Machine Learning Engineer Assignment

Introduction

I have developed an LSTM model for predicting stock price movements using historical financial data. I have done EDA and statistical analysis to better understand the data and get insights to model it better. I used Pfizer (PFE) stock for the demonstration.

The parameters of the model are stored in a config file, consisting of three main categories. These can be adjusted to run the model for different scenarios.

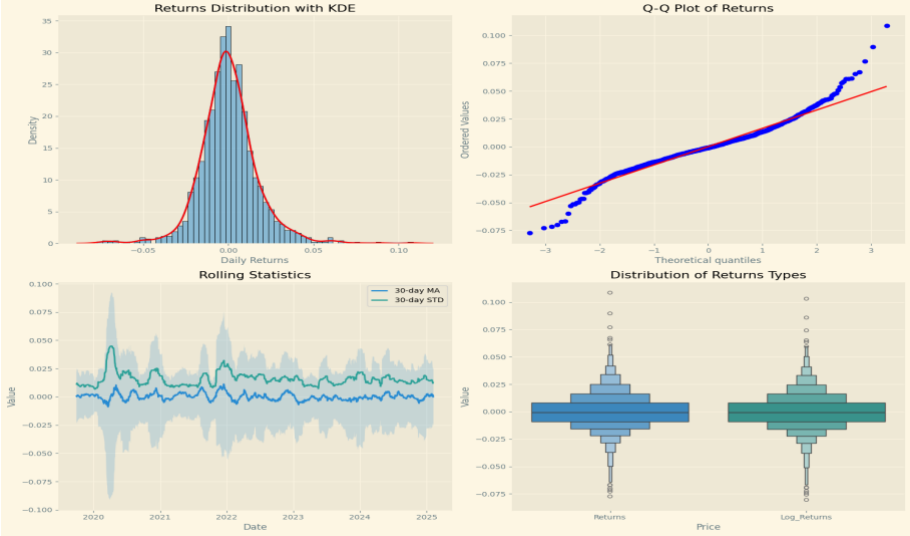
1) Data: Sets sequence length, dataset splits, history window, and target stock.
2) Feature Engineering: Specifies indicator windows, thresholds, and lookback periods to compute technical signals like momentum, volatility, and trends.
3) Model: Configures the LSTM architecture for things like layer sizes & dropout and training settings (batch size, epochs, learning rate, early stopping).

Data Collection

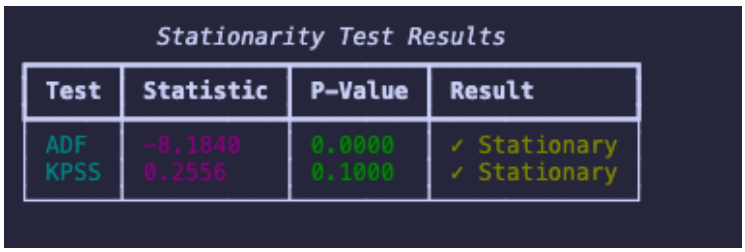
1) Extracted stock market data using Yahoo finance.
2) Selected key financial features: Open, High, Low, Close, and Volume.

Exploratory Data Analysis

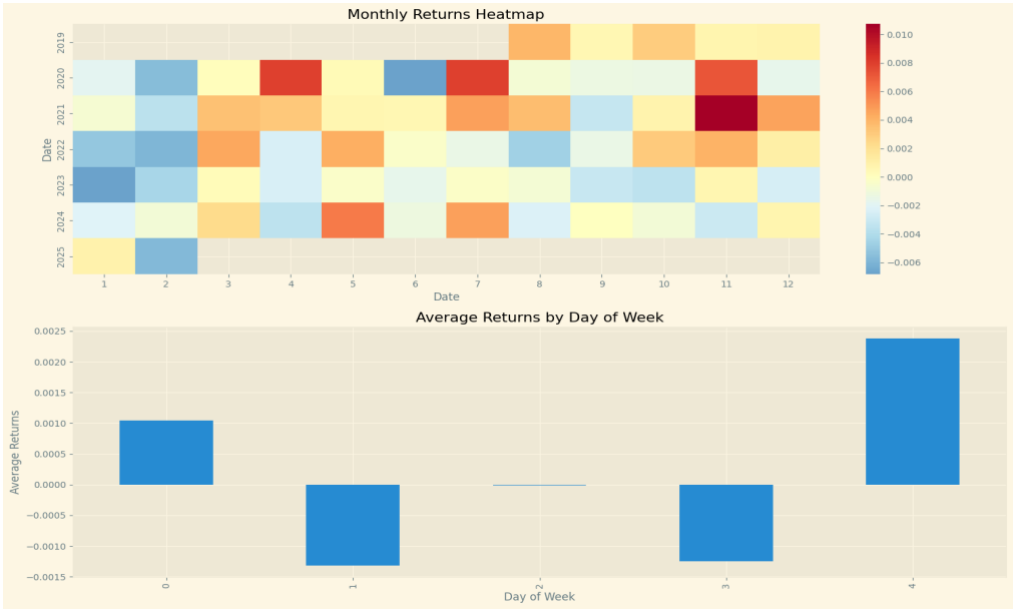
1) Visualized stock volatility over time, indicating periods of high variance. Distribution plots showed almost normal distribution.



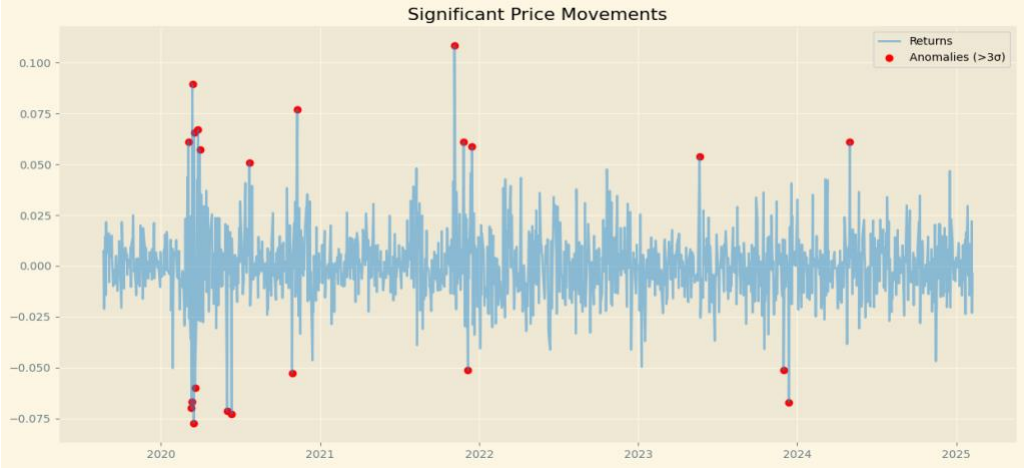
2) Confirmed stationarity which led me to use raw closing prices.



3) Analyzed seasonal trends in stock returns to consider using time-based features.



4) Identified significant market movements and anomalies



Feature Engineering

1) I calculated technical indicators like RSI, MACD, Bollinger Bands, ATR, ADX, OBV, Stochastic Oscillato which capture momentum, volatility, and trend strength.
2) I also calculated volume & fundamentals indicators which calculate a rolling volume average and volume momentum as well as integrating basic fundamental data such as EPS and dividend yield.
3) I then applied Recursive Feature Elimination to select the most relevant features.

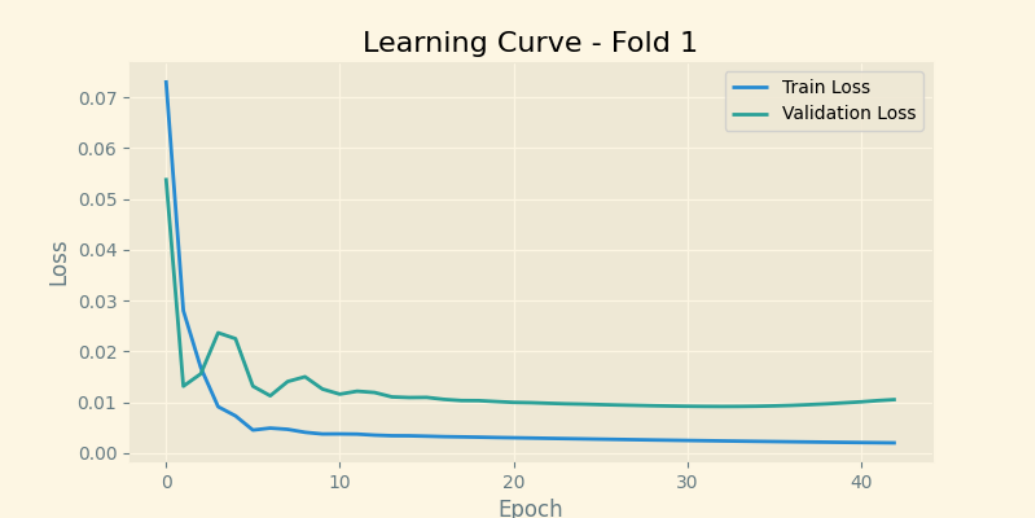
Data Preparation

1) Scaled features using Robust Scaler to handle outliers and maintain distribution.
2) Created sequences for LSTM input.
3) Split data into training, validation, and test sets while maintaining time-series nature to prevent data leakage.
4) The target is defined as the next day's closing price for this case as target is stationary.

Modeling Approach

1) Created an LSTM model and applied time series cross validation.
2) Implemented dropout layers and L2 regularization to prevent overfitting on the training set.
3) Used early stopping and learning rate reduction strategies to optimize convergence.
4) Utilized the Adam optimizer, because of hardware restrictions and convergence speed.

The learning curve shows early convergence for both train and validation. Train and validation loss also both are low indicating no overfitting.



Metric Selection

RMSE was prioritized as it penalizes large errors more than MAE, and it is often the choice in financial forecasting especially since they have a lot of anomalies. R^2 score was also analyzed to evaluate how well the model explains variance.

Results & Insights

The model performs well in minimizing prediction errors, with RMSE below 1 and an R2 score of 75%.

Metrics:

1) **MAE**: 0.5347
2) **MSE**: 0.4604
3) **RMSE**: 0.6785
4) **R2**: 0.7522