Predictive Modeling of Stock Prices Using Machine Learning

-Mihir Gupta

Abstract

This document outlines the usage of a machine learning model to analyze and predict stock prices, specifically focusing on the S&P 500 Exchange Traded Fund (ETFs) VOO. Utilizing the Scikit-learn module in Python, the project demonstrates a robust approach to data preprocessing, model selection, and evaluation, achieving an accuracy rate of 87% in stock price predictions. The findings indicate that certain derived features can contribute significantly to prediction accuracy, making them valuable for investors and analysts. However, further addition of features that closely correlate to fluctuations in daily prices such as crude oil prices and a real-time index measuring positivity of media activity regarding the S&P 500 are a work in progress and can add significantly to the accuracy of the model.

Introduction

Predicting stock prices presents both a significant challenge and a critical need in financial markets. Accurate predictions can inform investment strategies and risk management practices. This project aims to develop a predictive model using historical stock price data for S&P 500 ETFs, leveraging machine learning techniques to enhance the effectiveness of traditional financial analysis. It is important to note that this project does not consider 2020 and 2021 that were affected by the pandemic in its training or testing datasets.

Methodology

The project utilized Python's machine learning libraries, including Scikit-learn, to develop the predictive model. The code was structured as follows:

1. **Data Acquisition and Preparation**: Historical stock price data for VOO was obtained using the yfinance library. The dataset included key variables, such as opening, closing, high, and low prices.

2. **Feature Engineering:**
    - A new target variable was created: target, which indicates whether the closing price of a given day will be higher than that of the following day.
    - Additional features were engineered:
        - open-close: The difference between opening and closing prices.
        - low-high: The difference between the day's low and high prices.
        - is_quarter_end: A binary variable indicating whether the date is the end of a financial quarter.
3. **Data Preprocessing:**
    - The date column was converted to a datetime format to facilitate further analysis.
    - Features were standardized using StandardScaler to improve model performance.
4. **Data Splitting:** The dataset was divided into training and testing sets, with the training data encompassing records up to 2019 and the testing data from 2022 onward.
5. **Model Selection:**
    - Three machine learning models were selected for training:
        - Logistic Regression
        - Support Vector Classifier (SVC) with a polynomial kernel
        - XGBoost Classifier
6. **Model Training and Evaluation:** Each model was trained on the training set and evaluated on the test set, measuring training and validation accuracy using the area under the ROC curve (AUC).

Results

The results indicated varying levels of model performance:

- Logistic Regression:

    - Training Accuracy: 86.64%
    - Validation Accuracy: 88.75%
- Support Vector Classifier (SVC):
    - Training Accuracy: 83.99%
    - Validation Accuracy: 86.41%
- XGBoost Classifier:
    - Training Accuracy: 98.90%
    - Validation Accuracy: 86.30%

The Logistic Regression model yielded the highest validation accuracy, indicating its effectiveness in this application context. The high training accuracy of the XGBoost model suggests potential overfitting, necessitating careful model tuning for practical implementations.

Discussion

The results of this project demonstrate the importance of feature selection and engineering in predictive modeling. The derived features, particularly the open-close and low-high differentials, provided essential insights into market behavior. Throughout the training and evaluation process, challenges such as overfitting were noted, emphasizing the need for rigorous validation techniques. The addition of more highly correlated features and implementation of real-time prediction systems will improve the model’s score. It will be interesting to observe if the Logistic Regression model will continue to outperform the other models with additional features.

This project further demonstrates the ability of the models to maintain high accuracy and showcases the utility of machine learning in finance.