UNIVERSITÀ
DI PAVIA

Dipartimento di Scienze Economiche e

Aziendali Corso di Laurea magistrale in

FINANCE

SAFE AI IN CREDIT RATING SYSTEMS

SICURA NEI SISTEMI DI CREDIT SCORING

Relatore: Paolo Stefano Giudici

Chiar.mo Prof.

Tesi di Laurea
di Mina Sahim

Matr. N 513751

UNIVERSITÀ
DI PAVIA

Anno Accademico 2023-2024

Table of Contents

Chapter 1: Introduction

1.1. Background and Purpose of the Study **p.6**

1.2. Problem Statement and Research Questions **p.7**

1.3. Significance and Contribution of the Study **p.8**

Chapter 2: Literature Review

2.1. Artificial Intelligence and Safety Principles **p.9**

2.1.1 Artificial General Intelligence...**p.9**

2.1.2 The Concept of Safe AI **p.13**

2.1.3 Explainability **p.13**

2.1.4 Fairness **p.13**

2.1.5 Reliability and Robustness **p.14**

2.1.6 Accuracy...**p.14**

2.2 Credit Rating Systems and ESG Factors in Credit Ratings **p.14**

2.3 Performance Metrics **p.15**

UNIVERSITÀ
DI PAVIA

2.3.1 Confusion Matrix..**p.15**

2.3.2. Evaluating Models with ROC-AUC Metrics **p.16**

2.4 Explainability Metrics: Gini and SHAP **p.18**

Chapter 3: Methodology

3.1. Dataset and Features **p.20**

3.1.1. ESG Factors (Environmental, Social, Governance) **p.20**

3.1.2. Financial Variables... **p.24**

3.1.3. Credit Rating Data (sco, s22) **p.25**

3.2 Data Preprocessing Steps **p.25**

3.2.1 Analysis Objective **p.25**

3.2.2 Splitting Data into Training and Testing Sets **p.28**

3.3 Applied Models **p.28**

3.4 Explainability Metrics: SHAP and Gini-Based Analyses **p.29**

3.5 Performance Metrics **p.29**

3.5.1 Confusion Matrix **p.29**

3.5.3 ROC-AUC Analysis... **p.29**

Chapter 4: Analysis and Results

4.1 General Evaluation of Model Performance **p.31**

4.2 Explainability Analyses **p.34**

4.2.1 Shapley Values.. **p.35**

4.2.2 Beta Coefficient Analysis **p.39**

4.2.3 Mean Decrease Gini Results **p.42**

4.3 Impact of ESG Factors on Credit Ratings **p.45**

4.4 Fairness and Bias Tests **p.47**

4.4.1 False Positive and False Negative Analysis **p.47**

4.4.2 Class-Based Performance Comparisons **p.52**

4.5 ROC and AUC Analysis **p.55**

Chapter 5: Discussion and Evaluation

5.1. Interpretation of Model Performance **p.64**

5.2. Role of ESG Factors **p.64**

5.3. Limitations of the Study and Overcoming Data Limitations **p.65**

Chapter 6: Conclusion and Recommendations

6.1. Key Findings of the Study **p.68**

6.2 Recommendations and Future Work **p.68**

Appendices and References

- Code Used for Analysis **p.70**
- **References** **p.77**

UNIVERSITÀ
DI PAVIA

Lo scopo di questa tesi è esaminare come i principi dell'Intelligenza Artificiale Sicura (Safe AI) vengano applicati nei sistemi di valutazione del credito e, in questo processo, investigare l'impatto dei fattori ESG, oltre ai tradizionali dati finanziari, per creare un modello di valutazione del credito più trasparente e affidabile utilizzando le tecniche di Safe AI. Utilizzando i dati forniti da Modefinance, i principi di Safe AI saranno applicati con i dati disponibili, mentre l'impatto dei fattori ESG sarà esaminato in conformità con questi principi. Questa ricerca fornirà un'idea di quali principi necessitano di essere applicati meglio e, alla fine dello studio, si farà riferimento anche a come il dataset possa essere migliorato.

UNIVERSITÀ
DI PAVIA

Chapter 1: Introduction

1.1 Background and Purpose of the Study

The increasing use of artificial intelligence within the financial sector is giving rise to credit scoring models, with the potential to play a critical role in risk assessment. The application of Safe AI in financials will be critical for an effective process. In such a process, one will have to make sure, particularly in sensitive use cases such as credit rating, that artificial intelligence-powered systems deliver robustness,explainability,fairness, and accuracy.Credit rating systems provide a basis for individuals and organizations to make financial decisions by determining their risk profiles. Furthermore, the increasing importance of ESG – namely environmental, social and governance – factors has created new challenges and opportunities for credit rating systems. Sustainability, ethical governance and social responsibility, which are increasingly required within the scope of ESG factors, are good examples of this. However, to date, there has been little exploration of how ESG factors impact the fairness and explainability of machine learning models.

The aim of this thesis is to examine how Safe AI principles are applied in credit rating systems and in this process, to investigate the impact of ESG factors in addition to traditional financial data to create a more transparent and reliable credit rating model using Safe AI techniques.

Using the data provided by Modefinance, SAFE AI principles will be applied with the data available while using classification methods and the impact of ESG factors will be examined in accordance with these principles. This research will provide an idea of which principles need to be better applied and at the end of the study will also make reference to how the dataset can be improved.

1.2. Problem Statement and Research Questions

The increasing use of AI in financial decision-making raises critical questions about the trustworthiness of these systems. Despite advanced technology, the lack of adequate reliability metrics on AI models makes it difficult to understand how decisions such as credit scores are assigned and can lead to incorrect assignments. Lack of explainability, potential biases and fairness concerns, uncertainty about the role of ESG factors in improving credit scores despite their role as indicators of sustainable financial performance, the need for credit systems to remain reliable and robust across different scenarios and data distributions (such as when ESG factors are included), and accuracy are all very important issues to consider and are essential for the reliability of the use of AI in the financial sector.

This thesis will estimate ESG factors' contribution when mixed with financial information in a Random Forest model and reveal ESG factors' contribution through interpretability techniques. It will evaluate credit rating model sensitivities with regard to changing input information, for example, not including ESG factors, and, in such a way, validate the system under changing scenarios.

It will investigate whether AI-powered credit rating models will be fair for all classes and with whose fairness metrics, for example, with values of false positive and false negative in various credit classes.

1.3. Significance and Contribution of the Study

This research is important in the practical world of finance.

By combining the principles of safe AI such as explainability, fairness, robustness, and accuracy, with credit rating systems, how environmental, social and governance (ESG) factors can play a role in making these systems fairer and more reliable is investigated. The aim is to make advanced AI techniques more understandable and accessible, helping see how decisions are made in financial modeling and how reliable it is to provide this through artificial intelligence.

UNIVERSITÀ
DI PAVIA

Chapter 2: Literature Review

2.1. Artificial Intelligence and Safety Principles

2.1.1 Artificial General Intelligence

"Artificial General Intelligence" (AGI) is a type of intelligent agent that has the ability to learn and adapt, and can learn new information and perform intellectual tasks thanks to the ability to learn from experience.

There are variable estimates regarding when development of AGI can be achieved. Some believe it can happen in a matter of years, but others believe it will require a long period of time.

How much of a threat AGI could possibly present to humanity is also debated, and it is considered in regard to whether uncontrolled creation of AGI can have severe repercussions

There are a number of concepts used to describe systems with different capabilities and features in the field of artificial intelligence.

While strong AI is considered to be capable of performing a variety of cognitive tasks like a human,weak AI is an AI system designed for a specific task and is usually specialized in a particular field without possessing cognitive abilities. Artificial superintelligence differs from humans' perception and cognitive abilities. The concept of transformative artificial intelligence refers to types of artificial intelligence that can significantly change business processes, lifestyles or other areas, similar to an industrial revolution.

Tests for Human-level AGI

The Turing Test

The Turing test, which has more than one version, is a behavioral test rather than measuring actual intelligence and consciousness. The test is based on determining how successful the computer is in behaving like a human. If the evaluator cannot distinguish the program from a real person, the computer program passes the Turing test.

The Robot College Student Test

The Robot College Student Test is a test for assessing high-level achievement in artificial intelligence. In this test, an artificial intelligence system seeks admission in a university, studies the same subjects as humans, and seeks a degree through successful completion of these subjects.

The purpose of this test is to assess the capability of AI in performing such complex mental operations. On the other hand, the Robotics College Student Test aims at testing the capability of AI in similar performances of complex mental operations and can assess complex information processing capabilities. Passing through this will mean that the AI possesses deep learning capabilities and problem-solving capabilities, and in that case, it approaches the actual objective of AGI. But for testing general cognitive capabilities, such complex and comprehensive tests would be required. ChatGPT, most common in current times, works well with some types of tests but less effectively with other types of tests.

The Employment Test (Nilsson)

Artificial intelligence is becoming widespread in many areas of the labor market, from marketing to restaurants, even couriers and security guards. Companies like Knightscope even have robots that monitor and report on security missions.This test tries to measure the ability of artificial intelligence systems to do a certain job with the same efficiency as humans, and since efficiency is a criterion that can be used when evaluating whether it can replace human workers, if this test is passed, it can be considered that the system is suitable for that job.

Ikea Test (Marcus)

In this test scenario, an AI system follows instructions to properly assemble an Ikea flat-pack piece of furniture. The successful assembly of furniture by an artificial intelligence system may demonstrate its ability to combine various abilities such as problem solving, image processing, and robot control.

Coffee Test (Wozniak)

The purpose of the test involves following the steps of the coffee-making process to evaluate an artificial intelligence's ability to perform a practical task and demonstrating its intelligence by determining all the ingredients in the right amount in this process. Since this test is not fully completed, it is a good opportunity to observe the development of artificial intelligence.

<u>Brain Simulation and Emulation</u>

Brain simulation and emulation can play a role in many areas such as developing artificial intelligence and robotic systems. Kurzweil argues that the most important way to create machine intelligence is to understand human intelligence and to image the brain is the first step for this.

A brain simulation is created by biologically starting with a detailed scan and mapping of the brain,copying that map to a computing device and creating a model of the brain that is simulated,sufficiently faithful to the original that it behaves in much the same way as the original brain.

Considering the number of synapses in the brain for such simulations, serious computing power is required. Therefore, these researches develop with the continuous development of computer technology.

Benefits and Risks

Benefits

Artificial intelligence has led to serious advances in many areas, especially education and training,industry and production, automation and improvement in healthcare, and it has the potential for more. It has benefits in many areas, from fighting against cancer to improving educational quality.

While it improves efficiency in many tasks, it can also reduce some risks with the use of technologies such as nanotechnology. Artificial intelligence can play an important role in increasing our quality of life when used efficiently.

Risks

Artificial intelligence experts express their concerns about the potential risks brought by artificial intelligence technology. Some of these concerns are unemployment concerns, data privacy and security, bias and injustice, addiction and loss of control, technological conflict and security.These concerns are important issues that should be taken into consideration by experts and the public.

2.1.2 The Concept of Safe AI

Safe AI is the development of artificial intelligence (AI) systems determined by sustainability, accuracy,explainability,fairness and robustness.These ideas are very applicable to credit scoring systems.

2.1.3. Explainability

Explainability is the facility of artificial intelligence models, to explain how a model arrives at the output, to online interpret the reasoning used to arrive at the output. There are various models for interpretation, depending on the level of complexity of the model and the type of task that is to be solved.

2.1.4. Fairness

Fairness guarantees do not allow any biasing to be inserted via the AI models. Discrimination in credit scoring can be defined as differential false positive rates (FPR) and false negative rates (FNR) for the classes

Fairness means that the model can work impartially on all classes and groups.

Class specific: Metrics such as precision, recall and balanced accuracy calculated for each class. For example, if the "D" class has lower precision or recall values than any other class, then it indicates insecurity about the model performance in producing D. These analyzes will reveal the possibility that the model discriminates in certain classes.False positive and false negative rates were calculated and visualized. These metrics determine in which classes the model is likely to make more errors.In the interest of fairness, having very high FP or FN in a class may hint at potential bias of the model.

2.1.5 Reliability and Robustness

Reliability creates consistent output, but robustness tests how well models can manage with variation or unpredictable variation in the data.

2.1.6. Accuracy

Accuracy represents the predictive ability of the model and is evaluated using metrics such as confusion matrix and ROC-AUC. It reports how similar the predicted results are to the direct result.

2.2 Credit Rating Systems and ESG Factors in Credit Ratings

Environmental, social, and governance (ESG) factors, which measure long-term sustainability risk, are now incorporated in credit rating models.ESG represents an assessment of creditworthiness more comprehensive than the standard financial one.

Modefinance is a fintech company that focuses on credit risk while combining traditional financial metrics and the latest machine learning techniques. It provides a detailed dataset that will be considered a comprehensive credit rating atmosphere to examine the integration of ESG factors in traceability and to test the performance of the model according to appropriate safe artificial intelligence principles.

As a framework for potentially more sustainable and socially responsible financial practices, ESG factors are increasingly important to investors and regulators. To elaborate on these factors, we can include:

Environmental; carbon footprint, resource efficiency and climate change minimization,

Social; employee well-being, diversity and community engagement,

Corporate governance practices, in addition to governance; transparency and ethical behavior.

Considering that these factors are included in the Modefinance data set, the effects of ESG factors on credit scores can be measured. In this way, it is checked whether these factors contribute to model accuracy, explainability and fairness in credit rating systems and understand robustness.

2.3 Performance Metrics

2.3.1 Confusion Matrix

Describes in detail the procedure for forecasting each class of credit rating which included considering false positives and negatives, respectively. Accuracy is measured both mean and class-conditional predictive power, so that the model is robust for decision-making.

2.3.2 Evaluating Models with ROC-AUC Metrics

AUC-ROC (Area Under the Curve - Receiver Operating Characteristic)

Receiver Operating Characteristic (ROC) plots are an effective tool for visualizing and organizing the performance of classifiers. These plots show the balance between the true positive rate and the false positive rate. ROC analysis is also widely used in medical decision-making processes and has gained importance in machine learning and data mining research in recent years. Measures such as area under ROC curve (AUC) have been utilized in an attempt to represent performance of classifiers in a scalar value. Misinterpretation of receiver operating characteristic plots can, nevertheless, produce incorrect values in comparisons between classifiers. In multi-class cases, receiver operating characteristic analysis can become cumbersome, and careful consideration must, therefore, be taken in evaluation. Overall, use of receiver operating characteristic plots helps in a deeper analysis of performance of classifiers and in arriving at correct conclusions in evaluations.

The curve of ROC plots TPR in relation to FPR at different thresholds, and the AUC (Area Under Curve) plots overall performance capability of model in distinguishing between classes since AUC = 1 is Ideal model while AUC = 0.5 is random chance guessing. The AUC is defined by the Receiver Operating Characteristic (ROC) curve and measures the quality of the classification model. The curve is constructed by comparing the cumulative distribution functions of the predicted probabilities for class 0 and class 1. A good classifier has its curve in the upper left triangle within the unit square. This indicates that class 1 points have lower prediction probabilities than class 0 points. A curve that runs along the left and upper edges indicates perfect separation. A classification performance equal to chance is indicated by the curve following the diagonal. If the curve lies below the diagonal line, it indicates that the classification model is poor. The AUC is defined as the area under the ROC curve and is often used to summarize the overall performance of the model. Some authors plot the ROC curve by changing the true positive and false positive rates on the axes. However, both representations give similar results and are only a conventional approximation.

Metrics Derived from Confusion Matrix and AUC-ROC

Recall (Sensitivity) measures the rate at which true positive samples are correctly predicted by predicted positive samples.

It represents the rate at which true positive samples are predicted.As a well-developed feature, it is important to know how to predict the relevant samples. Precision measures the rate at which predicted positive samples are correctly predicted by true positive samples.

It represents the accuracy rate of predicted positive samples.

Both metrics have been calculated excluding samples with a negative label when comparing positive samples with predicted samples.

Both predicted samples and positive samples have been compared in evaluation, excluding both use of samples with a negative label in both calculation of both AUC-ROC and confusion matrices.

AUC-ROC and confusion matrices have, in addition, helped in communicating model performance measures to stakeholders in terms of explainability, a safe AI practice.

2.4 Explainability Metrics: SHAP,Gini and Beta

The measurement of feature importance consists of determining how much model prediction error goes up after values of a feature are shuffled. If shuffling a feature causes an increase in model error, the feature is deemed important, meaning the model relied on it for its predictions. If shuffling doesn't change the model's error, then such a feature is not relevant in that model did not utilize it. For a feature to have importance, it is most critical for the model to produce correct and reliable output.

SHAP (Shapley Additive Explanations)
SHAP provides local explainability by measuring feature contributions for individual predictions.It assigns a consistent “credit” to features based on their marginal impact on model predictions.
The best model for SHAP is the Random Forest. SHAP makes predictions interpretable by explaining them as feature contribution.
SHAP revealed individual-level explanations that conventional importance metrics couldn’t provide.

Gini Importance

Gini importance is a global explainability metric. It identifies each feature's contribution towards decreasing uncertainty (or impurity) in all decision trees in ensemble models.Random Forest, Classification Trees, and Bagging are the best models for gini importance, as gini importance measures one variable's ability to split data in model training.Comparing with SHAP, Gini provides feature importance at a composite level, while SHAP provides individual-dependent insights.

Beta Coefficients

Beta coefficients evaluate both direction and strength of association between each of the predictor variable and the target variable. It is most suited for Logistic Regression.

UNIVERSITÀ
DI PAVIA

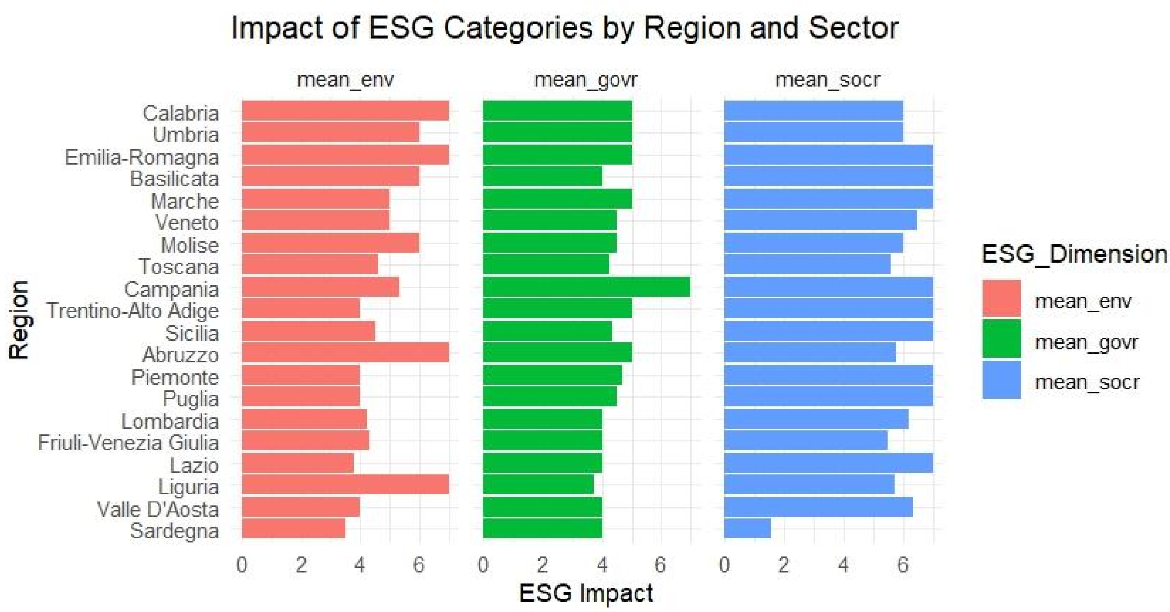
Chapter 3: Methodology

3.1. Dataset and Features

3.1.1 Esg Factors

In this study, it is shown through two scenarios whether these ESG factors are added or not. Changes due to subtraction and/or addition of ESG, accuracy, Gini importance, and statistics have been taken into consideration.

Application of ESG (Environmental, Social, and Governance) factors in Random Forest model will allow model performance evaluation through examination of such factors' contribution.



ESG (Environmental, Social and Governance) Variables

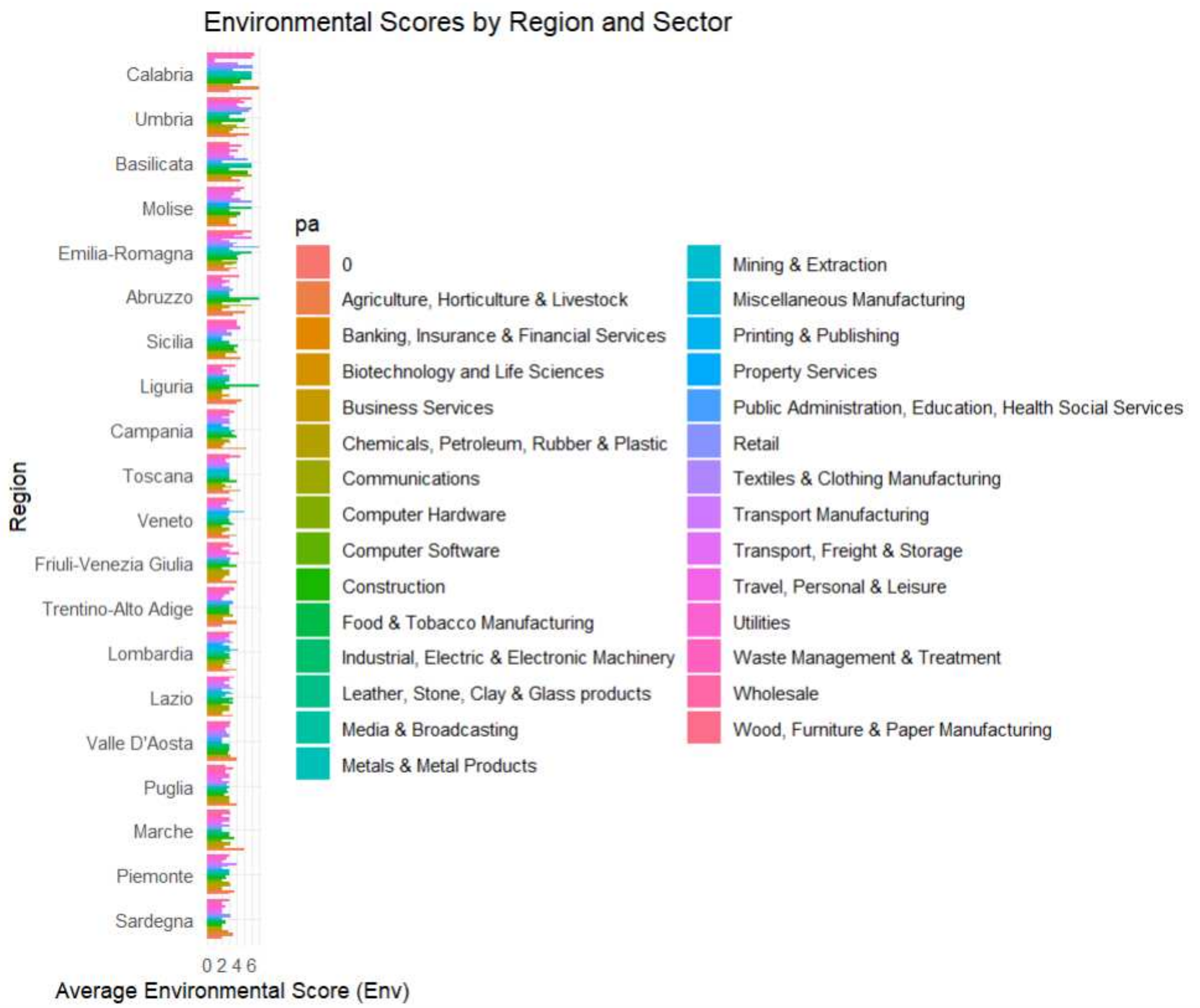
These are categorical variables that gives an allowance for the evaluation of the firm in regards to ESG performance.ESG risks are starting to gain relevance and even serve to the evaluation of companies' sustainability risk related to the credit risk.

Environmental: A measure of the degree of environmental stance and also measures environmental threats.

Social: This pertains to how the company has fared in social responsibility matters.

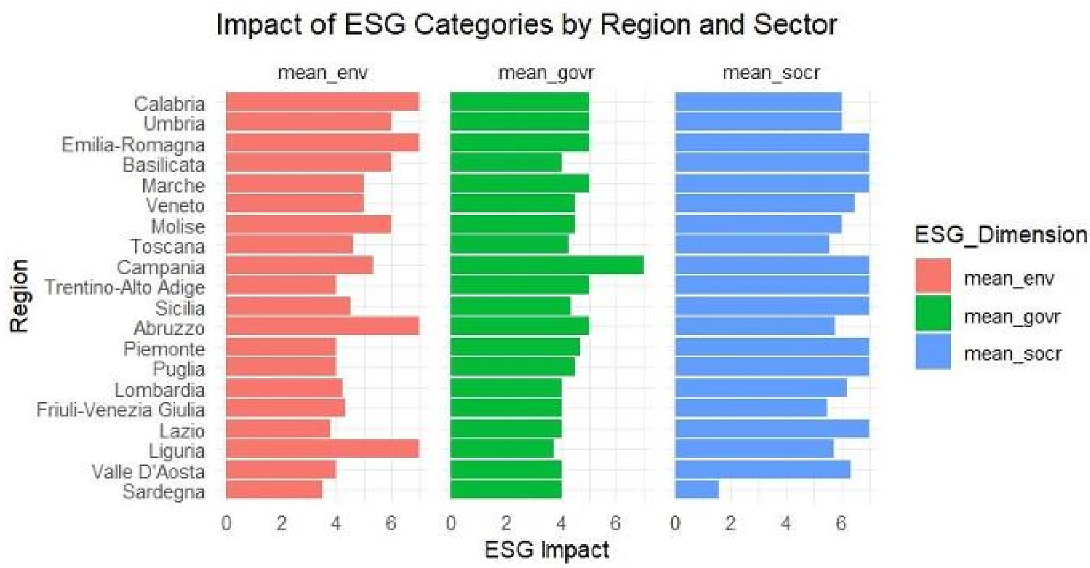
Governance: Shows evidence of the existence of good governance principles.

UNIVERSITÀ
DI PAVIA



This chart shows environmentally high-scoring Italian regions and sectors.

High-scoring regions and sectors, such as Basilicata, Umbria, and Calabria, and low-scoring regions and sectors, such as Sardegna and Piemonte, represent regions with room for improvement in environment policies and practice.



This chart plots three categories' mean ESG for Italian regions: social, governance, and environment.

Environmental Impact :

The environmental dimension varies somewhat, with locations like Calabria, Abruzzo and Emilia-Romagna having a positive rating. That would mean that these locations have a high concern for, or are doing a good job in, environmental statistics.

Governance Impact :

The governance is fairly balanced, but regions such as Campania clearly outshine others, and one can have a sensation that effective governance works in such a region.

Social Impact :

Social dimension scores are also relatively even across all regions, although some divergence exists among them.

3.1.2 Financial Variables (by Year)

These variables are utilized to calculate companies' financial performance on an annual basis. Financial information can be directly correlated with companies' credit ratings and plays a key role in credit risk analysis

- ta22, ta21, ta20: Displays total assets for 2022, 2021 and 2020 years respectively. These variables, representing the size and assets base of the company, utilize to evaluate the financial solidity of the company.

- ca22, ca21, ca20: Current assets reveal the liquidity position of the company and represent its capability to settle short-term liabilities

- shf22, shf21, shf20: Represent the stock price

- cl22, cl21, cl20: Displays current liabilities. It is a key indicator of the company's capability in terms of short-term payments and financial risk position

- opr22, opr21, opr20: It is operational income for 2022, 2021 and 2020 years respectively. It is an indicator of the level of income earned by the company through its key operations

- ebit22, ebit21, ebit20: It is profit before interest and taxes (EBIT). It is utilized to evaluate operational efficiency through representing the gain earned by the company through its key operations

- ni22, ni21, ni20: Net income reflects the gain earned by the company after tax. Net income fluctuations over years represent an indicator of the profitability performance of the company

- ebda22, ebda21, ebda20: It is referred to as income before depreciation (EBITDA). It is utilized to reveal the operational cash flow and cost structure of the company

These variables form inputs to the machine learning algorithms utilized in the study, and performance is monitored in analysis conducted for enhancing the accuracy of the model in estimating credit scores. Employing financial statistics at an annual level and ESG factors helps assess credit ratings not only in relation to past financial statistics but even in relation to long-term performance markers such as sustainability.

3.1.3 Credit Rating Data

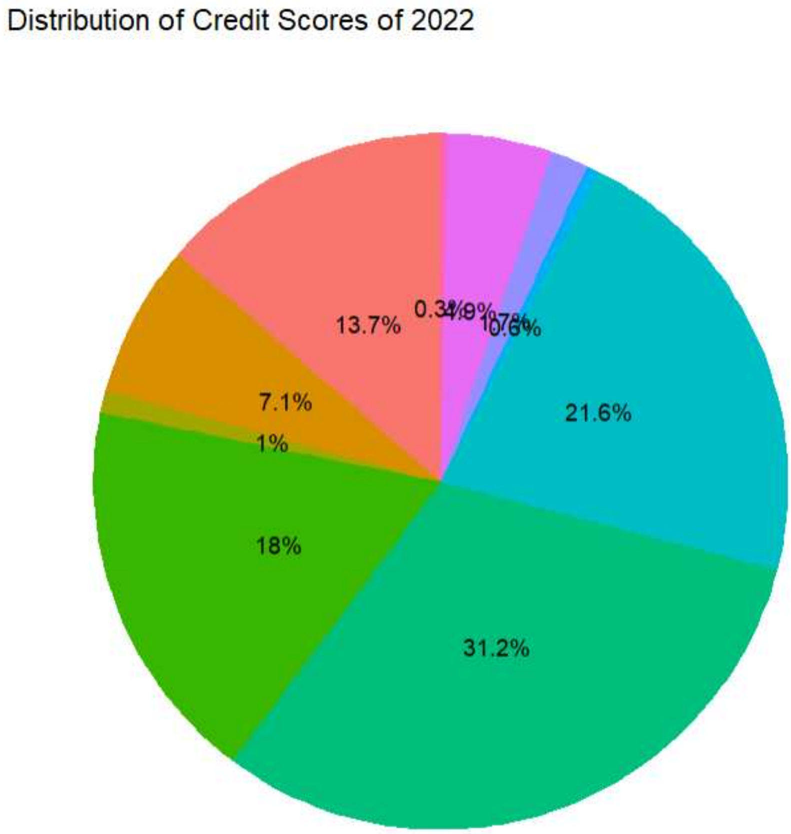
The sco variable is the main target variable used in the study and is represented by scoring the credit score. It is used categorically from AAA to D and this is detailed in the following sections of the study. s22 represents the credit scores for 2022.

3.2 Data Preprocessing

3.2.1 Analysis Objective

Objective is to identify the optimal explanatory factors and the find best-fitting function affecting credit ratings in 2022. By finding the most accurate model with the data, tests can be performed on the reliability and accuracy of the model. After testing safe AI metrics in all models, the effect of ESG factors will be tested in the best model.

The illustration below demonstrates the relative dispersion of credit scores by year in the year 2022, the test data. CC has the highest distribution among speculative classes, BB among non-speculative classes, and non-speculative are dominant in ratios.



As can be seen, CC has the highest distribution among speculative classes, BB has the highest distribution among non-speculative classes, and non-speculative classes are dominant in terms of ratio.

UNIVERSITÀ
DI PAVIA

3.2.2 Splitting Data into Training and Testing Sets

The dataset is partitioned into training and testing datasets for model training and evaluation in follow-up sessions. Since the data is represented in terms of years, training and testing datasets are distinguished according to respective years. The data is partitioned in terms of years.

While checking the dataset, confirmed no missing values, outliers, and other defects in the dataset.

The data is partitioned in terms of years, a single column is extracted, column headings are reabeled in an apt manner, and datasets for single years are arranged. "select" function is utilized in creating datasets for years in view: 2022, 2021, and 2020. Two datasets for 2021 and 2020 years are considered for training, and 2022 years' information is considered for testing.

3.3 Applied Models

The study aims to choose the best model before ESG factors are involved and after that measuring effect of ESG while using SAFE AI metrics.Forward & Backward stepwise selection,multinomial regression,lasso regression,bagging model,random forest model,classification tree model are applied. Since random forest was chosen as the best model, it was used in the study on ESG factors. Random forests are a fusion of tree predictors based on the assumption that the value of each tree is affected by a random vector that is independently sampled and follows the same distribution for all trees in the forest.

The Random Forest model is one of the most commonly used and effective models when it comes to making predictions, one of its benefits being that it eliminates overfitting because the theorem of large numbers governs it.

3.4 Explainability Metrics:Shap and Gini-Based Analyses

Before adding ESG factors, measurement was made on financial variables from all models. In this way, it was understood how important each financial variable was. After adding ESG factors, the importance order of ESG factors was determined according to the Random forest model with Gini-based measurement.

3.5 Performance Metrics

3.5.1 Confusion Matrix

In this study, accuracy and also fairness was calculated using a confusion matrix and accuracy changes within ESG were examined.

Metrics such as FPR, FNR, and precision were used to assess fairness across credit score categories.The confusion matrix indicates potential biases in false positive/negative rates across credit rating classes and showed the balance between true positives and false positives for each credit rating class in this study. The ability of the models to predict all classes is tested.

3.5.2 ROC-AUC Analysis

Accuracy was analyzed by confusion matrices and ROC-AUC metrics to provide a detailed assessment of performance within credit score classes and to denote discriminatory power on an overall scale.

This means that the AUC-ROC measures were addressed to discriminate between speculative and non-speculative rating. The model was capable of distinguishing between the rating of speculative and non-speculative nature, and for individual performance at a single class level. The Speculative vs. Non-Speculative Rating is an ROC-AUC classifier performance measuring the ability of the model to distinguish between the speculative classes and non-speculative classes. High value for AUC means that the model effectively differentiates between the two groups, while low value for AUC means there is no discrimination between the two rating types.In terms of the separation of each single class, ROC-AUC measured the performance of the model, such as separating one class from all others. In the case of this study, the speculative rating classes earned lower AUC points primarily due to data scarcity, and higher points for non-speculative classes.

UNIVERSITÀ
DI PAVIA

Chapter 4: Analysis and Results

4.1 General Evaluation of Model Performance

In this study, while choosing the ideal model for credit scoring, observations were made between multinomial logistic regression, forward selection,backward selection,random forest,classification tree,bagging and lasso models

Now entering the analysis phase, AIC (Akaike Information Criterion) is chosen as the parameter to evaluate the regression results. The following regression models have been selected for analysis:

Forward & Backward Stepwise Selection:

```
Coefficients:
    (Intercept)            ta            ca           shf            cl
AA    -0.5836405 -2.066317e-04  3.000021e-05  0.0001910497  1.239033e-04
AAA   -2.4439542 -2.364760e-04  4.276900e-05  0.0001922482 -6.305787e-05
B      0.7167334  2.270616e-04 -8.394725e-05 -0.0002965648 -4.185274e-06
BB     0.9433394  1.792633e-04 -3.014593e-05 -0.0002170880 -3.873033e-05
BBB    0.5667726  9.599216e-05 -1.006038e-05 -0.0001086940 -2.286276e-05
C     -1.7826974  3.036090e-04 -1.894824e-04 -0.0005554547  1.749850e-05
CC    -0.5188463  3.019161e-04 -1.548591e-04 -0.0005103803 -2.082170e-06
CCC   -0.1036268  2.779564e-04 -1.017926e-04 -0.0004508877 -9.743834e-06
D     -3.3805695  2.751002e-04 -3.090702e-04 -0.0005276105  7.676738e-05
              opr          ebit            ni          ebda
AA   6.409189e-07 -3.311557e-05  9.671702e-05  1.618592e-04
AAA  3.032783e-06  6.540978e-04  3.157793e-04 -3.761257e-04
B   -5.643362e-06 -6.844117e-04 -5.470572e-04 -1.806618e-04
BB  -4.359063e-06 -2.422925e-04 -3.176149e-04 -1.600820e-04
BBB -2.423834e-06 -1.477532e-04 -1.076715e-04 -4.771054e-05
C   -1.270507e-05 -1.609232e-03 -6.460016e-04 -4.506206e-04
CC  -1.089266e-05 -1.547165e-03 -7.067766e-04 -4.305920e-04
CCC -9.688264e-06 -1.485973e-03 -6.766115e-04 -2.941004e-04
D   -1.459663e-05 -1.694920e-03 -7.572187e-04 -5.390922e-04

Residual Deviance: 7789.46
AIC: 7951.46
```

The calculated AIC (Akaike Information Criterion) for the Backward Stepwise Selection method is 7951.56

Multinomial Regression

```
Coefficients:
    (Intercept)            ta            ca           shf            cl
AA    -0.5836405 -2.066317e-04  3.000021e-05  0.0001910497  1.239033e-04
AAA   -2.4439542 -2.364760e-04  4.276900e-05  0.0001922482 -6.305787e-05
B      0.7167334  2.270616e-04 -8.394725e-05 -0.0002965648 -4.185274e-06
BB     0.9433394  1.792633e-04 -3.014593e-05 -0.0002170880 -3.873033e-05
BBB    0.5667726  9.599216e-05 -1.006038e-05 -0.0001086940 -2.286276e-05
C     -1.7826974  3.036090e-04 -1.894824e-04 -0.0005554547  1.749850e-05
CC    -0.5188463  3.019161e-04 -1.548591e-04 -0.0005103803 -2.082170e-06
CCC   -0.1036268  2.779564e-04 -1.017926e-04 -0.0004508877 -9.743834e-06
D     -3.3805695  2.751002e-04 -3.090702e-04 -0.0005276105  7.676738e-05
              opr          ebit            ni          ebda
AA    6.409189e-07 -3.311557e-05  9.671702e-05  1.618592e-04
AAA   3.032783e-06  6.540978e-04  3.157793e-04 -3.761257e-04
B    -5.643362e-06 -6.844117e-04 -5.470572e-04 -1.806618e-04
BB   -4.359063e-06 -2.422925e-04 -3.176149e-04 -1.600820e-04
BBB  -2.423834e-06 -1.477532e-04 -1.076715e-04 -4.771054e-05
C    -1.270507e-05 -1.609232e-03 -6.460016e-04 -4.506206e-04
CC   -1.089266e-05 -1.547165e-03 -7.067766e-04 -4.305920e-04
CCC  -9.688264e-06 -1.485973e-03 -6.766115e-04 -2.941004e-04
D    -1.459663e-05 -1.694920e-03 -7.572187e-04 -5.390922e-04

Residual Deviance: 7789.46
AIC: 7951.46
```

The calculated AIC (Akaike Information Criterion) for the Multinomial Regression method is 7951.46

Lasso Regression

UNIVERSITÀ
DI PAVIA

```
Coefficients:
    (Intercept)            ta            ca           shf            cl
AA  -0.58283436 -2.061204e-04  2.995593e-05  0.0001905662  1.234650e-04
AAA -2.43988166 -2.361308e-04  4.289659e-05  0.0001918090 -6.359381e-05
B    0.71131370  2.269723e-04 -8.386150e-05 -0.0002964392 -4.170528e-06
BB   0.94356495  1.792729e-04 -3.016412e-05 -0.0002171019 -3.872065e-05
BBB  0.56426289  9.601311e-05 -1.005116e-05 -0.0001087160 -2.287518e-05
C   -1.73205416  3.041584e-04 -1.917143e-04 -0.0005560743  1.761860e-05
CC  -0.51307908  3.020935e-04 -1.551706e-04 -0.0005108658 -2.134898e-06
CCC -0.09595114  2.779820e-04 -1.019079e-04 -0.0004510750 -9.615220e-06
D   -3.42957068  2.772490e-04 -3.068681e-04 -0.0005294814  7.436801e-05
              opr          ebit            ni          ebda
AA   6.369868e-07 -3.320513e-05  9.617372e-05  1.619157e-04
AAA  3.063731e-06  6.514189e-04  3.163408e-04 -3.745156e-04
B   -5.637127e-06 -6.844998e-04 -5.473051e-04 -1.802487e-04
BB  -4.356357e-06 -2.419874e-04 -3.178737e-04 -1.601724e-04
BBB -2.422465e-06 -1.475628e-04 -1.079392e-04 -4.776118e-05
C   -1.294764e-05 -1.609147e-03 -6.460988e-04 -4.523840e-04
CC  -1.090616e-05 -1.547989e-03 -7.067100e-04 -4.313871e-04
CCC -9.739746e-06 -1.486398e-03 -6.766970e-04 -2.939932e-04
D   -1.491127e-05 -1.689885e-03 -7.555703e-04 -5.417295e-04

Residual Deviance: 7789.48
AIC: 7951.48
```

The calculated AIC (Akaike Information Criterion) for the Lasso method is 7951.48

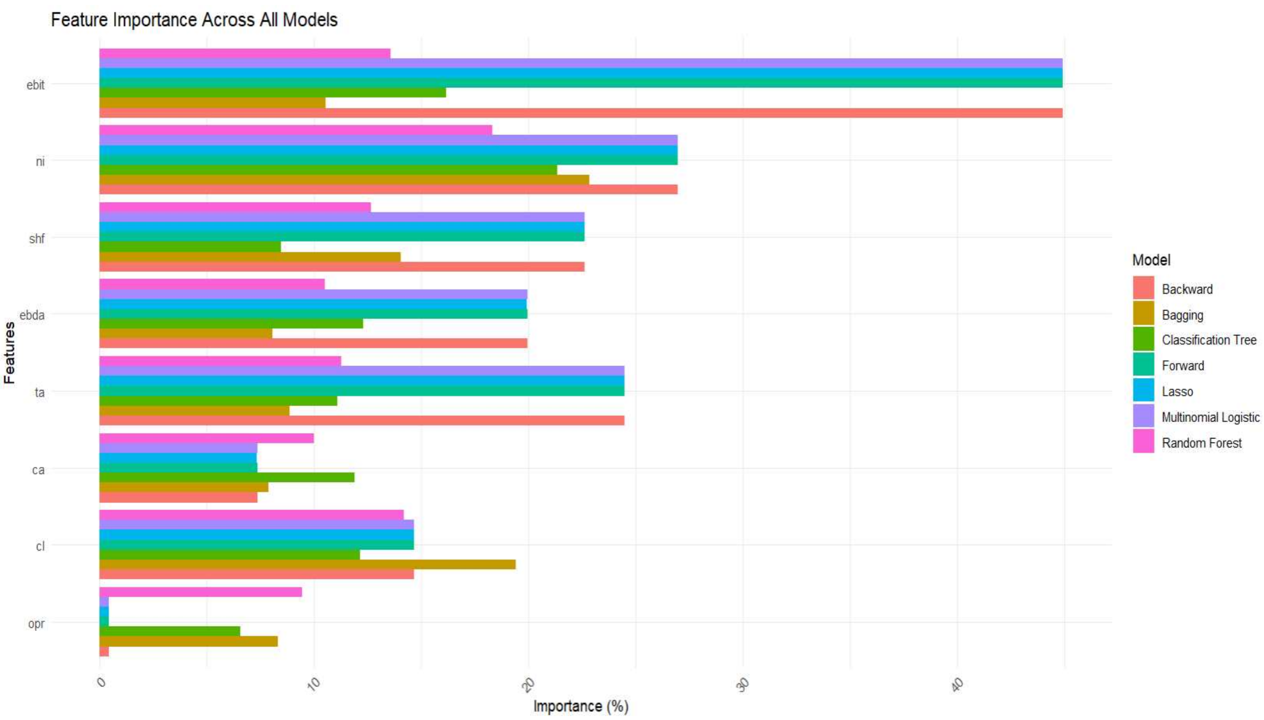
4.2 Explainability Analyses

The explainability of each model will be analyzed by the percentage distribution of variable importance. These analysis steps are as follows:

1. Calculating Variable Importances: Variable importance will be calculated for each model. Beta coefficients is used for Logistic regression and Lasso, and Mean Decrease Gini values is used for Random Forest, Bagging and Classification Tree.

2. Percentage Distribution: The impact of each variable on the model will be converted into a percentage scale and the relative weights of the variables on the model will be shown.

3. Visualization: The explainability of each model will be made more concrete by preparing bar charts showing the variable importance as a percentage.

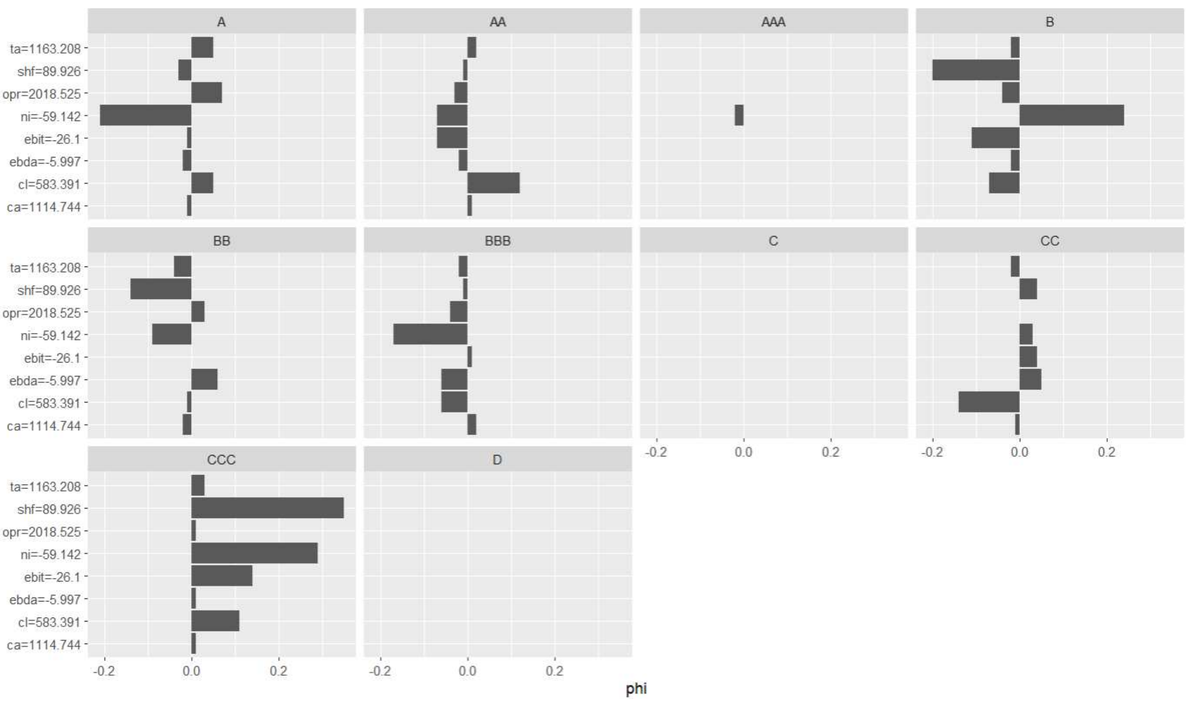


Random forest and bagging models identified current liabilities and net income to be the most influential variables amongst all independent variables, pointing to the strong influence credit ratings exert onto these variables. Clearly, focusing on variables like current libalities and net income improves the model's performance and consistency of results.

4.2.1 Shapley Values

Thanks to the Shapley value, it can explain the positive or negative contribution of each feature to the model estimate, which increases the explainability of the model. SHAP values are used to measure how important each feature is in a model. These values indicate the contribution of features in affecting the expected value of a prediction. SHAP values indicate that the order in which features are added is important and are calculated by taking the average over all possible orders. This helps us understand which features have the most impact on the prediction, giving a better understanding of how the model works.

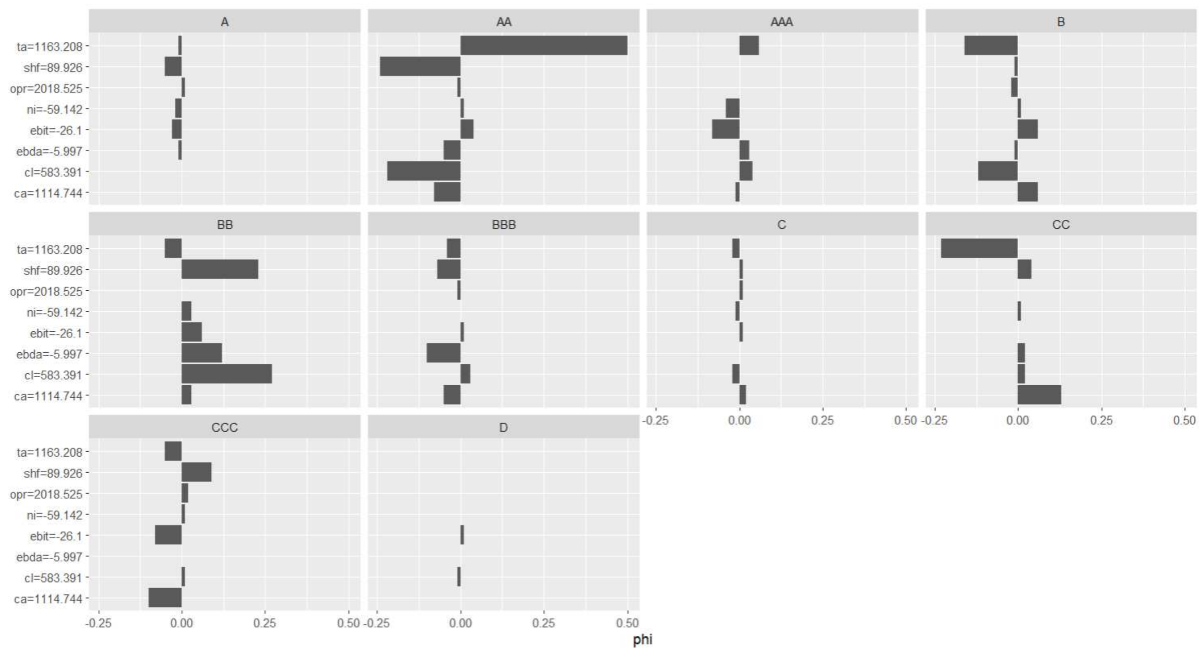
Random Forest Model



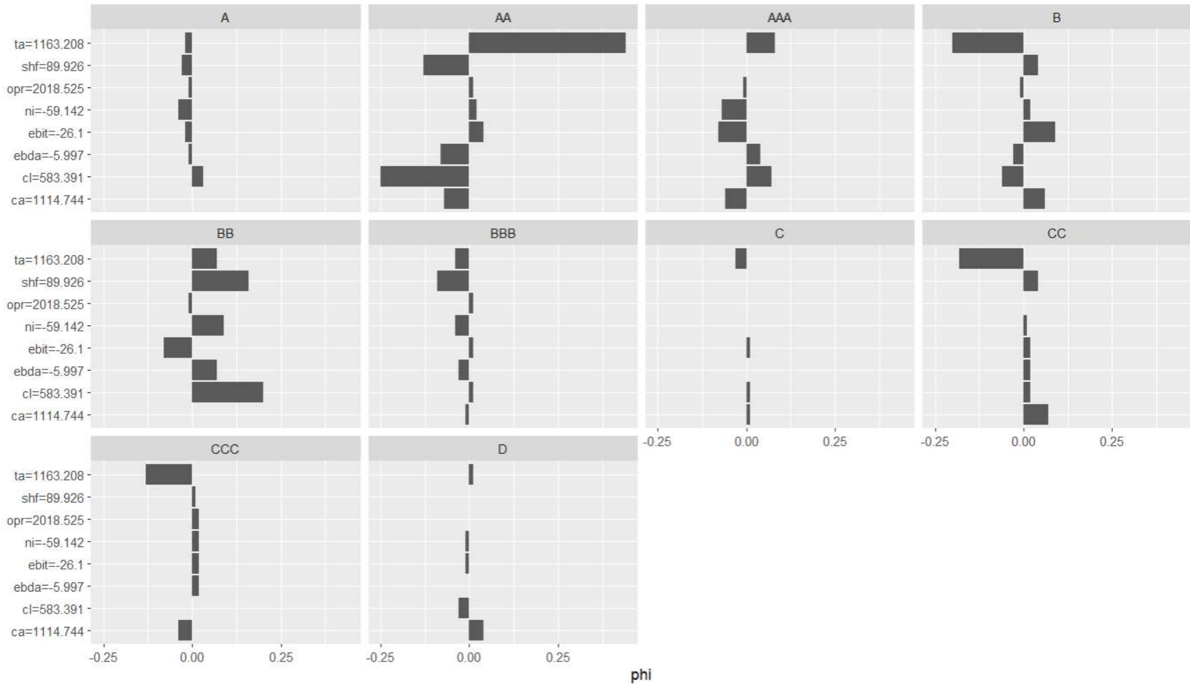
Net income,current liabilities and stock prices appear to be the most effective tools in determining whether the credit score is speculative.

The most important point to note here is that the reason why variables such as net income, which are considered positive in the finance world, appear to have a strong positive factor in classes being speculative like CCC, is that the net incomes in that class are generally negative in the 2022 data. In other words, the contribution of a negative net income to the class being CCC is positive.

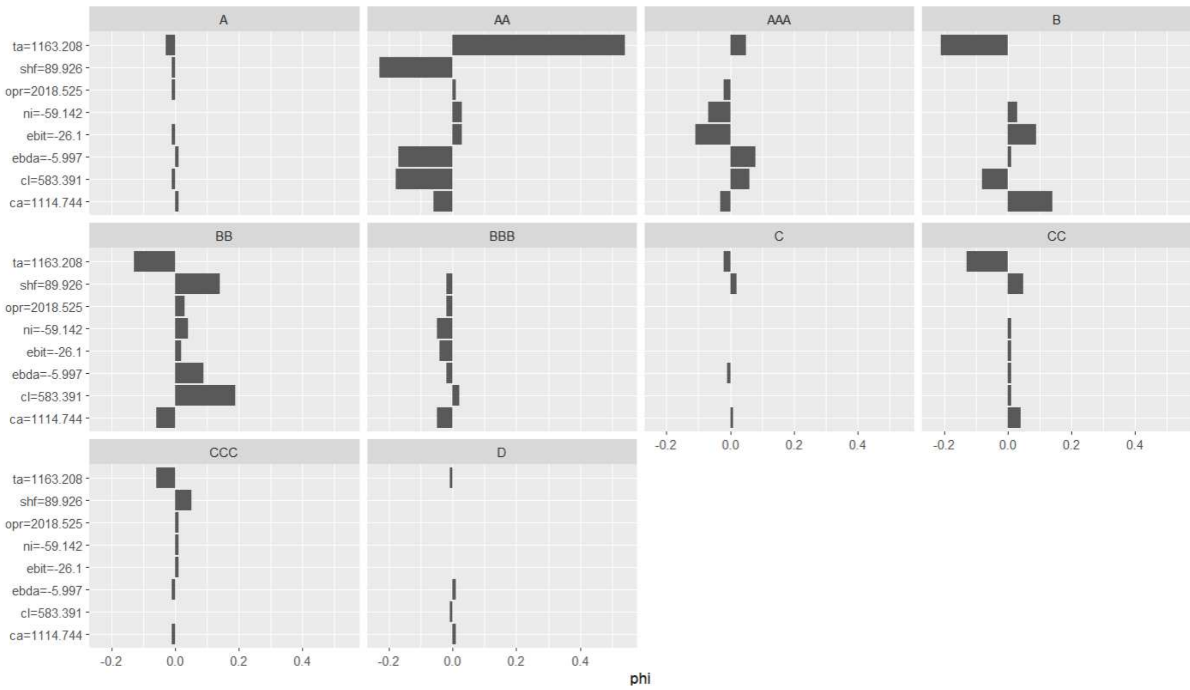
Multinomial Logistic Regression Model



Forward Model



Backward Model

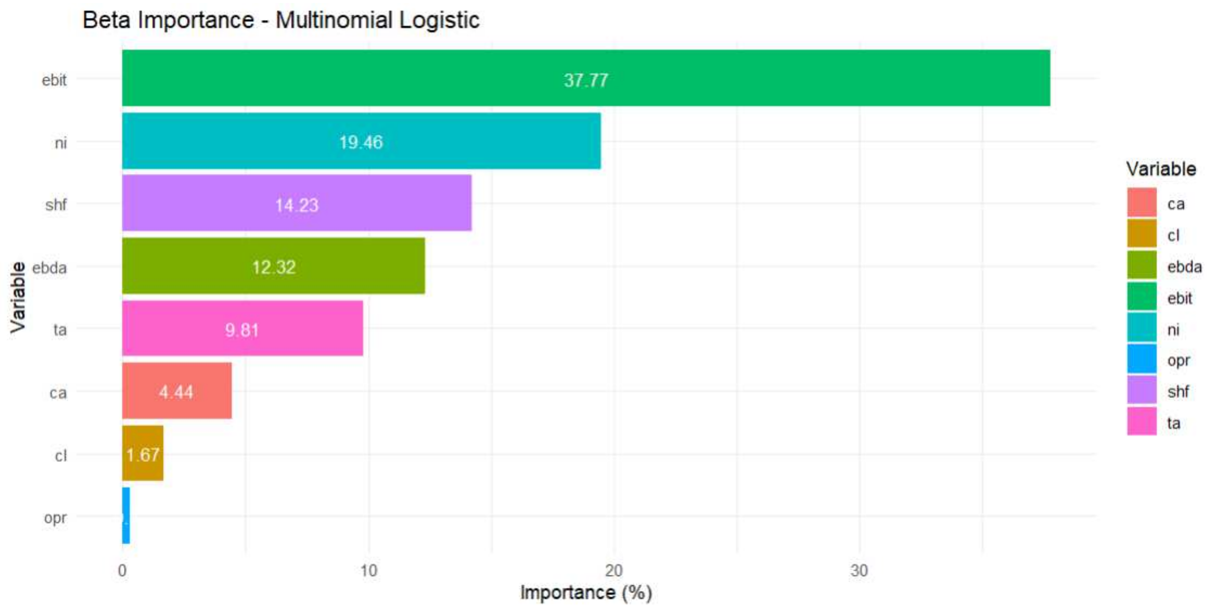


In the logistic regression, forward and backward model sections, it is seen that the most important contribution is determined by total assets.

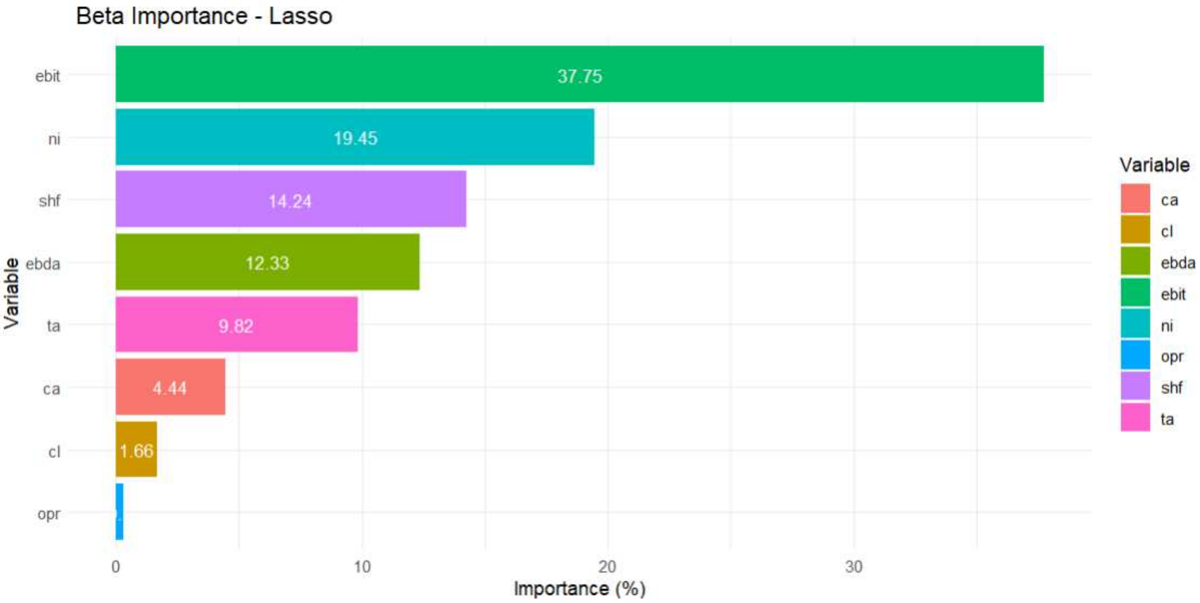
4.2.2 Beta Coefficient Analyses

The graphs show how effective financial data is compared to others according to the results we obtained in multinomial logistic, lasso, forward and backward models. In general, very close results were obtained in all models since variables with higher beta coefficients exhibit stronger predictive power in determining credit scores. There is no difference between the models in terms of importance order.

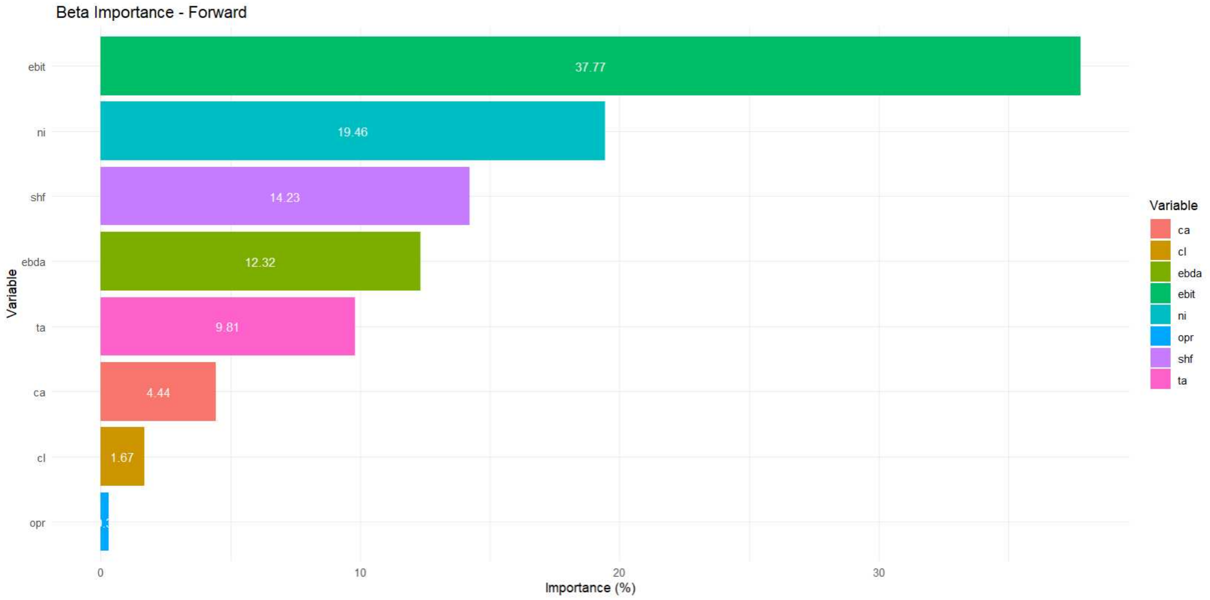
Relative Contribution of Variables in Logistic Regression

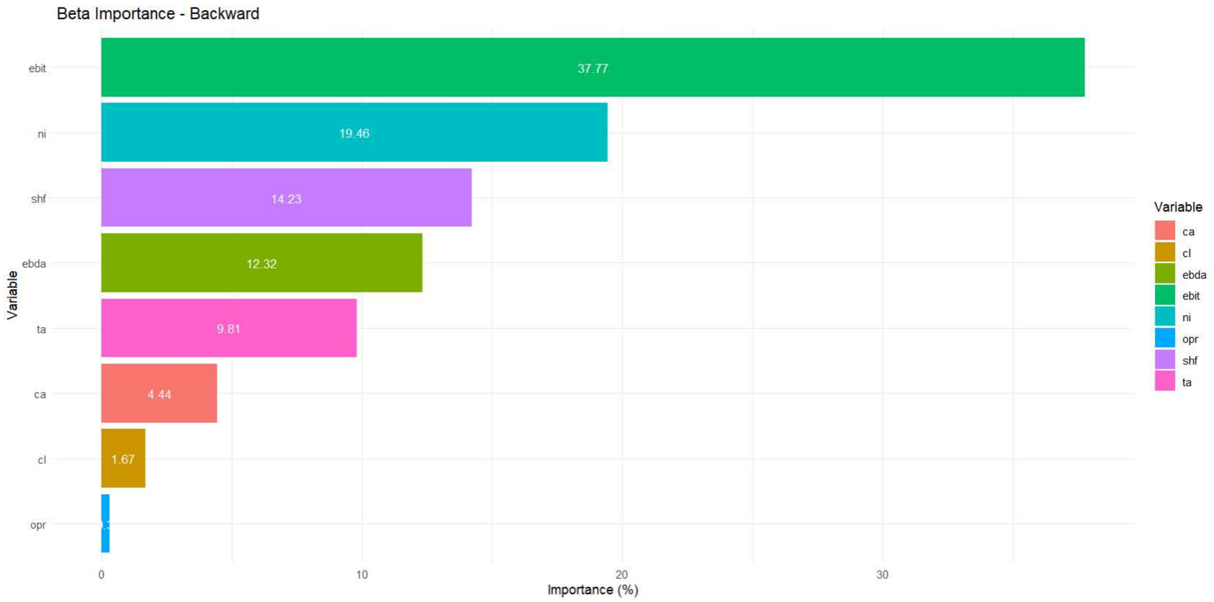


Relative Contribution of Variables in Lasso

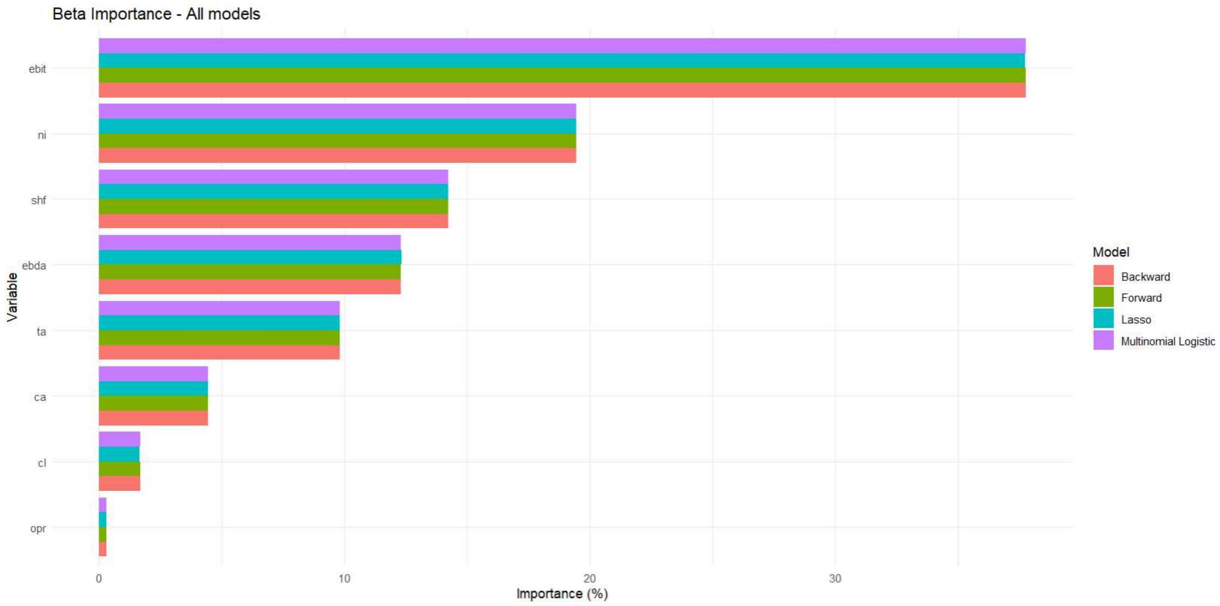


Relative Contribution of Variables in Forward and Backward





How close the results are to each other can be seen in the comparison graph below.

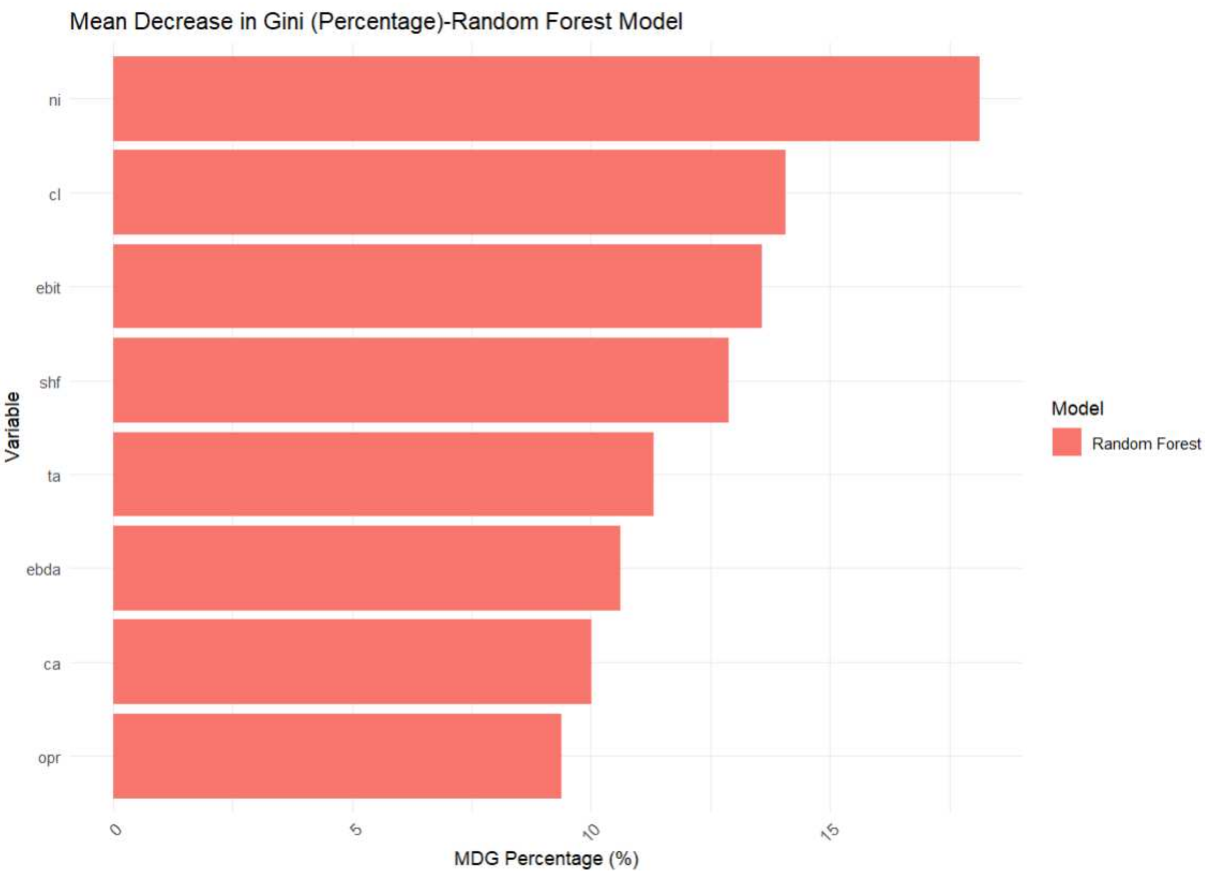


According to Beta Importance, variable EBIT is most significant in contributing towards model predictive performance and most important feature. On the other hand, operational income possesses least beta value of importance and least contributing feature towards model.

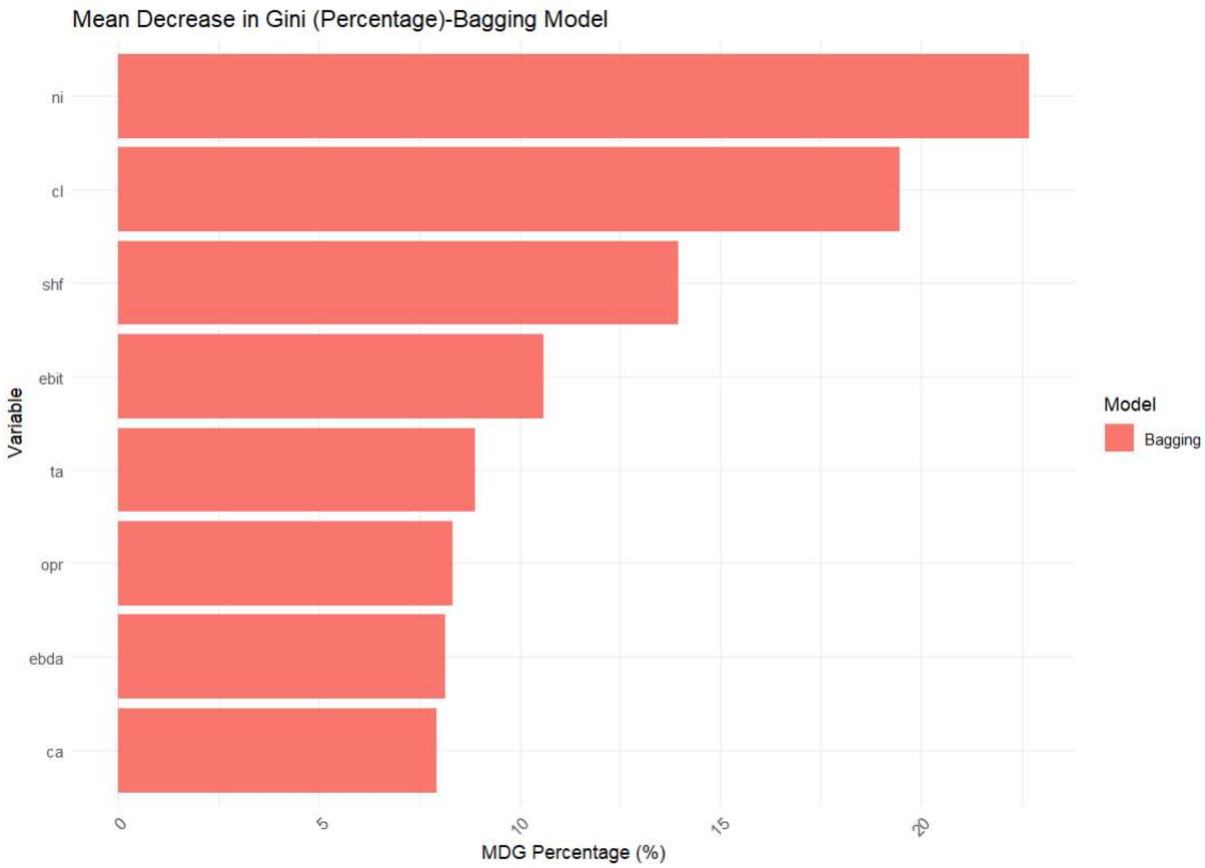
4.2.3 Mean Decrease Gini results

The following findings were obtained:

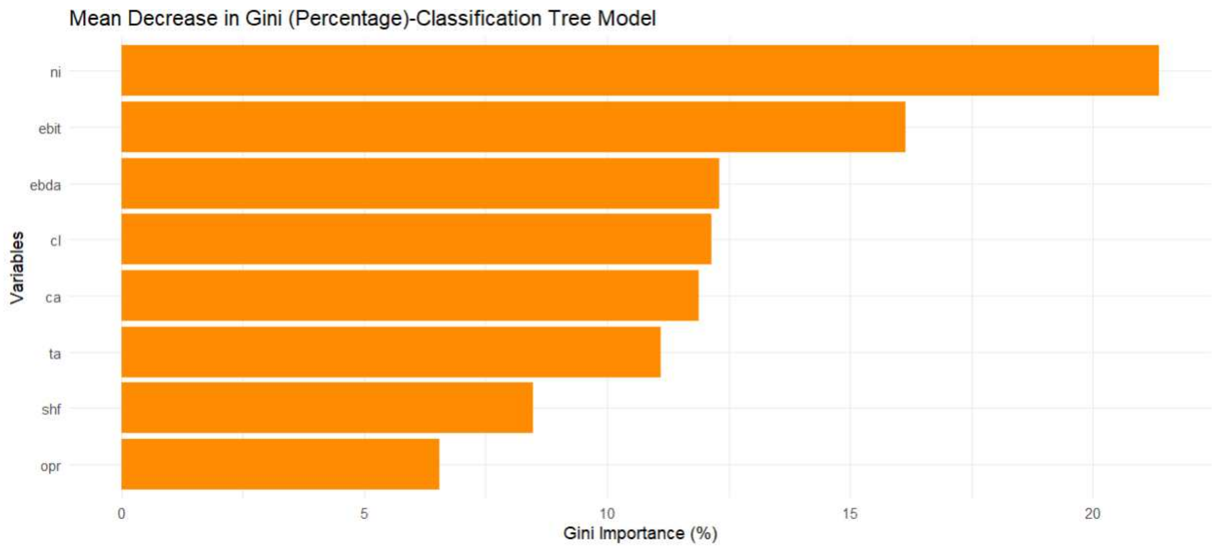
Relative Contribution of Variables in Random Forest Model



Relative Contribution of Variables in Bagging



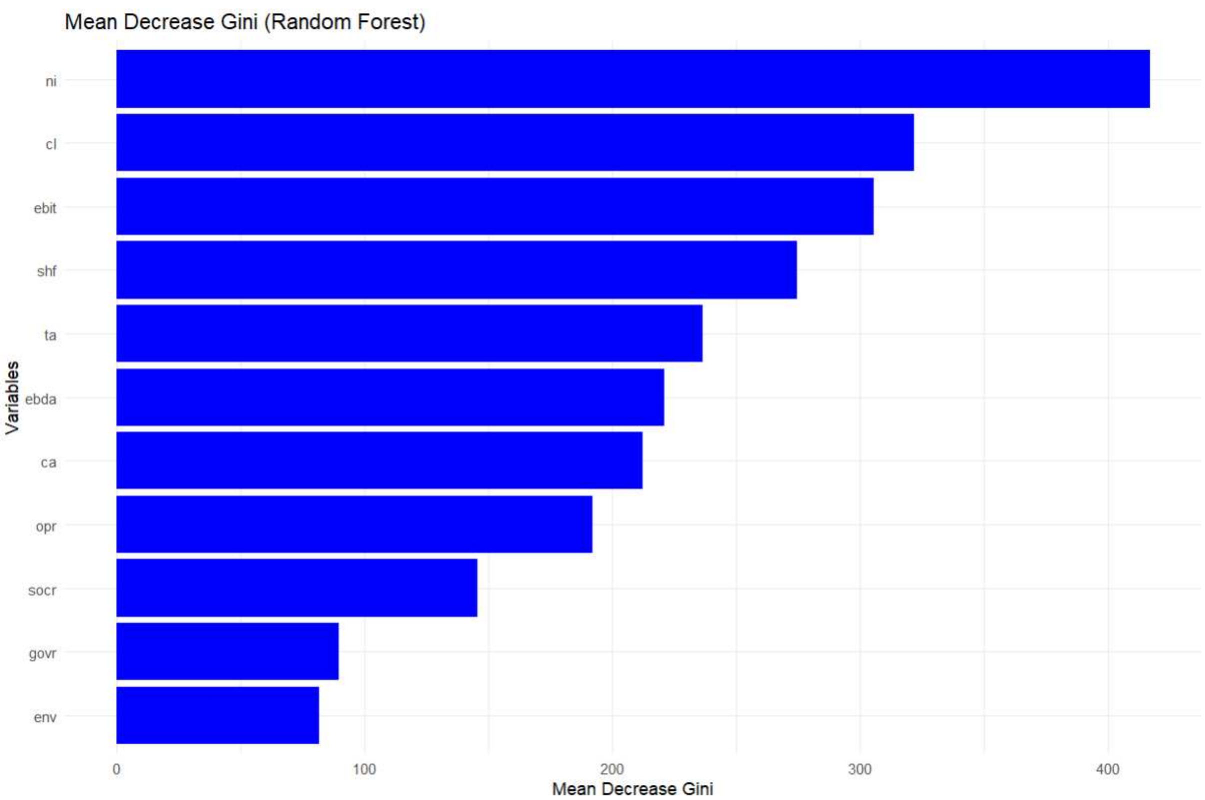
Relative Contribution of Variables in Classification Tree Model



According to the results of the Mean Decrease Gini metric, the net income was identified as the most important variable in all models. On the other hand, while the variable with the lowest importance in the Classification Tree and Random Forest models was operational income , this role belongs to the current assets variable in the Bagging model.

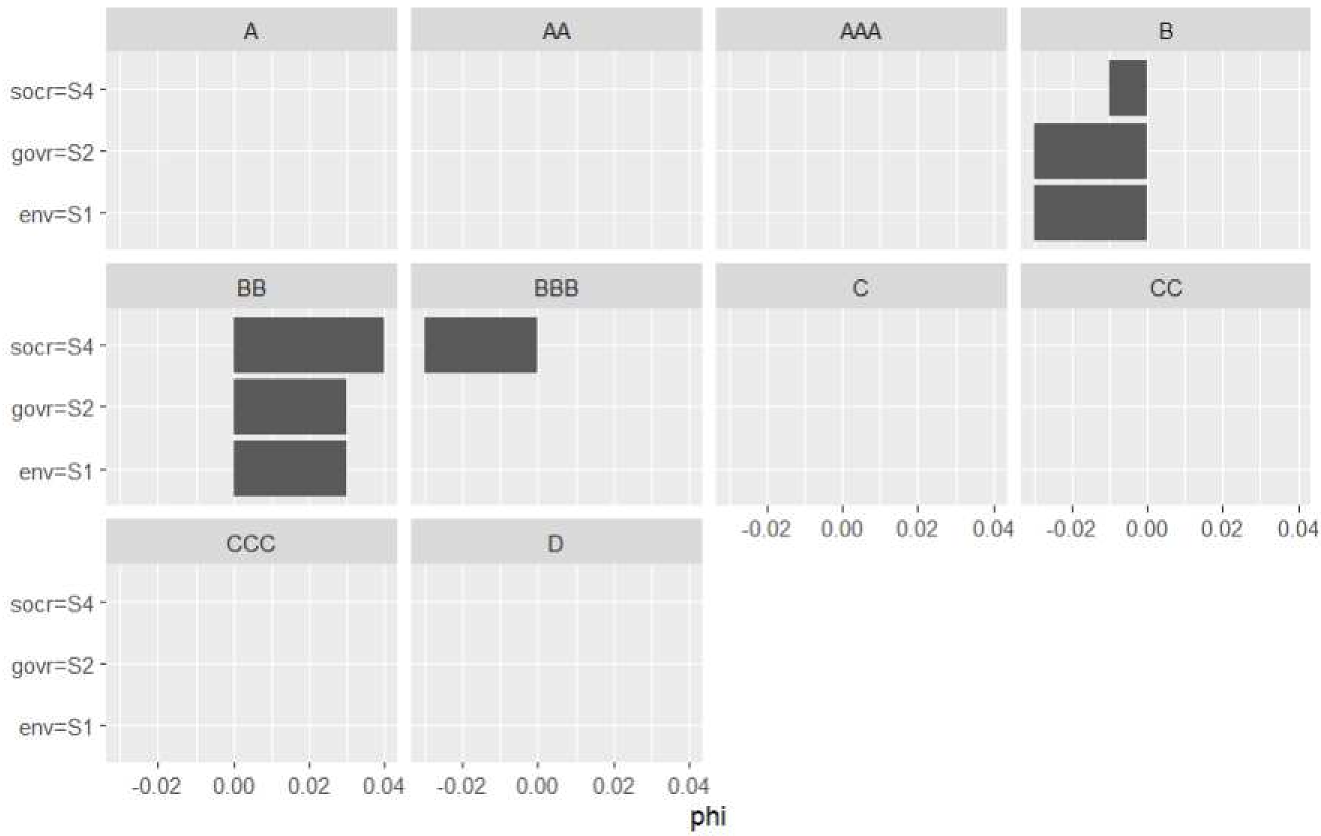
Unlike the previous observation, changes were observed in the order between the values in the Mean Decrease Gini measure.

4.3 Impact of ESG Factors on Credit Ratings



This plot shows the importance of variables for the Random Forest model using the Mean Decrease Gini method. First, social impact (socr) is the most important variable that contributes a lot to the model in terms of prediction in the Esg factors. The governance impact (govr)

variable is the second most important variable, which indicates that it holds a moderate explanatory power. The environmental (env) impact variable, is less important to the model when compared to the other two variables.
That would suggest that among ESG dimensions, social impact have the most predictive power in the Random Forest model and might be a leading indicator of sustainability or performance. The graph reflects that ESG factors have played a minor role in comparison with financial information. What this infers is that, even though ESG factors increasingly receive acceptance for being important drivers, their role, in terms of basic financial factors, in the Random Forest model is relatively less significant.



By looking at the graph where the importance of ESG factors is analyzed according to the

Shapley value, it can be stated that, as can be seen clearly in the above graph created using the Random forest model, ESG factors are quite ineffective in speculative classes. It is very important that no effect is observed other than B, BB and BBB values.

4.4 Fairness and Bias Tests

In these analyses, the confusion matrix was first used to understand how fairly the model behaved on a class basis. Since ESG models were included and used, the Random Forest model was preferred in this analysis. It was determined that the model could not make an equally successful prediction for every class, and the most important reason for this was that there was not enough data in the training data for some classes. This is especially true for speculative credit ratings.

4.4.1 False Positive and False Negative Analysis

These analyses were implemented by providing confusion matrix data in which ESG factors were included and the Random Forest model was used. True negative points at the number of cases whereby the model has correctly predicted a class not belonging, while false negative is the number of classes the model has predicted as negative but actually not.

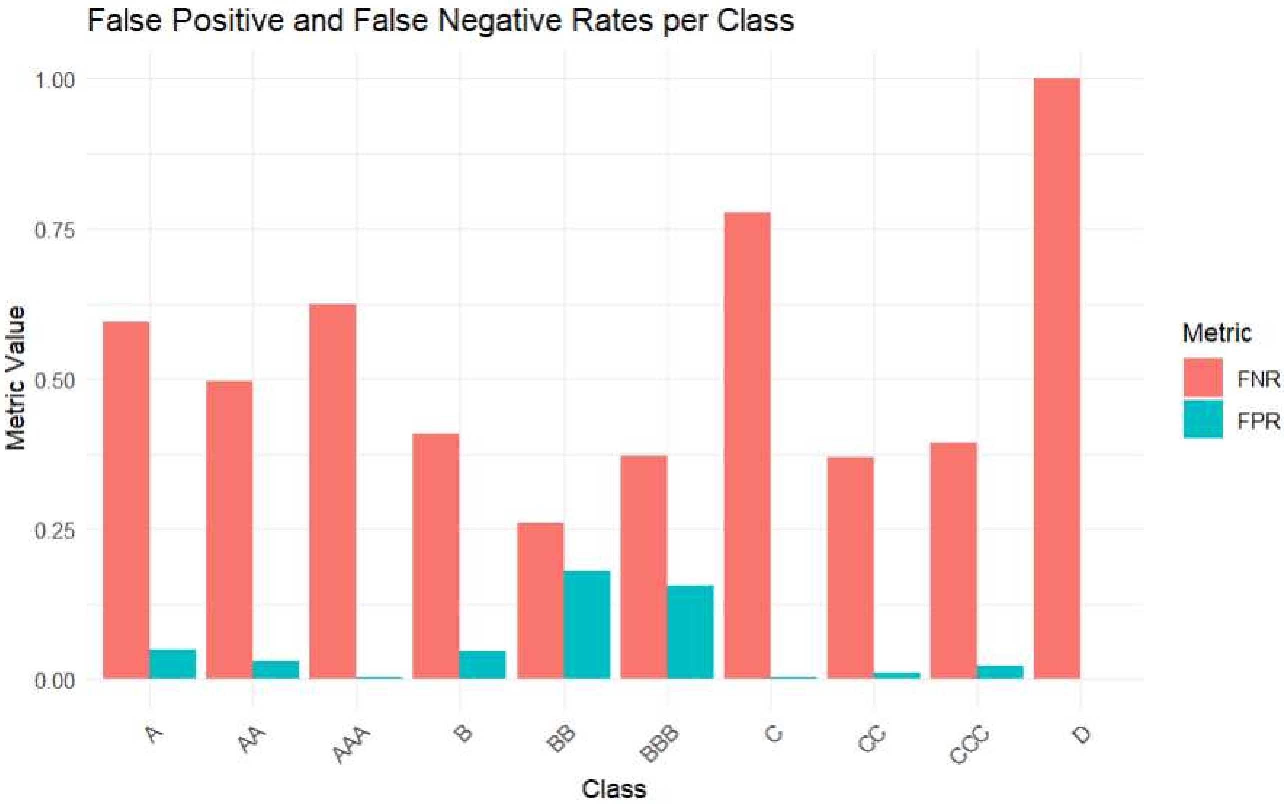
	Class	TP	FP	FN	TN	FP_Rate	FN_Rate
A	A	86	67	127	1279	0.049777117	0.5962441
AA	AA	56	41	55	1407	0.028314917	0.4954955
AAA	AAA	6	5	10	1538	0.003240441	0.6250000
B	B	166	60	114	1219	0.046911650	0.4071429
BB	BB	359	194	127	879	0.180801491	0.2613169
BBB	BBB	211	190	125	1033	0.155355683	0.3720238
C	C	2	3	7	1547	0.001935484	0.7777778
CC	CC	17	15	10	1517	0.009791123	0.3703704
CCC	CCC	46	35	30	1448	0.023600809	0.3947368
D	D	0	0	5	1554	0.000000000	1.0000000

The TPR and Precision rates are erratic between classes, though some classes have achieved high accuracy and precision in their models' performances, which remain very low, especially for classes AAA, C, and D.

Since most observations were made between BB and BBB classes, the Random Forest model was trained better in these classes.

The general accuracy has been achieved at high rates for most classes, which itself is not enough to make the model good in all classes.

The false positive rates are mostly low, reflecting a low misclassification rate.Further improvements could be attempted to enhance the model's predictive performance, particularly for classes C and D.

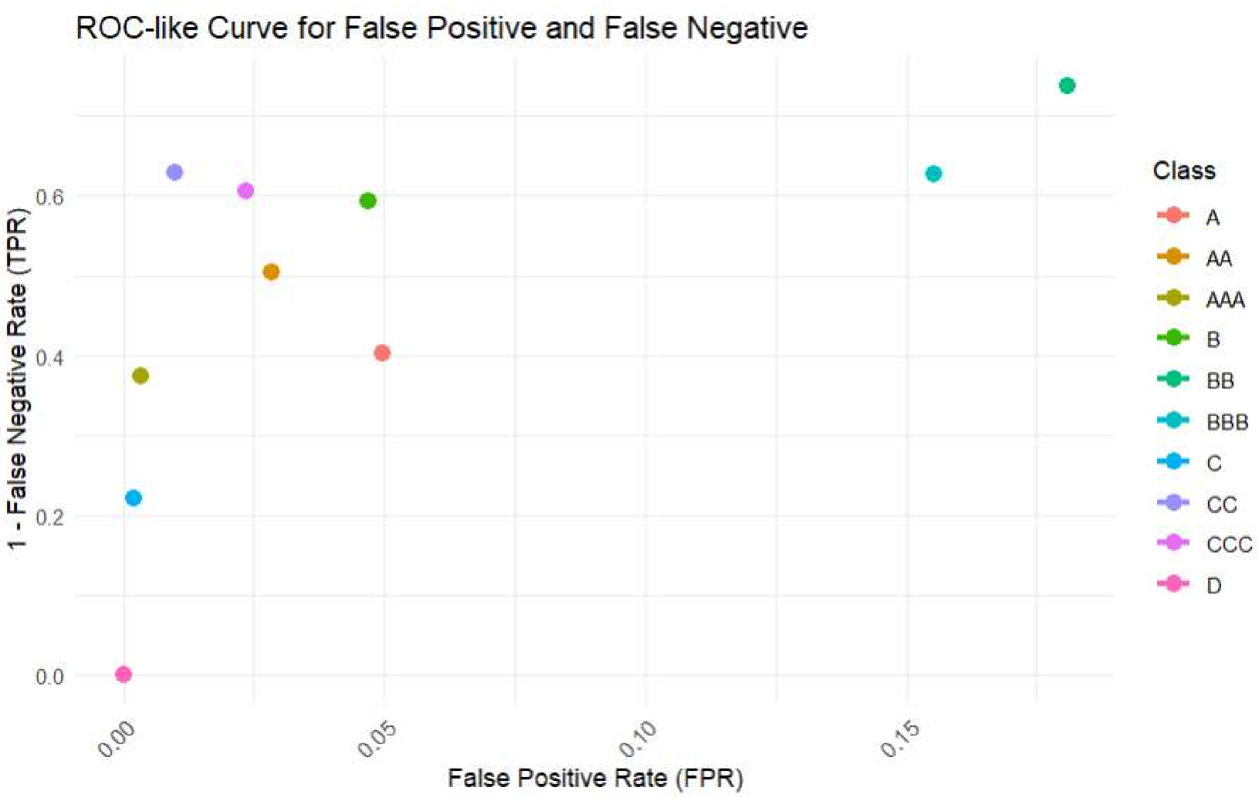


This graph shows the False Negative Rate and False Positive Rate for each class. The fact that the false negative rates are generally higher than the false positive rates indicates that the model is generally more prone to false negative classifications. This shows that the model tends to miss positive examples rather than falsely classifying class as positive. False positive ratios were quite low among all classes. Since there were few observations made in the D class, it also affected the model's predictions. Although there are actually 5 D classification of credit scores, the false negative rate was as high as 100% because it was never predicted.

The fact that the false negative ratio was so high in some classes indicates that the model performed unbalanced between classes. Overall, it can be concluded improvement to better discriminate between classes is needed.

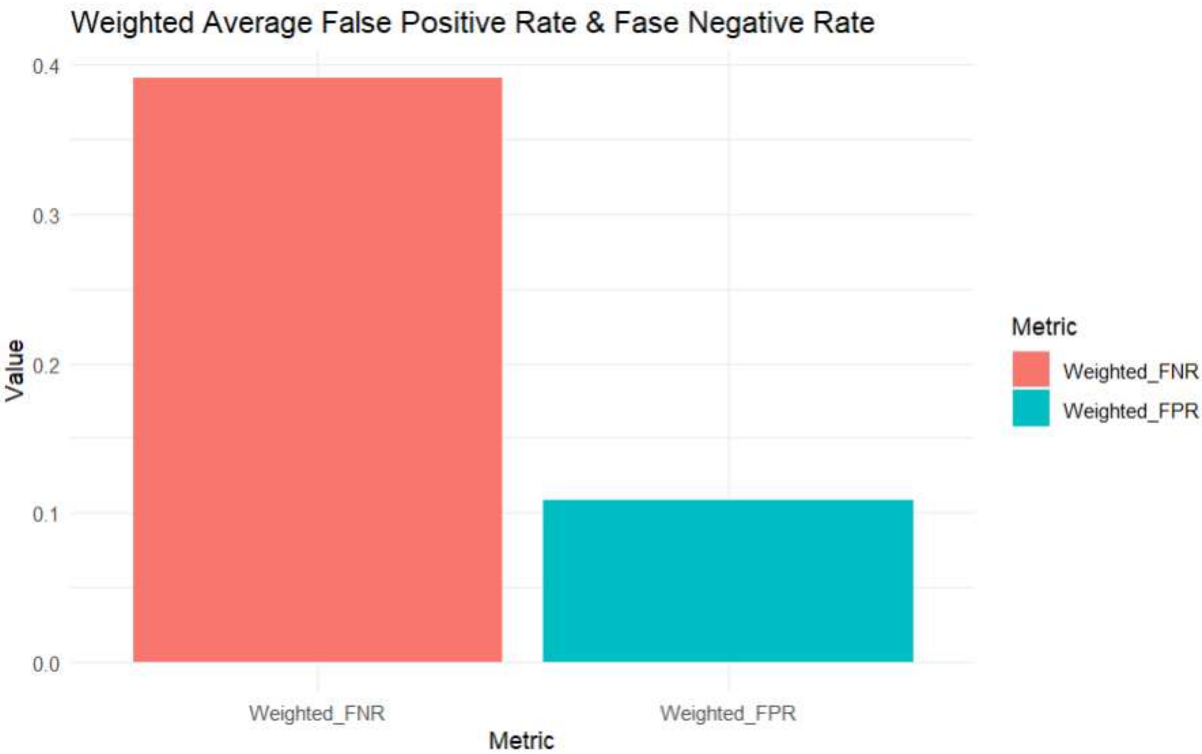
UNIVERSITÀ
DI PAVIA

As can be seen in this bar chart, if the necessary improvements are not made to the data set, there will be problems in terms of fairness.



This graph helps to better understand the model's accuracy and false positives for each class. Imbalances between classes are clearly visible.

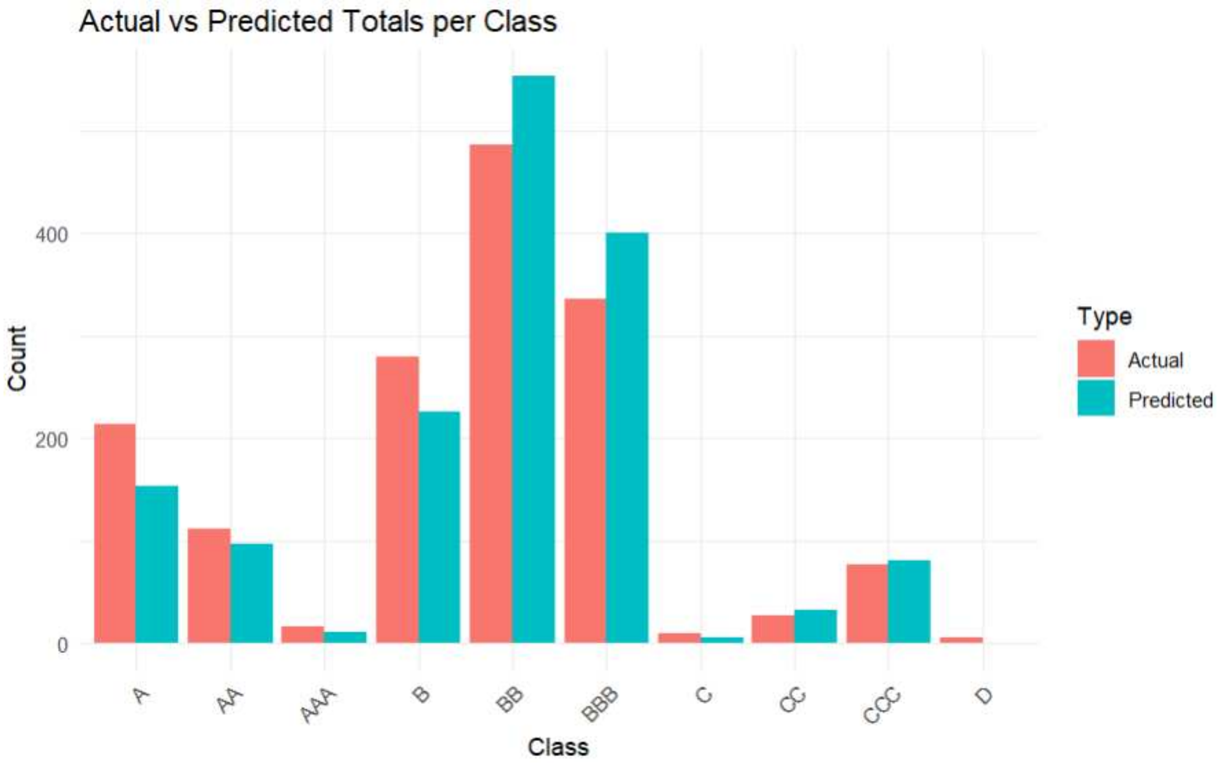
Since the number of observations in the classes is not equal, taking a weighted average provides a more logical perspective. In the graph below, false positive and false negative analysis was performed by taking the weighted average.



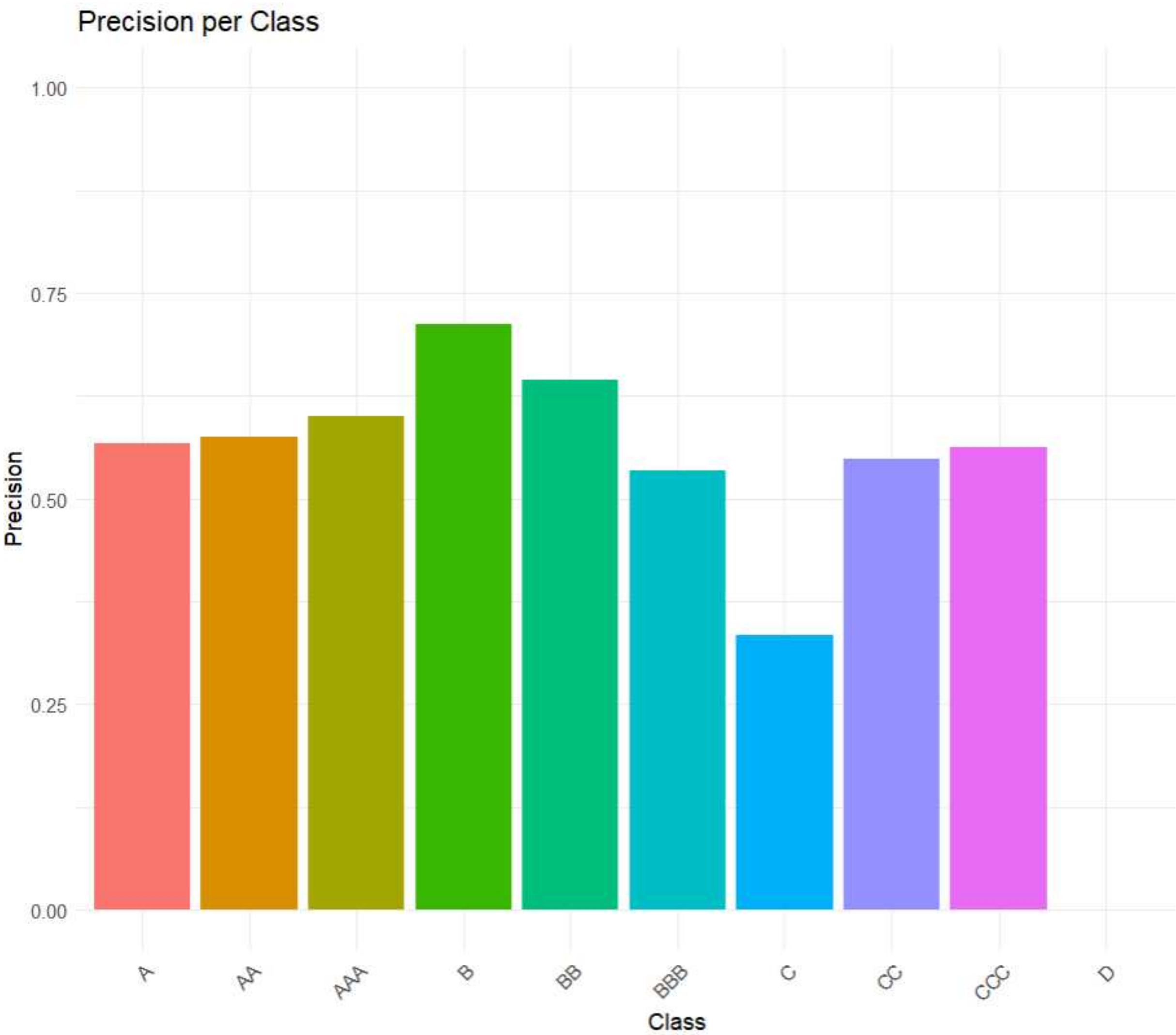
The low number of observations in some classes (especially class D) seriously affected the result. The fact that there was no D among the predictions but there were actually 5 D results caused the false negative result to be complete. The reason for calculating it as a weighted average is that the number of observations must also be taken into account for an objective evaluation. It would be misleading to assume that they had the same effect since there were many observations in class BB and very few in class D.

UNIVERSITÀ
DI PAVIA

4.4.2 Class-Based Performance Comparisions



In this graph, the actual and predicted values are better understood. As can be seen, in reality, the AAA, C and D values are quite low in the actual. This causes the predictive power of the models to vary across classes. The model could not predict with the same degree of fairness for each class. This is especially important in classes Class D.

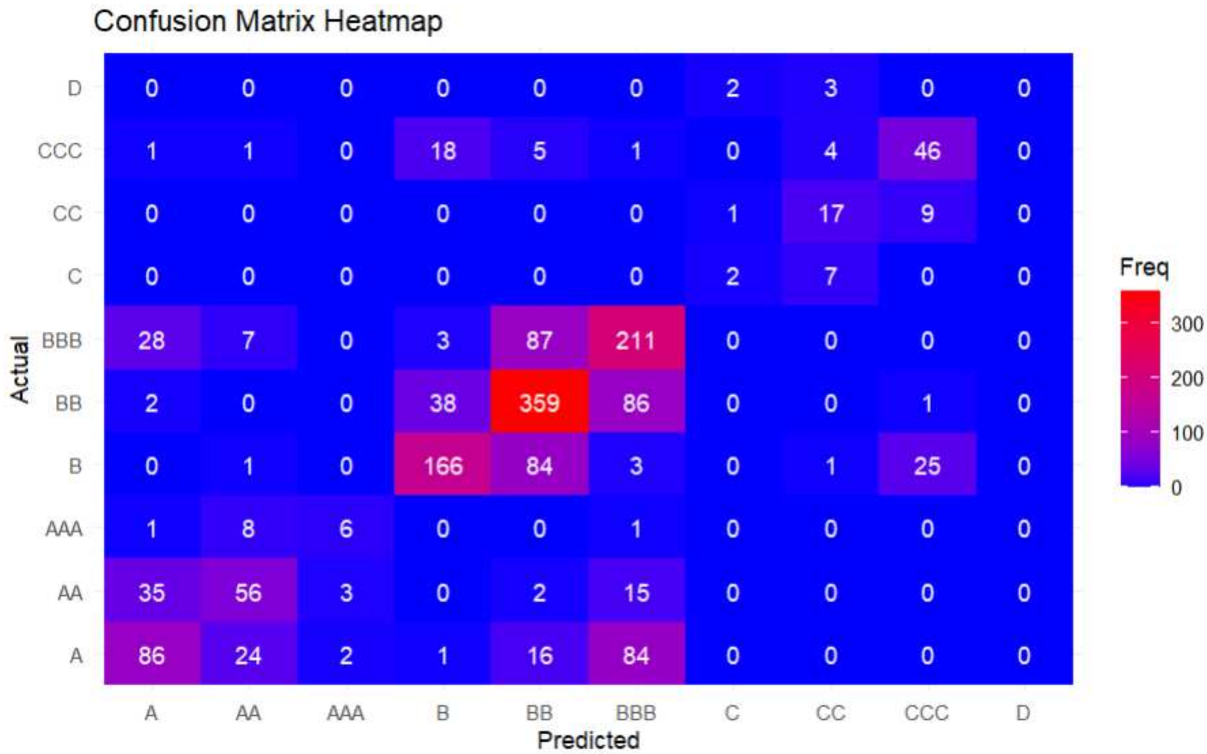


By retraining with ESG factors, its performance can be seen in a positive direction.

The target is to see model prediction and actual distribution:

Where actual and predicted values fall in similar range, then model is working well.

Where actual and predicted values have a high variation, it means model is predicting wrongly in certain classes and must be trained better.



The Random Forest model made the most predictions for the BB class and the frequencies for the other classes are quite low. The inadequacy of the data set mentioned in the previous sections can be understood more clearly in this matrix.

For more detailed analysis the following statistics are shown:

UNIVERSITÀ
DI PAVIA

Statistics by Class:

	Class: A	Class: AA	Class: AAA	Class: B	Class: BB	Class: BBB	Class: C
Sensitivity	0.40376	0.50450	0.375000	0.5929	0.7387	0.6280	0.222222
Specificity	0.95022	0.97169	0.996760	0.9531	0.8192	0.8446	0.998065
Pos Pred Value	0.56209	0.57732	0.545455	0.7345	0.6492	0.5262	0.400000
Neg Pred Value	0.90967	0.96238	0.993540	0.9145	0.8738	0.8921	0.995495
Prevalence	0.13663	0.07120	0.010263	0.1796	0.3117	0.2155	0.005773
Detection Rate	0.05516	0.03592	0.003849	0.1065	0.2303	0.1353	0.001283
Detection Prevalence	0.09814	0.06222	0.007056	0.1450	0.3547	0.2572	0.003207
Balanced Accuracy	0.67699	0.73809	0.685880	0.7730	0.7789	0.7363	0.610143

	Class: CC	Class: CCC	Class: D
Sensitivity	0.62963	0.60526	0.000000
Specificity	0.99021	0.97640	1.000000
Pos Pred Value	0.53125	0.56790	NaN
Neg Pred Value	0.99345	0.97970	0.996793
Prevalence	0.01732	0.04875	0.003207
Detection Rate	0.01090	0.02951	0.000000
Detection Prevalence	0.02053	0.05196	0.000000
Balanced Accuracy	0.80992	0.79083	0.500000

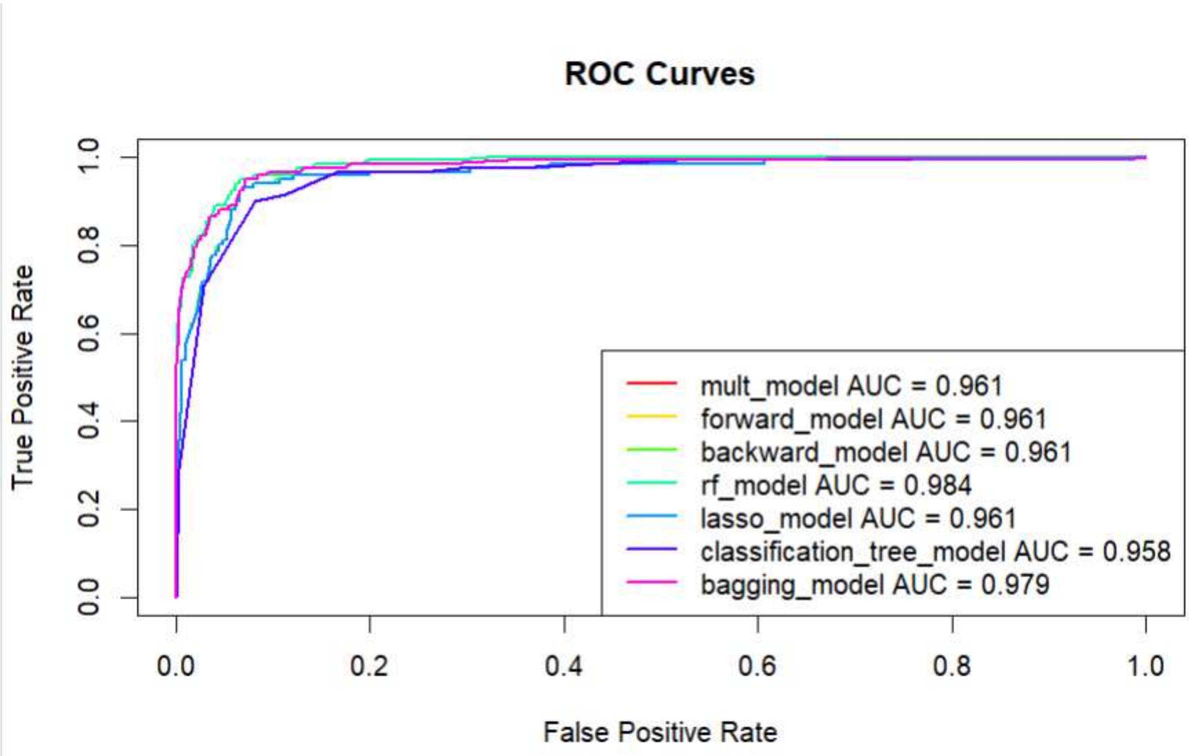
4.5 ROC and AUC Analysis

A model's ability to balance true positive rate (TPR) and false positive rate (FPR) is demonstrated by the ROC curve. By calculating the area under this curve, AUC provides a summary of the model's overall performance. Through the analysis of these flaws, ROC and AUC can assist in enhancing the model's reliability.

Speculative classes (CCC,CC,C,D) are generally those that carry higher risk and are more uncertain for investors.Non-speculative classes (AAA,AA,A,BBB,BB,B) are classes that carry lower risk and are considered safer. The model made probability estimates separately for these two groups and calculated the sum of the probabilities for the speculative classes.

Samples that are mistakenly projected to belong to the negative class are known as false negatives, and samples that are mistakenly predicted to belong to the positive class are known as false positives. Low mistake rates in both of these areas are characteristics of a Safe AI system.

ROC and AUC analysis is performed for all models before activating ESG factors and running them.

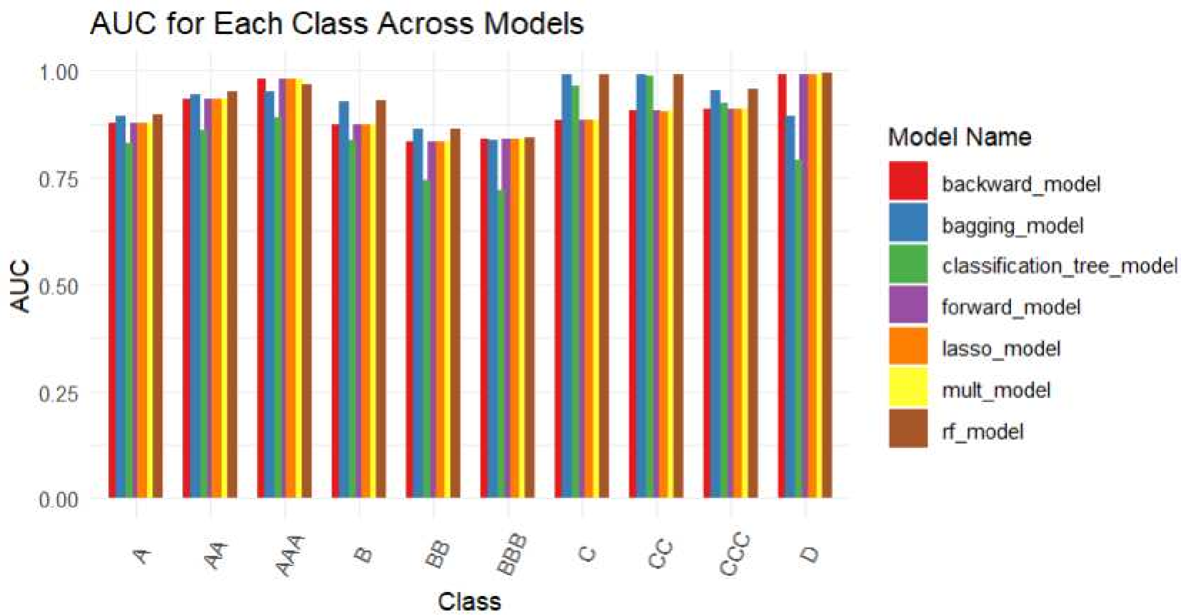


While the Random Forest model gave the best results, the Classification Tree model was weak in terms of classification performance with a lower AUC. In every model, speculative and non-speculative models can be distinguished from each other quite successfully. The probability of models creating such confusion is very low.

Models with AUC = 0.961 showed similar performance; This shows that the ROC curves of these models are close to each other.

Then, the probabilities of each class are separated one by one and their individual performances are analyzed.

This approach allows assessing not only the overall discriminatory power of the model, but also the prediction accuracy of each class. The results are evaluated by comparing the ROC curve and AUC values on a class basis.



Performances of the Models:

RF Model and Bagging Model have better or higher AUC values in many classes. These models showed higher performance than other models, especially in high-risk classes (such as C,CC,D).

The classification tree model cannot generalize for BB, BBB, and CCC classes with poor values for AUC in comparison with other models.

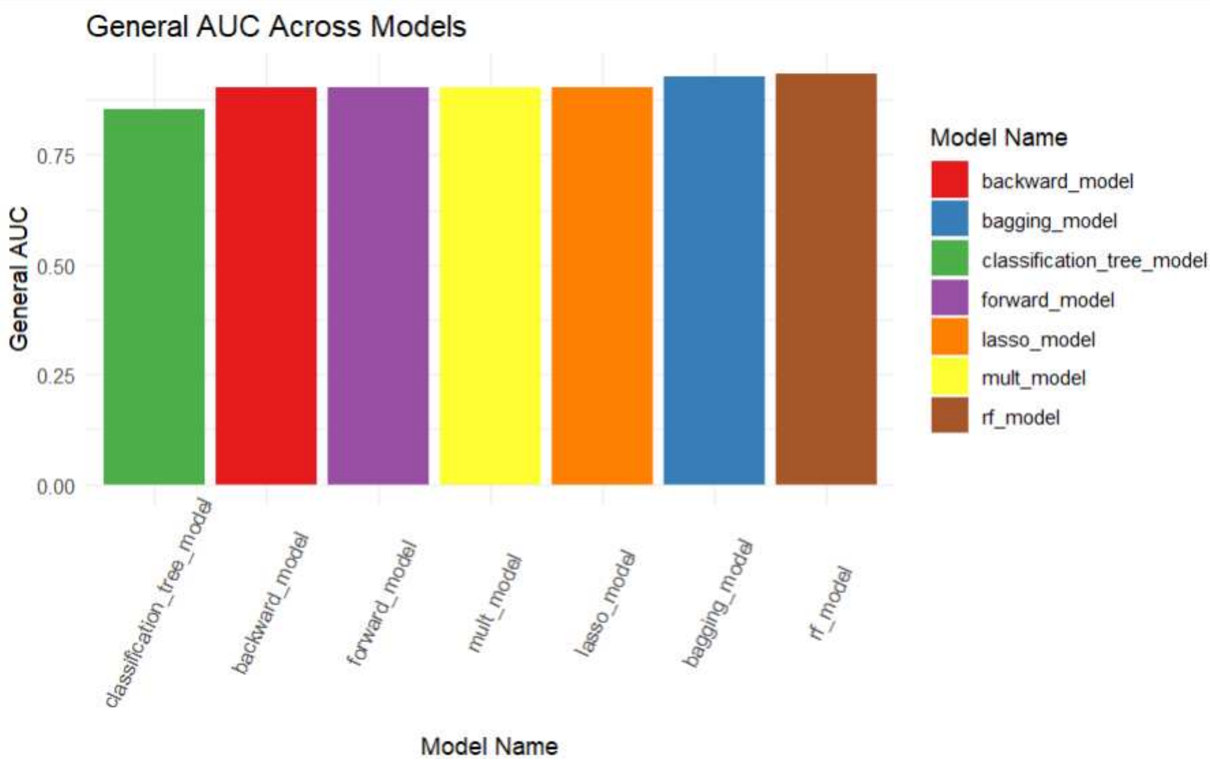
The values for AUC for all other models in general have almost same values, and similar performance for them can thus be noticed.

Difference Between Classes:

AAA, AA, and A classes: High values for AUC in these classes can be noticed.

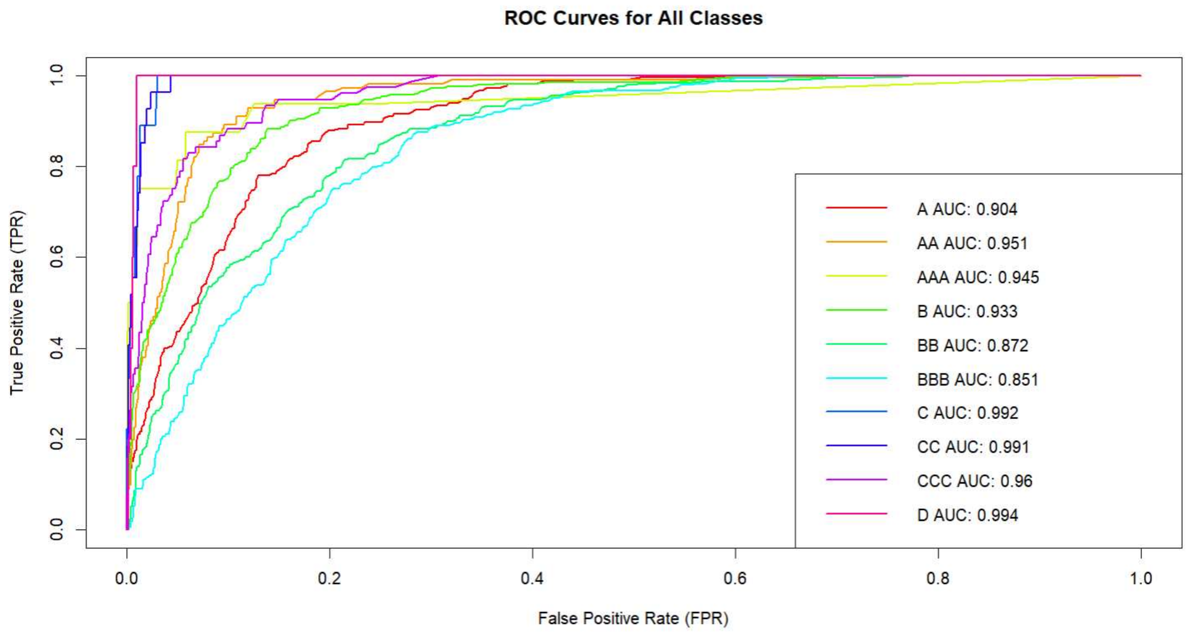
Classes B, BB, BBB, C, CC, and CCC: There is variation in between these classes. In specific, poor performance for classification tree model in these classes can be noticed.

D class: There is variation in between models in D class, but overall good performance can be noticed.



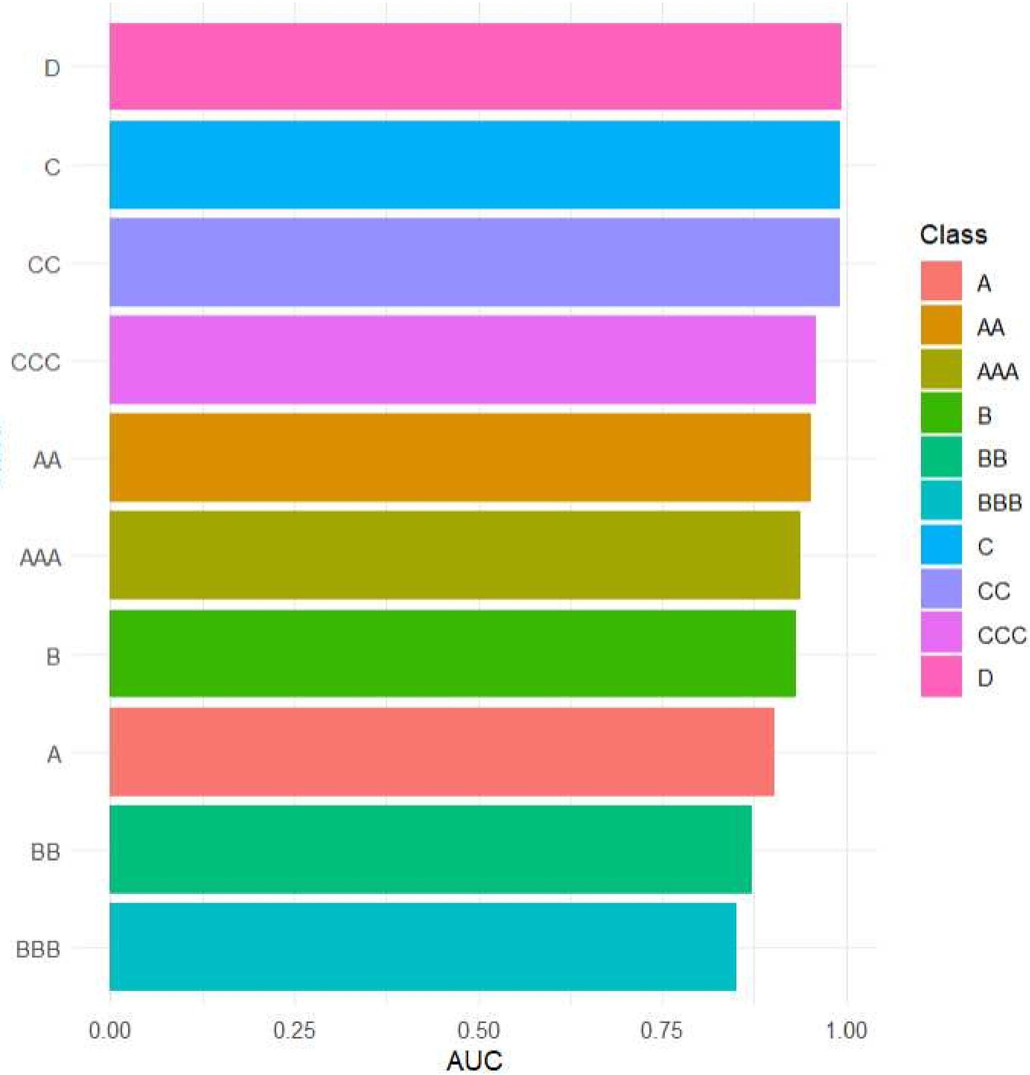
According to the general AUC value, the Random Forest model was the most successful in distinguishing between classes and therefore showed the best performance. On the other hand, the Classification Tree model had the lowest AUC value and was the model that performed the weakest in distinguishing between classes.

AUC Values of Each Class With ESG by Using Random Forest model



These analyses were conducted by considering each class of the model. The exact values more clearly with the numbers under the bar chart.

UNIVERSITÀ
DI PAVIA



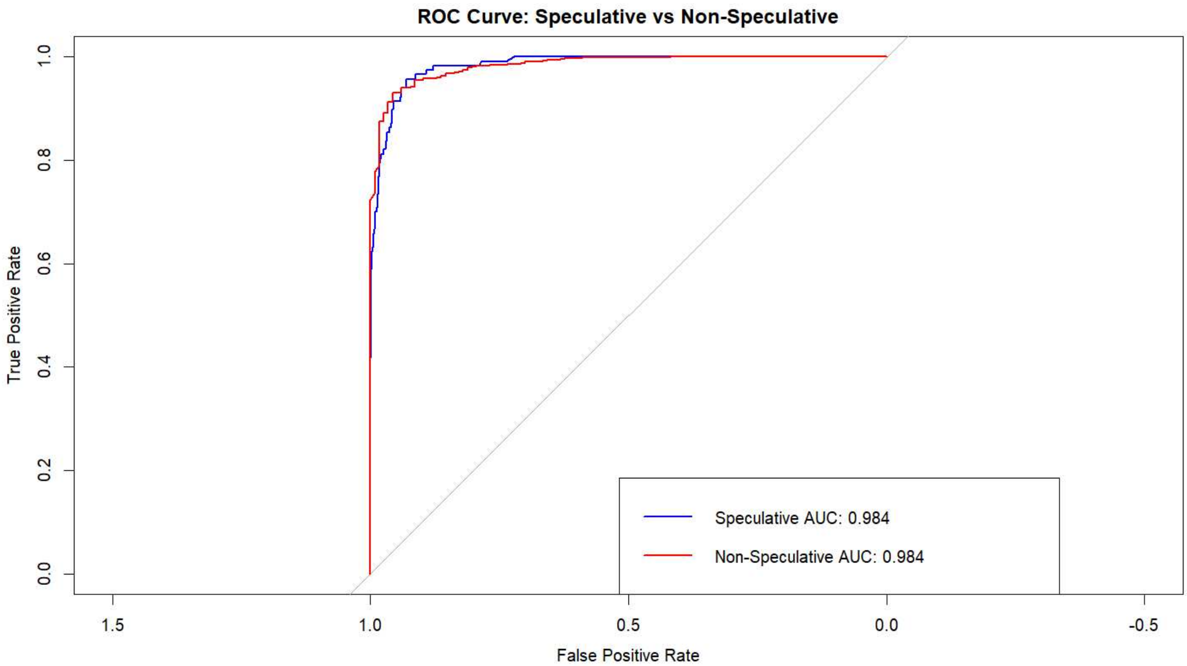
UNIVERSITÀ
DI PAVIA

	Class	AUC
A	A	0.9043017
AA	AA	0.9510508
AAA	AAA	0.9445480
B	B	0.9332081
BB	BB	0.8721461
BBB	BBB	0.8507257
C	C	0.9919355
CC	CC	0.9913572
CCC	CCC	0.9596080
D	D	0.9941441

After the random forest model was trained with the data, it was seen that the model had the ability to distinguish between different classes. Although the AUC values were high, the addition of ESG values did not show a significant improvement.Model does not tend to confuse model classes with other classes.

Note that the AUC is quite high in class D. Since the model does not predict D at all, the ROC curve may produce falsely high AUCs since the TPR and FPR calculations are based on few observations.

The following graphs were created using the R code using the ggplot2 package to create a ROC curve between the classes in the dataset.



In this plot, "Speculative" and "Non-Speculative" classes have been contrasted in relation to the curve for ROC under a model for a Random Forest.

The model is satisfactory in performance with regard to discrimination, for its value for AUC is approximately one. It signifies that the model can discriminate between both classes in a satisfactory manner.

The fact that values for both classes have approximately one value in terms of AUC signifies that the model is satisfactory in performance.

UNIVERSITÀ
DI PAVIA

Chapter 5: Discussion and Evaluation

5.1. Interpretation of Model Performance

The Random Forest model provided high accuracy in both speculative and non-speculative classes, as well as in the analyses where we evaluated the classes one by one. Although it was followed by the Bagging model, Random Forest manages the risk of overfitting better compared to Bagging, and this feature is quite successful in managing the effect of low-impact variables such as ESG on the model. Due to these advantages compared to the Bagging method, Random Forest was used to analyze the effect of ESG factors among the two most successful models.

5.2. Role of ESG Factors

This study showed that the role of ESG factors is limited when compared to financial variables. This was confirmed by Gini significance analysis and when these factors were added to the model, no significant improvement was observed in model performance, and no significant increase was achieved in AUC values. It was observed that ESG was less decisive, especially in the distinction between "speculative" and "non-speculative" classes, compared to the weight of financial data. This may be due to various reasons, as well as the insufficient number of observations for some classes. Model results showed that ESG can provide an additional source of information affecting credit ratings, but it could not overshadow the importance of financial data, and although it played an additional complementary role in the model, it could not be decisive. According to this study, more data years, expanded ESG indicators, and better data quality are needed to significantly improve the performance of ESG data.

5.3 Limitations of the Study and Overcoming Data Limitations

In this study, ESG models are used as constants, not as a time series. Only the same variables are used as predictors for 2020, 2021 and 2022 data.

In case of ESG factors data for 2020, 2021, 2022 can be analyzed separately.Grouping according to special situations is also important in adjusting ESG factor values according to timing. The COVID-19 period can be given as an example. It is effective to make distinctions based on various crises. Since the data is limited, data collection and analysis processes should be strengthened in order to create a more comprehensive ESG assessment. When evaluating companies' ESG data, trend analyses can be conducted to predict future trends and possible developments. Based on past data, it may be useful to develop predictions about how certain ESG indicators may change in the coming years.

In addition, a "shock scenario analysis" can be conducted to evaluate how companies have developed a risk management strategy against potential risks.

Importance of ESG factors may vary by sector and geography. Environmental factors may be more effective in the energy sector. It can also be divided into more specific sub-factors according to sectors. If we consider it from a geographical perspective, developed countries may give more importance to social factors. Studies on ESG factors among different groupings provide a more accurate and deep examination of how important a predictor ESG factors are.

Financial data was limited in this study but it can be expanded and debt ratios such as Debt-to-Equity Ratio, Debt-to-Total Assets Ratio, profitability ratios such as Net Profit Margin, Operating Profit Margin and Return on Equity - ROE can also be taken into account as predictors for better analysis. Sales regarding the company's future growth potential Factors such as growth rate or equity growth rate can also be used as a positive factor. In this way, it is possible to evaluate from a broader perspective. Increasing the number of variables to be used in the credit scoring process contributes to obtaining more precise and reliable estimates. For this, Relative valuation and Discounted Cash Flow valuation techniques can be analyzed in terms of usability. It is assured that FCFF and FCFE of Discounted Cash Flow valuation techniques represent actual value of the company. Thus, these values can be taken into consideration in the credit scoring.

Also, values including EV/EBITDA and P/E ratio of Relative valuation techniques represent relation between operational profit and value in marketplace of the company. Thus, these values can be examined in the credit scoring.

Credit Scoring techniques are conducted with statistics techniques and machine learning and determine credit risk of the company.

After analyzing these variables, other metrics used to increase the performance of the credit scoring algorithm can be examined. Such metrics may include elements such as Adjusted Present Value, Economic Value Added and Black-Scholes Model.

By incorporating market volatility it can be directly linked to credit risk and allow for better forecasts, especially for speculative classes.

UNIVERSITÀ
DI PAVIA

The low quantity of observations, in particular, for speculative classes, had an adverse influence on the performance of the model in the prediction process on these classes. All this has been the reason for the high rate of false negatives or positives for these classes. On top of that, SMOTE (Synthetic Minority Oversampling Technique) can be used as a sequence of data augmentation to increase the dataset size of speculative classes.

UNIVERSITÀ
DI PAVIA

Chapter 6: Conclusion and Recommendations

6.1. Key Findings of the Study

In this study, the SAFE AI principles of explainability, robustness, accuracy and fairness were examined using different methods. With the aim of understanding how the model works and which factors are more effective in predictions, SHAP, gini importance and bete coefficiens used for explainability principle were used and according to the results obtained from various models, the most important variables were EBIT, net income and current liabilities. The fact that there was no significant increase in the predictive power of the model with the addition of ESG factors to the model showed that the model was robust on financial data. Using ROC curves and AUC values, it was shown that the highest explainability values were applied with random forest and bagging values, and that the model was quite successful in predicting speculative and non-speculative classes. In the fairness part, false positive and false negative rates were calculated for each class and the high negative rates in speculative classes were determined in these groups. indicated poor performance and potential fairness issues.

6.2. Recommendations and Future Work

Macroeconomic indicators can become part of a strong and accurate model with increased usage, forexample, of, for example, market volatility. In addition,the expected return calculated with CAPM can be used as a dependent variable in the credit scoring model, and the expected return based on the risk premium of a particular asset can be a helpful indicator in determining the credit rating of that asset. Thus, in the Random Forest model, the numbers calculated with CAPM are used as independent variables to support the model created together with other financial data. All model and data expansion techniques, including weighted models, can be utilized, particularly in disadvantaged classes such as in instances of speculative rating of credit. In terms of policy recommendations, financial entities can have less transparent and uniform policies when rating ESG factors and laws can become promulgated for fairer evaluations. With increased usage of complex models, ESG factors can become understood in a better way for rating relations in credit. Once again, in current times, this work validates that SAFE AI fundamentals have a role in finance and new and additional factors and policies must become developed for strengthening these fundamentals.

UNIVERSITÀ
DI PAVIA

Code Used for Analysis

1. Check and Install Packages

```
check_and_install <- function(packages) {
  for (pkg in packages) {
    if (!require(pkg, character.only = TRUE, quietly = TRUE)) {
      install.packages(pkg, dependencies = TRUE)
      library(pkg, character.only = TRUE)
    }
  }
}
```

2. Load and Preprocess Data

```
load_and_preprocess_data <- function(filepath) {
  data <- read.csv(filepath) %>%
    mutate(across(where(is.character), as.factor)) %>%
    mutate(across(starts_with("ta") | starts_with("ca") | starts_with("shf") | starts_with("cl") |
starts_with("opr") | starts_with("ebit") | starts_with("ni") | starts_with("ebda"),
~as.numeric(gsub("[^0-9.-]", "", .)), .names = "{.col}")) %>%
    mutate(across(where(is.numeric), ~ifelse(is.na(.), mean(., na.rm = TRUE), .)))

  return(data)
}
```

3. Prepare Data by Year

```
prepare_data_by_year <- function(data, year_suffix) {
  selected_columns <- dplyr::select(data, ends_with(year_suffix))
  new_column_names <- c("sco", "ta", "ca", "shf", "cl", "opr", "ebit", "ni", "ebda")

  renamed_data <- setNames(selected_columns, new_column_names)

  return(renamed_data)
```

```
}
```

4. Clean and Bind Data

```
clean_and_bind_data <- function(data_2020, data_2021, data_2022) {
  training_data <- bind_rows(data_2020, data_2021) %>% na.omit()
  testing_data <- data_2022 %>% na.omit()

  testing_data$sco <- factor(testing_data$sco, levels = levels(training_data$sco))
  training_data$sco <- factor(training_data$sco, levels = unique(c(levels(training_data$sco),
levels(testing_data$sco))))

  return(list(training_data = training_data, testing_data = testing_data))
}
```

5. Evaluate Models

```
evaluate_models <- function(models, testing_data) {
  accuracies <- list()

  for (model_name in names(models)) {
    model <- models[[model_name]]
    tryCatch({
      predicted_labels <- predict(model, newdata = testing_data, type = "class")
      predicted_labels <- factor(predicted_labels, levels = levels(testing_data$sco))
      cm <- confusionMatrix(predicted_labels, testing_data$sco)

      cat("\nModel Name:", model_name, "\n")
      print(cm)

      if (model_name %in% c("rf_model", "bagging_model")) {
        cat("Variable Importance:\n")
        print(importance(model))
      } else {
        cat("Variable importance not applicable or not directly available for this model type.\n")
      }
```

```
    accuracies[[model_name]] <- cm$overall['Accuracy']
  }, error = function(e) {
    message(paste("Error in model", model_name, ":", e$message))
    accuracies[[model_name]] <- NA # Assign NA for models that fail
  })
 }

 accuracies_vector <- unlist(accuracies)
 return(accuracies_vector)
}
```

6. Calculate AUC for Each Class

```
calculate_auc_per_class <- function(model, data, target_col) {
 classes <- unique(data[[target_col]])
 auc_results <- list()

 for (class in classes) {
  data$binary_response <- as.numeric(data[[target_col]] == class)
  prob_predictions <- predict(model, newdata = data, type = "prob")[, class]
  roc_result <- roc(response = data$binary_response, predictor = prob_predictions)
  auc_value <- auc(roc_result)
  auc_results[[class]] <- auc_value
 }

 return(auc_results)
}
```

7. Calculate Feature Importance

```
calculate_feature_importance <- function(model, data, method = "shapley") {
 if (method == "shapley") {
  X <- data[, -which(names(data) == "sco")]
  predictor <- Predictor$new(model, data = X, y = data$sco)
  shapley <- Shapley$new(predictor, x.interest = X[1, ])
  return(shapley$results)
 } else {
```

UNIVERSITÀ
DI PAVIA

```
    if (inherits(model, "randomForest")) {
      return(importance(model))
    } else {
      stop("Feature importance is not available or not supported for this model.")
    }
  }
}
```

8. General AUC Calculation

```
calculate_general_auc <- function(auc_data, weights = NULL) {
  if (!is.null(weights)) {
    general_auc <- auc_data %>%
      group_by(model_name) %>%
      summarise(General_AUC = sum(AUC * weights) / sum(weights))
  } else {
    general_auc <- auc_data %>%
      group_by(model_name) %>%
      summarise(General_AUC = mean(AUC))
  }
  return(general_auc)
}
```

9. Plotting ROC Curves

```
plot_roc_curves <- function(model_results) {
  plot(0, 0, type = "n", xlim = c(0, 1), ylim = c(0, 1),
       xlab = "False Positive Rate", ylab = "True Positive Rate",
       main = "ROC Curves")

  colors <- rainbow(length(model_results))
  legends <- vector("character", length(model_results))

  for (i in seq_along(model_results)) {
    roc_curve <- model_results[[i]]$ROC
    auc_value <- model_results[[i]]$AUC
    lines(roc_curve@x.values[[1]], roc_curve@y.values[[1]], col = colors[i], lwd = 2)
```

```
    legends[i] <- paste(names(model_results)[i], "AUC =", round(auc_value, 3))
  }

  legend("bottomright", legend = legends, col = colors, lty = 1, lwd = 2)
}
```

10. Calculate Metrics for Confusion Matrix

```
calculate_class_metrics <- function(conf_matrix, class) {
  cm <- as.matrix(conf_matrix$table)
  tp <- cm[class, class]
  fp <- sum(cm[, class]) - tp
  fn <- sum(cm[class, ]) - tp
  tn <- sum(cm) - (tp + fp + fn)

  return(c(TP = tp, FP = fp, TN = tn, FN = fn))
}
```

11. Visualize Accuracies

```
plot_accuracies <- function(accuracies) {
  accuracies_df <- enframe(accuracies, name = "Model", value = "Accuracy")
  accuracies_df$Model <- gsub("\\.Accuracy", "", accuracies_df$Model)
  sorted_accuracies <- accuracies_df %>%
    arrange(desc(Accuracy)) %>%
    mutate(Model = factor(Model, levels = .$Model))

  ggplot(sorted_accuracies, aes(x = Model, y = Accuracy, fill = Model)) +
    geom_bar(stat = "identity") +
    coord_flip() +
    labs(title = "Model Accuracy Comparison", x = "Model", y = "Accuracy") +
    theme_minimal() +
    theme(legend.position = "none")
}
```

UNIVERSITÀ
DI PAVIA

12. Generate and Plot Confusion Matrix Heatmap

```
plot_confusion_matrix_heatmap <- function(conf_matrix) {
  conf_matrix_table <- as.data.frame(conf_matrix$table)

  ggplot(data = conf_matrix_table, aes(x = Prediction, y = Reference, fill = Freq)) +
    geom_tile() +  # Heatmap cells
    geom_text(aes(label = Freq), color = "white") +  # Cell values
    scale_fill_gradient(low = "blue", high = "red") +
    labs(title = "Confusion Matrix Heatmap", x = "Predicted", y = "Actual") +
    theme_minimal()
}
```

13. Resampling Function for Robust AUC Estimates

```
perform_resampling  <- function(model, data, n_resamples = 100) {
  auc_results <- numeric(n_resamples)

  for (i in 1:n_resamples)  {
    resampled_indices <- sample(1:nrow(data),  replace = TRUE)
    resampled_data <- data[resampled_indices, ]

    prob_predictions <- predict(model, newdata = resampled_data, type = "prob")
    if (!is.null(prob_predictions))  {
      roc_curve <- roc(resampled_data$sco_binary,  prob_predictions[, class])
auc_results[i]  <- auc(roc_curve)
    }
  }

  return(auc_results)
}
```

14. Aggregate ESG Data

```
 esg_data <- data %>%
   dplyr::select(env, socr, govr) %>%
   na.omit() %>%
   esg_data$env <- as.factor(esg_data$env) esg_data$socr <- as.factor(esg_data$socr)
esg_data$govr <- as.factor(esg_data$govr)if (!require(dplyr)) { install.packages("dplyr")
library(dplyr) }

add_esg_to_year_data <- function(data_year, esg_data) {

cbind(data_year, esg_data) }

data_2020 <- add_esg_to_year_data(data_2020, esg_data) data_2021 <-
add_esg_to_year_data(data_2021, esg_data) data_2022 <- add_esg_to_year_data(data_2022,
esg_data)
```

15. Visualize ESG Impact

```
plot_esg_impact <- function(data_grouped_long) {
  ggplot(data_grouped_long, aes(x = reorder(prov, Impact), y = Impact, fill =
ESG_Dimension)) +
   geom_bar(stat = "identity", position = "dodge") +
   coord_flip() +
   labs(title = "Impact of ESG Categories by Region and Sector", x = "Region", y = "ESG
Impact") +
   theme_minimal() +
   facet_wrap(~ESG_Dimension)
```

Sources to Cite

https://en.wikipedia.org/wiki/Artificial_general_intelligence https://www.yahoo.com/now/6-jobs-artificial-intelligence-already-150339825.html

https://analyticsindiamag.com/5-ways-to-test-whether-agi-has-truly-arrived/

Kurzweil, Ray (2005), The Singularity is Near, Viking Press 30 Hickey, Alex. "Whole Brain Emulation: A Giant Step for Neuroscience". Tech Brew. Retrieved 8 November 2023.

https://www.safe.ai/statement-on-ai-risk

https://philosophynow.org/issues/132/Artificial_Consciousness_Our_Greatest_Ethical_Challenge

Molnar, C. (2022). *Interpretable Machine Learning.*

https://klimavest.de/en/knowledge/guide/esg-criteria/

Sustainalytics (2022). *ESG Risk Ratings Methodology.*

Lundberg, S. M., & Lee, S.-I. (2017). *Advances in Neural Information Processing Systems.*

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities.*

https://run.unl.pt/bitstream/10362/115436/1/TEGI0505.pdf

Breiman, L. (2001). "Random Forests." *Machine Learning*

Fawcett, T. (2006). "An Introduction to ROC Analysis." *Pattern Recognition Letters, 27(8), 861–874*

Friede, G., Busch, T., & Bassen, A. (2015). "ESG and Financial Performance: Aggregated Evidence from More than 2000 Empirical Studies." *Journal of Sustainable Finance & Investment, 5(4), 210–233.*

Hand, D. J., & Till, R. J. (2001). "A Simple Generalization of the Area Under the ROC Curve for Multiple Class Classification Problems." *Machine Learning, 45(2), 171–186.*

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling.* Springer.

Powers, D. M. W. (2011). "Evaluation: From Precision, Recall, and F-Measure to ROC, Informedness, Markedness, and Correlation.

https://mpra.ub.uni-muenchen.de/103027/1/MPRA_paper_103027.pdf

https://lutpub.lut.fi/bitstream/handle/10024/167125/Gradu%20Arttu%20Pirinen%2003.04.2024.pdf?sequence=1

UNIVERSITÀ
DI PAVIA