JPMC Quant Challenge'23

Machine Learning

Piyush Jain
Bachelor of Technology in Computer Science
Delhi Technological University

Overview

Data Analysis

1. Problem Statement
    a. Problem Statement
    b. Solving methodology
2. Data Overview
3. Data Preprocessing
    a. Feature Creation
    b. Feature Selection

Model Building

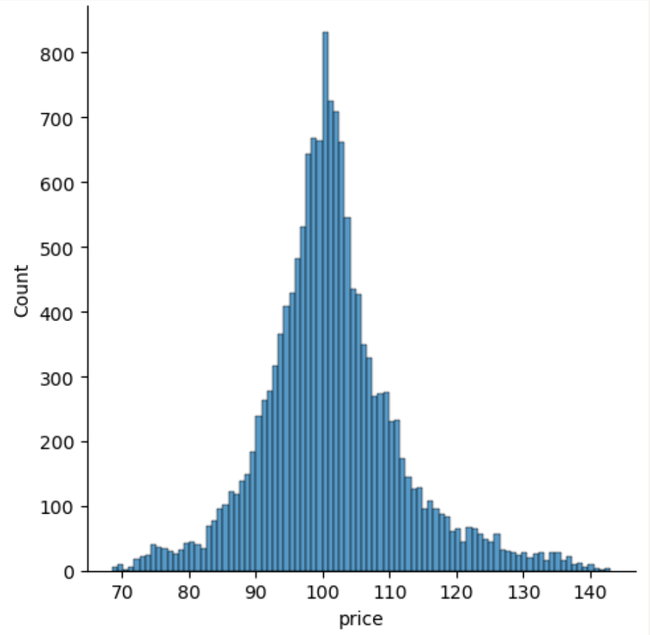
1. Modeling
    a. Modeling Methods
    b. Proposed solution
    c. Ensemble Technique
2. Conclusion

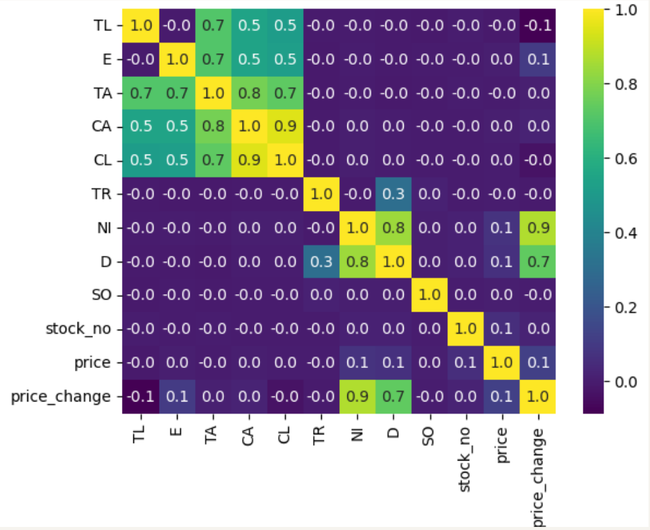
Problem Statement

We have access to monthly fundamental indicator values for 100 stocks. For each stock, we have data available for 9 financial metrics, namely Current Liabilities, Total Liabilities, Equity, Total Assets, Current Assets, Total Revenue, Net Income, Dividend, and Shares Outstanding. These values are recorded on the last day of each month. Each stock's data spans 200 months, and we have price information for the past 150 months. Our task is to predict the prices for these stocks as accurately as possible, i.e., aiming to minimize the error in our predictions. The evaluation metric we used are root mean square error and mean absolute error.

Problem Solving Methodology

To enhance human interpretability, we will utilize the available data to create financial ratios. These ratios will help us gain insights and understand the financial performance of the stocks. Additionally, instead of directly analyzing stock prices, we will focus on modeling price changes. This approach aligns with the random walk theory, which suggests that future price movements are best predicted by changes in prices rather than the actual price values themselves.





Data Overview

The training data comprises historical data for the past 150 months for each of the 100 stocks. It includes stock-specific information for 9 financial metrics. Notably, all stocks began at an initial price of 100, facilitating straightforward price comparisons. Additionally, the stock prices in the dataset adhere to a normal distribution with a mean value of 101. Importantly, the data is complete, with no missing values present.

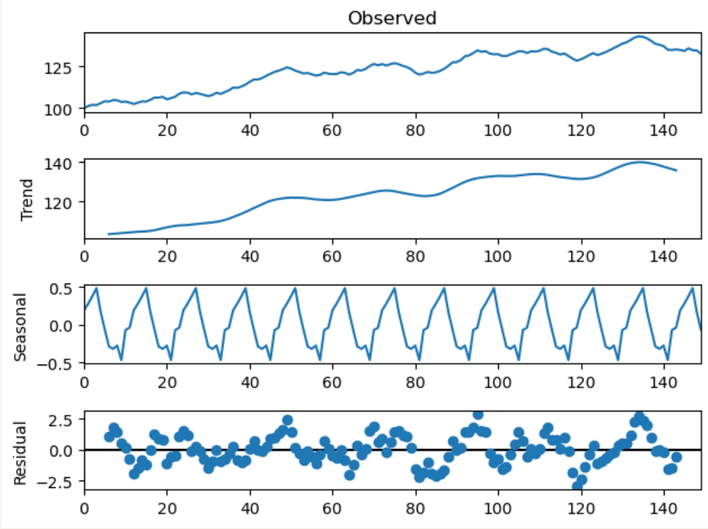
Financial metrics and their impact on share price and company insights:

1. **Total Liabilities(TL)** - Combined debts owed by the company.
2. **Equity(E)** - Ownership value for shareholders, positively influencing share price.
3. **Total Assets(TA)** - The overall worth of a company's possessions, decrease may lead to lower share price. Given by TA=TL+E
4. **Current Assets(CA)/Current Liabilities(CL)** - Short-term assets and obligations, usually exhibit a strong correlation.
5. **Total Revenue(TR)** - Overall sales or generated revenue on a monthly basis.
6. **Net Income(NI)** - Profitability after deducting expenses and taxes, highly correlated with price change.
7. **Dividend(D)** -Earnings distributed to shareholders, provided when income is positive.
8. **Shares Outstanding(SO)** - Total number of issued and held shares by investors.

Feature Creation

Key financial ratios for feature creation:

- **Debt-to-Equity (D/E) Ratio** : Given by TL/E . Higher D/E ratio signifies higher financial risk.
- **Profit Margin**: NI/ TR . Higher profit margin indicates better profitability.
- **Dividend Payout Ratio**: D/NI. Higher ratio indicates larger dividend distribution to shareholders.
- **Earnings per Share (EPS)**: NI/SO . Higher EPS implies greater profitability per share.
- **Current Ratio**: CA/CL . Measures a company's ability to pay short-term obligations.



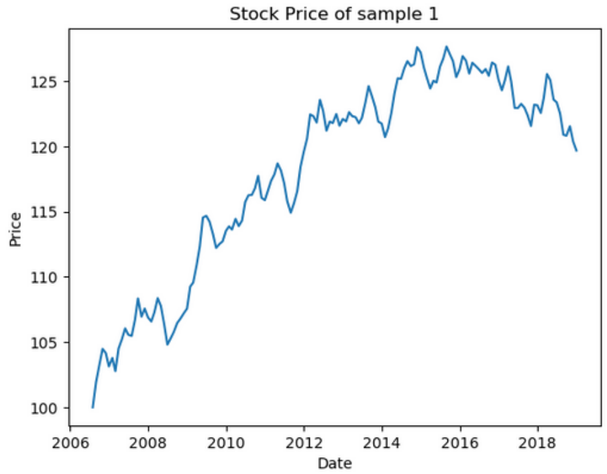
Additional miscellaneous features considered for modeling:

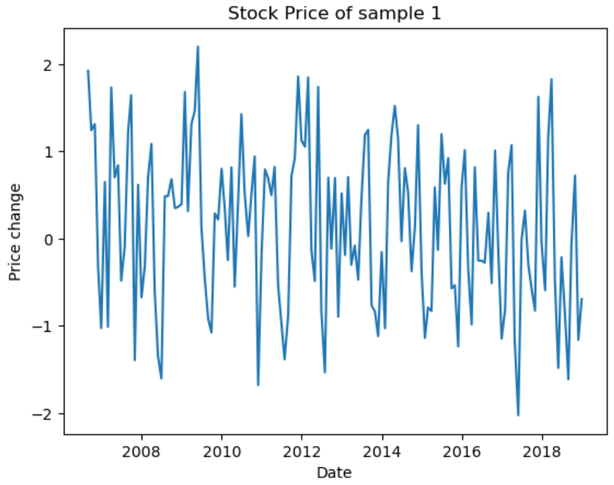
- **Number of Days**: The number of business days in a month can potentially impact the share price as different market dynamics may exist during varying lengths of time.
- **Quarter**: The quarter of the year can influence the investment patterns of investors, as certain quarters may have specific events or reporting periods that affect market sentiment.
- **Change Feature**: Changes in the stock price can be influenced by changes in company finances, such as fluctuations in assets or other financial metrics.
- **Total Income**: The total income of a company, calculated as the sum of dividends and net income. This metric captures both the returns distributed to shareholders and the profitability of the company.

Addressing Key Questions

What are the reasons for choosing to predict price changes and subsequently taking the cumulative sum, rather than directly predicting the price values?

To address the issue of non-stationarity in the stock price and align with the principles of the random walk model, we employ first-order differencing. This approach helps make the series stationary and acknowledges that the change in share price does not rely on its past values.





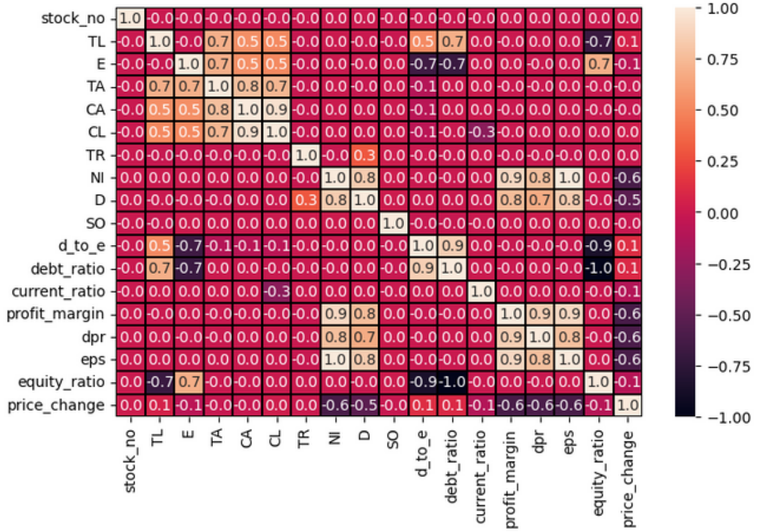
What is the reason behind shifting the features by one entry?

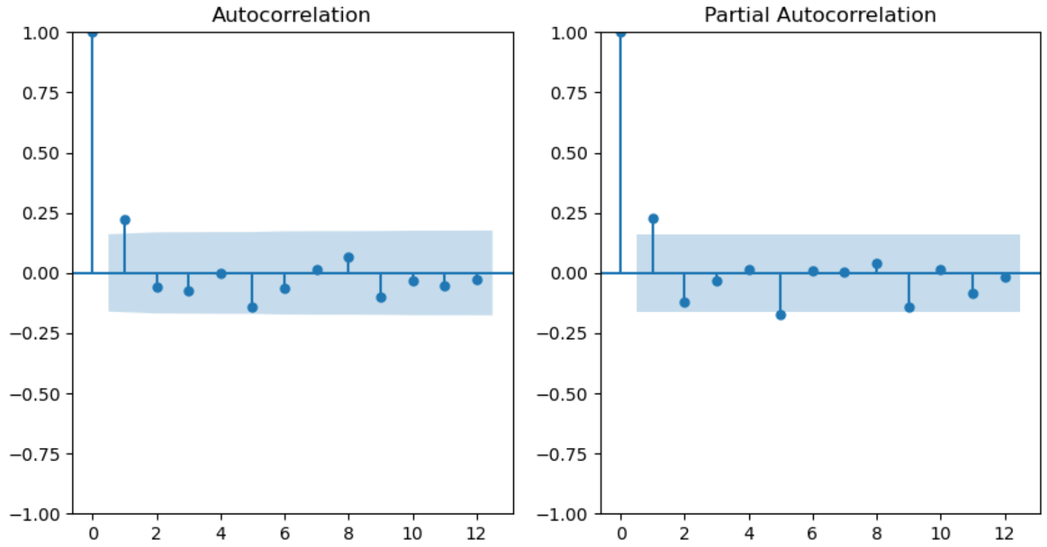
The reason for shifting the features by one entry is that investors can only respond to the company's financial information after it has been released. Therefore, any changes in the company's finance are reflected in the subsequent month's stock price change.

Feature Selection

We employed the following feature selection techniques to filter out the most useful features:

Feature Selection Technique	Result
Pearson Correlation Coefficient	This technique was used to identify and eliminate variables that exhibit multicollinearity
Autocorrelation Plot and Partial Autocorrelation Plot	As expected from the random walk model, no significant autocorrelation or partial autocorrelation was observed
Sequential Feature Selector	Identified the best subset of features for predicting the price change





Alternative Methods

Method 1: Direct Share Price Forecasting with Single Model for All Stocks

Pros:

- Simplicity and Speed: Provides a simple and fast baseline model.
- Uniform Error Propagation: Prediction errors do not accumulate in future forecasts.

Cons:

- Non-Stationary Price: Challenges in modeling and making the price trend stationary.
- Limited Correlation: Poor correlation between share price and other input features.

Model	RMSE	MAE
ElasticNet+RandomForest	6.55	5.00
ElasticNet+XGBoost	7.56	5.86
PolynomialRegression+RandomForest	13.14	9.48

Method 2: Predicted Price Change with Separate Model for Each Stock (Multiple Time Series)

Pros:

- Predicting price change aligns with the random walk theory and make the series stationary.
- High Correlation: Price change exhibits a strong correlation with input features.

Cons:

- Error Accumulation: Cumulative sum operation introduces non-uniform error accumulation.
- With only 150 data points per series, models struggle to capture the true patterns in the series.

Model	RMSE	MAE
LinearRegression	2.54	1.86
RandomForest	2.79	2.10
FBProphet	7.99	6.16

Method 3: Predicting Price Change with Single Model for all Stocks

Pros:

- Capturing True Patterns: Big dataset enables the model to capture the true patterns in the data.
- Faster Prediction: Single model approach allows for faster predictions.

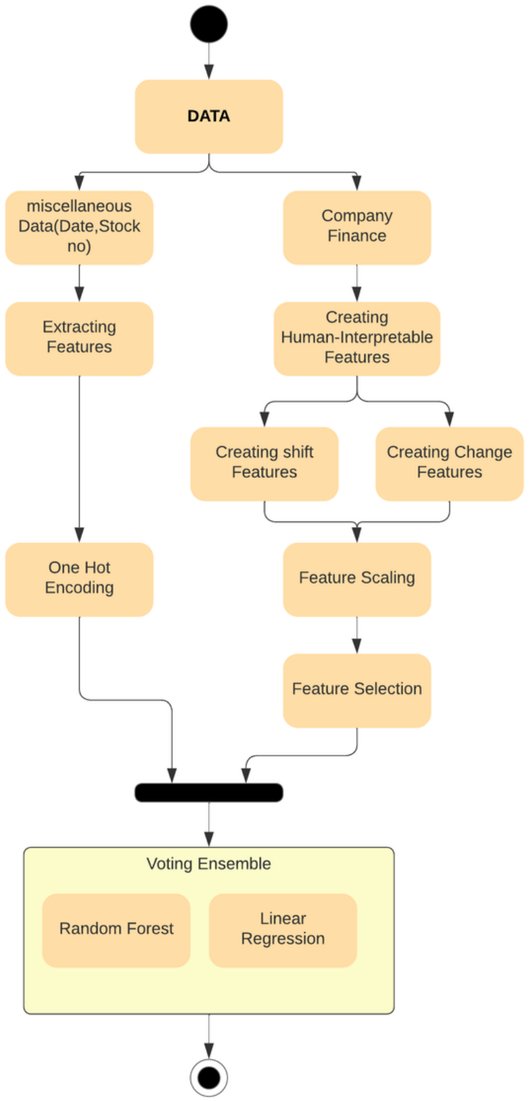
Cons:

- Non-Uniform Error Accumulation due to Cumulative sum operation.

Model	RMSE	MAE
RandomForest	2.52	1.57
Neural Network	2.99	2.23
FBProphet	7.34	5.67

Proposed Solution

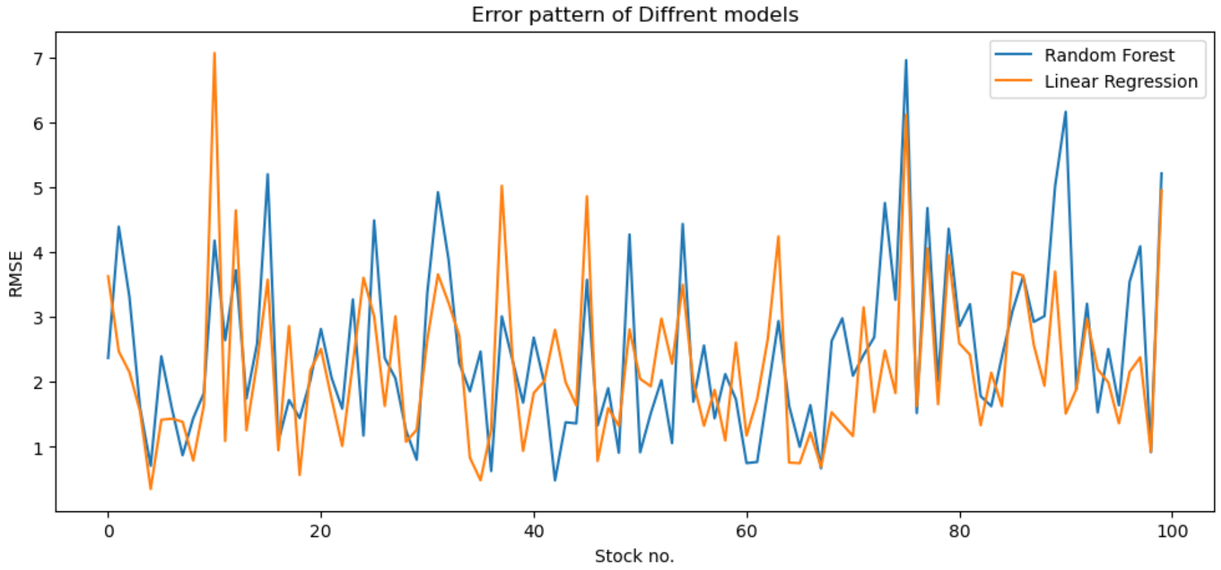
- Financial ratios are derived from the given data to enhance interpretability.
- The data is shifted by one timestamp to account for the delayed impact of financial metrics.
- Scaling is applied to the data for stable training and to ensure consistent ranges.
- One-hot encoding is performed on the stock number to differentiate between different stocks.
- A voting ensemble is utilized, combining linear regression and random forest models.
- Linear regression contributes a high bias, low variance component to the ensemble.
- Random forest contributes a low bias, high variance component to the ensemble.
- The ensemble model aims to strike a balance between bias and variance, resulting in improved predictions.



Ensemble Technique

To leverage the distinct error patterns of different models, a voting ensemble approach was employed, combining Random Forest and Linear Regression. By combining these models, which exhibit diverse error patterns, we aimed to capitalize on their respective strengths and weaknesses. This ensemble approach allows for improved prediction accuracy, as each model contributes its unique insights to different stocks in the dataset

model	RMSE	MAE
Voting ensemble Random Forest+ Linear Regression	2.11	1.50



Possible Improvements

- **Accounting for Holidays**: Consider incorporating holidays in the calculation of the number of working days to account for potential effects on stock prices during those periods.
- **Special Financial Events**: Include a feature that captures special financial events, such as stock splits, as they can significantly impact stock prices.
- **Error Accumulation Mitigation**: Explore alternative modeling approaches that minimize error accumulation or cancellation, ensuring more accurate predictions.
- **Hybrid Approach**: Adopt a hybrid approach by combining models for predicting both price and price change. This approach leverages the uniform error distribution of price predictions and captures the patterns of price change for more accurate forecasts, especially in the initial days.
- **Hyperparameter Tuning and Model Exploration**: Conduct hyperparameter tuning for existing models or consider experimenting with different models to improve prediction accuracy and performance.

These improvements aim to enhance the accuracy and robustness of the model by accounting for specific factors, reducing error accumulation, and exploring alternative modeling strategies.

Conclusion

Based on the data and model analysis, the following conclusions can be drawn:

- Stock market follows the Random Walk model, indicating that past stock prices do not have a significant impact on future prices. Therefore, price changes are better predictors of future stock prices.
- Return on investment (ROI) proves to be a more reliable and effective approach for predicting stock prices. It provides valuable insights into the profitability and financial performance of a company.
- The stock price is strongly influenced by the income of the company, with profit margin and earnings per share (EPS) emerging as the most relevant financial ratios. These metrics reflect the company's financial health and profitability, which have a direct impact on stock prices.
- The stock price exhibits diverse patterns, with some stocks following simple patterns and others exhibiting complex patterns. This highlights the need for ensemble techniques, such as combining multiple models, to capture and interpret the diverse patterns in stock price movements effectively.

By considering these conclusions, investors and analysts can make informed decisions and predictions regarding stock prices based on the underlying data and model analysis.