Improving Investment Returns with Machine Learning

Yunjing Ma, Jie Wang, Qianhui Zhong

October 28, 2024

1. Introduction

The stock market is crucial for personal and institutional investment strategies, and accurate stock price predictions offer valuable insights. However, forecasting is challenging due to the market's complex, non-stationary data. This project leverages machine learning to predict prices, helping investors optimize returns. It targets individual and institutional investors aiming to improve decision-making by analyzing future trends.

The LSTM machine learning model developed in this project aims to predict stock prices by analyzing historical stock data and other external factors like public interest, as observed through Google Trends. The primary goal is to create a system that can effectively predict trends and improve the overall accuracy of stock price forecasts.

2. Technical details

2.1 Data collection

The project only uses publicly available data, and there are no privacy concerns related to individual-level data. In this project, we selected Tesla's stock data from January 2021 to September 2024. Data was collected from two primary sources:

1. **iFind API**: This API is an extension of EXCEL from Hithink RoyalFlush Information Network Co., Ltd, which provides raw data, including various stock prices and technical indexes.

2. **Google Trends**: Public interest in Tesla stocks was measured using the Google Trends index, based on search frequency for relevant keywords. This index was not included in our initial Project Canvas, but during our exploration, we discovered that it somewhat

reflects market sentiment and could have a notable impact on stock prices. Additionally, the data is accessible, so we decided to include this index as our original variable.

The collected data will need preprocessing and cleaning, such as addressing recurring zero values, sharp price drops caused by stock splits, and other issues.

2.2 Data preprocessing

The preprocessing stage aimed to transform the raw data into a clean and structured format suitable for our further machine learning model. We carried out the following steps:

1. **Data Integration**: The iFind API and Google Trends data had different time spans, which were aligned to cover a consistent period. Daily data was chosen from the overlapping period (January 2021 - September 2024). Then, the data was combined using the date as the key, creating a comprehensive dataset for the model training. This allowed us to examine both stock prices and market sentiment, providing a holistic view of factors influencing Tesla stock prices.

2. **Dealing with Missing Data**: To get an overview of the data, we went through the dataset, and we found several rows showing zero values at regular intervals, which were found to correspond to weekends and holidays. To address this, the Python holidays library was employed to create a new column marking whether a given day was a trading day. After processing, a re-examination of the data revealed a few remaining 0 values, which occurred on non-trading days not recognized as national holidays. These instances were manually handled, resulting in a cleaner dataset.

3. **Feature Engineering**: We added a binary feature, "Is a trading day or not," to handle days when the stock market was closed. The categorical values were converted to numerical (1 for trading days and 0 for non-trading days).

4. **Data Normalization**: To ensure consistency across variables and to meet the requirements of input values in the LSTM model, the numerical features were scaled to a common range (0 to 1). This was applied to stock prices and the public interest index from Google Trends.

2.3 Exploratory data analysis and visualization

Just like the process in Project Canvas, the analysis began with an exploration of the dataset to gain a comprehensive understanding of its structure and contents.

1. We looked at the processed data, and descriptive statistical analyses were then performed using functions such as df.head(), df.describe(), and df.info(). These methods provided information into the distribution of the data and data types, offering a clearer view of its overall structure.
2. Next, various visualizations were created to enhance the interpretation of the data. During this phase, abnormal fluctuations in stock prices were observed, particularly sharp declines. Further investigation revealed that Tesla had recently undergone a stock split, which explained the sudden price drop. We adjusted the stock price by multiplying it by the split factor, ensuring the accuracy of the visual representations and removing the anomalies in stock price fluctuations. Additionally, we recalculated stock-related indicators, such as MA10, as they were also affected by the stock split.
3. Following this, Pearson and Spearman correlation analysis was conducted to understand the linear and nonlinear relationships between variables. Principal Component Analysis (PCA) and Factor Analysis were applied to reduce the dimensionality of the dataset. The Kaiser-Meyer-Olkin (KMO) test was performed at first, confirming the data's suitability for Factor Analysis. PCA analysis identified three principal components that explained a significant portion (over 85%) of the variance in the dataset and enabled the reduction of dimensionality. At last, by Factor Analysis, we identified the main explanatory variables for each principal component (Opening price, RSI, and Public Interest), these will be the key indicators for analyzing and predicting the stock price trend.

2.4 Prediction model for stock price

After a thorough comparison of different approaches in Project Canvas, the Long Short-Term Memory (LSTM) model was selected for the prediction task as it effectively captures long-term dependencies in time series, handles nonlinear patterns, and exhibits robustness to noise. Based on an analysis of the existing data and the properties of the LSTM model, two potential strategies were identified: the first involved using three indicators we get from prior PCA process as inputs and the closing price as the output for multi-step prediction; the second focused on using the closing price as the core input for self-prediction, i.e., the model progressively uses previous predictions as new inputs for the next step's prediction to forecast the stock price iteratively.

Both methods were implemented and evaluated. The results indicated that the multi-step prediction method yielded poor and unstable performance, while the self-prediction approach provided significantly better accuracy. Consequently, the latter strategy was chosen for further development. The LSTM model successfully achieved accurate stock price predictions, offering valuable insights for investors by enhancing predictive precision.

Based on our current knowledge and discussions, we expected the first approach to perform better. However, the results turned out to be the opposite, leading us to investigate the reasons behind the differences in model performance between the two approaches. Unfortunately, while the reasons for the poor performance of the multi-step prediction method were investigated, a definitive explanation was not found due to limitations in current knowledge. To address this gap, further studies in machine learning are being pursued, emphasizing explainability in machine learning techniques (like the "Introduction to Machine Learning" course in Period 2). It is hoped that future learning will provide deeper insights into the challenges encountered and allow for optimization of the model to improve its predictive capabilities.

3. Results

Through data processing, analysis, and LSTM modeling, we provided investors with two key insights:

1. Among numerous initial variables, we identified three metrics with the most significant impact on stock price fluctuations, enhancing analysis efficiency for investors.
2. We generated a relatively accurate forecast of short-term stock price movements, offering additional investment information that could help investors achieve excess returns.

In our mini-project, we used Tesla as an example. Ideally, we aim to make our model adaptable for all stocks using just a stock symbol. However, we have not yet found an effective way to retrieve all necessary data in real-time. Therefore, we envision our mini-project as an analytical and predictive approach using data science, presented in a blog format on our webpage https://jiewang0313.github.io/Data-Science-Power-in-Improving-Investment-Returns/. Investors can view Tesla's short-term performance clearly, and for similar analyses on other companies, they can collect data from a financial terminal and apply our code. All the resources about the project are at https://github.com/JieWang0313/Data-Science-Power-in-Improving-Investment-Returns.

4. Future steps

In future iterations, incorporating additional data sources could also strengthen the model. Exploring more technical indicators and other financial metrics related to stock prices and integrating sentiment analysis from social media, financial news, or economic indicators, the model could capture a broader range of factors influencing stock prices. This can enhance the model's robustness and improve its ability to account for market behaviors.

Further refinement of the machine learning models could significantly enhance predictive accuracy. Exploring more advanced architectures, such as Transformer models or hybrid approaches that combine statistical techniques with machine learning, presents a promising direction for model improvement. What’s more, hyperparameter tuning for models like LSTM, using methods such as grid search or Bayesian optimization, could optimize performance and further refine the model’s predictive capabilities.

Additionally, other financial analytical methods such as value analysis could be integrated into the machine learning model to explore the potential for predicting long-term stock price changes and trends, offering investors more comprehensive reference information. By combining short-term and long-term forecasts, it may also be possible to minimize prediction errors, enhancing the value of forecast information for investors.

5. Learning outcomes and conclusions

This project demonstrates the feasibility of using machine learning techniques to predict stock prices and improve investment returns. By leveraging historical stock data and external factors such as public interest from Google Trends, a robust machine learning model was developed to forecast stock price movements. The Long Short-Term Memory (LSTM) model, in particular, proved effective in capturing short-term fluctuations in stock prices.

The project provides significant value to investors by helping them make informed decisions based on data-driven predictions. It highlights the potential of machine learning to address real-world financial challenges, enabling investors to optimize their portfolios and improve their overall returns. It is a step toward applying machine learning to real-world financial problems, offering a practical solution for investors to enhance their market analysis capabilities.