Degree project in Technology

First cycle, 15 credits

ESG Integrated Machine Learning for Credit Risk Assessment

SOPHIA OLSSON & TUVA GUNNARSSON

ESG Integrated Machine Learning for Credit Risk Assessment

ESG-integrerad maskininlärning för kreditriskbedömning

Sophia Olsson and Tuva Gunnarsson

***Abstract*—Credit risk assessment is often based on statistical models using financial indicators, but such models may have limited ability to capture nonlinear relationships between financial and sustainability related variables. This thesis investigates whether the integration of Environmental, Social, and Governance (ESG) factors into machine learning models improves corporate credit risk prediction within the Nordic corporate bond market.**

The study combines financial and ESG data for listed Nordic bond issuers and evaluates Logistic Regression, Naive Bayes, Random Forest, and XGBoost across several feature configurations. Credit ratings are mapped into three risk classes, and model performance is assessed using Macro-F1 and ROC-AUC over repeated train-test splits. The results show that tree-based ensemble models outperform the statistical baselines. The best performing configuration is a Random Forest model using ESG and enriched financial features, achieving a mean Macro-F1 score of 0.655 and a mean ROC-AUC of 0.862. Adding ESG variables improves performance for several models, although the improvement is moderate relative to the variation across splits.

The findings suggest that ESG variables can provide complementary predictive information when combined with traditional financial indicators. Feature importance analysis indicates that leverage and repayments capacity measures remain central, while ESG related variables such as carbon intensity and board composition add additional signal. However, the study is limited by sample size, class imbalance, missing value imputation, and potential point-in-time leakage in enriched financial variables. Further research should evaluate ESG integrated credit risk models using larger and time aligned datasets.

***Sammanfattning*—Kreditriskbedömning baseras ofta på statistiska modeller som använder finansiella indikatorer, men sådana modeller kan ha begränsad förmåga att fånga ickelinjära samband mellan finansiella och hållbarhetsrelaterade variabler. Detta arbete undersöker om miljömässiga, sociala och bolagsstyrningsrelaterade faktorer, ESG, kan förbättra maskininlärningsbaserad klassificering av kreditbetyg, som används som en approximation för företags kreditrisk, på den nordiska företagsobligationsmarknaden.**

Studien kombinerar finansiell data och ESG-data för noterade nordiska företagsobligationsutgivare och utvärderar Logistisk Regression, Naive Bayes, Random Forest och XGBoost över flera variabeluppsättningar. Kreditbetyg delas in i tre riskklasser, och modellernas prestande utvärderas med Macro-F1 och ROC-AUC över upprepade tränings- och testdelningar. Resultaten visar att trädbaserade ensemblemodeller presterar bättre än de statistiska baslinjemodellerna. Den bäst presterande konfigurationen är en Random Forest modell med ESG-variabler och kompletterande finansiella variabler, med ett genomsnittligt Macro-F1 på 0.655 och ett genomsnittligt ROC-AUC på 0.862. ESG-variabler förbättrar prestandan för flera modeller, men förbättringarna är måttliga i relation till variationen mellan datadelningarna.

Resultaten indikerar att ESG-variabler kan bidra med kompletterande prediktiv information när de kombineras med traditionella finansiella indikatorer. Analys av attributen visar att mått kopplade till skuldsättning och återbetalningsförmåga fortfarande är centrala, medan ESG relaterade variabler såsom koldioxidintensitet och styrelsesammansättning bidrar ytterligare signal. Studien begränsas dock av stickprovsstorlek, klassobalans, imputering av saknade värden och potentiellt tidsmässigt dataläckage i de kompletterande finansiella variablerna. Framtida forskning bör därför utvärdera ESG-integrerade kreditriskmodeller med större och tidsmässigt anpassade dataset.

***Index Terms*—ESG, Machine Learning, Credit Risk, Industrial Engineering (IE).**

I. Introduction

Credit risk assessment plays a central role in financial systems, determining how capital is allocated, priced, and regulated. Financial institutions rely on quantitative models to estimate the probability of default for corporations, which directly affects lending decisions, capital requirements, and portfolio risk management. Traditionally, these models are based on financial ratios and linear statistical methods. Although these approaches are well-established and transparent, they are often limited in their ability to capture complex, nonlinear relationships between variables. As financial markets become increasingly data-driven and interconnected, there is growing interest in more flexible modeling approaches that better reflect real-world dynamics [1], [2].

Simultaneously, Environmental, Social, and Governance (ESG) factors have become increasingly important in financial markets and regulatory frameworks. Sustainable finance initiatives, climate-related disclosure requirements, and ESG-oriented investment strategies have expanded significantly in recent years [3]. Financial institutions are increasingly required to account for sustainability-related risks, including climate exposure and governance practices, in their risk assessments. Despite this shift, ESG information is often treated as a qualitative complement rather than a quantitatively integrated component in credit risk models.

The primary objective of this study is to evaluate whether integrating ESG factors into machine learning models improves the prediction of corporate credit risk within the Nordic corporate bond market. The study compares traditional statistical methods, Naive Bayes and logistic regression, with modern decision-tree based machine learning approaches: Random Forest and XGBoost. Model performance is assessed using established evaluation metrics such as accuracy, F1-score, and ROC-AUC. By examining differences in predictive performance, robustness, and interpretability, the study aims to

determine whether nonlinear models can capture interactions between ESG variables and financial indicators that traditional methods may overlook. The results contribute to ongoing research in AI-driven financial risk modeling [1] and provide empirical evidence on the role of ESG factors in credit assessment.

The thesis addresses the following research questions:

1) *Computer Science:* To what extent does the inclusion of ESG variables improve predictive performance, measured by classification metrics such as F1-score and ROC-AUC, in machine learning models for corporate credit risk assessment compared to traditional statistical models?
2) *Industrial Engineering – IE:* How does the adoption of ESG-integrated credit risk models influence capital allocation efficiency, portfolio risk exposure, and regulatory considerations in a Nordic financial context?

A. Economic and Societal Context – IE

Beyond technical performance, integrating ESG factors into credit risk models has important economic and societal implications. These models influence how capital is allocated within the economy and shape long-term investment patterns. If ESG-related risks are underestimated, financial institutions may misprice credit risk and increase systemic vulnerability. More accurate models may therefore contribute to improved risk pricing, enhanced financial stability, and support for the transition towards a more sustainable economy.

However, integrating ESG variables into machine learning models also raises ethical concerns related to transparency, accountability, and potential algorithmic bias. ESG data may vary in quality and coverage across companies and industries, which could contribute to systematic bias in predictive models.

From a broader societal perspective, the study relates to the role of financial systems in promoting sustainable and resilient economic development. By improving the integration of sustainability-related risks into quantitative credit risk models, the report may support more informed capital allocation and long-term financial stability. In this sense, the study is connected to the United Nations Sustainable Development Goals (SDGs) [4], particularly SDG 8 (Decent Work and Economic Growth), SDG 9 (Industry, Innovation and Infrastructure), and SDG 13 (Climate Action), since it addresses the incorporation of climate and sustainability-related risks into financial decision-making processes.

II. Previous Studies

Recent findings from the Stanford AI Index Report [2] conclude that AI's influence on society has never been more pronounced, suggesting it is poised to become the most transformative technology of the 21st century. Record levels of investment and business adoption illustrate the scale of this shift. As AI becomes more efficient and accessible, governments are responding with increased regulation and strategic investments.

Financial services sit at the intersection of these regulatory and investment trends, since both the data-intensity and the consumer-protection sensitivity of finance make it one of the priority sectors for emerging AI governance frameworks. Bahoo et al. [1] demonstrate that applications of AI in finance have expanded accordingly, identifying ten main research streams. Within banking credit risk, they argue that future research should incorporate both market and accounting data while exploring machine learning's role in modeling systemic risk.

Advanced algorithms in credit risk assessment are further discussed by Sriram [5], who finds that generative AI models significantly enhance assessment compared to traditional models like logistic regression. The ability of AI to process complex datasets at high speed represents a transformative approach to scoring. Wallquist and Nilsson [6] provide further evidence, demonstrating that machine learning models exhibit higher predictive accuracy than traditional statistical models when predicting ratings based on financial metrics. Similar conclusions are reached by Chang et al. [7], identifying Random Forest and Gradient Boosting as particularly effective. These models are also successfully employed by Kurniawan [8]. Specifically, Gradient Boosting is highlighted as a relevant and explainable model for credit risk by Abedin et al. [9] and Mushava and Murray [10]. Across these studies, the consensus is that machine learning models, tree-based ensemble methods in particular, consistently outperform linear and probabilistic baselines for credit rating prediction on heterogeneous tabular data.

The importance of explainable AI (XAI) is emphasized by Chang et al. [7] through the use of precision-recall AUC and feature importance analysis. The challenges of AI in credit decision-making are discussed by Rudin [11], who illustrates that complex machine learning models, including ensemble methods and neural networks, are often regarded as "black box" systems. Given the institutional pressure to justify credit decisions, model interpretability is critical for trust and regulatory compliance.

Beyond methodological choice, the literature also identifies several channels through which ESG variables carry credit-relevant information: environmental factors affect future cash flows via transition and physical risk, governance variables proxy for management quality, and disclosure practices signal operational maturity. The role of Environmental, Social, and Governance (ESG) factors is examined by Lim [3], who identifies credit risk management as one of the most cited streams within AI-driven finance. ESG factors are presented as essential for promoting a sustainable financial sector. Hajek et al. [12] demonstrate that incorporating ESG report data improves the accuracy of corporate credit rating predictions. These findings align with Nordic-specific studies: Nyström and Skog [13] conclude that creditors offer lower borrowing costs to firms with stronger ESG profiles, while Hentilä [14] and Berg and Persson [15] find positive correlations between ESG disclosure/scores and credit ratings across Nordic industry groups.

Although these three findings [13][14] [15] support the relevance of investigating ESG disclosure and profiles in connection to credit risk, they are done on a small scale, and illustrate the gap in which this thesis aims to operate. The

European context for ESG-integrated credit risk modeling is established most directly by Stellner, Klein, and Zwergel [16], who examine the relationship between corporate sustainability performance and corporate bond risk across Europe using panel regression on a sample of Eurozone corporate bond issuers.

Using Asset4 ESG performance scores alongside Standard & Poor's credit ratings, they document a significant association between ESG performance and both credit ratings and bond yield spreads. Their central finding is that the ESG–credit-quality relationship is moderated by country-level sustainability. They conclude that ESG performance translates into superior credit ratings primarily when firms are situated in highly sustainable countries. Their study omits non-Eurozone countries, and therefore most of the Nordics are not included in this influential study. Since the Nordics are highly ranked in country-wide sustainability, the findings in Stellner et al. [16] support that a Nordic-focused study could exhibit a particularly detectable ESG–credit-risk signal.

The integration of AI and ESG is further explored by Zhang and Ma [17], who present a framework for green bond risk assessment. Their study builds on Wang et al. [18], who show that green innovation positively influences ratings. Zhang and Ma find that deep learning effectively captures nonlinear relationships, with feature importance analysis identifying the Debt-to-Equity Ratio, Carbon Intensity, and ESG Score as influential variables. This is consistent with a broader observation that disaggregated feature importance analysis can isolate which dimensions of ESG performance carry the strongest predictive signal, which is interpretively useful when underlying ESG components vary in measurement quality.

The Nordic corporate bond market reached EUR 135 billion in 2025 [19], underscoring the relevance of this research. While literature establishes that ensemble models like Gradient Boosting frequently outperform traditional techniques and that ESG factors influence creditworthiness, limited research integrates XAI frameworks with ESG-driven modeling in the Nordic corporate bond market. This combination is particularly relevant in the Nordic context, where high country-level sustainability rankings suggest that ESG signal in credit risk should be most detectable, while the prevalence of mid-cap issuers without external ratings raises the practical demand for transparent internal modeling. This thesis seeks to address these gaps by developing an explainable machine learning framework incorporating both financial and ESG metrics.

III. Theory

A. Data Theory

1) Supervised Learning for Structured Financial Data: Predicting credit ratings for listed Nordic companies constitutes a supervised classification problem. Given a dataset

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}, \tag{1}$$

where $x_i \in \mathbb{R}^p$ represents financial and ESG features (e.g., leverage ratios, profitability metrics, and carbon scores) and $y_i \in \mathcal{Y}$ represents a categorical credit rating class, the objective is to learn a function

$$f : \mathbb{R}^p \to \mathcal{Y} \tag{2}$$

that minimizes a loss function $L(y, f(x))$ on unseen data. Formally, the model aims to approximate the conditional probability

$$P(Y = y \mid X = x), \tag{3}$$

or directly learn a discriminative decision boundary between rating categories.

The models applied in this study span linear probabilistic approaches and nonlinear ensemble-based learners. This provides both an interpretable statistical baseline and flexible models capable of capturing complex interactions present in multi-dimensional financial and ESG data.

2) Logistic Regression: Logistic regression is a linear discriminative model commonly used as a benchmark in classification problems [20]. For binary classification, the model estimates:

$$P(Y = 1 \mid X = x) = \sigma(w^\top x + b), \tag{4}$$

where w is the weight vector, b is the bias term, and $\sigma(z) = (1 + e^{-z})^{-1}$ is the logistic (sigmoid) function. For multi-class problems, the model is extended using the softmax function:

$$P(Y = k \mid X = x) = \frac{\exp(w_k^\top x)}{\sum_{j=1}^{K} \exp(w_j^\top x)}. \tag{5}$$

Parameters are estimated via maximum likelihood estimation. While interpretable, the model is sensitive to multicollinearity, which may lead to unstable parameter estimates [21].

3) Naive Bayes: Naive Bayes is a generative probabilistic classifier based on Bayes' theorem:

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}. \tag{6}$$

The "naive" assumption states that features are conditionally independent given the class:

$$P(X \mid Y) = \prod_{j=1}^{p} P(x_j \mid Y). \tag{7}$$

Despite this simplifying assumption, Naive Bayes serves as a computationally efficient probabilistic baseline for high-dimensional data.

4) Decision Trees and Ensemble Methods: Decision trees recursively partition the feature space by selecting splits that maximize information gain. Common impurity measures include Gini impurity:

$$G = 1 - \sum_{k=1}^{K} p_k^2. \tag{8}$$

Both Random Forest and Gradient Boosting rely on decision trees as base learners to reduce the high variance associated with individual trees.

Random Forest: Introduced by Breiman [22], this is an ensemble method based on bootstrap aggregation (bagging). The ensemble prediction is defined as:

$$\hat{f}(x) = \text{majority vote}\{T_b(x)\}_{b=1}^{B}, \tag{9}$$

where T_b denotes the b-th decision tree.

Gradient Boosting: This is a sequential ensemble method that minimizes a loss function via gradient descent in function space. The iterative model update is:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x), \tag{10}$$

where $h_m(x)$ is a weak learner and γ_m is the step-size. For classification, the model typically minimizes the log-loss:

$$L = -\sum_{i=1}^{n} y_i \log(\hat{p}_i). \tag{11}$$

5) Evaluation Metrics: Credit ratings exhibit ordinal structure and class imbalance. Accuracy is defined as the ratio of correct predictions to total observations. To account for imbalance, we utilize Precision, Recall, and the F1-score:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{12}$$

To measure the discriminatory power of each model independently of the chosen classification threshold, we report the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC).

The ROC curve plots the true positive rate against the false positive rate as the decision threshold varies across its range, and the AUC summarizes this curve as the probability that a randomly chosen positive instance receives a higher score than a randomly chosen negative one.

An AUC of 0.5 corresponds to random guessing, while 1.0 indicates perfect ranking. Since credit ratings are a multi-class problem, we adopt a one-vs-rest (OvR) formulation [23], in which the K-class task is decomposed into K binary problems by treating each class in turn as the positive class and the remaining classes as the negative class. Reporting AUCs per class keeps detail that aggregate metrics would omit, allowing us to identify which credit rating class a model discriminates well and which it confuses. This is particularly informative when we have class imbalance, where overall accuracy can mask poor performance on minority classes, such as high yield classifications.

B. Economics Theory – IE

1) Credit Risk and Expected Loss Framework – IE: Credit risk is commonly decomposed into Probability of Default (PD), Loss Given Default (LGD), and Exposure at Default (EAD). Expected Loss (EL) is defined as [24]:

$$EL = PD \times LGD \times EAD. \tag{13}$$

Modern credit analysis also incorporates market-based measures such as bond spreads to capture risk premiums [25].

2) Governance, Agency Theory, and Default Risk – IE: Agency theory [26] suggests that conflicts between managers and capital providers may increase agency costs and default risk. Governance-related ESG indicators have theoretical relevance in credit risk modeling, as they influence managerial risk-taking incentives, monitoring quality, and long-term firm stability. In this context, board related variables can be interpreted as indicators of a firm's governance structure, since the board is responsible for monitoring management and representing shareholder interests.

However, governance variables should be interpreted with caution in predictive models. A variable such as board composition may capture governance quality, but it may also reflect the firm's characteristics, such as firm size, listing maturity, sector composition, or general quality of ESG reporting. This means that an important governance related feature in machine learning models should not automatically be interpreted as a direct causal driver of lower default risk. Instead, it may partly act as a proxy for broader company characteristics that are also related to creditworthiness [27].

3) Institutional Theory and Market Structure – IE: The global credit rating industry is an oligopoly characterized by high concentration. Oligopoly theory explains this structure through entry barriers and reputational capital [28]. Institutional theory further suggests that established practices persist due to regulatory embedding and market legitimacy [29].

Adoption of AI-driven ESG models depends on a trade-off between predictive accuracy and institutional feasibility, including explainability, regulatory compliance, and market trust. In credit risk assessment, this trade off is particularly important because model outputs can influence lending decisions, bond pricing, and firms' access to capital. More complex models may improve predictive performance, but they also increase the effort required to explain, validate, and justify model decisions to regulators, investors, and internal decision makers. This practical cost can be understood as a transparency tax: the additional burden created when more accurate but less interpretable models are used in high stakes financial decisions [7].

IV. Methodology

The methodology is structured to address both the technical and economic research questions. It follows a dual-path approach: a quantitative experimental design for credit risk modeling and a qualitative-analytical framework for evaluating market and business implications.

A. Quantitative Modeling Framework

The technical component of the study empirically evaluates the predictive performance of the machine learning models described in Section III when integrated with Environmental, Social, and Governance (ESG) factors.

B. Data acquisition and preprocessing

The dataset consists of financial and ESG data for listed Nordic corporate bond issuers. The data was acquired from Stamdata, with permission, and included five financial measurements and 15 ESG measurements. To supplement the limited financial ratios from Stamdata we also acquired data from Yahoo finance, which will be referenced as enriched financial data, consisting of 13 additional financial ratios. While Stamdata provides historical snapshots, the enriched data from Yahoo Finance consists of current snapshot ratios. This introduces a potential point-in-time leakage risk, where

Algorithm 1 Data Preparation and Splitting

function PREPAREDATA(dataset, config)
  $df \leftarrow$ LOADCSV($dataset$)
  for each fs in $config.feature_sets$ **do**
    $view \leftarrow df[fs.cols + target + gid]$
    validate no missing column definitions
    store $view$
  end for
  return $views$
end function
function SPLITDATA(df, strategy, test_size, seed)
  if $strategy =$ "stratified" **then**
    $train, test \leftarrow$ STRATIFIEDSPLIT($df,\ stratify{=}target,\ test_size,\ seed$)
  else if $strategy =$ "grouped" **then**
    $train, test \leftarrow$ GROUPSHUFFLESPLIT($df,\ groups{=}org_num,\ test_size,\ seed$)
  end if
  return $train_idx,\ test_idx$
end function

Algorithm 2 Model Initialization

function BUILDMODEL(name, params, cw_mode)
  if $name =$ "logistic_regression" **then**
    $est \leftarrow$ LOGISTICREGRESSION($C,\ max_iter,\ solver{=}$"lbfgs", $multi_class{=}$"multinomial")
  else if $name =$ "naive_bayes" **then**
    $est \leftarrow$ GAUSSIANNB()
  else if $name =$ "random_forest" **then**
    $est \leftarrow$ RANDOMFORESTCLASSIFIER($n_estimators{=}500$)
  else if $name =$ "xgboost" **then**
    $est \leftarrow$ XGBCLASSIFIER($n_estimators{=}500,\ lr{=}0.05,\ depth{=}4$)
  end if
  if $cw_mode =$ "balanced" **and** supported **then**
    $est.class_weight \leftarrow$ "balanced"
  end if
  return est
end function

Algorithm 3 Metric Evaluation

function EVALUATE($y_t,\ y_p,\ y_{prob}$)
  $acc \leftarrow$ ACCURACYSCORE($y_t,\ y_p$)
  $prec \leftarrow$ PRECISIONSCORE($y_t,\ y_p,\ avg{=}$"mac")
  $rec \leftarrow$ RECALLSCORE($y_t,\ y_p,\ avg{=}$"mac")
  $f1_M \leftarrow$ F1SCORE($y_t,\ y_p,\ avg{=}$"macro")
  $f1_W \leftarrow$ F1SCORE($y_t,\ y_p,\ avg{=}$"weighted")
  $y_b \leftarrow$ ONEHOTENCODE($y_t,\ classes$)
  $auc \leftarrow$ ROCAUCSCORE($y_b,\ y_{prob},\ mc{=}$"ovr", $avg{=}$"mac")
  return $\{acc,\ prec,\ rec,\ f1_M,\ f1_W,\ auc\}$
end function

Algorithm 4 Feature Importance Extraction

function EXTRACTIMPORTANCE(model, feat_names)
  if model has `feature_importances_` **then**
    $scores \leftarrow model.feature_importances_$
  else if model has `coef_` **then**
    $scores \leftarrow$ mean($|$`coef_`$|,\ axis{=}0$)
  else ▷ NaiveBayes: not available
    return $\emptyset$
  end if
  return SORTDESC(ZIP($feat_names,\ scores$))
end function

Note. Algorithms 1–4 summarize the experimental workflow used in the quantitative part of the study. Algorithm 1 describes how feature views are constructed and split into training and test sets. Algorithm 2 specifies the model families used in the experiments. Algorithm 3 defines how model predictions are evaluated, while Algorithm 4 describes how feature importance scores are extracted for model interpretation.

future financial information is used to predict past ratings. This limitation is further addressed in the discussion.

To ensure the models can train on the data effectively, the data had to be cleaned and preprocessed before running the experiment. Data was collected from Stamdata and Yahoo Finance API feeds and merged into a table. To ensure maximum coverage of relevant data points, it was confirmed to have non-null coverage for `siI_RatingNormalized`, removed rows missing required target/feature inputs, checked and resolved duplicate issuer-year rows.

We combine the traditional financial metrics with ESG variables to construct a multi-dimensional feature space with the input values grouped into categories to allow for training in multiple feature configurations, illustrated in table 1.

Core financial features include enterprise value, revenue, book value of equity and debt, and the debt-to-equity ratio. Enriched ratios comprise liquidity measures (current, quick, and cash ratios), leverage measures (debt-to-assets, debt-to-equity, equity ratio, and interest coverage), profitability measures (gross, operating, and net margin, as well as ROA and ROE), and asset turnover. The environmental ESG features cover total greenhouse gas emissions, Scope 1 and Scope 2 (location-based) emissions, carbon intensity, the presence of a carbon reduction target, flags for high-impact climate sectors and fossil fuel sectors, and two biodiversity indicators capturing whether a company reports on biodiversity and whether its operations negatively affect it. Finally, the social and governance features capture exposure to controversial weapons, controversial products, and debt collection or predatory lending activities, along with board composition variables: the number of female and male board members and the resulting female board share.

The output variable of credit scoring had a range from AAA to NR, which meant that the output space was relatively complex. To improve stability and applicability of the trained models we therefore decided to constrain the solution space to three classes. The class mapping was done using `siI_RatingNormalized` (CQS (Credit Quality Steps)-like scale) where 1-2 -> Investment Grade Low Risk, 3 -> Investment Grade High Risk, 4-6 -> High Yield.

Model Evaluation and Interpretation: The implementation is performed in Python, utilizing standard libraries for machine learning. We follow a comparative approach, training the linear models (Logistic Regression, Naive Bayes) alongside the nonlinear ensemble models (Random Forest, XGBoost).

After preprocessing, the dataset is transformed into several

Table I: Feature sets used across model configurations.

Feature group (#)	Fin.	Fin.+	ESG	ESG+Fin.	ESG+Fin.+
Core financial (5)	✓	✓		✓	✓
Enriched ratios (13)		✓			✓
ESG environmental (9)			✓	✓	✓
ESG social & gov. (6)			✓	✓	✓
Total features	**5**	**18**	**15**	**20**	**33**

feature views corresponding to the configurations in Table I. This procedure is summarized in Algorithm 1, where each feature view contains the selected feature columns, the target variable, and the group identifier used for grouped splitting.

The data is then divided into training and test sets using two strategies. The first is a group split, where the same company is not allowed to appear in both the training and test set. The second is a stratified split, where the class distribution is preserved as far as possible across the training and test sets. These two split strategies are also shown in Algorithm 1.

- **Grouped split:** Data points do not appear twice in the test and training data set, and any given row is either only in the training or test data set.
- **Stratified:** This ensures representation of key subgroups [30] , namely strata, which leads to more precise and relevant estimates. We split rows such that training data has roughly the same mix of classes as the full data, but since rows are chosen independently the same company can appear in both training and test splits.

For each feature configuration and split strategy, the models are initialized according to Algorithm 2. The initialized estimators are trained on the training split and evaluated on the corresponding test split. This process is repeated over ten independent random seeds for both the grouped and stratified split strategies.

Hyperparameters were deliberately not tuned against the test data, to avoid optimistic bias on the small sample; we adopt standard literature-typical settings, 500 trees for both ensembles so the bagging/boosting estimate is stable, and a deliberately conservative XGBoost configuration (learning rate 0.05, maximum depth 4) to limit overfitting given $N = 382$.

The mapping of company ID to Yahoo Finance ticker was not 1-to-1, which meant that we had to handle missing data points when training models. The code does not drop the data points which only have partial enriched features. It imputes missing enriched values and keeps all rows in training/testing. This is done when feature views are created, which is done by column selection only, with no missing-value filtering in `build_feature_views.py:19`.

Training always starts with a median imputer in the pipeline in `train_models.py:35`. The model is fit on split rows directly (not a filtered subset) in `run_experiments.py`. This means, in practice, that if a row is missing some enriched features, those missing cells are replaced by the training-set median for each feature; and if a row is missing all enriched-only features, it still stays in the dataset and gets median-filled values for those columns.

Since the imputer is inside the `sklearn` pipeline and fit during model training, test-set missing values are also filled using medians learned from training data only, ensuring good leakage hygiene.

To account for the small sample size, every reported metric in this thesis is the mean and standard deviation across ten train-test splits generated under independent random seeds. This protocol allows the reported point estimates to be interpreted with their corresponding uncertainty bands, and supports the robustness analysis presented in VI-D

Metrics and Feature Importance: Performance will be assessed using the metrics defined in Section 3.7, specifically F1-score and ROC-AUC, to account for class imbalance. To address the "black box" nature of the ensemble models, we will use feature importance rankings to interpret how individual ESG factors influence the final credit score. Evaluation of metrics is performed as in Algorithm 3.

Feature importance is evaluated based on the model type. Random Forest and XGBoost use Gini importance (mean decrease in impurity): each time a feature is used to split a tree node, the impurity decrease is accumulated and averaged across all trees. This is computed internally by `scikit-learn`/`XGBoost` and exposed via `feature_importances_`. For Logistic Regression, we use the absolute coefficient magnitude averaged across classes. Since the model is multinomial with 3 classes, `coef_` is a $(3 \times n_{features})$ matrix; taking mean($|coef|$, $axis$=0) yields a single importance score per feature. Features are standardized beforehand (`StandardScaler` in the pipeline), so coefficient magnitudes are comparable. For Naive Bayes, no importance scores are extracted since `GaussianNB` does not expose coefficients or importances in a directly comparable form.

C. Economic and Managerial Analysis – IE

The Economics component evaluates the implications of the modeling results for capital markets and business strategy within a Nordic context.

1) Analytical Framework – IE: The analysis is grounded in the credit risk decomposition formula introduced in Eq. (13). The primary focus is on how the inclusion of ESG variables affects the estimated PD.

Specifically, the analysis examines whether ESG-integrated models contribute to more efficient capital allocation by identifying risks that traditional models overlook. It further evaluates how improved predictive accuracy may influence credit spread pricing and the cost of debt for Nordic corporations.

2) Institutional Reflection – IE: While the quantitative modeling provides empirical evidence of predictive performance, the Industrial Engineering analysis critically evaluates the institutional feasibility of these models.

First, ESG data quality varies significantly across sectors, raising concerns regarding measurement error. Second, the complexity of nonlinear models may introduce a "transparency tax," where predictive accuracy comes at the cost of reduced explainability. Adoption depends not only on predictive superiority but also on regulatory compliance and market trust.

V. Result

A. Quantitative Results from Experiment Pipeline

Table II: Top 5 model configurations ranked by mean Macro-F1 across 10 random seeds.

Split	Feature set	Model	Macro-F1	AUC
Grouped	ESG + Financial Enriched	Random Forest	0.6552 ± 0.0625	0.8622 ± 0.0395
Stratified	ESG + Financial Enriched	XGBoost	0.6527 ± 0.0488	0.8539 ± 0.0174
Grouped	ESG + Financial Enriched	XGBoost	0.6377 ± 0.0695	0.8477 ± 0.0355
Stratified	Financial Enriched	XGBoost	0.6287 ± 0.0765	0.8074 ± 0.0324
Stratified	ESG + Financial Enriched	Random Forest	0.6268 ± 0.0510	0.8853 ± 0.0119

Table II reports the five best models ranked by mean Macro-F1 across 10 random seeds. The best performing configuration was grouped ESG and Financial Enriched by Random Forest, which received a mean Macro-F1 of 0.6552 and mean AUC of 0.8622. The second best configuration is stratified ESG + Financial Enriched XGBoost model, with a mean Macro-F1 0.6527 and mean AUC 0.8539. Overall the best configurations per Macro-F1 score were dominated by the most enriched dataset and the tree models. In Appendix A Table VI and VII it is evident that the first spot that a non-tree model occupies is nr 9 within the grouped split and 10 in the stratified split measured by Macro-F1 descending. Naive Bayes ranked last in all regards for both splits. All reported figures and tables are computed on held-out test sets of $n \approx 77$ issuers per seed, drawn from a full sample of $N = 382$ issuer-years (37 High Yield, 157 Investment Grade High Risk, 188 Investment Grade Low Risk).

In Table VII we see that adding ESG features to the financial baseline improves F1 across all models:

- **XGBoost:** $0.629 \rightarrow 0.653$ (+0.024)
- **Random Forest:** $0.564 \rightarrow 0.626$ (+0.062)
- **Logistic Regression:** $0.474 \rightarrow 0.516$ (+0.042)

The same ordering holds for AUC, e.g. Random Forest $0.840 \rightarrow 0.885$.

Figure 1 reports the top 15 most important features for logistic regression, random forest and XGBoost, which is calculated as mentioned in Model Evaluation and Interpretation.

Figure 2 illustrates the one versus rest OvR, one class compared to both the others, for the receiver operating characteristic for all three classes and the best performing data set (ESG + Financial Enriched). The tree based models dominate across all classes, logistic regression falls somewhat short of both XGBoost and Random Forest, while Naive Bayes lags significantly behind. The investment grade curves show the most consistent classification patterns, while the high yield curve is jagged.

Table III shows that the best configuration per model was ESG + Financial Enriched for all models except Naive Bayes, for which it was the pure financial enriched. Since we used multiple seeds to evaluate the models, grouped and stratified splits did not differ greatly. When only using one seed and training once, grouped and stratified splits can cause quite large deviations, especially with a limited dataset.

B. Economic Implications of Predictive Lift – IE

To translate the observed predictive lift into an economic interpretation, we examine how the addition of ESG features changes risk-class assignments at the issuer level and what this implies under a standard expected loss framework, introduced in Section III-B. The analysis is illustrative rather than prescriptive: it uses indicative probability of default (PD) anchors per rating class together with a Basel-style loss given default (LGD) assumption, in order to convert a classification difference into a credit-risk magnitude that institutional investors would recognize.

Table III: Best Configuration Per Model (10 Seeds)

Model	Feature Set	Split	$F1_{Macro}$ mean $\pm$ std	AUC mean $\pm$ std
Random Forest	ESG + Fin. Enriched	grouped	0.6552 ± 0.0625	0.8622 ± 0.0395
XGBoost	ESG + Fin. Enriched	stratified	0.6527 ± 0.0488	0.8539 ± 0.0174
Logistic Reg.	ESG + Fin. Enriched	grouped	0.5317 ± 0.0446	0.7560 ± 0.0529
Naive Bayes	Financial Enriched	grouped	0.3844 ± 0.0581	0.6925 ± 0.0500

We compare the Random Forest predictions under the Financial Enriched configuration with those under the ESG + Financial Enriched configuration on the held-out test set, averaged across the ten random seeds under the grouped split. The results are shown in Appendix B, Table IX.

On average, approximately 18.8% of test issuers per seed receive a different risk-class assignment when ESG features are added. Of these reclassifications, the majority occur between Investment Grade Low Risk and Investment Grade High Risk, while approximately 2.2% of test issuers move from High Yield to an Investment Grade class and 2.1% move in the opposite direction. The cross-boundary flow is thus close to symmetric, suggesting that ESG features primarily refine gradation within investment grade rather than moving a company across the High Yield boundary.

To assign economic magnitudes, we adopt indicative one-year PD anchors per class based on long-run rating-to-default mappings reported in the credit risk literature [24]: $PD_{IG\text{-}Low} \approx 0.10\%$, $PD_{IG\text{-}High} \approx 0.50\%$, and $PD_{HY} \approx 4.00\%$. We use $LGD = 45\%$, a senior unsecured benchmark commonly applied in Basel-aligned credit modeling [24], and normalize EAD to one monetary unit per issuer. Expected loss is computed as $EL = PD \times LGD \times EAD$, yielding the per-class EL values reported in Table IV.

Under these anchors, a single reclassification across the High Yield / Investment Grade-High boundary corresponds to an EL difference of approximately 158 basis points of exposure ($(0.040 - 0.005) \times 0.45$), and a reclassification across the High Yield / Investment Grade-Low boundary to approximately 176 basis points. Reclassifications within the investment grade band correspond to a smaller EL difference of approximately 18 basis points.

Aggregated gross across all reclassifications, the average seed reassigns roughly 4.3% of test issuers cross the boundary at ≈160 bps each, and roughly 14.5% are reclassified within investment grade at ≈18 bps each, for a gross cumulative exposure-weighted EL difference on the order of several hundred basis points per seed.

These figures should be read as order-of-magnitude indicators rather than precise estimates. The PD anchors are not estimated from Nordic data, the rating-class mapping collapses substantial within-class heterogeneity, and the sym-

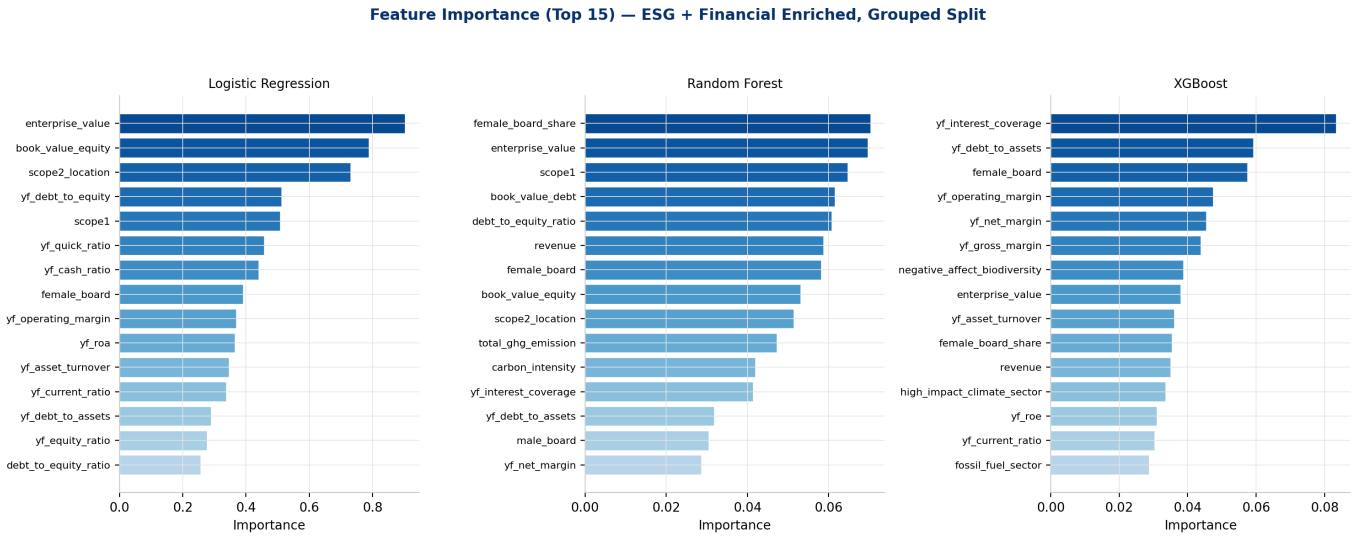

Figure 1: Feature Importance (Top 15) for best scoring data configuration

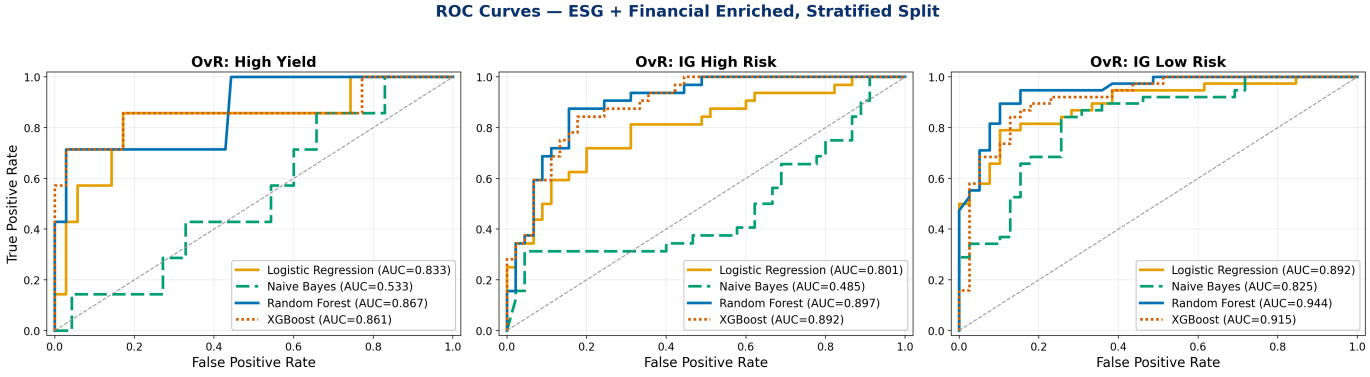

Figure 2: ROC Curves for best scoring data configuration

metric nature of High Yield / Investment Grade flows means the net portfolio-level EL difference is smaller than the gross reclassification activity suggests.

Table IV: Indicative expected loss per issuer by risk class, assuming LGD $= 45\%$ and EAD $= 1$. PD anchors based on long-run rating-to-default mappings [24].

Risk class	PD	LGD	EL
Investment Grade Low Risk	0.10%	45%	0.045%
Investment Grade High Risk	0.50%	45%	0.225%
High Yield	4.00%	45%	1.800%

VI. Discussion

Across three of the four model families, integrating ESG with financial features improved mean Macro-F1, with directional consistency: $+0.062$ for Random Forest, $+0.024$ for XGBoost, and $+0.042$ for Logistic Regression; for Naive Bayes the change was negative (-0.038, grouped), consistent with its violated independence assumption (Sec. VI-A). A paired test across the ten shared splits (Table V) shows the Random Forest ESG lift is statistically significant for Macro-F1 under both splits and for AUC under the stratified split, while XGBoost shows a significant AUC gain under both splits; the XGBoost Macro-F1 gain is not significant. Because the seeds resample one finite dataset, this demonstrates stability under resampling rather than out-of-sample generalisation, and the absolute lift stays modest. The results indicate that ESG scores enrich the predictive power of the models, though the lift is relatively modest and highlights the complexity of corporate credit rating prediction. Using standard machine learning algorithms as a baseline further illustrates the dominance of tree-based learning approaches when dealing with multi-class classification problems with multidimensional feature spaces.

A. Performance of Individual Models

The results from the Naive Bayes (NB) model are the first ones worth noting. In our results, it is effectively worse than a majority class baseline on accuracy (though it performs better on F1-score), which indicates that the independence assumption and Gaussian conditional distribution required for Naive Bayes are not fulfilled by this dataset. This being said, under a balanced class weight, accuracy is expected to drop below majority baseline, because the model trades majority precision for minority recall.

Since many of the financial ratios are directly dependent on each other, the poor performance of NB comes as no surprise,

but serves as a useful sanity baseline to illustrate that this problem of multi class credit rating prediction is not trivially solvable.

Tree ensemble methods dominate the results in predictive capability with insignificant differences between the two (differences fall within one standard deviation across seeds). This indicates that the problem is not linearly separable or trivially solvable. This is in accordance with the hypothesis derived from the literature review regarding the predictive power of nonlinear models. The application of such models on Nordic datasets and ESG factors strengthen the hypothesis that nonlinear models have superior ability to classify problems that are not linearly separable.

Logistic regression performed well as a baseline that loses cleanly across the metrics. While being able to classify somewhat accurately in comparison with NB, its consistent 0.05–0.13 F1 gap behind trees demonstrates that the dataset requires nonlinear interactions or feature combinations to separate the three classes.

In terms of practical tradeoffs, Logistic Regression remains the most interpretable and the cheapest to retrain, but is bound by its linearity assumption. Random Forest and XGBoost capture nonlinear interactions and achieve the highest ranking quality (AUC 0.85–0.88) at the cost of opacity and a larger explainability burden. Naive Bayes is the fastest to train but is unsuitable here due to violated independence assumptions. This positions tree ensembles as the preferred candidate when predictive accuracy is the priority and a XAI layer (e.g. feature importance) is acceptable for compliance.

B. Ranking Quality and Metric Interpretation

In addition to classification metrics, the results include ROC-AUC, which measures the models' ability to rank issuers by risk independently of a fixed classification threshold. These metrics underscore the superiority of tree ensemble methods, where model ranking is continually consistent: Random Forest $\geq$ XGBoost > Logistic Regression $\gg$ Naive Bayes across all three classes. The jagged line of the NB in all of the three OvR curves in Figure 2 illustrates that the model is conditionally broken for this type of classification due to near-linearly dependent features.

The best model achieves a Macro-F1 of 0.65 but a significantly higher AUC of 0.85–0.88, meaning that the models are better at ranking issuers by risk than at assigning hard class labels. Since credit applications in corporate finance and debt issuing often utilize risk thresholds rather than discrete labels, the high probability quality of tree ensembles is of practical value. The class difficulty ordering indicates that the models are picking up financial health signals that separate safe from not safe, which in turn could be subject for further research.

Table 2 also illustrates that Investment Grade Low Risk is easiest to predict across the models, for which a likely contributor is both the consistency of "good" features for a large, stable institution or corporation, but also the size of the dataset. The opposite holds for High Yield: it was the smallest class, so test sets contained roughly seven High Yield issuers per seed, which both limits learnable signal and produces the visibly jagged High Yield ROC curves.

C. Issuers Crossing Class Boundaries

Of the 260 ESG-driven reclassifications pooled over all seeds and splits, 60 cross the High Yield / Investment Grade boundary and 200 are refinements within the Investment Grade band (Table **??**). The two behave very differently. The 60 HY/IG crossings involve only 29 distinct issuer-years and are strongly directional: ESG moves the prediction toward the true class in 40 cases against only 15 away (5 wrong under both), a net of $+25$ correct decisions ($\approx$2.7:1). These crossers are disproportionately High Yield, 13 of the 37 High Yield issuer-years (35%) cross at least once, versus 7 of 188 Investment-Grade-Low issuer-years (3.7%) and they differ from the rest precisely on the ESG variables Fig. 1 ranks highest for the Random Forest: a lower mean female board share (0.27 vs. 0.40) and far higher concentration in high-impact climate sectors (0.79 vs. 0.51).

The 200 within-IG moves are, by contrast, spread across 97 distinct issuers, roughly balanced in direction (98 toward higher risk, 102 toward lower), and only weakly correct 111 toward the true class against 74 away (1.5:1). The crossers here are almost indistinguishable from non-crossers on ESG features (carbon intensity 984 vs. 1065; female board share 0.38 vs. 0.40). ESG therefore primarily reshuffles gradation inside investment grade where the distinction is least consequential for mandates and capital treatment while its sharp, reliable, ESG-driven effect is concentrated at the High Yield boundary that actually governs lend/no-lend and eligibility decisions.

D. Robustness of Performance Estimates

A methodological observation worth highlighting is that the first attempt to train the models was done with only one random seeded split per grouped/stratified. This caused the stratified and grouped results to differ, which raised alarms about the robustness of the results. To ensure more robust and relevant end metrics, we decided to run multiple seeded splits, which showed that the single-seed estimate was more than 2.4 standard deviations above the repeated-seed mean. This gap is a direct consequence of the dataset's class imbalance.

E. Statistical significance of the ESG lift

To test whether adding ESG features improves performance beyond split-to-split noise, we compare the financial enriched baseline with the 'esg financial enriched configuration on a paired basis: for each of the ten seeds both feature sets are trained and evaluated on the identical train-test split, so the difference in a metric isolates the contribution of the ESG features. For each model we report the mean paired difference Δ, its 95% confidence interval, and two-sided p-values from a paired t-test and from the Wilcoxon signed-rank test (the latter as a non-parametric robustness check appropriate for $n = 10$).

The 95% confidence intervals for the Macro-F1 lift exclude zero for the Random Forest (grouped $[+.005, +.139]$; stratified $[+.035, +.089]$) and Logistic Regression (grouped $[+.006, +.103]$; stratified $[+.007, +.078]$), but include zero for XGBoost; for the AUC lift they exclude zero for both tree

Table V: Paired significance of the ESG lift (`esg_financial_enriched` vs. `financial_enriched`) across the ten shared splits. Δ is the mean paired difference (positive favours ESG); p_t/p_W are the F1 paired t-test / Wilcoxon p-values; "Wins" counts seeds where ESG improved F1. Models: RF Random Forest, XGB XGBoost, LR Logistic Regression, NB Naive Bayes.

Split	Model	ΔF1	p_t/p_W	ΔAUC	Wins
Grouped	RF	$+.072$	$.039/.049$	$+.034$	8/10
	XGB	$+.033$	$.253/.375$	$+.049$	6/10
	LR	$+.055$	$.031/.084$	$+.017$	7/10
	NB	$-.038$	$.008/.014$	$-.015$	2/10
Stratified	RF	$+.062$	$< .001/.002$	$+.045$	10/10
	XGB	$+.024$	$.312/.432$	$+.047$	10/10
	LR	$+.043$	$.025/.049$	$+.037$	8/10
	NB	$-.008$	$.414/.492$	$+.016$	8/10

ensembles under the stratified split. For the Random Forest the ESG lift is statistically significant for both Macro-F1 (stratified $+0.062$, $p < 0.001$; grouped $+0.072$, $p = 0.04$) and ROC-AUC, improving the metric in 8/10 (grouped) and 10/10 (stratified) seeds. Logistic Regression shows a significant Macro-F1 gain under both splits. For XGBoost the Macro-F1 gain is not significant, yet its AUC gain is ($p \leq 0.014$): ESG sharpens XGBoost's risk ranking even where it does not move enough hard labels to register in F1. Naive Bayes is the informative counter-case, adding ESG features significantly degrades its Macro-F1 (-0.038, $p = 0.008$, grouped), consistent with its conditional-independence assumption, which correlated ESG variables violate. Caveat: the ten seeds are repeated resamples of the same 382-issuer sample, so these tests establish that the lift is stable under resampling rather than a single fortunate split; they do not by themselves establish generalisation to issuers outside the sample, and the absolute magnitude of the lift remains modest.

F. Feature Importance and Economic Implications – IE

The feature importance analysis in Figure 1 provides insight into which variables the models rely on when classifying credit risk in the Nordic market. Several of the most influential variables are traditional financial indicators, particularly leverage related measures such as debt-to-equity, debt-to-assets, equity ratio, and interest coverage. From an economic perspective this is expected, since the variables reflect capital structure, solvency, and repayment capacity, which are central components of credit risk.

However, these leverage related variables are not fully independent measures, since they are all based on related balance sheet and income statement values. Their importance should therefore not be interpreted as four separate economic effects, but rather as an indication that the models identify a common leverage and solvency signal. This also helps explain why Naive Bayes performs poorly in the experiments. As discussed earlier, Naive Bayes assumes that features are conditionally independent given the class label, an assumption that is clearly weakened when several important financial ratios are derived from the same underlying accounting values. Logistic Regression may also be affected by the same feature structure , since multicollinearity can make coefficient estimates less stable [21]. In contrast, Random Forest and XGBoost are less affected by this problem, since they do not rely on estimating one linear coefficient for each feature in the same way. Instead, they use related variables through different splits and capture nonlinear pattern in the data.

ESG-related variables, such as female board share and carbon intensity, also emerge as influential. This suggests that ESG information contains additional predictive signal beyond traditional financial ratios. The importance of carbon intensity may reflect transition risk, where firms with higher emission exposure could face higher future costs, regulatory pressure, or changing investor preferences. Similarly, the importance of female board share can be interpreted in relation to the agency theory perspective introduced in section III, where board composition may capture aspects of monitoring quality and governance structure. However, the interpretation should be made cautiously, as female board share may also be a proxy for firm size, listing maturity, sector composition, or broader ESG disclosure quality rather than acting as a causal driver of credit risk [27].

G. Nordic Bond Market – IE

From a market-structure perspective, in the Nordic bond market, credit ratings are traditionally dominated by an oligopoly of global agencies. Oligopoly theory explains this through high entry barriers and reputational capital. However, our research suggests that machine learning driven ESG models can provide a quantitative alternative to these subjective ratings. While established rating agencies provide market legitimacy, their ratings are not necessarily 'correct' in a purely predictive sense. As the market for private debt grows, there is an increasing need for flexibility, data-driven assessment tools that can bypass traditional agency reliance.

H. Portfolio Management and Capital Allocation – IE

From a portfolio management perspective, more accurate models contribute to improved capital allocation efficiency. Credit risk assessment is integral to building a balanced, long-term viable investment portfolio, particularly for institutional investors who need to deploy capital without introducing unnecessary risk or drawdowns to portfolio value. Accurate credit-spread estimation is central to this, since spreads drive both pricing and risk-adjusted return at the portfolio level. In the framework illustrated by Eq. 13, better ranking maps directly onto better PD estimation and expected-loss provisioning. The practical value of ESG integration is therefore at the spread-pricing layer, not at the binary lend/no-lend decision.

It is also common for short-term mutual bond funds to be used by institutional investors operating under mandates that simultaneously prohibit high-yield paper and require sustainable investment. This thesis shows that, in the Nordic market, these two requirements can be reconciled rather than traded off against each other. Since many Nordic corporate bond issuers do not have ratings from the global rating oligopoly, internal

rating estimates, such as the ML approach explored here, can help investors manage risk in this segment of the market.

However, the use of "black-box" models such as Random Forest in finance poses challenges for transparency and regulatory compliance. Adoption depends on a trade-off between predictive accuracy and institutional feasibility - the "transparency tax".

I. Contribution to Research Landscape

This thesis provides Nordic-specific evidence for the Stellner et al. moderation hypothesis, particularly for the non-Eurozone Nordic countries that fall outside their original sample. It also demonstrates that the ESG–credit signal survives the move from panel regression on composite ESG scores to ML classification on disaggregated ESG components.

J. Limitations and Future Research

A primary limitation is the risk of point-in-time leakage. While Stamdata provides historical snapshots, Yahoo Finance returns current ratios, meaning future financial information may have been used to predict past ratings.

Furthermore, the sample size (N=382) with only 37 High Yield observations limits statistical power. Additionally, 248 of the 382 rows required median imputation, meaning the observed lift is partly mediated by imputation behavior, since 65% of rows have at least one imputed enriched cell. We also note the absence of ESG composite scores and climate policy uncertainty variables common in larger literature.

Furthermore, 65% of rows received at least one imputed enriched feature; the median imputation strategy, while standard, compresses cross-sectional variance in the enriched ratios and likely attenuates their feature-importance scores. The reliance on Yahoo Finance for enriched ratios also means that companies with thin coverage (typically smaller mid-cap Nordic issuers) are disproportionately imputed, which may bias the models toward signal from larger, better-covered firms.

Future studies should explore larger datasets and consider splitting the Investment Grade category further to capture its inherent diversity.

VII. Conclusion

This study investigates whether ESG factors can improve machine learning based credit rating classification in the Nordic corporate bond market. The results indicate that ESG variables can provide complementary predictive information when combined with traditional financial indicators. Tree based ensemble models, particularly Random Forest and XGBoost, outperform the statistical baselines across most configurations, with the best model achieving a mean Macro-F1 of 0.655 and a mean ROC-AUC of 0.862.

The improvement from adding ESG variables is directionally consistent for several model families, but moderate relative to the variation across train-test splits. This suggests that ESG information may contribute useful signal, but should be viewed as a complement to, rather than a replacement for, traditional financial indicators. The feature importance analysis further indicates that leverage and repayment capacity variables remain central, while ESG variables such as carbon intensity and board composition add additional predictive information.

From an industrial engineering perspective, the results suggest that ESG integrated machine learning models may support more informed credit risk assessment, portfolio analysis, and capital allocation. However, practical adoption is constrained by the transparency tax associated with complex models, as well as by ESG data quality and regulatory explainability requirements.

VIII. Author Contributions

S. Olsson was responsible for data collection and the ESG framework, while T. Gunnarsson developed the machine learning models. Both authors contributed equally to the writing of the report.

IX. AI Assistance

Generative AI tools were used to support code generation, debugging, and language refinement during the preparation of this thesis. The research idea, problem formulation, methodological choices, analysis, and interpretation of results were developed and evaluated by the authors. All AI-assisted material was reviewed, modified, and approved by the authors, who take full responsibility for the content of the report.

References

[1] S. Bahoo, M. Cucculelli, X. Goga, and J. Mondolo, "Artificial intelligence in finance: A comprehensive review through bibliometric and content analysis," *SN Business & Economics*, vol. 4, no. 2, p. 23, 2024.

[2] Stanford Institute for Human-Centered AI, "The 2025 AI index report," Stanford University, Tech. Rep., 2025.

[3] T. Lim, "Environmental, Social, and Governance (ESG) and artificial intelligence in finance: State-of-the-art and research takeaways," *Artificial Intelligence Review*, vol. 57, no. 4, 2024.

[4] United Nations General Assembly, "Transforming our world: The 2030 agenda for sustainable development," 2015.

[5] H. K. Sriram, "AI neural networks in credit risk assessment: Redefining consumer credit monitoring and fraud protection through generative AI techniques," *Migration Letters*, vol. 19, no. 6, 2022.

[6] C. Wallquist and O. Nilsson, "Enhancing credit rating prediction with machine learning and explainable AI," Master's thesis, University of Gothenburg, 2025.

[7] V. Chang, Q. A. Xu, S. H. Akinloye, V. Benson, and K. Hall, "Prediction of bank credit worthiness through credit risk analysis: An explainable machine learning study," *Annals of Operations Research*, 2024.

[8] R. Kurniawan, "Application of random forest algorithm on credit risk analysis," *Procedia Computer Science*, vol. 245, pp. 740–749, 2024.

[9] M. Z. Abedin, C. Guotai, P. Hajek, and T. Zhang, "Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk," *Complex & Intelligent Systems*, vol. 9, pp. 3559–3579, 2023.

[10] J. Mushava and M. Murray, "Flexible loss functions for binary classification in gradient-boosted decision trees: An application to credit scoring," *Expert Systems with Applications*, vol. 238, p. 121876, 2024.

[11] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[12] P. Hajek, J.-M. Sahut, and R. Myskova, "Predicting corporate credit ratings using the content of ESG reports," *Annals of Operations Research*, 2024.

[13] K. Nyström and E. Skog, "ESG and the cost of debt: Evidence from nordic public firms," Master's thesis, University of Gothenburg, 2022.

[14] D. Hentilä, "Impact of ESG disclosure on credit ratings in the nordics," Master's thesis, Hanken School of Economics, 2022.

[15] J. Berg and J. Persson, "ESG-betygets påverkan på kreditbetyget: En kvantitativ studie på 311 publika nordiska företag," Bachelor's thesis, University of Gävle, 2023.

[16] C. Stellner, C. Klein, and B. Zwergel, "Corporate social responsibility and Eurozone corporate bonds: The moderating role of country sustainability," *Journal of Banking & Finance*, vol. 59, pp. 538–549, 2015.

[17] D. Zhang and X. Ma, "Machine learning-based credit risk assessment for green bonds: Climate factor integration and default prediction analysis," *Journal of Sustainability, Policy, and Practice*, vol. 1, no. 2, pp. 121–135, 2025.

[18] Y. Wang, J. Feng, R. Shinwari, and E. Bouri, "Do green finance and green innovation affect corporate credit rating performance? evidence from machine learning approach," *Journal of Environmental Management*, vol. 360, p. 121212, 2024.

[19] Nordic Trustee and Ocorian, "Nordic corporate bond market report 2025," Nordic Trustee, Tech. Rep., 2025.

[20] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ: Wiley, 2013.

[21] A. F. Lukman, S. Mohammed, O. Olaluwoye, and R. A. Farghali, "Handling multicollinearity and outliers in logistic regression using the robust Kibria–Lukman estimator," *Axioms*, vol. 14, no. 1, p. 19, 2025.

[22] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[23] D. J. Hand and R. J. Till, "A simple generalisation of the area under the ROC curve for multiple class classification problems," *Machine Learning*, vol. 45, no. 2, pp. 171–186, 2001.

[24] C. Bluhm, L. Overbeck, and C. Wagner, *Introduction to Credit Risk Modeling*, 2nd ed. Boca Raton, FL: Chapman & Hall/CRC, 2010.

[25] D. Lando, "Credit risk modeling," in *Handbook of Financial Time Series*. Springer, 2009, pp. 787–798.

[26] M. C. Jensen and W. H. Meckling, "Theory of the firm: Managerial behavior, agency costs and ownership structure," *Journal of Financial Economics*, vol. 3, no. 4, pp. 305–360, 1976.

[27] S. Drempetic, C. Klein, and B. Zwergel, "The influence of firm size on the esg score: Corporate sustainability ratings under review," *Journal of Business Ethics*, vol. 167, no. 2, pp. 333–360, 2020.

[28] J. Tirole, *The Theory of Industrial Organization*. Cambridge, MA: MIT Press, 1988.

[29] P. J. DiMaggio and W. W. Powell, "The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields," *American Sociological Review*, vol. 48, no. 2, pp. 147–160, 1983.

[30] S. K. Ahmed, "How to choose a sampling technique and determine sample size for research: A simplified guide for researchers," *Oral Oncology Reports*, vol. 12, p. 100662, 2024.

X. Appendix

A. Summary of all Results

Table VI: All configurations — Grouped Split (10 seeds).

Feature set	Model	Accuracy	$\text{Precision}_{\text{macro}}$	F1_{macro}	AUC
ESG + Financial Enriched	Random Forest	0.7349 ± 0.0494	0.7210 ± 0.0758	0.6552 ± 0.0625	0.8622 ± 0.0395
ESG + Financial Enriched	XGBoost	0.6975 ± 0.0551	0.6517 ± 0.0912	0.6377 ± 0.0695	0.8477 ± 0.0355
Financial Enriched	XGBoost	0.6566 ± 0.0666	0.6152 ± 0.0959	0.6052 ± 0.0869	0.7988 ± 0.0556
ESG + Financial	Random Forest	0.6576 ± 0.0556	0.6769 ± 0.1057	0.5846 ± 0.0654	0.8155 ± 0.0507
Financial Enriched	Random Forest	0.6889 ± 0.0775	0.6451 ± 0.1420	0.5831 ± 0.0941	0.8279 ± 0.0636
ESG + Financial	XGBoost	0.6369 ± 0.0748	0.5922 ± 0.0798	0.5740 ± 0.0836	0.7945 ± 0.0528
ESG Only	XGBoost	0.5996 ± 0.0513	0.5654 ± 0.0609	0.5535 ± 0.0712	0.7429 ± 0.0577
ESG Only	Random Forest	0.6163 ± 0.0543	0.6407 ± 0.0836	0.5491 ± 0.0650	0.7654 ± 0.0559
ESG + Financial Enriched	Logistic Regression	0.5970 ± 0.0414	0.5391 ± 0.0472	0.5317 ± 0.0446	0.7560 ± 0.0529
Financial Only	XGBoost	0.5784 ± 0.0821	0.5294 ± 0.1000	0.5109 ± 0.0848	0.7331 ± 0.0568
Financial Only	Random Forest	0.5976 ± 0.0860	0.5767 ± 0.1255	0.5107 ± 0.0902	0.7445 ± 0.0546
ESG + Financial	Logistic Regression	0.5471 ± 0.0541	0.4992 ± 0.0506	0.4930 ± 0.0564	0.7471 ± 0.0437
ESG Only	Logistic Regression	0.5396 ± 0.0557	0.4989 ± 0.0637	0.4877 ± 0.0654	0.7229 ± 0.0545
Financial Enriched	Logistic Regression	0.5337 ± 0.0552	0.5160 ± 0.0749	0.4769 ± 0.0673	0.7393 ± 0.0713
Financial Only	Logistic Regression	0.4608 ± 0.0517	0.5423 ± 0.0426	0.4416 ± 0.0498	0.7135 ± 0.0547
Financial Enriched	Naive Bayes	0.3871 ± 0.0574	0.5341 ± 0.0500	0.3844 ± 0.0581	0.6925 ± 0.0500
ESG + Financial Enriched	Naive Bayes	0.3529 ± 0.0565	0.5258 ± 0.0622	0.3464 ± 0.0565	0.6778 ± 0.0323
Financial Only	Naive Bayes	0.3233 ± 0.0682	0.5121 ± 0.0679	0.3263 ± 0.0704	0.6609 ± 0.0394
ESG + Financial	Naive Bayes	0.3209 ± 0.0551	0.5177 ± 0.0723	0.3183 ± 0.0555	0.6525 ± 0.0494
ESG Only	Naive Bayes	0.2063 ± 0.1237	0.3851 ± 0.1315	0.1890 ± 0.1019	0.6260 ± 0.0829

Table VII: All configurations — Stratified Split (10 seeds).

Feature set	Model	Accuracy	$\text{Precision}_{\text{macro}}$	F1_{macro}	AUC
ESG + Financial Enriched	XGBoost	0.7351 ± 0.0393	0.6823 ± 0.0526	0.6527 ± 0.0488	0.8539 ± 0.0174
Financial Enriched	XGBoost	0.6909 ± 0.0644	0.6581 ± 0.0896	0.6287 ± 0.0765	0.8074 ± 0.0324
ESG + Financial Enriched	Random Forest	0.7403 ± 0.0437	0.7642 ± 0.0922	0.6262 ± 0.0512	0.8849 ± 0.0117
ESG + Financial	XGBoost	0.6805 ± 0.0246	0.5995 ± 0.0649	0.5743 ± 0.0443	0.8033 ± 0.0185
ESG + Financial	Random Forest	0.6974 ± 0.0362	0.6716 ± 0.1280	0.5714 ± 0.0623	0.8285 ± 0.0268
ESG Only	Random Forest	0.6753 ± 0.0362	0.6785 ± 0.1220	0.5689 ± 0.0709	0.8034 ± 0.0344
ESG Only	XGBoost	0.6610 ± 0.0375	0.5947 ± 0.0691	0.5658 ± 0.0587	0.7639 ± 0.0354
Financial Enriched	Random Forest	0.7065 ± 0.0637	0.6578 ± 0.1576	0.5643 ± 0.0763	0.8397 ± 0.0239
Financial Only	XGBoost	0.6195 ± 0.0578	0.5493 ± 0.0875	0.5351 ± 0.0719	0.7499 ± 0.0322
ESG + Financial Enriched	Logistic Regression	0.6026 ± 0.0383	0.5190 ± 0.0411	0.5163 ± 0.0413	0.7570 ± 0.0500
Financial Only	Random Forest	0.6364 ± 0.0561	0.5557 ± 0.1128	0.5149 ± 0.0696	0.7828 ± 0.0403
ESG + Financial	Logistic Regression	0.5584 ± 0.0659	0.5031 ± 0.0548	0.4922 ± 0.0600	0.7406 ± 0.0444
ESG Only	Logistic Regression	0.5545 ± 0.0662	0.4964 ± 0.0535	0.4859 ± 0.0625	0.7193 ± 0.0473
Financial Enriched	Logistic Regression	0.5234 ± 0.0494	0.5058 ± 0.0451	0.4738 ± 0.0458	0.7205 ± 0.0558
Financial Only	Logistic Regression	0.4390 ± 0.0577	0.5214 ± 0.0398	0.4187 ± 0.0506	0.6995 ± 0.0465
Financial Enriched	Naive Bayes	0.3364 ± 0.0601	0.4708 ± 0.0595	0.3326 ± 0.0548	0.6432 ± 0.0529
ESG + Financial Enriched	Naive Bayes	0.3364 ± 0.0738	0.4940 ± 0.0598	0.3243 ± 0.0546	0.6592 ± 0.0455
ESG + Financial	Naive Bayes	0.3117 ± 0.0490	0.5089 ± 0.0712	0.2994 ± 0.0373	0.6438 ± 0.0468
Financial Only	Naive Bayes	0.2922 ± 0.0455	0.4502 ± 0.0493	0.2918 ± 0.0449	0.6373 ± 0.0334
ESG Only	Naive Bayes	0.3117 ± 0.1751	0.4681 ± 0.0751	0.2652 ± 0.1283	0.6673 ± 0.0512

B. Reclassification Summary (Random Forest)

This appendix reports the issuer-level reclassification analysis underlying the discussion. Predictions from the Random Forest model trained on the `financial_enriched` feature set (baseline, A) are compared with predictions from the same model trained on the `esg_financial_enriched` feature set (comparison, B), evaluated on the same held-out test rows across ten random seeds $(0, \ldots, 9)$. A reclassification is counted whenever a test-row prediction differs between A and B. HY$\rightarrow$IG denotes a row predicted as High Yield by A and as any Investment Grade class by B; IG$\rightarrow$HY denotes the reverse direction. High Yield detection uses the case-insensitive substring high yield on the predicted class label. Beyond the aggregate counts, we additionally classify each reclassification by which boundary it crosses High Yield $\leftrightarrow$ Investment Grade, or within Investment Grade (High Risk $\leftrightarrow$ Low Risk) map it back to its issuer via the row index, and record whether the move is *toward* or *away from* the issuer's true class. Results are pooled across the ten seeds and both split strategies.

Table VIII: Average reclassification counts per seed, by split strategy. All averages computed over $n = 10$ seeds.

Split	Seeds	Avg. n	Avg. recl.	HY→IG	IG→HY	Rate
Grouped	10	77.4	14.5	1.7	1.6	0.188
Stratified	10	77.0	11.6	1.2	1.6	0.151

Table IX: Seed-level reclassification details for the Random Forest model, comparing predictions from the `financial_enriched` baseline against the `esg_financial_enriched` configuration on identical test rows. The *Dropped* column reports rows excluded due to test-set misalignment between the two configurations (zero in all cases).

Split	Seed	n Compared	Dropped	Total Reclassified	HY→IG	IG→HY	Reclass. Rate
Grouped	0	76	0	19	4	1	0.250
Grouped	1	76	0	14	4	0	0.184
Grouped	2	75	0	15	0	1	0.200
Grouped	3	77	0	12	1	2	0.156
Grouped	4	77	0	15	2	2	0.195
Grouped	5	77	0	16	1	2	0.208
Grouped	6	81	0	13	0	1	0.160
Grouped	7	81	0	14	3	1	0.173
Grouped	8	76	0	10	0	2	0.132
Grouped	9	78	0	17	2	4	0.218
Stratified	0	77	0	10	2	1	0.130
Stratified	1	77	0	12	3	1	0.156
Stratified	2	77	0	15	0	1	0.195
Stratified	3	77	0	12	1	1	0.156
Stratified	4	77	0	10	0	1	0.130
Stratified	5	77	0	9	2	1	0.117
Stratified	6	77	0	9	2	1	0.117
Stratified	7	77	0	13	0	6	0.169
Stratified	8	77	0	10	0	1	0.130
Stratified	9	77	0	16	2	2	0.208

Table X: Issuers whose Random Forest prediction crosses a class boundary when ESG features are added, ranked by crossings (pooled over 10 seeds $\times$ 2 splits). "Net" is correct minus incorrect boundary moves (positive = ESG helps).

Issuer	Rating	True class	Crossings	Net
HY ↔ IG boundary (29 issuers, 60 crossings)				
Hörmann Industries GmbH	BB	High Yield	7	+7
Ziklo Bank AB	A3	IG Low Risk	6	+6
Ahlstrom-Munksjö Holding 3 Oy	B–	High Yield	5	+5
Nortura SA	BB+	High Yield	4	−4
Tryg Forsikring A/S	A2	IG Low Risk	4	+4
Akelius Residential Prop. AB	BBB–	IG High Risk	3	−3
Preem Holding AB	Ba3	High Yield	3	+3
Within-IG boundary (97 issuers, 200 crossings)				
Jernhusen AB	A	IG Low Risk	9	+9
Danfoss A/S	BBB	IG High Risk	6	+6
Jernhusen AB (2025)	A	IG Low Risk	6	+6
Suomen Hypoteekkiyhdistys	BBB	IG High Risk	6	−6
Norwegian Property ASA	BBB–	IG High Risk	5	+5
Rikshem AB	A3	IG Low Risk	5	+5
Fjellinjen AS	AA	IG Low Risk	5	−5

TRITA-EECS-EX-2026:409
Stockholm, Sweden 2026
www.kth.se