6.3730 Pset 5

Problem 1

Part A

Compute the % change in wind output across hourly intervals on the train data split. That is, the percentage change in the time series from time t to t+ 1hr. Plot the histogram of this value, and comment on whether you think autoregressive models are appropriate for modeling this data.

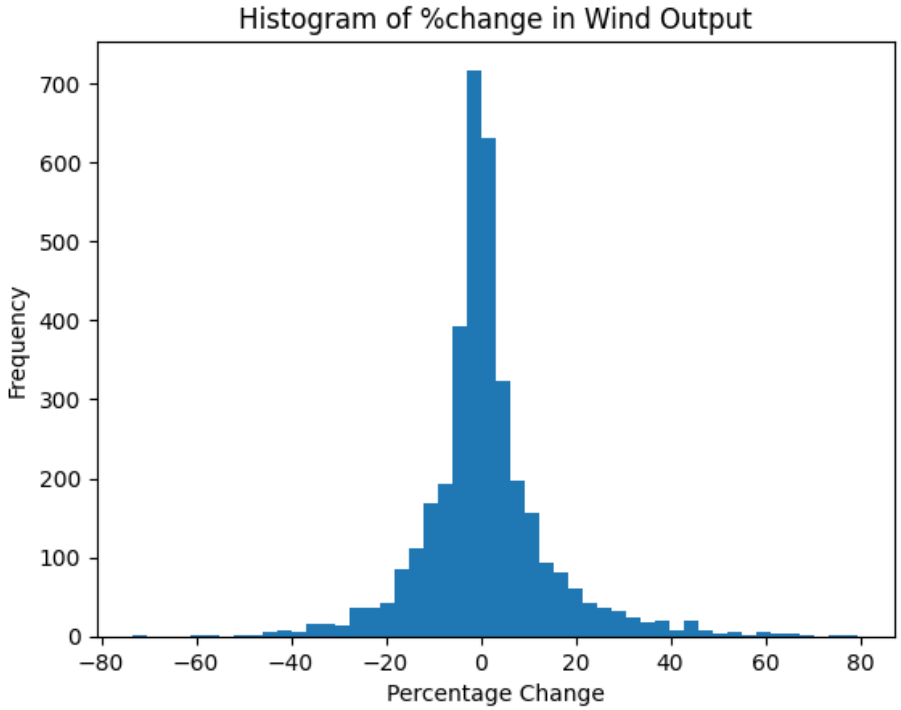


Figure 1: HIstogram of % Change Frequencies

The histogram of the percentage change in wind output is relatively stationary and centered around 0, it suggests that there might not be a strong trend or seasonality in the data. In this case, autoregressive models could be appropriate for modeling and forecasting the percentage change in wind output.

Autoregressive models assume that the current value of a time series is a linear combination of past values. Since the distribution appears to be stationary, an autoregressive model can capture the relationship between past and present values effectively.

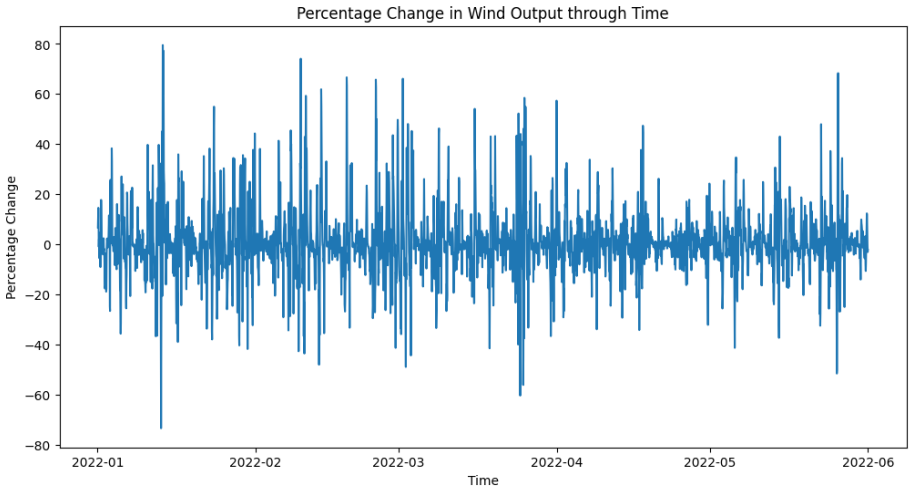


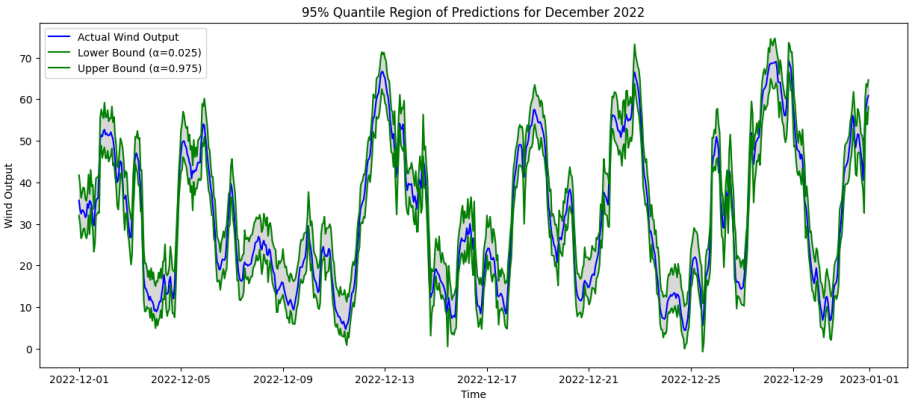
Figure 2: Percent Change in Wind Output through Time

Part B

After fitting the data to a AR model using a quantile regression with an $\alpha = .5$ I got the following MSE and proportions under the actual value

Mean Squared Error: 3.9455757473638973 Proportion of Predictions Below Actual Value: 0.4333907056798623

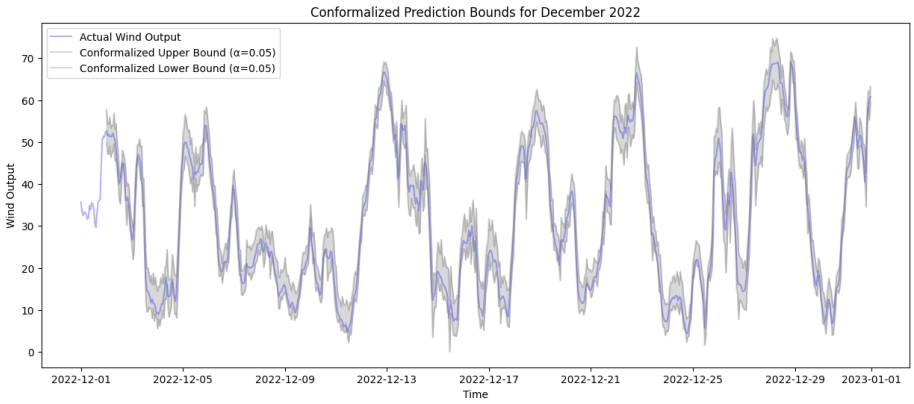
Part C



The resulting plot shows the actual wind output, the lower and upper bounds of the 95% quantile region, and the shaded region representing the uncertainty. The coverage and width of the intervals provide a measure of the model's confidence in its predictions. The intervals can vary in width, indicating that the model's

confidence changes depending on the time. In some cases, the intervals are narrow, suggesting that the model is more confident in its predictions, while in other cases, the intervals are wider, indicating a higher degree of uncertainty.

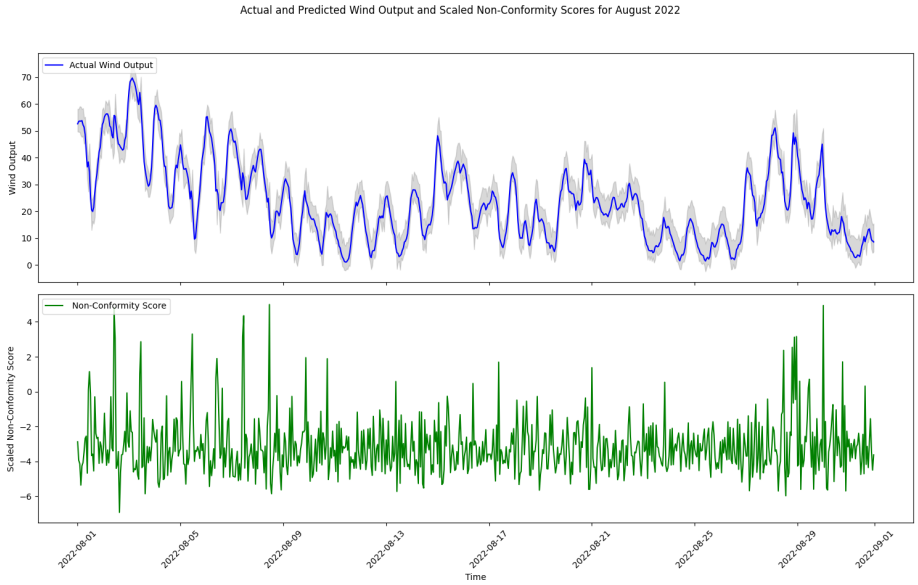
Part D and E

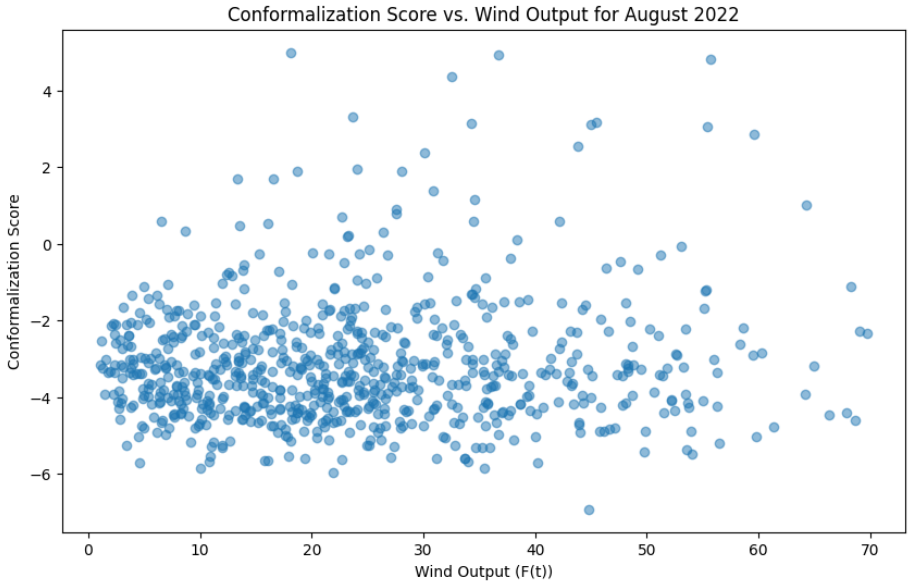


1. Interval Width: The weighted conformal produce slightly different interval widths compared to just the quantile regression. At the extreme changes there seem to be even more certainty when comparing to the nonconformalized graph from earlier

2. Coverage: The weighted conformal technique is expected to achieve the desired coverage level (e.g., 95%) more consistently than quantile regression because it takes into account the dependencies between adjacent data points, whereas quantile regression does not. The conformalized bounds are more adaptive to the local properties of the time series.

Problem 1.2





I was not too sure as to how to draw the the conformalization score as a function of wind output% at time t so I decided to plot it both as a scatter plot as well as separately to show the bounds and the conformalization score on a separate graph to show the changes as t changes.

The conformalization score captures how much the actual value deviates from

the initially predicted quantile bounds. When the actual value lies within the bounds, the non-conformity score will be small, indicating that the actual value is not very surprising given the predictions. When the actual value lies outside the bounds, the non-conformity score will be larger, suggesting that the prediction was not accurate enough to capture the true value.

Thresholding the conformalization scores allows us to control the coverage of prediction intervals. By selecting an appropriate threshold, you can create prediction intervals that contain the true values with a specified level of confidence. The prediction intervals will adapt to the local distribution and changing patterns in the data.

Problem 2

Part A

Here are the formulas in LaTeX:

Bayes' theorem:

$$P(d = T|x) = \frac{P(x|d = T) * P(d = T)}{P(x)}$$

Prior probabilities:

$$P(d = T) = 3 * P(d = S)$$

Marginal probability of x:

$$P(x) = P(x|d = T) * P(d = T) + P(x|d = S) * P(d = S)$$

$$P(x) = PT(x) * P(d = T) + PS(x) * P(d = S)$$

Substitute the prior probabilities and rewrite the Bayes' theorem equation:

$$Q(d = T|x) = \frac{PT(x) * 3 * P(d = S)}{PT(x) * 3 * P(d = S) + PS(x) * P(d = S)}$$

Simplify the equation:

$$Q(d = T|x) = \frac{3 * PT(x)}{3 * PT(x) + PS(x)}$$

Part B

The covariate shift assumes that the conditional probability of y|x remains unchanged between the source and target domains:

$$P_S(y|x) = P_T(y|x)$$

The target risk of h is the expected loss over the target domain:

$$R_T(h) = \mathbb{E}_{x\sim P_T(x), y\sim P_T(y|x)}[Loss(y, h(x))]$$

We can rewrite this expression using the joint probability distribution of the target domain:

$$R_T(h) = \int_x \int_y Loss(y, h(x)) P_T(x, y) dy dx$$

Now, let's use the fact that the conditional probability of y|x is the same for both domains and express the joint probability distribution in the target domain in terms of the source domain:

$$P_T(x, y) = P_T(x) P_S(y|x)$$

Substitute this into the target risk expression:

$$R_T(h) = \int_x \int_y Loss(y, h(x)) P_T(x) P_S(y|x) dy dx$$

$$Q(d = T|x) = \frac{3 * P_T(x)}{3 * P_T(x) + P_S(x)}$$

We can solve for P_T(x):

$$P_T(x) = \frac{Q(d = T|x) * P_S(x)}{3 - Q(d = T|x)}$$

Substitute this into the target risk expression:

$$R_T(h) = \int_x \int_y Loss(y, h(x)) \frac{Q(d = T|x) * P_S(x)}{3 - Q(d = T|x)} P_S(y|x) dy dx$$

Now, we have an expression for the target risk of h as a function of P_S(x), P_S(y|x), h, Loss(y, h(x)), and the friend's domain classifier Q(d = T|x):

$$R_T(h) = \int_x \int_y Loss(y, h(x)) \frac{Q(d = T|x) * P_S(x)}{3 - Q(d = T|x)} P_S(y|x) dy dx$$