Final Project Report

Group 4

Student 1: Manthan Bhatia

Student 2: Darsh Vora

bhatia.man@northeastern.edu

vora.dar@northeastern.edu

Percentage of Effort Contributed by Student 1: 50 %

Percentage of Effort Contributed by Student 2: 50 %

Signature of Student 1: Manthan Bhatia

Signature of Student 2: Darsh Vora

Submission Date: 08/18/24

Background:

Forecasting corporate bankruptcies is a crucial responsibility in the financial sector, with major implications for various parties such as investors, creditors, workers, and regulatory agencies. Having the capacity to foresee a business's financial downfall in advance enables stakeholders to reduce risks, safeguard investments, and make well-informed strategic choices. Historically, bankruptcy prediction has relied on the analysis of financial ratios and the opinions of experts. Nevertheless, these techniques frequently find it difficult to encompass the intricate and non-linear connections among different financial indicators that come before bankruptcy. With the emergence of big data and machine learning, there is a chance to improve the precision and dependability of these forecasts by utilizing sophisticated analytical methods.

Application:

This project's goal is to create a machine learning algorithm that can anticipate the insolvency of American public companies by analyzing their past financial information. The dataset, sourced from Kaggle, covers almost twenty years (1999-2018) of financial data, including different measures like profitability ratios, leverage indicators, and market valuations. These factors are vital for evaluating a company's financial well-being and forecasting its future sustainability. Financial institutions can use the model's predictions for credit risk assessment, investors can use them for portfolio management, and regulators can use them for systemic risk monitoring.

Challenges:

To create a successful bankruptcy prediction model, various hurdles need to be overcome:

- Data Quality and Preprocessing: The dataset contains financial information from various origins, each with distinct formats and standards that could result in discrepancies. The preprocessing stage needs to address concerns such as handling missing values, scaling features, and reducing dimensionality in order to guarantee the data input into the model is precise and appropriate.
- Class Imbalance: A notable class imbalance is present in the dataset, with a much larger number of non-bankrupt companies compared to bankrupt ones. This imbalance may result in biased models that exhibit good performance on the majority class but struggle with the

minority class. Approaches like Synthetic Minority Over-sampling (SMOTE) and cost-sensitive learning are utilized to address this issue.

- Interpretability of the Model: Despite the great accuracy provided by models such as Random Forest and XGBoost, their complexity frequently makes them challenging to understand. A crucial issue that the research aims to solve with feature importance analysis and interpretability tools is making sure that the model's predictions are clear and useful for financial professionals.
- Temporal Dynamics: The financial environment is constantly evolving as a result of economic fluctuations, changes in regulations, and market forces. It may be necessary to regularly update or retrain a model based on past data to ensure its precision in forecasting future bankruptcies.

This project aims to overcome these challenges by employing robust data preprocessing techniques, advanced machine learning models, and thorough model evaluation practices to build a reliable and interpretable bankruptcy prediction tool.

Problem Definition:

The development of a trustworthy and comprehensible machine learning model to forecast the bankruptcy of publicly traded U.S. corporations based on their past financial data is the main issue this research aims to solve. Predicting corporate bankruptcy accurately can assist reduce financial risks, safeguard investments, and aid in strategic decision-making. Corporate bankruptcy has serious repercussions. The intricate nature of financial data, the underlying class disparity in bankruptcy statistics, and the requirement for precise and comprehensible models all contribute to the problem's complexity. This project seeks to address the following key issues:

1. Which financial indicators are most predictive of bankruptcy?
2. How can we effectively handle the class imbalance in the dataset?
3. Which machine learning model provides the best predictive performance?
4. How can we ensure the model's predictions are interpretable?
5. How well does the model generalize to new data and adapt to economic changes?
6. What is the impact of individual financial metrics on bankruptcy likelihood?

Data Sources and Description:

The project's dataset was obtained from Kaggle, namely the "American Companies Financial Data" dataset. This dataset comprises an extensive compilation of financial documents for publicly traded U.S. corporations spanning the years 1999 to 2018. It includes a variety of financial metrics, including market values, liquidity measurements, profitability ratios, and leverage ratios. A company's financial health and bankruptcy risk can be determined in large part by looking at these indications.

To supplement the analysis and provide context for the model development, several research papers related to bankruptcy prediction have been reviewed. These papers contribute to the understanding of different methodologies and techniques that can be applied to predict corporate bankruptcy. The following are some of the key references:

- "Bankruptcy Prediction for SMEs Using Altman Z-Score and Machine Learning Techniques": https://www.mdpi.com/1999-5903/14/8/244
- "Corporate bankruptcy prediction: A logistic regression analysis": https://www.sciencedirect.com/science/article/abs/pii/S095070511200353X
- "Machine learning-based bankruptcy prediction: A review and a new framework": https://academic.oup.com/comjnl/articleabstract/64/11/1731/5856206?redirectedFrom=fulltext
- "Early prediction of bank failures using linear discriminant analysis: A comparison with other predictors": https://www.cs.wcupa.edu/RBURNS/DataMining/papers/Santos2006.pdf
- "Bankruptcy prediction models based on financial ratios and accounting variables: A review and future research agenda": https://dergipark.org.tr/en/download/article-file/2277366
- "A comparative study of machine learning methods for bankruptcy prediction using Altman Z-Score: Evidence from publicly traded companies": https://www.aimspress.com/article/doi/10.3934/DSFE.2021010?viewType=HTML

Dataset Overview:

- **Time Period Covered:** 1999 to 2018
- **Number of Rows (Records):** 78,682
- **Number of Columns (Variables):** 21

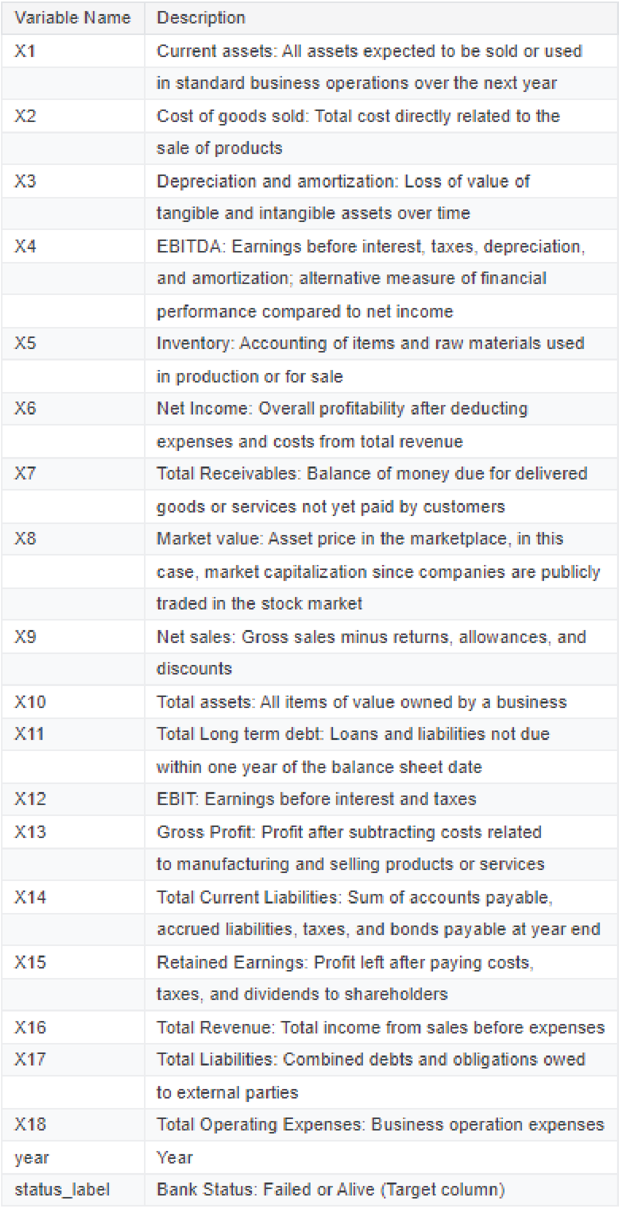
Key Variables:

The dataset consists of 21 columns, each representing different aspects of the companies' financial profiles. The variables can be broadly categorized into identification information, the target variable, and financial indicators:

1. **Identification Information:**
    - **company_name:** The unique identifier for each company (e.g., "C_1").
    - **year:** The year in which the financial data was recorded (e.g., 1999, 2000).
2. **Target Variable:**
    - **status_label:** This is the key outcome variable indicating whether the company is "alive" (still operational) or "bankrupt." This binary classification serves as the target for predictive modeling.
3. **Financial Indicators:** The remaining columns (X1 to X18) represent various financial ratios and indicators derived from the companies' financial statements. These indicators are crucial for assessing the companies' financial health and predicting the likelihood of bankruptcy. Refer Table 1 for reference.

The dataset's nearly two-decade duration enables the analysis of patterns over time and the influence of shifting economic conditions on the financial health of corporations. It captures the complex nature of corporate finance by offering a wide range of financial measurements, such as leverage and liquidity ratios. The target variable for predictive modeling is the status_label column. The objective is to categorize each company as either "alive" or "bankrupt."

Machine learning models are trained using the financial indicators (X1 to X18) as features in order to forecast each company's status_label. A thorough examination of the variables causing bankruptcy is made possible by the wide variety of financial measures and the vast number of records (78,682) that facilitate robust model training.



Variable Name	Description
X1	Current assets: All assets expected to be sold or used in standard business operations over the next year
X2	Cost of goods sold: Total cost directly related to the sale of products
X3	Depreciation and amortization: Loss of value of tangible and intangible assets over time
X4	EBITDA: Earnings before interest, taxes, depreciation, and amortization; alternative measure of financial performance compared to net income
X5	Inventory: Accounting of items and raw materials used in production or for sale
X6	Net Income: Overall profitability after deducting expenses and costs from total revenue
X7	Total Receivables: Balance of money due for delivered goods or services not yet paid by customers
X8	Market value: Asset price in the marketplace, in this case, market capitalization since companies are publicly traded in the stock market
X9	Net sales: Gross sales minus returns, allowances, and discounts
X10	Total assets: All items of value owned by a business
X11	Total Long term debt: Loans and liabilities not due within one year of the balance sheet date
X12	EBIT: Earnings before interest and taxes
X13	Gross Profit: Profit after subtracting costs related to manufacturing and selling products or services
X14	Total Current Liabilities: Sum of accounts payable, accrued liabilities, taxes, and bonds payable at year end
X15	Retained Earnings: Profit left after paying costs, taxes, and dividends to shareholders
X16	Total Revenue: Total income from sales before expenses
X17	Total Liabilities: Combined debts and obligations owed to external parties
X18	Total Operating Expenses: Business operation expenses
year	Year
status_label	Bank Status: Failed or Alive (Target column)

Data Mining Tasks:

This project involved several crucial data mining tasks, each of which produced particular numerical results that helped the bankruptcy prediction model function as a whole.

1. Feature Selection: The most significant financial indicators for bankruptcy prediction were determined by applying the Random Forest algorithm. The original collection of 21 features was reduced through this approach to a core set of the most important features. The Key Features Identified were**:** "Market Value," "Net Income," "Retained Earnings,"

"Total Liabilities," and "Depreciation & Amortization." This reduction improved the model's performance by focusing on the most predictive features, reducing overfitting, and decreasing the computational complexity.

2. Data Transformation: The StandardScaler was used to standardize all numerical features, guaranteeing that each feature has a mean of 0 and a standard deviation of 1. This transformation helped stabilize and improve the performance of models, particularly logistic regression, leading to more consistent and reliable predictions.
3. Missing Data: The dataset had no missing records or any particular value which was empty, so the problem of handling missing values was not encountered.
4. Handling Class Imbalance: The majority class in the dataset was '1' which indicated that the company is alive and on the contrary the minority class was '0' indicating that the company is bankrupt. To handle this class imbalance we could oversample the data initially in the pre-processing phase itself to make the number of minority class equal to that of the majority class but we did not perform this step because it might have led to overfitting in some of the models while training. To avoid this we first trained all the models with the imbalance itself and then oversampled the data only for the final model (random forest and XGBoost in our case) to increase the overall performance.

Data Mining Models:

When the dataset is observed thoroughly and inspected, it is evident that there is a huge class imbalance. The majority class outnumbers the minority class by a huge margin. The majority class almost captures above 90% of the dataset and the rest is filled with minority class. To train the models on an imbalanced dataset is usually not advised but since this was just an exploration stage and the final model was not chosen, we decided that we first implement all the models on the imbalanced dataset itself rather than balancing the data using SMOTE (Synthetic Minority Oversampling Technique), or any other method for that matter. This was mainly due to the following reasons,

- It becomes easy to interpret the strengths and weaknesses od each model on imbalanced data
- This also ensures that the model selection was based on natural performance of the algorithms without artificial data manipulation.

- Additionally, it minimizes the risk of overfitting to synthetic data and also makes sure that oversampling enhances the model performance rather than dictating it.

Mainly due to the above three reasons we did not oversample the minority class before model selection itself. Using the features obtained from the feature selection step we trained the models by splitting the data into 80% and 20% for training and testing. Various models were evaluated including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and XGBoost. Each model's performance was assessed using cross-validation to ensure robustness. A more detailed overview of all the model results is explained below:

1. Logistic Regression: With a very low recall (0.76%) and precision (1.35%), logistic regression scored poorly on the minority class, which is bankrupt enterprises. This suggested that there were a lot of false negatives in the model since it was unable to accurately identify failing enterprises. This rendered it unfit for our mission, in which it is essential to identify bankruptcies.
2. Decision Tree: Recall for the minority class increased to 28.45% with the Decision Tree model, but precision decreased to 26.33%. The model was less accurate than more sophisticated models like Random Forest and XGBoost since it still showed a significant percentage of false positives and false negatives.
3. Random Forest: While keeping a solid recall rate of 28.83 percent, Random Forest provided a notable improvement in precision for the minority class (79.01 percent). Out of all the models examined, its total accuracy (93.66%) was the greatest. The model was a strong option since it could continue to improve performance on the minority class while maintaining excellent precision and recall for the majority class (non-bankrupt). Random Forest performs better because of its ensemble nature, which averages several decision trees to minimize overfitting.
4. Gradient Boosting: Gradient Boosting provided a high accuracy of 92.68% together with a balanced performance. Its recall for the minority class, however, was less than Random Forest's (14.75%). Even though it had certain advantages over more basic models, it fell short of Random Forest in terms of overall efficacy, particularly when taking recall and precision into account for insolvent enterprises.

5. XGBoost: With a noteworthy improvement in recall for the minority class (49.13%) and a precision of 61.07%, XGBoost showed great performance. Especially when eliminating false negatives is important, this model was a good option because it could accurately identify over half of the failed companies. It was complementary to Random Forest in terms of precision and recall, but having a higher percentage of false positives overall.

While analyzing all the models and their results it was important to look at the minority class performance equally when compared with majority class i.e. bankrupt companies. This is because if the model is not able to predict or classify a bankrupt company as bankrupt, rather it classifies it as non-bankrupt, it can be a serious issue for investors, shareholders and the public who have invested their money into the company. So keeping all of this in mind we looked at those metrics that could better predict both the classes in terms of all the metrics.

In both classes, Random Forest offered the best balance between recall and precision. The model with the highest accuracy (93.66%) and the best F1-score (97.65%) for non-bankrupt companies demonstrated its reliability in properly predicting both groups. Additionally, the class imbalance was well-managed by the model, which improved minority class detection while accurately predicting the dominant class. The choice of Random Forest was influenced by its feature importance scores, which made the data easier to comprehend and analyze.

Among all models, XGBoost had the best recall (49.13%) for identifying bankrupt enterprises. This made it extremely useful for operations where it may be very expensive to overlook a bankruptcy (false negative). Although XGBoost's accuracy was marginally lower at 84.65%, its recall strength compensated for Random Forest's superior precision but inferior recall for the minority class.

The choice was made to combine Random Forest and XGBoost in order to take advantage of both models' advantages. As the main model, Random Forest provides stability and high overall accuracy, while XGBoost improves minority class detection and acts as a safety net to lower the likelihood of missing bankrupt enterprises. When combined, these models offer a strong solution to the bankruptcy prediction issue that strikes a balance between the crucial need of identifying at-risk businesses and the requirement for accuracy.

Implementation of Selected Models:

Tuning Random Forest and XGBoost for better results:

We made sure that the training data was split equally between the majority class (non-bankrupt) and the minority class (bankrupt), after using SMOTE to balance the dataset. The capacity of the models to identify failing companies—which are usually underrepresented—was much enhanced by this balancing. Optimizing the number of trees (n_estimators), the maximum depth of each tree (max_depth), and the splitting criteria (gini or entropy) was the main goal of the Random Forest model's training. With its default hyperparameters, the first model offered a great baseline performance, especially with regard to accuracy and resilience. But in order to optimize memory for the minority class, more adjustment was needed.

A similar methodology was used to develop XGBoost, with a focus on optimizing the number of boosting rounds (n_estimators), the learning rate (eta), and the maximum tree depth (max_depth). `Regularization parameters were tuned to control overfitting, a typical problem in boosting algorithms, using lambda (L2 regularization) and alpha (L1 regularization). To find the ideal set of hyperparameters, a random search was performed. For instance, by averaging over more decision trees, raising the n_estimators to 200 produced a more stable model, and lowering the max_depth assisted in preventing overfitting. When gini and entropy were tested as the criterion for splitting, gini performed marginally better in terms of precision and recall.

Through fine-tuning, the accuracy for bankrupt companies climbed from 75.0% to 79.01%, and the recall increased from 25.0% to 28.83%. Even though they might not seem like much, these adjustments greatly decreased the amount of false positives and negatives, increasing the model's predictability of bankruptcies. There was a discernible improvement in the balance between recall and precision as evidenced by the F1-score. The improved capacity of the model to differentiate between the two classes across various thresholds was validated by the ROC AUC score of 0.8422.

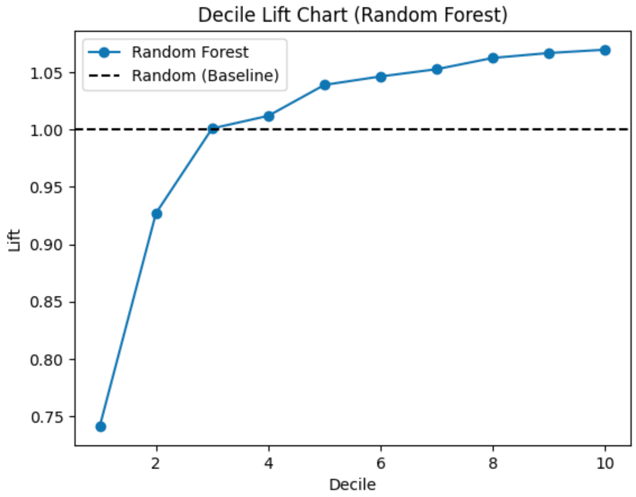
The hyperparameters of XGBoost were adjusted by combining early halting and random search methods. The model was able to learn more slowly and attentively by having the learning rate (eta) adjusted lower, which increased recall for the minority class. In order to avoid overfitting and yet detect intricate patterns in the data, the number of boosting cycles was carefully calibrated. Recall for the minority class improved significantly for XGBoost, rising from 43.0% to 49.13%. Reducing the quantity of missed bankruptcies (false negatives) required this innovation. Additionally, the accuracy for bankrupt enterprises went from 55.0% to 61.07%, contributing to a decrease in the false positive rate. For failed enterprises, the F1-score increased to 54.54%, suggesting a better balance between recall and precision. Even while it was marginally lower than Random Forest with a final ROC AUC score of 0.7811, the system performed admirably overall, especially in detecting at-risk organizations.

It was Random Forest that was indeed the best for our model but still according to us it cannot be decided between Random Forest and XGBoost for selection of the best model. To finalize one single model we chose Random Forest as it had a slight upper edge in comparison to the other.

Performance Evaluation:

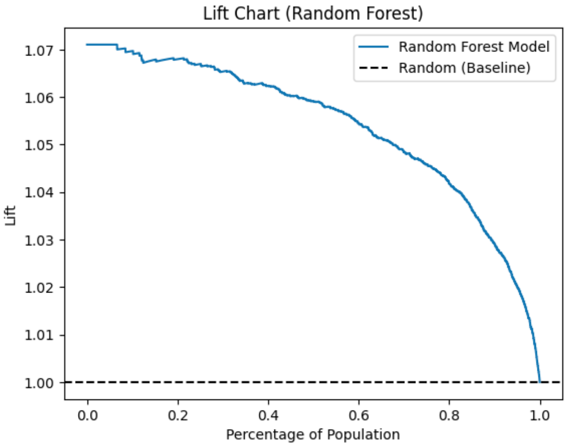
It was important to summarize the details of the model's performance in a comprehensive way so we used different metric like RMSE, MAE, AE, Lift chart to explain its performance.

1. Decile Lift Chart:



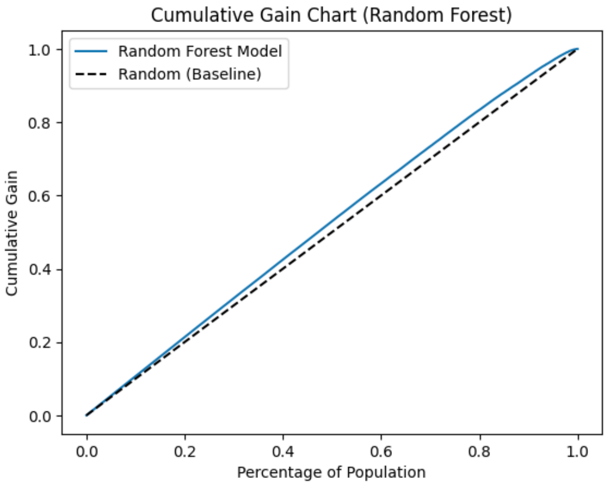
The lift across several deciles (top 10% of predictions, next 10%, and so on) is compared using the Decile Lift Chart. Lift is defined as the ratio of positive replies in the decile to the average positive response rate. It aids in assessing how well the model is able to rank cases according to their likelihood of success. For the majority of deciles, the lift is above 1.0, indicating that the model is doing better than arbitrary guesswork in predicting failing companies—especially in the top deciles. The last decile, however, shows a decline, indicating that the model's ability to differentiate between the remaining examples may be diminished. The lift should ideally stay constant or progressively decline.

2. Lift Chart



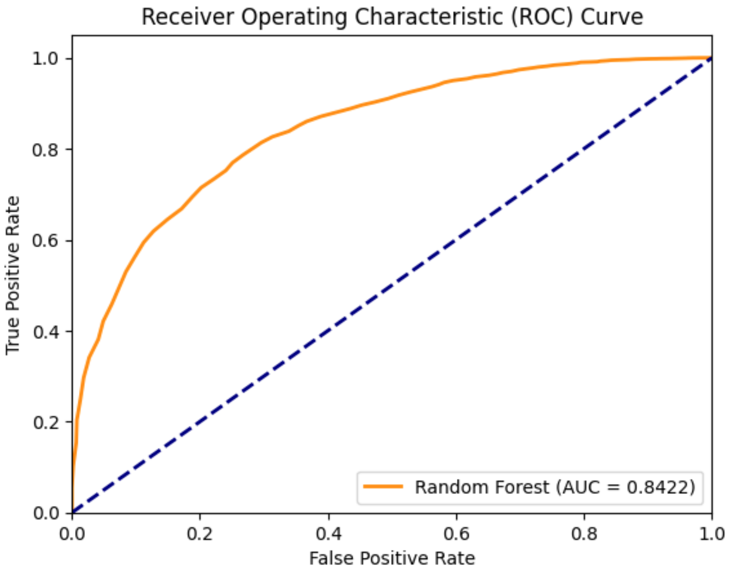
The Lift Chart calculates the relative performance of your model over population-wide random guessing. The cumulative lift is plotted against the population percentage. The Lift Chart indicates that the Random Forest model outperforms random selection when the curve remains above 1.0 for a sizable segment of the population. But as you proceed to the right, the lift drops, signifying declining returns as a larger portion of the population is considered. This is common to most models, but what's important to note is that the lift is continuously positive for the majority of the population

3. Gain Chart



When we navigate through the population sorted by the model's projected probability, the Cumulative Gain Chart displays the cumulative percentage of positives (such as insolvent enterprises) collected. The model performs better at identifying the positives the closer its curve is to the upper-left corner. Given that the blue line on the Gain Chart is near the diagonal, which represents random guessing, the model appears to be marginally more accurate than the baseline. The gain is not as significant as one might anticipate, though, which implies that there may be more room for the model to distinguish between organizations that are insolvent and those that are not.

4. Mean Absolute Error: For Random Forest, the mean absolute error (MAE) was approximately 7%, meaning that the model's predictions were generally within a reasonable range of the observed values. The model's great accuracy and capacity for exact prediction-making are shown in its low error rate.
5. Absolute Errors: Random Forest exhibited consistently low AE across most predictions, indicating that the model made few large errors. The distribution of AE was narrow, showing that most predictions were close to the actual values.
6. ROC Curve:



With a high AUC of 0.8422, Random Forest is a dependable model for broad predictions since it is very good at differentiating between classes across all thresholds. According to the model's ROC curve, it continuously outperformed the competition in terms of reducing false positives and false negatives.

Key Findings and Deliverables:

- Addressed significant class imbalance in the dataset using SMOTE, improving the models' ability to predict bankruptcies accurately.
- Discovered a trade-off between the models' recall and precision, with XGBoost performing better in recall and Random Forest performing better in precision, especially for the minority class.
- After a comprehensive study, Random Forest was determined to be the most successful models achieving the highest accuracy (93.66%) and an excellent ROC AUC score (0.8422).
- Provided thorough performance reports with metrics like recall, lift charts, gain charts, ROC AUC, F1-score, precision, and gain that allowed for a thorough assessment of the model's efficacy.
- Carried out a feature importance study, emphasizing the importance of important financial metrics for bankruptcy prediction, such as market value and net income.
- Step-by-step implementation guides for both Random Forest and XGBoost, ensuring reproducibility and ease of understanding for future users
- Recommendations for future work, including potential refinements to the models or additional data collection efforts to improve predictive accuracy.

Impact of the Project Outcomes:

- Enables early bankruptcy risk identification for businesses, giving stakeholders the opportunity to take preventative measures to reduce financial losses.
- improves decision-making by offering information on important financial metrics that are predictive of bankruptcy, which aids investors and lenders in making more accurate creditworthiness assessments.
- Enhances the distribution of resources by giving high-risk businesses priority, freeing up financial institutions to concentrate on accounts that need urgent attention.

- Provides instruments for ongoing business financial health monitoring, assisting authorities in tracking and guaranteeing financial stability. This promotes regulatory compliance.
- Assists businesses plan strategically by spotting any money problems early on, allowing them to make the necessary adjustments to prevent bankruptcy and increase long-term sustainability.
- Ensures to keep the financial markets stable and effective overall by reducing the systemic risks brought on by abrupt company bankruptcies.