StockWell

Measuring the Financial Wellness of the S&P 500

KWK Machine Learning x Finance Challenge

By: Hailey Muñiz

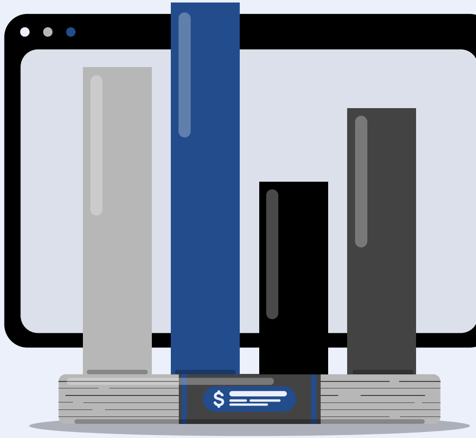
Problem Statement

Investors, students, and the public often struggle to understand if a company is financially "healthy" or stable from raw stock data. Traditional financial indicators can be confusing or inaccessible, and existing scoring systems are proprietary or not accessible to the public.

Can we use machine learning to predict a company's near-future financial stability using key financial indicators like EBITDA, revenue growth, market capitalization, and current stock price?

01

Methodology Documentation

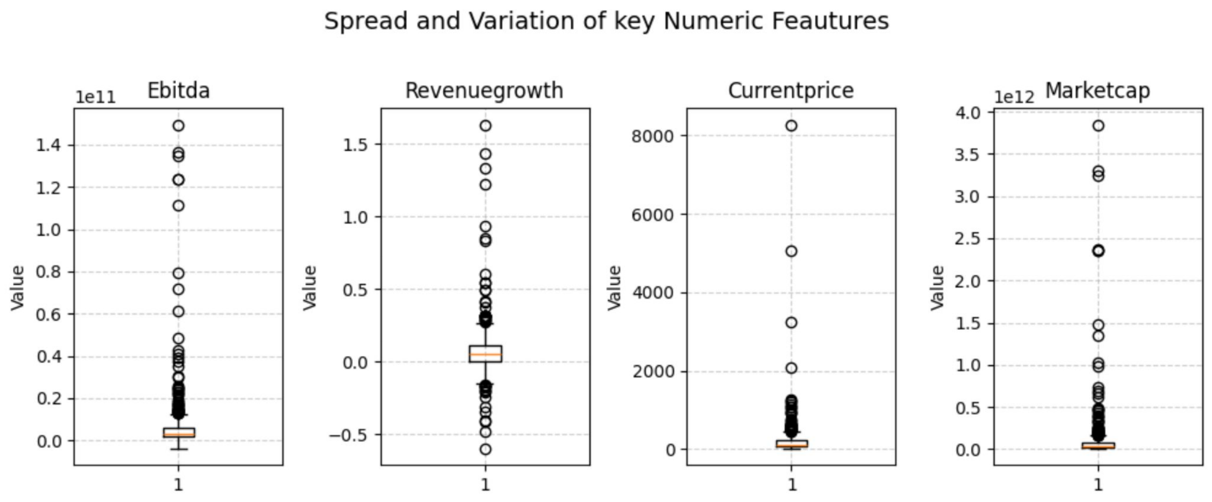


Data Sourcing

- Dataset: S&P 500 Stocks (from Larxel) on Kaggle
- Features: EBITDA, Revenue Growth, Current Price, Market Cap
- Clean, structured, publicly available

Exploratory Analysis

- Noticed skewed distributions, outliers & correlations
- Initial Patterns: Market Cap & EBITDA dominated the financial health indicators



From the spread and variation of the EBITDA, revenue growth, current price, and market cap of the companies in the S&P 500, there are differences in scale; the data is very skewed, and there are many outliers. This allows me to know that I need to scale and choose the right features to build a fair and accurate Financial Health Score.

Cleaning & Features

- Fill in the missing values for EBITDA, Revenue Growth and Full-time employee values with the median
- Drop the columns: Longname, Shortname, City, Country, Exchange, and State
- Convert the sector industry to numeric values before modeling

Financial Health Score (FHS)

1. Inspiration:

- Inspired by the **Altman Z-Score**, a classic metric for predicting bankruptcy risk
- Designed to give a **single, easy-to-compare financial health metric**

2. Metrics Used:

- **EBITDA** → profitability
- **Revenue Growth** → growth potential
- **Current Price** → valuation
- **Market Cap** → size/stability

3. Standardization:

- Each metric standardized using **Z-Score normalization**
- Ensures one metric doesn't dominate the score

$$Z = \frac{X - \mu}{\sigma}$$

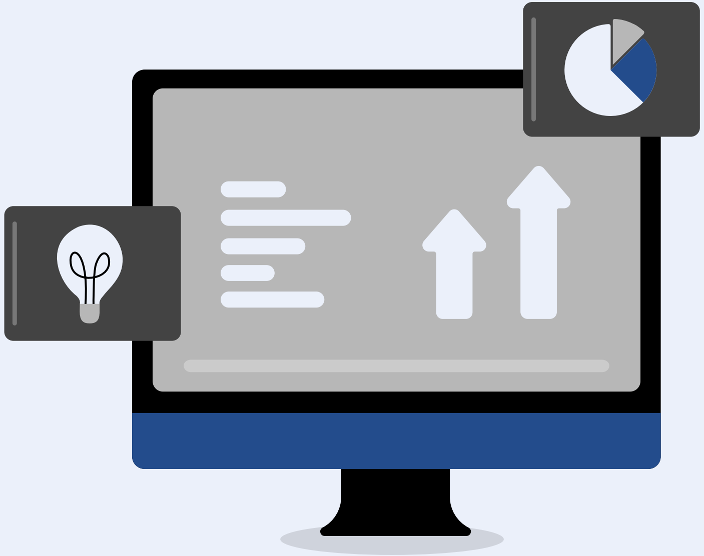
4. Combining Metrics:

- **Higher FHS → stronger financial health**
  Serves as the **target for your Random Forest Regressor**

$$FHS = \frac{E - \mu_E}{\sigma_E} + \frac{R - \mu_R}{\sigma_R} + \frac{P - \mu_P}{\sigma_P} + \frac{M - \mu_M}{\sigma_M}$$

5. Why It Works:

- Provides a **comparative measure** across companies
- Simplifies analysis while capturing multiple dimensions of financial health



02

Model
Random Forest

Model Selection

1. Model Choice:

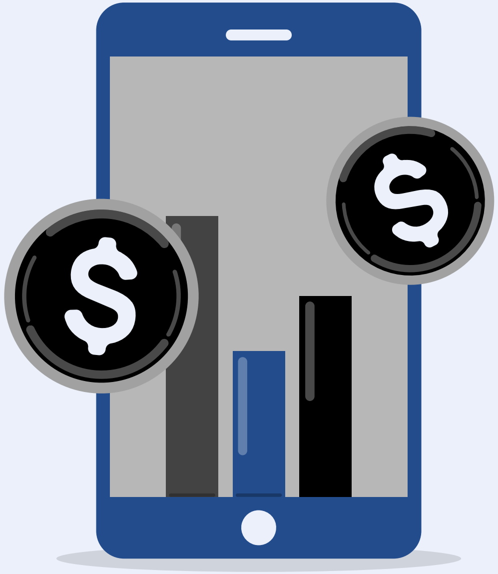
- Used **Random Forest Regressor**
- Handles **non-linear relationships** between features and FHS
- Resistant to overfitting on small datasets

2. Training & Testing

- Split the dataset into **training and test sets**
- Features: EBITDA, Revenue Growth, Current Price, and Market Cap (all standardized)
- Target: **Financial Health Score (FHS)**

3. Evaluation Metrics:

- **Mean Absolute Error (MAE)** → average prediction error
- **Root Mean Squared Error (RMSE)** → penalizes large errors
- **R^2 Score** → proportion of variance explained by model



Model Implementation & Analysis

Three models were tuned to predict the FHS Score

- **MAE (Mean Absolute Error):** 0.879 - 0.921
- **RMSE (Root Mean Squared Error):** 2.147 - 2.153
- **R^2 (Coefficient of Determination):** 0.549 - 0.551.

All models performed similarly, but Model 1 (100 trees, depth 4) outperformed the others. Overall, the models captured about half of the variation in FHS and had good predictive accuracy.

Feature Importance

Feature	Importance
Market Capitalization	0.905
Current Price	0.041
EBITDA	0.029
Revenue Growth	0.025

Key Findings:

- Market capitalization is the strongest predictor of future financial health, where larger companies tend to be more stable.
- Short-term performance, like current price, EBITDA, and revenue growth, contributes less.
- For this dataset, company size is an indicator compared to short-term financial metrics for FHS.

Recommendations

Based on the model results, organizations or an analyst using the Financial Health Score (FHS) approach would prioritize company size, like market capitalization, when assessing financial stability. Because the market cap was the strongest predictor of a company's long-term standing, it will give a more reliable estimation of future health compared to short-term financial movements like daily price changes.

As well, since the simple model performed best, users should not assume that increasing the model's complexity will improve accuracy unless we use a larger data set.

Limitations

The data set is small, which limits the model's capability to assess financial performance.

As well, the features were limited, where additional indicators such as debt ratios, liquidity ratios, cash flow history, and macroeconomic conditions were not included.

Market cap dominated the model, where predictions may lean too heavily on company size.

The dataset only included S&P 500 companies; the model is based on large firms. These findings may not apply to smaller emerging companies, whose financial health patterns can behave differently.

Users should be cautious not to overgeneralize the results or assume that the size alone guarantees stability.

Next Steps

Future work could explore how adding wider sets of financial metrics will improve prediction accuracy. Especially leverage, liquidity, profitability, and historical financial trends. As well, having more historical data will help capture the models' variability.

Also to compare different models like gradient boosting, neural networks, or additional models to see if the relationship is consistent.

Lastly, exploring scenario-based predictions like market downturns, interest rate changes, or shifts in revenue growth could affect future FHS.

RESOURCES

1. Kenton, Will. “S&P 500 Index: What It’s for and Why It’s Important in Investing.” Investopedia, Investopedia, 20 Nov. 2025, www.investopedia.com/terms/s/sp500.asp.
2. Mulani, Safa. “Using StandardScaler() Function to Standardize Python Data.” DigitalOcean, DigitalOcean, 3 Aug. 2022, www.digitalocean.com/community/tutorials/standardscaler-function-in-python.
3. Kartik. (2025, August 2). Logistic regression in machine learning. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/understanding-logistic-regression/
4. Tardi, C. (n.d.). Composite index: Definition, types, and examples. Investopedia. https://www.investopedia.com/terms/c/compositeindex.asp
5. Altman’s Z-score model. Corporate Finance Institute. (2023, October 27). https://corporatefinanceinstitute.com/resources/commercial-lending/altmans-z-score-model/
6. Alexzap. (2025, October 4). Navigating the Financial Maze in an Era of Volatility — 1. Track These 60+ Fundamental Measures in Python. Medium. https://wire.insiderfinance.io/navigating-the-financial-maze-in-an-era-of-volatility-1-bb7c1c6659b7
7. Stock price prediction using machine learning in Python. GeeksforGeeks. (2025, August 6). https://www.geeksforgeeks.org/machine-learning/stock-price-prediction-using-machine-learning-in-python/
8. GeeksforGeeks. (2025a, September 3). Gradient boosting in ML. https://www.geeksforgeeks.org/machine-learning/ml-gradient-boosting/

THANKS!