**REPORT\_BUSINESS CASE #1**

| **Data Transformation**    **Missing Values Handling**    **Price changes**    **Distribution of Price and Volume**    **Open Price vs Volume**    **Outliner**    **Quarter Comparison**    **Relative Strength Index(RSI)**  **Moving Average (MA**  **Moving Average Convergence Divergence**    **Bollinger Band**      **Check Stationarity and Dickey-Fuller test:**      **ACF & PACF to find the suitable p,d,q**      **AIC and min-max scaling data for the best ARIMA model**      **LSTM Model** | **1. DATA CLEANING** - We could see that data for some of the dates was missing the reason for that was on **weekends** and **holidays** Stock Market remains closed hence no trading happens on these days  - Most of the columns were in the **wrong data type** when they were currently in object => Team processes to convert them to the correct type:   * **Date -> date time** * **Price, Open, High, Low, Vol, Change % -> float**   - The dataset has **no missing values.** - Because this was a table that stores transactions in time, the Date feature would have unique values => Dataset **did not have any data with duplicate dates.**  **=>** Dataset had **1308 rows × 6 columns** **2. INITIAL FINDINGS/ OVERVIEW****2.1 Market situation**  - The price chart is quite gloomy in the period 2018 - Q1 2020. From Q2 2020, the price is showing an **upward trend** and **peaks** around the end of 2021 at 43895.8.Then the price tends to decrease again in the next 1 year.   * 2020: Vietnam's economy in the first 11 months of 2020 continued to maintain recovery momentum under new normal conditions. → Conditions for steel industry growth. * 2021: The steel industry grew strongly because Vietnam businesses boosted exports and output. HPG Group has seized the opportunity to conquer foreign markets. * 2022: The impact of Covid caused raw material prices to increase, interest rates to increase, and banks to tighten credit, causing HPG Group's import-export output to decrease.  **2.2 Component:** 2 Price levels peaked (10.000 - 15.000 and around 35.000-40000) despite small Volume Trading fluctuations   * Shows the **reliability** and **reputation** of stock prices without any bad stock market activity (bear traps, bull traps, stock market bubbles...) * Or it can be interpreted that there are 2 customer groups with different "investment taste" of HPG stock price * Shows the participation of many transactions before or after the Price peaked.  **2.3 Open Price** - Open Price could be influenced by many macro and micro factors (GDP, CPI...), not necessarily from Trade Volume.  * In Q3-2022 When the transaction volume reached its second peak => the Price bottomed and immediately after that, when the price increased, the transaction volume immediately decreased significantly.  **2.4 Outliner:** Price is concentrated on 1 point, but for Volume it's the opposite: spread evenly over **a long distance** - Stock shark is a term that refers to large investors with **strong financial potential** and advantages in information in the market  * The steel industry's trade situation changes significantly (when the industry's trade situation peaks in 2021-2022) → Potential in stock Price would be profitable  **2.5 Quarter Comparison**: HPG Group also was paying attention to this and trying to had good results at the **end of the quarter.**  * Prices was higher in the months which were quarter end as compared to that of the non-quarter end months. * The volume of trades was lower in the months which were at the quarter end.  **2.6 Relative Strength Index**: RSI is above the **overbought or overvalued** level in the period 2020-2022.  * It was ​indicated that the Price was overbought, suggesting it may be overvalued during in 2018-2020   **2.7/ Moving Average (MA):** The **longer** the milestone, the **smoother** the MA line will be, but MA suitable with the **Bollinger Band** line, and parameter for the **ARIMA** model   * MA line 20 was the default index to look at the chart without too much volatility, is smoothing does not cause more noise than other indicators  **2.8 Moving Average Convergence Divergence(MACD)**  * MACD was significant in economics as it helped traders understand whether the **bullish** or **bearish momentum** in the price was **strengthening** or **weakening**. * A signal line (9-day EMA of the MACD) was then plotted on top of the MACD to function as a trigger for buy and sell signals.  **2.9 Bollinger Band**: Price fluctuations were **not exceed** compared to the Rollinger band line  * When the stock price reached the **Upper Band** (i.e., it was 2 times the standard deviation above the MA-20), it indicated that the stock might be **overbought** and had a tendency to **decrease** in price. * Conversely, when the stock price reached the **Lower Band** (i.e., it was 2 times the standard deviation below the MA-20), it suggests that the stock might be **oversold** and had a tendency to **increase** in price.  **3. FEATURE SELECTION AND ENGINEERING** From the correlation matrix, we observe that the open, high, low, and **Price** are highly correlated with each other, while the Vol. has a very low correlation with the other features. Given the equivalent level of impact of open, high, low, and close, we decide to select the **Price** as the feature for our stock price prediction model.  => In conclusion, due to the high correlation among Open, High, Low, and Price, and the similar impact on the stock price prediction, we choose **Price** as the primary feature for our predictive modeling. **4. ARIMA MODEL:****4.1/Check Stationarity and Dickey-Fuller test:**  * The Dickey-Fuller test results with a test statistic of -1.186 and p-value of 0.679 indicate we failed to reject the null hypothesis of a unit root, suggesting the time series **data is non-**stationary**.** * Using the **Differencing method,** The Dickey-Fuller test results with a test statistic of -7.711 and a p-value of 1.26e-11 indicate a strong rejection of the null hypothesis of a unit root. This suggests the time series data is v**ery likely stationary.**   **4.2/ACF & PACF to find the suitable p,d,q** **ACF plot:** While the Dickey-Fuller test initially indicated stationarity, the autocorrelation function (ACF) plot reveals strong evidence of non-stationarity in the time series data. The high and slowly decaying autocorrelations suggest the presence of a trend or other non-stationary components that were not captured by the Dickey-Fuller test.  **PACF plot**: A potential ARIMA model for this time series could be an AR(2) model. This model would include two autoregressive terms, corresponding to the significant spikes at lags 1 and 2.  **4.3/ AIC and min-max scaling data for the best model**  * After scaling the data using min-max scaling, an automated search for the **best ARIMA model was conducted**, evaluating various parameter combinations based on the Akaike Information Criterion (AIC).   => The ARIMA model with parameters (2, 1, 0) was identified as the best fit for the data, achieving the lowest AIC value of -6116.61. **4.4/Predict the price** The analysis reveals a time series with a clear downward trend in the "load" variable over the observed period.   1. While the forecasting model demonstrates good accuracy in the short term, its performance deteriorates as the forecast horizon extends, suggesting limitations in capturing long-term dynamics. 2. The discrepancy between the initial Dickey-Fuller test results, indicating stationarity, and the autocorrelation function plot, suggesting non-stationarity, necessitates further investigation into the data's underlying properties and potential model refinements.   Overall, while the model shows promise for short-term forecasting, it requires further improvement to ensure accurate and reliable long-term predictions. **4.5/Compare predictions to the actual load**  * The combination of a low MAE (0.0243), a relatively low RMSE (0.0326), and a high R² (0.9638) indicates that the model performs well in predicting the time series. The predictions are generally accurate, with a good fit to the test data * However, the slightly higher RMSE compared to the MAE suggests that there might be a few larger errors that are influencing the overall model performance. It's worth investigating these outliers to understand if they are due to specific events or patterns in the data that the model could potentially be improved to capture.  **5. LSTM MODEL:** The LSTM model was implemented to predict prices by following these steps. First, the price data was filtered and split into a training set (before March 1, 2022) and a test set (from March 1, 2022, to March 31, 2023).  The data was then normalized to the range **[0, 1]** to improve the model's training efficiency. Next, time series sequences of **60** steps were created for the training data, using the price at step 61 as the target value. The data was converted into numpy arrays and reshaped to fit the input requirements of the LSTM.  The LSTM model was built with two LSTM layers (**128 and 64 units**) and two Dense layers, then compiled with the Adam optimizer and trained on the training data for **1 epoch**. Subsequently, time series sequences for the test data were created and the trained model was used to predict prices. The predictions were inversely transformed back to the original scale. The model was evaluated using **RMSE, MAE, and R²** score metrics, with results of **1440.2883**, **1181.5377** and **0.7185**, respectively, indicating that the model effectively learned the data trends. **6. MODEL COMPARISON**  | **Metric** | **ARIMA** | **LSTM** | | --- | --- | --- | | Mean Absolute Error (MAE) | **0.0243** | **1181.5377** | | Test RMSE (Root Mean Squared Error) | **0.0326** | **1440.2883** | | Test R-squared (R²) | **0.9638** | **0.7185** |   **In conclusion:** ARIMA outperforms LSTM across all three metrics, demonstrating significantly better accuracy, lower error magnitudes, and a stronger ability to explain the variance in the data |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |