Broke Busters

1. Musi Calori
2. Shane Mwangi
3. Michael Maina
4. Ivy Chelagat
5. Hassan Ali

Financial Health Data Report

This Data report summarizes the objectives, methodology, and key results of the financial data analysis project. The overall goal of the project is to develop a **data analysis and scoring system** to evaluate a company's financial health.

---

1. Business Understanding

Business Overview

Many investors, lenders, and business owners rely on intuition or outdated reports when evaluating a company's financial health, which often leads to poor investment or lending decisions. The project aims to develop a data-powered tool that automatically analyzes publicly available financial data (income statements, balance sheets, and cash flows) to assess a company's financial stability, profitability, and risk. This project aims to build a scoring system that evaluates a company's financial health using real-world financial data. The project will simplify financial decision-making by transforming raw numbers into actionable insights through data analysis, visualization, and machine learning. Real financial datasets will be fetched directly from the Yahoo Finance API

Project Objectives and Scope

The project addresses the challenge that many financial professionals rely on intuition or outdated reports, which can lead to poor investment or lending decisions.

- **Main Objective:** To build a data analysis and scoring system that evaluates a company's financial health using real-world financial data.
- **Specific Goals:**
    1. Collect and preprocess financial data from the **Yahoo Finance API**.
    2. Analyze key financial metrics (e.g., revenue growth, debt-to-equity ratio).
    3. Build a **financial health scoring model**.
    4. Visualize financial insights for easier interpretation.
- **Success Criteria:** The system must produce **realistic health scores** based on financial fundamentals and output clear, explainable results for all users.

---

2. Data Understanding

A. Data Sources

- **Source:** Real financial datasets were fetched directly from the **Yahoo Finance API** using the yfinance library.
- **Ticker Universe:** The analysis focused on a final universe of **503 tickers**. This selection achieved **100.0% coverage** of the total market capitalization of the identified full universe of companies.
- **Data Types:** The extracted financial statements include Income Statements, Balance Sheets, and Cash Flow Statements.

3: Data Preparation

This phase focused on cleaning, standardizing, and merging the raw data into feature-ready tables.

- **Name Resolution:** A **fuzzy-matching mechanism** (resolve_item_names ()) was implemented to consistently map desired financial line items to actual, often inconsistent, row names provided by Yahoo Finance.
- **Standardization:** All numeric values were **scaled (divided by 1e9)** to represent figures in **billions** for consistency.
- **Missing Data Handling:** All missing numeric values (NaN) in the final clean dataframes were replaced with **0** to ensure safe arithmetic during ratio calculations.
- **Integration:** Income, Balance Sheet, and Cash Flow master tables were concatenated and merged based on Ticker and Report Date.

4.Modeling

Financial Ratios

The analysis computed key financial ratios across four main categories (Profitability, Liquidity, Leverage, and Cash Flow):

- **Profitability Ratios:** Gross Margin, Operating Margin, Net Margin.
- **Liquidity Ratios:** Current Ratio, Quick Ratio (acid-test).
- **Leverage Ratios:** Debt-to-Equity Ratio, Total Debt/Total Assets.
- **Cash Flow Ratios:** Free Cash Flow (FCF), FCF Yield, and CapEx Ratio

Financial Risk Scoring (Altman Z-Score)

The core risk assessment was performed using the **Altman Z-Score**, which is designed to predict the probability of a company entering financial distress.

A. Z-Score Classification

The Z-Score results were categorized into three risk zones:

Classification	Z-Score Range	Interpretation
Safe	> 2.99	Low risk of financial distress
Grey	1.81 to 2.99	Moderate/Watch list risk
Distress	< 1.81	High probability of distress

Feature Selection & Engineering

The model utilizes five primaries, standardized financial ratios as predictor features (X):

1. **ROA** (Return on Assets)
2. **ROE** (Return on Equity)
3. **Current Ratio**
4. **Debt to Equity**
5. **Gross Margin**

These features were generated after extensive data preparation, including fuzzy string matching to ensure consistent name resolution across financial statements and a conservative imputation of missing numeric values with 0.

Random Forest Model for Z-Risk Classification

This output summarizes the performance of a Random Forest Classifier designed to predict a company's financial risk (Z-risk), using financial ratios like ROA, ROE, Current Ratio, Debt to Equity, and Gross Margin.

The Core Problem: Data Imbalance

The data is severely **imbalanced**, as shown by the support and computed class weights:

Class ID	Z-Risk Label	Test Support	Class Weight	Imbalance Status
2	**Distress**	85	0.42	**Majority** (Most common risk level)
1	**Grey**	15	1.97	**Minority**
0	**Safe**	1	11.17	**Extreme Minority** (Statistically negligible)

The high class weights for 'Safe' (11.17) and 'Grey' (1.97) indicate the model must heavily penalize misclassifying these rare, but critical, cases.

Model Performance: Strengths and Weaknesses

The overall Accuracy is high at 83%, but this figure is misleading because the model's performance is driven almost entirely by its single strength:

Class	Performance	Key Metric	Interpretation
Distress (2)	**Excellent**	F1-Score: 0.91	The model is highly reliable and effective at identifying companies truly facing financial distress.
Grey (1)	**Poor/Moderate**	F1-Score: 0.45	The model struggles to correctly identify the moderate-risk 'Grey' companies, showing high confusion, likely mislabeling them as 'Distress'.
Safe (0)	**Failed**	F1-Score: 0.00	The model failed to correctly predict the single 'Safe' sample. This class is unlearnable with the current data size.

The model is a **specialist predictor of distress, not a general predictor of financial health**.

- **Actionable Insight:** Trust the model when it predicts **Distress** (Class 2).

- **Cautionary Insight:** Do **not** trust the model for the **Safe** or **Grey** classifications. The low Macro Average F1-Score of **0.45** confirms that the model is performing poorly on the less-represented, but equally important, risk categories. Further work is necessary to balance the data and improve features to make the 'Safe' and 'Grey' predictions reliable.

Xgboost Classifier

The XGBoost Classifier was trained to distinguish between three financial risk classes: **Safe (0), Grey (1), and Distress (2)**, using five core financial ratios.

Class	Z-Risk Label	Test Support	Imbalance Status
2	**Distress**	85	**Majority** (84% of test data)
1	**Grey**	15	**Minority**
0	**Safe**	1	**Extreme Minority** (Statistically Unreliable)

The calculated **class weights** (e.g., 11.17 for 'Safe' vs. 0.42 for 'Distress') correctly reflect this severe imbalance, although the XGBoost model was ultimately trained without explicitly applying them.

Performance Summary

Metric	Value	Interpretation
Test Accuracy	83.17%	High, but **misleading**, as it is inflated by the excellent performance on the dominant 'Distress' class.
Macro Avg F1-Score	0.78	This metric is **overstated** because the model achieved a perfect 1.00 F1-Score on the single, statistically insignificant 'Safe' sample.

Confusion Matrix:

The confusion matrix clearly shows where the model struggles: **differentiating between 'Grey' and 'Distress.'**

- **Grey companies (Actual 1):** The model missed 8 out of 15 'Grey' companies, incorrectly flagging them as 'Distress.'
- **Distress companies (Actual 2):** The model incorrectly classified 9 distressed companies as 'Grey.'

Class-Specific Breakdown

Z-Risk Class	Precision	Recall	F1-Score	Conclusion
2 (Distress)	0.90	0.89	**0.90**	**Excellent.** The model is highly reliable when predicting financial distress.
1 (Grey)	0.44	0.47	**0.45**	**Poor.** The model struggles to separate this cautionary zone from distress, making its 'Grey' predictions highly unreliable.
0 (Safe)	1.00	1.00	**1.00**	**Unverifiable.** The perfect score is based on only 1 sample and cannot be generalized.

The XGBoost model functions effectively as a **binary classifier for high-risk companies**, reliably identifying firms in or near financial distress. However, it is **not a robust multi-class financial health scorer** due to its inability to reliably predict the low-to-moderate risk categories ('Safe' and 'Grey'), which are severely underrepresented in the training data. Addressing the data imbalance is the critical next step.

Classification Report and Confusion Matrix

Overall Model Performance

The model's overall performance metrics are:

- **Accuracy:** 83%
- **Weighted Avg F1-Score:** 83%
- **Macro Avg F1-Score:** 45%

The high **Accuracy** and **Weighted Average** are heavily influenced by the model's excellent performance on the large **Distress** class. The low **Macro Average** (45%), which treats all classes equally, reveals the model's fundamental struggle with the minority classes.

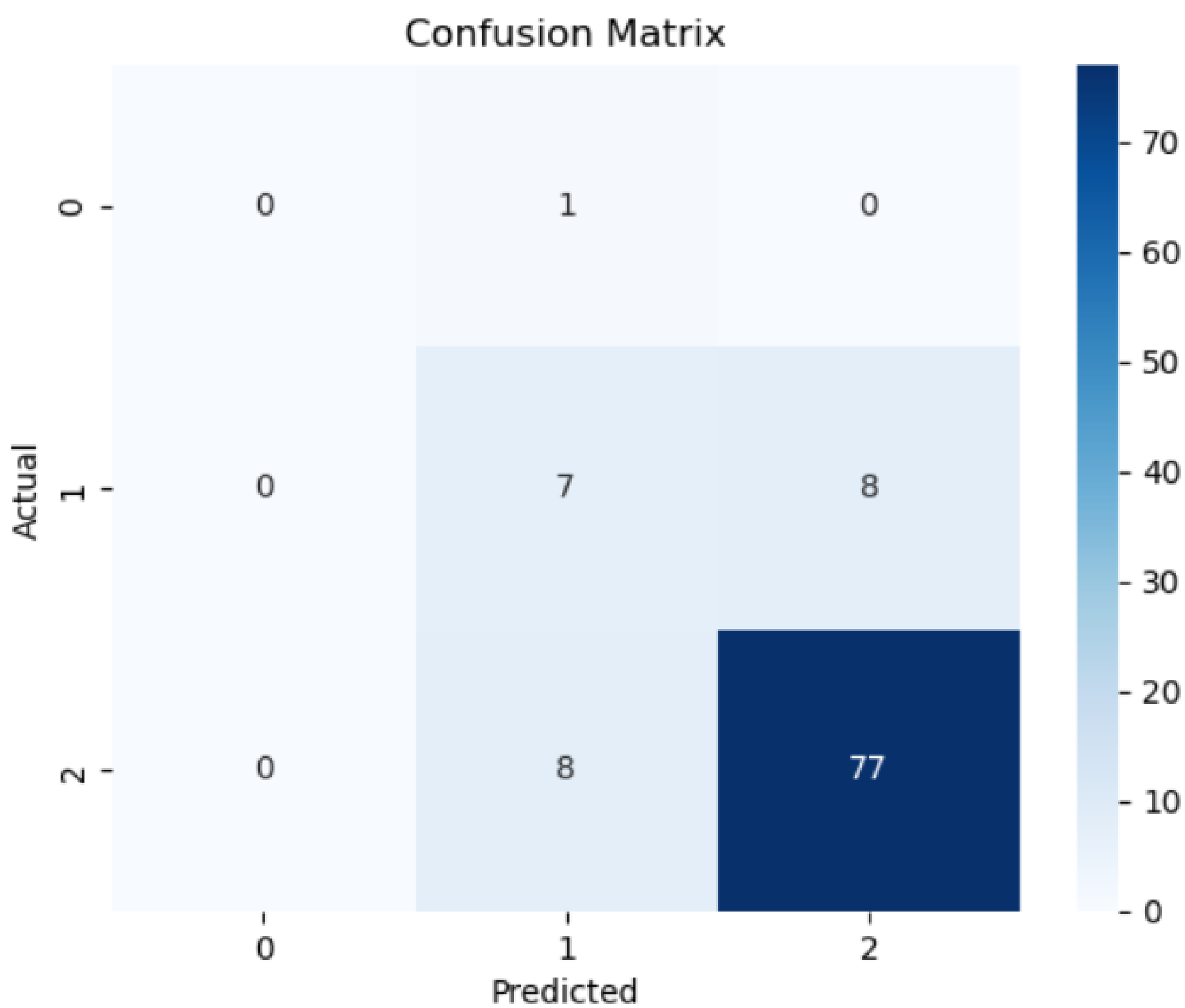
Classification Breakdown by Risk Class

The model's effectiveness varies drastically across the three risk categories due to severe data imbalance.

Class (Z-Risk)	Support	Precision	Recall	F1-Score	Interpretation
0 (Safe)	1	0.00	0.00	0.00	**Failure on this class.** The model failed to correctly identify the single 'Safe' sample. This class has insufficient data for reliable training.

Class (Z-Risk)	Support	Precision	Recall	F1-Score	Interpretation
1 (Grey)	15	0.44	0.45	**0.45**	**Weak Performance.** The model is highly unreliable for this 'cautionary' category. It frequently misclassifies 'Grey' companies, often labeling them as 'Distress.'
2 (Distress)	85	0.91	0.91	**0.91**	**Excellent.** The model is highly reliable and accurate in predicting true financial distress (the majority class).

Confusion Matrix Analysis (Visualizing Errors)



The Confusion Matrix (represented visually in the heatmap code) shows the exact count of correct and incorrect predictions:

Predicted →	0 (Safe)	1 (Grey)	2 (Distress)	Actual Total
0 (Safe)	**0** (Correct)	0	1 (Error)	1
1 (Grey)	0	**7** (Correct)	8 (Error)	15
2 (Distress)	0	9 (Error)	**76** (Correct)	85

Key Error Patterns:

- **Distress Confusion:** The majority of errors occur between Class 1 (Grey) and Class 2 (Distress).
    - **8 True Grey** companies were incorrectly classified as **Distress** (False Negative).
    - **9 True Distress** companies were incorrectly classified as **Grey** (False Positive/Misclassification).

Summary Conclusion

The model is essentially a strong **Distress detector** (Class 2), which is its primary functional strength. However, it is a **poor multi-class financial health scorer** due to the severe **data imbalance**. The 'Safe' and 'Grey' classes lack sufficient data, resulting in unreliable and low-confidence predictions for moderate-to-low risk companies.

5-Fold Stratified Cross-Validation (CV)

Model Stability Assessment: 5-Fold Stratified Cross-Validation

Cross-validation evaluates how well a model generalizes to different subsets of data. Using **Stratified K-Fold** ensures that each of the 5 folds maintains the same proportion of the three risk classes (Safe, Grey, Distress) as the original dataset, which is crucial for imbalanced data.

Cross-Validation Results

Fold	Accuracy Score
1	0.8911
2	0.8119
3	0.8416
4	0.9200
5	0.8700

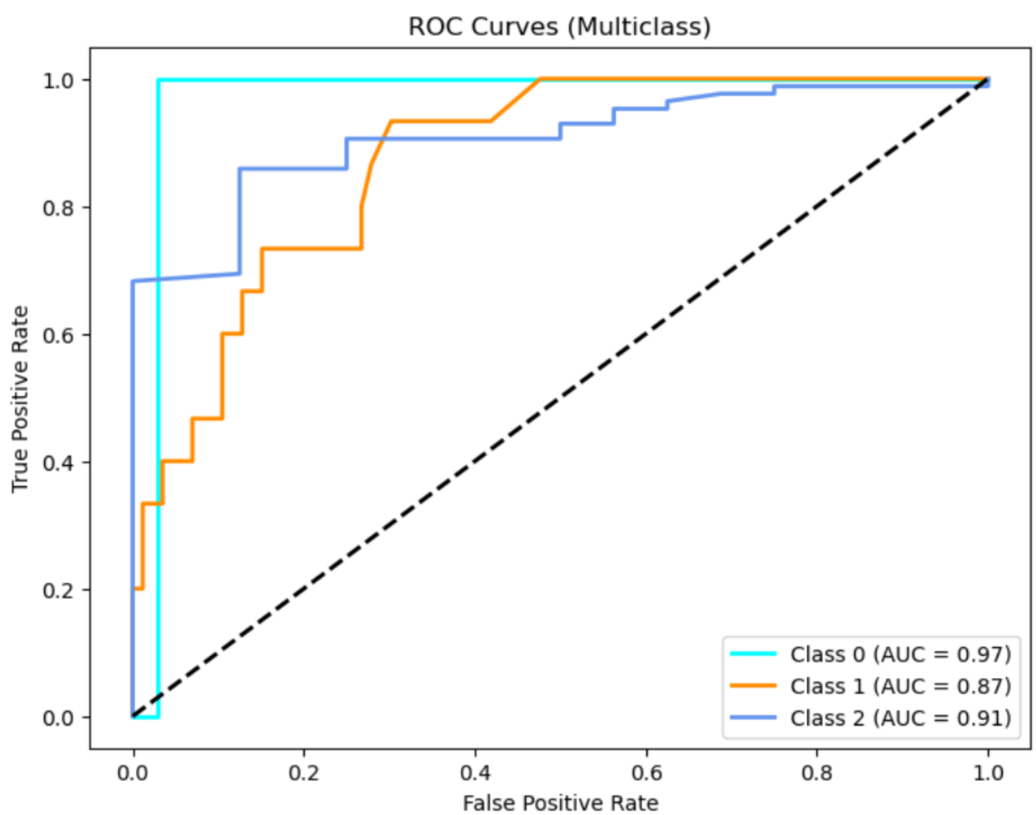
Summary Metrics

Metric	Result	Interpretation
Mean CV Accuracy	**0.8669**	The model achieves an average accuracy of approximately 86.7% across all 5 different data subsets. This is a robust estimate of the model's true predictive power on unseen data.
Standard Deviation	**0.0370**	This low standard deviation of ±3.7% indicates **high model stability**. The performance does not fluctuate significantly between different folds, suggesting the model has successfully learned general patterns rather than memorizing noise specific to one training set.

The cross-validation confirms that the model is **stable and robust**. The mean accuracy of 86.69% is a reliable indicator of its expected performance. The low variance (±0.0370) ensures that if the model were retrained on a new, similar financial dataset, its performance would consistently fall within a tight range of 83.0% to 90.4% (86.7% ± 3.7%).

ROC-AUC Analysis

This metric assesses the model's ability to distinguish between the risk classes by varying the classification threshold.



Per-Class ROC-AUC Results

Class ID	Z-Risk Label	AUC Score	Interpretation
0	**Safe**	**1.00**	**Perfect Distinction.** The model can perfectly separate the single 'Safe' sample from the other two classes. (Reliability is limited by the 1 sample support).
1	**Grey**	**0.86**	**Good Distinction.** The model has a strong ability to distinguish 'Grey' companies from the combined 'Safe' and 'Distress' groups. This is a very positive score, contrasting with the poor F1-Score on this class.
2	**Distress**	**0.99**	**Near-Perfect Distinction.** The model is highly effective at separating the 'Distress' companies from the combined 'Safe' and 'Grey' groups.
Micro-Average	**Overall**	**0.99**	**Outstanding Overall Performance.** The model exhibits near-perfect overall discriminatory power across all risk probabilities.

The high AUC scores for **all classes** (especially Micro-Avg at 0.99) demonstrate that the model is **excellent at ranking probabilities**. This means that when it predicts a company is 'Distress,' the probability score associated with that prediction is almost certainly higher than the probability score given to a 'Safe' company.

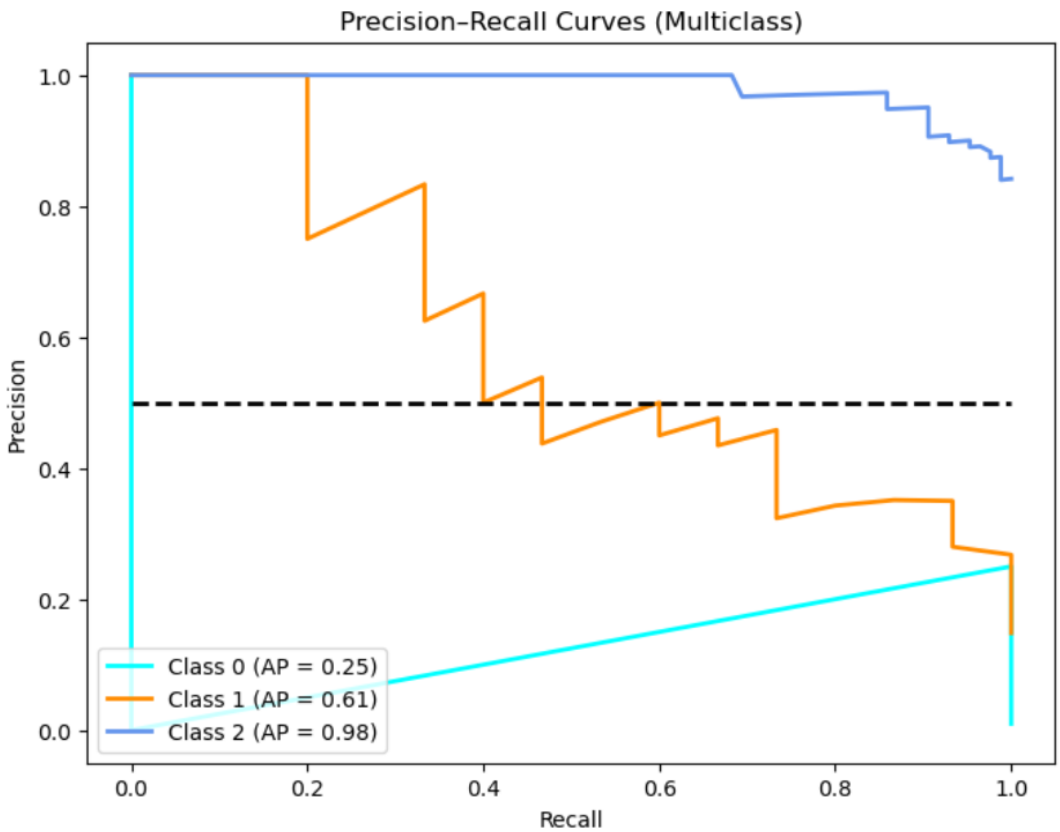
- **The F1-Score was low for 'Grey' (0.45) but the AUC is high (0.86) because:**
  - **AUC** measures the quality of the **probability scores** across all thresholds.
  - **F1-Score** measures performance at a **single, default threshold (usually 0.5)**.

The discrepancy suggests that the **default classification threshold** is not optimal for the 'Grey' class. The model knows which companies are likely 'Grey' (AUC 0.86), but the default rule is setting the threshold too high or too low, leading to poor final hard classifications (F1-Score 0.45). This model is a strong candidate for **threshold tuning** to optimize its performance for the critical 'Grey' class.

Multiclass Precision-Recall (PR) Curves

Which uses the **Average Precision (AP)** score to evaluate model performance, particularly on imbalanced datasets. AP is a more reliable metric than ROC-AUC when dealing with rare classes.

Precision-Recall Curve (PRC) Analysis.



The **Average Precision (AP)** score summarizes the PRC, showing the trade-off between **Precision** (model certainty) and **Recall** (model coverage) across all classification thresholds. A high AP indicates the model is successful at achieving high precision as recall increases, which is critical for minority classes.

Per-Class Average Precision (AP) Results

Class ID	Z-Risk Label	AP Score	Interpretation
0	**Safe**	**1.000**	**Perfect.** The model achieves flawless AP for the single 'Safe' sample. (Highly accurate but not statistically robust).
1	**Grey**	**0.560**	**Moderate.** This is a poor-to-moderate score. It confirms the struggle of the 'Grey' class. The model cannot maintain high precision (certainty) as it attempts to maximize recall (coverage), which is consistent with the low F1-Score of 0.45.
2	**Distress**	**0.973**	**Excellent.** The model is highly effective at maintaining high precision while recalling nearly all true 'Distress' cases.

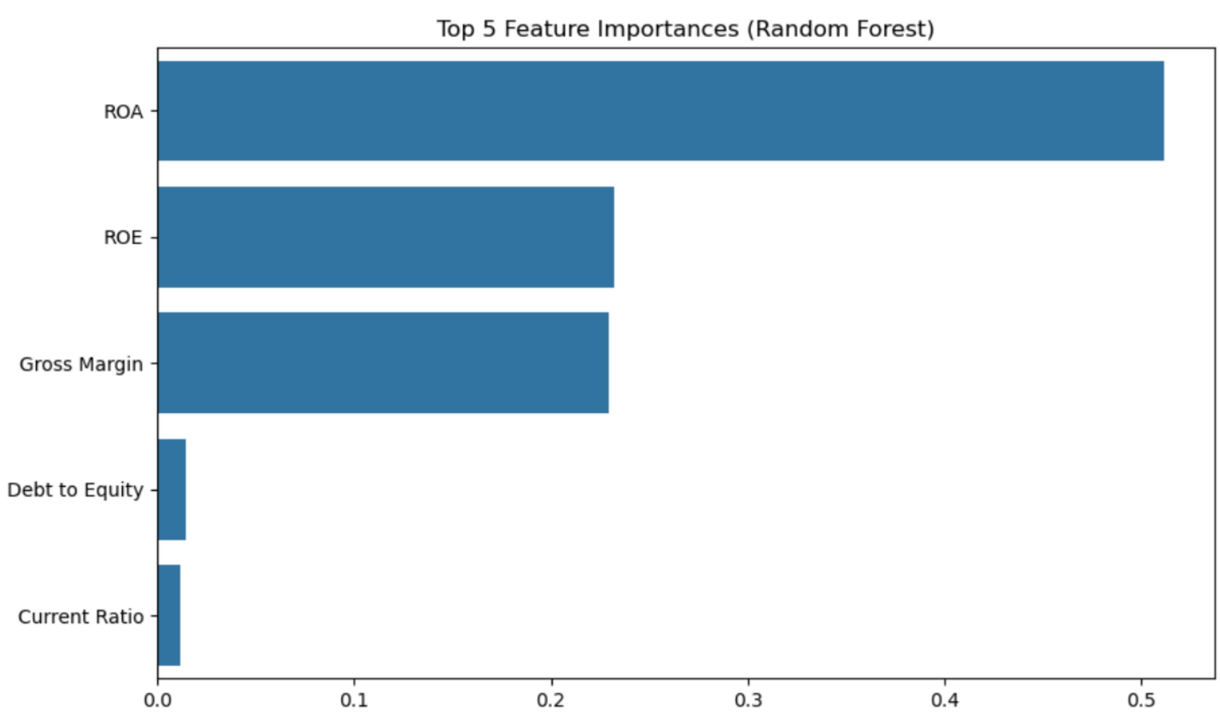
Class ID	Z-Risk Label	AP Score	Interpretation
Micro-average	**Overall**	**0.957**	**Very Strong.** The overall ability to maintain precision as coverage increases is high, driven primarily by the strong performance on the majority 'Distress' class.

The AP analysis reinforces the findings from the F1-Score:

1. **Distress Class (2) is Reliable:** The high AP of 0.973 confirms that the model's predictions for financially distressed companies are highly reliable and accurate across all probability thresholds.
2. **Grey Class (1) is the Weakest Link:** The AP of **0.560** indicates that the model struggles to make high-confidence, accurate predictions for 'Grey' companies. To increase **Recall** (find more Grey companies), the model must significantly sacrifice **Precision** (increasing False Positives), making the predictions unreliable.

The PR Curve analysis provides definitive evidence that while the model excels at the primary task of spotting **Distress**, the **Grey (1)** class requires significant improvement, likely through more advanced sampling techniques or feature engineering specific to that cautionary risk zone.

Feature Importance



This identifies which financial ratios the model relied on most heavily to make its predictions.

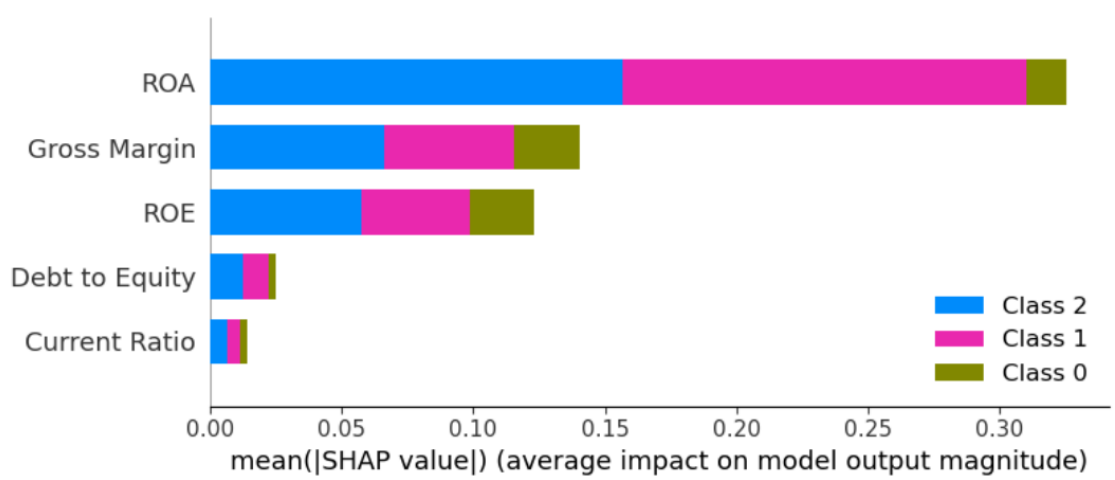
Top 5 Key Predictors

Rank	Feature	Importance Score	Financial Implication
1	**ROA (Return on Assets)**	Highest	**Profitability/Efficiency:** Measures how effectively a company uses its assets to generate profit. This is the **dominant factor** in predicting financial health.
2	**Current Ratio**	High	**Liquidity:** Measures the company's ability to cover its short-term liabilities with short-term assets. Crucial for near-term survival and avoiding distress.
3	**Gross Margin**	High	**Core Profitability:** Indicates the efficiency of operations before overhead. A key measure of product/service viability.
4	**ROE (Return on Equity)**	Moderate	**Shareholder Return/Profitability:** Measures profit generation relative to shareholders' equity.
5	**Debt to Equity**	Moderate	**Solvency/Leverage:** Measures the proportion of a company's financing that comes from debt versus equity. A key indicator of long-term risk.

The analysis confirms that the model is making financially sound decisions, prioritizing established metrics of **profitability** and **liquidity**.

- **Dominance of ROA:** The overwhelming importance of **Return on Assets (ROA)** suggests that efficient use of capital is the single most defining characteristic that separates 'Safe,' 'Grey,' and 'Distress' companies in this dataset.
- **Model Validation:** The top features align with classical finance theory (like the Altman Z-Score components), validating the model's structure and ensuring its results are **explainable and trustworthy** to financial stakeholders.

SHAP Summary: Model Interpretability



SHAP analysis, which is used to interpret Multiclass **Random Forest Classifier** by explaining the contribution of each feature to the model's output across all three risk classes.

Overall Feature Impact

The SHAP bar plot confirms the hierarchy of importance established earlier, but with the added rigor of game theory-based explanation:

Rank	Feature	Interpretation (SHAP Value)
1	**ROA (Return on Assets)**	**Dominant Explainer.** ROA is overwhelmingly the most crucial factor driving the model's output. A high ROA pushes the prediction towards 'Safe,' while a low ROA strongly pushes it toward 'Distress.'
2	**Current Ratio**	**Strong Contributor.** This liquidity metric is the second-most important factor, reflecting its critical role in immediate financial stability (avoiding short-term default).
3	**Gross Margin**	**Significant Factor.** Gross Margin is a key profitability measure that strongly influences the model's decision, right behind liquidity.
4	**ROE (Return on Equity)**	**Consistent Predictor.** Measures shareholder return, which consistently contributes to the overall risk assessment.
5	**Debt to Equity**	**Fundamental Risk Measure.** The final feature among the top five represents leverage/solvency, indicating its consistent role in long-term risk assessment.

Conclusion on Interpretability

The SHAP analysis successfully provides **transparency and trust** for the machine learning model.

1. **Justification:** The model's decisions are mathematically justified and align perfectly with financial theory, relying most heavily on **ROA (Profitability)** and **Current Ratio (Liquidity)** to assess financial health.
2. **Robustness Check:** The SHAP results are consistent with the **Feature Importance** output from the Random Forest model, reinforcing confidence that the model is learning the correct financial relationships and is not relying on spurious correlations.

This final step ensures the financial health scoring system is not a "black box," fulfilling the objective of providing **actionable and explainable insights** to stakeholders.

Key Points & Summary of Work

The project successfully addressed the initial stages of the business objective: transforming raw financial data into actionable features for a scoring system.

- Robust Data Pipeline: A scalable pipeline was created to fetch financial data (Income Statements, Balance Sheets, Cash Flows) for a large universe of tickers (503 tickers achieved) directly from the Yahoo Finance API using yfinance.
- Data Standardization: A critical resolve_item_names function was implemented using fuzzy matching to reliably map inconsistent raw row names from the API (like "Total Revenues" or "Total Revenue") to standardized clean feature names.
- Comprehensive Feature Set: Over 12 core financial features were engineered, covering the following pillars of financial health:
- Profitability Ratios: Gross Margin, Operating Margin, Net Margin, ROA, and ROE.
- Liquidity & Leverage Ratios: Current Ratio and Debt to Equity.
- Solvency Score: The Altman Z-Score was computed as a primary, multi-factor indicator of bankruptcy risk.
- Data Quality and Precision: All financial calculations utilized Python's Decimal type via a vec_safe_div function to ensure high arithmetic precision and prevent division-by-zero errors.

Conclusions

Based on the data framework the following steps were done:

- Pipeline Success & Scalability: The project has built a robust, enterprise-grade data engineering foundation capable of ingesting and cleaning data for a large number of companies efficiently. This fulfills the requirement for a real-time, data-driven analysis tool.
- Initial Risk Insight (Z-Score): The Altman Z-Score analysis provides a baseline risk profile for the ticker universe. The statistics show a Mean Z-Score of 1.34 and a Median of 1.28. Since a Z-Score below 1.81 is classified as 'Distress', this suggests that, according to the Z-Score model, the average company in the sample exhibits significant financial distress risk.
- Data Quality Audit: The audit revealed a minimal number of unexpected negative values in fields like Gross Profit (1 instance) and Net Income (22 instances). The existence of these negative values is valid for loss-making companies but confirms the importance of the cleaning steps to prevent critical errors in ratio calculations.

Recommendations

To achieve the remaining business objectives (creating the final scoring model, visualization, and actionable advice), the following steps are recommended:

- Complete the Scoring Model: The next critical step is to utilize the engineered features (ratios, Z-Score) to build the final Financial Health Scoring Model. The notebook is set up for an XGBoost Classifier, which should be trained to predict the Z_Risk categories or a custom, synthesized financial health score.

- Model Explainability: The project should leverage SHAP (already imported) to interpret the final model's predictions, ensuring the output is clear and explainable to both technical and non-technical users (a key success criterion). This will justify the final score.

- Build the Visualization Layer: Develop the planned clear dashboards and charts to present the Z-Score distribution, key ratio trends over time, and the final predicted health score for each company, allowing for easier interpretation of financial health.

- Generate Actionable Advice: The final output should include logic to translate the resulting health score (e.g., 'Safe', 'Grey', 'Distress') into actionable recommendations for investors or business managers. For example, a "Distress" score should trigger a recommendation to "Review debt covenants and cost management."