Bankruptcy Prediction Using Machine Learning

Introduction

Corporate bankruptcy prediction is a crucial task in finance, impacting investors, creditors, and stakeholders. Leveraging machine learning, we aim to forecast whether a company will go bankrupt two years into the future using financial data. The dataset includes key variables such as EPS, Liquidity, Profitability, etc., with the target variable "BK" indicating bankruptcy (1) or non-bankruptcy (0) status. Our approach involves spot-checking models like Random Forest, SVM, Logistic Regression, LDA/QDA, and Gradient Boosting, selecting those with the highest F-1 Macro score for further development. This report outlines our experimental setup, covering data preprocessing, model methodologies, hyperparameter tuning, and evaluation criteria. We'll compare model performance and offer recommendations for model selection, discussing practical applications in the financial world and their stakeholder implications.

Section 1: Exploratory Data Analysis (EDA)

Our initial exploration of the dataset involved several steps to gain insights into the data's characteristics, patterns, and potential issues.

Data Overview: We began by examining the structure and contents of the dataset. This included checking for missing values and identifying the data types of variables. Moreover, we removed any duplicated rows from the dataset to ensure the integrity of our analysis and prevent biases that may arise from duplicate observations.

Statistical Summary: This provided us with an overview of the dataset's numerical characteristics. We compared both Bankrupted companies and non-bankrupted companies. We found that Bankrupted companies have relatively lower EPS and Return on Equity than non-bankrupted companies. This could potentially be the statistical difference between those two kinds of companies. Apart from this, both companies are not much different in other areas.

Correlation Matrix: We constructed a correlation matrix to examine the relationships between variables. There were low to moderate correlations between independent variables, however, none of the independent variables correlated with dependent variables. As shown in **Figure 1**

Distribution of Bankruptcy Status: We analyzed the distribution of bankrupted and non-bankrupted companies in the dataset. We found that this dataset is highly imbalanced between bankrupted and non-bankrupted companies where 99.4 percent of all companies were non-bankrupted while only 0.6% of all companies were bankrupted companies. Therefore, we need to apply resampling to our models.

Shape and Outliers: Lastly, we conducted a histogram as shown in **Figure 2** to see a distribution of the data and a box plot as shown in **Figure 3** to identify the outliers. The shape of the data for all variables was highly stuck together which was possibly influenced by the high number of outliers. This gave us an insight that when we conducted the data processing step, we needed to take care of this issue.

By conducting comprehensive exploratory data analysis, we gained valuable insights into the dataset's characteristics, identified potential patterns and trends, and highlighted any data issues that needed to be addressed during the modeling process. This foundational understanding will inform our subsequent modeling steps and help us develop accurate and robust bankruptcy prediction models.

Figure 1: Heat Map

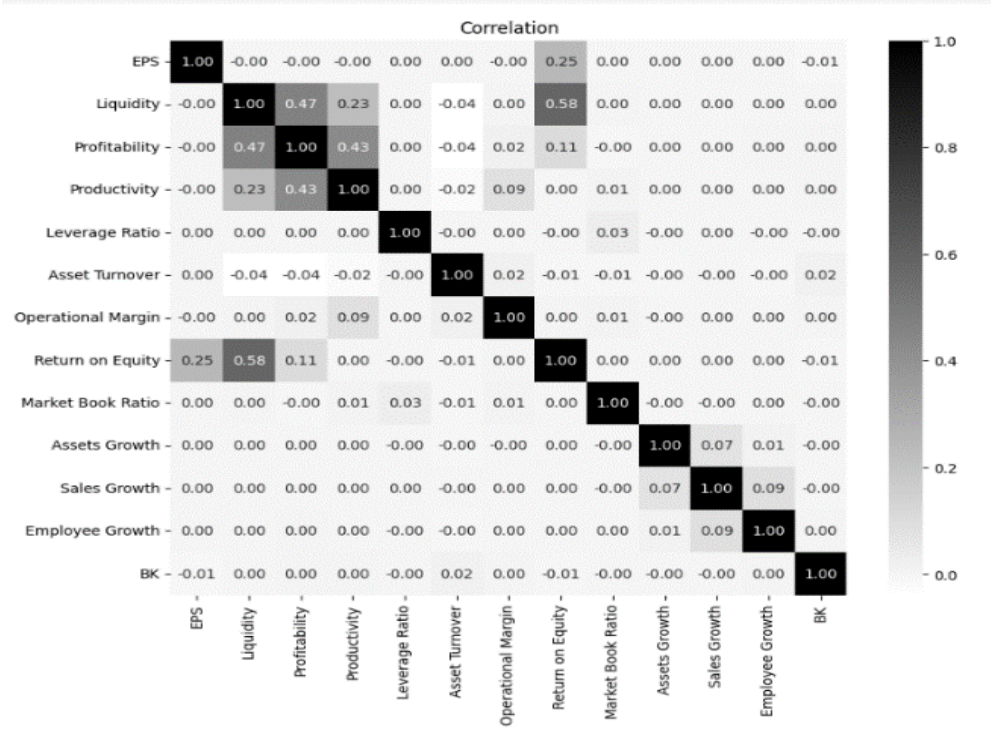


Figure 2: Distribution Of the Data

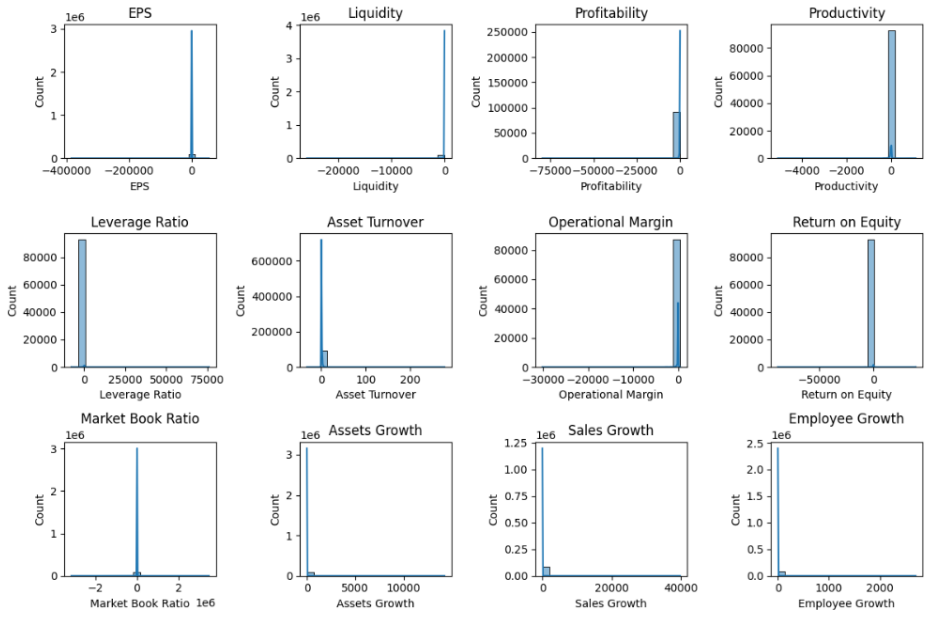
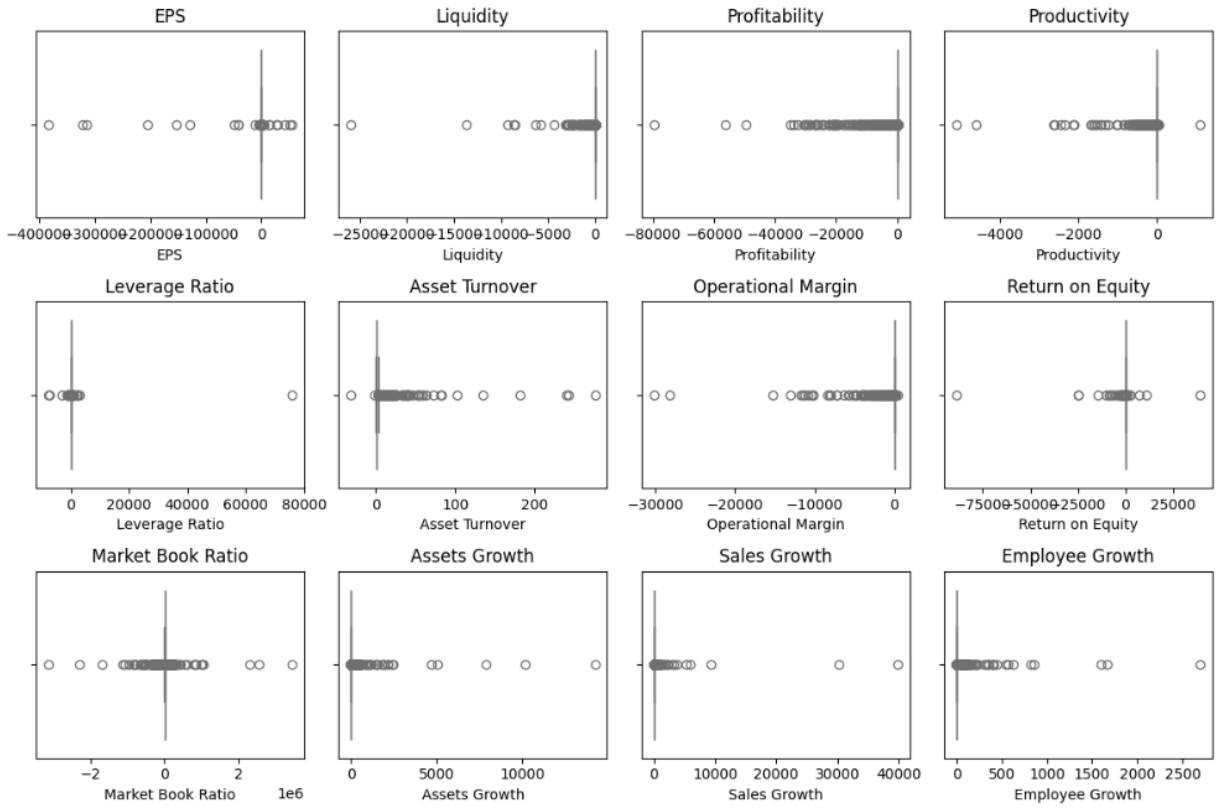


Figure 3: Box Plot



Section 2: Data Processing

In this section, we detail the steps taken to preprocess and transform the data before feeding it into our machine-learning models. The data processing stage is crucial for ensuring the quality and reliability of our predictive models.

Handling Missing Data: We addressed missing values in the dataset using the IterativeImputer technique. This method is suitable for datasets with complex relationships between variables and effectively handles missing data while preserving the underlying data structure.

Standardize: To address the presence of outliers and non-normal data in the dataset, RobustScaler was utilized. RobustScaler is robust to outliers, meaning that it scales features in a way that is less influenced by extreme values. This was crucial for maintaining the integrity of the data and preventing outliers from unduly influencing the performance of our machine-learning models.

Resampling: Since bankruptcy prediction datasets typically exhibit class imbalance, with a significantly larger number of non-bankrupt instances than bankrupt instances, we addressed this issue using the Synthetic Minority Over-sampling Technique (SMOTE).

By employing a pipeline to orchestrate these preprocessing steps, we ensure the data transformation process is seamless, reproducible, and integrated into our machine-learning workflow. Each technique used serves a specific purpose in enhancing the quality and effectiveness of our predictive models, ultimately leading to more accurate and reliable bankruptcy predictions.

Section 3: Model Development

3.1 Model Selection

In this section, we conducted model development using a two-step approach. Firstly, we performed a spot check for five different models: Random Forest, Support Vector Machine (SVM), Logistic Regression, Linear Discriminant Analysis (LDA), and Gradient Boosting. Subsequently, we refined our approach by utilizing a pipeline incorporating data preprocessing techniques and model selection. The chosen preprocessing techniques were tailored to address the characteristics of the dataset, aiming to improve model performance and generalization ability.

First Attempt: For the initial attempt, we evaluated the models using basic preprocessing techniques and aimed for an F1-macro score. The preprocessing steps included data imputation using IterativeImputer and data scaling with RobustScaler. We employed the following models: RandomForestClassifier, SVC, LogisticRegression, Linear Discriminant Analysis (LDA), and Gradient Boosting Classifier. The results are shown below in **Figure 4.** The Gradient Boosting Classifier (.55) performed the best followed by the Logistic Regression Classifier (.52) and Random Forest Classifier (.51).

Figure 4: First Spot Check F1-Macro Scores Comparison.

Model	F1-macro Score	Reason for Selection
RandomForestClassifier	0.51	Robust to outliers, handles non-linear relationships well
SVC	0.50	Effective in high-dimensional spaces, potential for complex decision boundaries
LogisticRegression	0.52	Simple and interpretable, assumes a linear relationship between features and target
LinearDiscriminantAnalysis	0.50	Preserves class separability, useful for classification tasks, fast to perform and develop.
GradientBoostingClassifier	0.55	Ensemble method that combines weak learners to improve accuracy

3.1.2 Second Attempt: In the second attempt, we aimed to enhance model performance by resampling the minority class. The preprocessing steps included IterativeImputer for data imputation, RobustScaler for data scaling, and SMOTE for handling class imbalance. We employed the same models. The results are shown below in **Figure 5**. F-1 Macro for Random Forest Classifier (.55) and SVC (.54) were increased compared to the first attempt while Logistic

Regression (.34), Linear Discriminant Analysis (.42), and Gradient Boosting (.51) were dropped. Especially for Logistic Regression.

Figure 5: First Spot Check F1-Macro Scores Comparison Added PCA and Resampling.

Model	F1-macro Score	Reason for Selection
RandomForestClassifier	0.58	Robust to outliers, handles non-linear relationships well
SVC	0.53	Effective in high-dimensional spaces, potential for complex decision boundaries
LogisticRegression	0.34	Simple and interpretable, assumes a linear relationship between features and target
LinearDiscriminantAnalysis	0.42	Preserves class separability, useful for classification tasks
GradientBoostingClassifier	0.51	Ensemble method that combines weak learners to improve accuracy

Based on the information provided and considering the F1-macro scores and practical considerations, three models were selected for further development: Random Forest, Linear Discriminant Analysis (LDA), and Gradient Boosting. These models were chosen based on their performance in terms of F1-macro score, which is a suitable metric for evaluating model performance on imbalanced datasets. Additionally, the decision to exclude the Support Vector Classifier (SVC) from further development was made due to its longer training time. Therefore, focusing on Random Forest, Linear Discriminant Analysis, and Gradient Boosting allows us to allocate resources efficiently and prioritize model development efforts on algorithms that have demonstrated competitive performance and are more feasible to optimize within the project constraints.

3.2. Model Development Process

3.2.1 Random Forest Classification

In this section, we explore the development of Random Forest models to predict bankruptcy based on the provided dataset. We experimented with different configurations and preprocessing techniques to optimize the model's performance.

Model 1: Original Features

We started by training a Random Forest classifier using the original features of the dataset. The preprocessing steps are included. Random Forest model was trained using a pipeline incorporating these preprocessing steps. Hyperparameter optimization was conducted using GridSearchCV to determine the optimal number of estimators and maximum tree depth. The model achieved an F1-score of 16% for bankrupt instances, a macro F1-score of 57%, and 27 true positives.

Model 2: Combined Growth Features

In this iteration, we combined three growth-related features into one, aiming to simplify the feature space and potentially improve model performance. The preprocessing steps remained the same as in Model 1**.** Despite combining features, the model's performance remained similar, with an F1-score of 13% for bankrupt instances, a macro F1-score of 56%, and 25 true positives.

Model 3: Backward Stepwise Feature Selection

For the third model, we employed backward stepwise feature selection using Recursive Feature Elimination (RFE) with Random Forest as the base estimator. This technique selects the most relevant features, reducing the dimensionality of the dataset and potentially enhancing model interpretability and performance**.** The model achieved an F1-score of 14% for bankrupt instances with a macro F1-score of 56%, and 29 true positives.

In summary, all three models achieved similar macro F1-scores around 56-57%. However, there were slight differences in their performance regarding the F1-score for bankrupt instances and the number of true positives identified. Model 3, utilizing feature selection, showed a slight improvement in the number of true positives compared to the other models. The results were summarized in **Figure 5** below.

Figure 5: Random Forest Classifier Result Summary

Model Description	F1-macro Score	F1-score for Bankrupt Instances (1)	False Positives (FP)	False Negatives (FN)	True Positives (TP)	True Negatives (TN)
Model 1: Original Features	0.57	0.16	203	87	27	18258
Model 2: Combined Growth Features	0.56	0.13	245	89	25	18216
Model 3: Backward Stepwise Feature Selection	0.56	0.14	280	85	29	18181

3.2.2 Linear Discriminant Analysis

In this section, we delve into the exploration of Linear Discriminant Analysis (LDA) models to predict bankruptcy based on the provided dataset. Through various configurations and preprocessing techniques, we aimed to optimize the model's performance. We made a minor adjustment to the preprocessing step by adding PCA as it could significantly improve the model's performance for a linear-based algorithm.

Model 1: Original Features

We initiated the modeling process by training an LDA classifier using the original features of the dataset. The preprocessing steps mirrored those applied in the Random Forest section with PCA. The LDA model was then trained via a pipeline integrating these

preprocessing steps. Hyperparameter tuning via GridSearchCV determined the optimal number of components for PCA and the solver for LDA. The resulting model achieved an F1-score of 9% for bankrupt instances, a macro F1-score of 54%, and 27 true positives.

Model 2: Combined Growth Features

In this iteration, we consolidated three growth-related features into one, aiming to simplify the feature space and potentially enhance model performance. The preprocessing steps remained consistent with Model 1. Despite the feature combination, the model's performance demonstrated marginal improvement, yielding an F1-score of 9% for bankrupt instances, a macro F1-score of 54%, and 27 true positives..

Model 3: Backward Stepwise Feature Selection

For the third model, we employed backward stepwise feature selection using Recursive Feature Elimination (RFE) with LDA as the base estimator. This technique aimed to select the most relevant features, reducing dataset dimensionality and potentially enhancing model interpretability and performance. Despite these efforts, the model's performance closely resembled that of the previous models, achieving an F1-score of 2% for bankrupt instances, a macro F1-score of 42% and, 27 true positives.

Overall, all three models demonstrated similar macro F1-scores around 42-54%, with marginal differences in the F1-score for bankrupt instances and the number of true positives identified. Despite employing feature selection techniques, Model 3 did not show significant performance improvement compared to the other models. The summarized results are presented in **Figure 6** below.

Figure 6: Linear Discriminant Analysis Results Summary

Model Description	F1-macro-Score	F1-score for Bankrupt Instances (1)	False Positives (FP)	False Negatives (FN)	True Positives (TP)	True Negatives (TN)
Model 1: Original Features	0.54	0.09	462	87	27	17999
Model 2: Combined Growth Features	0.54	0.09	462	87	27	17999
Model 3: Backward Stepwise Feature Selection	0.42	0.02	462	87	27	17999

3.2.3 Gradient Boosting Classification

In this section, we explored the development of gradient-boosting models to predict bankruptcy based on the provided dataset. We experimented with various configurations and preprocessing techniques to optimize the models' performance.

Model 1: Original Features

We initiated the modeling process by training a gradient-boosting classifier using the original features of the dataset. The preprocessing steps mirrored those applied in the Random Forest section. Hyperparameter tuning via GridSearchCV determined the optimal number of estimators, maximum tree depth, and learning rate for the Gradient Boosting classifier. The resulting model achieved an F1-score of 25% for bankrupt instances, a macro F1-score of 62%, and 41 true positives.

Model 2: Combined Growth Features

In this iteration, we consolidated three growth-related features into one, aiming to simplify the feature space and potentially enhance model performance. The preprocessing steps remained consistent with Model 1. Despite the feature combination, the model's performance demonstrated marginal improvement, yielding an F1-score of 26% for bankrupt instances, a macro F1-score of 63%, and 48 true positives.

Model 3: Backward Stepwise Feature Selection

For the third model, we employed backward stepwise feature selection using Recursive Feature Elimination (RFE) with Gradient Boosting as the base estimator. This technique aimed to select the most relevant features, reducing dataset dimensionality and potentially enhancing model interpretability and performance. Despite these efforts, the model's performance closely resembled that of the previous models, achieving an F1-score of 25% for bankrupt instances, a macro F1-score of 62%, and 53 true positives.

To sum up, all three models demonstrated similar macro F1-scores around 62-63%, with marginal differences in the F1-score for bankrupt instances and the number of true positives identified. Despite employing feature selection techniques, Model 3 did not show significant performance improvement compared to the other models. The summarized results are presented in **Figure 7** below.

Figure 7: Gradient Boosting Classifier Results Summary

Model Description	F1-macro Score	F1-score for Bankrupt Instances (1)	False Positives (FP)	False Negatives (FN)	True Positives (TP)	True Negatives (TN)
Model 1: Original Features	0.62	0.25	174	73	41	18287
Model 2: Combined Growth Features	0.63	0.26	204	66	48	18257
Model 3: Backward Stepwise Feature Selection	0.62	0.25	256	61	53	18205

Section 4: Model Comparison

In the model comparison phase, our evaluation primarily focused on the results generated by each model. Among the models assessed, Gradient Boosting consistently outperformed others, displaying the highest F1-macro scores across all experiments. This superior performance was particularly notable in accurately identifying instances of bankruptcy, making Gradient Boosting the standout choice for predictive modeling in our context. Despite its overall success, it's essential to note the individual performance of each model. Random Forest, while robust and reliable, fell slightly short compared to Gradient Boosting in terms of F1-macro scores and identifying bankrupt instances. Linear Discriminant Analysis (LDA), on the other hand, showed respectable performance but lagged in accurately predicting bankruptcy instances, reflecting its limitations in handling complex relationships within the data.

The results underscore the efficacy of Gradient Boosting as the preferred choice for bankruptcy prediction tasks, given its consistent and superior performance metrics. However, it's crucial to acknowledge the strengths and weaknesses of each model to make informed decisions about their practical deployment in real-world financial scenarios.

Future Steps

Looking ahead, despite limitations in time, computing power, and resources, several strategies could enhance our model's performance. Firstly, more advanced feature engineering techniques could be employed to extract additional insights from the data, potentially improving the model's predictive power (**1**). Secondly, exploring alternative ensemble methods like XGBoost, LightGBM, or CatBoost could offer performance gains. These algorithms are known for their efficiency and ability to handle large datasets, potentially capturing more complex relationships in the data compared to Gradient Boosting (**2**). Lastly, optimizing model size and complexity through techniques like pruning, quantization, and model distillation could facilitate deployment in real-world settings with resource constraints (**3**). By prioritizing and implementing these strategies, we can continue to refine our bankruptcy prediction model and deliver valuable insights to stakeholders in the financial sector.

Section 5: Assignment Questions

In addressing the imbalanced nature of bankruptcy prediction datasets, we prioritized techniques like the Synthetic Minority Over-sampling Technique (SMOTE) to ensure a balanced representation of both bankrupt and non-bankrupt instances. Our analysis revealed that Gradient Boosting consistently outperformed other models, showcasing superior performance metrics even in the presence of class imbalance. Its ability to handle complex data relationships and effectively classify minority instances makes it the recommended choice for bankruptcy prediction tasks. However, computational resources and model interpretability must also be considered. Therefore, while Gradient Boosting is the primary recommendation, Random Forest emerges as a reliable alternative due to its robustness and ease of implementation. Linear Discriminant Analysis (LDA) may complement these models, especially in scenarios prioritizing interpretability over predictive accuracy.

In the financial world, accurate bankruptcy prediction models are essential for assessing risk and making informed decisions. Our recommended models, particularly Gradient Boosting and Random Forest, offer valuable insights for investors, creditors, and financial institutions. By analyzing historical financial data, these models help stakeholders identify companies at risk of bankruptcy, allowing them to adjust their strategies accordingly. For investors, these models provide predictive analytics to identify potentially risky investments and optimize portfolios. By flagging companies at high risk of bankruptcy, investors can minimize losses and maximize returns. Similarly, creditors can use these models to evaluate the creditworthiness of borrowers, reducing the likelihood of default and financial losses. Financial institutions can integrate these models into their risk management frameworks to assess portfolio stability and comply with regulations. By proactively identifying companies at risk of bankruptcy, institutions can implement preventive measures to mitigate financial risks and protect their assets.

References

(1) Corporate Bankruptcy Prediction Models: A Comparative Study for the Construction Sector in Greece. https://www.mdpi.com/2079-3197/12/1/9

(2) Estimating corporate bankruptcy forecasting models by maximizing discriminatory power. https://link.springer.com/article/10.1007/s11156-021-00995-0

(3) Bankruptcy or Success? The Effective Prediction of a Company's Financial Development Using LSTM**.** https://www.mdpi.com/2071-1050/12/18/7529