Slide 5 (Data: quick overview | Definition)

- Stocks froms various sectors: technology, energy (primarily oil & gas), food, real estate, some others
- "High" is considered as Y label (probably better to use "Close" as it should help deal with outliers)
- stock data is from 2010
- stock data does not exhibit attrition (the process of reducing something's strength or effectiveness through sustained attack or pressure).

Slide 6 (Data: quick overview | Preprocessing)

Random Forest and XGBoost can be used for important feature selection by ranking features based on their importance scores, which are derived from the frequency and impact of their use in splitting decision trees within the models.

Standardization: enables different features to be compared on the same scale and improving the performance and convergence rate of many machine learning algorithms.

Slide 7 (Data: quick overview | First Differencing)

- Tool used by statisticians.
- Since data is a combination of daily and quarterly points, data moments are very important for medium-term forecasting.

Slide 8 (Feature selection, Splitting & Scaling)

- With these algorithms, we compute feature importance scores based on the contribution of each feature to the model's predictive performance.
- Decrease in node impurity for Forest and Total Gain (improvement in accuracy brought by a feature to the branches it is on) for XGBoost.

Slide 9 (Train/Validation splits)

By implementing a novel data splitting technique that includes multiple validation splits throughout the time series, we anticipate an increase in model generalizability and prediction quality. This approach aims to provide the model with diverse evaluation samples at different temporal contexts.

Slide 10 (Sliding window)

The Sliding Window technique in the PatchTST (Patch-based Time Series Transformer) model involves sequentially dividing a time series into overlapping

windows of fixed size. Each window serves as an input patch, enabling the model to capture temporal dependencies and extract local patterns from the time series data while maintaining spatial context.

Ways to imporove

Data

- Probably better to use "Close" as it should help deal with outliers
- Probably makes sense to include more indicators (since feature selection puts big weights)
- First-difference is a tool of statistical analysis (stationarity is a necessity for econometric models), makes sense to try to fit the model with data as is