**Cross Validation Image**  
  
  
  
The implementation of a manual cross-validation technique tailored for time series data analysis of Bitcoin prices, using a rolling window approach rather than random splitting due to the sequential nature of the dataset. Specifically, the script establishes a series of temporal folds, each encompassing 363 days, against a substantial training set size of 3000 days, ensuring that a complete integer number of folds is utilized without exceeding the available data.

The image shows five distinct cross-validation folds, each divided into training and test segments—the training segment is illustrated with a blue line, while the test segment is depicted in red. This clear segmentation is indicative of the model's training phase on historical data (blue) before being subjected to the prediction phase (red), providing a pragmatic framework to evaluate the model's predictive accuracy on data it has not previously encountered.

A close examination of the visualized data across the five folds reveals notable price volatility within the Bitcoin market, characterized by drastic fluctuations and several prominent peaks, each of which is captured within different folds. This variability is crucial for assessing the model's robustness, as it ensures that the model's predictive capabilities are tested against a diverse array of market conditions and behaviors, ranging from steady inclines to abrupt declines. The ability of the model to adapt to these conditions can significantly inform its reliability and effectiveness in real-world applications.

**Random Forest image**  
  
The evaluation of a Random Forest model applied to Bitcoin price prediction is presented, illustrating the model's varying degree of success across different temporal segments. The graph in the study delineates a comparison of actual and predicted Bitcoin prices, with the model achieving an average Mean Absolute Percentage Error (MAPE) of 26.57, indicating a moderate prediction accuracy, and an average Mean Squared Prediction Error (MSPE) of 12.88, reflecting the variance of the predictions from the actual values. Notably, the model's Akaike Information Criterion (AIC) averages at 34613.21, serving as a gauge for the model's relative quality. The predictive trajectory closely mirrors the actual price movements, though it struggles with the market's pronounced peaks and troughs, underscoring the inherent volatility and unpredictability of cryptocurrency markets.

A detailed cross-validation, represented in the accompanying table, reveals that the model's performance fluctuates across folds; Fold 3 exhibits the highest MAPE, indicating lower predictive accuracy, whereas Fold 5 boasts the lowest MAPE and AIC, suggesting a superior model fit and predictive precision for that segment. These variations underscore the challenges in forecasting financial time series data, where external factors often induce significant predictive discrepancies.

**XG Boost Image**

the XG Boost model applied to Bitcoin price predictions indicate that the model has a close following of the actual price movements, denoted by the red line, with its predictions. The model records an average Mean Absolute Percentage Error (MAPE) of 25.99, reflecting a moderate level of accuracy in its predictions. The Mean Squared Prediction Error (MSPE) stands at 12.68, which points to the model's ability to predict with a reasonable degree of precision, and the average Akaike Information Criterion (AIC) is 34660.88, suggesting the model's goodness of fit to the data.

The table provides a breakdown of performance metrics across five different folds, with Fold 3 showing a notably higher MAPE, indicating that the predictions for this particular fold were less precise relative to the actual values. In contrast, Fold 5 demonstrates the lowest MAPE and a relatively lower AIC, indicating a more accurate set of predictions and a better fit model for the data within this fold.

The graph reflects the model's challenge in capturing the extreme peaks and troughs of the Bitcoin price, a common difficulty in the volatile cryptocurrency market. These findings emphasize the complexities of financial time series forecasting and the need for robust modeling techniques to handle the unpredictable nature of such data.

**Lagged Information**

In the construction of the time series predictive model for Bitcoin pricing, the introduction of lagged features constitutes a critical enhancement to the analytical framework. Specifically, a lagged feature was generated by shifting the price data by one day, thus creating a new column Lag\_1d in the dataset. This methodological step is imperative for encapsulating the temporal autocorrelation characteristic of financial time series, where previous values bear predictive power over future values.

Subsequent to the introduction of Lag\_1d, the dataset was purged of rows containing NA values, which resulted from the lagging process. These NA entries, indicative of initial time points lacking historical data, were removed to preserve the data's integrity. This cleansing step ensures that the model's inputs are devoid of potential biases or errors that could compromise the validity of subsequent predictive insights.

**Random Forest lagged Features image**

In the assessment of two distinct Random Forest models applied to the task of predicting Bitcoin prices, the incorporation of lagged features has demonstrably enhanced the model's forecasting precision. The initial model, devoid of lagged features, presented a Mean Absolute Percentage Error (MAPE) of 26.57, a Mean Squared Prediction Error (MSPE) of 12.88, and an Akaike Information Criterion (AIC) of 34613.21. These values serve as a baseline against which the performance of subsequent modeling refinements can be measured.

The introduction of lagged features, signifying the inclusion of the previous day's price data in the model, resulted in a notable reduction in all three key performance metrics. The adjusted model exhibited a MAPE of 22.25, a reduction that indicates an enhanced accuracy in the percentage error rate of the predictions. The MSPE decreased to 10.86, suggesting a tighter clustering of the model's price forecasts around the actual observed values. Most telling was the reduction in AIC to 27784.42, implying a substantially improved fit of the model to the historical price data.

This comparative analysis substantiates the premise that temporal dynamics, captured through lagged features, are essential for the accurate modeling of financial time series. By leveraging the previous day's price data, the model gains critical insights into the immediate trends and fluctuations within the market, thereby refining its predictive capability. The discernible improvements across all metrics reinforce the methodological value of incorporating lagged features into the model, validating their role in achieving a more profound and accurate forecast of Bitcoin prices.

**XG Boost with Lagged Features**

Building on the earlier findings with the Random Forest model, the application of XGBoost to Bitcoin price prediction further underscores the significance of lagged features in time series forecasting. The comparative analysis reveals a stark contrast in the performance of the XGBoost model before and after the inclusion of lagged features. Initially, the XGBoost model—absent of these temporal indicators—reported a Mean Absolute Percentage Error (MAPE) of 25.99, a Mean Squared Prediction Error (MSPE) of 12.68, and an Akaike Information Criterion (AIC) of 34660.88, setting the stage for subsequent enhancement.

The integration of lagged features into the XGBoost model marks a pivotal improvement. The MAPE sees a dramatic reduction to 13.61, indicating a notable decrease in the percentage error of predictions. The MSPE follows suit, plummeting to 6.73, which signifies a greatly improved precision in forecasting the variance of Bitcoin prices. Additionally, the model's AIC drops to 27334.89, reflecting a refined model fit that better captures the complexities of the data.

This progression from the Random Forest to the XGBoost model, with the strategic incorporation of lagged features, demonstrates a consistent theme: temporal data points are invaluable for enhancing the accuracy and reliability of predictive models in financial time series. By effectively leveraging the information embedded in the preceding time steps, both models achieve a deeper level of analytical rigor, yielding forecasts that are not only more aligned with the actual market movements but also provide a stronger basis for decision-making in the volatile cryptocurrency domain.

**Model comparison:**The comparative analysis of Random Forest and XGBoost models for Bitcoin price prediction reveals a distinct enhancement in performance through the inclusion of lagged features. Initially, the Random Forest model presented a MAPE of 26.57 and an AIC of 34613.21. With lagged features, these metrics improved to a MAPE of 22.25 and an AIC of 27784.42. The XGBoost model showed even more substantial improvements post-lagged feature inclusion, with the MAPE dramatically reduced to 13.61 and the AIC to 27334.89. This comparison unequivocally demonstrates the value of incorporating temporal data, with the XGBoost model achieving superior predictive accuracy and model fit over the Random Forest model. The integration of lagged features emerges as a pivotal factor in enhancing the models' ability to navigate the complexities of financial time series data.

**Question 2**

The Random Forest and XGBoost models were applied to estimate the carbon emissions from Bitcoin mining, using a robust cross-validation method. The performance metrics across five folds highlighted significant variability, with the Random Forest model exhibiting a MAPE range from 2.58 to 25.11 and MSPE from 0.16 to 7.38. The AIC scores varied from 49.405 to 1658.34, suggesting sensitivity to the data segment being trained on. Notably, Fold 1 of the Random Forest model achieved the lowest MAPE and MSPE, indicating the most accurate predictions for that data subset. Conversely, Fold 2 showed the highest errors, pointing to the least accurate predictions.

**Model Performance Without Lag Features:**

The initial Random Forest model, which did not incorporate lag features, yielded a MAPE of 10.53, an MSPE of 2.19, and an AIC of 6483.71. This model's predictions, while moderately accurate, were outperformed by the enhanced model that included lag features.

**Inclusion of Lag Features:**

Incorporating lag features into the Random Forest model resulted in significant improvements, with a reduced MAPE of 8, MSPE of 1.61, and a notably lower AIC of 5071.91. This affirmed the importance of past values as significant predictors for future trends in carbon emissions related to Bitcoin mining.

**XGBoost Model Enhancement:**

The XGBoost model also saw an uptick in performance with lag features. This was evidenced by improvements in the forecasting metrics, confirming the predictive value of historical data points in this context.

**Comparative Models:**

When comparing the two models with the inclusion of lag features, the Random Forest model demonstrated superior performance over the XGBoost model, particularly in terms of MAPE and AIC. The improved accuracy and parsimony suggest that the Random Forest model, with its lagged data inputs, is more adept at capturing the structure of the carbon emissions data, leading to more precise forecasts.

In conclusion, for the objective of estimating carbon emissions from Bitcoin mining, the Random Forest model with lagged features emerged as the more effective forecasting tool, leveraging temporal dependencies to provide more accurate and reliable predictions.

**Reference observations :  (initial results just for our reference)**

The Random Forest model and Xgboost model have  been used to predict the estimated carbon emissions (MtCO2e) related to Bitcoin mining.

Random Forest

These tables shows the performance metrics with five different folds used in cross-validation. The metrics vary significantly across folds, with MAPE ranging from 2.58 to 25.11, MSPE from 0.16 to 7.38, and AIC from 49.405 to 1658.34.The variation in MAPE, MSPE, and AIC across the folds indicates fluctuating performance, which might suggest that the model's predictive performance could be sensitive to the particular data segment it is being trained on, which is common in time series forecasting.

The lowest MAPE and MSPE are observed in Fold 1, suggesting that the predictions for that particular fold were closest to the actual values which means the model performed best on this subset of data. Conversely, Fold 2 shows the highest errors, indicating the model’s predictions were least accurate for this segment. The AIC values, which penalize model complexity, are also quite varied, with Fold 5 showing the lowest (most favorable) AIC.

These  images shows a time series plot comparing the actual and predicted values of the estimated carbon emissions.

Without Lag Features: The first plot on the left displays the actual and predicted values of the Random Forest model without using lag features. The average Mean Absolute Percentage Error (MAPE) is 10.53, the average Mean Squared Prediction Error (MSPE) is 2.19, and the average Akaike Information Criterion (AIC) is 6483.71.

With Lag Features: The second plot on the right shows the performance of the Random Forest model that includes lag features. The inclusion of lag features has improved the model's accuracy, as indicated by the lower average MAPE of 8, MSPE of 1.61, and AIC of 5071.91.

The introduction of lag features has led to a noticeable improvement in the model's performance. This is evident from the reduction in all three metrics (MAPE, MSPE, AIC), suggesting that past values of the series (lags) are significant predictors for future values in the context of carbon emissions from Bitcoin mining.

The substantial reduction in AIC when using lag features suggests that the more complex model provides a better fit to the data, even after accounting for the increase in the number of parameters.

The comparison of these models indicates that for time series forecasting, especially with data that likely has autocorrelation like carbon emissions data, incorporating lag features can capture temporal dependencies better than a model without them.

Observations:

Model Performance with Lag Features: Both Random Forest and XGBoost models improved in predictive performance when lag features were included. This suggests that for time series forecasting in this context, historical data points are valuable predictors.

Comparison of MAPE: The Mean Absolute Percentage Error (MAPE) decreased for both models when lag features were included, indicating a higher accuracy of predictions. The Random Forest model with lag features outperformed the XGBoost with lag features in terms of MAPE.