Week ML5 – Transformer NN Model

H. Majeed, S. Bukhari, S. Gohar

December 2, 2025

1. Result Presentation and Interpretation of Model Training and Results

1.1 Model Training

The model was trained using Adam optimizer and Mean Squared Error as the primary loss function. Early stopping was incorporated to ensure there is no overfitting, and the model training can be stopped once the model reached optimal generalization performance.

As shown in the Fig 1. Both the training loss and validation loss converge smoothly for univariate data, showing stable learning.

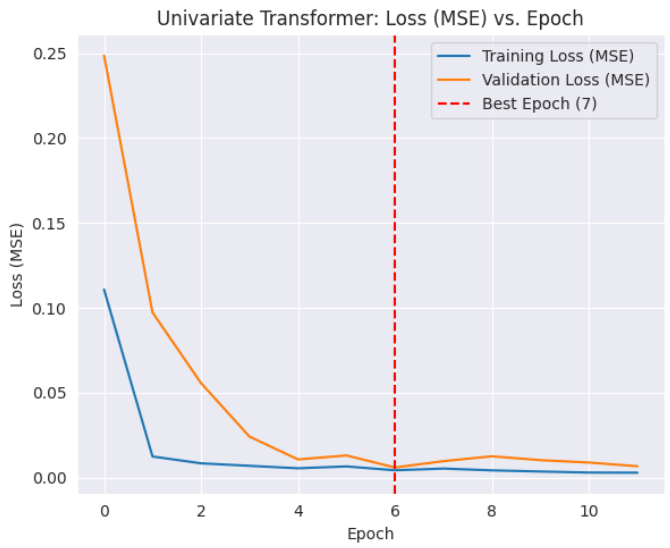

Figure 1: Transformer training and validation loss comparison on univariate data to epochs

To further evaluate the model robustness we extended the datasetto a multivariate setting by incorporating trading volume as an additional feature. A log transformation and Minmax scaling were applied to both features. The resulting curves are shown in Fig 2.

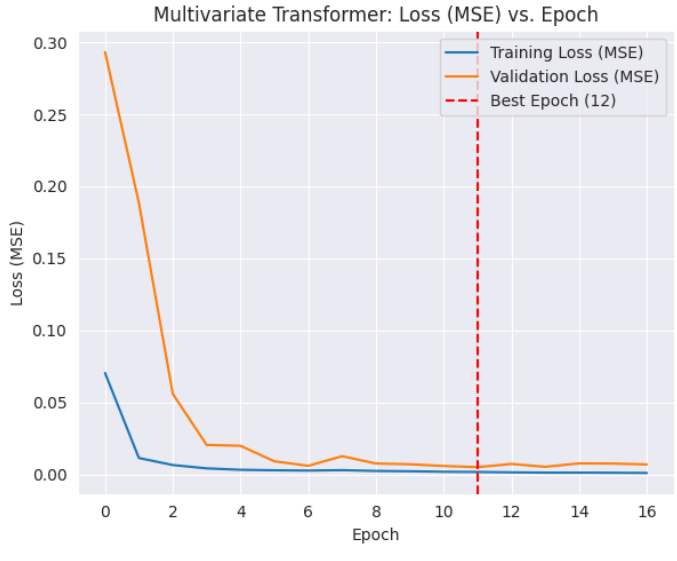

Figure 2: Transformer training and validation loss comparison on multivariate data to epochs

1.2 Forecasting Results

The performance of the univariate was evaluated against the Naïve Baseline prediction (predicting the last known value) as well as LSTM. The univariate transformer provided clean, accurate, forecast as shown in Fig 3. which successfully tracked most of the major volatile swings even beating the LSTM and Naïve Baseline.

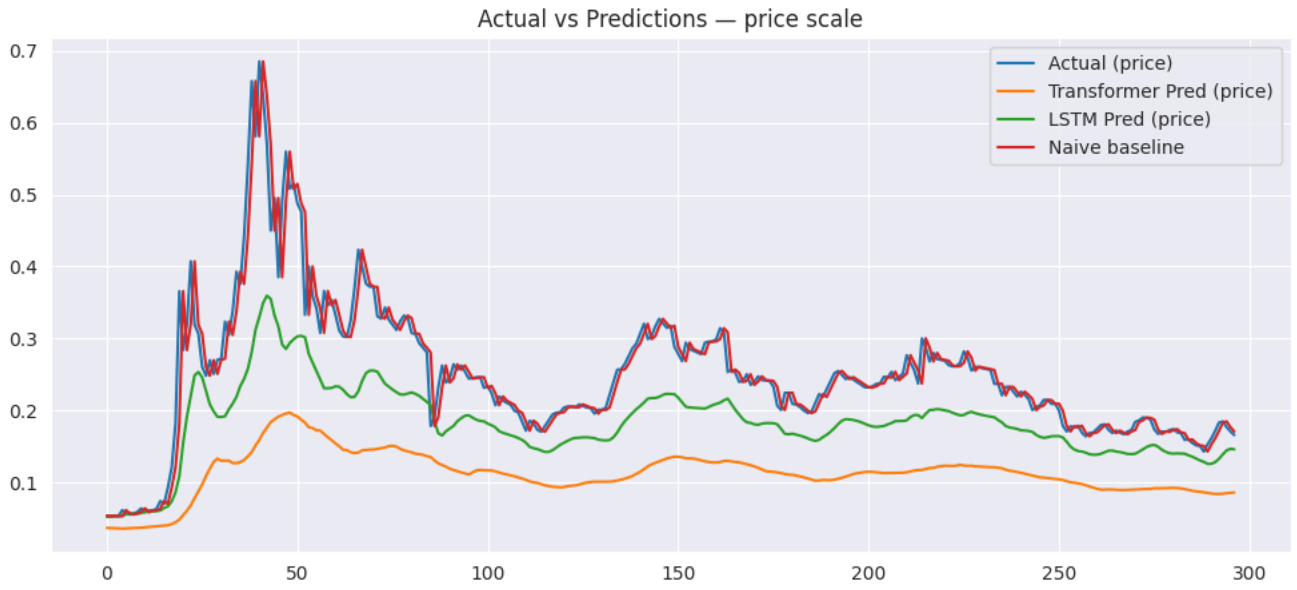

Figure 3: Transformer prediction comparison on univariate data with different machine learning techniques

In contrast, the multivariate transformer struggled in the prediction and consistently underestimated the true magnitude and volatility of closing price as shown in Fig 4.

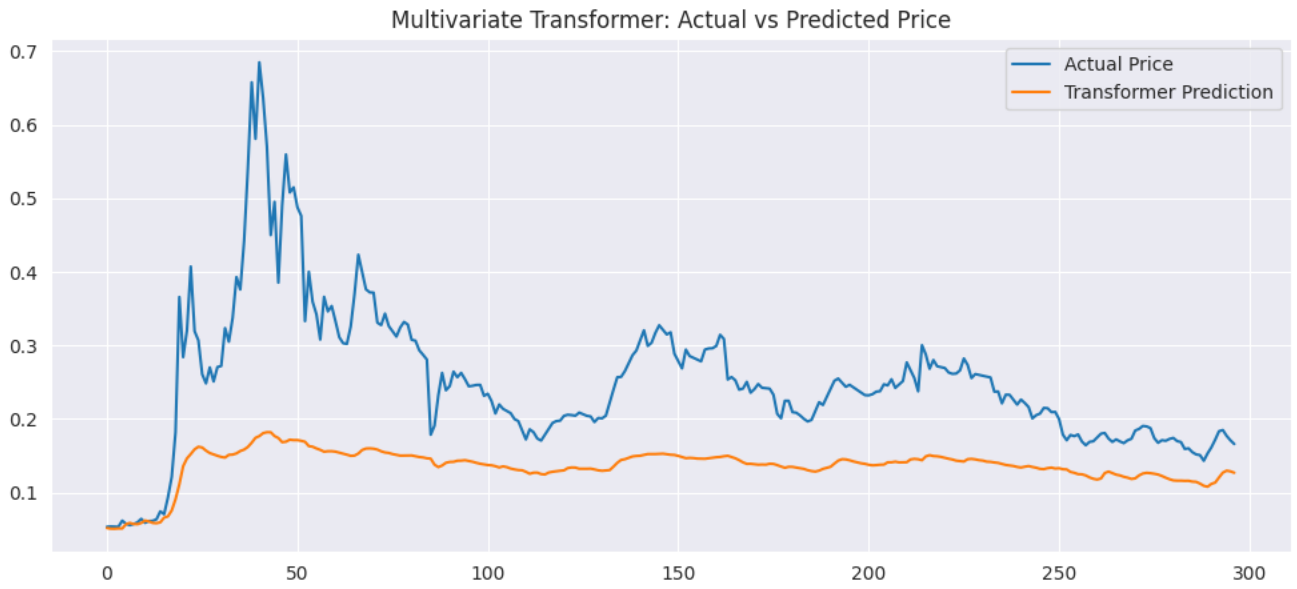

Figure 4: Transformer prediction comparison on multivariate data

A likely explanation to this poor performance is since the model's input feature doubled the transformer's attention mechanism must now learn both the temporal relationships and the cross-feature relationship. The weak and noisy signal from volume feature may have caused feature dilution, in which attention weights diffused across less predictive volume data. As the complexity of the model increased, but the training size of the data remained the same which led to model underfitting in case of multivariate data.

2. Plan for Model Optimization

The primary cause of the smooth output in multivariate could be linked to scaling logic. A dedicated Minmax scaler must be fitted only on the target variable closing price for the output predictions. The current use of the 2-feature scaler with zero placeholder for the second feature introduces systematic bias during the inverse

transformation, that causes observed magnitude underestimation. To combat the underfitting of model we could increase the model capacity by increasing *d_model* from 64 to 128 and potentially adding a second transformer encoder block. Since the model is already underfitting we could decrease the dropout rate from 0.1 to 0.05 or even 0.0. Lastly, a correlation analysis could be performed to confirm if the next day price is correlated with volume or not. We could then propose a feature engineering strategy to use moving average for the volume or some other calculated column instead of raw volume to get possible better results.