**MACHINE LEARNING ASSIGNMENT**

Raghu Ram Jupudi

School of Business, St. Lawrence College, Kingston

ADMN5006: Financial Analytics

Professor Maverick Ramsaran

November 19th, 2023

# Introduction and Business Problem

In the realm of financial analytics, one of the most pressing challenges is the prediction of corporate bankruptcy. The ability to accurately forecast when a company is on the brink of financial collapse is invaluable for various stakeholders, including investors, creditors, and market regulators. The complexity of this task is heightened by the myriad of financial factors that can influence a company's trajectory towards bankruptcy. In this report, we focus on leveraging machine learning techniques to tackle this challenge, utilizing a comprehensive dataset that captures a range of financial indicators.

The core objective of this report is to develop and evaluate two distinct machine learning models – Logistic Regression and Random Forest, to predict corporate bankruptcy. The choice of these models is grounded in their proven efficacy in classification tasks and their suitability for handling complex, nonlinear relationships inherent in financial data. The specific aims of this analysis are as follows:

1. **Model Development and Comparison:**

* *Logistic Regression*: As a linear model, Logistic Regression will serve as a baseline, offering a straightforward interpretation of the relationship between financial indicators and bankruptcy risk.
* *Random Forest*: As a more complex ensemble model, Random Forest will provide a contrast to Logistic Regression in terms of handling non-linear relationships and feature interactions.

1. **Performance Evaluation and Interpretability:**

* We will assess each model's performance using metrics such as accuracy, precision, recall, and F1-score. These metrics will provide insights into the models' effectiveness in correctly classifying companies as bankrupt or non-bankrupt.
* The interpretability of the models will be discussed, examining how the features influence the prediction of bankruptcy and the practical implications of these findings.

# Exploratory Data Analysis

The dataset for this analysis comprises financial ratios and indicators of 92,872 companies, spanning from 2000 to 2023. It includes 12 features and one target variable. The features are as follows:

* *EPS (Earnings Per Share):* A measure of a company's profitability on a per-share basis.
* *Liquidity:* Working Capital divided by Total Assets.
* *Profitability:* Retained Earnings divided by Total Assets.
* *Productivity:* EBIT (Earnings Before Interest and Taxes) divided by Total Assets.
* *Leverage Ratio:* The sum of Total Long-term Debt and Debt in Current Liabilities divided by Stockholders' Equity.
* *Asset Turnover:* Sales divided by Total Assets.
* *Operational Margin:* EBIT divided by Sales.
* *Return on Equity:* Net Income divided by Stockholders' Equity.
* *Market Book Ratio:* (Price Close Annual Fiscal \* Common Shares Outstanding) divided by Book Value Per Share.
* *Asset Growth:* Year-over-year change in assets.
* *Sales Growth:* Year-over-year change in sales.
* *Employee Growth:* Year-over-year change in employee count.

The target variable is:

* *BK:* Indicator of whether a company is bankrupt or not (0 for non-bankrupt, 1 for bankrupt).

Here are some findings for the exploratory data analysis:

1. **Class Imbalance:**

The dataset is highly imbalanced with a considerable skew towards non-bankrupt cases. There are 92,314 non-bankrupt and only 558 bankrupt cases, comprising approximately 99.4% and 0.6% of the dataset, respectively. This imbalance necessitates the use of specialized techniques for balancing classes, such as SMOTE in the model training process.

1. **Outliers and Extreme Values:**

Several features exhibit extreme values, indicative of outliers. For instance, the Leverage Ratio shows a maximum value of 82,000, substantially higher than its average. To mitigate the impact of outliers on model performance winsorization is applied to the data.

1. **Missing Values:**

The dataset has missing values in several features, including Market Book Ratio, Assets Growth, Sales Growth, and Employee Growth. We can address these missing values by replacing them with the median values.

1. **Correlation Among Features:**

A correlation analysis reveals relationships among various features. Notably, Profitability and Return on Equity have a strong positive correlation, suggesting a tendency to move together. Awareness of these correlations is essential to avoid issues like multicollinearity in model development.

The following are some of the data visuals that we used for the exploratory data analysis:

* *Correlation Heatmap:* The Correlation Heatmap effectively visualizes the strength and direction of relationships between the various financial indicators in the dataset. This visualization is particularly useful for identifying how closely related different variables are to each other. For instance, a strong positive correlation indicates that as one variable increases, the other tends to increase as well, and vice versa for a negative correlation. By highlighting these interdependencies, the heatmap aids in understanding which variables might influence each other, an insight that is crucial for feature selection and model interpretability.
* *Bankruptcy Distribution (Count Plot):* The Bankruptcy Distribution Count Plot starkly illustrates the class imbalance present in the dataset, with a significant skew towards non-bankrupt cases compared to bankrupt ones. This visualization is critical as it underscores the challenge of dealing with imbalanced data in machine learning models. It highlights the necessity for employing specific strategies, such as oversampling the minority class or using specialized algorithms, to ensure that the model does not become biased towards predicting the majority class and can accurately identify the rarer event of bankruptcy.
* *Box Plot of Profitability:* The Box Plot of Profitability offers a clear depiction of the distribution of the Profitability metric across companies. It provides a visual summary of the central tendency and spread of this financial indicator, including the median, quartiles, and the presence of outliers. The plot's ability to show potential extreme values or outliers in the Profitability metric is particularly valuable. It indicates instances where companies have exceptionally high or low profitability, which are crucial for understanding the financial health and risks associated with different companies.

# Data Processing

**Handling Outliers**

The data processing began with addressing the outliers, which are extreme values that differ significantly from other observations. The chosen method for this was 'Winsorizing,' a technique that caps extreme values at specified percentiles, specifically the lower and upper 5% in this case. This approach is preferred over trimming, as it retains the data's overall structure by replacing extreme values with the nearest ones at the defined percentiles. Winsorizing was applied to various financial metrics like EPS, Liquidity, and Profitability, among others. This method effectively normalized the data distributions, making them more conducive for accurate modeling.

**Handling Missing Values**

The next critical step in data processing was managing missing values. Given the presence of nulls across different features, median imputation was employed to address this issue. Median imputation involves replacing missing values with the median of the respective feature, a robust measure less influenced by outliers. This technique ensures the preservation of the overall data distribution while filling in missing values. Applying median imputation across all features resulted in a dataset free from null values, enhancing its quality and suitability for machine learning models.

**Resulting Dataset**

Post-processing, the dataset was free from outliers and missing values. The implementation of Winsorizing and median imputation led to a more standardized and normalized dataset. Such preprocessing is essential, especially for financial data, to avoid skewed predictions and increase the robustness of bankruptcy prediction models. The refined dataset, now ready for modeling, is expected to yield more reliable and accurate results, enhancing the effectiveness of the predictive models to be developed.

# Model Development

1. **Model Selection**

Two models were selected for this analysis: Logistic Regression and Random Forest Classifier. These models were chosen for their distinct characteristics and suitability for binary classification problems.

***Logistic Regression:*** This model is known for its simplicity and interpretability. It's particularly useful in understanding the relationship between the independent variables and the binary dependent variable, making it an excellent choice for gaining insights into which factors contribute most to bankruptcy. Logistic Regression was used for its interpretability and ability to provide insights into the influence of different predictors.

***Random Forest Classifier:*** An ensemble learning method that operates by constructing multiple decision trees during training and outputting the class that is the mode of the classes of the individual trees. It is chosen for its ability to handle non-linear relationships and interactions between variables effectively, offering a more robust performance on complex datasets. Random Forest was chosen for its robustness and ability to capture complex patterns in the data, which is often essential in handling real-world financial datasets.

1. **Model Development Process**

**Logistic Regression**

*Data Preparation:* The features were first standardized using StandardScaler to ensure that all variables contribute equally to the model. The dataset was then split into training and test sets, with the training set further balanced using SMOTE to address the class imbalance issue.

*Training:* The Logistic Regression model was trained on the balanced training dataset.

**Random Forest Classifier**

*Training:* The Random Forest Classifier was also trained on the same balanced training set used for the Logistic Regression model.

**Hyperparameter Tuning for Random Forest**

As part of the model development process, hyperparameter tuning was conducted on the Random Forest Classifier using GridSearchCV. This process is crucial to optimize the model's performance by systematically varying key parameters and identifying the best combination.

*Parameter Grid:* A range of values for five key hyperparameters were defined:

* *n\_estimators:* Number of trees in the forest, with values [100, 200, 300].
* *max\_features:* Number of features to consider at every split, with options ['auto', 'sqrt'].
* *max\_depth:* Maximum number of levels in the tree, with values [10, 20, 30].
* *min\_samples\_split:* Minimum number of samples required to split a node, with values [2, 5, 10].
* *min\_samples\_leaf:* Minimum number of samples required at each leaf node, with values [1, 2, 4].

*Execution:* The grid search algorithm evaluated all possible combinations of these parameters across 3-fold cross-validation. It was set to optimize for accuracy (scoring='accuracy') and ran on multiple jobs in parallel (n\_jobs=-1) for computational efficiency.

1. **Model Results**

**Logistic Regression**

* *Confusion Matrix:* Showed 15757 true negatives, 2706 false positives, 33 false negatives, and 79 true positives.
* *Classification Report:* Indicated a high recall but low precision for the bankrupt class (1). The model could identify most of the bankrupt cases (71% recall) but had a high rate of false positives, as seen in its precision (3%).

**Random Forest Classifier**

* *Confusion Matrix:* Showed 18305 true negatives, 158 false positives, 101 false negatives, and 11 true positives.
* *Classification Report:* Revealed a high accuracy and precision for the non-bankrupt class (0) but low precision and recall for the bankrupt class (1). The model was highly accurate in identifying non-bankrupt cases but struggled with the minority bankrupt class.

**Grid Search**

* *Best Parameters:* The best-performing model used the parameters: max\_depth of 30, max\_features set to 'sqrt', min\_samples\_leaf of 1, min\_samples\_split of 2, and n\_estimators of 300.
* *Best Score:* The highest accuracy achieved in the cross-validation process was approximately 99.36%

# Model Comparison

The evaluation and comparison of the Logistic Regression and Random Forest models on the bankruptcy prediction dataset reveal distinct performance characteristics of each model.

**Logistic Regression Model:**

* *High Recall, Low Precision:* The Logistic Regression model demonstrated high recall (71%) but low precision (3%) for the bankrupt class. This indicates that while the model is quite effective at identifying most of the bankrupt cases, it also incorrectly labels many non-bankrupt cases as bankrupt (high false positives).
* *Interpretability:* A significant advantage of the Logistic Regression model is its interpretability. The model coefficients can provide insights into how each feature influences the likelihood of bankruptcy.
* *Suitability:* This model is more suitable when the cost of missing a bankrupt case (false negative) is higher than incorrectly identifying a non-bankrupt case as bankrupt (false positive). It's ideal for scenarios where identifying potential bankruptcy risks is prioritized, even at the expense of accuracy.

**Random Forest Classifier:**

* *Overall Accuracy and Precision for Non-Bankrupt Class:* The Random Forest model showed high accuracy (99%) and precision (99%) for the non-bankrupt class but struggled with the bankrupt class, achieving only 10% recall and 7% precision.
* *Handling Complex Interactions:* The model's ability to handle non-linear relationships and feature interactions is a significant advantage, especially in datasets with complex patterns.
* *Suitability:* This model is more appropriate in scenarios where accurately identifying the majority class is crucial, and the cost of false positives is relatively high.

**Comparative Insights**

* *Trade-off Between Recall and Precision:* The Logistic Regression model's high recall rate makes it valuable for identifying most bankrupt cases, but its low precision leads to a high number of false alarms. On the other hand, the Random Forest model, with its high overall accuracy, is more reliable in predicting non-bankrupt cases but misses a significant portion of the bankrupt ones.
* *Class Imbalance Impact:* The difference in performance metrics reflects the challenges inherent in modeling highly imbalanced datasets. While SMOTE was used to balance the classes, the models' responses to the minority class (bankrupt cases) vary considerably.
* *Model Selection Considerations:* Choosing between these models depends on the specific requirements and constraints of the application. If missing a bankrupt prediction has severe consequences, the Logistic Regression model might be preferable despite its lower precision. However, if avoiding false positives is critical, the Random Forest model would be a better choice.

# Conclusion

**Practical Applications in the Financial World**

The models developed in this study, Logistic Regression and Random Forest, offer valuable tools for various applications in the financial sector. Their practical implementation can be tailored to different aspects of financial decision-making and risk management.

* *Credit Risk Assessment:* Financial institutions can utilize these models to evaluate the bankruptcy risk of companies when considering lending or extending credit. The Logistic Regression model, with its high recall for bankrupt cases, can be particularly useful in early warning systems where identifying potential defaulters is crucial.
* *Investment Decisions:* Investors can leverage these models to assess the financial health of potential investment targets. The Random Forest model, with its high overall accuracy, can aid in creating a risk profile of companies, helping investors to make informed decisions.
* *Regulatory Compliance and Monitoring*: Regulatory bodies can use these models for monitoring the financial stability of companies, especially in sectors prone to high volatility. The models can aid in identifying firms at risk of bankruptcy, enabling timely intervention or regulatory actions.
* *Portfolio Management:* Asset managers and financial advisors can integrate these models into their portfolio management strategies, using them to mitigate risks by avoiding companies with high bankruptcy probability.

**Future steps**

To enhance the performance of your Logistic Regression and Random Forest models, considering the constraints of time, computational power, and resources, the following techniques and tools could be beneficial:

* *Advanced Ensemble Techniques:* Implement Gradient Boosting Models like XGBoost or LightGBM for better handling of imbalanced datasets and efficiency.
* *Deep Learning Approaches:* If resources allow, experiment with Neural Networks to capture complex patterns in data more effectively.
* *Improved Data Preprocessing:* Engage in advanced feature engineering and consider dimensionality reduction techniques like PCA to simplify the feature space.
* *Optimization and Automation:* Utilize Bayesian optimization for more efficient hyperparameter tuning and explore AutoML platforms to automate various steps of the machine learning process.
* *Handling Class Imbalance:* Implement cost-sensitive learning to focus on the cost of misclassification and explore alternative resampling techniques like Borderline-SMOTE or ADASYN.
* *Incorporating External Data:* Consider adding macroeconomic indicators to provide additional context to your models, potentially improving their predictive accuracy.