Introduction

This document outlines the methodology and findings of the stock market prediction project aimed at forecasting future stock prices using historical data. The project incorporates machine learning techniques to develop model that assist in making informed investment decisions.

Dataset Overview

The dataset used in this project is sourced from Kaggle and contains daily trading data such as open, high, low, close prices, and volume of stocks. Additionally, it includes several derived technical indicators which are essential for the analysis.
Dataset : https://www.kaggle.com/datasets/luisandresgarcia/stock-market-prediction

Exploratory Data Analysis (EDA):

- **Data Cleaning:** Handled missing values, corrected data types, and removed duplicates.
- **Visualization:** Used plots like time series for stock prices, histograms for distribution of variables, and correlation heatmaps to understand relationships between features.

Feature Selection

Based on the correlation heatmap analysis, features demonstrating high multicollinearity were either transformed or removed to enhance model performance. Key features used include 'open', 'high', 'volume', along with exponential moving average and other technical indicators like stochastic.

Model Building

A Linear Regression model was chosen for its straightforward implementation and interpretability, especially valuable in financial datasets where understanding the relationship between features and target is crucial. This model is well-suited for continuous output prediction, which aligns with predicting stock prices.

- **Training:** The model was trained on a subset of the dataset features selected based on the exploratory data analysis. These features included 'open', 'high', 'volume', and technical Indicators like exponential moving average and stochastic. The target variable in this context is the 'close' price of the stock.
- **Parameters:** The initial Linear Regression model was used with default parameters. Given the nature of the dataset, no regularization was applied initially to keep the model simplicity and focus on understanding the basic linear relationships in the data.

Model Evaluation

- **Metrics Used:** The primary metric used to evaluate the model's performance was the Root Mean Squared Error (RMSE). RMSE is particularly effective in quantifying the magnitude of error between predicted values and actual values, making it suitable for regression tasks where you need to understand the error in terms of the same units as the target variable.
- **Results:** The Linear Regression model achieved an RMSE of 0.86. This indicates that, on average, the model's predictions deviate from the actual closing prices by approximately 0.86 units. Given the scale of stock price movements, this level of RMSE suggests that the model has a moderate level of accuracy in predicting stock prices.

Conclusions and Recommendations

The Linear Regression model exhibited moderate effectiveness in predicting stock prices, leveraging key features such as open, high, low, volume, and moving averages. Despite the simplicity of the model, it managed to provide valuable insights into the factors influencing stock price movements, proving its utility in the financial analysis domain.

References

- Kaggle for the dataset.
- Scikit-Learn documentation for modeling techniques.
- Various academic papers and books on financial time series forecasting.