# Inaccess - Assignment

Case study for the Machine Learning Engineer position

George Iosifidis

Libraries used for this project:

1. pandas
2. numpy
3. sklearn
4. statsmodels
5. seaborn

My first thought was to perform a time series analysis since there is a “timestamp” column in the dataset, so I can understand the way “ActivePowerToUtility” changes over time. From the autocorrelation plot it is obvious that seasonality exists with a period of roughly 90 timesteps. After removing the seasonal component of the series with Lagged Differencing transformation method, I examined the trend of the data, where it appeared to be a polynomial trend. I used seasonal_decompose from statsmodel library to verify this.

To continue with, I used Pearson’s r correlation coefficient to detect any linear relationship between our features and “ActivePowerToUtility”. The results implied strong positive linear correlation between “ActivePowerToUtility” and “AverageIrradianceCleansed”. Also both “AverageModuleTemperatureCleansed” and “Solar Elevation” had positive values (>0.5).

Therefore, I used a Linear Regression model which was trained on “AverageIrradianceCleansed” and “ActivePowerToUtility” input-target pairs. A train/test split of test_size = 0.25 and standardization (with StandarScaler of sklearn) was performed on data. This model achieved a score of 0.958 ($R^2 = 95.8\%$).

In order to examine if other features could contribute to better accuracy, I also trained a Multiple Regression model with “AverageIrradianceCleansed”, “AverageModuleTemperatureCleansed” and “Solar Elevation” as the predictors.

With a train/test split of test_size = 0.5 and standardization, the model had a score of 0.96 ($R^2 = 96\%$).

Multiple Regression did not perform particularly better than Linear Regression, so I would choose the latter because of its simplicity and high accuracy.