## Preprocessing and Feature Engineering for Stock Price Prediction

This report outlines the process of preprocessing and engineering features from time series data for stock price prediction. These steps are crucial for preparing the data for machine learning models to effectively learn patterns and make predictions about future stock prices.

### 1. Data Acquisition

The first step involves obtaining historical stock price data for a specific company or index. This data can be retrieved from various sources such as financial data providers ([https://pypi.org/project/yfinance/](https://pypi.org/project/yfinance/" \t "_blank)), online APIs, or financial websites.

**Important considerations:**

* **Timeframe:** Decide on the historical period you want to analyze (e.g., daily data for the past year, hourly data for the past month).
* **Data points:** Select the relevant data points typically including Open, High, Low, Close, and Volume for each time step.

### 2. Data Preprocessing

Once you have the raw data, it's essential to prepare it for further analysis. Here are some common preprocessing tasks:

* **Handling missing values:** Address missing data points by filling them with appropriate strategies like interpolation, forward fill (using the previous value), or dropping rows with missing values (if a small portion).
* **Outlier detection:** Identify and handle outliers that might skew the data distribution. You can use techniques like z-scores or interquartile ranges (IQR) to detect outliers and decide on removal or winsorization (capping extreme values).
* **Data consistency:** Ensure consistency in data format and units. Standardize date formats and convert currency values if necessary.

### 3. Feature Engineering

Feature engineering involves creating new features from the existing data to improve the model's ability to learn and predict future prices. This step is crucial for extracting meaningful insights from the time series data.

**Here are some common feature engineering techniques for stock price prediction:**

* **Technical indicators:** These are mathematical calculations based on past price and volume data that aim to capture trends and momentum. Examples include:
  + **Moving Average Convergence Divergence (MACD):** Measures the relationship between two moving averages of closing prices.
  + **Relative Strength Index (RSI):** Indicates whether a stock is overbought or oversold based on recent price changes.
* **Lag features:** Lag features involve incorporating past values of existing features (e.g., past closing prices, volume) into the model. This helps the model learn from historical trends.
* **Volatility measures:** Features like standard deviation or Average True Range (ATR) capture the volatility of the stock price over a specific period.

**Additional Considerations:**

* **Normalization or standardization:** Scaling features to a common range (e.g., 0-1 or with zero mean and unit variance) can improve model performance.
* **Feature selection:** Analyze the engineered features and select the most informative ones for your model. This can be done using techniques like correlation analysis or feature importance scores from machine learning models.