Reese Dychiao
rkd9714
Data Bootcamp
Final Project

Predicting Global Carbon Dioxide Emissions

I. Introduction to Problem & Data

A. Problem Statement

The 21st century faces many challenges which vary in severity, duration, and impact. Among these, climate change stands out as a critical issue as it affects all living beings, presents long-term challenges, and has profound effects on our ecosystems. A key source of global warming is carbon dioxide (CO2) emissions–a metric every individual contributes to. In order to approach this issue, scientists must take measures to determine CO2 emissions in the future and sequentially draw informed solutions and strategies to effectively address this problem. This project aims to make predictions of future CO2 emissions based on historical data. By exploring a range of time series models, this project will select the most accurate projections, providing valuable insights for developing effective climate change mitigation strategies.

B. Dataset

The dataset used in this project is taken from Kaggle. It presents CO2 emissions in kilotons (kt) from 2019 to 2024. For each date, there is an entry for major countries and each sector within those countries. Additionally, the dataset provides daily global CO2 emissions for each sector. Since a time series model requires a single entry for each date, this project will utilize the global data by averaging emissions across all sectors. This approach will enable the model to generate broad predictions that span various industries and regions.

II. Data Preprocessing & Preliminary Exploration

The initial dataset included columns for the country, date, sector, value, and timestamp. To prepare the data, the following transformations were applied. The resulting data frame is shown in Figure 2.1.

1. The index was set to the date column, cast as a DateTime object.
2. The data was filtered to include only the global emissions data.
3. The average emissions across all sectors were calculated.

This processed data can be used to plot a time series, as displayed in Figure 2.2. The plot reveals seasonal patterns, with CO2 emissions peaking at the beginning of each year, trending downward in the middle months, and then rising again towards the end of the year. This cyclical behavior suggests a recurring annual trend, which may be influenced by seasonal activities and industrial cycles.

date	value
2019-01-01	16.190175
2019-01-02	17.193493
2019-01-03	17.605170
2019-01-04	17.817105
2019-01-05	17.114223

Figure 2.1 Processed Data Frame

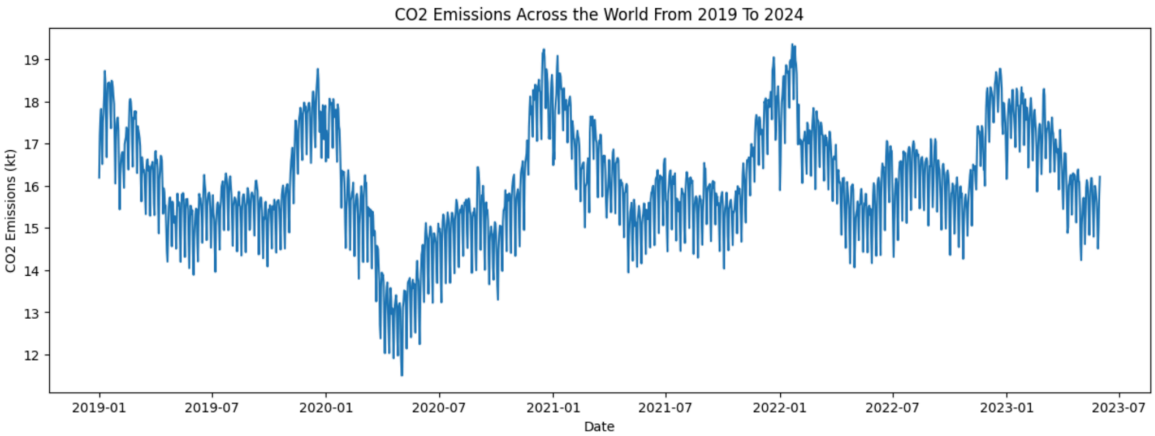

Figure 2.2 CO2 Emissions Across the World From 2019 to 2024

In addition, we can create autocorrelation and partial autocorrelation plots, as illustrated in Figure 2.3. These plots help analyze the relationships between lagged values and the time series data. The autocorrelation plot shows that the lagged values have a positive correlation with the series, as indicated by the bars extending beyond the significance bounds. Furthermore, the gradual decline of the bars suggests a long-term dependency in the data. The partial autocorrelation plot, on the other hand, indicates which specific lags are significant by showing the direct relationships between the time series and its lags.

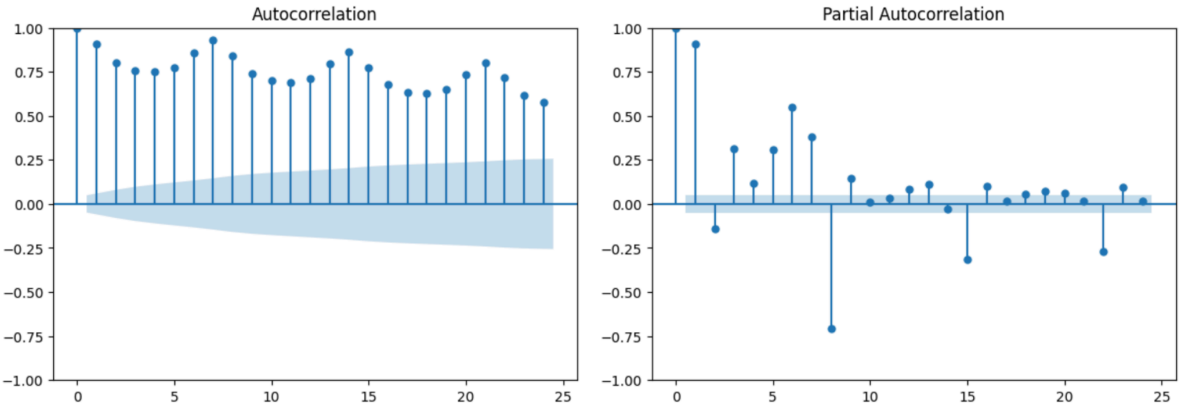

Figure 2.3 Autocorrelation and Partial Autocorrelation Plot

Lastly, before creating models to predict the data, we need to split the dataset into training and testing subsets. This allows us to evaluate the accuracy and performance of each model. Figure 2.4 shows the plot of the training-test split.

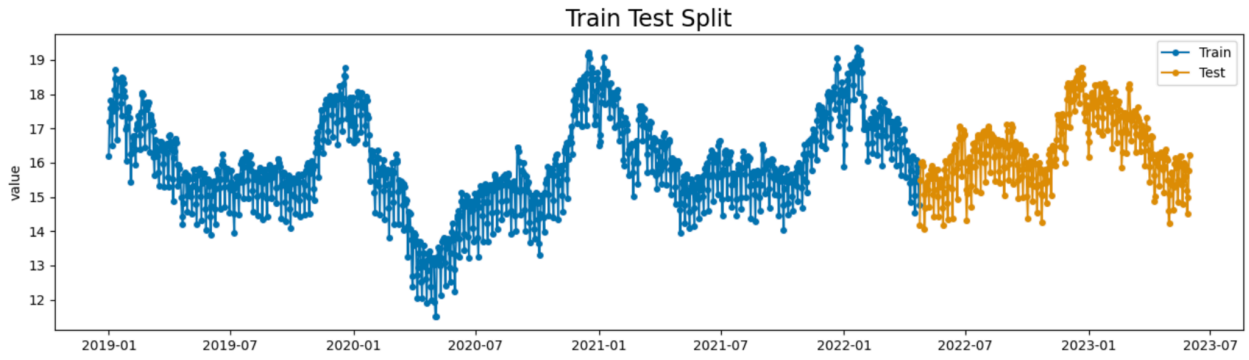

Figure 2.4 Plot of CO2 Emissions Data Train-Test-Split

III. Models & Analysis

A. Naive Forecasting

Naive forecasting relies on the assumption that the data from the previous period will be the same as the data in the next period. This approach is often used as a baseline model, against which other more complex models can be compared. A visualization of the naive forecasting model is shown in Figure 3.1.1. The model's performance will be evaluated using the Mean Absolute Percentage Error (MAPE), which measures the average error of the predictions. For this model, the MAPE is 0.1133. Moving forward, models that demonstrate a MAPE lower than this prove to have better performance.

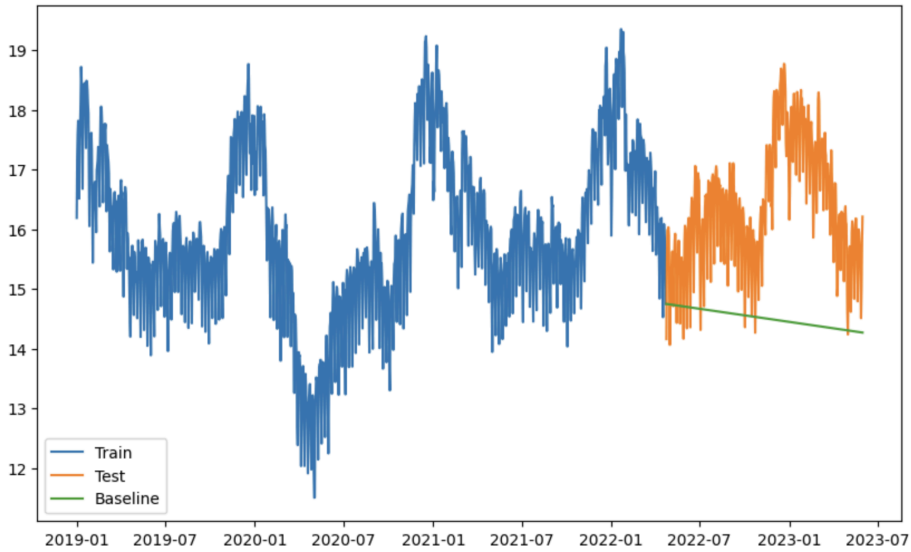

Figure 3.1.1 Naive Forecasting Model Plot

B. Simple Exponential Smoothing

The simple exponential smoothing model makes predictions by calculating a weighted linear sum of past values, with more recent observations receiving higher weights than those further in the past. Figure 3.2.1 illustrates the predictions generated by this model. With a MAPE of 0.0994, it outperforms the naive forecaster, indicating improved accuracy. However, the plot shows that the model still fails to account for patterns in the actual data.

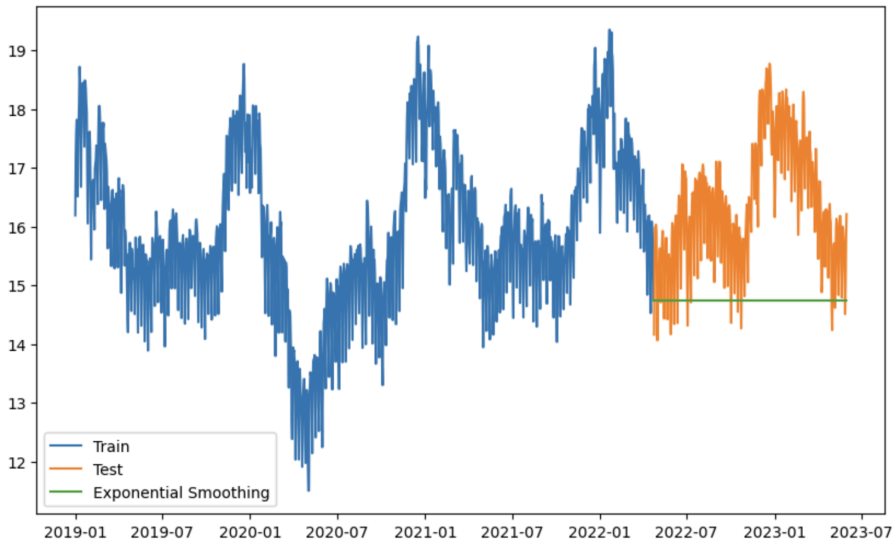

Figure 3.2.1 Simple Exponential Smoothing Model Plot

C. Holt-Winters'

The Holt-Winters model generates predictions by capturing three key data characteristics: trend, seasonality, and the seasonal period. The trend used in the model is 'additive' because the data appears to exhibit a linear trend without significant growth or decline over time. The seasonal period is set to 365 because the data is observed daily, indicating daily seasonality. The plot of the CO2 emissions data does not clearly indicate whether seasonality is additive or multiplicative, so we will create models using both approaches and compare their performance.

Figure 3.3.1 shows the Holt-Winters' additive model, while Figure 3.3.2 displays the multiplicative model. The additive model's MAPE is 0.0392, while the multiplicative model's MAPE is 0.0405. Although the difference in performance between the two models is minimal, we can conclude that the additive model performs slightly better. Notably, both models outperform the naive forecasting and simple exponential smoothing models, as they account for the seasonal patterns present in the data, leading to more accurate predictions.

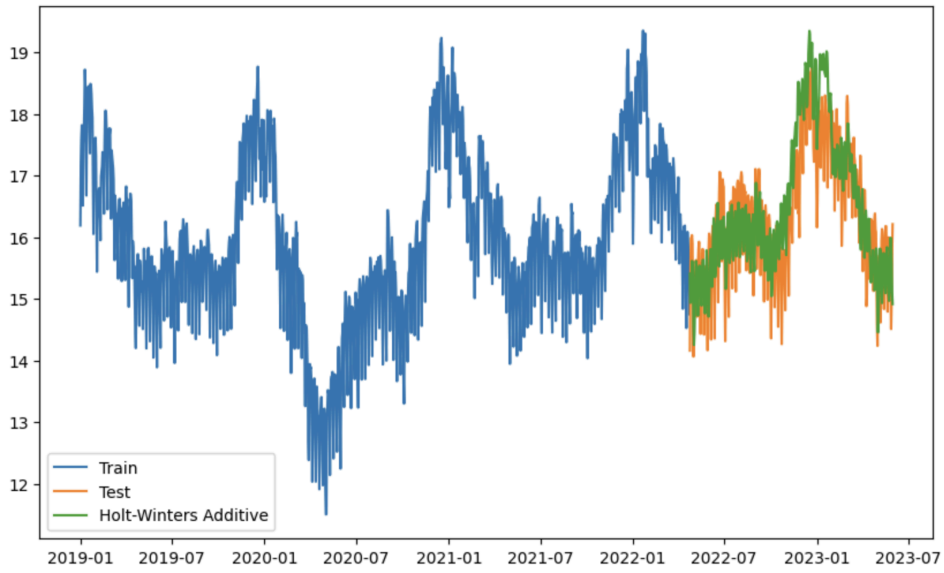

Figure 3.3.1 Holt-Winters' Additive Model Plot

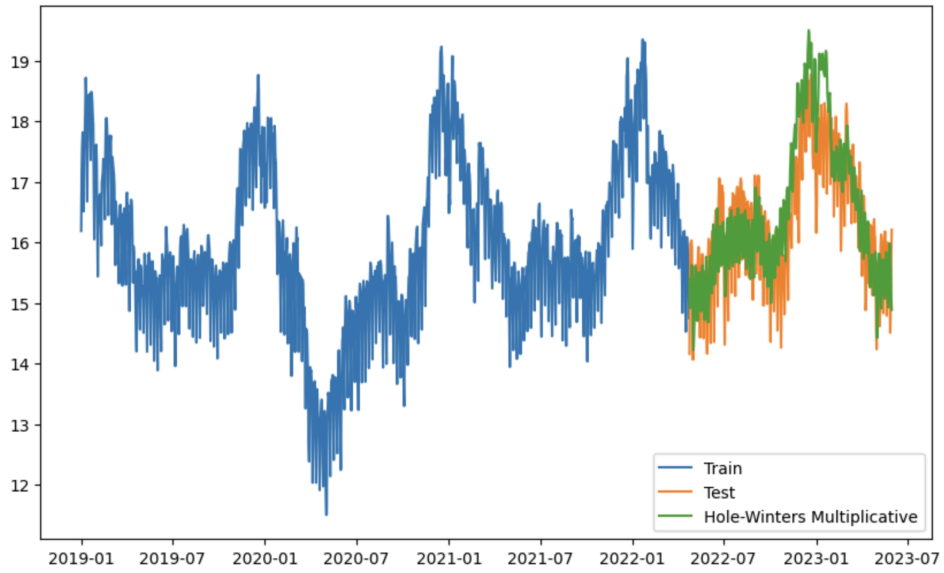

Figure 3.3.2 Holt-Winters' Multiplicative Model Plot

D. Autoregression (AR)

The AR model predicts future data points by using lagged values as inputs in a formula where each lagged value is multiplied by an autoregressive coefficient, which determines its influence on the prediction. The predictions made by the AR model are shown in Figure 3.4.1, with a MAPE of 0.0544. This indicates that the AR model performs better than both naive forecasting and simple exponential smoothing. However, it is outperformed by the two Holt-Winters' models. Based on the plot, this is likely because the AR model was unable to detect the seasonal patterns of the data,

whereas the Holt-Winters' models are designed to handle data with seasonal patterns, and therefore generate more accurate predictions.

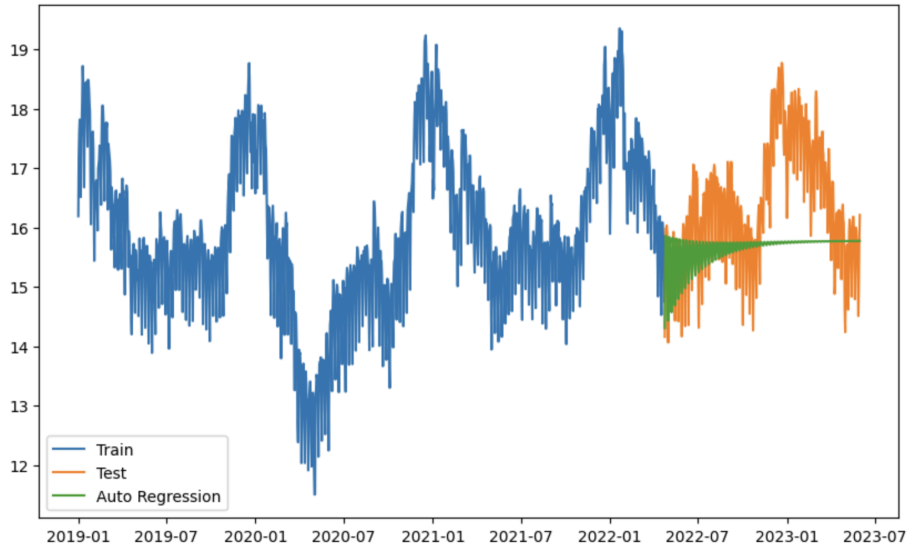

Figure 3.4.1 AutoRegression Model Plot

E. Autoregressive Integrated Moving Average (ARIMA)

Before fitting the ARIMA model to the training data, we must first ensure that the data is stationary. Stationarity means that the data's mean, variance, and autocorrelation structure do not change over time. To determine if the data is stationary, we can use the Augmented Dickey-Fuller (ADF) Test. The resulting p-value from the test on our data is 0.016. Since this is less than the significance level of 0.05, we reject the null hypothesis and conclude that the dataset is stationary. Therefore, we can directly apply the ARIMA model without differencing the data.

The ARIMA model is an extension of the autoregressive (AR) model. It uses past values to predict future data points and incorporates moving averages to improve prediction accuracy by accounting for forecast errors. The model is characterized by three parameters: the number of autoregressive terms (p), the number of differencing steps (d), and the number of moving average terms (q). In this case, we use ARIMA(1, 1, 1), indicating one autoregressive term, one differencing step, and one moving average term. Figure 3.5.1 shows the ARIMA model predictions, with a MAPE of 0.0804. This indicates that the ARIMA model performs better than naive forecasting and simple exponential smoothing. However, similar to the AR model, it did not capture the seasonality and patterns of the data. Consequently, the Holt-Winters' model presented a lower MAPE than the ARIMA model, demonstrating its strong ability to handle seasonality in the CO2 emissions dataset.

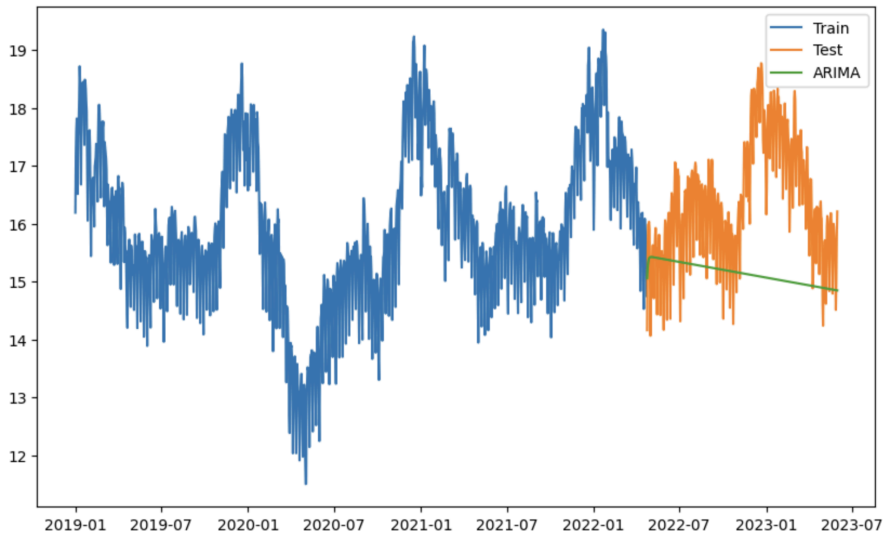

Figure 3.5.1 ARIMA Model Plot

F. AutoARIMA

The AutoARIMA model automatically determines the best parameters for the ARIMA model and applies them. The plot of this model is shown in Figure 3.6.1, with a MAPE of 0.0498, yielding results similar to those of the ARIMA model. The determined ARIMA parameters are (p, d, q) = (5, 1, 4), indicating five autoregressive terms, one differencing operation to ensure stationarity, and four moving average terms. However, as seen in the figure, the AutoARIMA model was unable to predict the nuanced patterns of the dataset, instead forecasting a constant middle value. This indicates that the AutoARIMA model may not be the ideal prediction model for this dataset, especially when compared to the Holt-Winters' model. The AutoARIMA model is more suitable for datasets with less seasonality and a linear relationship with past values. In contrast, the CO2 emissions dataset exhibits strong seasonal patterns without a clear linear trend, making the Holt-Winters' model a better fit for capturing its seasonality and complex patterns.

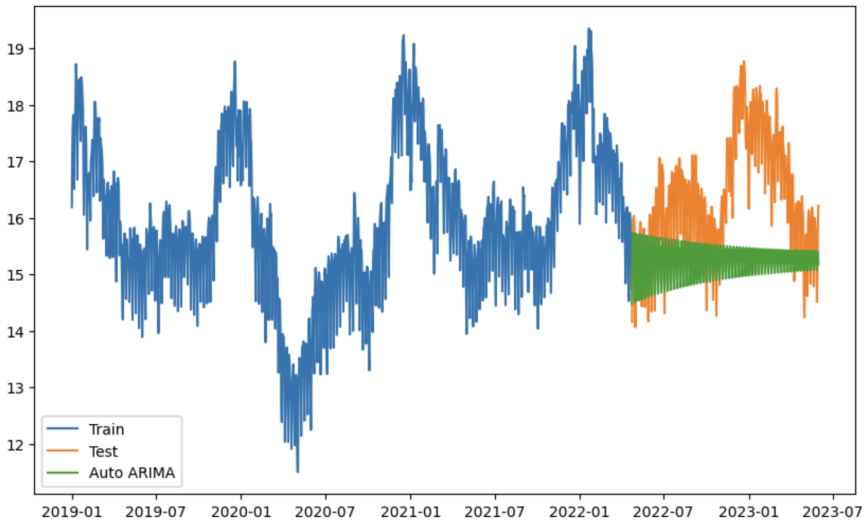

Figure 3.6.1 Auto ARIMA Model Plot

G. Long Short-Term Memory (LSTM)

Before applying the LSTM model, it is important to perform feature engineering to capture the underlying patterns in the data. Feature engineering involves transforming raw data into meaningful features that can enhance model performance. This process typically includes creating a forecasting framework with lagged data and time-related features. From this framework, the most relevant features are selected based on their potential to improve the model's predictions.

After the feature engineering and selection process, the data is prepared for training the LSTM model. LSTM is a type of Recurrent Neural Network (RNN), specializing in capturing both short-term and long-term dependencies in the data. In this application, normalized features were used to ensure that the model was not biased by the scale of the input data. In doing so, min-max scaling brought all features into the same range. Following this, the LSTM model was optimized using the Adam optimizer, which adapts the learning rate during training, and mean squared error was used as the loss function to measure the model's accuracy. Once the model was trained, the test data was fed into the LSTM, and the predictions were made. The predictions and actual values are demonstrated in Figure 3.7.1. The plot demonstrates that the model largely outperformed the other models. The model's MAPE of 0.0187 aligns with this observation, yielding the lowest MAPE out of all the models.

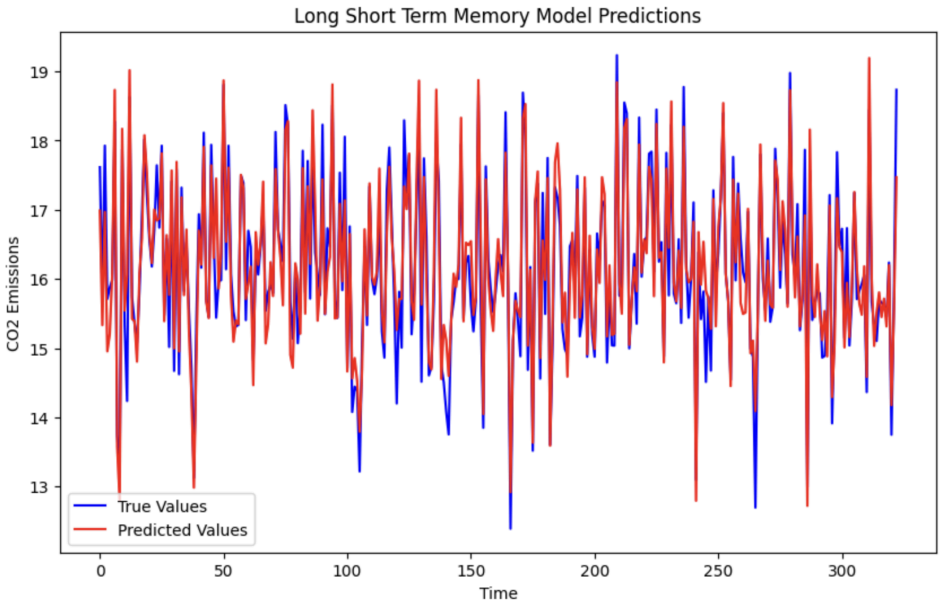

Figure 3.7.1 Long Short Term Memory Model Predictions Plot

IV. Interpretation

With both visual and quantitative results from each of the models applied, there is sufficient data to compare their effectiveness in predicting CO2 emission trends. The least accurate models were the naive forecasting and simple exponential smoothing models. Both models produced predictions that were overly steady and consistent, which was a poor fit for the volatile and seasonal nature of the actual CO2 emission data. These models failed to capture the fluctuations, resulting in a significant misalignment between the predicted and actual values.

Next, the ARIMA and AutoARIMA models offered slight improvements over the aforementioned models but still struggled to capture the true behavior of the test data. These models are best suited for stationary data, which presents a relatively stable mean and variance over time. However, the CO2 emission data is far from stationary, exhibiting significant variation and seasonal spikes that ARIMA could not adequately model. This mismatch highlights the limitations of ARIMA in the presence of non-stationary and highly fluctuating data.

Similarly, the autoregression model performed somewhat better than ARIMA but still faced challenges in capturing the random and unpredictable spikes in CO2 emissions. Autoregressive models work well when the underlying time series follows linear patterns, but in the case of CO2 emissions, the data exhibits non-linear behaviors. The autoregression model, like ARIMA, struggles with non-stationary data and is less effective at dealing with irregular, abrupt changes.

On the other hand, several models demonstrated much better performance in predicting the CO_2 emission data. One of the better models was the Holt-Winters' model, which incorporates both trend and seasonality components. The additive and multiplicative seasonalities effectively captured the seasonal patterns present in the data. It is worth noting that the additive seasonal model performed slightly better than the multiplicative model. This is likely due to the nature of the CO_2 emission data, where the seasonal fluctuations appear to remain roughly constant over time, making the additive model more appropriate for this dataset.

The best-performing model was the Long Short Term Memory network. Not only did the LSTM produce the lowest mean absolute percentage error, but it also captured both the short-term and long-term dependencies in the data. This is particularly important for time series data like CO_2 emissions, where short-term volatility and long-term trends both play crucial roles in predicting future values. By leveraging its memory capabilities, the LSTM effectively accounted for both recent fluctuations and deeper historical patterns, making it the most accurate model for this specific dataset.

To directly compare the effectiveness of the Holt-Winters' and LSTM models in predicting CO_2 emissions, we can examine their coefficients of determination, commonly referred to as R-squared values. This statistic measures how well the model's predictions replicate the actual data, essentially indicating the proportion of the variance in the dependent variable that is predictable from the independent variables. The Holt-Winters' additive model has an R-squared value of 0.4212, meaning it explains 42.12% of the variability in the CO_2 emissions data. Similarly, the Holt-Winters' multiplicative model has an R-squared value of 0.3796, explaining 37.96% of the data variability. These values suggest that while the Holt-Winters' models capture some of the seasonal patterns in the data, a significant portion of the variability—over half—remains unexplained. This indicates limitations in their ability to fully model the complexity of CO_2 emissions.

In contrast, the LSTM model achieves an R-squared value of 0.8984, meaning it explains 89.84% of the data variability. This is more than double the Holt-Winters' models' R-squared values, highlighting the LSTM model's exceptional ability to capture the underlying patterns and dependencies in the dataset. The high R-squared value of the LSTM model suggests that it effectively models CO_2 emissions without overfitting the data. With this, the LSTM model emerges as the most suitable for predicting CO_2 emissions in this dataset.

V. Conclusion & Next Steps

Given these observations, we can conclude that predicting CO_2 emissions requires careful consideration of the appropriate model. The CO_2 emissions recorded in this dataset exhibit

clear seasonal trends on a yearly basis, which means that any predictive model must account for this seasonality to be considered effective. The Long Short Term Memory model has demonstrated its ability to leverage short-term and long-term data, effectively capturing the seasonal patterns and nuances within the data to generate accurate predictions that explain most of the data's variability.

The success of this model in predicting CO2 emissions has significant implications. Accurate predictions can empower environmental scientists, policymakers, and organizations to make informed decisions regarding climate change mitigation strategies. For example, knowing when CO2 emissions are likely to peak during the year could help optimize emissions reduction policies and create targeted environmental interventions. By anticipating periods of high emissions, proactive measures can be taken to reduce their environmental impact.

Moving forward, more comprehensive models can be developed using this dataset, as it includes additional variables such as country and sector. A promising next step would be to build multivariate time series models. These models would not only predict CO2 emissions over time but also incorporate the influence of different sectors and regions. Additionally, since climate-related predictions are inherently uncertain, incorporating uncertainty quantification techniques, such as Bayesian modeling or bootstrapping, into the predictive process could help understand the range of possible outcomes and improve decision-making. These techniques are more advanced and are therefore a step toward more robust and informed policy decisions.