Machine Learning–Based Classification of Financially Strong Firms Using Comprehensive Fundamental Analysis Metrics

Chandan Maka
Marketing and Data Science
Anderson College of Business and Computing
Regis University, Denver, CO, USA
cmaka@regis.edu

Abstract

There are lots of stocks that are listed in various stock exchanges and that number can reach in thousands. For most of the investors, investment firms and analysts the main challenge is to find out the stocks that can be classified as fundamentally strong companies within the market. With an attempt to find the solution to the above problem, previously multiple works have been done to classify the stocks or select the stocks of a companies whose financial status is strong. This practicum project seeks to classify the stocks with machine learning models and plenty of the fundamental metrics associated with those stocks collectively. This approach utilizes the fundamental financial metrics related to profitability, liquidity, leverage, cash flow quality, efficiency, capital investment, cost structure and shareholder-related measures derived from publicly disclosed financial reports. The goal is to find out the stock's financial status with the help of machine learning model that has been trained under supervision of collected and prepared data. This practical approach will find its potential in the financial domain of equity, portfolio and risk.

I. Introduction/Background

Fundamental financial analysis of an equity securities has been one of the important aspect in the domain of financial management and risk analysis. These are done by analyzing the financial metrics of those companies published in their financial report which are publicly disclosed by those companies. Different metrics like profitability, liquidity, leverage, cash flow and many more have always been important aspect to any financial analysts who is trying to gain the financial status of that equity firm. In this modern financial world, lots of those financial data are increasing in large numbers which might be a bad aspect to traditional manual financial analysis while integrating different financial variables related to that corresponding equity firms. In recent modern financial world, unbiased decision is required and should incorporate multiple dimensional financial data in a systematic way. This approach is made possible with the help of stable data pipelining and machine learning, which can identify the relationship between these metrics and improve the classification. This practicum integrates different data science approaches like data preparation, feature engineering, data visualizations, machine learning models and classification to reach its goal which is stock classification(financially stable or not). The machine learning model will be trained on financial data that has been pre processed which belongs to various companies and learning of the model falls under supervised learning which is trained on labeled data. The models learns the relationship between different financial metrics and use that relationship to further classify the equity firm. This project revolves around the advanced data science domain and financial aspect and focus on actual problem in the relevant field.

II. Problem Statement

The main challenge that has been addressed by this practicum is to create and establish a foundational model which classifies the publicly traded companies according to their fundamental financial metrics.

Before in traditional manual financial analysis approach one used to analyze all the fundamental metrics of a company and decide if it is financially stable or not. But with increasing amount of data, time taken to analyze those data became very complex and biased. At the end, investors and analysts may overlook financially strong company or misclassify the companies that has fundamentally weak characteristics.

In order to overcome the problem we have to consider different approaches of modern data science and encompass the stable data pipeline which helps in data acquisition and data preparation, exploration of data visually, training machine learning models through supervised approach and model evaluation. The complexity of the problem is seen in financial and data science domain where improved equity firm classification would help in reduced financial risk, improved capital allocation and unbiased financial decision among investors. At the end, solving the problem would showcase the use of data science in financial domain.

III. Related Work

Historically, B. Graham's work [1] has significantly influenced in stock selection and investment decision-making. Graham's principles influenced lots of domain like value investing, fundamental analysis, behavioral finance, risk management and different investment philosophies. One of the scoring method which is Altman Z-score model which scores the company based on its financial state had been explored in-depth in one of the article [2]. It explores all the modifications of that model made by Altman on various public and private enterprises. Each version is reviewed in detailed manner explaining how each of its element helped in the assessment of financial stability. One of the study used Altman scoring method to analyze the financial performance of five manufacturing comapanies—Hindustan Unilever, Colgate Palmolive, Nestlé, ITC, and P and G with their secondary financial data from 2013 to 2017 [3]. Altman Z-score is also one of the scoring method that is widely used to find if the comapany is financially distress or not.

Machine Learning technique have been used in different stock selection domain for investment strategies. In 2008, T.S. Quah [4] used neural networks on fundamental financial data for stock selection in the Dow Jones Industrial Average (DJIA). This research's objective was to enhance decision-making processes, offering valuable insights for portfolio management. XingYu Fu et al. [5] had proposed a framework that classifies stocks as good or bad using 244 technical and fundamental features. Based on their return-to-volatility ratio, stocks were labeled. Several machine learning models such as Logistic Regression, Random Forest, Deep Neural Networks, and Stacking were trained for the classification tasks. Namdari and Li [6] in 2018 utilized a FeedForward Neural Network for predicting stock trends using 12 financial ratios. The dataset contained companies listed on Nasdaq between 2012 and 2017. The deep learning model using fundamental analysis outperformed the one using historical prices. In 2021, Huang Y., Capretz L.F., and Ho D. [7] used FeedForward Neural Network(FNN), RandomForest (RF) and Adaptive Neural Fuzzy Inference System (ANFIS) to predict stock returns. They prepared 22 years worth of stock quarterly financial data and analyzed them. The result was concluded as substantial excess returns in their portfolio selections. This showed that machine learning models could be used to help financial analysts with decision-making regarding stock investment. Using synchronized financial reports from the Taiwan stock market, several machine learning models : Random Forest, Feedforward Neural Networks, GRU, and FinGAT—were applied to predict stock returns based on financial ratios. The return from those portfolios was significantly good with top 10 stocks giving more than 100 % relative return with TW50 index.This research by Tsai et al. shows machine learning–based fundamental analysis for stock selection works pretty well [8].

IV. Methodology

The methodology of this project is divided into 2 main phases. First phase is data preparation & model development and the next phase is implementation where we implement our model with real-world data.

A. Data preparation and Model development

In this first phase, we use Financial Modeling Prep (FMP) API to collect the necessary structural financial data which was extracted from the SEC filings of different publicly traded companies. The structured data are obtained in JSON format. Data like income statements, cash-flow statements and balance-sheet statements are collected from the API. We convert the JSON data into csv for further processing. All these data are merged into one unified dataset. During data cleaning process null values were handled and irrelevant columns were removed to simplify our work. Important financial features were retained for further preprocessing.

Feature Engineering was done with the available features present in our dataset. We computed different financial ratios with several financial categories such as profitability, liquidity, leverage, efficiency, cash flow quality, cost structure, capital investment and shareholder value [5], [8] so that it covers diverse aspects of financial health. Exploratory Data Analysis (EDA) was performed for visualization of the distribution of the data in dataset. Boxplot and histograms were plotted. Since financial data contained large range of numbers we applied winsorization method to limit the extreme values. Additionally, signed logarithmic transformation was also applied to visualize the distribution in non distorted way. Missing values that were generated during ratio calculation was imputed with K-nearest Neighbors (KNN) imputation method.

For model training, we needed to complete the data labeling procedure. Two methods : K-means clustering and binary quantile labeling method were used. K-means clustering didn't work well hence, binary quantile method was used for data labeling of the companies listed in the dataset. Each financial feature was assigned a score based on whether its value was above or below the median depending on condition whether its higher or lower values indicated better financial performance. These scores were averaged to compute overall financial score of that company and the labeling was done on the basis of the median financial score. Below is the pseudo code of the labeling procedure :

1) **Input:** Dataset D with financial features $F = \{f_1, f_2, \ldots, f_n\}$.
2) Compute the median M_i for each feature f_i.
3) For each record $r \in D$:
   a) For each feature f_i:
      - If f_i is positively related to financial strength:
$$score_{fi} = \begin{cases} 1 & \text{if } r[f_i] \geq M_i \\ 0 & \text{otherwise} \end{cases}$$
      - If f_i is negatively related to financial strength:
$$score_{fi} = \begin{cases} 0 & \text{if } r[f_i] \geq M_i \\ 1 & \text{otherwise} \end{cases}$$
   b) Compute
$$financial\_score(r) = \text{average}(score_{f1}, score_{f2}, \ldots, score_{fn})$$
4) Compute threshold
$$T = \text{median}(financial\_score)$$
5) Assign label for each record r:
$$label(r) = \begin{cases} \text{Strong} & \text{if } financial\_score(r) \geq T \\ \text{Weak} & \text{otherwise} \end{cases}$$
6) **Output:** Dataset with labels *Strong* or *Weak*.

A Random Forest classification model was used for training the dataset. Fu et al. showed that machine learning models such as Random Forest model can effectively be used to distinguish between strong and weak stocks using financial and technical indicators in a stock selection framework with better result [5]. The performance of the model was evaluated on metrics such as accuracy, preciion, and recall.

B. Implementation

In second phase which is the implementation phase, the trained model was applied to real-world data that was extracted from annual reporting of companies (10-k form). A parser is developed using a library called EdgarTools where it parses the relevant financial information like income, balancesheets and cashsflow statements from the xml file of 10-k annual report. After that with the extracted information, the same financial ratios are computed which was used during training phase to maintain consistency in the model input features.

Finally, an API system was developed using the trained model and annual report parser where it takes the input as url of the SEC filing report of any company listed in US stock exchange where it extracts the required financial statements, generates the required financial ratios and applies the same data processing procedures to maintain consistency and is fed to model and the model outputs the label for the given company. This result is returned by API along with probability score enabling automated financial classification based on the company's annual report.

V. Data Analysis

A. Data Collection and Description

The data was obtained from Financial Modeling Prep API which contained the financial statements like **income**, **cash flow**, and **balance sheets** statements. Using API we collected the financial statements from the latest annual report of **17513** companies. These data contained fundamental financial metrics such as revenues, operating income, assets, liabilities, equity and so on. After merging all those statements into one unified dataset we obtained dataset that had **132** columns. However, many of those features were not needed so, we selected some handful features that would be helpful for our project. Only a subset of features that are relevant to financial ratio computation was selected. The selection was made on 8 categories of financial metrics [5], [8]. This includes :

- Profitability metrics
- Liquidity metrics
- Leverage metrics
- Cash flow quality metrics
- Efficiency metrics
- Capital investment metrics
- Cost structure metrics
- Shareholder value metrics

B. Handling missing values

561 records had missing values which was comparatively low to our total data so, we decided to drop the rows that had missing values. After dropping the rows with missing values, the dataset size was reduced to **16,952 records**.

C. Feature Engineering

To better represent the finiancial status of a company new features were derived from the existing one on the basis of 8 fundamental metrics. A total of **29 financial ratios** were computed using financial statement features. It is shown in Table I below.

Category	Example Metrics
Profitability Metrics	gross_margin, operating_margin, net_margin, ebit_margin, ebitda_margin, ROA, ROE, ROIC
Liquidity Metrics	current_ratio, quick_ratio, cash_ratio, working_capital
Leverage / Solvency Metrics	debt_to_equity, debt_to_assets, equity_ratio, net_debt_to_ebitda, interest_coverage
Efficiency / Activity Metrics	asset_turnover, inventory_turnover, receivables_turnover
Cash Flow Metrics	ocf_ratio, ocf_to_revenue, fcf_margin, cf_to_net_income, capex_to_revenue, capex_to_ocf
Shareholder Value Metrics	free_cash_flow_per_share, book_value_per_share
Cost Structure Metrics	opex_ratio

TABLE I: Categories of Financial Ratios Used in the Study

During ratio calculation, we had some denominators having value zero which would produce infinite values, which were not suitable so, we replaced it with null values. We later handle those values through imputation.

D. Exploratory Data Analysis (EDA)

It was done to find out the distribution of the data. We plotted various boxplots and histograms. In the boxplot we found the extreme outliers but we didn't remove them as they were valuable information of the financial data. We winsorize the data so, that the extreme values were clipped at higher and lower bound which was 1st and 99th percentile. Below figure 1 is the boxplot of the data features :

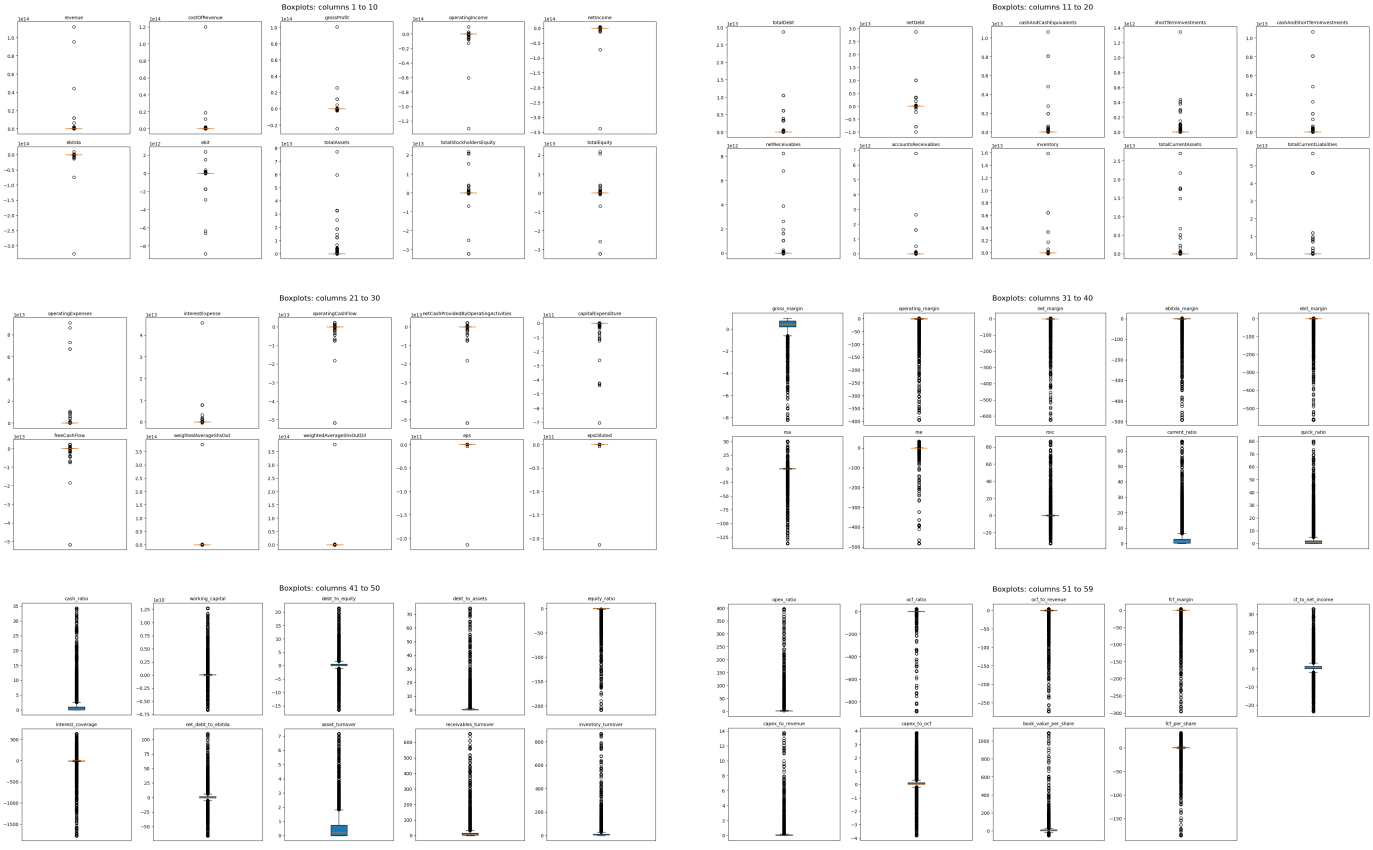

Fig. 1: Boxplot of financial features

We also plotted the signed logarithmic transformation histogram because before that due to extreme range of values it created a single spike histogram which was distorted hence, it was necessary to use signed log transformation to view the distribution of features clearly. Below figure 2 shows the signed logarithmic transformation histogram :

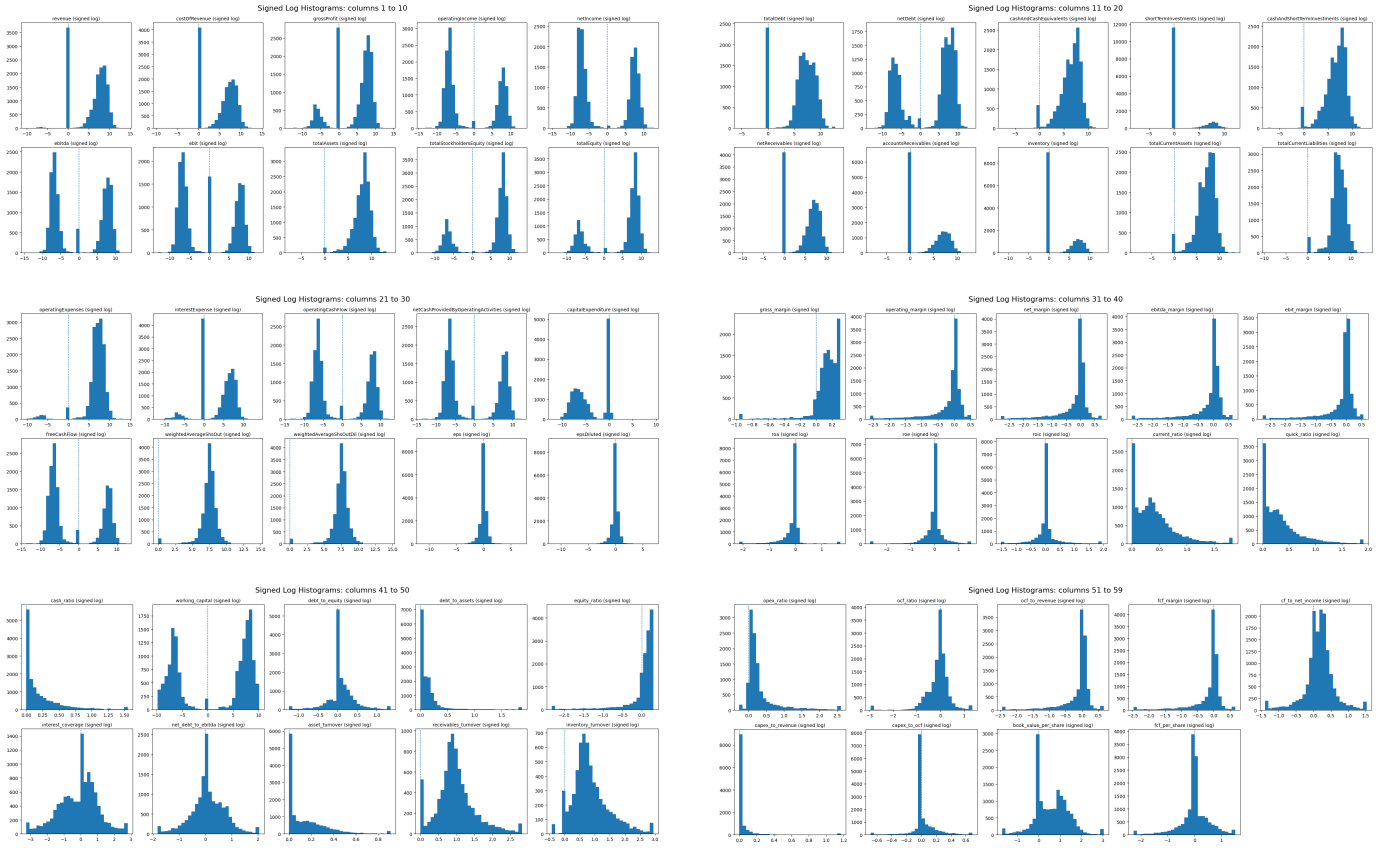

Fig. 2: Signed Log transformed histogram of financial features

From the distribution we can find that the data has highly skewness and binormal distribution. During feature engineering we had lots of features used as denominator as zero which led to creation of null values. So, imputating null values with mean or median will be inappropraite so, we use KNN impuation to fill the null values that were created during feature engineering.

Data labeling is one of the important process that was performed in data preparation phase. Our machine learning algorthim needs some sort of label of those companies which distinguish weak and strong companies. We started the process first with K-means clustering an unsupervised learning, with k = 2 but the cluster was hugely biased.

Cluster	Number of Records
Cluster 0	14784
Cluster 1	180

TABLE II: Distribution of records across clusters using K-means clustering with k = 2

Below is the visualization as well in figure.

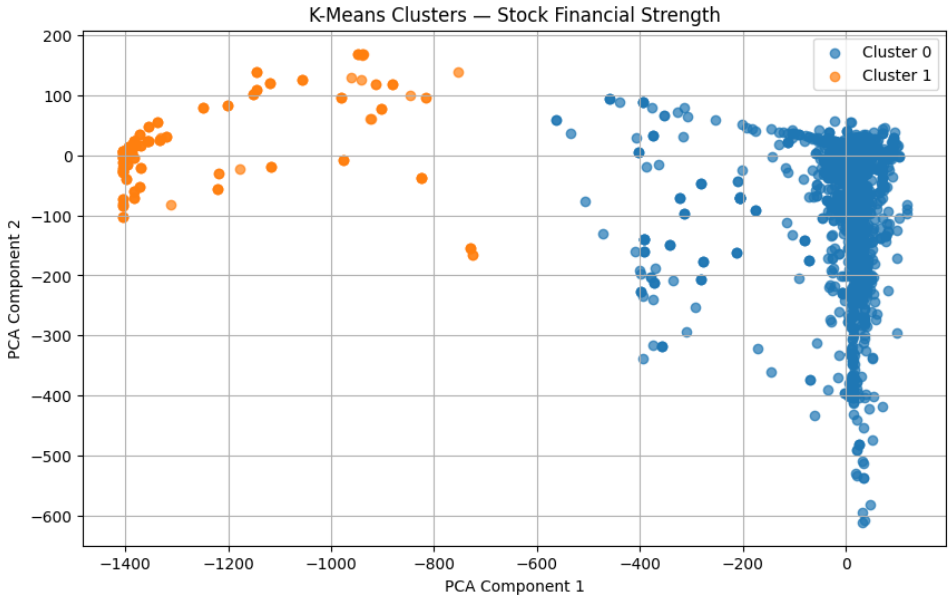

Fig. 3: Distribution of records across clusters using K-means clustering with k = 2

K-means clustering didn't worked for labeling so we used **Binary Quantile** method to label and it worked well. We can see the result below in table III:

Stock Label	Number of Records
Strong	7574
Weak	7390

TABLE III: Distribution of stock labels in the dataset using Binary Quantile Method

After all the data preparation and analysis task we were ready to train the model for classification. We splitted the 80% of the data as training data and 20% as test data. We used Random Forest moidel to train on the data. Result of the model can be find below :

Metric	Value
Accuracy	0.9703

TABLE IV: Random Forest model Accuracy

Class	Precision	Recall	F1-score	Support
0	0.96	0.98	0.97	1478
1	0.98	0.96	0.97	1515

TABLE V: Classification report of the model

	Predicted 0	Predicted 1
Actual 0	1444	34
Actual 1	55	1460

TABLE VI: Confusion matrix of the classification Random Forest model

We can observe the top 20 feature importance selected by Random Forest model for this project in figure below:

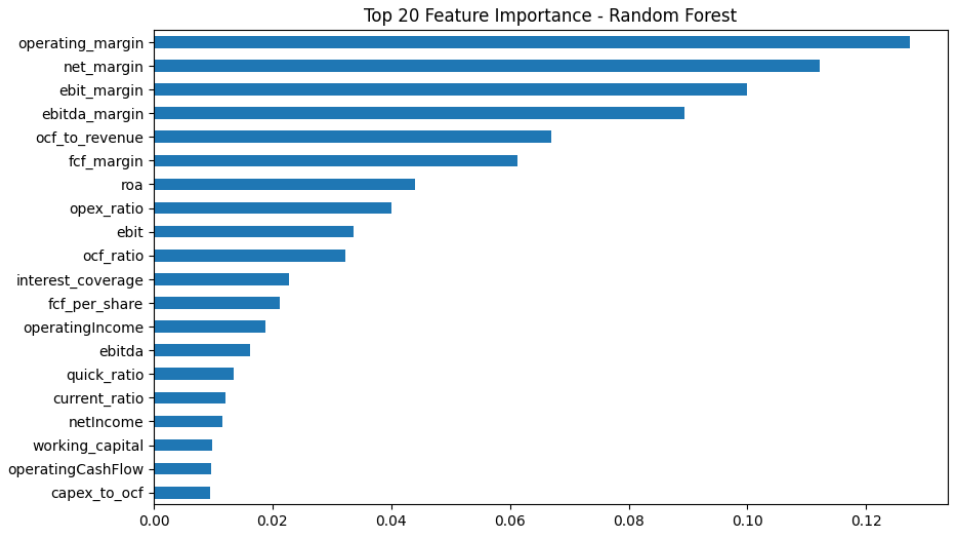

Fig. 4: Top 20 Feature importance - Random Forest

VI. Outcomes

There were 2 important phases for the whole project one was the data preparation and model training phase and the other one was implementation phase.

For the data preparation and model training phase we collected the data, prepared the data according to our need did feature engineering and labeled the data. Labeled data was then used for model training and we trained the model with that data and developed the machine learning model which did quite good in our test data so, the trained model was quite acceptable for our next phase. We used Random Forest model for training on the dataset we prepared. We achieved a model that can classify quite well on the test data.

During the implementation phase, we developed an annual report parser which takes in the URL of the company's annual report which is the link to xml instance of the annual financial report (10k-k form). It parses the financial information and statement from that xml instance that can be used for classification task. We used that data and generate the same financial ratios used during training phase, for the model's input. Then the model uses that input and classifies the selected comapany as 0 or 1 along with its probability score. O means financially weak and 1 means financially strong company. We created an API that takes the company's annual report sec filing url as input and gives API response as classification label of the selected company along with probability score.

We selected two random company for testing : Apple Inc. and IONQ Inc. The outcome from our model is shown below in table VII:

Company	Year	Prediction	Weak Probability	Strong Probability
IONQ	2025	Weak (0)	0.8967	0.1033
AAPL	2025	Strong (1)	0.0005	0.9995

TABLE VII: Model Prediction Outcomes for Selected Companies

VII. Timeline

The project was expected to complete over an eight- week period time in a systematic approach. The structured timeline for the project is outlined below in Table VIII with weekly phases, key activities and expected deliverables for the project.

TABLE VIII: Eight-Week Practicum Project Timeline

Week	Phase	Key Activities and Deliverables
Week 1	Proposal Development	Finding the project scope and finalizing it, writing project proposal
Week 2	Data Acquisition	Access API and collect the data from the API and store them
Week 3	Data Cleaning and Preparation	Handling missing values, normalize and scale financial metrics, feature engineering and labelling
Week 4	Exploratory Data Analysis	Statistical analysis, visualizing relationship of features, correlations among variables, exploratory data visualization
Week 5	Model Development	Train model, perform hyper-parameter tuning, evaluate model
Week 6	Model Interpretation and Validation	Validate model and check robustness
Week 7	Results Analysis	Analyzing findings and insights, identifying limitations
Week 8	Final Report and Documentation	Complete the practicum report and submit final deliverables

VIII. Conclusion

This project shows how we can leverage use of machine learning to classify the stock as strong or weak in its financial status using the financial information of the company. Integration of financial statement data, feature engineering of key financial ratios, and a Random Forest classification model was done successfully to built a system that is capable of classifying if the publicly traded company is financially strong or weak. The model achieved an accuracy of 97% indicating that financial ratios combined with machine learning can provide meaningful insights for stock analysis. The implementation of an API that is capable of processing SEC 10-k filings and generates realworld classification shows the practical implementations of the system in financial domain and investment decision. Overall, the project demonstrates the potentiality of data science techniques in automated financial decision and in evaluating corporate financial health

References

[1] B. Graham, D. L. Dodd, and S. Cottle, *Security Analysis: Principles and Technique*, 4th ed. New York: McGraw-Hill, 1962.

[2] F. Rashid, R. Khan, and I. Qureshi, "A comprehensive review of the altman z-score model across industries," *SSRN Electronic Journal*, vol. 27, 12 2023.

[3] JMRA, "Predicting bankruptcy of selected firms by applying altman's z-score model," 2017. [Online]. Available: https://www.academia.edu/37626255/PREDICTING_BANKRUPTCY_OF_SELECTED_FIRMS_BY_APPLYING_ALTMANS_Z_SCORE_MODEL

[4] T.-S. Quah, "Djia stock selection assisted by neural network," *Expert Systems with Applications*, vol. 35, no. 1-2, pp. 50–58, 2008.

[5] X. Fu, J. Du, Y. Guo, M. Liu, T. Dong, and X. Duan, "A machine learning framework for stock selection," 2018. [Online]. Available: https://arxiv.org/abs/1806.01743

[6] A. Namdari and Z. S. Li, "Integrating fundamental and technical analysis of stock market through multi-layer perceptron," in *2018 IEEE technology and engineering management conference (TEMSCON)*. IEEE, 2018, pp. 1–6.

[7] Y. Huang, L. F. Capretz, and D. Ho, "Machine learning for stock prediction based on fundamental analysis," in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2021, pp. 01–10.

[8] P.-F. Tsai, C.-H. Gao, and S.-M. Yuan, "Stock selection using machine learning based on financial ratios," *Mathematics*, vol. 11, p. 4758, 11 2023.