Model Evaluation

Data Preprocessing + Exploration

The data cleaning and exploration steps involved checking for missing values and converting the data to the appropriate formats, such as floats for numerical columns and datetime for date columns. Pairplots were created to analyze the relationships between features, with the hue set to the "Close" column, revealing positive linear relationships for most features except for the "Volume" column. Additionally, the correlation matrix was computed to quantify the correlation between features, confirming the high positive correlations observed in the pairplots. Kernel density estimate (KDE) plots were used to visualize the distribution of closing prices, showing that the majority of prices hover around the $100 mark or less. Histograms were also employed to provide a more detailed analysis of price distribution, indicating clustering around the range of 50 to 100.

Random Forest:

The dataset was split into features (Volume, Open, High, Low) denoted as X, and the target variable (Close/Last) denoted as Y, for the purpose of training, validation, and testing. In order to ensure normalization and minimize discrepancies between features, the data was standardized using a StandardScaler. A Random Forest model was then constructed to predict the closing price. Hyperparameters such as max_depth, min_samples_leaf, and n_estimators were selected to optimize the model's performance. Grid Search and K-fold cross-validation techniques were employed to search for the best combination of hyperparameters. The optimal values obtained were a max_depth of None, min_samples_leaf of 1, and n_estimators of 100.

The model's performance was evaluated using several metrics. The Mean Squared Error (MSE), the Mean Absolute Error (MAE), and the R-squared (R^2) score. The evaluation metrics for the Random Forest model were as follows: MSE of 0.89, MAE of 0.61, and R^2 of 0.99. These metrics suggest that the model performs well in predicting the closing prices, with low errors and high explanatory power.

To assess the model's accuracy, the actual and predicted prices were graphed and compared. The graph showed that the predicted prices closely aligned with the actual prices, indicating a good fit between the model and the observed data. Lastly, using the trained Random Forest model, the predicted closing price for the year 2021 was determined to be $265.48.

LSTM (Neural Network):

The dataset was divided into training, validation, and testing sets in a similar manner to the Random Forest approach. To preprocess the data for the LSTM model, the MinMaxScaler was used to scale the values between 0 and 1, ensuring consistent ranges across the features. For the LSTM model, a sequence length of 30 was defined, which represented a month's worth of data. The create_sequences function was then implemented to generate input sequences for both X (input features) and Y (target variable). The data was trained and tested using these sequences.

The LSTM model consisted of four layers: three LSTM layers with dropout regularization and one Dense layer. The model was compiled with the mean squared error (MSE) loss function and the Adam optimizer. Early stopping was applied during the training process to prevent overfitting by monitoring the validation loss. After training the model, the loss value and validation loss were printed, indicating the performance of the model. The loss value was 0.0011, indicating a low error on the training data, while the validation loss was 0.03, suggesting good generalization performance on the unseen validation data.

To visualize the training progress, a function was created to plot the loss history of the model. The plot showed that as the number of epochs increased, both the training loss and validation loss decreased, indicating that the model was learning and improving over time. Lastly, the trained LSTM model was used to make predictions for future prices. The predicted prices for year 1, year 2, and year 3 were 152.59, 143.35, and 154.75, respectively. These predictions provide insights into the potential future trends in the stock market and can assist in making informed investment decisions.

SQL

SQL was implemented to query data from the recent Apple Stock Prices from the years 2022 to 2023. In the SQL implementation, the maximum price in 2022 was found to be 174.55, while the minimum price was 126.04. This resulted in a range of 48.51. Additionally, the 25th percentile was determined to be 137.29, and the 75th percentile was 152.68. To calculate the lower and upper boundaries for potential outliers in 2022, a nested query was used. By applying the formula lower boundary = Q1 - ((Q3 - Q1) * 1.5), and upper boundary = Q3 + ((Q3 - Q1) * 1.5), the lower boundary for 2022 was set at 114.20, and the upper boundary was set at 175.77.

Similarly, the SQL implementation was extended to analyze the data for the year 2023. The maximum price for 2023 was 193.97, the minimum price was 125.02, and the range was calculated as 68.95. The lower boundary was determined to be 98.30, and the upper boundary was set at 210.39. By utilizing nested queries and calculating the quartiles, these boundaries were established to identify potential outliers in the dataset.