, which is useless for practical bankruptcy forecasting. As such, any statements that end within this 8 month period will be excluded from any following prediction horizon, as well as any statements that end after the bankruptcy filing date.

After this 8 month period preceding bankruptcy, the following prediction horizons are defined using a 12-month interval. Prediction horizon 1 is composed of sequences with final statements ending between 8 and 20 months before bankruptcy filing. This translates to a practical forecasting window of 1 day to 20 months before bankruptcy. This is due to the variability in filing delays, as a statement ending 8 months and 1 day before bankruptcy, that takes 8 months to be filed with PRH, would be publicly available just 1 day before bankruptcy, while a statement ending 20 months prior, and hypothetically filed instantly, would already be publicly available 20 months before bankruptcy.

Each prediction horizon also includes all sequences from preceding prediction horizons, excluding horizon 0. If a company has multiple financial statements ending within the correct period before bankruptcy and a sufficiently long sequence of statements, it can be used to generate multiple bankrupt sequences. Given this setting, for prediction horizon X value, the model, in practice, answers the question:

'Based on this financial sequence, will bankruptcy occur within the next X accounting periods?'

3.2 Dataset Description

The dataset for this study consists of financial statements of Finnish companies from the interval between 2010 and 2025, which are publicly available for a fee through PRH. Financial statements are mostly only available in a PDF format, which means values have to be extracted through optical character recognition or similar means. All companies included in the dataset for this study are either public or private limited liability companies. A limited liability company is a business structure that separates the owner's personal assets from the company itself. This means that shareholders are not personally liable for their company's debts. A limited liability company cannot be informally closed. It must be closed through liquidation proceedings, bankruptcy or through a merger or demerger.

The full dataset contains a total of 2,229,512 financial statements from 327,509 companies, out of which 323,829 belong to the non-bankrupt class and 3680 belong to the bankrupt class. Each non-bankrupt company has on average roughly 7 financial statements and for bankrupt companies that number is roughly 5.5. Many feature columns contain a large number of null values, with most features having over 400,000. This is due to a combination of errors in optical character recognition, errors of companies in reporting their numbers, and the fact that many small companies do not need to report net values, only gross ones.

In addition to raw financial metrics extracted directly from the financial statements, common financial ratios are also calculated. The only non-numerical data point is the industry in which the company operates. The industries correspond to the TOL 2008 Standard Industrial Classification, used by the Finnish National Statistics Institution, Statistics Finland [65]. TOL 2008 is based on the EU's classifi-

cation of economic activities called NACE, which is in turn derived from the UN's International Standard Classification of All Economic Activities called ISIC [65]. The full industry classification code is a five-digit code, where each digit specifies a subcategory inside the given parent category. To control cardinality, only the first two digits of the industry codes will be used. In total, there are 83 industry codes present in the dataset at this level of cardinality.

Tables A.1 and A.2 in the appendix list the numerical features extracted from the financial statements. Variables from the income sheet, which provides information on the company's profit or loss are included in Table A.1. In Table A.2 are variables from the balance sheet, which show the company's assets and liabilities on the date of the filing. [10] Tables A.3 and A.4 contain common financial indicators, which are derived from the other features in the dataset. These indicators are split into two categories: profitability and solvency indicators in Table A.3 and liquidity, size, and efficiency indicators in Table A.4.

3.3 Data Considerations and Validity

All insolvent companies would necessarily fall under any definition of a company in financial distress. However, all companies in financial distress will not end in bankruptcy, as if there are no debts to default on, there can be no bankruptcy. Companies experiencing financial distress can cease operations without declaring bankruptcy if they do not have outstanding debts. Consequently, the non-bankrupt class in the dataset will inevitably hold companies that are "semantically" insolvent. Companies that mirror the economic characteristics of bankrupt companies where their liabilities are larger than their future revenues, but that might avoid official bankruptcy through means like equity financing, where a company raises money by selling shares in the business. Therefore, for all non-bankrupt companies, the final 3 financial statements will always be removed to reduce the amount of "semantically"

bankrupt companies in the non-bankrupt class.

The legislation stipulates that financial data should be reported up to the date of the bankruptcy filing, however these financial statements are either often not filed, do not find their way to the PRH database, are in such irregular forms that the optical character recognition algorithm routinely fails to extract the information from them or their dating is incorrect. [66] In the dataset for this study, out of all bankrupt companies, only roughly 10% of them have their final financial statement dated as ending within 3 months before, or any time after, the date the company filed for bankruptcy. Therefore, we cannot confidently assume that these statements precede immediate bankruptcy.

For the objectives of this study, we must be able to distinguish whether the dating of financial statements around bankruptcy is routinely incorrect or if the statements are simply missing. In Figure 3.1, we have the kernel density estimate (KDE) plots of operating profits of the final financial statements of bankrupt companies grouped by the relationship between the statement's end date and the official bankruptcy filing date. The operating profit from the final financial statement from non-bankrupt companies has also been plotted for reference. The KDE plot is a smoothed representation of the probability distribution of a continuous variable. For readability, only values that fall within the interquartile range (IQR), the range between the 1st and 3rd quartiles, have been included. The Figure shows that the final statements of companies that are further away from the true bankruptcy date tend to have higher operating profits. This demonstrates that while the dating of statements in relation to bankruptcies in the data does not follow the legislation, the relationship is still reliable, as further away statements clearly exhibit fewer signs of financial distress.

Given this observation, the remainder of the study will proceed under the assumption that for most bankrupt companies in the data, the financial statements from

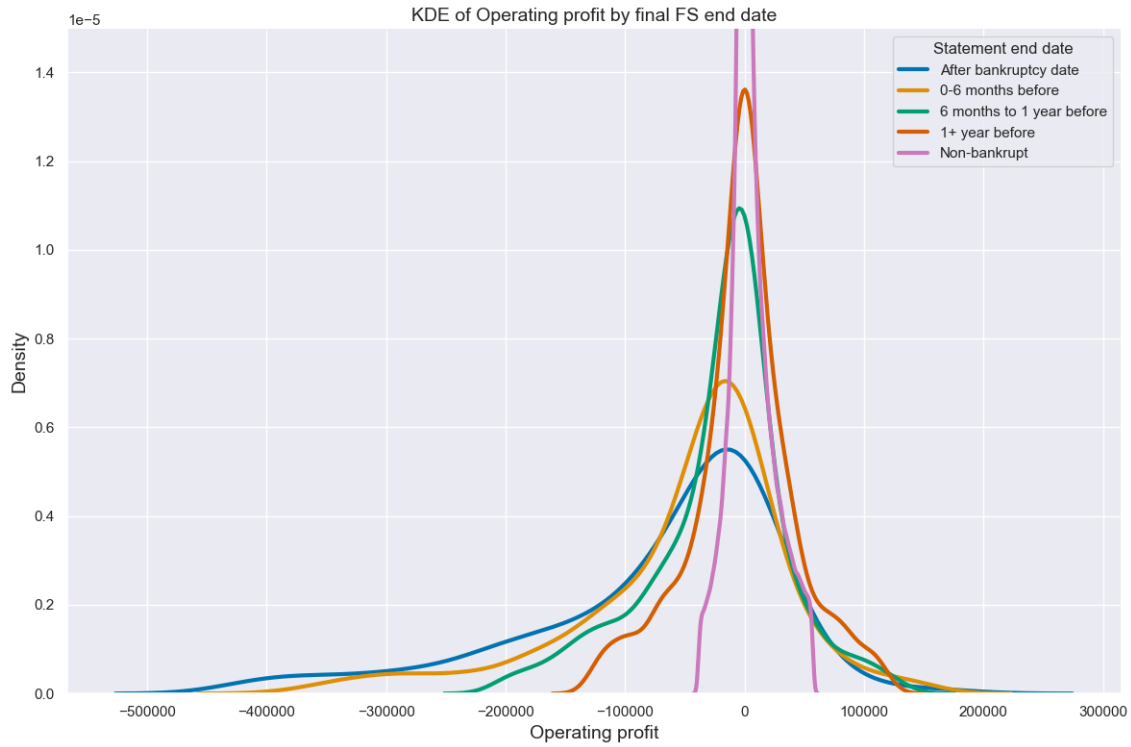

Figure 3.1: KDE Plot of Final FS Operating Profit IQR Grouped by End Date

accounting periods preceding the bankruptcy are simply missing. This drastically reduces the amount of bankrupt data that is available for the shorter prediction horizons.

3.4 Data Preprocessing

The main challenges in preprocessing for the dataset used in this study are gaps between the financial statements and varying lengths of accounting periods. Gaps must be handled to make sure the temporal relations are representative of the true distribution. Varying lengths of accounting periods affect the scaling of features that are dependent upon the length of the time series. Features such as Gross profit and Operating expenses are nearly monotonically increasing, meaning that a longer accounting period almost necessarily means a higher value. If financial statements are of varying lengths, the relation between features that are not dependent on the

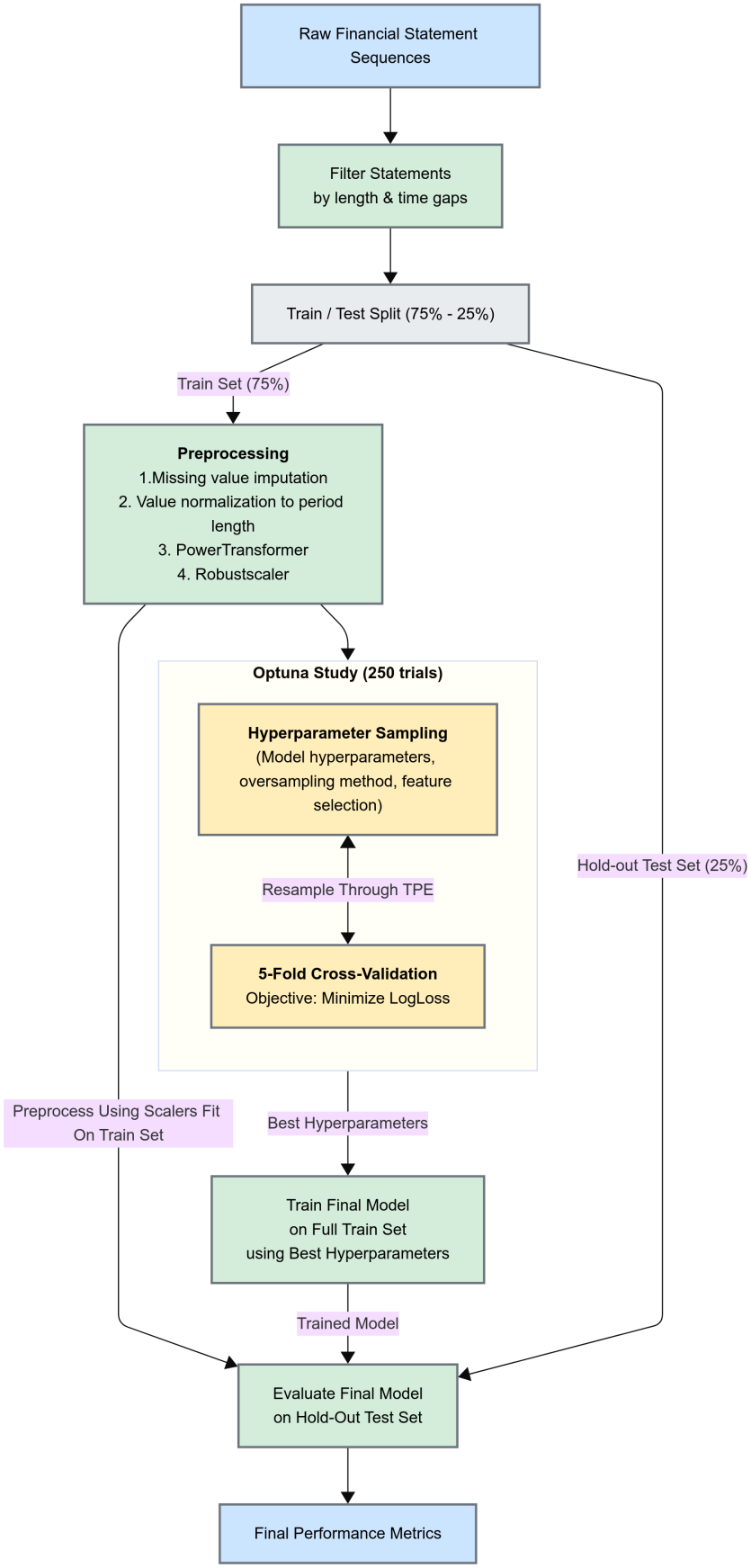

Figure 3.2: Data Processing and Model Evaluation Pipeline

length and those that are, will not be represented correctly.

Figure 3.2 visualizes the data preprocessing and model evaluation pipeline for a single prediction horizon and sequence length combination. The features from all financial statements of a company are first gathered into a time series. To handle the gaps, all financial statements that are prior to a gap in the time series that is longer than 360 days will be removed. Any statements that are after the gap will be retained. This will reduce the overall size of the dataset, but only a fraction of the removed statements are from sequences that would become bankrupt training examples in the end, as only the final statements of a bankrupt company can be used as bankrupt sequences.

The time series are then filtered based on whether there is enough financial statements remaining in the series after removing the gaps. For each sequence length, any company with less than the required sequence length number of statements remaining after gaps have been handled will be dropped from the dataset. If a bankrupt company has enough financial statements, it can be used to create a number of sequences corresponding to the prediction horizon value. For non-bankrupt companies, the final prediction horizon number of financial statements cannot be used, as its future bankruptcy status is indeterminate. For safety, an extra three financial statements from the end of the sequences are ignored to make sure the companies in the non-bankrupt class are representative of healthy companies. If the company has enough financial statements, it can be used to create multiple sequences, but for non-bankrupt companies, only non-overlapping sequences will be created, unlike bankrupt companies.

The dataset is then split into train and test sets using a stratified split, which preserves the original proportion of bankrupt companies in both the sets. During the split, due to resource limitations, the non-bankrupt class is also undersampled for more reasonable training times. The undersampling is done randomly into a dataset

with 5000 non-bankrupt companies. For consistency, the same seed is used each time, such that the non-bankrupt companies in both the train and the test sets remained fixed across all runs for a given sequence length and prediction horizon combination. However, the specific bankrupt samples, whether real or synthetically generated, could differ between the sequence length and prediction horizon combinations.

Next, the train set goes through preprocessing. Missing values are interpolated using linear interpolation. If the first financial statements of a sequence have missing values, they are filled with zeros and if a sequence ends with missing values, they are forward filled from previous ones. If all statements have a specific feature value missing, it is filled with zeros. After this, the features that are dependent on the length of the accounting period are normalized to align with a accounting period length of 365 days. This is done by dividing the feature value by the statements accounting period length in days and multiplying by 365. The features that go through this normalization are Revenue, Raw materials and consumables, External services, Raw materials and services total, Gross profit, Personnel costs, Depreciation, Operating expenses, Operating profit, Financial income and expenses, Extraordinary items, Appropriations, Income taxes and Net earnings.

All numerical features are then processed using a Yeo-Johnson power transformation to approximate normality, followed by RobustScaler from scikit-learn to mitigate the influence of outliers. The transformer and scaler are fit solely on the train set, and will be used to scale the final test set at the end of the pipeline.

While technically a preprocessing steps, the correlation-based filter feature selection method and the oversampling techniques are not applied before the Optuna study, as they are treated as hyperparameters. When a version of the dataset with feature selection or oversampling is required by the study, it will be applied dynamically inside the training loop.

The correlation-based feature selection method takes in candidate features sorted

by the amount of null values in ascending order and whenever a feature is accepted, any features which have a Pearson correlation coefficient that is above the threshold of 0.95 are removed from candidacy. The oversampling is done using one of the techniques: M-SMOTE and a modified version of baseline SMOTE. Synthetic samples are generated until the train set has a 1:1 ratio between non-bankrupt and bankrupt samples, so the absolute amount of synthetic samples depends on the prediction horizon and sequence length for that particular run.

3.5 Model Implementation and Training

Hyperparameter optimization is done on the training set using a 5-fold cross validation approach. The optimization employs Bayesian optimization via Optuna, a hyperparameter optimization framework for Python. For each trial in the Optuna study, hyperparameter values are sampled from the value ranges displayed in Tables 3.1, 3.2, and 3.3, using the Tree-structured Parzen Estimator (TPE) [67]. TPE is a model optimization algorithm based on modeling the probability distribution of optimal hyperparameter values, instead of modeling the objective function itself. In Optuna, the TPE sampler implements this algorithm by first running random trials, splitting them into high and low-performing subsets of hyperparameters and then probabilistically selecting them for future trials in a way that efficiently guides the hyperparameters towards optimal performance. Each Optuna study consisted of 250 trials, with the objective of minimizing the average logarithmic loss achieved across the five CV folds. All training and optimization procedures were executed on an RTX 4070S GPU and an R9 7900X CPU with the total time taken per Optuna study varying between 1 to 24 hours based on the model, sequence length and prediction horizon combination.

Upon completion of hyperparameter optimization, the configuration yielding the best average cross-validated performance is identified. A final model is then trained

using these optimal hyperparameters on the full training dataset. Before evaluation, the hold-out test set undergoes the same preprocessing steps as the training data, using transformers and scalers fitted exclusively on the training set to prevent data leakage. Feature selection is applied if it was part of the optimal configuration. Importantly, no oversampling is performed on the test data, regardless of whether it improved training performance, ensuring the test set remains representative of the true data distribution. The final model is then evaluated on this preprocessed hold-out set, yielding the final unbiased performance metrics

3.5.1 XGBoost

Table 3.1: XGBoost Hyperparameter Search Space

Hyperparameter	**Value Range**	**Scale**
learning_rate	$[0.01, 0.3]$	Logarithmic
max_depth	$[3, 10]$	Integer
min_child_weight	$[1, 10]$	Integer
gamma	$[0, 1.0]$	Linear
colsample_bytree	$[0.5, 1.0]$	Linear
subsample	$[0.5, 1.0]$	Linear
reg_alpha	$[10^{-8}, 5.0]$	Logarithmic
reg_lambda	$[10^{-8}, 5.0]$	Logarithmic
n_estimators	1000 (with early_stopping_rounds = 50)	-
Fixed Parameters		
scale_pos_weight	Calculated per fold (bankrupt/non-bankrupt ratio)	-
grow_policy	lossguide	-

Table 3.1 presents the search space used for XGBoost hyperparameters.

Learning Rate

The `learning_rate` is a common hyperparameter present in most machine learning models. It generally determines the magnitude of each change done to the model parameters during training, and in the context of gradient boosting models, it specifically controls the contribution that each tree makes to the final output. Lower

values therefore mean that more trees will be created.

Tree Structure Parameters

The hyperparameters controlling the structure of each generated tree are `max_depth`, `min_child_weight` and `gamma`. `Max_depth` sets a cap to the depth of each decision tree. `Min_child_weight` and `gamma` influence the splitting of the feature space at each tree. `Min_child_weight` sets the minimum sum of instance weight for each child node in a tree, which prevents the split from happening on highly specific feature values like outliers. `Gamma` controls the complexity of each tree by setting the minimum loss reduction required for each split, which prevent splits that do not improve the model enough to justify their addition to the model's complexity. [68]

Sampling Parameters

`Colsample_bytree` and subsample control fraction of features and observations randomly sampled for each tree, respectively. By controlling the number of features available for each split, `colsample_bytree` notably reduces the correlation between trees and increases the model's robustness against multicollinearity among the features. The subsample parameter helps to prevent overfitting by sampling the observations on each split, which is particularly relevant when dealing with imbalanced data. [68]

Regularization Parameters

`Reg_alpha` and `reg_lambda` are the L1 and L2 regularization terms. These terms also go by the names lasso regression and ridge regression, respectively. The purpose of regularization terms is to control model complexity, which is done by adding the penalty terms to the objective function. The penalty terms penalize large weight values in the leaf nodes, which serves a similar function to the `gamma` parameter by

producing simpler trees and reducing the chance of overfitting on small subsets of feature values. [68]

Estimator and early stopping

The `n_estimators` parameter sets the maximum number of trees to be built in the boosting process. `Early_stopping_rounds` stops the building of trees if the validation performance during cross-validation does not improve over the specified number of consecutive iterations, which in this case was set to 50. If this happens, the model .reverts to the version from 50 iterations ago. [68]

Class imbalance parameter

The `scale_pos_weight` parameter is used specifically in binary classification. It scales the gradient of samples belonging to each class based on the factor, which functionally increases the importance of bankrupt sequences. It is explicitly set as the ratio of bankrupt to non-bankrupt sequences in the training set of a given cross-validation fold, which is dependent upon whether either of the SMOTE implementations was applied on that particular set.

3.5.2 LightGBM

Table 3.2 presents the search space used for LightGBM hyperparameters. The parameters `learning_rate`, `num_iterations` and `early_stopping_rounds` function identically to their counterparts in XGBoost.

Tree Structure Parameters

Unlike XGBoost, LightGBM uses a leaf-wise tree growth strategy. The tree structure in LightGBM is governed by the parameters `num_leaves`, `min_data_in_leaf`, and `min_gain_to_split`. `Num_leaves` defines the maximum number of leaves for each

Table 3.2: LightGBM Hyperparameter Search Space

Hyperparameter	**Value Range**	**Scale**
learning_rate	$[0.01, 0.3]$	Logarithmic
num_leaves	$[16, 256]$	Integer
min_data_in_leaf	$[10, 100]$	Integer
lambda_l1	$[10^{-8}, 5.0]$	Logarithmic
lambda_l2	$[10^{-8}, 5.0]$	Logarithmic
min_gain_to_split	$[0.0, 1.0]$	Linear
bagging_fraction	$[0.5, 1.0]$	Linear
bagging_freq	$[1, 7]$	Integer
feature_fraction	$[0.5, 1.0]$	Linear
feature_fraction_bynode	$[0.5, 1.0]$	Linear
max_bin	$[50, 500]$	Integer
num_iterations	1000 (with early_stopping_rounds = 50)	-
Fixed Parameters		
max_depth	-1	-
boosting_type	gbdt	-

tree. `Min_data_in_leaf` controls the minimum number of data points required to form one leaf node, which serves a similar purpose to `min_child_weight` in XGBoost, but is instead based on the data point count rather than their weights. `Min_gain_to_split` sets the minimum loss reduction required for each split to limit complexity. Additionally, `max_depth`, which controls the maximum depth of each tree, is set at -1 to reduce the size of the search space, as tuning `num_leaves` already serves a similar purpose. [69]

Sampling Parameters

Notably, because we are not setting the `data_sample_strategy` parameter, Gradient-Based One-Side Sampling is not used. This is because using GOSS over the default random sampling strategy slowed down the fits significantly, which is not unexpected, as sampling based on the gradient is significantly less useful for smaller datasets. The sampling in LightGBM, when not using GOSS, is controlled by `bagging_fraction` and `bagging_freq`, which control the sampling of the data, as well as `feature_fraction` and `feature_fraction_bynode`, which control the sam-

pling of the features. `Bagging_fraction` specifies the fraction of data points sampled and `bagging_freq` specifies how often the sampling is performed, where a value of X means sampling is done every X iterations. `Feature_fraction` specifies the number of features selected each time a tree is built, equal to `colsample_bytree` in XGBoost. `Feature_fraction_bynode` expands on this by enabling feature sampling on a per-node level, not just a per tree level. `Max_bin` controls the maximum number of bins used in the histogram-based splitting algorithm of LightGBM. [69]

Regularization Parameters

`Lambda_l1` and `lambda_l2` implement L1 (lasso) and L2 (ridge) regularization, respectively. These terms penalize large leaf node values to control complexity and reduce overfitting, equal to reg_alpha and reg_lambda in XGBoost.

3.5.3 Random Forest

Table 3.3: Random Forest Hyperparameter Search Space

Hyperparameter	**Value Range**	**Scale**
n_estimators	$[100, 1000]$ (step 50)	Integer
max_depth	$[5, 50]$	Integer
min_samples_split	$[2, 20]$	Integer
min_samples_leaf	$[1, 20]$	Integer
max_features	$[0.1, 1.0]$	Linear
class_weight	{None, 'balanced'}	Categorical

Table 3.3 details the search space used for the Random Forest Classifier hyperparameters. Parameters `n_estimators`, `max_depth` function the same as their equivalents from XGBoost and LightGBM. `Class_weight` is equal to `scale_pos_weight` from XGBoost. [70]

Sampling Parameters

`Max_features` controls the number of features that are considered for each split. It takes a random sample of features for each split in the tree, whereas the equivalent parameters from XGBoost and LightGBM sample the features on a per tree basis. `Min_samples_split` controls the minimum number of samples present in a node to initiate further splitting and `min_samples_leaf` controls the minimum number of samples present in a leaf node of an already completed split to keep in the tree. [70]

4 Results

This chapter details the results of the experimental section of this study. The Chapter begins with tables for basic performance metrics for each sequence length and prediction horizon combination alongside the ROC curves. Then the effect of prediction horizon is further examined through plotting the test set XGBoost metrics grouped by prediction horizon. Feature importance is then reviewed using the built-in feature importance metrics for each model. Finally, the nature of the predictions is further analyzed through grouping the test set XGBoost metrics by the revenue of the companies.

4.1 Basic Metrics

Table 4.1 presents the basic model performance metrics for sequence length 3 across prediction horizons and Table 4.2 the metrics for sequence length 5. The best values for each metric are bolded. All models generally exhibit higher precision than recall for the bankrupt class. This suggests that while predictions of bankruptcy are often correct when made, the models tend to miss a significant number of actual bankruptcies. The more advanced XGBoost and LightGBM models produced better results than Random Forest across all sequence length and prediction horizon combinations. XGBoost appears to be the most robust and consistently best-performing model followed closely by LightGBM. Additionally, though not listed in the metrics, its training and inference times were on par with LightGBM. XGBoost

Table 4.1: Model Performance Comparison for Sequence Length 3 Across Prediction Horizons

Model	LogLoss	Brier	AUC ROC	Precision	Recall	F1-Score
Prediction Horizon 1						
XGBoost	**0.154**	**0.044**	**0.961**	**0.854**	**0.710**	**0.775**
LightGBM	0.156	0.045	0.960	0.838	0.702	0.764
Random Forest	0.186	0.053	0.948	0.838	0.655	0.735
Prediction Horizon 2						
XGBoost	**0.254**	**0.074**	0.946	**0.829**	0.752	0.789
LightGBM	**0.254**	0.075	**0.948**	0.823	**0.759**	**0.790**
Random Forest	0.283	0.086	0.932	0.802	0.701	0.748
Prediction Horizon 3						
XGBoost	**0.292**	**0.089**	**0.942**	0.798	**0.810**	**0.804**
LightGBM	0.294	**0.089**	0.941	**0.828**	0.781	**0.804**
Random Forest	0.315	0.097	0.932	0.810	0.756	0.782
Prediction Horizon 4						
XGBoost	**0.372**	**0.114**	**0.918**	0.798	**0.774**	**0.786**
LightGBM	0.380	0.115	**0.918**	**0.807**	0.752	0.779
Random Forest	0.385	0.122	0.899	0.789	0.726	0.756

Table 4.2: Model Performance Comparison for Sequence Length 5 Across Prediction Horizons

Model	LogLoss	Brier	AUC ROC	Precision	Recall	F1-Score
Prediction Horizon 1						
XGBoost	**0.139**	**0.040**	**0.971**	**0.861**	**0.759**	**0.807**
LightGBM	0.146	0.042	0.968	0.853	0.731	0.787
Random Forest	0.180	0.052	0.951	0.811	0.692	0.747
Prediction Horizon 2						
XGBoost	**0.227**	**0.066**	**0.959**	**0.865**	**0.793**	**0.827**
LightGBM	0.229	**0.066**	0.957	0.850	0.789	0.819
Random Forest	0.249	0.075	0.950	0.838	0.762	0.798
Prediction Horizon 3						
XGBoost	0.291	0.086	0.945	**0.844**	**0.807**	**0.825**
LightGBM	**0.287**	**0.085**	**0.947**	0.843	0.796	0.819
Random Forest	0.308	0.095	0.935	0.825	0.762	0.792
Prediction Horizon 4						
XGBoost	**0.319**	**0.097**	**0.939**	0.834	0.802	**0.818**
LightGBM	0.320	0.098	**0.939**	0.826	**0.807**	0.816
Random Forest	0.338	0.104	0.928	**0.837**	0.769	0.801

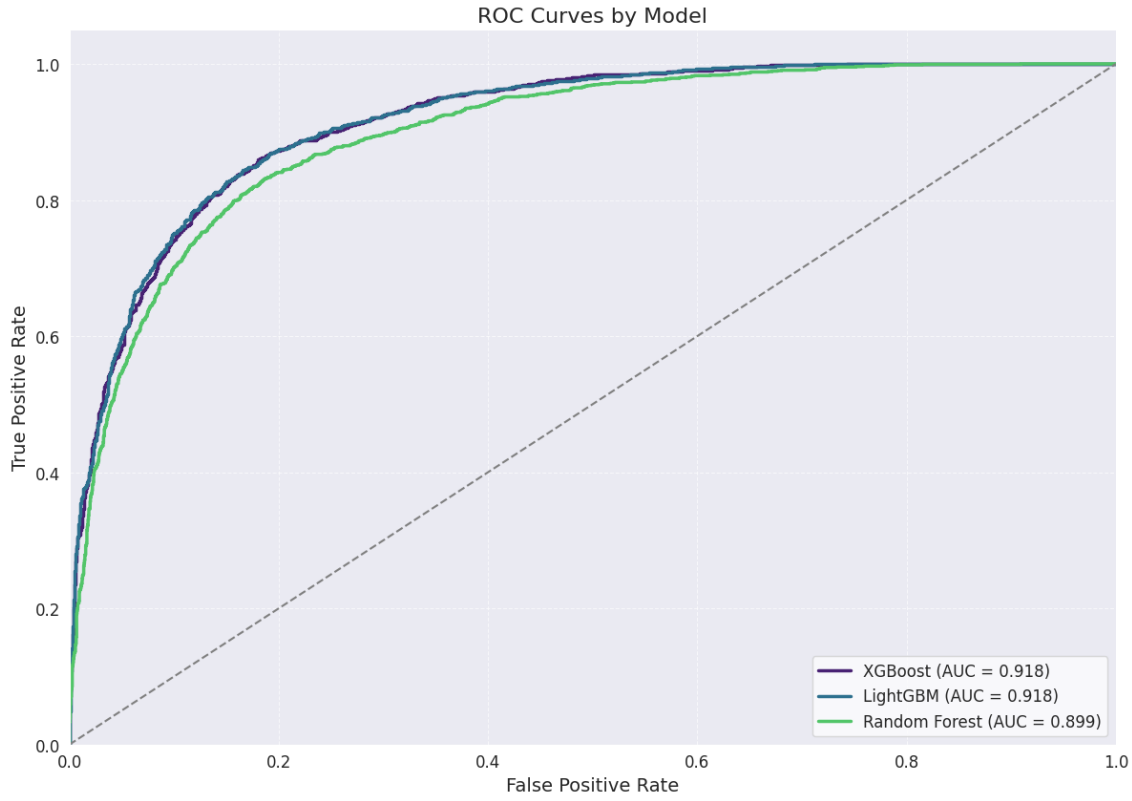

Figure 4.1: Receiver Operating Characteristic Curves

produced the most accurate predictions when the correlation based feature selection method was applied, whereas for LightGBM, the best metrics were from runs without any feature selection. All three models produced the best results when trained using M-SMOTE oversampled training data, though on rare occasions, the training data with no oversampling had stronger results.

Figure 4.1 depicts the ROC curves of each model for sequence length 3 and prediction horizon 4. This sequence length and prediction horizon combination is used because it has the largest dataset, which means it's likely to be the most robust in terms of the effect of singular test set samples. The ROC curves for XGBoost and LightGBM are nearly identical with Random Forest having slightly worse classification overall.

4.2 Sequence Length and Prediction Horizon

Figure 4.2 presents the logarithmic loss values of test set predictions generated by XGBoost, binned by the time difference in days between the bankruptcy filing and the end date of the last financial statement of the series with a two month interval between the bins. The same prediction horizon of 4 and sequence length of 3 is used. The purple line represents the average log loss value for bankrupt companies in the bin and the green line represents the overall log loss value for all non-bankrupt companies in the test set. Figure 4.3 shows the recall value for each bin in purple and the overall specificity in green.

Both figures reveal a consistent trend, the predictive performance of the model deteriorates as the time difference between the bankruptcy and the last financial statements increases. The performance metrics for the bankrupt class only begin to converge with the non-bankrupt line in the rightmost bins. An inflection point can be observed around the 1100 day, or approximately 3 year, mark. Predictions made with a prediction horizon shorter than 3 years begin to show marked improvements in their accuracy, whereas for horizons longer than 3 years, the accuracy of predictions remains relatively poor and there is little trend in the plots.

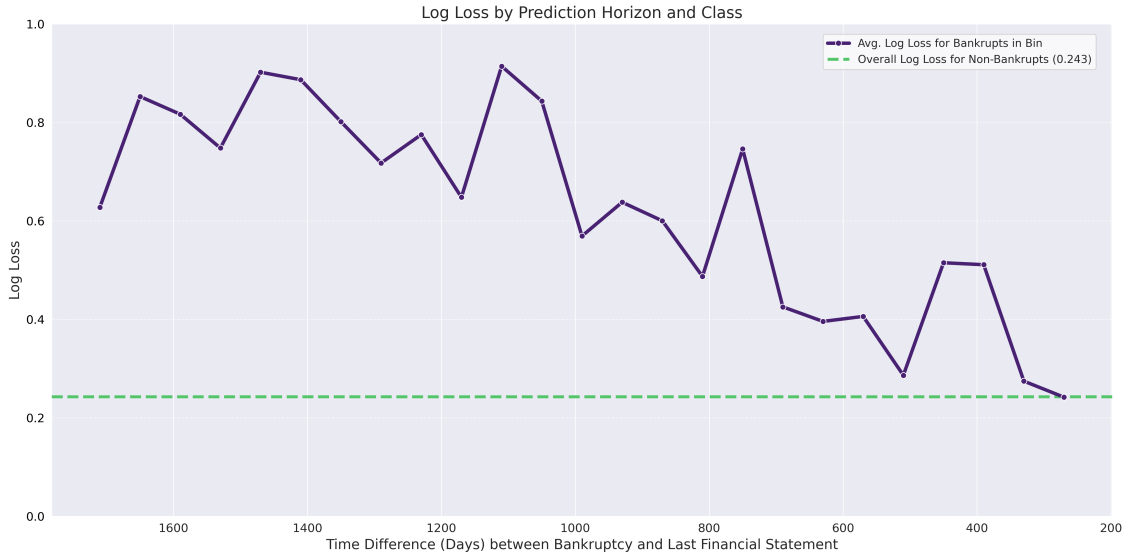

Figure 4.2: XGBoost Logarithmic Loss Grouped by Prediction Horizon for Sequence Length 3

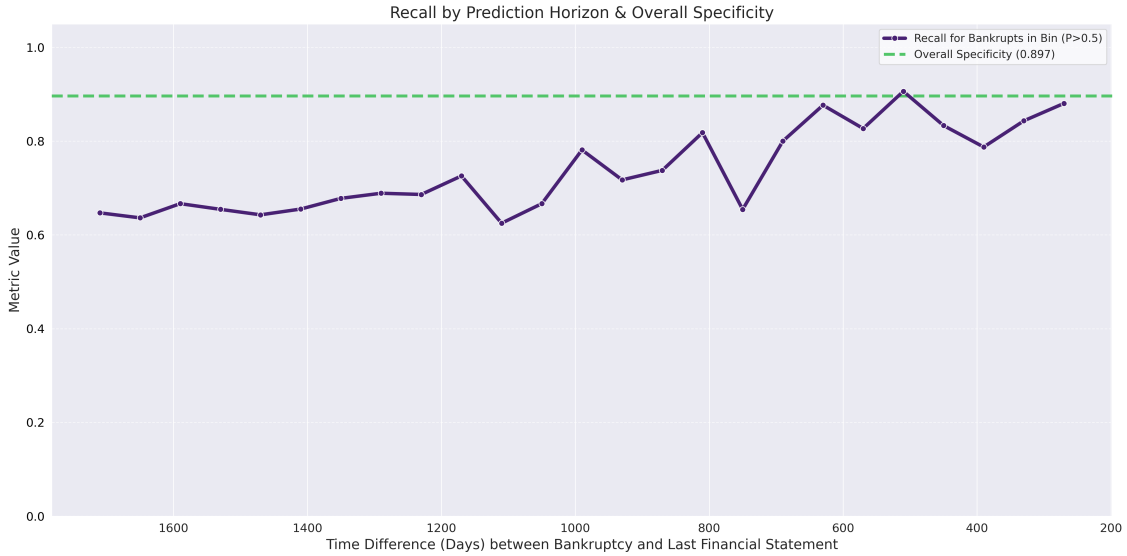

Figure 4.3: XGBoost Recall Grouped by Prediction Horizon for Sequence Length 3

4.3 Feature Importance

Figures 4.4, 4.5 and 4.6 visualize the test set's average feature importance across all hyperparameter combinations from sequence length 3 and prediction horizon 4. The feature importance metrics are extracted using the built-in feature importance capabilities of each of the models. For the boosting models, XGBoost and LightGBM, the importance score of a feature reflects how much the feature contributed to reducing the overall prediction error of the model. The more splits the feature is selected for and the more the error decreases when building the next tree with that split, the higher the feature importance score is going to be. In contrast, the feature importance values from Random Forest don't directly relate to the prediction errors, but rather how well the feature can be used to cleanly divide the data into distinct groups based on the bankruptcy label within a single tree. The more the feature contributes to reducing Gini impurity of the data, the higher the feature importance value.

Based on the figures, the features prioritized by LightGBM differ significantly from the other two models. Particularly, the categorical industry feature was weighted as considerably more influential than the other models. For each model, most of the important features come from the latest timestep, but features from earlier timesteps are also included. Financial income and expenses, Net financial expenses to revenue and Gross profit maintain a high importance score across all models.

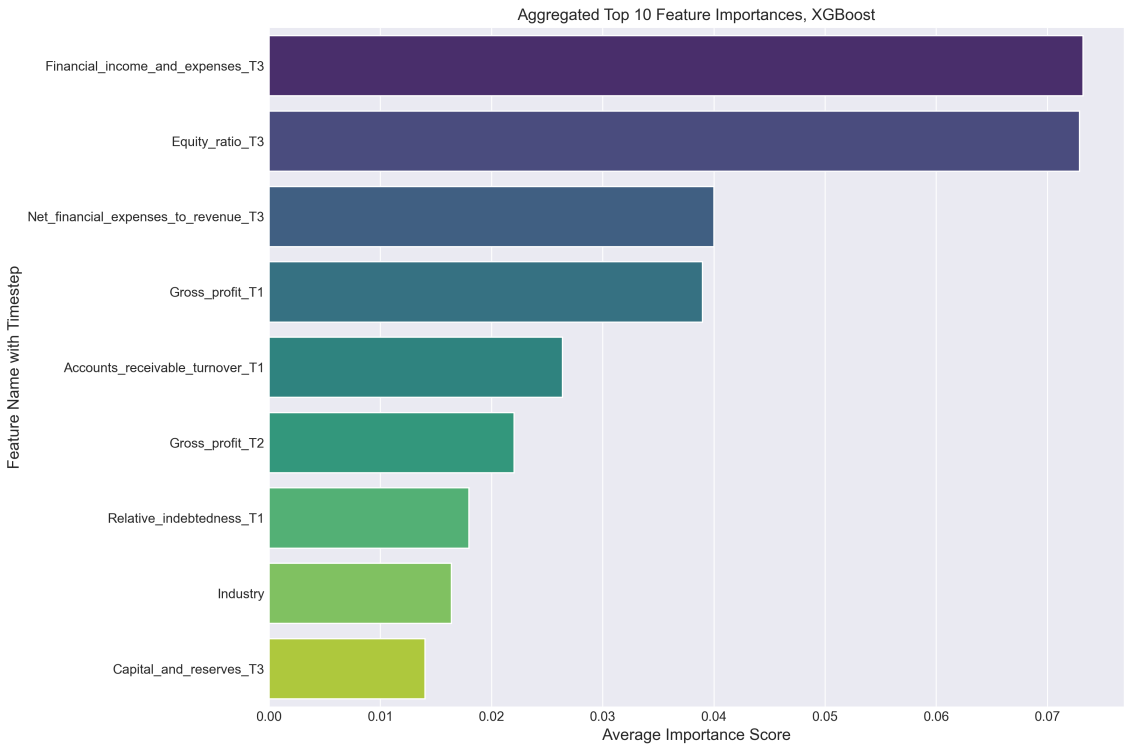

Figure 4.4: XGBoost Feature Importance

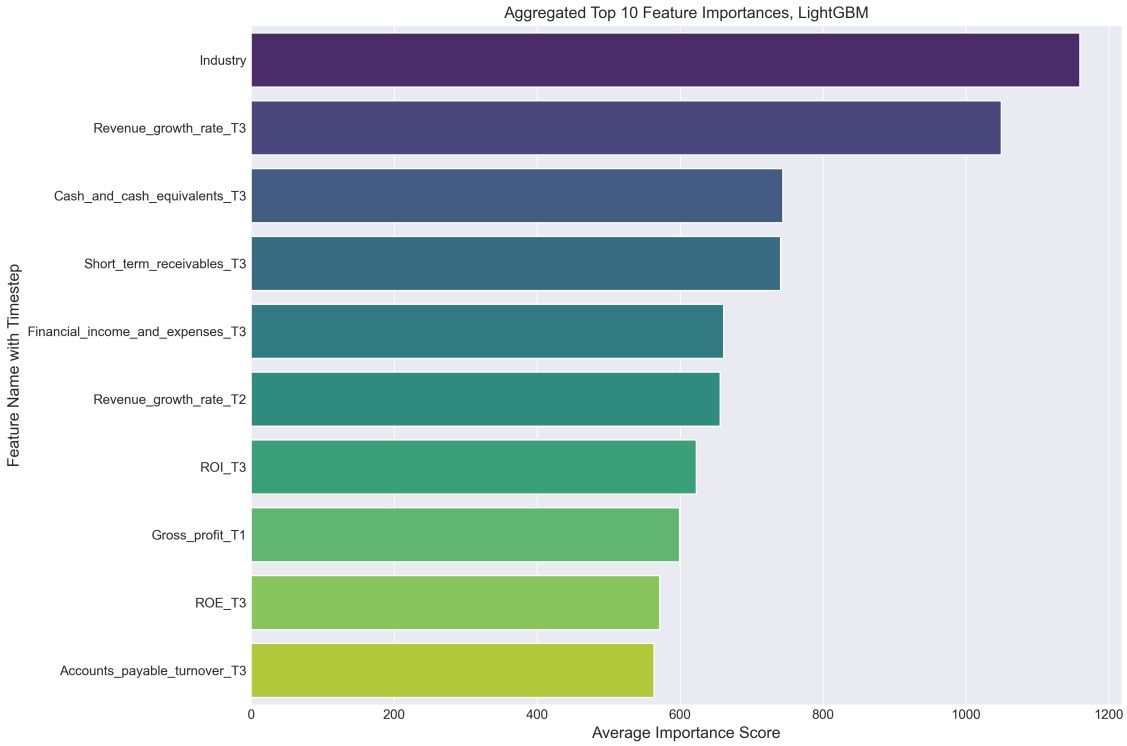

Figure 4.5: LightGBM Feature Importance

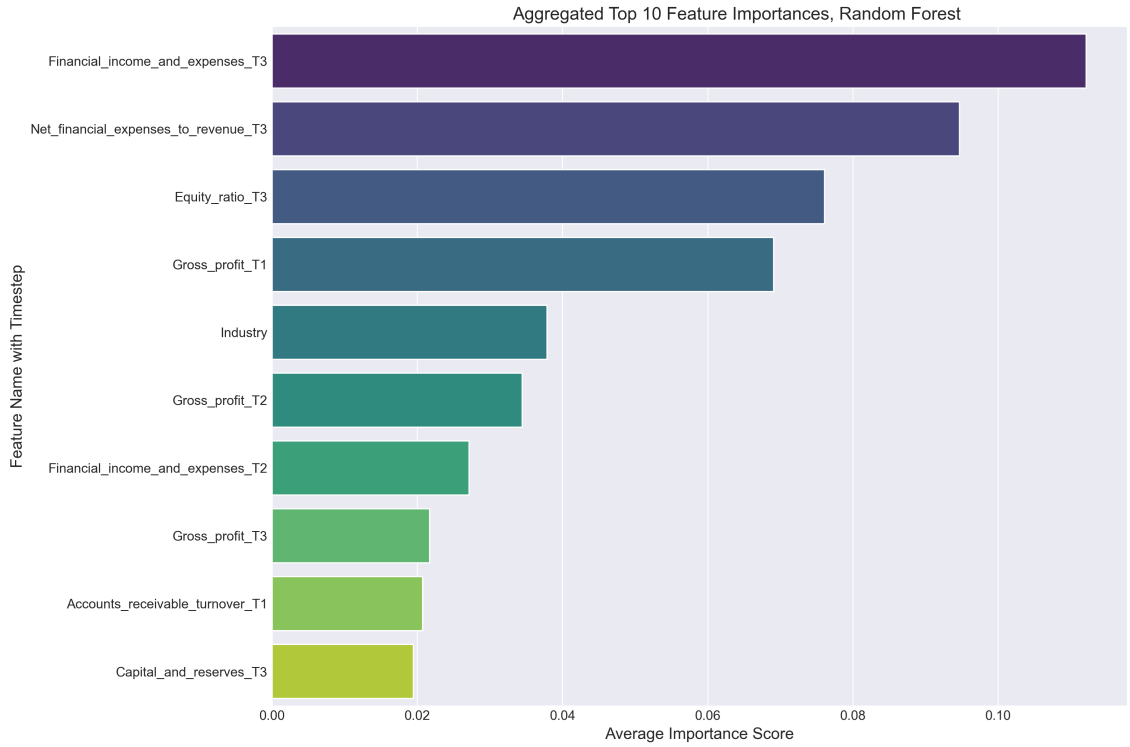

Figure 4.6: Random Forest Feature Importance

4.4 Further Analysis

To further analyze the prediction errors, Figure 4.7 visualizes the logarithmic loss values of test set predictions from XGBoost binned by the revenue value from the final financial statement of the sequence. The same prediction horizon of 4 and sequence length of 3 is used with the bankrupt predictions in purple and non-bankrupt predictions in green. It reveals that the lower the revenue, the more incorrect the average prediction is for the bankrupt class and vice versa for the non-bankrupt class. For larger companies, ones with over 3 million euros in revenue, the bankrupt predictions are more accurate on average than the non-bankrupt ones.

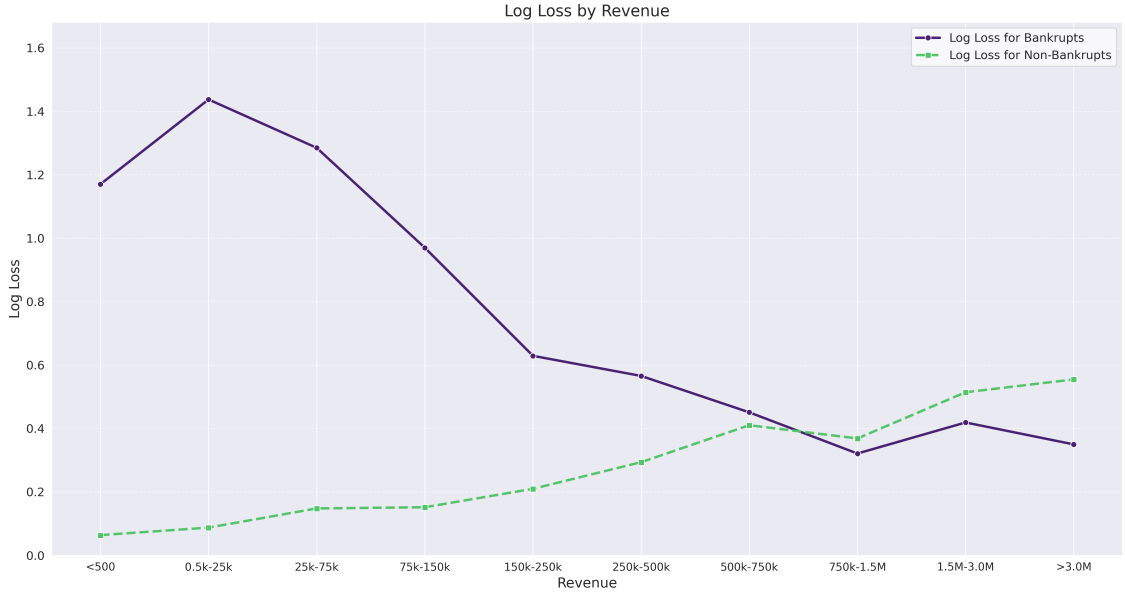

Figure 4.7: XGBoost Logarithmic Loss Grouped by Revenue

5 Discussion

In this chapter, the research questions are directly answered, limitations of the study are presented and directions for future research are suggested.

5.1 Research Question 1

Table 4.1 indicates XGBoost as the best performing model across prediction horizons and sequence lengths, followed closely by LightGBM. The difference between the two models is marginal, but XGBoost consistently produced slightly more accurate predictions. The ROC curves of XGBoost and LightGBM are also identical. Random Forest demonstrated significantly lower prediction performance across the board. The extent of difference is substantial between the boosting algorithms, XGBoost and LightGBM and the bagging algorithm, Random Forest, but only marginal between the two boosting algorithms themselves. Though not listed in the Table, XGBoost training and inference times were also on par with LightGBM. However, this is mostly due to the relatively small dataset. On a larger dataset, using Gradient-Based One-Side Sampling, LightGBM would most likely be considerably faster to train. Notably, on most sequence length and prediction horizon combinations, LightGBM produced better results without the correlation based feature selection method, whereas with XGBoost, the results were always stronger when feature selection was applied.

5.2 Research Question 2

Based on the metrics in Tables 4.1 and 4.2, it would seem that the longer sequence length of 5 financial statements is beneficial, with the metrics for sequence length 5 being better across the board than those of length 3. However, this comes with a caveat. The sequence length of 5 necessitates that all bankrupt companies must have a minimum of 5 financial statements filed. It is therefore possible that this small improvement in metrics is actually due to some more difficult 3 or 4 length sequences not being included in the test set. Younger companies are on average smaller than older ones, and Figure 4.7 indicates that for younger companies, the bankrupt class is more difficult to predict correctly. The feature importances in Figures 4.4, 4.5 and 4.6 contain features from timesteps other than the latest one, which means that the earlier timesteps do seem to have descriptive power. However, the average importance of features from the last two timesteps are significantly higher than for the earlier ones, with the features from the earliest timestep for sequence length 5 being selected for just a few splits in the total tree building process of both XGBoost and LightGBM. Due to these factors, it is unlikely that sequence lengths longer than 3 have much effect on the prediction performance.

The metrics for each prediction horizon in Tables 4.1 and 4.2 indicate that predicting bankruptcy further out is significantly harder. The models maintain a good AUC score, but the increasing Log Loss scores indicate a drop in the models ability to produce correct high confidence predictions. This also comes with the caveat, that the proportion of bankrupt to non-bankrupt companies in the test sets of longer prediction horizons is higher. Bankrupt companies are more difficult to correctly predict on average, which means including a higher proportion of them will lead to worse metrics even if the performance of the models remains the same. However, this is controlled for in the Figures 4.2 and 4.3. Based on the Figures, it is clear that the prediction of non-bankrupt companies is significantly more accurate

than the bankrupt ones, with the purple bankrupt line only converging with the non-bankrupt line at the rightmost bins. The apparent inflection point around the 3 year mark in the Figures also indicates that 3 years might be a cutoff point for the performance of the predictions, with a prediction horizon any longer than it producing consistently poor results. In short, increasing the length of the prediction has a negative effect on performance up to a three-year prediction, after which performance remains consistently poor.

Intuitively, this three year cut off point is sensible. Companies that eventually go bankrupt must almost necessarily exhibit losses in their primary business activities. The only way such companies can stay solvent is through already existing capital within the company or outside funding through credit. If predicting bankruptcy was reliably possible a long period into the future, it would be accounted for in the activities of creditors. Companies exhibiting features of this hypothetically predictable long term bankruptcy would then cease to get funding from creditors, causing them to go bankrupt sooner, until some equilibrium is reached. If long term bankruptcy was predictable and it was not accounted for by nearsighted creditors, it would lead to the creditors going consistently bankrupt, as credit would be given out to future insolvent companies.

5.3 Research Question 3

The feature importance Figures 4.4, 4.5 and 4.6 indicate the features prioritized by LightGBM differ significantly from the other two models. The categorical industry feature received a particularly higher importance score compared to the other two models. This indicates that LightGBM could be more capable of dealing with categorical features, though it did not produce better prediction results than XGBoost. LightGBM can handle categorical features directly using a special algorithm, whereas XGBoost and Random Forest require all input features to be numeric.

The highest importance features for each model are from the latest timestep, but features from earlier timesteps are also included in the Figures. Financial income and expenses, Net financial expenses to revenue and Gross profit are consistently identified as the most important features. While each of the three features relate to a company's profitability, they depict distinct aspects of its financial performance. Gross profit represents the profitability of a company's primary business activities whereas Financial income and expenses captures the impacts of non-operational financial activities like interest income or expenses and investment returns. Net financial expenses to revenue contextualizes the costs from these non-operational financial activities against sales. It represents how much of the company's generated revenue is consumed by financing related activities.

5.4 Limitations of the Study

A primary limitation of this study is its exclusive focus on Finnish limited liability companies. The dataset is specific to the Finnish accounting standards and regulatory filing deadlines. Therefore, the generalizability of the findings to other countries is not guaranteed. Different national regulations regarding financial reporting, economic conditions and even the general business cultures of countries could significantly alter which features are predictive and how well the models perform.

Another limiting factor discovered during the research is that the imbalance of the dataset does not only pertain to the binary bankruptcy label, but also to factors like the size of the companies. Figure 4.7 visualizes the impact of this through binning the Log Loss metric by revenue from the XGBoost model with prediction horizon 4 and sequence length 3. It reveals that the lower the revenue, the more incorrect the average prediction is for the bankrupt class and vice versa for the non-bankrupt class. While this does mean that the trained model struggles to correctly classify small bankrupt companies, it does not directly imply that small bankrupt companies are

more difficult to classify. The proportion of bankrupt to non-bankrupt companies in the first three bins is roughly 1:9, whereas in the last three bins it is 2:3. This does not mean that smaller companies are less likely to fail, but simply that the debts of low revenue companies are likely to also be smaller, which means that the ability for the debt to be handled outside of court through funds from family members, for example, is far higher. The models might learn lower revenue values as indicators of financial well-being, causing the bankrupt companies in that bin to get probability predictions that are consistently inaccurate. This shows a distinct limitation of handling the class imbalance by simply applying an oversampling method. The imbalance is more granular and can lead to undesirable results for subsets of companies if such a model is applied in real-life credit rating systems for example.

Finally, the models were trained on data from a specific economic period. Their predictive power may decrease in future if changes in the economic landscape alter the relationships between financial variables and bankruptcy risks. The economic period used in this study was not devoid of then unforeseen macroeconomic shocks. Primarily, the COVID-19 pandemic likely had a major impact on the bankruptcy risks of companies, but the war in Ukraine also impacted Finnish companies, especially ones in industries that did the most business with Russian companies, like energy and forestry. These factors were not controlled for in this study as all data from the 2010-2025 period was included.

Numerous studies have examined the effects of COVID-19 on company insolvency, however only those involving specifically Finnish companies provide relevant insights for the purposes of this study. In Finland, the legislation around civil and company insolvency was temporarily amended in the years following the onset of COVID-19. From May 1, 2020 to April 30, 2021, the ability for creditors to initiate corporate bankruptcies was restricted. Creditors could no longer demand the bankruptcy of a debtor for overdue payments from after March 2020. [71] This represents a significant

deviation from the norm and means that bankruptcies from this era might not be representative of the baseline economic landscape.

In [72], the characteristics of bankruptcies in Finland during COVID-19 were studied. It was concluded that there were differences in insolvency rates between the demographics of companies. Specifically, non-financial factors such as industry or location were highlighted as informative features. However, the analysis did not explicitly examine how COVID-19 influenced these insolvency patterns and whether the pandemic changed insolvency risks in the company demographics.

An additional consideration for this study is that when using, for example, a prediction horizon of 1 and a sequence length of 3, only non-bankrupt sequences that end in or before the year 2019 will be included in the data due to the filing lag and the removal of the three latest financial statements from non-bankrupt companies. This means that the non-bankrupt class did not include sequences that ended in the COVID-19 era, whereas the bankrupt class did. This might have had a negative effect on the prediction performance for the bankrupt class.

5.5 Future Research

The dataset for this study is relatively small, with the largest training set containing approximately 5000 non-synthetic bankrupt sequences As noted earlier, the distribution of bankrupt companies is also not evenly spread across characteristics like revenue. However, despite these issues, in this particular context, it is practically the definitive dataset. All financial statements filed by Finnish limited liability companies are available through PRH and this dataset holds all those that are feasibly extractable through OCR. Therefore, further research in this domain should examine employing non-financial data. This could include elements such as market sentiment through sentiment analysis from news articles or operational analysis through extracting textual features from director's reports, which are often con-

catenated to the financial statements filed to PRH. Additionally, future research should particularly investigate features that are publicly available immediately, unlike financial statements. An example of this is debt judgments available through the Finnish Legal Register Centre (ORK). Debt judgments are official court decisions legally confirming an outstanding debt and are typically made public shortly after the summary judgment is issued. The existence and size of such judgments could be indicative of potential bankruptcy risks. If so, they could provide crucial up-to-date information about the companies finances and help address the misalignment in the availability times of bankruptcy filings and financial statements.

A valuable extension to this research would also be a comparative study applying the models trained on companies of just one country to financial data from other countries to gauge the generalizability of such models. In the literature, models are most often trained and tested subsets of the same, or at least very similar datasets. However, a significant limitation in such study would be that rigorousness of financial statements can differ between countries. This means that the models would most likely have to be trained using just the variables that are filed in statements from both countries, which is less of a problem in the Nordic context, as the financial statement filing regulations are quite similar. The model trained in this study could therefore be applied on datasets from other Nordic countries to examine the extent to which bankruptcy risk drivers are universal versus region-specific.

6 Conclusion

This thesis investigated the prediction of bankruptcy for Finnish limited liability companies using ensemble machine learning models applied to financial time series data. The study focused on comparing model performance, the impact of input data sequence length and prediction horizon, and identifying key predictive financial features, while also addressing class imbalance through oversampling techniques. A key methodological consideration was the explicit handling of the disjointed public availability times of financial statements relative to bankruptcy filings, aiming for a more practically relevant assessment of predictive performance. The accurate and realistic assessment of insolvency risks is critical, as corporate failure carries substantial economic consequences.

In recent years, machine learning models, particularly ensemble methods, have gained popularity in corporate insolvency prediction and in the domain of time series forecasting in general. The ensemble boosting models, XGBoost and LightGBM, demonstrated superior predictive capabilities over Random Forest. XGBoost, in particular, consistently achieved the highest performance across various metrics, though the differences with LightGBM were often marginal. The study confirmed that shorter prediction horizons yield more accurate predictions, with a notable decline in predictive power beyond a three-year horizon. While the longer input sequence length of 5 offered slight improvements over sequence length 3, the most recent financial data proved most influential. Key financial indicators such as

Financial income and expenses, Net financial expenses to revenue, and Gross profit consistently emerged as important predictors. The application of M-SMOTE was beneficial in mitigating the class imbalance problem. However, a key limitation regarding dataset imbalances beyond the target variable was identified.

Future research could explore more sophisticated imbalance handling or integrating non-financial data. In the context of Finnish companies specifically, future research could assess the inclusion of immediately accessible indicators like debt judgments from the Finnish Legal Register Centre. Such data, available without the reporting lags inherent in financial statements, could further address the challenge of temporal misalignment and provide more up-to-date information, especially for companies that routinely file their financial statements late.

References

[1] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[2] G. Ke, Q. Meng, T. Finley, *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree", in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[3] L. Breiman, "Random Forests", *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324.

[4] K. Udofia, "Establishing Corporate Insolvency: The Balance Sheet Insolvency Test", Available at SSRN 3355248, 2019.

[5] Finlex, *Bankruptcy Act 120/2004*, Chapters 7-8, 2004. [Online]. Available: https://finlex.fi/fi/lainsaadanto/2004/120 (visited on 06/08/2025).

[6] Y.-C. Lee, "Application of Support Vector Machines to Corporate Credit Rating Prediction", *Expert Systems with Applications*, vol. 33, no. 1, pp. 67–74, 2007. DOI: 10.1016/j.eswa.2006.04.012.

[7] P. Gogas, T. Papadimitriou, and A. Agrapetidou, "Forecasting Bank Credit Ratings", *The Journal of Risk Finance*, vol. 15, no. 2, pp. 195–209, 2014. DOI: 10.1108/JRF-07-2013-0051.

[8] A. Usman, "Credit Ratings and Stock Price Crash Risk", *Applied Economics Letters*, vol. 31, no. 20, pp. 2199–2206, 2024. DOI: 10.1080/13504851.2023.2212267.

[9] U.S. Securities and Exchange Commission, *SEC Financial Report Manual, Topic 5: Smaller Reporting Companies*, 2022. [Online]. Available: https://www.sec.gov/about/divisions-offices/division-corporation-finance/financial-reporting-manual/frm-topic-5 (visited on 06/08/2025).

[10] Digital and Population Data Services Agency, *Financial Statements*, 2025. [Online]. Available: https://www.suomi.fi/company/financial-management-and-taxation/accounting-and-financial-management/guide/financial-statements-and-audit/financial-statements (visited on 06/08/2025).

[11] W. H. Beaver, "Financial Ratios As Predictors of Failure", *Journal of Accounting Research*, vol. 4, pp. 71–111, 1966. DOI: 10.2307/2490171.

[12] E. I. Altman, "Financial Ratios, Discriminant Analysis and Prediction of Corporate Bankruptcy", *The Journal of Finance*, vol. 23, no. 4, pp. 589–609, 1968. DOI: 10.1111/j.1540-6261.1968.tb00843.x.

[13] J. A. Ohlson, "Financial Ratios and the Probabilistic Prediction of Bankruptcy", *Journal of Accounting Research*, vol. 18, no. 1, pp. 109–131, 1980. DOI: 10.2307/2490395.

[14] M. E. Zmijewski, "Methodological Issues Related to the Estimation of Financial Distress Prediction Models", *Journal of Accounting Research*, vol. 22, pp. 59–82, 1984. DOI: 10.2307/2490859.

[15] J. Begley, J. Ming, and S. Watts, "Bankruptcy Classification Errors in the 1980s: An Empirical Analysis of Altman's and Ohlson's Models", *Review of Accounting Studies*, vol. 1, no. 4, pp. 267–284, 1996. DOI: 10.1007/BF00570833.

[16] F. Barboza, H. Kimura, and E. Altman, "Machine Learning Models and Bankruptcy Prediction", *Expert Systems with Applications*, vol. 83, pp. 405–417, 2017. DOI: 10.1016/j.eswa.2017.04.006.

[17] C. Clement, "Machine Learning in Bankruptcy Prediction–A Review", *Journal of Public Administration, Finance and Law*, no. 17, pp. 178–196, 2020.

[18] H. Kim, H. Cho, and D. Ryu, "Corporate Bankruptcy Prediction Using Machine Learning Methodologies with a Focus on Sequential Data", *Computational Economics*, vol. 59, no. 3, pp. 1231–1249, 2022. DOI: 10.1007/s10614-021-10126-5.

[19] X. Brédart, "Bankruptcy Prediction Model Using Neural Networks", *Accounting and Finance Research*, vol. 3, no. 2, pp. 124–128, 2014. DOI: 10.5430/afr.v3n2p124.

[20] S. Shetty, M. Musa, and X. Brédart, "Bankruptcy Prediction Using Machine Learning Techniques", *Journal of Risk and Financial Management*, vol. 15, no. 1, p. 35, 2022. DOI: 10.3390/jrfm15010035.

[21] P. Golbayani, I. Florescu, and R. Chatterjee, "A Comparative Study of Forecasting Corporate Credit Ratings Using Neural Networks, Support Vector Machines, and Decision Trees", *The North American Journal of Economics and Finance*, vol. 54, p. 101 251, 2020. DOI: 10.1016/j.najef.2020.101251.

[22] A. Dasilas and A. Rigani, "Machine Learning Techniques in Bankruptcy Prediction: A Systematic Literature Review", *Expert Systems with Applications*, vol. 255, 2024. DOI: 10.1016/j.eswa.2024.124761.

[23] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M4 Competition: 100,000 Time Series and 61 Forecasting Methods", *International Journal of Forecasting*, vol. 36, no. 1, pp. 54–74, 2020. DOI: 10.1016/j.ijforecast.2019.04.014.

[24] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "M5 Accuracy Competition: Results, Findings, and Conclusions", *International Journal of Forecasting*, vol. 38, no. 4, pp. 1346–1364, 2022. DOI: 10.1016/j.ijforecast.2021.11.013.

[25] S. Makridakis, E. Spiliotis, R. Hollyman, F. Petropoulos, N. Swanson, and A. Gaba, "The M6 Forecasting Competition: Bridging the Gap Between Forecasting and Investment Decisions", *International Journal of Forecasting*, 2024. DOI: 10.1016/j.ijforecast.2024.11.002.

[26] W. Liu, Y. Suzuki, and S. Du, "Ensemble Learning Algorithms Based on Easyensemble Sampling for Financial Distress Prediction", *Annals of Operations Research*, 2025. DOI: 10.1007/s10479-025-06494-y.

[27] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting", *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[28] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine", *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001. DOI: 10.1214/aos/1013203451.

[29] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. DOI: 10.1613/jair.953.

[30] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning", *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 2, pp. 539–550, 2008.

[31] C. Cortes and V. Vapnik, "Support-vector networks", *Machine learning*, vol. 20, pp. 273–297, 1995.

[32] C.-F. Tsai, K.-L. Sue, Y.-H. Hu, and A. Chiu, "Combining Feature Selection, Instance Selection, and Ensemble Classification Techniques for Improved Financial Distress Prediction", *Journal of Business Research*, vol. 130, pp. 200–209, 2021. DOI: 10.1016/j.jbusres.2021.03.018.

[33] G. Lombardo, M. Pellegrino, G. Adosoglou, S. Cagnoni, P. M. Pardalos, and A. Poggi, "Machine Learning for Bankruptcy Prediction in the American Stock Market: Dataset and Benchmarks", *Future Internet*, vol. 14, no. 8, pp. 244–, 2022. DOI: 10.3390/fi14080244.

[34] A. El-Qadi, M. Trocan, T. Frossard, and N. Díaz-Rodríguez, "Credit Risk Scoring Forecasting Using a Time Series Approach", in *Proceedings of the 41st International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, 2022, p. 16. DOI: 10.3390/psf2022005016.

[35] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[36] B. Siswoyo, Z. A. Abas, A. N. C. Pee, R. Komalasari, and N. Suyatna, "Ensemble Machine Learning Algorithm Optimization of Bankruptcy Prediction of Bank", *IAES International Journal of Artificial Intelligence*, vol. 11, no. 2, pp. 679–688, 2022. DOI: 10.11591/ijai.v11.i2.pp679-688.

[37] X. Tang, S. Li, M. Tan, and W. Shi, "Incorporating Textual and Management Factors into Financial Distress Prediction: A Comparative Study of Machine

Learning Methods", *Journal of Forecasting*, vol. 39, no. 5, pp. 769–787, 2020. DOI: 10.1002/for.2657.

[38] T. M. Alam, K. Shaukat, M. Mushtaq, *et al.*, "Corporate Bankruptcy Prediction: An Approach Towards Better Corporate World", *The Computer Journal*, vol. 64, no. 11, pp. 1731–1746, 2021. DOI: 10.1093/comjnl/bxab003.

[39] J. Liu, C. Li, P. Ouyang, J. Liu, and C. Wu, "Interpreting the Prediction Results of the Tree-Based Gradient Boosting Models for Financial Distress Prediction with an Explainable Machine Learning Approach", *Journal of Forecasting*, vol. 42, no. 5, pp. 1112–1137, 2023. DOI: 10.1002/for.2929.

[40] G. Perboli and E. Arabnezhad, "A Machine Learning-Based DSS for Mid and Long-Term Company Crisis Prediction", *Expert Systems with Applications*, vol. 174, p. 114 758, 2021. DOI: 10.1016/j.eswa.2021.114758.

[41] O. Liashenko, T. Kravets, and Y. Kostovetskyi, "Machine Learning and Data Balancing Methods for Bankruptcy Prediction", *Ekonomika*, vol. 102, no. 2, pp. 28–46, 2023. DOI: 10.15388/Ekon.2023.102.2.2.

[42] T. Le, "A Comprehensive Survey of Imbalanced Learning Methods for Bankruptcy Prediction", *IET Communications*, vol. 16, no. 5, pp. 433–441, 2022. DOI: 10.1049/cmu2.12357.

[43] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning", in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, Ieee, 2008, pp. 1322–1328.

[44] E. Nylén-Forthun, M. Møller, and N.-G. B. Abrahamsen, "Financial Distress Prediction Using Machine Learning and XAI: Developing an Early Warning Model for Listed Nordic Corporations", M.S. thesis, Norwegian University of Science and Technology (NTNU), 2022.

[45] O. W. Bøe-Waal, "The Applicability of Machine Learning Models in Corporate Bankruptcy Prediction", M.S. thesis, Copenhagen Business School, 2023. [Online]. Available: https://research-api.cbs.dk/ws/portalfiles/portal/98729089/1593199_Master_thesis_Final_15_05.pdf (visited on 06/08/2025).

[46] M. Oinonen, "Classification of Financial Distress in Nasdaq Helsinki Companies Using Decision Tree", M.S. thesis, LUT University, 2024. [Online]. Available: https://urn.fi/URN:NBN:fi-fe2024061048453 (visited on 06/08/2025).

[47] F. Paraschiv, M. Schmid, and R. R. Wahlstrøm, "Bankruptcy Prediction of Privately Held SMEs Using Feature Selection Methods", Available at SSRN 4588767, 2023.

[48] L. Soares de Melo Junior, F. M. Nardini, C. Renso, and J. A. Fernandes de Macêdo, "An Empirical Comparison of Classification Algorithms for Imbalanced Credit Scoring Datasets", in *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2019, pp. 747–754. DOI: 10.1109/ICMLA.2019.00133.

[49] T. Cover and P. Hart, "Nearest neighbor pattern classification", *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[50] H. Qu and Z. Zhang, "A Time Series Data Augmentation Method based on SMOTE", in *2024 36th Chinese Control and Decision Conference (CCDC)*, 2024, pp. 5336–5341. DOI: 10.1109/CCDC62350.2024.10587816.

[51] B. V. Dasarathy and B. V. Sheela, "A Composite Classifier System Design: Concepts and Methodology", *Proceedings of the IEEE*, vol. 67, no. 5, pp. 708–713, 1979. DOI: 10.1109/PROC.1979.11327.

[52] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting", *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997. DOI: 10.1006/jcss.1997.1504.

[53] L. Breiman, "Bagging Predictors", *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996. DOI: 10.1007/BF00058655.

[54] R. E. Schapire, "The Boosting Approach to Machine Learning: An Overview", in *Nonlinear Estimation and Classification*, D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, and B. Yu, Eds., Springer, 2003, pp. 149–171. DOI: 10.1007/978-0-387-21579-2_9.

[55] P. Bühlmann and B. Yu, "Analyzing Bagging", *The Annals of Statistics*, vol. 30, no. 4, pp. 927–961, 2002. DOI: 10.1214/aos/1031689014.

[56] S. Nembrini, I. R. König, and M. N. Wright, "The Revival of the Gini Importance?", *Bioinformatics*, vol. 34, no. 21, pp. 3711–3718, 2018. DOI: 10.1093/bioinformatics/bty373.

[57] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Wadsworth, 1984, ISBN: 0-534-98053-8.

[58] O. Rainio, J. Teuho, and R. Klén, "Evaluation Metrics and Statistical Tests for Machine Learning", *Scientific Reports*, vol. 14, no. 1, p. 6086, 2024. DOI: 10.1038/s41598-024-56541-1.

[59] J. A. Hanley and B. J. McNeil, "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve", *Radiology*, vol. 143, no. 1, pp. 29–36, 1982. DOI: 10.1148/radiology.143.1.7063747.

[60] M. Grandini, E. Bagli, and G. Visani, *Metrics for Multi-Class Classification: An Overview*, 2020. arXiv: 2008.05756.

[61] K. Rufibach, "Use of Brier Score to Assess Binary Predictions", *Journal of Clinical Epidemiology*, vol. 63, no. 8, pp. 938–939, 2010. DOI: 10.1016/j.jclinepi.2009.11.009.

[62] Finlex, *Finnish Accounting Act 1336/1997*, Chapter 3, Section 9, 1997. [Online]. Available: https://www.finlex.fi/en/legislation/1997/1336 (visited on 06/08/2025).

[63] Finlex, *Limited Liability Companies Act 624/2006*, Chapter 5, Section 3, 2006. [Online]. Available: https://www.finlex.fi/fi/lainsaadanto/2006/624 (visited on 06/08/2025).

[64] Finlex, *Limited Liability Companies Act 624/2006*, Chapter 8, Section 10, 2006. [Online]. Available: https://www.finlex.fi/fi/lainsaadanto/2006/624 (visited on 06/08/2025).

[65] Statistics Finland, *Standard Industrial Classification TOL 2008*, 2008. [Online]. Available: https://stat.fi/en/luokitukset/toimiala/toimiala_1_20080101 (visited on 06/08/2025).

[66] Finlex, *Bankruptcy Act 120/2004*, Chapter 9, Section 3, 2004. [Online]. Available: https://finlex.fi/fi/lainsaadanto/2004/120 (visited on 06/08/2025).

[67] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for Hyper-Parameter Optimization", in *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, Eds., vol. 24, 2011, pp. 2546–2554.

[68] XGBoost Developers, *XGBoost Parameters*, 2022. [Online]. Available: https://xgboost.readthedocs.io/en/stable/parameter.html (visited on 06/08/2025).

[69] Microsoft Corporation, *LightGBM Parameters*, 2025. [Online]. Available: https://lightgbm.readthedocs.io/en/stable/Parameters.html#parameters (visited on 06/08/2025).

[70] scikit-learn Developers, *sklearn.ensemble.RandomForestClassifier*, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (visited on 06/08/2025).

[71] Finlex, *Hallituksen esitys eduskunnalle laiksi konkurssilain väliaikaisesta muuttamisesta*, 2020. [Online]. Available: https://finlex.fi/fi/hallituksen-esitykset/2020/46 (visited on 06/08/2025).

[72] L. Ervo, "Insolvency Law and COVID-19: The Finnish Example on Tackling the Pandemic", in *Finance, Law, and the Crisis of COVID-19: An Interdisciplinary Perspective*, Springer, 2022, pp. 123–137. DOI: 10.1007/978-3-031-01127-9_8.

Appendix

Table A.1: Income Sheet Variables

Variable	**Description**
Revenue	The total income generated from the company's principal sales activities before any cost deductions.
Raw materials and consumables	Cost of materials and supplies utilized in manufacturing.
External services	Costs of services procured from third-party providers.
Raw materials and services total	Sum of costs from manufacturing of goods or provision of services.
Gross profit	Difference between total revenue and cost of production.
Personnel costs	Expenditure in employee salaries, wages and benefits.
Depreciation	Decline in value of a tangible fixed assets.
Operating expenses	Expenditures from business activities not directly tied to production, e.g., leasing costs, or merger-related losses.
Operating profit (EBIT)	Profit from a company's core business operations.
Financial income and expenses	Net earnings from returns on investments, shareholdings, interest and debt servicing costs.
Extraordinary items	Net earnings from activities that are not part of its regular business operations.
Appropriations	Distribution of net earnings for purposes such as dividend payments.
Income taxes	Amount of taxes paid for net earnings.
Net earnings	Total earnings after all operating expenses, taxes, and other costs have been subtracted from gross profit.

Table A.2: Balance Sheet Variables

Variable	**Description**
Intangible assets	Non-physical assets that hold value, e.g., patents or trademarks.
Tangible assets	Physical assets, leases and company shares.
Investments	Expenditure on procuring long-term assets.
Short-term receivables	Debts owed to the company by other parties that are due within one year of statement date.
Long-term receivables	Debts owed to the company by other parties that are due more than one year from statement date.
Securities	Financial instruments owned by the company.
Cash and cash equivalents	Cash or highly liquid assets.
Non-current assets	Long-term assets used in business operations e.g., property or equipment.
Current assets	Short-term assets used in daily operations, e,g, cash or inventory.
Retained earnings	Net earnings less paid out dividends.
Subscribed capital	Value of shares shareholders have formally committed to buy.
Other reserves	Funds not subject to distribution to shareholders.
Capital and reserves (Equity)	The net assets after paying off all creditors.
Long-term liabilities	Debts owed by the company by other parties that are due within one year of statement date.
Short-term liabilities	Debts owed by the company by other parties that are due more than one year from statement date.
Liabilities	Liabilities in total.
Total assets	Total net worth of the company.

Table A.3: Profitability and Solvency Indicators

Indicator	**Derivation**
Gross margin	$\frac{\text{Gross profit}}{\text{Revenue}}$
EBITDA	Operating profit − Amortization and depreciation
EBITDA margin	$\frac{\text{EBITDA}}{\text{Revenue}}$
Operating profit margin	$\frac{\text{Operating profit}}{\text{Revenue}}$
Free cash flow	Net earnings+Amortization and depreciation
Free cash flow to revenue	$\frac{\text{Free cash flow}}{\text{Revenue}}$
Net profit margin	$\frac{\text{Net earnings}}{\text{Revenue}}$
Return on equity (ROE)	$\frac{\text{Net earnings}}{\text{Average equity}}$
Return on investment (ROI)	$\frac{\text{Net earnings} + \text{Financial expenses} + \text{Taxes}}{\text{Average investments}}$
Return on assets (ROA)	$\frac{\text{Net earnings} + \text{Financial expenses} + \text{Taxes}}{\text{Average assets}}$
Equity ratio	$\frac{\text{Interest}}{\text{Total assets} - \text{Noncurrent assets}}$
Net gearing	$\frac{\text{Interest bearing debt} - \text{Cash and cash equivalents}}{\text{Total capital}}$
Relative indebtedness	$\frac{\text{Liabilities}}{\text{Revenue}}$
Net financial expenses to revenue	$\frac{\text{Financial expenses}}{\text{Revenue}}$
Total liabilities payback period	$\frac{\text{Interest bearing debt}}{\text{Free cash flow}}$

Table A.4: Liquidity, Size, and Efficiency Indicators

Indicator	**Derivation / Description**
Quick ratio	$\frac{\text{Short-term receivables} + \text{Cash \& equiv.} + \text{Long-term receivables}}{\text{Short-term liabilities}}$
Current ratio	$\frac{\text{Inventories} + \text{Short-term receivables} + \text{Cash \& equiv.} + \text{Long-term receivables}}{\text{Short-term liabilities}}$
Value added	Operating profit − Sales profit from fixed assets + Amortization and depreciation + Personnel costs
Value added margin	$\frac{\text{Value added}}{\text{Revenue}}$
Revenue growth rate	$\frac{\text{Revenue}_{\text{current}} - \text{Revenue}_{\text{previous}}}{\text{Revenue}_{\text{previous}}}$
Invoicing	Revenue + Short-term advances received − Previous long-term advances received
Personnel growth rate	$\frac{\text{Number of employees}_{\text{current}} - \text{Number of employees}_{\text{previous}}}{\text{Number of employees}_{\text{previous}}}$
Value added per personnel costs	$\frac{\text{Value added}}{\text{Personnel costs}}$
Working capital	Inventories + Trade receivables + Trade receivables from related parties − Trade payables − Advances received
Working capital to revenue	$\frac{\text{Working capital}}{\text{Revenue}}$
Net working capital	Cash & cash equivalents + Securities + Short-term receivables + Inventory − Short-term liabilities
Net working capital to revenue	$\frac{\text{Net working capital}}{\text{Revenue}}$
Inventories turnover rate	$\frac{\text{Inventories} - \text{Advance payments of inventories}}{\text{Raw materials and consumables}}$
Accounts receivable turnover	$\frac{\text{Trade receivables}}{\text{Revenue}}$
Accounts payable turnover	$\frac{\text{Accounts payable}}{\text{Raw materials and consumables}}$
Capital turnover rate	$\frac{\text{Revenue}}{\text{Average total assets}}$
Number of employees	Raw count of employees for the period.
Industry	First two digits of TOL 2008 Standard Industrial Classification.