Project Proposal

Machine Learning for Stock Price Forecasting

Stock price forecasting is a challenging task due to the complex, dynamic, and non-linear nature of financial markets. Traditional models for Time Series Forecasting (such as ARIMA) struggle to capture these intricacies, and may lead to suboptimal investment decisions in a variety of scenarios, especially when the data has no clear seasonality and features non-linear patterns. In this project we investigate machine learning techniques for stock price forecasting, with the aim to reproduce complex patterns in stock price movements and deliver robust forecasts with improved accuracy over traditional methods. We will focus on training Random Forests and Long-Short-Term-Memory (LSTM) neural networks for stock price forecasting, evaluate their accuracy and compare their performance. We will focus on two main stocks: Coca-Cola and Tesla. The former has a history of solid growth, whereas the latter is a newer and more volatile one, and this sharp contrast will allow us to evaluate our models on datasets reflecting a different behavior over time. Ultimately this project aspires to contribute to enhancing the reliability of stock price forecasting and facilitate a more informed, confident and profitable decision-making for investment firms in an ever increasingly data-centric world.

Problem Statement

The traditional statistical models for stock price forecasting have limitations in capturing the complex and dynamic nature of stock markets. The problem we want to address is to enhance the accuracy and reliability of stock price forecasting by using advanced machine learning models, concretely Random Forests and Long Short-Term Memory (LSTM) networks.

Context

Potential clients may be investment firms making critical investment decisions based on stock price predictions. Inaccurate forecasts can result in significant financial losses. By improving the accuracy of these predictions, we can make impactful contributions to their strategic investment decisions, leading to higher profitability.

Criteria for Success

Success in this project would be significantly improving the accuracy of the firm's stock price predictions compared to their current models. For instance, one could aim for a minimum 15% improvement in accuracy metrics such as Mean Squared Error (MSE).

Scope of Solution Space

We will focus on investigating the applicability of Random Forest and LSTM models for stock price predictions.

Constraints

We are bound by data availability and the amount of historical data that can be feasibly processed. Another constraint is time: the financial market is highly volatile and waiting for very accurate predictions can result in missed opportunities, so our models should also be relatively fast in their training and performance.

Stakeholders

The primary stakeholders are the firm's investment decision-makers. Secondary stakeholders include the firm's investors who stand to benefit from improved decision-making, and the data science team that would be implementing the ML models.

Data sources

We will use historical stock price data from Yahoo Finance and Alpha Vantage. From the former we will obtain datasets providing open, close, high, low prices, and trading volumes, whereas the earnings per share will be obtained from the latter.

Approach

Our approach will involve the following steps:

1. **Data Acquisition & Preprocessing**: Acquire historical stock price data and preprocess it to create a structured time-series dataset suitable for our ML models.
2. **Exploratory Data Analysis (EDA)**: Analyze the data to identify patterns, anomalies, or other insights that may inform our model training.
3. **Feature Engineering**: Generate meaningful features that can improve the predictive performance of our models.
4. **Model Building and Validation**: Develop Random Forest and LSTM models and validate them using techniques like cross-validation.
5. **Evaluation**: Compare the accuracy and computational efficiency of the two models and determine which is most suited to our needs.