Financial Health Score Prediction using MLR

Submitted by: Pragya Gupta

PGDM – Research & Business Analytics

Abstract

This project focuses on predicting the Financial Health Score of companies using Multiple Linear Regression (MLR). The primary objective is to identify the most significant variables influencing financial health across industries. The project employs both Excel-based regression analysis and Python implementation, ensuring consistency and validation across tools. The findings highlight consistently significant predictors of financial health, offering valuable insights into corporate performance evaluation.

Introduction

In today's competitive business landscape, financial stability is a critical measure of an organization's long-term sustainability. Predicting the financial health of companies enables stakeholders—including investors, regulators, and management—to make informed decisions.

This project leverages Multiple Linear Regression (MLR) to model financial health scores based on several independent variables. The analysis not only identifies significant predictors but also tests their consistency across industries. By combining Excel and Python-based statistical modeling, the project ensures a robust methodology and cross-validation of results.

Dataset Description

The dataset used for this project contains financial indicators of companies across 7 industry groups (industry code 0–6).

- Target Variable (Dependent): Financial Health Score

- Independent Variables: A mix of financial ratios, performance metrics, and categorical variables (encoded for analysis).

- Size & Structure: [Add no. of rows/columns from Excel here]

- **Data Preprocessing:**
  - Removed insignificant columns (IDs)
  - Treated missing values & duplicates
  - Outliers identified using box plots and adjusted where appropriate
  - Label encoding applied to categorical variables

Methodology

Exploratory Data Analysis (EDA)

1. Data Understanding: Statistical summary, data dictionary, unique values, distribution analysis
2. Data Cleaning: Removal of nulls, duplicates, and irrelevant columns
3. Outlier Treatment: Box plots to detect and adjust extreme values
4. Linearity Check: Scatterplots between dependent and independent variables
5. Multicollinearity Detection: Correlation matrix and VIF analysis

Regression Analysis

1. Built baseline MLR model across all independent variables
2. Identified significant predictors using p-values (<0.05)
3. Iteratively removed insignificant variables (Stepwise Regression)
4. Validated significance of predictors using VIF (Variance Inflation Factor)
5. Performed ANOVA to check model-level significance
6. Conducted Residual Analysis to confirm assumptions (linearity, homoscedasticity, independence of errors)

Cross-Industry Analysis

- The regression steps were repeated for each industry code (0–6).
- Significant predictors were compared across industries to highlight consistent drivers of financial health.

Results & Analysis

- Most variables were found to be statistically insignificant in predicting financial health scores.

- Excel regression results and Python MLR outputs were consistent, validating the methodology.

- Across industries, certain predictors (e.g., profitability ratios, liquidity measures) emerged as consistently significant.

- Diagnostic checks confirmed that the final models satisfied regression assumptions:

  - No severe multicollinearity

  - Residuals were evenly distributed (homoscedasticity)

  - Strong model significance (F-test)

Discussion

- The findings suggest that a subset of financial indicators strongly influences financial health across industries.

- Industry-specific differences exist, but certain financial metrics remain universally significant.

- The project demonstrates the value of combining Excel and Python analysis: Excel provides transparency in regression calculations, while Python allows automation and advanced diagnostics.

Conclusion

This project successfully applied Multiple Linear Regression to predict financial health scores and identify significant variables. The final model revealed consistently important predictors across industries, confirming the robustness of MLR for financial modeling.

The dual implementation in Excel and Python strengthened the analysis, ensuring cross-validation. The insights can support investment decisions, financial risk assessments, and strategic planning.

Future Work: Exploring non-linear models (e.g., Random Forests, Gradient Boosting) could further enhance predictive accuracy.