ECE 597DE/697DE - Data-Enabled Electricity Networks – Spring 2021
Final Project

Assigned: April 30th, 2021
Due: Wednesday, May 12th, by 5pm (to be submitted via Moodle)
To be completed INDIVIDUALLY (though discussions with classmates are **encouraged**)

Timeseries datasets are very common in the energy and electricity field, and it is common to develop techniques for predicting future values in those timeseries. Improved prediction of future values enables better planning and allocation of resources.

The purpose of this assignment is to compare the performance of three different techniques for timeseries prediction of solar generation. When evaluating a new data-driven technique, it is important to compare its performance to established (and often simpler) techniques. This is called *establishing a baseline*. In this assignment, you will implement and evaluate three different forecasting techniques on a dataset of solar generation for 10 separate installations. Here is a description of the three techniques:

Persistence. A persistence model simply uses the most recent value of a time series as the prediction of the next value. This technique is most effective when values in the dataset change slowly relative to the sampling rate (e.g., solar generation variation each minute) or when variation or noise in the dataset is difficult to model (e.g., local variations in wind generation).

Linear Regression. A linear regression model uses the trend over a window of recent data to predict the next datapoint. Using a fixed window (either of time or of number of samples) over recent data, a line of best fit is computed and extended to predict the value of the next datapoint into the future. The ideal length of the window depends on the variations in the dataset. This technique is effective when a dataset has consistent trends.

Autoregressive Integrated Moving Average (ARIMA). ARIMA models use combinations of historical data from the timeseries to try to explain variation and forecast future values. An ARIMA model takes three parameters: p, d, and q, each of which is explained below:

p: the order (or lag) of the autoregressive term, or the number of lags to use as predictors
d: the order of the differencing term (how many differencing terms are needed to make the dataset "stationary")
q: the order of the moving average term, or the number of forecast errors to consider

In practice, you would need to analyze what value of *d* is needed using some statistical tests (for example, the Augmented Dickey-Fuller test). This informs you how many times the dataset needs to be differenced for it to become "stationary", such that it will be appropriate to forecast with an ARIMA model. For this assignment, *<u>you should use the value d = 0 for all ARIMA models</u>*, as the dataset is already stationary.

The dataset you will use to evaluate these techniques is of solar generation for 10 separate installations over 11 months at a one-minute resolution (roughly 5 million datapoints in total). The data file is a CSV that provides the UTC timestamp, Eastern timezone timestamp, and solar production (in kW) for each of the 10 installations.

Real data will often have some anomalies that you will need to figure out how to work around. In this case, I have cleaned the data to make it easier for you. The original dataset included one file for each house for each day, did not include any readings between 9pm and 5am (i.e., assumed that they were zero), and also had a number of negative values for power measurements. I have created a single file with all of the data, filling in the gaps. Daylight savings time and timezones are also perpetually a challenge to work with. The reason I am telling you these things is so that you are prepared to put in substantial effort into data "wrangling" and cleaning when dealing with real-world data in the future!

For this assignment, you will need to implement forecasting models on the solar generation data using each of the three techniques (persistence, linear regression, and ARIMA). You will evaluate the parameters of each model. Then, you will compare the performance of the three models to understand the improvement from using more historical data and different techniques. I recommend using the *pandas* and *statsmodels* libraries with Python. There are links to tutorials for Python and pandas on the course syllabus, as well as a wide range of relevant available materials on the Internet to help you. Additionally, here are two resources on timeseries analysis and ARIMA with Python:

[1] https://www.machinelearningplus.com/time-series/time-series-analysis-python/
[2] https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/

Additionally, you can use the newly-created course Piazza website to interact with your classmates.

Your submission (due via Moodle) should include well-commented code that shows where you computed forecasts and created graphs. Additionally, you should include a write-up that answers the below questions and includes the graphs that you have generated in order to rationalize your answers.

Questions to Answer – Linear Regression:

1) What is the ideal window for the linear regression technique (from the below list)? Please create a graph that compares the average forecasting performance for all ten installations against the window size. Choose from among window sizes (in minutes) of [2,5,10,15,30,60,120].

Questions to Answer – ARIMA:

2) To select parameters for the *p* and *q* parameters of the ARIMA model, you should produce a graph of the autocorrelation function (ACF) and partial autocorrelation function (PACF) for each of the ten solar installations. From these plots, determine the values of *p* and *q* by seeing which values lie outside the significance region (see [2] above for a demonstration). To answer this question, explain how you chose the ideal values of *p* and *q*, and include at least one plot of the ACF and PACF in your write-up (you do not need to include all of the plots you generate).

Questions to Answer – Performance Comparison:

3) In class, we discussed three different metrics for assessing errors in a forecasting problem: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). Please create three graphs (one for each metric) that compare the performance of the three forecasting models (persistence, linear regression, and ARIMA) averaged over the 10 installations. You should use the ideal parameters for the models that you found in the earlier problems. Considering the three metrics (RMSE, MAE, and MAPE) - which is/are best for comparing the performance of these techniques? Please provide reasons for your selection.

4) What differences did you find among the performance of the algorithms? What do you think causes those differences?

5) How does each of these three techniques perform as the resolution of the input data changes? Please create a graph that shows the performance of each technique averaged over the ten installations with data resolutions (in minutes) of [1,5,15,30,60]. Note that in order to create a lower resolution dataset, you should sample from the dataset, rather than averaging together values. For linear regression, choose a window size that matches the duration of history you found in question 1 (note that this will change the number of samples your window will include as the input data resolution changes).