UNIVERSITÀ
DI PAVIA

Università degli Studi di Pavia

Dipartimento di Scienze Economiche e Aziendali

Corso di Laurea Magistrale in Finance

APPLICATION OF MACHINE LEARNING MODELS IN CREDIT RATING ANALYSIS

Relatore:

Chiar.mo Prof. Paolo Giudici

Tesi di Laurea di

Zhang Caoxutong

Matr. n.513754

Anno Accademico 2024/2025

APPLICATION OF MACHINE LEARNING MODELS IN CREDIT RATING ANALYSIS

ABSTRACT

This paper investigates the application of machine learning in credit rating. Through an extensive review of literature, it is found that the credit rating process necessitates the integration of machine learning techniques. Utilizing data from 1,559 Italian companies, the study addresses the issue of data imbalance through oversampling. Subsequently, seven active learning models are trained, with the Random Forest model identified as the most accurate predictor. It is also observed that deep learning algorithms generally yield favorable results, although artificial neural network models are susceptible to the influence of hardware performance, necessitating further enhancement.

The paper further incorporates a feature importance analysis, highlighting the critical role of credit ratings and shareholders' funds in the credit assessment process in 2021. Additionally, a novel interpretability method, namely Rank Gradient Explainability (RGE), is introduced to enhance model transparency and credibility. The RGE method reveals that shareholders' funds are the most influential feature, and financial data from 2022 is the most critical across all financial years. Based on these findings, targeted recommendations are provided for credit rating agencies, operating companies, and financial researchers, emphasizing the importance and practical value of applying machine learning models to the credit rating system.

Keywords: Machine Learning, Credit Rating, Multiclass Classification Models, Oversampling, Explainability

APPLICAZIONE DEI MODELLI DI APPRENDIMENTO AUTOMATICO NELL'ANALISI DEL RATING CREDITIZIO

RIASSUNTO

Questa ricerca esamina l'applicazione dell'apprendimento automatico nella valutazione del credito. Attraverso una revisione estensiva della letteratura, si è constatato che il processo di valutazione del credito richiede l'integrazione di tecniche di apprendimento automatico. Utilizzando dati provenienti da 1.559 aziende italiane, lo studio affronta il problema dell'imbalance dei dati attraverso il sovracampionamento. Successivamente, vengono addestrati sette modelli di apprendimento attivo, identificando il modello Random Forest come il predittore più accurato. Si osserva inoltre che gli algoritmi di deep learning producono generalmente risultati favorevoli, sebbene i modelli di reti neurali artificiali siano suscettibili all'influenza delle prestazioni hardware, richiedendo ulteriori miglioramenti.

La ricerca include inoltre un'analisi dell'importanza delle caratteristiche, evidenziando il ruolo cruciale dei rating del credito e dei fondi dei soci nel processo di valutazione del credito nel 2021. Inoltre, viene introdotto un nuovo metodo di interpretazione, denominato Rank Gradient Explainability (RGE), per migliorare la trasparenza e la credibilità del modello. Il metodo RGE rivela che i fondi dei soci sono la caratteristica più influente e che i dati finanziari del 2022 sono i più critici tra tutti gli anni finanziari. Sulla base di queste scoperte, vengono fornite raccomandazioni mirate per le agenzie di rating del credito, le aziende operative e i ricercatori finanziari, sottolineando l'importanza e il valore pratico dell'applicazione dei modelli di apprendimento automatico al sistema di valutazione del credito.

Parole chiave: Apprendimento Automatico, Rating Creditizio, Modelli di Classificazione Multiclasse, Sovracampionamento, Spiegabili

ACKNOWLEDGEMENTS

The results of a research paper are undoubtedly significant, but the acknowledgments in a thesis are equally important, especially in a graduation thesis, as they mark the conclusion of a phase in one's life. As this article signifies the end of my academic journey in Italy, I feel the need to express gratitude to many individuals.

Firstly, I would like to thank my thesis supervisor, Professor Giudici. His lectures were exceptionally clear, making them very accessible to international students. I will always remember his gentle greetings and patient explanations, which he offered as if he were a friend rather than a teacher. His passion for machine learning and artificial intelligence inspired me greatly, and his exceptional academic prowess greatly assisted me in completing this thesis.

Next, I wish to express my appreciation to all the staff at Pavia, and the Università degli Studi di Pavia for the valuable two years they have accompanied me. They provided an environment of an esteemed institution and enabled me to acquire substantial knowledge within a short span of two years.

I also want to thank my friends in Italy: Mina, Letizia, Clara, Devanchi, Konghao, Youpeng, and others. They not only supported me in my studies but also stood by me as dear friends, ensuring that I never felt lonely while abroad.

Additionally, I wish to thank my friends in China who have been waiting for my return, particularly Sally Gao, Yunuo, and Dongzhi, who were with me through many tough times, and our shared cat, Tangyuan.

I must also thank my parents, who have supported me in every way. I am grateful to have been born into a happy and warm family, and I hope that with the completion of this article, it brings additional joy to my mother.

Lastly, I believe I should thank myself for persevering and accomplishing numerous tasks. This article is a lifelong medal of honor for me.

Thank you to the beautiful city of Pavia, for making Italy my second home.

Carrying sincere apologies, I would like to add some further acknowledgments after the completion of my thesis, as many unforeseen events have taken place. First, I wish to express my gratitude to Shandong Energy Group for the job opportunity they have offered me. I will embrace this new role with the best of my abilities.

Secondly, I must thank Xue Yiyi, who has put in tremendous effort in the process of submitting my re-entry visa application, aiding me in obtaining the necessary documents with remarkable speed.

Thirdly, I am grateful to my university for its efficient administrative processes, which have helped me quickly complete my graduation application and various other documentations. This has been of great importance to me, and I look forward to completing my defense on schedule, graduating on time, and starting my new job as planned.

I also wish to thank my friends who have silently supported me. Although I cannot mention each one by name, their efforts are engraved in my heart.

Contents

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

1.1 Literature Review

With the maturation of artificial intelligence, big data technology, and machine learning, the application of machine learning in the financial sector has become increasingly widespread, particularly in the area of credit rating assessment and prediction, which holds significant potential. Following the post-pandemic recovery of the financial markets, intensified competition has highlighted the crucial role of credit ratings in loan decision-making. The integration of artificial intelligence to enhance the efficiency of credit rating technologies has emerged as a key research direction for credit rating agencies. Concurrently, corporations are also focusing on improving their credit ratings as an essential aspect of enhancing their competitiveness.

Pedregosa et al. (2011) have created and enriched the scikit-learn (sklearn) library, establishing a comprehensive functional foundation for machine learning applications in Python. Existing research has offered extensive insights into the application of machine learning in credit rating. The utilization of machine learning in this domain primarily employs three types of technologies. The first is classification techniques, where models like decision trees are common due to their interpretability and high predictive accuracy. Additionally, support vector machines, which project data into a higher-dimensional space to find the maximum margin hyperplane, and artificial neural networks, which simulate the connections between human brain neurons to handle more complex nonlinear relationships, are also widely used.

Another extensively applied approach is ensemble methods, which include tree-based models like decision trees and random forests, as well as ensemble algorithms such as gradient boosting

and XGBoost. The third type is clustering techniques, with k-nearest neighbors being a primary method for analyzing similar data.

Researchers have conducted comparative studies on different machine learning models. Random forests, XGBoost, and deep neural networks are notably superior models. Dai et al. (2021) found that random forests performed best in predicting bank credit ratings, while Alonso & Carbó (2021) discovered that XGBoost had higher accuracy and calibration capabilities in predicting consumer defaults.

Investigators have also explored the practical application of machine learning models in the field of credit risk assessment. Alonso & Carbó (2021) used the XGBoost model to predict consumer defaults and assessed its economic impact, finding that it could save up to 17% in regulatory capital. Dai et al. (2021) employed survival analysis to study the relationship between a company's daily income and its credit rating, revealing that the number of days with zero or negative income affects the company's credit rating.

The "black box" nature of machine learning models has led to a demand for Explainable AI (XAI) methods. Babaei et al. (2023) used the Shapley value method to explain the credit scores predicted by the random forest model and guided feature selection to balance predictive accuracy and interpretability. Building on this, Giudici, P., & Raffinetti, E. (2024) proposed the Rank Graduation Explainability (RGE) method for interpretable evaluation. RGE is a rank-based gradient explainability method that calculates the impact of each feature on the model's predictive outcomes, taking into account the ranking information of the features.

1.2 Data Interpretation

This dataset is utilized to analyze credit rating conclusions from Modifinance, based on credit

rating data of 1,559 Italian companies. It encompasses significant financial indicators for the years 2022, 2021, and 2020, including Total Assets, Shareholders' Funds, Net Income, EBIT, EBITDA, among others. Additionally, the dataset incorporates details about the companies' business sectors, geographical locations, ESG scores, and includes credit rating outcomes for the years 2020 and 2021 as references. This comprehensive dataset provides a thorough overview of the companies' operational and financial performance over the past three years. The aim is to use this data to construct a superior machine learning model for credit rating utilization.

This paper employs credit rating conclusions data provided by Modefinance for analysis.

Utilizing the information provided by Ahelegbey, D., Giudici, P., & Pediroda, V. (2023), Modefinance is a financial technology company registered with the ESMA (European Securities and Markets Authority) and acts as a credit rating agency. Although the company does not operate as a peer-to-peer platform, it offers scoring services to investors.

Modefinance assesses credit ratings by extracting financial information from companies' publicly disclosed balance sheets and income statements.

The company provides credit rating analyses not only for Italy but also for other European countries, employing methods to reduce fiscal legislative differences and maintain consistency in supervisory data across European nations.

The credit assessment model used by Modefinance is the Multi-Objective Rating Evaluation (MORE) model. The core principle of this model involves observing and analyzing a company's financial statements to assess its financial behavior and operational status, followed by the allocation of a credit grade. To achieve a higher rating category, a company should maintain

balanced development in key financial variables such as profitability, liquidity, solvency, and coverage ratios.

1.3 Practical Significance

This project aims to investigate a machine learning approach for predicting credit ratings. By comparing different credit rating models for the data, the optimal machine learning model is selected. To enhance model explainability, an innovative explainability assessment method (RGE value) is creatively applied to explain different features, providing new insights for companies to improve their credit ratings and offering a perspective in the field of artificial intelligence for credit rating work.

1.4 Analysis Process

The first step of this paper involves interpreting the data by examining the naming and specific meanings of various factors, followed by a visual interpretation of some data. By categorizing and visualizing the basic data, we can increase familiarity with it, which provides insights for future data preprocessing. For instance, visualizing the target variable categories not only confirms the construction of a multi-class machine learning model but also reveals issues of data imbalance. This lays the foundation for the subsequent analysis and the selection of appropriate model construction and solutions.

The second step is data preprocessing after understanding the data and variables. This includes steps such as identifying and supplementing missing data, encoding non-numeric data, etc. By preprocessing the data, we address the identified issues, preparing the data for subsequent model construction.

The third step is model construction. This paper employs seven models: Logistic Regression,

Decision Tree, Random Forest, Gradient Boosting Machine, Support Vector Machine, Artificial Neural Network, and k-Nearest Neighbors. Initially, I optimize the models using techniques like random search, grid search, and cross-validation to obtain the best model configurations. Then, I compare the accuracy and explainability of these models. The optimal model is selected, and its construction method is evaluated using metrics such as accuracy, confusion matrix, classification report, learning curves, and feature importance (or explainability).

Finally, the entire process is summarized. The optimal model is chosen by screening for the one with the best accuracy, and an innovative explainability assessment method (RGE) is used to evaluate the feature importance of the model, enhancing its explainability.

1.5 ODF Measures of Performance

The analytical tools employed in this study are based on the scikit-learn packages developed by Pedregosa et al. (2011) (http://scikit-learn.org/). These tools are integrated into the Python code via the Anaconda suite (https://www.anaconda.com/download).

1.5.1 Accuracy Tool

The metric for accuracy assessment in this paper is the accuracy_score function, which is part of the sklearn.metrics module in the scikit-learn (sklearn) library. The fundamental concept of this function is to calculate the proportion of correctly predicted samples by the model relative to the total number of samples. The formula for accuracy is as follows:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Samples}$$

Therefore, for the accuracy_score function, the printed accuracy will be a number between 0 and 1. The closer the value is to 1, the higher the probability of correct predictions, indicating that the model has a more precise predictive capability.

1.5.2 Classification Report Tool

This paper investigates classification models and also employs a classification report to intuitively display the prediction outcomes for each category. The output of the classification report is implemented through the classification_report function in scikit-learn.

In the scikit-learn library, the classification_report function provides the precision, recall, and F1-score for each category. By default, the F1-score is calculated using a threshold of 0.5. Generally, the default threshold of 0.5 for the F1-score in the classification_report is a reasonable choice under balanced distribution circumstances. This paper utilizes oversampling to balance the data, addressing the relative balance in the data distribution, hence continues to use the threshold of 0.5.

1.5.3 Explainability Tool

Additionally, this paper employs the compute_rge_values function from the safeaipackage library provided by Giudici, P., & Raffinetti, E. (2024), utilizing Rank Gradient Explainability (RGE) as a measure of explainability.

The paper contrasts the conclusions regarding the importance of features derived from using Rank Gradient Explainability (RGE) with those computed by the corresponding white-box model in terms of feature importance. This comparison aims to explore different standards for measuring feature importance across various patterns.

1.6 Innovation Points

This paper applies extensive grid search and cross-validation methods in model construction to enhance the generalizability of the models and reduce the risk of overfitting. Additionally, learning curves are utilized to determine if overfitting has occurred.

Furthermore, this paper innovatively adopts the Rank Gradient Explainability (RGE) as a standard for measuring explainability. This approach addresses the issue of black-box models being unable to utilize feature importance ranking, establishing a unified explainability criterion for evaluating data across models.

1.7 Shortcomings

The present study primarily investigates the differences among various models. Consequently, in the model training phase, all models were selected for training without further feature selection. Future work can target different models for more specific analysis.

Chapter 2 Data Introduction

2.1 Interpretation of Data Shape

The original dataset utilized data from 1,559 companies, encompassing a total of 33 columns. Excluding the company name, code, and the credit rating results for the year 2022, which were used as the target variable, there are 31 features in total. By assessing the data shape, one can gauge the scale of the data. This is a set of medium-sized original data. Given that it is real-world data, there is a high likelihood of encountering classification imbalance issues. Additionally, with a large number of features involved, a gradual analysis of each feature is required.

2.2 Feature Interpretation

In the interest of coding efficiency, the dataset utilized in this research employs abbreviated column names. Table 1 serves as a reference, detailing the original and abbreviated names of each feature. This table includes all the features that were incorporated during the model training process.

Firstly, the Company ID column, which serves as a substitute for the company name, is excluded as the numerical values in this column do not possess any meaningful order and may also affect the accuracy of the model's predictions.

Furthermore, I have not categorized the data by year additionally. Because rating companies need to consider a company's financial data from the past three years to comprehensively assess its operational status. However, to analyze the extent to which the time frame influences the company's credit rating results, I selected the top three models for calculating the RGE values with year group during the interpretability analysis. This is done to determine whether the

interpretability of any of these three better-performing models is affected by the combination of time periods.

Table 1 Column Name Comparison Chart

Column Name	ex-Column Name	Column Name	ex-Column Name
Company ID	Company ID	SF_22	Shareholders funds th EUR 2022
SECO_22	MORE evaluation - Score 2022	SF_21	Shareholders funds th EUR 2021
SECT	sectors	SF_20	Shareholders funds th EUR 2020
REG	Region in country	CL_20	Current liabilities th EUR 2020
ESG_CLA	ESG_class (S1 best / s7 worst)	ORT_22	Operating revenue (Turnover) th EUR 2022
ENV_R	ENV Rating	ORT_21	Operating revenue (Turnover) th EUR 2021
Social_R	Social Rating	ORT_20	Operating revenue (Turnover) th EUR 2020
Gov_R	Governence Rating	EBIT_22	Operating profit (loss) [EBIT] th EUR 2022
SECO_21	MORE evaluation - Score 2021	EBIT_21	Operating profit (loss) [EBIT] th EUR 2021
SECO_20	MORE evaluation - Score 2020	EBIT_20	Operating profit (loss) [EBIT] th EUR 2020
TA_22	Total assets th EUR 2022	NI_22	Profit (loss) for the period [Net income] th EUR 2022
TA_21	Total assets th EUR 2021	NI_21	Profit (loss) for the period [Net income] th EUR 2021
TA_20	Total assets th EUR 2020	NI_20	Profit (loss) for the period [Net income] th EUR 2020
CA_22	Current assets th EUR 2022	EBITDA_22	EBITDA th EUR 2022
CA_21	Current assets th EUR 2021	EBITDA_21	EBITDA th EUR 2021
CA_20	Current assets th EUR 2020	EBITDA_20	EBITDA th EUR 2020

2.3 Data Visualization

2.3.1 Descriptive Statistics of Raw Data Variables

Initially, descriptive statistics were conducted on all the existing data variables, and a portion of the results is presented in **Error! Reference source not found.**.

Table 2 Descriptive Statistics of Raw Data Variables (Partial）

	count	unique	top	freq
SECO_22	1559	10	BB	486
SECT	1559	29	Business Services	235
REG	1559	20	Lombardia	349
ESG_CLA	1559	7	S3	639
ENV_R	1523	7	S2	810
Social_R	1418	7	S4	302
Gov_R	1360	6	S3	506
SECO_21	1559	11	BB	486
SECO_20	1559	11	BB	453
TA_22	1559	1550	n.a.	9
TA_21	1559	1545	n.a.	15
TA_20	1559	1515	n.a.	45
CA_22	1559	1550	n.a.	9

The descriptive statistics of the variables indicate a significant imbalance in the credit rating outcomes (SECO_22, SECO_21, SECO_20), with a high frequency of the 'BB' rating. Additionally, the company locations are predominantly concentrated in the Lombardy region, suggesting a severe classification imbalance in the acquired real-world data.

For numerical data, the presence of missing values across all numerical features (the table above shows only a portion of these features) indicates the need for identifying and correcting these data gaps.

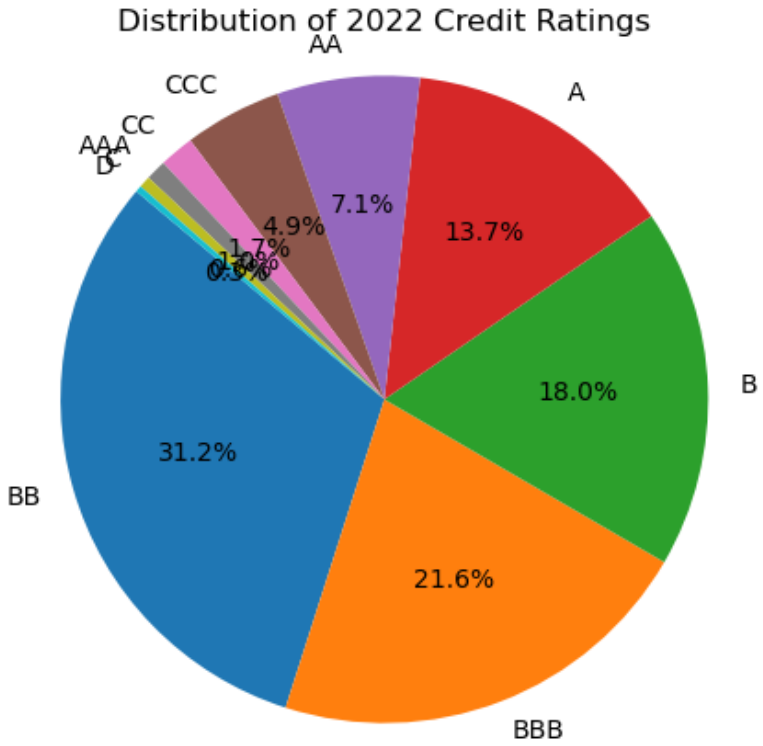
2.3.2 Visualization of Non-numeric Data

To gain a more intuitive understanding of the distribution of non-numeric data, I continued with a visual analysis of several non-numeric label features.

2.3.2.1 Visualization of the Distribution of Target Variables

From the descriptive statistics, we can observe an indication of classification imbalance, with a concentration of the 'BB' category in the target variables. **Error! Reference source not found.** presents the visualization of the classification of the target variable.

Fig 1 Distribution of Credit Ratings in 2022



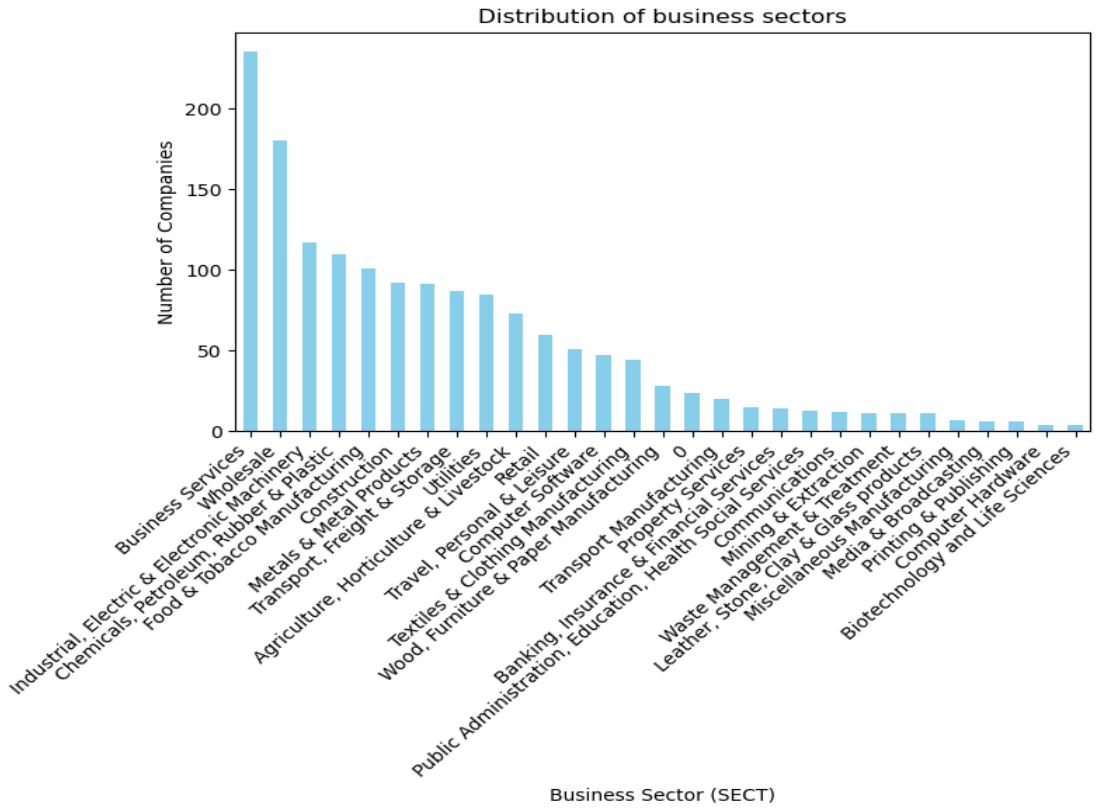
As evident from the image, over 30% of the rating results are 'BB'; the ratings 'A', 'B', 'BB', and 'BBB' constitute over 80% of the data's rating outcomes, indicating a severe classification imbalance in the dataset. Therefore, addressing this classification imbalance is essential before constructing the model.

3.2.2.2 Visualization of Company Business Areas

I also generated a bar chart for the company's business areas to assess the sectors in which the analyzed companies are concentrated. Fig 2 illustrates the distribution of company business areas.

Based on the image, it can be observed that the most prevalent business area is Business Services, while Biotechnology and Life Sciences have the lowest frequency of occurrence. Additionally, some companies do not have a detailed business area specified (represented as 0). However, the overall distribution of company sectors is relatively balanced, with a diversity that can provide a criterion for tree models, enhancing their generalization capabilities.

Fig 2 Distribution of Company Business Areas

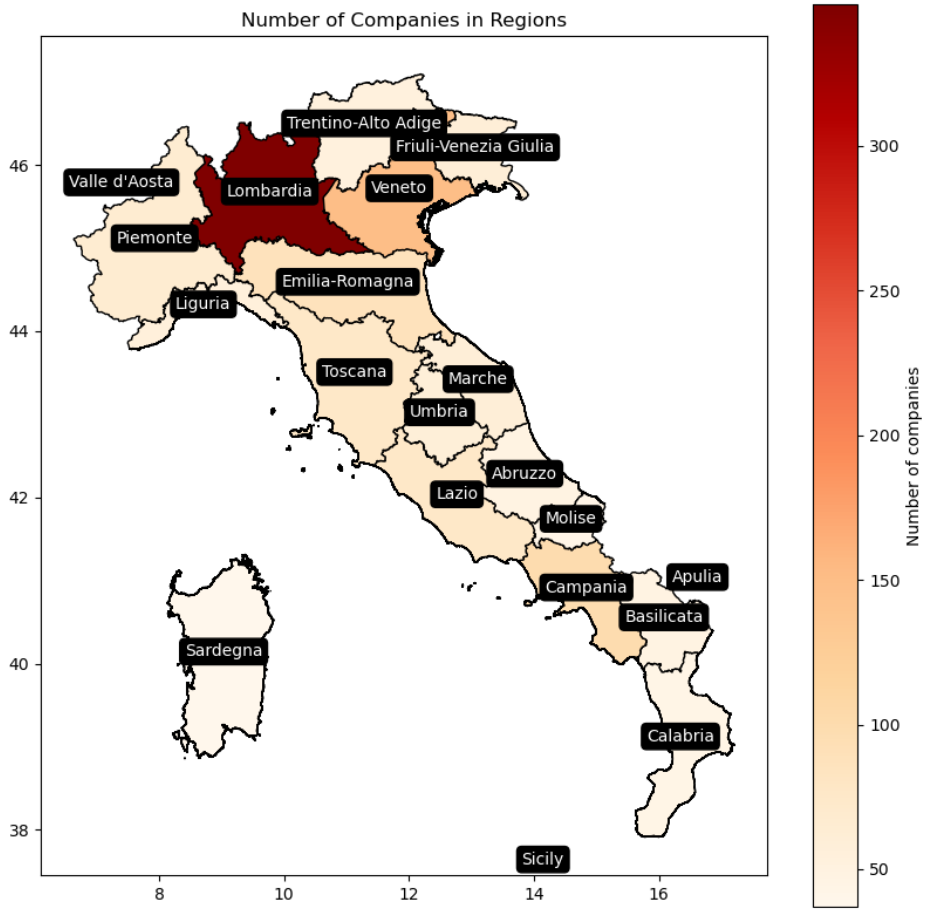


3.2.2.3 Frequency Distribution Heatmap by Region

Fig 3 presents a heatmap of the frequency distribution by company location, indicating that the data collected is predominantly concentrated in the Lombardia region. This also suggests that the REG variable representing the region has an imbalanced distribution. Consequently, it can be concluded that the original data exhibits classification imbalance in many non-numeric variables. Therefore, it is necessary to employ certain learning methods to address this issue of

imbalanced distribution.

Fig 3 Heatmap of Company Frequency Distribution by Region



2.3 Data Standardization and Transformation

2.3.1 Distinguishing Feature Columns

I divided the feature columns into numeric and non-numeric types, treating all features as individual factors influencing the target variable (SECO_22) without further differentiation by year. This facilitates subsequent data preprocessing and labeling efforts.

2.3.2 Identifying and Filling Outliers

Given the significant impact of every fluctuation in financial data on credit ratings, and considering the substantial differences in financial data due to varying company sizes, business areas, and regions, outliers were only addressed in terms of missing value imputation. If data

dimensionality reduction is required during model construction, it will be handled separately in that process.

For non-numeric data, missing values were exclusively found in ESG class ratings. Since ESG ratings have a hierarchical order, to maintain the data's rating trend, I chose the most neutral evaluation, S4, to fill in the missing values. This approach neither introduces additional categories nor affects the trend of the data ratings.

For handling missing values, I replaced occurrences of missing values and special characters that were not numbers with 0. The absence of financial data can significantly impact a company's credit evaluation. Using 0 to represent missing values simulates the scenario of missing financial data, accurately reflecting the influence of such missing data on credit ratings.

2.3.3 Data Labeling

I employed label encoding, first creating a dictionary and then labeling all non-numeric data. Using label encoding and dictionary creation helps to preserve the data's hierarchical trend and enhance the model's generalization capabilities.

2.4 Data Segmentation and Processing

2.4.1 Original Data Segmentation

Segmenting the original data was done to reiterate the issue of target function classification imbalance, facilitating the subsequent use of simulated data methods.

2.4.2 SMOTE Method

Following the approach of Pamuk, M., & Schumann, M. (2023), the SMOTE oversampling technique was used to address the issue of class imbalance. I created a new dataset (named df) using SMOTE oversampling for actual model construction. Table 3 presents the descriptive

statistics of the df dataset after oversampling. The results indicate that numeric data no longer contains special characters or missing values, allowing for normal descriptive calculations. The target variable's class distribution has reached a balanced state, making it a suitable dataset for machine learning. Using the df dataset, I randomly sampled 20% as the test set and the remaining 80% as the training set, employing this data classification method for machine learning model training.

Table 3 Descriptive Statistics of df Variables (Partial)

	count	mean	std	min	max
SECT	4860	13.75082305	8.2650232	0	28
REG	4860	9.567489712	4.532175869	0	19
ESG_CLA	4860	3.038065844	0.852336935	1	7
ENV_R	4860	2.128806584	1.132447867	1	7
SECO_21	4860	4.816872428	2.211264938	0	10
SECO_20	4860	4.859053498	2.078736796	0	10
TA_22	4860	140096.7684	453397.4025	0	14392422
CA_20	4860	56845.10205	149478.8463	0	2255696
SF_22	4860	42240.86076	183500.484	-49091	5336752
EBITDA_22	4860	13371.55434	117114.6028	-303716	2095592
SECO_22	4860	5.5	2.872576871	1	10
SECO_21	4860	4.816872428	2.211264938	0	3
SECO_20	4860	4.859053498	2.078736796	0	3

Chapter 3 Model Building & Assessment

The data I used is classification value, hence regression models suitable for time series cannot be applied. Among various classification models, I ultimately selected seven modeling approaches for analysis: Logistic Regression model, Decision Tree model, Random Forest model, Gradient Boosting Machine model, Support Vector Machine model, Artificial Neural Network model, and K-Nearest Neighbors model.

My overall approach to building the models was as follows: Initially, I created a basic model to preliminarily assess its accuracy. Subsequently, I optimized the model, attempting methods such as cross-validation and grid search to enhance model precision and test its generalization capability. Finally, I selected the model with the strongest generalization ability, which I refer to as the optimal model. For the optimal model under each approach, I will evaluate their accuracy, confusion matrix, classification report, learning curve, and feature importance (only for white-box models), to facilitate comparison between different types of models.

3.1 logistics Regression (GLM)

3.1.1. Building the Basic Model

The accuracy of the model trained on the training and test sets using the sklearn's Logistics Regression model is shown in **Error! Reference source not found.** It is evident that the accuracy of the original model is poor. Additionally, a warning was encountered during the execution of the code. This warning suggested increasing the maximum number of iterations to ensure model convergence.

Table 4 Accuracy of the Original GLM Model

Accuracy on the training set of pre-glm	0.5491255144032922
Accuracy on the test set of pre-glm:	0.5277777777777778

Furthermore, the warning indicated the need for scaling the data before training the model. Therefore, it is necessary to optimize the original model.

3.1.2. Model Optimization

To enhance model accuracy and ensure convergence while avoiding warnings, I adopted two approaches during the optimization process. The first involved increasing the maximum number of iterations for the model, and the second was the implementation of grid search. Additionally, I used the same scaler to scale the data before training the model.

I will apply the StandardScaler from the sklearn library for data scaling. To maintain consistency in the values, I have given a uniform name to the scaling results to prevent confusion with other scaler outcomes.

3.1.2.1. Increasing Maximum Iterations

I named the model with the increased maximum iterations as "logis2" and set the maximum number of iterations to 2000. The accuracy of this model is presented in **Error! Reference source not found.**

Table 5 Accuracy of the Model with Higher Maximum Iterations

Accuracy on the training set for glm2/higher-max_iter	0.6134259259259259
Accuracy on the test set for glm-2/higher-max_iter	0.5936213991769548

Upon comparing the optimized accuracy table, it was found that increasing the maximum number of iterations led to a certain improvement in the model's accuracy, without any additional warnings. However, the training results on the training set were not as expected, with only about 60% success rate in training. Considering the suboptimal training outcomes, I decided to employ a grid search approach to find the optimal parameters.

3.1.2.2 Grid Search

Due to the prolonged time required for the grid search of the maximum number of iterations in

logistic regression, I maintained the maximum number of iterations at 2000 and planned to introduce a first-order lasso parameter (l1). Among the parameters selectable in the optimization algorithm, the first is the solver. In logistic regression algorithms, the solver defaults to liblinear. However, for multi-class classification problems, I need to employ alternative labeling methods that facilitate classification across all categories, which liblinear does not support the use of first-order regularization parameters, hence the need to modify the solver. To obtain the answer as efficiently as possible, I chose to directly use the saga, a stochastic optimization algorithm with linear convergence.

Therefore, I set the object of the grid search to be the regularization parameter C, which is the reciprocal of the regularization strength. Additionally, I used 4-fold cross-validation to enhance generalization while reducing computation time.

Table 6 Results of Cross-Validated Grid Search

Best parameters of glm3/grid search	{'C': 10}
Accuracy on the training set for glm3/grid search	0.5979938271604939
Accuracy on the test set for glm3/grid search	0.581275720164609

Error! Reference source not found. displays the optimal parameters and accuracy information obtained from the grid search. Additionally, the model issued a warning of non-convergence. Based on the conclusion, it is evident that significant modifications are required for the logistic regression model to be applied to this dataset. The current computational process exhibits poor accuracy and fails to converge even after 2000 iterations, indicating that the conclusions from the grid search are not suitable for the model's predictions.

3.1.3 Model Assessment

3.1.3.1 Optimal Model

From the above optimization process, it is evident that the model during the cross-validation grid search did not converge in terms of parameters and exhibited poor accuracy. Therefore, it was decided to directly use a model with the number of iterations increased to 2000, namely GLM2/higher-max_iter.

The accuracy of the optimal model is shown in **Error! Reference source not found.** above.

3.1.3.2 Confusion Matrix

To visualize the predictive outcomes of the model more intuitively, I have utilized color-coded confusion matrices to depict the model's performance on both the training and test sets. Fig 4,Fig 5 represent the actual confusion matrices of the model on the training and test sets.

Based on the comparison of the confusion matrices, it can be observed that the model has poor generalization capabilities for classes B, BB, and BBB, leading to easy confusion among them.

Fig 4 Confusion Matrix of the Optimal GLM Model on the Training Set

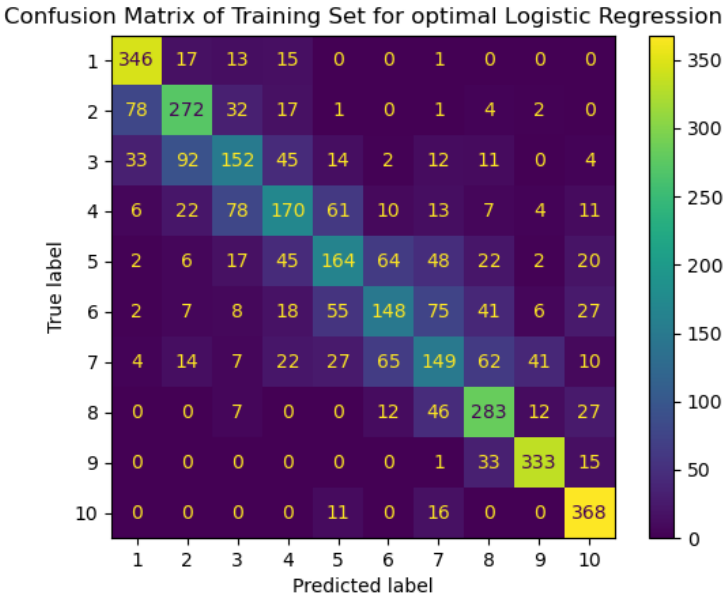
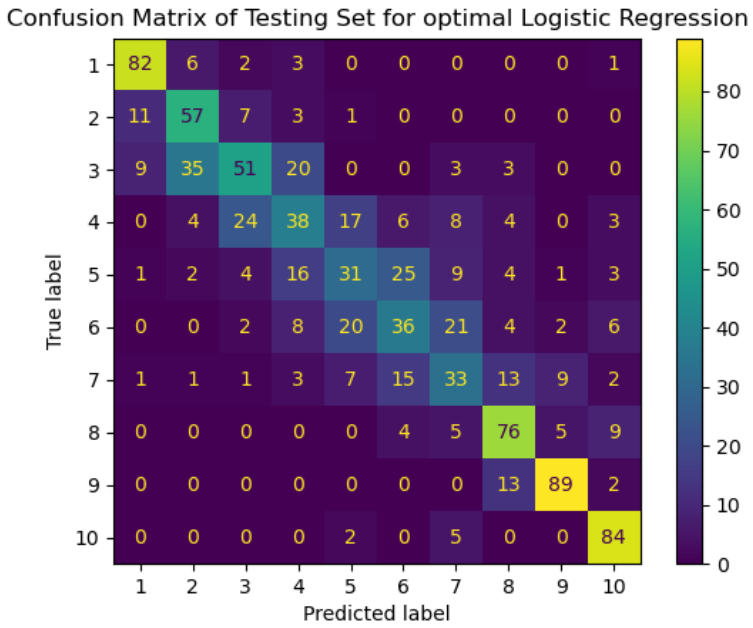


Fig 5 Confusion Matrix of the Optimal GLM Model on the Testing Set



Although the accuracy of this model is not ideal, with only 60% of the predictions being correct, the models that were not successfully predicted are mostly concentrated around the correct conclusions.

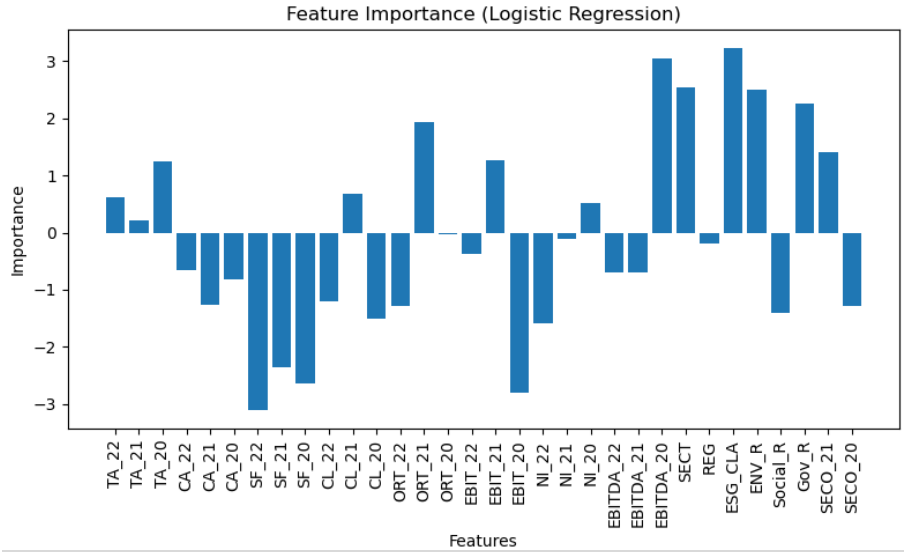
3.1.3.3 Feature Importance

The feature importance analysis for the GLM model is essentially an analysis of the logistic regression coefficients. By examining the model coefficients, one can assess the impact and importance of each feature. The magnitude of the coefficient's absolute value determines the level of importance of the feature. After sorting the coefficients by their absolute values in descending order, it is found that the top five most important features are: ESG_CLA, SF_22, EBITDA_20, EBIT_20, and SF_20.

The results indicate that under the absolute value analysis of the coefficients, the overall ESG score is the most influential feature and has a positive correlation. In the context of this model, Shareholders' Funds are also a significant influencing factor.

Additionally, the sign (positive or negative) of the coefficient can determine the relationship between the feature and the predicted outcome. A positive coefficient indicates a positive correlation, while a negative coefficient indicates the opposite.

Fig 6 GLM Feature Importance



To visualize the importance of feature impact, Fig 6 is a bar chart I created based on the coefficients. Features above the x-axis have positive coefficients, indicating a positive correlation with the target variable, while those below the x-axis represent negatively correlated variables.

The plot reveals that the ESG score is a significant factor with a strong positive correlation. Additionally, Shareholders funds exhibit a substantial negative correlation.

Furthermore, the balance of coefficients is not poor, with a minimal discrepancy in the number of positive and negative coefficients and no significant fluctuation, indicating that the model is relatively balanced.

3.1.3.4 Learning Curve

After model training, it is crucial to be vigilant about the possibility of overfitting, and it is

necessary to have some intuitive methods to test for it. A common visual method to check for overfitting in a model is the learning curve.

The principle behind the learning curve is to use cross-validation and grid search to determine whether the gap between the training and test sets gradually increases as the number of training samples increases.

Fig 7 GLM Model Learning Curve

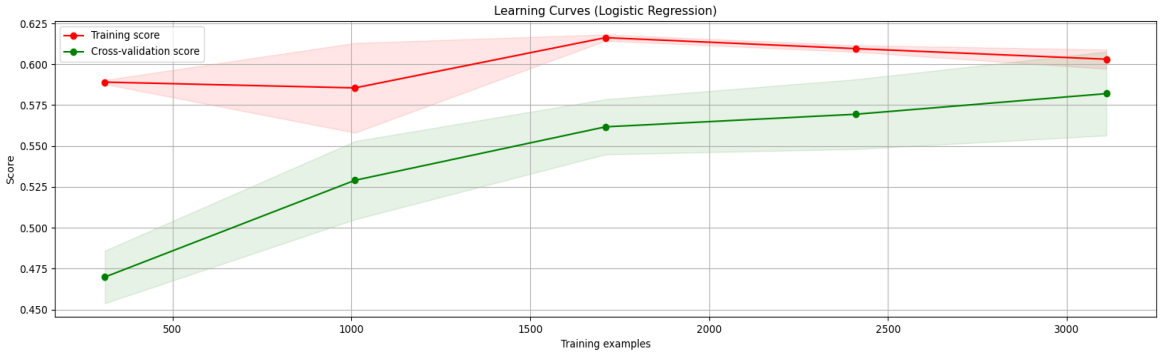


Fig 7 is the learning curve of the optimal GLM model. Since the model's accuracy is not very high, the GLM model I designed does not seem to exhibit signs of overfitting. As the amount of training data increases, the accuracy gap between the training and test sets gradually decreases.

In summary, although the GLM model's results in the confusion matrix and overfitting analysis are acceptable, the overall accuracy is low, and it does not possess the level of successful prediction. I will continue to explore other methods to improve the predictive capability.

3.2 Decision Tree Model

3.2.1 Constructing the Original Model

Firstly, I directly used the decision tree model constructor from the SKLEARN library to build an original decision tree model and printed its accuracy. As shown in Table 7, the original decision tree model has an accuracy of 1.0 on the training set and 0.8 on the test set. The

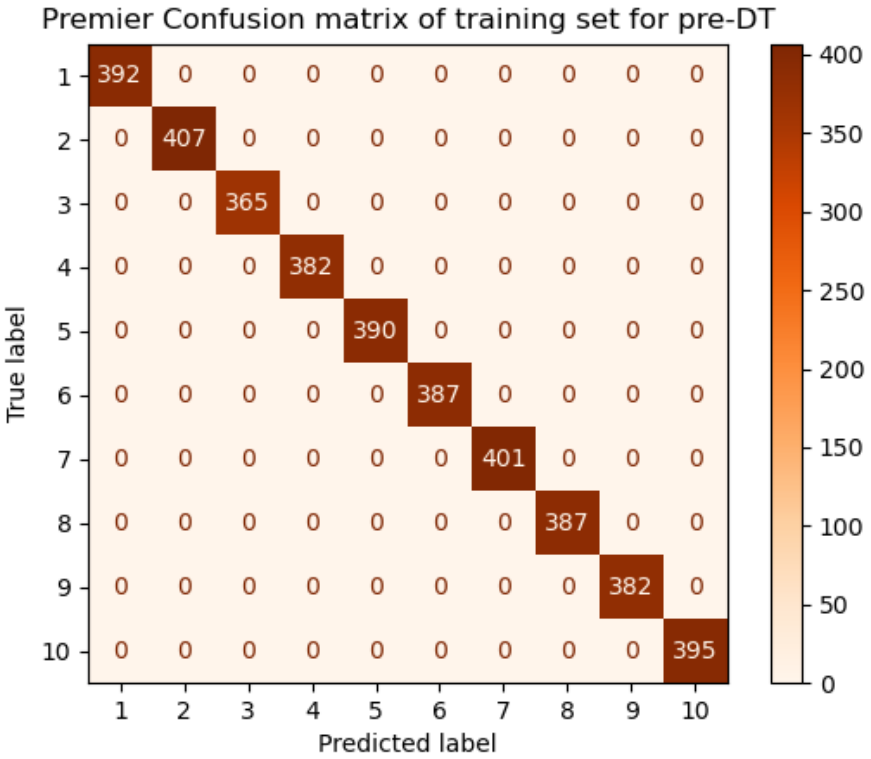
appearance of 1.0 accuracy suggests that the model is likely to be overfitted.

Table 7 Original Decision Tree Model Accuracy

Accuracy on the training set of pre-dicision tree	1.0
Accuracy on the test set of pre-dicision tree	0.8024691358024691

3.2.1.2 Confusion Matrix of Original Model

Fig 8 Premier Confusion matrix of training set for pre-DT



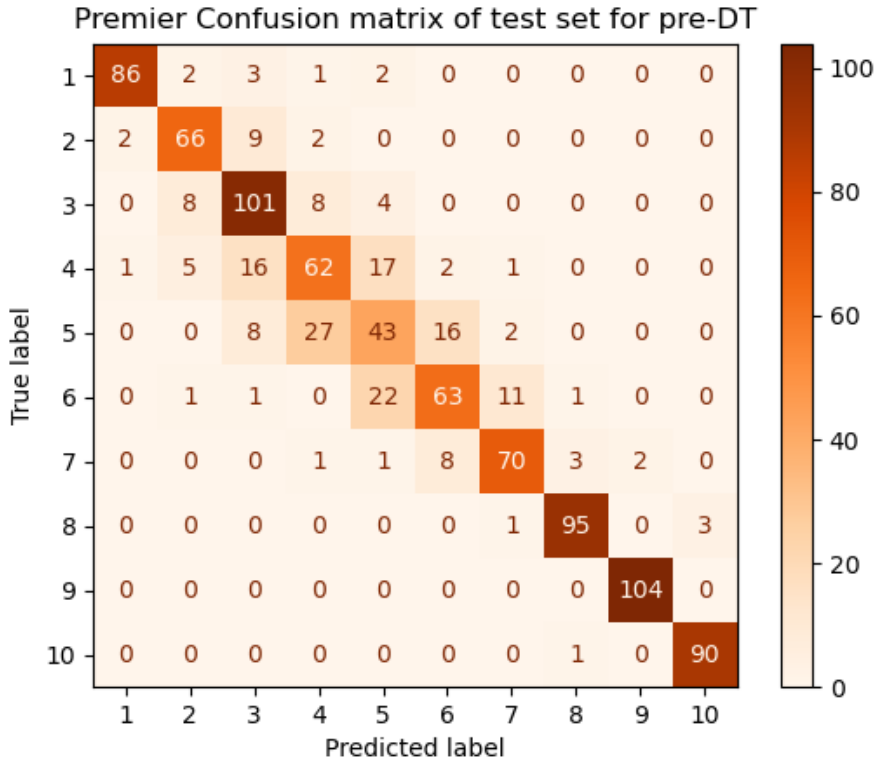
To make the analysis more intuitive, I also printed the confusion matrices of the original model on the training and test sets, as shown in Fig 8 and Fig 9.

The confusion matrices indicate that the original decision tree model is likely to be overfitted. To investigate whether the overfitting is due to the model's complexity, I visualized the tree of this model. From the visualization, it can be observed that the decision tree model is highly complex, with more than 15 levels.

The visualization of the model further confirms that the obtained decision tree model suffers

from significant overfitting issues.

Fig 9 Premier Confusion matrix of testing set for pre-DT

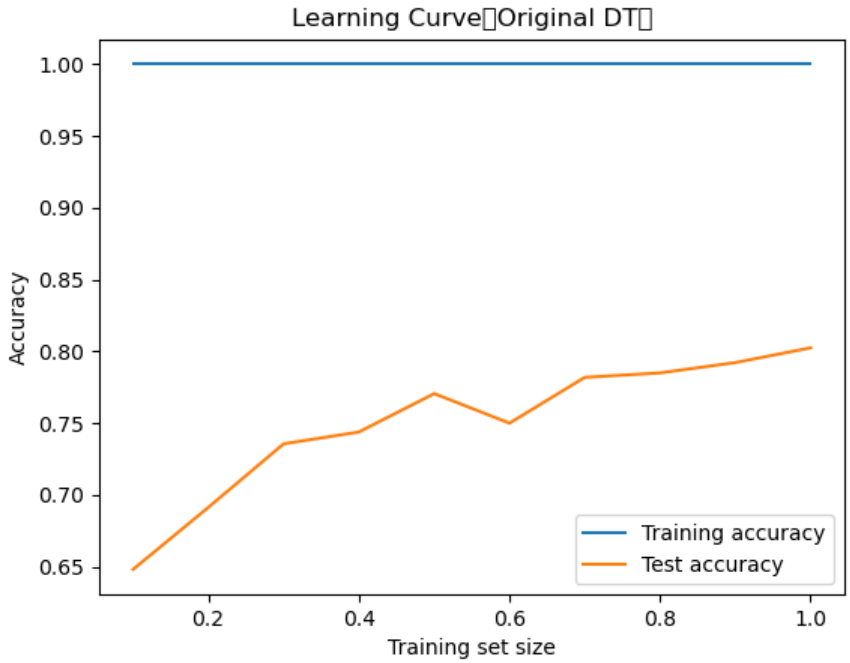


3.2.1.4 Plotting the Original Model Learning Curve

Most academic findings suggest that decision tree models are prone to overfitting issues. To make the overfitting problem more intuitive, I plotted the learning curve of the original decision tree model, as shown in Fig 10.

According to the learning curve, the accuracy of the training set remains at 1, which is significantly higher than the accuracy of the test set (around 0.75 to 0.8). Furthermore, as the amount of data increases, this gap does not decrease when the training set reaches 50% of the data; instead, it even increases. This indicates that the model is overfitting and needs optimization. In response to the common overfitting issue in decision tree models, the academic community has proposed various solutions. Here, I have chosen to use decision tree pruning methods to optimize the decision tree model.

Fig 10 Learning Curve(Original DT)



3.2.2.2 Model Optimization (Decision Tree Pruning)

For the pruning of decision tree models, there are two methods that can be used: one is the cost complexity method, and the other is grid search for parameters.

3.2.2.2.1 Cost Complexity Pruning

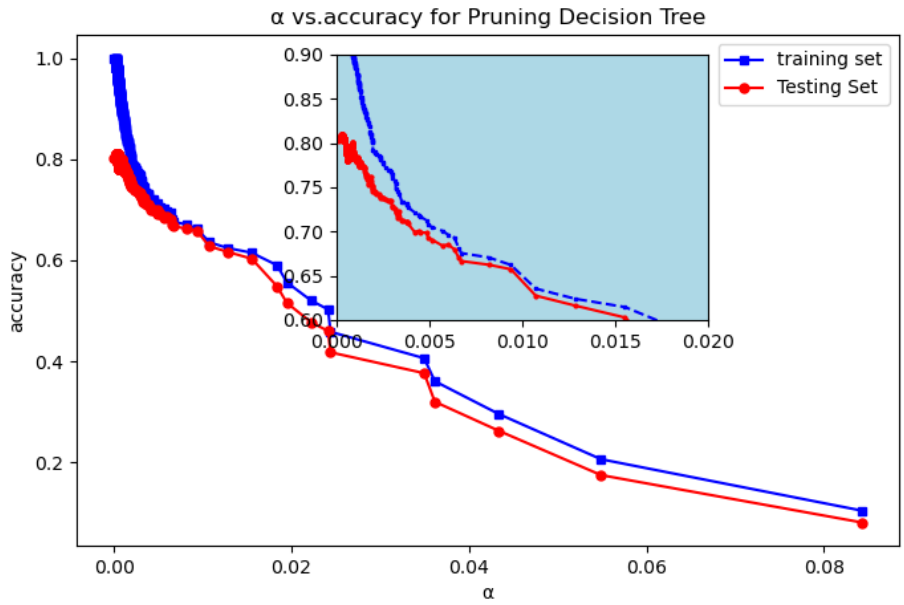
The cost complexity pruning method involves analyzing the complexity of the default decision tree model using the training set first. Then, the optimal decision tree model is established based on the value of the optimal complexity parameter (ccp-alpha), and its prediction accuracy on the training and test sets is calculated.

Table 8 CCP-Alpha Grid Search Results

	ccp_alpha_dtc2	train_acc_dtc2	test_acc_dtc2
16	0.000251	0.994856	0.8107
20	0.000257	0.99357	0.8107
19	0.000254	0.994342	0.8107
18	0.000254	0.994599	0.8107
17	0.000251	0.994856	0.8107
66	0.000365	0.980967	0.809671
15	0.000251	0.99537	0.809671
14	0.00025	0.995885	0.809671

Table 8 displays the top eight optimal results obtained from the algorithm. It can be observed that when ccp-alpha is equal to 0.000251, the pruned decision tree model achieves the highest accuracy on the test set.

Fig 11 α vs.accuracy



To visualize the changes in the accuracy of the decision tree regression model on the training and test sets under different constraints of the ccp-alpha parameter, I created a correlation plot between model accuracy and ccp-alpha. The results are shown in Fig 11.

According to the plot, it can be observed that the model accuracy decreases with the increase of model complexity, both on the training and test sets. I ultimately used the ccp-alpha value obtained from the grid search, which allows for the use of a simpler model while ensuring model accuracy.

Table 9 Optimal ccp_alpha Accuracy

Accuracy on the training dataset of Pruned DT1/Cost-Complexity	0.995627572
Accuracy on the testing dataset of Pruned DT1/Cost-Complexity	0.809670782

After training the model with the optimal ccp-alpha value, the accuracy results are shown in

Table 9 The accuracy of the model on the training set has decreased to some extent, alleviating the issue of overfitting.

To visualize the training results, Fig 12 and Fig 13 show the confusion matrices for the training and test sets, respectively, under the optimal ccp-alpha condition.

Fig 12 Confusion matrix of training set for Pruned DT1/Cost-Complexity

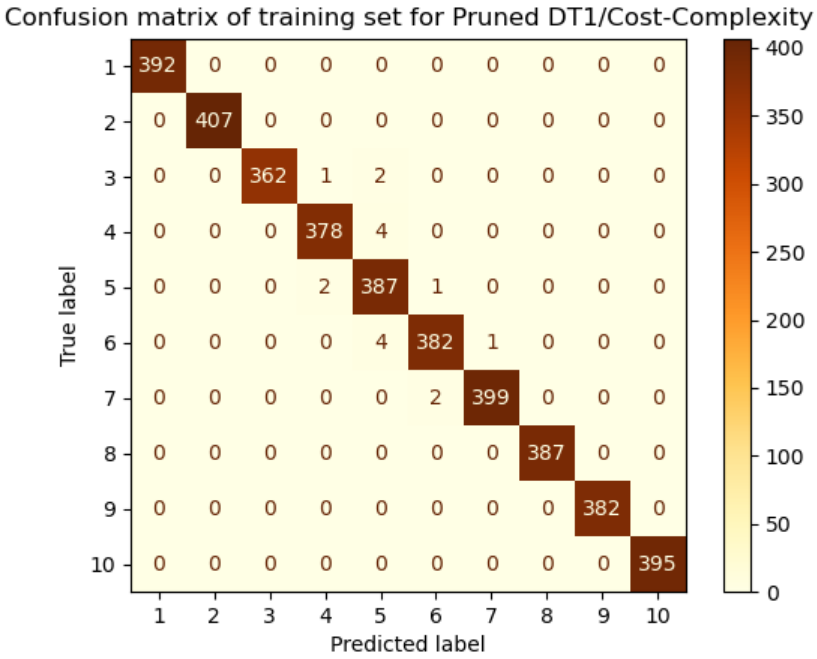
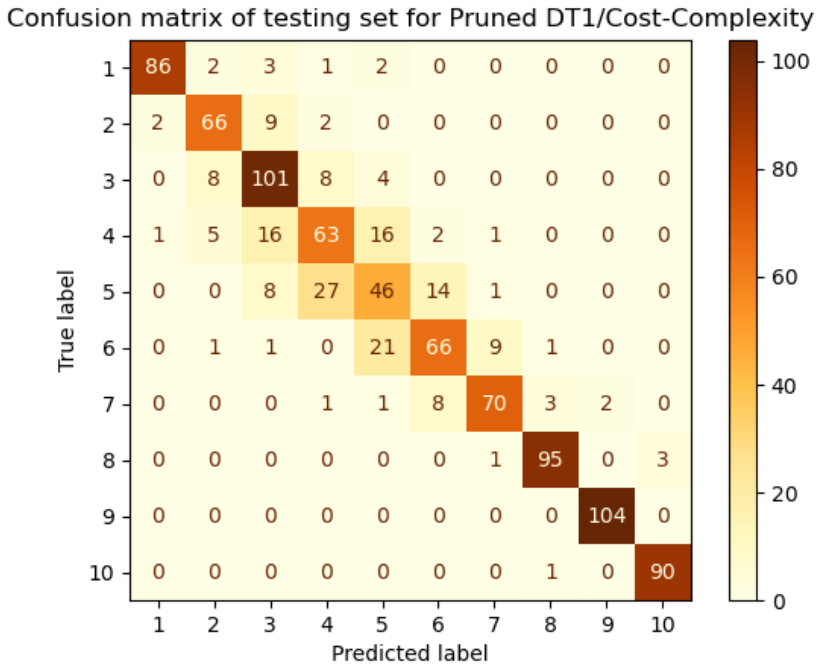


Fig 13 Confusion matrix of testing set for Pruned DT1/Cost-Complexity



Based on the results of the test set confusion matrix, it is evident that the model still has some potential for overfitting. Therefore, I decided to visualize the decision tree to determine if the model is still too complex. After visualization, the decision tree remains quite complex, which I believe is the cause of the overfitting on the training set. Consequently, I decided to use a cross-validation grid search method to reduce complexity by lowering the maximum depth to obtain the best model.

3.2.2.2.2 Parameter Grid Search Method

Table 10 Results of the Grid Search

	tree_depth_dtc3	tree_leafnode_dtc3	test_acc_dtc3
218	11	28	0.702675
278	14	28	0.702675
238	12	28	0.702675
158	8	28	0.702675
198	10	28	0.702675
178	9	28	0.702675
258	13	28	0.702675
179	9	29	0.701646
217	11	27	0.701646

To limit the maximum complexity, I have set the maximum depth to search for as 15 and the maximum number of leaf nodes as 30. The results of the grid search are presented in Table 10. From the output results, it can be observed that the optimal parameter combination consists of a maximum depth of 11 and a maximum number of leaf nodes of 28. Since a depth exceeding 15 increases the complexity of the model without a significant improvement in results, I have chosen a depth of 8. This choice ensures that the decision tree is not too complex while greatly improving the accuracy of the model. Using this parameter set (max_depth=8, max_leaf_nodes=28), a new pruned decision tree model is constructed, and the prediction accuracy on both the training and test datasets is presented.

Based on the results of the grid search, the accuracy of the optimal decision tree model

established is shown in Table 11.

Table 11 Decision Tree Accuracy from Grid Search

Accuracy on the training dataset for Pruned DT2/Grid Search	0.719135802
Accuracy on the test dataset for Pruned DT2/Grid Search	0.702674897

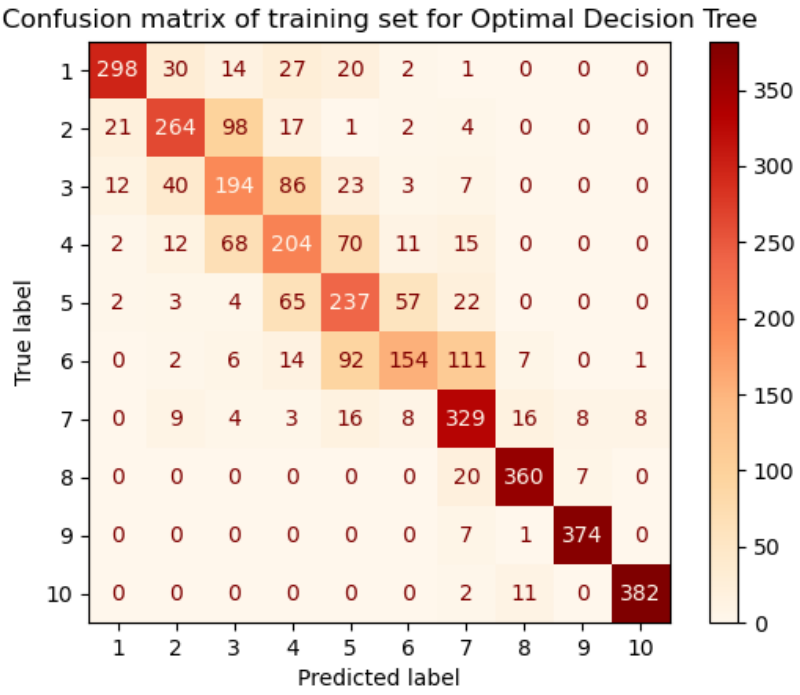
The result illustrates that the model's accuracy on the training and test sets is quite similar, with both exceeding 70%. This demonstrates that the pruning method effectively addresses the issue of overfitting. Consequently, the optimal decision tree model selected will be the one obtained through grid search with the maximum depth.

3.2.3 Model Assessment

Based on the previous model establishment, the optimal decision tree model is determined to be the one obtained from grid search with the maximum depth, as shown in Table 11

3.2.3.1 Confusion Matrix

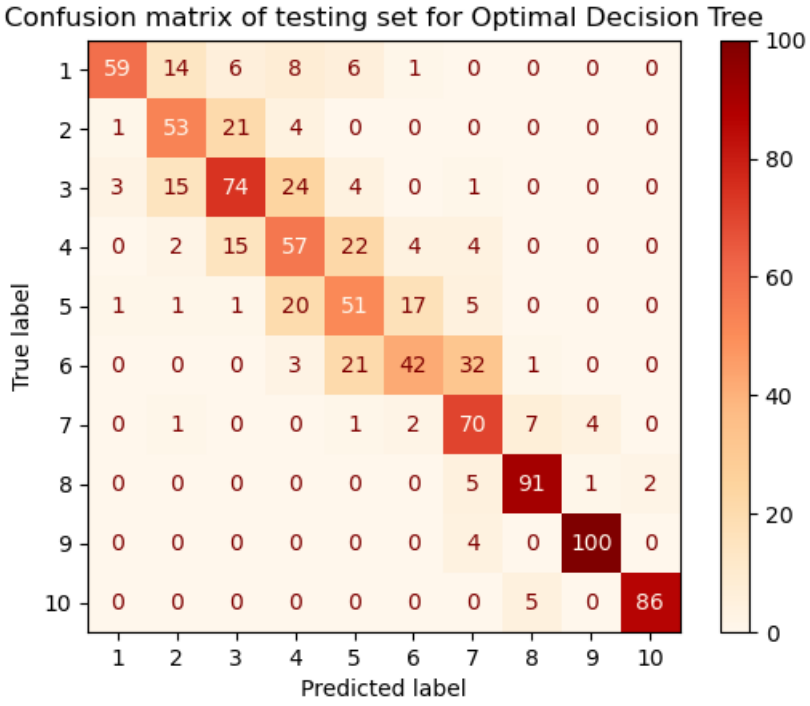
Fig 14 Confusion matrix of training set for Optimal Decision Tree



To render the model's predictive outcomes more intuitive, I have chosen to visualize the confusion matrix. Fig 14 and Fig 15 present the confusion matrices for the optimal decision tree model on the training and test sets.From the confusion matrix images, it can be observed that

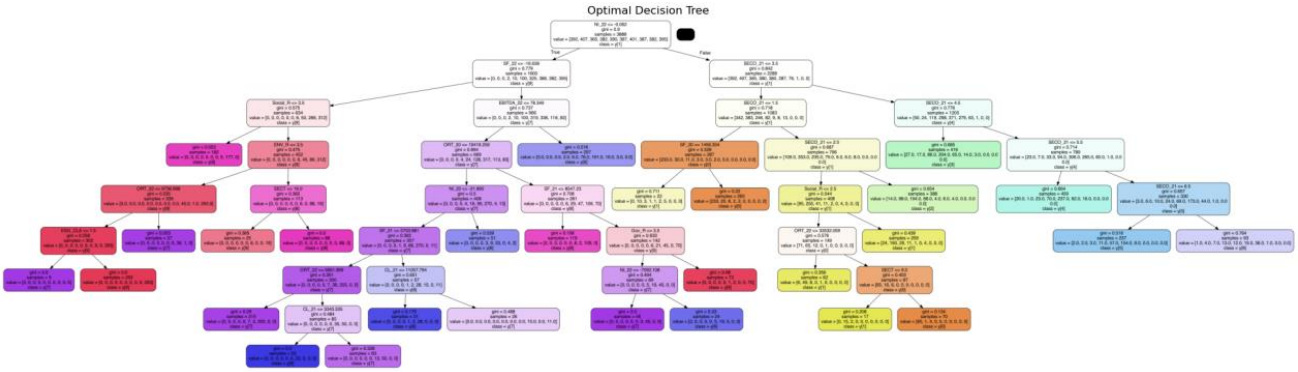
the model maintains a high accuracy rate on both the training and test sets. By reducing the complexity of the model while ensuring a high level of precision, the issue of overfitting has been successfully avoided.

Fig 15 Confusion matrix of testing set for Optimal Decision Tree



3.2.3.3 Visualizing the Optimal Decision Tree

Fig 16 Optimal Decision Tree

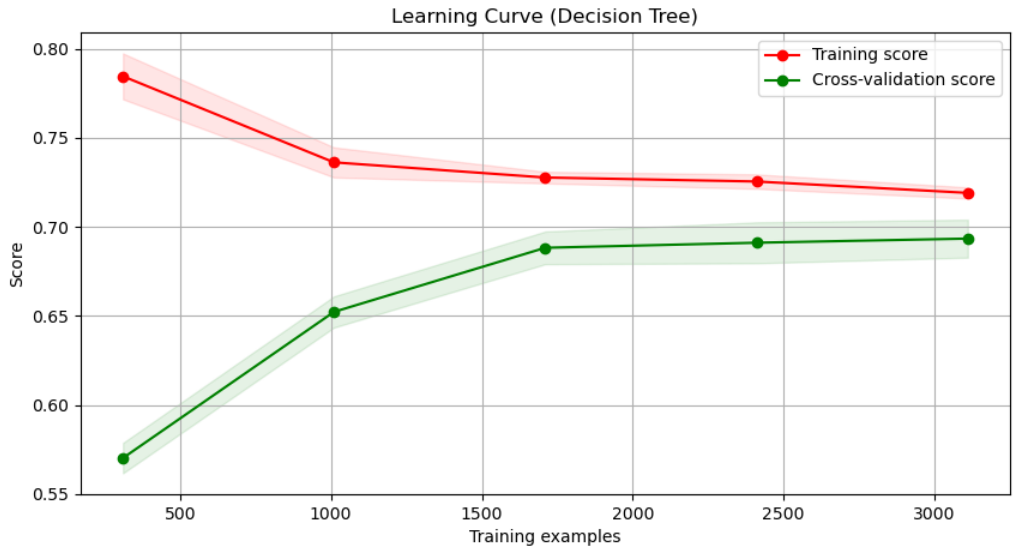


To further assess whether the model remains overly complex, I visualized the tree structure of the optimal model, as shown in Fig 16. It can be noted that after reducing the maximum depth, the model has been significantly simplified. The ability to maintain high precision with a simpler model demonstrates the substantial success of this model optimization.

3.2.3.4 Learning Curve

To verify whether overfitting still persists, I also generated learning curves for the model. As illustrated in Fig 17, as the number of training samples increases, the accuracy during training and testing gradually converges and stabilizes. This indicates that the method successfully addresses the issue of overfitting.

Fig 17 Learning Curve (Decision Tree)



The decision tree model is a crucial component in multi-classification strategies and serves as a fundamental learning model. However, due to its nature of having only a single tree, the conclusions drawn from it may have significant limitations. I intend to explore and compare other machine learning models with higher generalization capabilities for analysis.

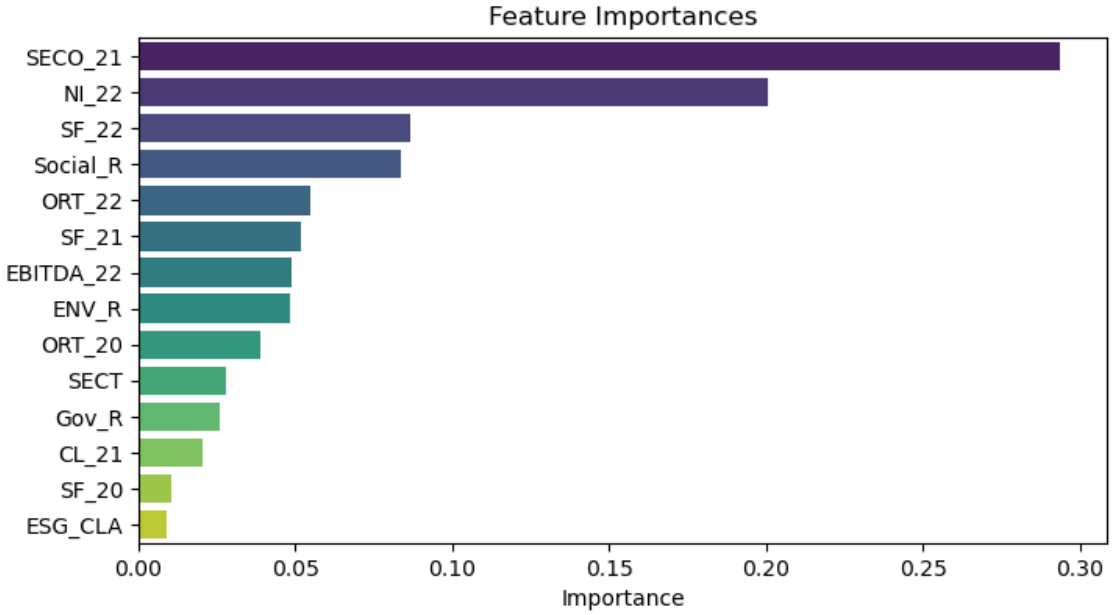
3.2.3.6 Feature Importance

Similarly, the decision tree model is also capable of outputting feature importance.

When training a model using a decision tree, the model automatically computes the importance score for each feature and stores it in the feature_importances_ attribute.

Consequently, I have generated a bar chart depicting the feature importance of the model.

Fig 18 Feature Importances（DT）



I have also listed the explanatory proportions of each feature in Table 12 to facilitate the interpretation of subtle differences.

Table 12 Feature Importances（DT）

rank	Feature	Importance	rank	Feature	Importance
1	SECO_21	0. 2939	8	ENV_R	0. 0482
2	NI_22	0. 2004	9	ORT_20	0. 0386
3	SF_22	0. 0866	10	SECT	0. 0277
4	Social_R	0. 0838	11	Gov_R	0. 0257
5	ORT_22	0. 0549	12	CL_21	0. 0202
6	SF_21	0. 0519	13	SF_20	0. 0103
7	EBITDA_22	0. 0487	14	ESG_CLA	0. 009

In the process of creating the charts, I removed features with an explanatory value of 0 for ease of presentation. The results indicate that the most influential feature for the decision tree model is the rating from the previous year (2021), accounting for nearly 30% of the entire model's impact. Additionally, other significant features include the Net Income and Shareholders' Funds from 2022. Moreover, four out of the top seven features in terms of importance are financial indicators from 2022. Social ratings also hold a considerable weight. This suggests the reference object for the decision tree model, which is to predict based on the 2021 rating, combined with

the 2022 financial indicators and social ratings. These features are also displayed in the visualized decision tree in Fig 16.

3.3 Random Forest (RF)

The Random Forest model is an efficient machine learning method applied to multi-classification models and can be regarded as a classifier that establishes multiple decision trees. The Random Forest algorithm is simple and easy to implement, capable of producing high-precision conclusions for various types of data, while also being able to quickly process a large number of input variables. Additionally, as the number of data features increases, the stability of the Random Forest model also improves. Given that my data possesses a considerable number of analyzable features, it is highly suitable for the Random Forest model.

3.3.1 Determining the optimal parameters.

According to previous academic findings, the diversity of the Random Forest model is determined by the maximum depth and the number of trees. Therefore, to obtain the optimal model, one only needs to select the best parameters through a parameter search method. Consequently, I employed a grid search method to find the optimal parameter combination and directly used the best data for modeling.

3.3.2 Model Assessment

3.3.2.1 Optimal Model

The optimal maximum depth obtained through grid search was found to be 13, and the optimal number of trees was 600.

Table 13 Accuracy of the model with the optimal parameters.

Accuracy on the training set of rf	0.997427984
Accuracy on the test set of rf	0.881687243

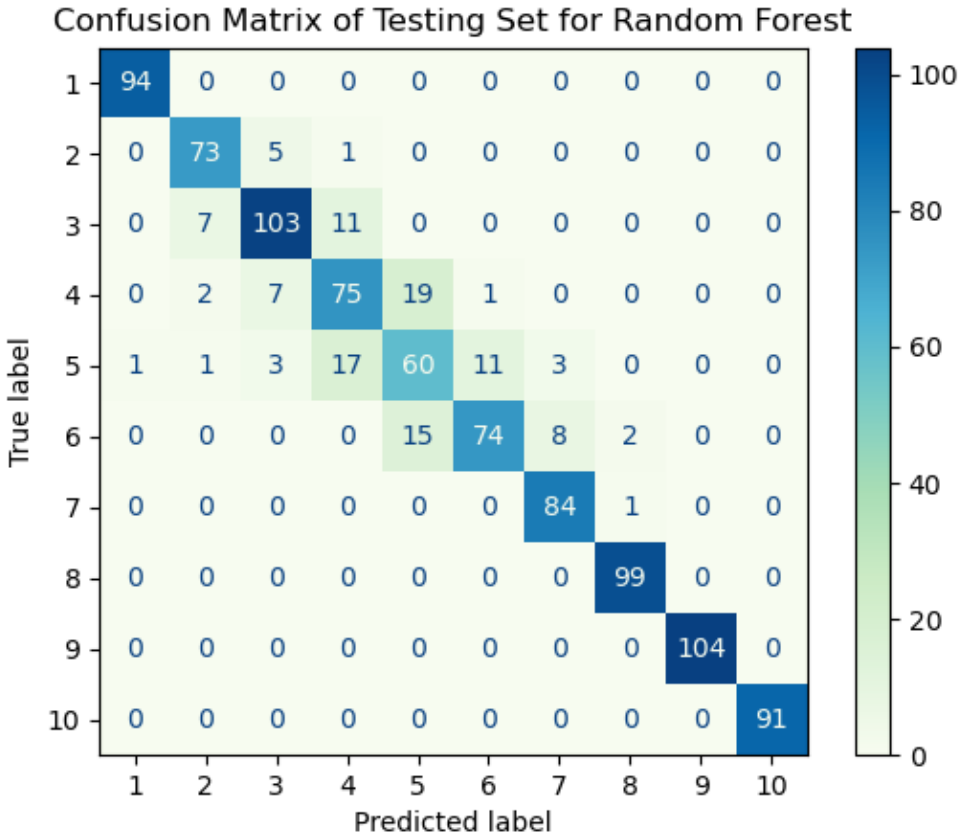
The accuracy of the model built using the results of the grid search is presented in Table 13.

Based on the optimal accuracy, it can be inferred that the model's training results are of high precision.

3.3.2.2 Confusion Matrix

Fig 19 Confusion Matrix of Training Set for Random Forest

Fig 20 Confusion Matrix of Testing Set for Random Forest



Similarly, I have outputted a confusion matrix to assess the model's accuracy,as shown in Fig 19 and Fig 20.

As depicted in Fig 19 and Fig 20, the model exhibits significant accuracy on both the training and testing levels, with a small discrepancy, indicating a strong generalization capability.

3.3.2.3 Out-of-Bag (OOB) Error

In addition to predictive accuracy, there is another type of error called Out-of-Bag (OOB) Error, which is used to measure the effectiveness of the model.

In random forest classifiers, the OOB error is a method used to measure the model's generalization ability. During the training process of random forests, due to the use of bootstrap sampling, a portion of the data is not selected for each tree constructed. This unselected data is referred to as "Out-of-Bag" (OOB) data. For each tree, the corresponding OOB data can be used

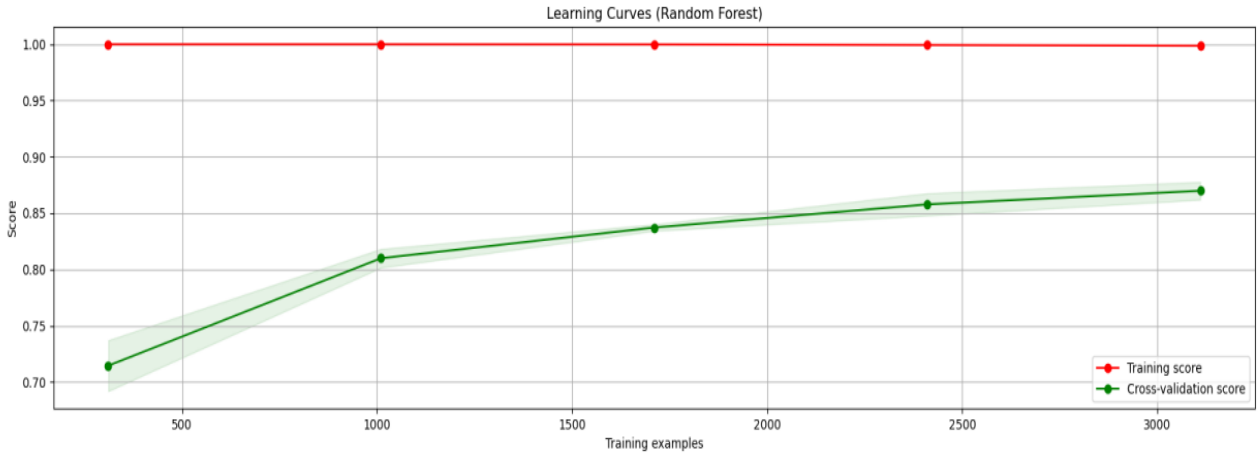
to estimate the model's error, which involves calculating the misclassification rate of this OOB data on that particular tree. By averaging the OOB errors across all trees, the overall OOB error for the entire random forest is obtained. The advantage of the OOB error is that it does not require a separate validation or test set to estimate the model's error; it can be directly evaluated during the training process. This makes the OOB error an effective and efficient method for model evaluation, especially when the dataset is limited.

The calculated OOB error value is 0.12217078189300412, which demonstrates that the model I designed still possesses excellent predictive effectiveness.

3.3.2.4 Learning Curve

To assess whether the model is at risk of overfitting, I continued to construct a learning curve. As observed in Fig 21, with the increase in the number of models, the accuracy of the training set gradually stabilizes at a higher level, which aligns with the characteristics of the Random Forest model and indicates a satisfactory performance.

Fig 21 Learning Curves (Random Forest)



3.3.3.5 Feature Importance

The feature importance of the Random Forest model is highly explanatory for the Random Forest model. In the implementation of the Random Forest algorithm, feature importance scores

are calculated and stored internally within the algorithm. When using the RandomForestClassifier from the scikit-learn library, the model automatically computes the importance of each feature during the training process.

I first outputted the Random Forest feature importance list, shown in Table 14.

Table 14 Feature importance（RF）

rank	Feature	Importance	rank	Feature	Importance
1	SECO_21	9.3298%	17	ORT_22	2.4259%
2	NI_22	8.8794%	18	ORT_20	2.4072%
3	EBIT_22	6.9641%	19	SF_20	2.3736%
4	SF_22	5.7728%	20	CA_20	2.3707%
5	EBITDA_22	5.7071%	21	ORT_21	2.2394%
6	SECO_20	3.9856%	22	TA_21	2.1384%
7	NI_21	3.6740%	23	EBIT_20	2.0808%
8	SF_21	3.1452%	24	TA_20	2.0513%
9	EBIT_21	3.1045%	25	EBITDA_21	1.8940%
10	CA_22	3.0827%	26	REG	1.8147%
11	CL_22	3.0238%	27	EBITDA_20	1.7434%
12	CA_21	2.6616%	28	Social_R	1.6528%
13	TA_22	2.5560%	29	Gov_R	1.6369%
14	NI_20	2.5350%	30	SECT	1.6004%
15	CL_20	2.4945%	31	ENV_R	1.4331%
16	CL_21	2.4890%	32	ESG_CLA	0.7325%

According to the feature importance results, it can be observed that the most significant feature affecting the model is the credit rating result from 2021. Following this, the features with a higher proportion are the financial data from 2022, particularly net income, EBIT, shareholders' funds, and EBITDA, followed by the credit rating from 2020.

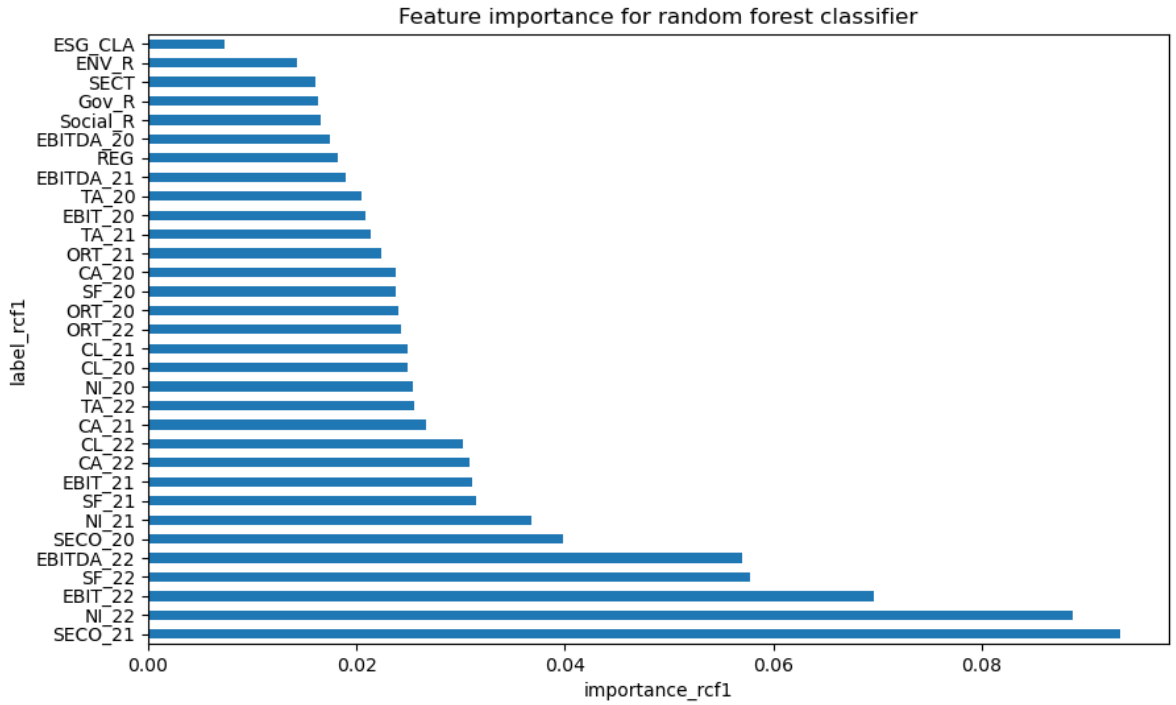
From the chart analysis, it can be inferred that net income, shareholders' funds, and EBIT are important financial indicators for referencing credit rating results, regardless of whether it is for 2022 or 2021. Additionally, the rating results from the previous two years serve as important reference indicators for the current year.

The feature with the least impact is the ESG Comprehensive Classification. Whether it is

environmental, social, or governance ratings, they seem to have little influence on the Random Forest model.

To make the conclusions more intuitive, I have also created a feature importance graph, as shown in Fig 22.

Fig 22 Feature importance（RF）



The Random Forest model has demonstrated a good performance. I will continue to explore other machine learning models in pursuit of more optimized model construction results.

3.4 Gradient Boosting Machine (GBM)

Gradient Boosting Machine (GBM) is a type of ensemble learning algorithm based on the boosting technique. It constructs models iteratively to minimize a loss function, with each tree attempting to correct the errors of the previous tree. GBM is built sequentially, with each tree depending on the previous one.

3.4.1 Building the premier Model

Using the model from the sklearn toolkit to construct the original model, the accuracy obtained

is presented in Table 15.

Table 15 Premier GBM Accuracy

Accuracy on the training set of pre-GBM	0.968878601
Accuracy on the test set of pre-GBM	0.836419753

Based on the accuracy, it can be inferred that the constructed original model is already sufficiently precise, with the training accuracy of the test set exceeding 80%.

3.4.2 Model Optimization

I attempted to use cross-validated grid search to select the optimal n_estimators, learning_rate, and max_depth, but the computation time was excessively long. The results from training the original model were also quite impressive, so I decided to choose the basic Gradient Boosting Machine that I had set up as the final model.

3.4.3 Model Assessment

3.4.3.1 Optimal Model Accuracy

The optimal accuracy of the Gradient Boosting Machine model I constructed is presented in Table 16.

3.4.3.2 Confusion Matrix

The optimal accuracy of the Gradient Boosting Machine model I constructed is presented in Table 16 To visualize the model's predictive results, I have also created confusion matrices for both the training and test sets, as shown in Fig 23 and Fig 24.

Fig 23 Confusion Matrix of Training Set for gbm

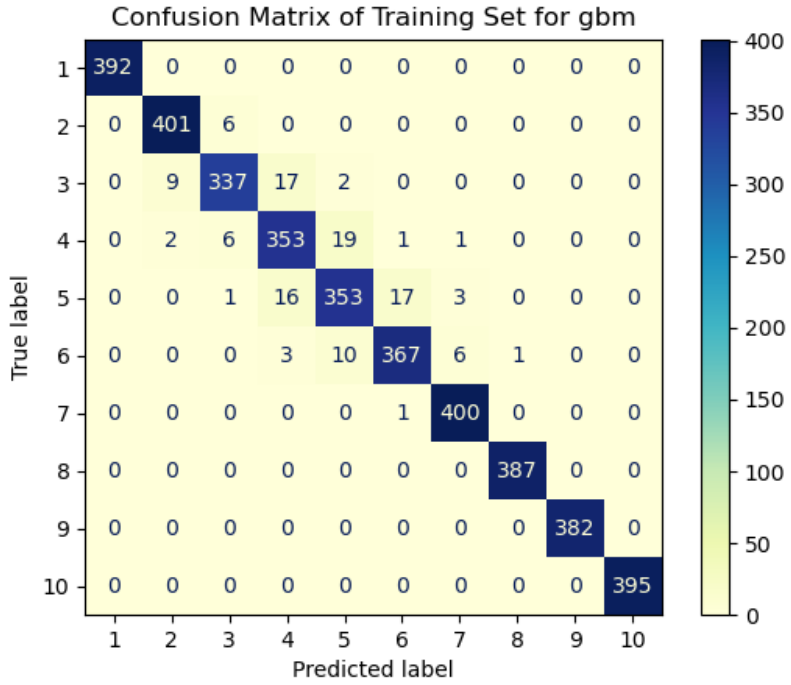
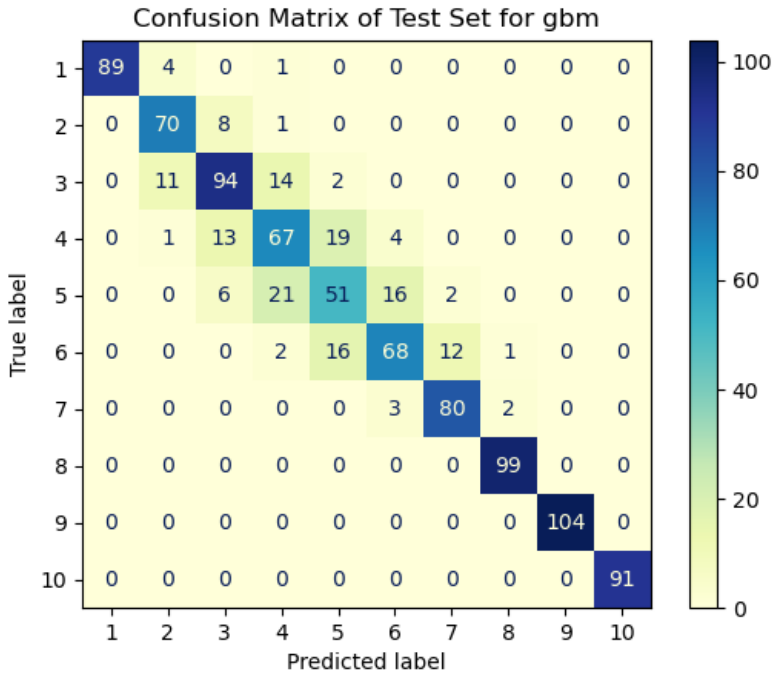


Fig 24 Confusion Matrix of Test Set for gbm

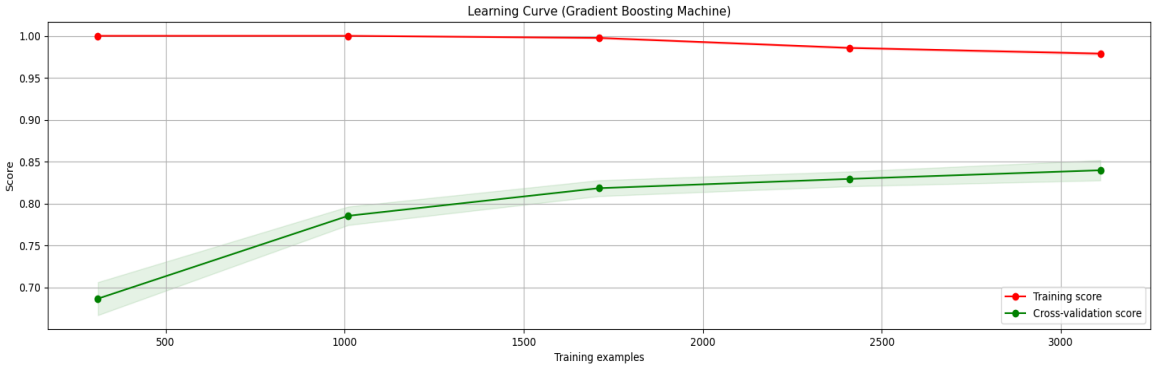


Based on the confusion matrices, it can be observed that the model exhibits high accuracy on both the training and test sets. The easily confused analysis method is limited to the vicinity of the target, with no significant errors, demonstrating that the model has strong generalization capabilities.

3.4.3.3 Learning Curve

Similarly, to verify the absence of overfitting, I continue to introduce a learning curve (Fig 25) for analysis.

Fig 25 Learning Curve (Gradient Boosting Machine)



According to the learning curve, it can be observed that once the number of training samples reaches 1500, the model's accuracy tends to stabilize, and the gap between the training and test sets gradually narrows, indicating that the model does not exhibit significant overfitting phenomena.

3.4.3.4 Feature Importance

Following the previous approach, Table 16 outputs the feature importance ranking list, and Fig 26 provides a more intuitive bar chart.

During training, Gradient Boosting Machines (GBM) calculate the importance scores of features based on their contribution to the model. These scores are determined after the model training is complete, through internal algorithms and logic within the model.

Based on the analysis of the charts, it can be concluded that the rating result from 2021 remains a crucial explanatory factor, with an explanatory degree of over 25%, indicating that the Random Forest model has given significant consideration to the rating result from the previous year. Following this, the relatively important factors are the financial data from 2022, especially

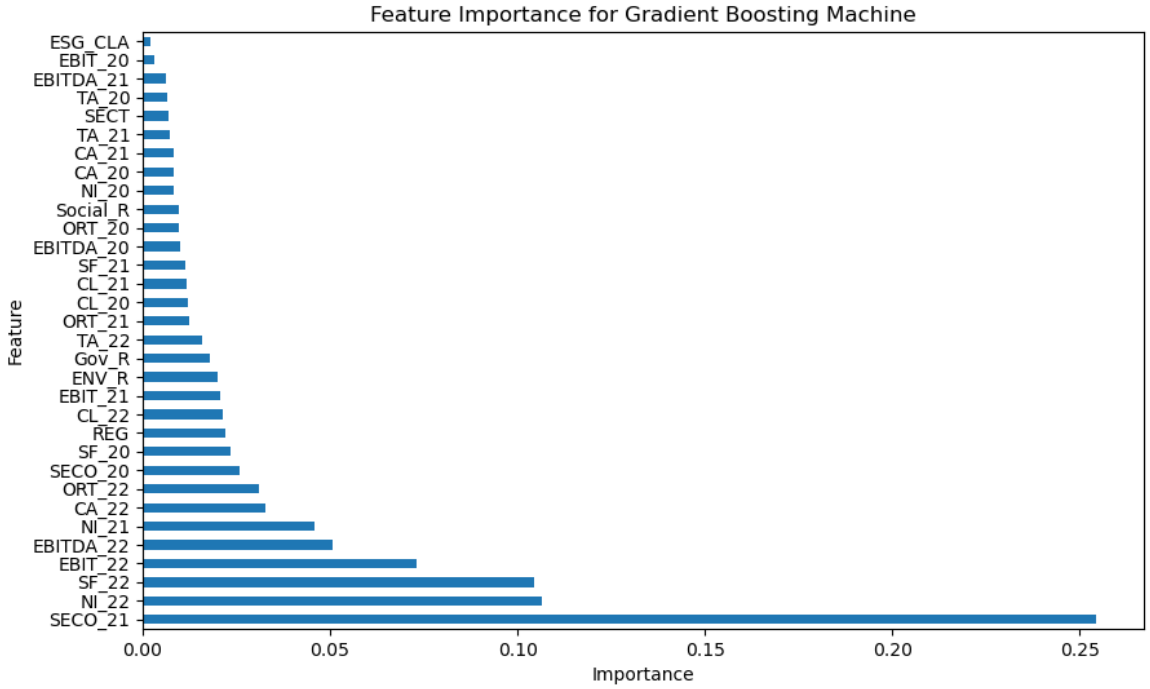
net income, shareholders' funds, EBIT, and EBITDA.

The financial indicators from 2022 are concentrated in the top 16 important rankings, which proves that the current year's financial indicators are significant factors affecting the rating.

Table 16 Feature Importance（GBM）

rank	Feature	Importance	rank	Feature	Importance
1	SECO_21	25.4341%	17	ORT_21	1.2538%
2	NI_22	10.6475%	18	CL_20	1.2230%
3	SF_22	10.4575%	19	CL_21	1.1643%
4	EBIT_22	7.2910%	20	SF_21	1.1283%
5	EBITDA_22	5.0785%	21	EBITDA_20	1.0016%
6	NI_21	4.5915%	22	ORT_20	0.9637%
7	CA_22	3.2823%	23	Social_R	0.9620%
8	ORT_22	3.1045%	24	NI_20	0.8366%
9	SECO_20	2.5746%	25	CA_20	0.8364%
10	SF_20	2.3600%	26	CA_21	0.8112%
11	REG	2.2173%	27	TA_21	0.7164%
12	CL_22	2.1232%	28	SECT	0.6927%
13	EBIT_21	2.0663%	29	TA_20	0.6633%
14	ENV_R	2.0043%	30	EBITDA_21	0.6105%
15	Gov_R	1.7886%	31	EBIT_20	0.3264%
16	TA_22	1.5895%	32	ESG_CLA	0.1991%

Fig 26 Feature Importance（GBM）



Overall, the model is most influenced by the credit rating result from 2021 (the previous year), and the least affected by the ESG class. The financial data from 2022 and the credit rating result from 2020 also serve as important references for analysis.

3.5 Support Vector Machine (SVM)

The Support Vector Machine is a widely used deep learning algorithm, a type of supervised learning model commonly employed for data classification and regression problems. According to existing research, the SVM model is less prone to overfitting and is a very practical black-box model.

3.5.1 Building the premier Model

Firstly, I used the linear kernel function as the model for the Support Vector Machine calculation method. The classification report and accuracy obtained are presented in Table 17.

The output of this model is highly dependent on the generated data, but the predictions of the original model were poor, as was the accuracy, indicating significant issues with the model.

Table 17 premier SVM Classification Report

	precision	**recall**	**f1-score**	**support**
1	0.83	0.91	0.87	94
2	0.62	0.7	0.66	79
3	0.62	0.54	0.58	121
4	0.5	0.53	0.52	104
5	0.5	0.46	0.48	96
6	0.46	0.38	0.42	99
7	0.53	0.54	0.53	85
8	0.83	0.89	0.86	99
9	0.96	0.95	0.96	104
10	0.89	0.98	0.93	91
accuracy			0.68	972
macro avg	0.68	0.69	0.68	972
weighted avg	0.68	0.68	0.68	972
Accuracy	0.684156379			

Accuracy: The overall accuracy is 0.684, meaning the model correctly classifies about 68.4% of all samples. Whether this accuracy is acceptable or needs improvement depends on your

specific application and expectations.

Macro Avg: The precision, recall, and F1-Score of the macro average are all around 0.68, indicating that the model's average performance across different classes is relatively consistent.

Weighted Avg: The weighted average metrics are similar to the macro average, taking into account the differences in sample sizes across classes. This metric better reflects the overall performance.

3.5.2 Model Optimization

3.5.2.1 Randomized Search Cross-Validation

This approach involves randomly sampling a specified number of parameter combinations from a given distribution for the search.

The results show that the model's overall accuracy is 0.80, indicating that the model correctly predicted 80% of all predictions. This is quite a good accuracy, especially for a multi-class classification problem.

The best parameters used by the model include the Radial Basis Function (RBF) kernel, an automatically calculated gamma value, a polynomial degree of 2, coef0 of 0.5, and a C value of 100. This indicates that the model chose the RBF kernel and adjusted the regularization parameter C appropriately.

Therefore, I decided to use this model as the optimal SVM model.

3.5.3 Model Assessment

3.5.3.1 Optimal Model and Accuracy

Based on the conclusions from the randomized search, the best accuracy obtained on both the training and test sets is presented in Table 18

Table 18 Accuracy for optimal SVM

Accuracy on the training set for optimal Support Vector Machine	0. 91255144
Accuracy on the test set for optimal Support Vector Machine	0. 797325103

The model resulting from the randomized search is of high quality, exhibiting good performance on both the training and test sets.

3.5.3.2 Confusion Matrix

Fig 27 Confusion Matrix of Training Set for Optimal SVM

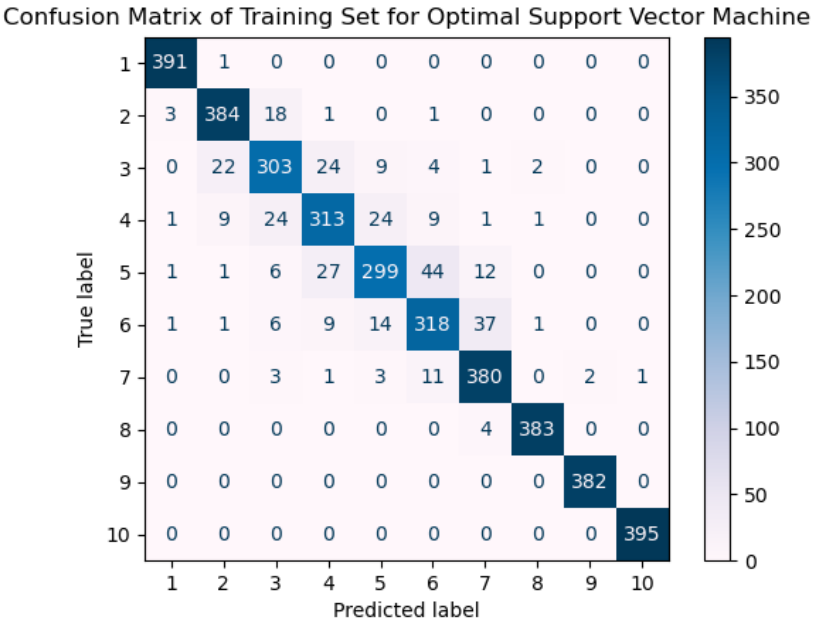
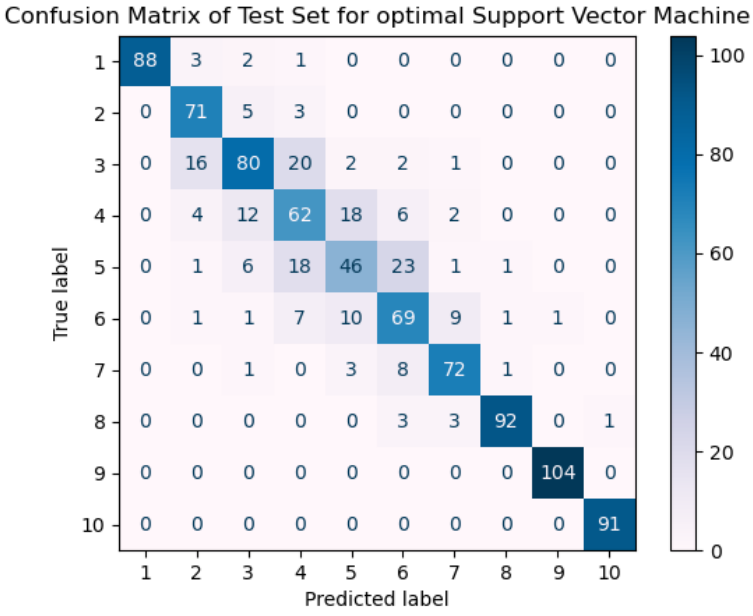


Fig 28 Confusion Matrix of Test Set for optimal SVM



Similarly, to make the model training results more intuitive, I have outputted the confusion

matrices for the training and test sets (Fig 27 and Fig 28).

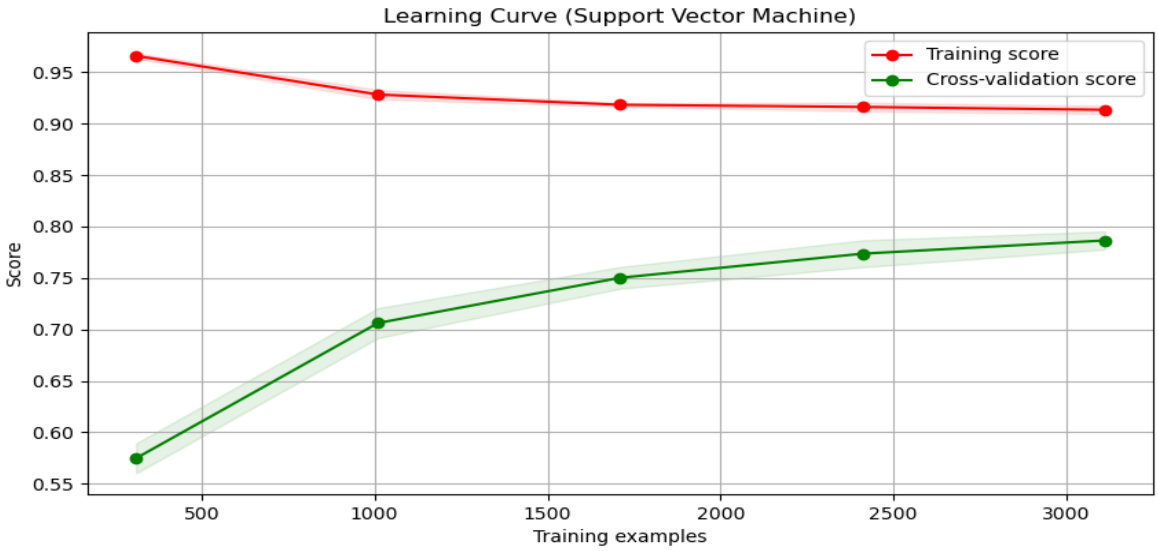
According to the confusion matrices, it can also be observed that the model's training results are good, and its performance on real data is also satisfactory. However, there are still some issues, such as the model performing better in training than in the test set. Categories 5 and 4 are easily confused, indicating that the model's performance on data without additional oversampling is not optimal.

Nevertheless, the model still has a relatively high number of correct predictions, suggesting that even after using oversampled training samples, the model can successfully predict most real data.

3.5.3.3 Learning Curve

Although the Support Vector Machine model is less prone to overfitting, I still plotted a learning curve in the Fig 29 It was found that the accuracy of the model on both the training and test sets quickly stabilized and the gap gradually narrowed, indicating no significant overfitting issues.

Fig 29 Learning Curve (Support Vector Machine)



3.5.3.5 Feature Importance

Feature importance can only be output when the kernel is set to 'linear'. The optimal model

obtained from cross-validation and random search uses the RBF function as the kernel, therefore, it is not possible to generate a feature importance plot.

Overall, the Support Vector Machine model is an excellent choice with high accuracy and less prone to overfitting. There are many new algorithms in black-box models, and I will attempt other black-box models next to test their performance.

3.6 Artificial Neural Network (ANN)

The Artificial Neural Network algorithm is a biomimetic neural network model in the field of machine learning and cognitive science, typically using fully connected neural networks, MLP, or known as the Multi-Layer Perceptron algorithm. It is an artificial neural network structure with a relatively simple connection method.

The neural network consists of an input layer, hidden layers, and an output layer. The factors affecting LP performance mainly include the activation function and network structure. The network structure covers the number of layers of neurons in the model, the number of neurons in each layer, and their connection methods. The network structure mainly consists of input layers, hidden layers, and output layers. The input layer is used to receive external signals without functional processing, while the hidden layers and output layer process the input signals and output the results through the output layer.

The differentiation of neural network learning capabilities also comes from the network structure they use. According to the number of hidden layers in the network structure, the model can be divided into two types: single-hidden-layer neural networks and multi-hidden-layer neural networks.

Next, I will construct single-hidden-layer neural network and multi-hidden-layer neural

network models separately to compare their effects.

3.6.1 Single Hidden Layer MLP Classifier

The single-hidden-layer neural network has only one hidden layer and is established using the fully connected neural network MLP in the SKLearn library.

3.6.1.1 Building the premier Model

The accuracy of the original model output by the single-hidden-layer neural network I built is presented, along with the confusion matrices on the test set and training set, and the classification report of the model.

Table 19 Single-layer MLP accuracy

Single-layer neural network training set accuracy	0. 36
Single-layer neural network test set accuracy	0. 38

Table 19 shows the accuracy results, and it can be observed that the model performs poorly on both the training and test sets.

Table 20 Single MLP Classification Report

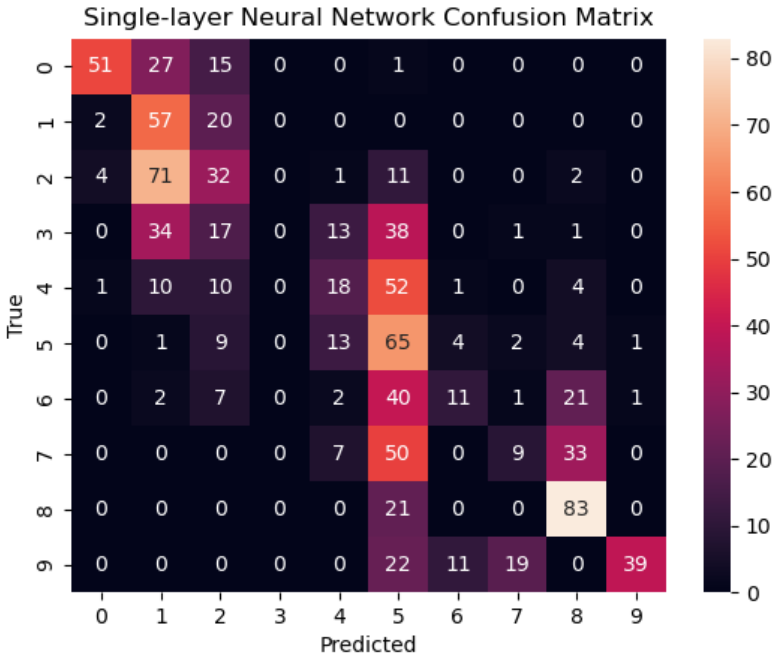
	precision	**recall**	**f1-score**	**support**
1	0.88	0.54	0.67	94
2	0.28	0.72	0.41	79
3	0.29	0.26	0.28	121
4	0	0	0	104
5	0.33	0.19	0.24	96
6	0.22	0.66	0.33	99
7	0.41	0.13	0.2	85
8	0.28	0.09	0.14	99
9	0.56	0.8	0.66	104
10	0.95	0.43	0.59	91
accuracy			0.38	972
macro avg	0.42	0.38	0.35	972
weighted avg	0.41	0.38	0.35	972

Additionally, I have outputted the classification report, and the results presented in Table 20 are very unsatisfactory.

For the BBB (classification code 4) rating, none of the predictions were correct.

To make the prediction results more intuitive, I created a confusion matrix for the test set, as shown in Fig 30.

Fig 30 Single-layer MLP Confusion Matrix（n=50）



The confusion matrix indicates that the model confuses the conclusions of various classifications, and the number of accurate predictions is even less than the number of errors. The result showed that 50 hidden neurons could not correctly predict the result, and too few neurons reduced the accuracy of the model, so I expanded the number of model neurons to 100 for analysis.

3.6.1.2 Model Optimization

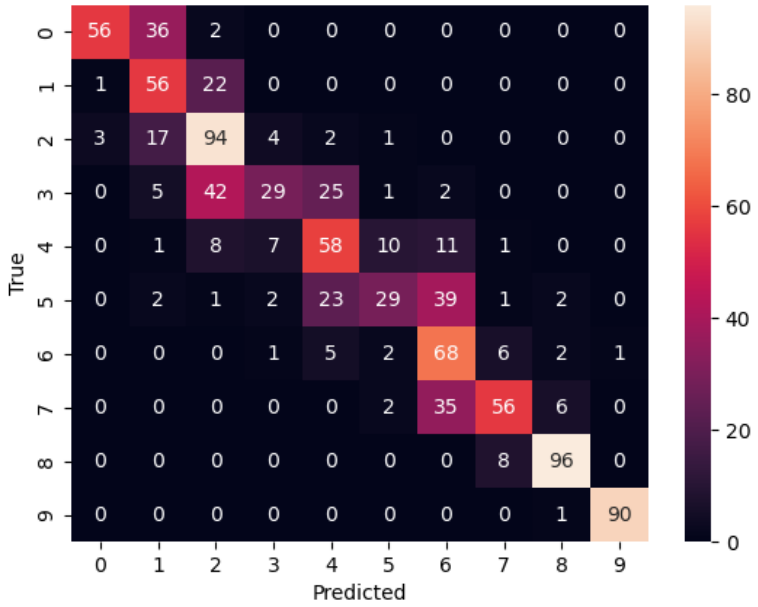
Table 21 Accuracy of ANN（n=100）

Training set accuracy of single-layer neural networks（n=100）	0.7
Testing set accuracy of single-layer neural networks（n=100）	0.65

The accuracy of the model after increasing the number of hidden layer neurons to 100 is shown in Table 21.It can be observed that after increasing the number of neurons, the accuracy of the model has significantly improved. To facilitate the observation of the prediction, I have

outputted the confusion matrix for the test set, as shown in Fig 31.

Fig 31 Confusion Matrix of ANN（n=100）



The improved model has a significantly better correct prediction rate compared to the previous model, with no longer a large number of incorrect predictions. This demonstrates that increasing the number of neurons can effectively improve the prediction results of the single hidden layer MLP classifier.

Afterward, I also attempted to obtain a higher accuracy optimized model through grid search cross-validation. However, due to limitations in machine performance, it took a long time and was unable to compute the results.

Subsequently, I tried the Multi-Hidden Layer MLP Classifier. Similarly, due to performance constraints, the Multi-Hidden Layer MLP Classifier takes a long time and can't produce results within 5 minutes. I decided to temporarily abandon the artificial neural network method and carry out follow-up calculations after the subsequent hardware update.

In search of an efficient and fast computing black-box model for assessment, I will choose a model with simple logic.

3.7 k-Nearest Neighbors (kNN)

The k-Nearest Neighbors algorithm is a widely used machine learning model. Its principle is to have the algorithm select the closest data points from the training set for prediction. The differentiation in the kNN model lies in the number of neighbor data points selected (n_neighbors).

3.7.1 Building the Premier Model

I created an original model with the number of neighbors set to 5. The accuracy of this original model is presented in Table 22.

Table 22 Accuracy pre-knn

Accuracy on the training set for pre-knn	0.857253086
Accuracy on the testing set for pre-knn	0.772633745

It can be observed that the model's prediction results are good, with high accuracy. However, since my data used oversampling methods, it is necessary to examine the model's performance across different classifications.

3.7.2 Confusion Matrix

To visually examine the model's performance across different classifications, I created confusion matrices for the original kNN model on both the training and test sets, as shown in Fig 32 and Fig 33.

Based on the distribution of the confusion matrix, it is evident that although the model achieved a high precision on the test set, its predictive performance on real-world data is unsatisfactory. The model tends to confuse classes BBB and A, indicating poor generalization capabilities in practical scenarios. Therefore, I have decided to optimize the original function to enhance the model's generalization ability.

Fig 32 Confusion matrix of training set for pre-knn

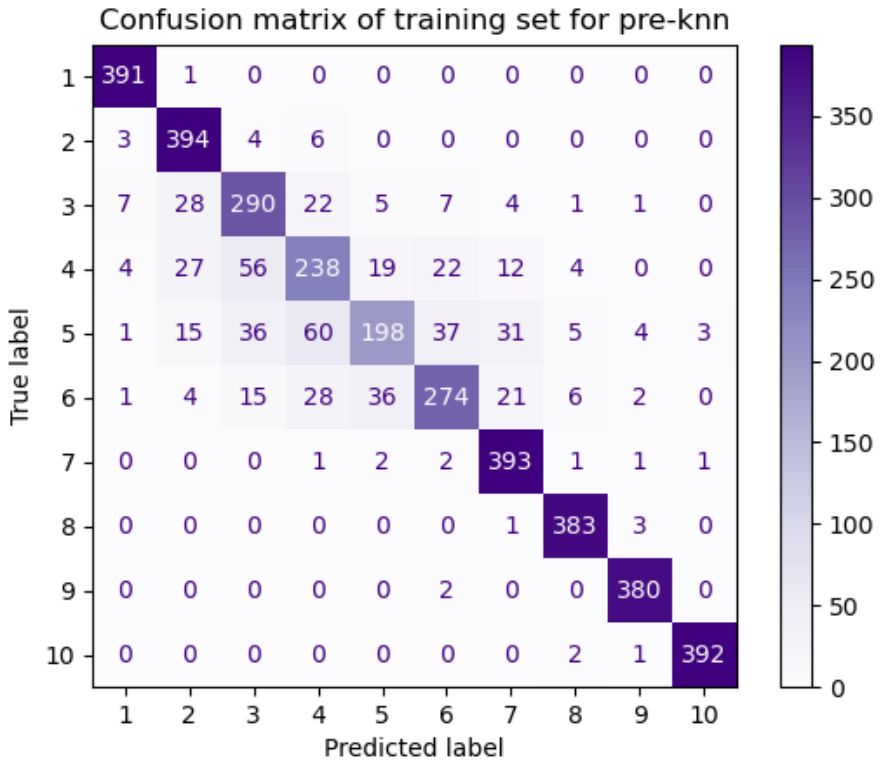
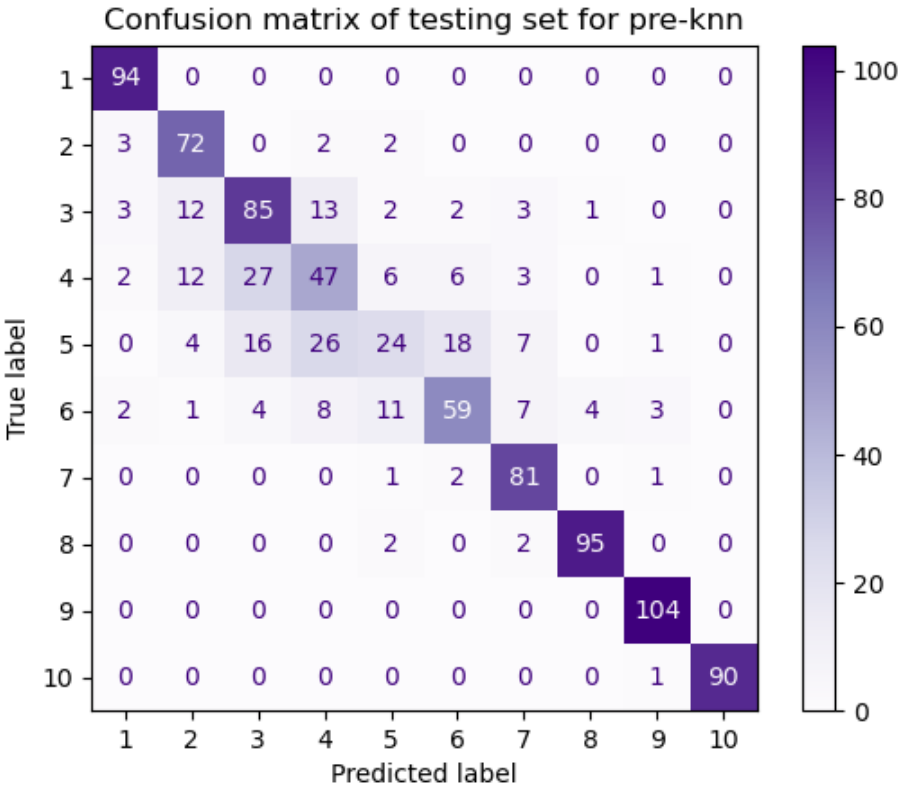


Fig 33 Confusion matrix of testing set for pre-knn

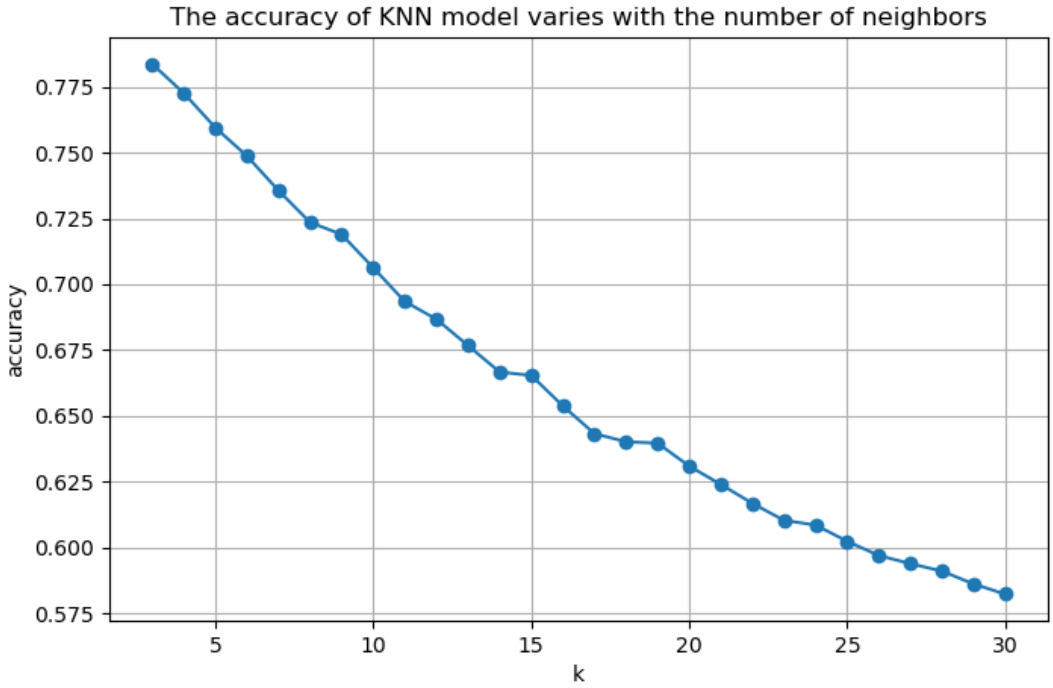


3.7.2 Model Optimization

In seeking to optimize the K-Nearest Neighbors (KNN) model, the primary task is to determine the optimal number of neighbors. For ease of analysis, a graph is presented illustrating the

correlation between the number of neighbors and model accuracy.

Fig 34 The accuracy of KNN model varies with the number of neighbors



The Fig 34 shows that as the number of neighbors (k) increases, the model's accuracy tends to decrease.

To prevent overfitting due to excessively high precision, I conducted a grid search with cross-validation starting from k=3. The results indicate that the optimal number of neighbors is 3.

3.7.2.1 Learning Curve Comparison

To verify whether the model's performance is affected by an excessively low number of neighbors leading to overfitting, I plotted learning curves for k=3 and k=5. Fig 35 shows the learning curve for k=3。And Fig 36 for k=5. Comparison reveals that the model did not change significantly, but the accuracy decreased by 10%, hence the decision to use 3 as the number of neighbors. Regardless of the value of k, the learning curves exhibit a similar trend, gradually approaching stability.

Fig 35 Learning Curves (KNN,k=3)

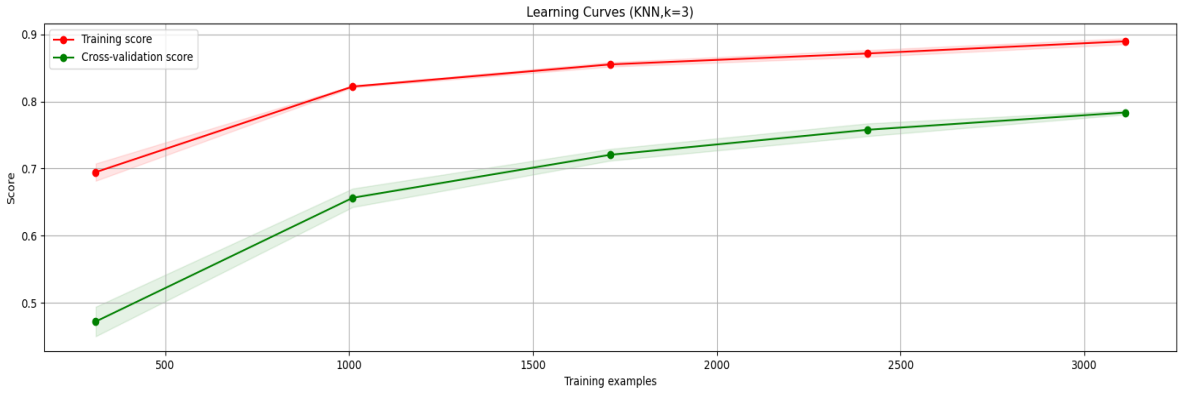
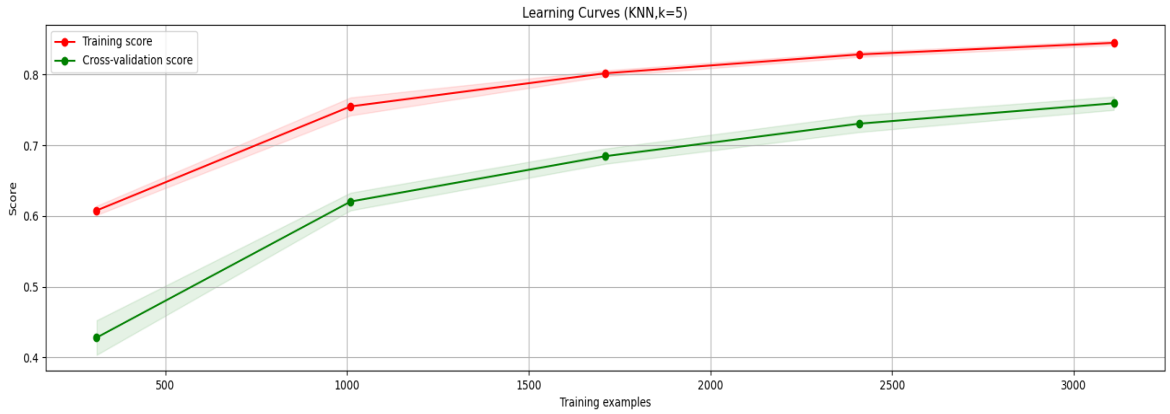


Fig 36 Learning Curves (KNN,k=5)



From this, it can be concluded that the optimal model is the one with k=3.

3.7.3 Model Assessment

3.7.3.1 Optimal Model and Accuracy

Table 23 Accuracy for optimal k-Nearest Neighbors

Accuracy on the training set for optimal k-Nearest Neighbors	0.901234568
Accuracy on the testing set for optimal k-Nearest Neighbors	0.804526749

Based on previous research, k=3 is identified as the optimal parameter for the KNN model. An assessment of this optimal model was conducted. The results for the optimal accuracy are presented in Table 23.

3.7.3.2 Confusion Matrix

Fig 37 Confusion matrix of training set for optimal k-Nearest Neighbors

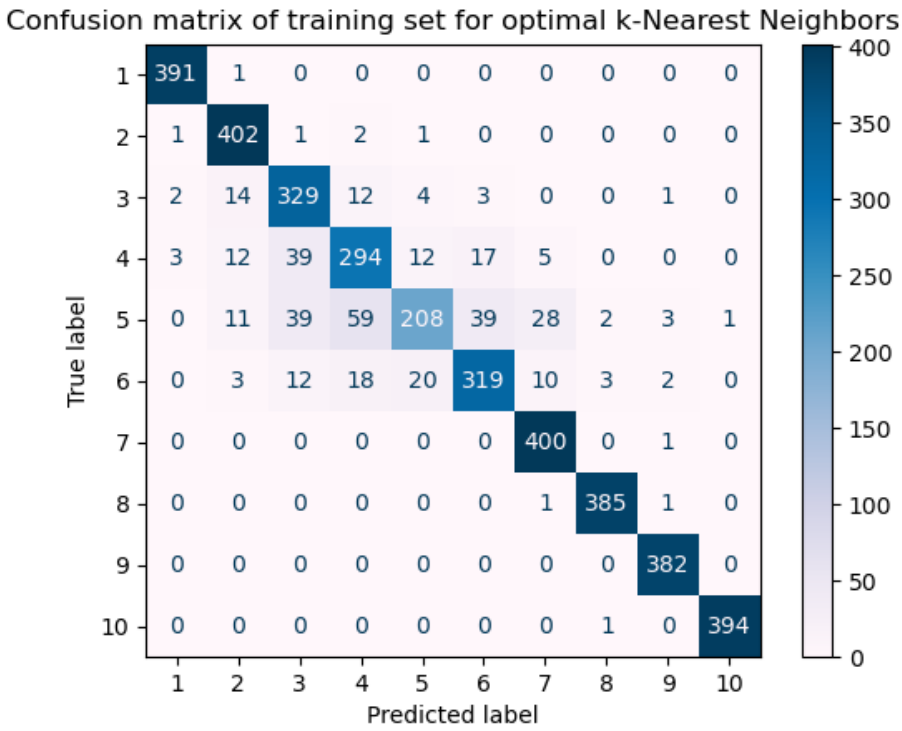
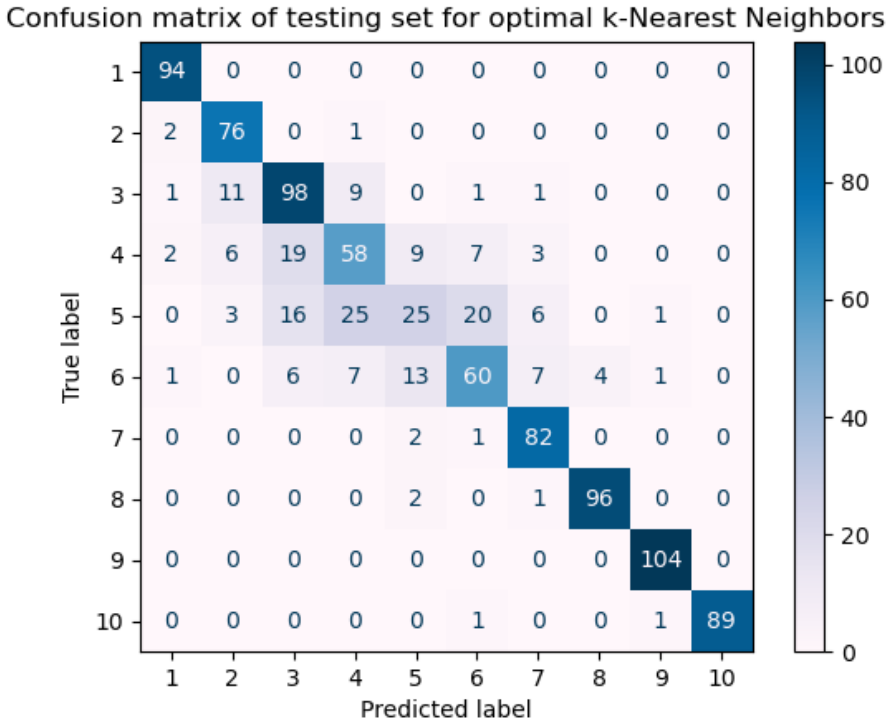


Fig 38 Confusion matrix of testing set for optimal k-Nearest Neighbors



Similarly, Fig 37 and Fig 38 display the confusion matrices for the training and test sets, respectively, to facilitate the observation of the model's predictive performance and generalization capabilities.

From the confusion matrix, it can be observed that the model has a strong generalization

capability, but it tends to confuse BBB and BB ratings (categories 4 and 5) in the test set. This is because the data for these two credit ratings are based on actual data, rather than being oversampled. In reality, BBB and BB ratings are particularly prone to confusion, and the algorithm tends to draw similar conclusions when selecting neighbors, thus causing the confusion.

Chapter 4 Summary

4.1 Summary of the Optimal Model

In the passed part, I selected seven models to evaluate my data and identified the most efficient and accurate model within each category using various methods.

Next, I will summarize each optimal model and compare them among the seven models to determine the one with the strongest generalization capability.

This section will follow the following steps:

- First, I will list the accuracy of each model on the test set to analyze the overall predictive capability of the model.
- Second, I will present the classification report for the model to analyze its predictive performance for each category and assess its generalization capability.
- Third, for the white-box models that have previously output feature importance, I will analyze the top five features with the greatest impact and their total explanatory ratio.
- Fourth, I will introduce the RGE (Random Gradient Boosting) analysis method to enhance the interpretability of the model and analyze the explanatory power of each feature under the RGE analysis logic.

Then, I will rank the test set accuracy of all models and select the top three models with the highest accuracy for additional interpretability analysis.

Based on the analysis of the interpretability of each RGE value, I will introduce the analysis of indicators grouped by year to assess the overall explanatory power of the model and determine if the model is influenced by time. For white-box models, I will compare the top five important features derived from the feature importance explanation logic with those from the RGE value

explanation logic. For black-box models, I will introduce the importance of the top five features under the RGE explanation, as well as their overall proportion.

4.1.1 Optimal Logistic Regression GLM Model

The optimal logistic regression model chosen is the one with only the original model's max_iter set to 2000. The model's accuracy on the test set is 59.36%.

4.1.1.1 Optimal Classification Report

The classification report is used to display the performance metrics of the training and test data, including precision, recall, and F1-score, as well as the support for each category (i.e., the number of samples in that category). This allows for a more intuitive demonstration of the model's predictive effectiveness for each category.

Table 24 displays the Classification Report of the GLM model on the training and test sets.

For the training set, categories 9 and 10 perform well, with F1 scores of 0.85 and 0.84, respectively, indicating that the model classifies these two categories effectively. However, due to the use of oversampling techniques in the training data, the sampling duplication rate for these two evaluations is high, making it easier for the model to predict.

Categories 3, 4, 5, 6, and 7 perform poorly, with F1 scores ranging from 0.39 to 0.45, suggesting that the model's classification effectiveness for these categories is not good.

Both the macro average (macro avg) and weighted average (weighted avg) F1 scores are 0.60, indicating that there is a certain degree of imbalance in the model's performance across different categories.

A comprehensive analysis reveals that the model's predictions are biased towards negative ratings (C, D ratings), which may be related to the results of oversampling. The overall accuracy

of the model is 61%, which is above average, proving that the model's performance during the training and learning process on the training set is acceptable.

Table 24 Classification Report of the GLM

training	precision	recall	f1-score	support
1	0.73	0.88	0.8	392
2	0.63	0.67	0.65	407
3	0.48	0.42	0.45	365
4	0.51	0.45	0.48	382
5	0.49	0.42	0.45	390
6	0.49	0.38	0.43	387
7	0.41	0.37	0.39	401
8	0.61	0.73	0.67	387
9	0.83	0.87	0.85	382
10	0.76	0.93	0.84	395
accuracy			0.61	3888
macro avg	0.6	0.61	0.6	3888
weighted avg	0.6	0.61	0.6	3888
testing	**precision**	**recall**	**f1-score**	**support**
1	0.79	0.87	0.83	94
2	0.54	0.72	0.62	79
3	0.56	0.42	0.48	121
4	0.42	0.37	0.39	104
5	0.4	0.32	0.36	96
6	0.42	0.36	0.39	99
7	0.39	0.39	0.39	85
8	0.65	0.77	0.7	99
9	0.84	0.86	0.85	104
10	0.76	0.92	0.84	91
accuracy			0.59	972
macro avg	0.58	0.6	0.58	972
weighted avg	0.58	0.59	0.58	972

For the test set, when faced with oversampled data, category 9 has the highest F1 score of 0.85, indicating that the model's classification effectiveness for category 9 is the best. When dealing with real data, categories 4 and 5 perform poorly, with F1 scores of 0.39 and 0.36, respectively, suggesting that the model's classification effectiveness for these two categories is not good.

Both the macro average and weighted average F1 scores are 0.58, close to the macro average

and weighted average F1 scores of the training data, indicating that the model's class balance on the test data is similar to that of the training data, but both face the problem of poor classification effectiveness.

The overall accuracy of the model is 59%, which is not significantly different from the accuracy of the training data (61%), suggesting that the model has good generalization capability on the test data.

4.1.1.2 Top 5 Important Features

In combination with the previous feature importance ranking results, we identified the top five features with the highest absolute values of coefficients and examined their degree of explanation.

The top five most important features obtained are 'TA_22', 'CL_22', 'EBIT_22', 'SECO_21', 'NI_22′.The total explanation percentage of these features is *32.60%.*

From this, we can infer that in the explanatory context of the absolute value ranking of coefficients, the financial data from 2022 and the scores from 2021 are relatively important explanatory features. Their combined explanatory power can reach over 30%.

To seek other explanations for the model, I also attempted other model interpretability methods.

4.1.1.3 Explainability

Following the method provided by Babaei, G., Giudici, P., & Raffinetti, E. (2024), I utilized the Rank Gradient Explainability (RGE) metric to evaluate these three optimal functions.

The feaipackage employed is a library of functions designed to assess the interpretability of features within models. By using this package, it is possible to evaluate the feature importance for black-box models, thereby enhancing the model's interpretability.

The Compute_rge_values function is used to measure the contribution of the given variables. The method for evaluating RGE values is not complex; a higher RGE value indicates a greater contribution of the variable.

By sorting the explanatory degree of each feature in the model using RGE values, the results are presented in Table 25.

Table 25 RGE Value（GLM）

rank	Feature	RGE	rank	Feature	RGE
1	TA_20	0.937767	17	REG	0.929088
2	NI_21	0.936313	18	ENV_R	0.928768
3	ORT_22	0.936	19	SF_20	0.927336
4	CL_22	0.934127	20	EBIT_21	0.925675
5	TA_21	0.933853	21	ESG_CLA	0.925031
6	EBITDA_21	0.933552	22	NI_20	0.922827
7	CA_20	0.933393	23	EBITDA_20	0.922714
8	CL_20	0.932707	24	EBIT_22	0.914711
9	SECO_20	0.932665	25	CA_22	0.9117
10	NI_22	0.932215	26	EBITDA_22	0.911109
11	Gov_R	0.931508	27	CA_21	0.910362
12	CL_21	0.931256	28	SF_21	0.907725
13	SECT	0.930498	29	ORT_21	0.90627
14	Social_R	0.930301	30	ORT_20	0.90627
15	TA_22	0.929758	31	SECO_21	0.894631
16	EBIT_20	0.929213	32	SF_22	0.893928

Based on the RGE (Rank Gradient Explainability) setting, it is evident that a higher RGE value indicates a greater degree of explanation. Analyzing each feature individually reveals that the variable with the highest explanatory power is TA_20, indicating that the glm model is most influenced by TA_20. Conversely, SF_22 has the lowest explanatory power under the RGE framework, suggesting that the explanation of shareholders' funds for the year 2022 is the weakest in this context.

Continuing with the RGE logic, the top-5 important features are identified as 'TA_20', 'NI_21',

'ORT_22', 'CL_22', and 'TA_21'. When combined, these features yield an RGE value of 0.952112, which signifies a very high degree of explanatory power.

4.1.2 Optimal Decision Tree Model

The selected optimal decision tree model is characterized by a maximum depth of 8 and a maximum number of leaf nodes set to 28. We will construct a model named op_dt for easier summarization. The accuracy of the optimal decision tree model on the test set is 70.267%.

4.1.2.1 Optimal Classification Report

Table 26 presents the classification reports for both the training and test sets.

Regarding the training data, the model's overall accuracy is 72%, indicating a decent performance on the training dataset. Categories 1, 8, 9, and 10 exhibit high F1 scores (all above 0.9), suggesting strong recognition capabilities for these categories, although there may be some overfitting due to oversampling.

Categories 3, 4, 5, and 6 have lower F1 scores (all below 0.65), indicating poorer performance in these categories. This suggests that the model's generalization ability on real data does not meet expectations.

Category 7 has a high recall rate (0.82), but a slightly lower precision (0.64), indicating that the model is good at identifying samples of this category but also tends to misclassify samples from other categories into this one. This suggests that the model is prone to confusion when identifying CCC ratings.

Table 26 Optimal Classification Report（DT）

training	precision	recall	f1-score	support
1	0.89	0.76	0.82	392
2	0.73	0.65	0.69	407
3	0.5	0.53	0.52	365
4	0.49	0.53	0.51	382
5	0.52	0.61	0.56	390
6	0.65	0.4	0.49	387
7	0.64	0.82	0.72	401
8	0.91	0.93	0.92	387
9	0.96	0.98	0.97	382
10	0.98	0.97	0.97	395
accuracy			0.72	3888
macro avg	0.73	0.72	0.72	3888
weighted avg	0.73	0.72	0.72	3888
testing	**precision**	**recall**	**f1-score**	**support**
1	0.92	0.63	0.75	94
2	0.62	0.67	0.64	79
3	0.63	0.61	0.62	121
4	0.49	0.55	0.52	104
5	0.49	0.53	0.51	96
6	0.64	0.42	0.51	99
7	0.58	0.82	0.68	85
8	0.88	0.92	0.9	99
9	0.95	0.96	0.96	104
10	0.98	0.95	0.96	91
accuracy			0.7	972
macro avg	0.72	0.71	0.7	972
weighted avg	0.72	0.71	0.7	972

The macro and weighted averages are similar and consistent with the overall accuracy, indicating that the model's performance across different categories is relatively balanced.

The report for the test set shows that the model's overall accuracy on the test data is 70%, slightly lower than that on the training data, suggesting some degree of overfitting, but not severe.

Categories 1, 8, 9, and 10 continue to perform well on the test data.

Categories 3, 4, 5, and 6 still perform poorly on the test data, with low F1 scores, consistent with their performance on the training data.

Category 7 has a high recall rate (0.82), but a lower precision (0.58), similar to the training data, but with a more pronounced discrepancy.

The macro and weighted averages are close and align with the overall accuracy, indicating that the model's performance across different categories is relatively balanced.

4.1.2.2 TOP-5 Important Features

Based on the previously calculated feature importance ranking, we will extract the top five features with the highest impact along with their respective explanatory proportions and sum up their degrees of explanation, as shown in Table 27.

Table 27 Top5 Important Features（DT）

SECO_21	NI_22	SF_22	Social_R	ORT_22	Total
29.39%	20.04%	8.66%	8.38%	5.49%	71.96%

The top five features account for more than 70% of the explanatory proportion, indicating that the model, in the context of feature importance, relies heavily on the rating results from the previous year. Under this explanatory logic, the credit rating from the previous year, along with the current year's net income, shareholders' funds, and social scores, carry significant weight.

4.1.2.3 Explainability

Continuing with the introduction of the RGE value for comparative analysis, we aim to uncover additional explanatory logics for the model.

The RGE models for each feature of the optimal decision tree model are presented in Table 28.

Table 28 RGE（DT）

rank	Feature	RGE	rank	Feature	RGE
1	Social_R	0.944911	17	CA_21	0.93722
2	EBITDA_20	0.93722	18	CA_22	0.93722
3	SF_22	0.93722	19	TA_20	0.93722
4	EBIT_20	0.93722	20	TA_21	0.93722
5	EBIT_21	0.93722	21	TA_22	0.93722
6	EBIT_22	0.93722	22	SECO_20	0.93722
7	ORT_20	0.93722	23	EBITDA_21	0.93722
8	ORT_21	0.93722	24	Gov_R	0.93722
9	EBITDA_22	0.93722	25	ENV_R	0.93722
10	CL_20	0.93722	26	ESG_CLA	0.93722
11	CL_21	0.93722	27	NI_20	0.93722
12	CL_22	0.93722	28	SF_20	0.93697
13	REG	0.93722	29	ORT_22	0.935986
14	SF_21	0.93722	30	SECT	0.935493
15	CA_20	0.93722	31	SECO_21	0.867723
16	NI_21	0.93722	32	NI_22	0.734429

The results of the RGE feature analysis indicate that the variable with the highest degree of explanation is Social_R, suggesting that the decision tree model is most influenced by Social_R. The variable with the lowest degree of explanation is NI_22, indicating that the net income for the year 2022 has the lowest explanatory power in the RGE framework. This demonstrates that under the RGE explanatory environment, changes in social evaluations have a significant impact on the model's analysis, while the influence of the 2021 ratings and changes in net income on the model's final assessment results has decreased.

Similarly, under the RGE logic, the top-5 important features are "Social_R", "EBITDA_20", "SF_22", "EBIT_20", and "EBIT_21". The combined RGE value obtained from these features is 0.944911, which is also very high, indicating that these features have a substantial combined impact on the model.

4.1.3 Optimal Random Forest Model

The optimal random forest (RF) model selected is the one with 600 trees, a maximum depth of 13, and the use of 'balanced' class weights for classification. The best model's test set accuracy is 88.17%, which is a very high level of precision.

4.1.3.1 Optimal Classification Report

For the training set, the precision, recall, and F1 scores for all categories are very high, almost all at 1.00, indicating that the model performs well on the training data with virtually no errors. Each category (from 1 to 10) has precision, recall, and F1 scores of 1.00 or close to 1.00, which means the model can accurately identify each category. The number of correct predictions for each model also remains above 380, with even distribution.

On the test set, the performance on the test data is not as good as on the training data, with an overall accuracy of 0.88, suggesting a decline in performance on unseen data. Categories 1, 8, 9, and 10 have very high precision, recall, and F1 scores, close to or equal to 1.00, indicating that the model's predictions for these categories are quite accurate based on oversampling.

Categories 2, 3, 6, and 7 perform secondarily, with precision, recall, and F1 scores ranging from 0.80 to 0.99.

Categories 4 and 5 perform the worst, with precision, recall, and F1 scores between 0.62 and 0.72, indicating that the model's ability to identify these categories is weaker, and there is room for improvement in actual predictions on real data.

Table 29 Classification Report（RF）

training	precision	recall	f1-score	support
1	1	1	1	392
2	1	1	1	407
3	1	0.99	1	365
4	0.99	0.99	0.99	382
5	0.99	0.99	0.99	390
6	1	0.99	1	387
7	1	1	1	401
8	1	1	1	387
9	1	1	1	382
10	1	1	1	395
accuracy			1	3888
macro avg	1	1	1	3888
weighted avg	1	1	1	3888
testing	**precision**	**recall**	**f1-score**	**support**
1	0.99	1	0.99	94
2	0.88	0.92	0.9	79
3	0.87	0.85	0.86	121
4	0.72	0.72	0.72	104
5	0.64	0.62	0.63	96
6	0.86	0.75	0.8	99
7	0.88	0.99	0.93	85
8	0.97	1	0.99	99
9	1	1	1	104
10	1	1	1	91
accuracy			0.88	972
macro avg	0.88	0.89	0.88	972
weighted avg	0.88	0.89	0.88	972

The support for each category (i.e., the number of samples in that category) ranges from 79 to 121, with relatively even distribution. The macro-average precision, recall, and F1 scores are all around 0.88, suggesting that the model has a good average performance across all categories. The weighted average precision, recall, and F1 scores are also similar to the macro-average, indicating that the model's overall performance remains good even after considering class imbalance.

The classification report for the model is presented as Table 29.

The model performs perfectly on the training data but shows a decline on the test data, which may indicate overfitting.

The model's performance is weaker on categories with more real-world data (such as categories 4 and 5), suggesting that more data may be needed for training.

Overall, the model has high accuracy and good generalization ability.

4.1.3.2 Top-5 Important Features

Using feature importance, we extract the top five most significant features of the model, and their degrees of explanation are presented in Table 30.

Table 30 Top-5 Important Features（RF）

SECO_21	NI_22	EBIT_22	SF_22	EBITDA_22	Total
9.33%	8.88%	6.96%	5.77%	5.71%	36.65%

The feature importance of the Random Forest model still primarily references the rating from 2021, followed by the integration of financial data from 2022, such as net income and EBIT. The total degree of explanation for the top five important features reaches over 35%.

4.1.3.3 Explainability

To enhance interpretability, we continue to use the RGE values.

The analysis reveals that despite the significant impact of NI_22, EBITDA_22, and EBIT_22 on the model in terms of feature importance, their performance under the RGE（Rank Gradient Explainability） logic is not impressive. Instead, total assets in 2022 emerge as the most influential factor under the RGE logic.

Again, under the RGE logic, the top-5 important features for the Random Forest model are "TA_22", "CL_22", "CL_21", "SECO_20", and "Social_R". The combined RGE value extracted from these features is 0.949004, which is very high, indicating that the combination

of these features is also crucial for the model.

The RGE values for each feature of the Random Forest model are presented in Table 31.

Table 31 RGE（RF）

rank	Feature	RGE	rank	Feature	RGE
1	TA_22	0.952703	17	SECO_21	0.933341
2	CL_22	0.949356	18	NI_21	0.933007
3	CL_21	0.947322	19	SF_21	0.932846
4	SECO_20	0.94618	20	CA_22	0.932034
5	Social_R	0.944307	21	ORT_22	0.929803
6	SF_20	0.942738	22	CL_20	0.928603
7	EBIT_20	0.942153	23	TA_20	0.925652
8	Gov_R	0.941725	24	EBITDA_20	0.922132
9	TA_21	0.9411	25	SF_22	0.91952
10	NI_20	0.940709	26	ESG_CLA	0.917433
11	CA_20	0.939898	27	ENV_R	0.910055
12	CA_21	0.939779	28	SECT	0.908282
13	REG	0.935987	29	EBIT_21	0.893656
14	ORT_20	0.935652	30	NI_22	0.893184
15	ORT_21	0.934844	31	EBITDA_22	0.891563
16	EBITDA_21	0.933705	32	EBIT_22	0.842155

4.1.4 Optimal Gradient Boosting Machine (GBM) Model

The chosen optimal gradient boosting machine (GBM) model is the original model. Now, we will construct a model named op_gbm for easier summarization. The accuracy of the obtained optimal model on the test set is 83.64%.

4.1.4.1 Optimal Classification Report

I continue to output the classification reports for the training and test sets, as shown in Table 32. The classification report for the training set shows that the overall accuracy (accuracy) of the training data is 0.97, indicating that the model performs exceptionally well on the training data.

Table 32 Classification Report（GBM）

training	precision	recall	f1-score	support
1	1	1	1	392
2	0.97	0.99	0.98	407
3	0.96	0.92	0.94	365
4	0.91	0.92	0.92	382
5	0.92	0.91	0.91	390
6	0.95	0.95	0.95	387
7	0.98	1	0.99	401
8	1	1	1	387
9	1	1	1	382
10	1	1	1	395
accuracy			0.97	3888
macro avg	0.97	0.97	0.97	3888
weighted avg	0.97	0.97	0.97	3888
testing	**precision**	**recall**	**f1-score**	**support**
1	1	0.95	0.97	94
2	0.81	0.89	0.85	79
3	0.78	0.78	0.78	121
4	0.63	0.64	0.64	104
5	0.58	0.53	0.55	96
6	0.75	0.69	0.72	99
7	0.85	0.94	0.89	85
8	0.97	1	0.99	99
9	1	1	1	104
10	1	1	1	91
accuracy			0.84	972
macro avg	0.84	0.84	0.84	972
weighted avg	0.83	0.84	0.83	972

The precision, recall, and f1-score for categories 1, 8, 9, and 10 are all 1.00, indicating that the model's recognition of these categories is almost perfect. The precision, recall, and f1-score for other categories are also above 0.90, suggesting that the model's recognition of these categories is quite accurate. This proves that the learning results from the oversampled data are very good. Facing the original real-world data, the f1-scores for categories 4 and 5 are relatively lower, at 0.92 and 0.91, but they still remain at a high level.

The classification report for the test set indicates that the overall accuracy (accuracy) of the test data is 0.84, which is a decrease compared to the training data but still decent. The precision, recall, and f1-score for categories 1, 8, 9, and 10 remain high, especially with precision and recall both being 1.00 for categories 8, 9, and 10, indicating that the model has a strong ability to recognize these categories.

The precision and recall for categories 2, 3, 6, and 7 are all above 0.75, suggesting that the model's recognition of these categories is also quite accurate. For categories with a higher proportion of resampled data, the model generally has better predictive results.

Facing categories 4 and 5, where real-world data account for a larger proportion, the predicted precision, recall, and f1-score are relatively lower, especially with precision and recall for category 5 both below 0.60, indicating that the model's ability to recognize these real-world categories is still not ideal.

Comparing the training data and test data, the model's performance on the training data is significantly better than on the test data, which suggests that the model still exhibits some degree of overfitting. However, the accuracy of the test data is still high, indicating that the model's generalization ability is overall good.

4.1.4.2 Top5 Important Features

Table 33 Top-5 Important Features

SECO_21	NI_22	EBIT_22	SF_22	EBITDA_22	Total
25.43%	10.65%	10.46%	7.29%	5.08%	58.91%

The most important top five features extracted by the feature importance model are listed in Table 33, along with their degree of explanatory power. The feature importance of the Gradient Boosting Machine (GBM) model primarily references the 2021 ratings as the most significant, followed by the integration of 2022 financial data, such as net income and EBIT. The

cumulative explanatory power of these top five features approaches 60%, demonstrating their efficiency in explaining the model.

4.1.4.3 Explainability

Table 34 RGE（GBM）

rank	Feature	RGE	rank	Feature	RGE
1	TA_22	0.943004	17	CL_20	0.925697
2	Social_R	0.938998	18	CA_20	0.924625
3	CL_22	0.938026	19	TA_20	0.924266
4	CA_21	0.937108	20	ORT_21	0.923862
5	EBITDA_20	0.935472	21	ORT_20	0.922525
6	CL_21	0.934992	22	SECO_20	0.922447
7	CA_22	0.933471	23	NI_21	0.919627
8	NI_20	0.933233	24	SF_20	0.919253
9	EBITDA_21	0.931658	25	ENV_R	0.916886
10	TA_21	0.930224	26	EBIT_21	0.908304
11	REG	0.928976	27	EBITDA_22	0.907567
12	SECT	0.928698	28	SECO_21	0.902624
13	EBIT_20	0.928103	29	NI_22	0.873194
14	Gov_R	0.927679	30	ORT_22	0.862994
15	SF_21	0.927647	31	EBIT_22	0.835412
16	ESG_CLA	0.926113	32	SF_22	0.798411

To enhance interpretability, the RGE values are continued to be used. The RGE values for each feature in the Gradient Boosting Machine model are presented in Table 34.

The analysis reveals that despite the significant impact of NI_22 and EBIT_22 on the model's explanatory power in terms of feature importance, their performance under the RGE (Rank Gradient Explainability) logic is not commendable. Conversely, the total assets in 2022 emerge as the most crucial factor under the RGE logic, with social ratings also holding significant importance. This conclusion is nearly identical to that of the Random Forest model.

To examine the overall impact mechanism, the top-5 important features extracted from the

Random Forest model under the RGE logic are "TA_22", "Social_R", "CL_22", "CA_21", and "EBITDA_20". The composite RGE value obtained from their combination is 0.963989, indicating a very high combined importance of these features in the model.

The aforementioned four models are considered white-box models in machine learning, as they allow for the provision of feature importance rankings. This enables a comparative analysis of the logic behind feature importance rankings versus the important features identified under the RGE logic. Moving forward, I will integrate three optimal black-box models, which do not facilitate the calculation of feature importance for ranking. Consequently, RGE values will be directly utilized for this purpose.

4.1.5 Optimal Support Vector Machine (SVM) Model

The chosen optimal SVM model employs the RBF kernel and sets the penalty parameter C for the error term to 100. We will now construct a model named op_svm for easier summarization. The best model achieved a test set accuracy of 79.73%.

4.1.5.1 Optimal Classification Report

Continuing with the standard procedure, we will export the classification report for the model, as presented in Table 35.

Table 35 Classification Report（SVM）

training	precision	recall	f1-score	support
1	0.98	1	0.99	392
2	0.92	0.94	0.93	407
3	0.84	0.83	0.84	365
4	0.83	0.82	0.83	382
5	0.86	0.77	0.81	390
6	0.82	0.82	0.82	387
7	0.87	0.95	0.91	401
8	0.99	0.99	0.99	387
9	0.99	1	1	382
10	1	1	1	395
accuracy			0.91	3888
macro avg	0.91	0.91	0.91	3888
weighted avg	0.91	0.91	0.91	3888
testing	**precision**	**recall**	**f1-score**	**support**
1	1	0.94	0.97	94
2	0.74	0.9	0.81	79
3	0.75	0.66	0.7	121
4	0.56	0.6	0.58	104
5	0.58	0.48	0.53	96
6	0.62	0.7	0.66	99
7	0.82	0.85	0.83	85
8	0.97	0.93	0.95	99
9	0.99	1	1	104
10	0.99	1	0.99	91
accuracy			0.8	972
macro avg	0.8	0.8	0.8	972
weighted avg	0.8	0.8	0.8	972

The overall accuracy of the model on the training dataset, which is 0.91, indicates a high level of accuracy.

The precision, recall, and f1-score for categories 1, 8, 9, and 10 are exceptionally high, nearing or reaching 1.00, suggesting that the model's identification of these categories is highly accurate.

The f1-scores for categories 2, 3, 4, 5, and 6 range between 0.82 and 0.84, indicating good performance with room for improvement.

Category 7 has an f1-score of 0.91, indicating that the model's predictions for this category are also relatively accurate. The overall accuracy on the test dataset is 0.8, showing a slight decline compared to the training data, but still a decent performance. The precision, recall, and f1-score for both macro avg and weighted avg are 0.80 on the test data, indicating stable overall performance of the model.

Categories 1, 8, 9, and 10 maintain high f1-scores on the test data, at 0.97, 0.95, 1.00, and 0.99 respectively, suggesting strong generalization capabilities of the model for these categories. The f1-scores for categories 2, 3, 4, and 5 are relatively lower, especially for categories 4 and 5, which are only 0.58 and 0.53 respectively, indicating weaker generalization capabilities for these categories. This suggests that the model may need further optimization, particularly when dealing with real-world data that has been more extensively added, highlighting the need for improved generalization through increased learning.

Categories 6 and 7 have f1-scores of 0.66 and 0.83 respectively, showing moderate performance, which could also be enhanced by increasing the learning volume. Overall, the model demonstrates excellent performance in terms of accuracy. For models that have not undergone sampling, it is essential to learn from a larger sample size to improve generalization capabilities and avoid overfitting. Overall, the performance of this SVM model is quite good.

4.1.5.2 Explainability

Since the SVM model is a black-box model without a ranking of feature importance, we directly calculate the RGE (Rank Gradient Explainability) values for each feature and then select the

top five for sorting.

The RGE values for all features are presented in Table 36.

Table 36 RGE（SVM）

rank	Feature	RGE	rank	Feature	RGE
1	TA_22	0.88975	17	EBITDA_22	0.86748
2	EBIT_21	0.889083	18	EBIT_20	0.866999
3	CA_20	0.888839	19	EBIT_22	0.866966
4	ORT_20	0.887728	20	CL_21	0.866078
5	ORT_21	0.887728	21	TA_20	0.865297
6	CA_21	0.887476	22	SF_20	0.86363
7	ORT_22	0.887382	23	REG	0.855575
8	CL_22	0.887183	24	CL_20	0.854744
9	EBITDA_20	0.878199	25	Gov_R	0.852046
10	Social_R	0.875917	26	NI_20	0.850104
11	TA_21	0.875525	27	SECO_20	0.849911
12	EBITDA_21	0.873513	28	SF_22	0.843884
13	CA_22	0.873367	29	ENV_R	0.818164
14	NI_21	0.872501	30	ESG_CLA	0.794052
15	NI_22	0.869685	31	SECT	0.788971
16	SF_21	0.867545	32	SECO_21	0.734782

The analysis of the RGE (Rank Gradient Explainability) values reveals that the SVM model performs slightly worse compared to the previous models in terms of RGE values. The earlier white-box models were almost all above 0.9, which may be related to the black-box nature of SVMs that prevents an accurate understanding of the internal mechanisms of the model.

From the analysis, it is evident that the most influential feature for the SVM model is the total assets in 2022, while the least influential feature is the credit rating in 2021.

Furthermore, the top-5 features affecting the SVM model are identified as "TA_22", "EBIT_21", "CA_20", "ORT_20", "ORT_21". Extracting these five features as a combination to examine the RGE value results in a value of 0.851609, which is even lower than the individual RGE values of the top-5 features. This suggests that combining these features in this instance actually

negatively impacted the model's performance, indicating the need for independent analysis of each feature to draw better conclusions.

4.1.6 Optimal Artificial Neural Network (ANN) Model

The optimal conclusion for the Artificial Neural Network (ANN) model is a single-hidden-layer MLP classifier with 100 neurons. The accuracy of the obtained model is 65.02%. The overall accuracy is relatively poor, which may be influenced by the performance of the equipment used for training the model, preventing further optimization.

4.1.6.1 Optimal Classification Report

Upon examining the classification report, it is evident that the prediction accuracy on the training dataset is 0.7, indicating that the performance on the training data is not particularly outstanding. Unlike previous classification results, the model demonstrates unusually accurate predictions for category 5 (BB) ratings, with an F1-score reaching 0.64. This suggests that the Artificial Neural Network (ANN) model is capable of correctly predicting some real-world data, and with better machine performance, the ANN model could yield even more precise results.

The classification report for the test set reveals that the overall accuracy of the model is 0.69, which is a slight decrease compared to the training data. However, the prediction accuracy for category 5 still exceeds expectations, indicating that the ANN model requires substantial time to improve its performance. This also suggests that the ANN model has the potential for better performance once its capabilities are enhanced.

The classification report for the optimal Artificial Neural Network model is presented in Table 37.

Table 37 Classification Report（ANN）

training	precision	recall	f1-score	support

1	0.94	0.7	0.8	392
2	0.62	0.79	0.69	407
3	0.49	0.73	0.58	365
4	0.72	0.3	0.43	382
5	0.59	0.69	0.64	390
6	0.73	0.34	0.46	387
7	0.51	0.85	0.64	401
8	0.85	0.59	0.7	387
9	0.91	0.95	0.93	382
10	0.98	0.99	0.99	395
accuracy			0.7	3888
macro avg	0.73	0.69	0.69	3888
weighted avg	0.73	0.7	0.69	3888
testing	**precision**	**recall**	**f1-score**	**support**
1	0.93	0.6	0.73	94
2	0.48	0.71	0.57	79
3	0.56	0.78	0.65	121
4	0.67	0.28	0.39	104
5	0.51	0.6	0.56	96
6	0.64	0.29	0.4	99
7	0.44	0.8	0.57	85
8	0.78	0.57	0.65	99
9	0.9	0.92	0.91	104
10	0.99	0.99	0.99	91
accuracy			0.65	972
macro avg	0.69	0.65	0.64	972
weighted avg	0.69	0.65	0.64	972

4.1.6.2 Explainability

Continuing with the standard procedure, we output the RGE values for all features in the Artificial Neural Network (ANN) model, as shown in Table 38. Unlike before, the totally asset feature no longer has the highest explanatory power; instead, the feature with the highest explanatory power is the EBIT (Earnings Before Interest and Taxes) from 2021. Additionally, the RGE analysis data is generally below 0.75, indicating that the model still needs further strengthening.

Table 38 RGE（ANN）

rank	Feature	RGE	rank	Feature	RGE
1	EBIT_21	0.721792	17	SF_21	0.653704
2	NI_22	0.721328	18	SF_20	0.652101
3	EBITDA_20	0.711076	19	CA_20	0.641916
4	NI_20	0.709795	20	CL_21	0.630581
5	Gov_R	0.694169	21	CA_22	0.59945
6	ENV_R	0.69415	22	ORT_21	0.594833
7	ESG_CLA	0.694144	23	EBIT_22	0.586952
8	Social_R	0.694103	24	CL_20	0.56541
9	REG	0.694029	25	TA_22	0.554782
10	SECO_20	0.693933	26	ORT_20	0.552512
11	SECO_21	0.693763	27	TA_20	0.546041
12	SECT	0.693675	28	TA_21	0.541852
13	NI_21	0.684108	29	ORT_22	0.531763
14	EBIT_20	0.68226	30	CL_22	0.530227
15	EBITDA_22	0.664419	31	SF_22	0.528559
16	EBITDA_21	0.659589	32	CA_21	0.513111

The model's feature ranking shows a relatively clear division by data category. For instance, the RGE values for the ratings from 2020 and 2021 are close, as are the RGE values for the EBITDA from 2022 and 2021, as well as the shareholders' funds and totally asset from 2020 and 2021. Among the top five features influencing the model, two are net income. This suggests that under the RGE value analysis logic, the ANN model has similar levels of importance for features from

the same data source.

Following the standard procedure, we extract the top five important features affecting model accuracy, which are: "EBIT_21", "NI_22", "EBITDA_20", "NI_20", "Gov_R". It is noteworthy that government ratings and financial data from 2020 and 2021 are included in the important features, and net income appears repeatedly.

Combining the top five important features, we obtain an RGE value result of 0.690527, which is a moderately high level but has weaker explanatory power compared to other models.

4.1.7 Optimal K-Nearest Neighbors (kNN) Model

The chosen optimal K-Nearest Neighbors (KNN) model is the one with the best number of neighbors set to 3. The best model achieved an accuracy of 80.45%.

4.1.7.1 Optimal Classification Report

Table 39 presents the classification report for the kNN model.The analysis of the classification report reveals that the overall accuracy on the training data is 0.9, indicating that the model performs well in terms of comprehensive data. The precision, recall, and f1-score for both macro avg and weighted avg are 0.90, suggesting that the model's performance is relatively even across all categories.

When examining individual categories, Category 5 has a lower recall rate of only 0.53, indicating that the model has a poor recognition ability for this category and faces several issues. However, the performance of other categories is good, with f1-scores above 0.83.

In the classification report for the test data, the overall accuracy drops to 0.8, which is still within an acceptable range. The precision, recall, and f1-score for both macro avg and weighted avg are around 0.80, indicating that the model's overall performance on the test data is good.

Table 39 Classification Report（KNN）

training	precision	recall	f1-score	support
1	0.98	1	0.99	392
2	0.91	0.99	0.95	407
3	0.78	0.9	0.84	365
4	0.76	0.77	0.77	382
5	0.85	0.53	0.66	390
6	0.84	0.82	0.83	387
7	0.9	1	0.95	401
8	0.98	0.99	0.99	387
9	0.98	1	0.99	382
10	1	1	1	395
accuracy			0.9	3888
macro avg	0.9	0.9	0.9	3888
weighted avg	0.9	0.9	0.9	3888
testing	**precision**	**recall**	**f1-score**	**support**
1	0.94	1	0.97	94
2	0.79	0.96	0.87	79
3	0.71	0.81	0.75	121
4	0.58	0.56	0.57	104
5	0.49	0.26	0.34	96
6	0.67	0.61	0.63	99
7	0.82	0.96	0.89	85
8	0.96	0.97	0.96	99
9	0.97	1	0.99	104
10	1	0.98	0.99	91
accuracy			0.8	972
macro avg	0.79	0.81	0.8	972
weighted avg	0.79	0.8	0.79	972

Upon analyzing individual categories, it is evident that models that cannot widely apply oversampling techniques generally have lower f1-scores. Categories 4 and 5 have f1-scores of 0.57 and 0.34, respectively, indicating that the model has poor generalization ability for these categories, especially Category 5, which has low precision and recall. This suggests that the kNN model requires a larger dataset for training to achieve better generalization ability.

4.1.7.2 Explainability

We continue to use RGE values to evaluate the explainability of the kNN model. The RGE values for each feature are presented in Table 40.

The RGE (Rank Gradient Explainability) values for the K-Nearest Neighbors (kNN) model are not ideal, with the highest proportion of feature importance scoring only 0.64. The analysis reveals that the most important explanatory factor for the kNN model is the shareholders' funds for the year 2020, while totally asset and current liabilities are among the least important features.

Table 40 RGE（kNN）

rank	Feature	RGE	rank	Feature	RGE
1	SF_20	0.643982	17	SECO_20	0.61084
2	CA_21	0.634078	18	EBIT_20	0.610003
3	SF_21	0.623716	19	NI_20	0.609024
4	ORT_22	0.623469	20	NI_22	0.607765
5	SF_22	0.616108	21	EBIT_22	0.592678
6	EBITDA_21	0.615989	22	EBITDA_22	0.586246
7	EBIT_21	0.614344	23	CA_20	0.581198
8	NI_21	0.61353	24	CA_22	0.579089
9	EBITDA_20	0.613411	25	TA_22	0.562846
10	Gov_R	0.61084	26	ORT_21	0.557762
11	SECO_21	0.61084	27	ORT_20	0.557762
12	ESG_CLA	0.61084	28	CL_22	0.551703
13	ENV_R	0.61084	29	TA_20	0.540783
14	Social_R	0.61084	30	TA_21	0.538106
15	REG	0.61084	31	CL_21	0.532115
16	SECT	0.61084	32	CL_20	0.505671

The top five features influencing the kNN model are identified as "SF_20", "CA_21", "SF_21", "ORT_22", and "SF_22". The fact that all shareholders' funds are present in the top five suggests that shareholders' funds contribute significantly to the clustering of samples, thus resulting in

higher accuracy.

When calculating the RGE value for the group consisting of these top five important features, the obtained value is 0.573421, which is lower than the values obtained from analyzing each feature individually. This indicates that these five features may not be suitable for combined analysis.

4.2 Accuracy Comparison Analysis

After individually evaluating all the optimal models, I have decided to compare their performances to identify the model with the best predictive capabilities.

Table 41 presents a list of all the optimal models ranked by their accuracy on the test set in descending order. By measuring the accuracy of the models on the test set, we can assess both the generalization ability of the models and their overall performance.

Table 41 Accuracy Ranking

Random Forest (RF)	Gradient Boosting Machine (GBM)	K-Nearest Neighbors (kNN)	Support Vector Machine (SVM)	Decision Tree (DT)	Artificial Neural Network (ANN)	Logistic Regression Model (GLM)
88.169%	83.642%	80.453%	79.733%	70.267%	65.021%	59.362%

From this, it is evident that the model with the highest accuracy is the Random Forest model, followed by Gradient Boosting Machine, Support Vector Machine, and k-Nearest Neighbors, all of which demonstrate good performance. Due to constraints on computational time and the performance of the training equipment, the Artificial Neural Network model could not be optimized and thus shows poorer results.

Previous research has shown that Random Forest models are widely used in the field of credit rating for learning purposes, and my research also confirms that Random Forest models are excellent white-box models in terms of performance.

Gradient Boosting Machine and Support Vector Machine models are also extensively used in the field of machine learning for predictive credit ratings. Typically, after increasing the number of training samples, the Support Vector Machine and Artificial Neural Network models can achieve very accurate prediction results. The analysis of previous classification reports also reveals that the Support Vector Machine and Artificial Neural Network models have high recall and precision rates in real-world data with high proportions and without the use of oversampling techniques. This suggests that, with extensive data training, black-box models should become efficient predictive models for credit ratings.

4.3 Explainability Comparison Analysis

My analysis not only selects the most suitable machine learning models for credit rating but also ranks the importance of the 32 features used.

Based on previous statistics, I have summarized the top five important features and their combined explanatory proportions for white-box models in feature importance analysis. Since black-box models cannot calculate feature importance, I have also added conclusions based on RGE (Rank Gradient Explainability) values for all models.

Using Table 42, we can compare the important features and their proportions under different feature analysis logics for white-box models, as well as analyze the most influential features under a unified measurement standard (RGE values) for different models.

Table 42 Explainability Comparison Analysis

MODEL	ITEM	TOP-1	TOP-2	TOP-3	TOP-4	TOP-5
Logistic Regression GLM Model	top 5 RGE group RGE	TA_20	NI_21	ORT_22	CL_22	TA_21
				0.952112		
	top 5 IF	TA_22	CL_22	EBIT_22	SECO_21	NI_22
	total IF			32.60%		
Decision Tree Model	top 5 RGE group RGE	Social_R	EBITDA_20	SF_22	EBIT_20	EBIT_21
				0.944911		
	top 5 IF	SECO_21	NI_22	SF_22	Social_R	ORT_22
	total IF			71.96%		
Random Forest RF Model	top 5 RGE group RGE	TA_22	CL_22	CL_21	SECO_20	Social_R
				0.949004		
	top 5 IF	SECO_21	NI_22	EBIT_22	SF_22	EBITDA_22
	total IF			36.65%		
Gradient Boosting Machine(gbm) Model	top 5 RGE group RGE	TA_22	Social_R	CL_22	CA_21	EBITDA_20
				0.963989		
	top 5 IF	SECO_21	NI_22	SF_22	EBIT_22	EBITDA_22
	total IF			58.91%		
Support Vector Machine (svm)	top 5 RGE group RGE	TA_22	EBIT_21	CA_20	ORT_20	ORT_21
				0.851609		
	IF	For black-box models, computation is not feasible.				

MODEL	ITEM	TOP-1	TOP-2	TOP-3	TOP-4	TOP-5
Artificial Neural Network (ANN) Model	top 5 RGE group RGE	EBIT_21	NI_22	EBITDA_20	NI_20	Gov_R
	RGE	0.690527				
	IF	For black-box models, computation is not feasible.				
K-Nearest Neighbors (kNN) Model	top 5 RGE group RGE	SF_20	CA_21	SF_21	ORT_22	SF_22
	RGE	0.573421				
	IF	For black-box models, computation is not feasible.				

Under the RGE value logic, among the seven models used, the most frequently and significantly featured attribute was the total assets from 2022. The EBITDA from 2020 also appeared as a recurring feature in social ratings. Considering the nature of the data to which the features belong, total asset data was most likely to be considered an important feature. Moreover, under the RGE value logic, the diversity in the years of the features used was observed.

In a comprehensive evaluation, the more important indicators identified were the rating result from 2021 (SECO_21), total assets from 2022 (TA_22), and social rating (Social_R), which are the key features that require close attention. While focusing on these specific features, it is also essential to pay close attention to the financial metrics of EBIT, EBITDA, net income, and shareholders' funds in the annual financial data. Operating revenue, current liabilities, and total assets from other years should be given some attention. The impact of regional distribution and business sector distribution was found to be minimal, and current assets were also identified as a financial data point that is easily overlooked.

Based on the conclusions regarding feature importance, it was found that the top five important features in terms of total feature importance accounted for a maximum proportion of 71.96%.

The model with the highest degree of explainability was the decision tree model, which achieved an explainability of over 70%. The model with the lowest degree of explainability was the logistic regression model, with an explainability of only 32.6%.

The group RGE score is used to indicate the impact of the feature combination on model variations. Group RGE offers better explainability, taking into account each model and its important features extensively.

Therefore, it can be observed that the selected top five important features have a significant explanatory power. This demonstrates that the chosen features can participate in the process of model fitting results with a larger proportion.

4.4 Extra Explainability Better Accuracy Models

Based on the accuracy rankings, the 4 models with the best accuracy are Random Forest(RF),Gradient Boosting Machine (gbm), K-Nearest Neighbors (knn), and Support Vector Machine (svm).

For the four high-quality models selected,whose accuracy greater than 80%, I plan to add additional explanatory analyses.

Additional explanatory method was added to preliminarily explore the impact of the group of data years in which the features are present on the model. Using the method of constructing combinations with RGE, the data was combined separately in chronological order to investigate the degree of influence of the data's year on the model under the optimal model. This serves as a preliminary assessment of whether time has a significant impact on the model.

4.4.1 Random Forest(RF)

Firstly, I analyzed the model with the best accuracy performance, which is the random forest

model. The annual analysis conclusions from this model are presented in Table 43 as follows.

Table 43 annual RGE（RF）

year	RGE
2022	0.773942
2021	0.934371
2020	0.938901

It can be observed that the random forest model performs the worst with the 2022 data and the best with the 2020 data, indicating that the time combination of 2020 has a relatively significant impact. However, since the conclusions for 2021 and 2020 are not substantially different, it suggests that the random forest model may not be greatly influenced by time group.

Regardless of the data from which year, the RGE values are relatively high, indicating that the data combination remains crucial for model analysis. This is in line with the algorithmic principles of the random forest model.

4.4.2 Gradient Boosting Machine (gbm)

Using the same process, Table 44 presents the RGE values for the Gradient Boosting Machine (GBM) model across different year group.

It can be observed that the GBM model performs optimally with the year group of 2021, indicating that the GBM model is significantly influenced by the data from 2021. However, similar to the random forest model, the conclusions for 2021 and 2020 are not substantially different, suggesting that the GBM model is also unlikely to exhibit significant variability in results due to temporal differences.

Table 44 Annual RGE（GBM）

year	RGE
2022	0.719141
2021	0.934465
2020	0.911236

Additionally, it can be noted that for any year's data group, the RGE values are high, indicating that the GBM model is relatively holistic, with each feature holding significant importance.

4.4.3 K-Nearest Neighbors (knn)

Table 45 displays the RGE values for the k-Nearest Neighbors (k-NN) model across different year group.

Table 45 Annual RGE（KNN）

year	RGE
2022	0.476548
2021	0.518257
2020	0.549494

The analysis of the k-NN model shows a significant difference from other models, as the k-NN model exhibits poor performance in terms of RGE values, with low explanatory power across different years. Since the explanatory power does not vary much and is not high, it can be inferred that the conclusions of the data model have little relationship with the year.

4.4.4 Support Vector Machine (svm)

Table 46 presents the RGE values for the Support Vector Machine (SVM) model across different year group

Table 46 Annual RGE（SVM）

year	RGE
2022	0.761142
2021	0.833973
2020	0.802096

Similar conclusions can be drawn for the SVM model. Firstly, the SVM model performs relatively well in terms of RGE values, with a high degree of explanatory power for various data sets. However, the differences between different years are not significant, suggesting that the model is not greatly influenced by temporal factors.

Chapter 5 Conclusions and Policy Recommendations

5.1 Model Conclusions:

By selecting the optimal test set accuracy, the longitudinal optimal model identified is the Random Forest model, which exhibits an accuracy of 88%, performing exceptionally well in comparison to other models in terms of predictive effectiveness. Additionally, models such as Gradient Boosting Machines, k-Nearest Neighbors, and Support Vector Machines have achieved an accuracy of around 80%, making them excellent choices for predicting credit ratings.

Furthermore, through the analysis of explainability, integrating all explanatory logic and models, it was found that SECO_21, TA_22, Social_R, NI_22, and SF_22 are among the five features with high frequency of occurrence. Other significant features include EBIT and EBITDA, which are secondary only to the aforementioned features.

The analysis also reveals that in terms of feature importance and the evaluation of analytical logic, the rating results from 2021 and the financial data from 2022 are crucial reference features affecting the model. From the perspective of the evaluation logic of RGE values, total assets and social evaluation carry higher explanatory weights.

Moreover, model evaluation has also highlighted a severe issue of class imbalance in the dataset used, comprising 1559 data points, which has an impact on machine learning training. For more in-depth and precise research, it would be necessary to increase the sample size and diversity for algorithmic training.

Additionally, this paper briefly investigated the impact of time combinations on several well-performing models. The conclusion drawn is that time combinations do not significantly affect

the evaluation in the models trained, suggesting that credit ratings are not influenced by the year in which the features are recorded.

5.2 Policy Recommendations

Based on the conclusions of this study, the following recommendations are proposed:

1. For companies seeking predictive credit ratings, it is essential to consider the credit rating results from the previous year and pay close attention to the current year's economic indicators. Additionally, companies should focus on enhancing their social evaluation within ESG scores. Moreover, during ongoing operations, companies should prioritize maintaining stability in financial indicators such as total assets, EBITDA, and net income to ensure the predictability of credit ratings.

2. Credit rating agencies aiming to simulate credit ratings through machine learning methods should opt for the Random Forest model, which offers the best evaluation results under limited computational resources. Followed by Gradient Boosting Machines, k-Nearest Neighbors, and Support Vector Machines. If computational resources are available, more generalized artificial neural network models can be considered.

3. For institutions looking to employ machine learning methods for analytical purposes, it is crucial to use high-performance computing devices for efficient operations. To avoid model overfitting, a more complex grid search cross-validation method should be chosen. Given the inevitability of class imbalance in credit rating datasets, it is necessary to increase the number of training samples and combine different sampling methods for data evaluation and algorithm training. It is also important to distinguish between feature importance explanations and RGE value explanations that cause feature variability. Future research on credit ratings can focus on

enhancing explainability, improving generalization capabilities, and reducing class imbalance. Additionally, deeper investigations can be conducted into the impact of time combinations on machine learning models for credit ratings.

REFERENCES

[1] Alonso, A., & Carbó, J. M. (2021). Understanding the performance of machine learning models to predict credit default: A novel approach for supervisory evaluation. Documentos de Trabajo. N.º 2105. Banco de España. https://ssrn.com/abstract=3774075

[2] Alonso, A., & Carbó, J. M. (2020). Machine Learning in Credit Risk: Measuring the Dilemma Between Prediction and Supervisory Cost. Documentos de Trabajo N.º 2032, 374. https://ssrn.com/abstract=3724374

[3] Babaei, G., & Giudici, P. (2024). GPT Classifications, with Application to Credit Lending. Machine Learning with Applications, 16, 100534. https://doi.org/10.1016/j.mlwa.2024.100534

[4] Babaei, G., Giudici, P., & Raffinetti, E. (2023). Explainable FinTech lending. Journal of Economics and Business, 125-126, 106126. https://doi.org/10.1016/j.jeconbus.2023.106126

[5] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Müller, A. C., Grisel, O., … & Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. arXiv preprint arXiv:1309.0238.

[6] Dai, Z., Li, A., Yuchen, Z., & Qian, G. (2021). The application of machine learning in bank credit rating prediction and risk assessment. In Proceedings of the 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE 2021) (pp. 986-989). IEEE.

[7] Dong, D., Lin, B., & Dong, X. (2024). Logistics financial risk assessment based on decision tree algorithm model. Procedia Computer Science, 243, 1095–1104. https://doi.org/10.1016/j.procs.2024.09.130

[8] Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear

[9] Golbayani, P., Florescu, I., & Chatterjee, R. (2020). A comparative study of forecasting corporate credit ratings using neural networks, support vector machines, and decision trees. North American Journal of Economics and Finance, 54, 101251.

[10] Khandani, E. A., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 34(8), 2767–2787. https://doi.org/10.1016/j.jbankfin.2010.06.001

[11] Lopez-Arevalo, I., Aldana-Bobadilla, E., Molina-Villegas, A., Galeana-Zapién, H., Muñiz-Sanchez, V., & Gausin-Valle, S. (2020). A memory-efficient encoding method for processing mixed-type data on machine learning. Entropy, 22(12), 1391.

[12] Pamuk, M., & Schumann, M. (2023). Opening a New Era with Machine Learning in Financial Services? Forecasting Corporate Credit Ratings Based on Annual Financial Statements. International Journal of Financial Studies, 11(3), 96. https://doi.org/10.3390/ijfs11030096

[13] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & Duchesnay, E. (2011). Journal of Machine Learning Research, 12, 2825-2830.

[14] Takawira, O., & Mwamba, J. W. (2020). Determinants of sovereign credit ratings: An application of the naïve Bayes classifier. Eurasian Journal of Economics and Finance, 8(4), 279-299. https://doi.org/10.15604/ejef.2020.08.04.008

[15] Tsai, C.-F., & Chen, M.-L. (2010). Credit rating by hybrid machine learning techniques. Applied Soft Computing, 10(3), 374–380.

[16] Ye, Y., Liu, S., & Li, J. (2008). A multiclass machine learning approach to credit rating prediction. In 2008 International symposiums on information processing (ISIP) (pp. 57–61).

[17] Zontul, M., Sönmez, F., Ajlouni, N., Hameed, A. A., Ajlouni, F., Dehghanian, K., & Moghimi, S. (2020). Customer Credit Rating Estimation Using Machine Learning Methods. International Journal of Economics and Management Sciences, 9(1), 36-38.