**The steps of handling the dataset:**

1. Receiving the dataset from Vattenfall.
2. Convert it to CSV file
3. The received dataset was slightly modified to facilitate working with the dataset. Firstly, the column names are changed- see Table 1.

|  |  |
| --- | --- |
| Original column name | Modified column name |
| Donut\_2\_WSSolarRad | SolarRad |
| Donut\_2\_WSWindDirection | WindDir |
| Donut\_2\_WSAmbTemp | AmbTemp |
| Donut\_2\_WSWindSpeed | WindSpeed |
| Donut2\_NormalDLR | NormalDLR |
| Donut2\_Tension | Tension |
| Donut2\_Sag | Sag |
| Donut2\_Inclinometer | Inclinometer |
| Donut2\_Low\_Point | LowPoint |
| Donut2\_Conductor\_1\_Temperature | CondTemp |
| Donut2\_Current | Current |
| Donut2\_CTM\_DLR | CTMDLR |
| ZL8WSHumidity | Hum |
| ZL8WSBarometricPressure | BarPress |

Table 1. Modifying the column names

CTMDLR, Tension, Sag, Inclinometer and LowPoint observations were not used for building the predictive models.

Hum, and BarPress were used to calculate the DLR values for the summer period of 2014.

**Steps of the analysis**

The R-codes are named accordingly to the order of running them.

**Processing**

The dataset was processed and there were applied various logic conditions on various variables.

**Exploratory Data Analysis (EDA)**

The EDA consisted in building the boxplot, scatterplot and correlograms of the dataset.

**Building the predictive models**

Steps in building the models on the training set and running them on the test sets:

1. Build the models on the train dataset.
2. Save the models
3. Load the models in the script where the testing is performed. Test the models on the test set.
4. Choose the best performant models
5. Use the best performant models for building the predictive models for the next variable in the training set.

For example, building the models for Current:

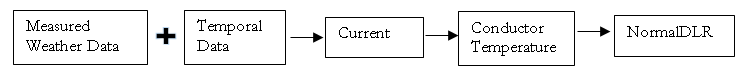
1. Input parameters:
   1. Temporal data
   2. Wind speed and ambient temperature
   3. Current parameter’s observations
2. Build the models in “7. PredCurrent.R”
3. Run the models in the test dataset “AprilTest.R”
4. Choose the best performant model according to the error rate profiles and R2 values achieved in the test dataset. “AprilTest.R”
5. Use the best performant model for building the predictive models for the next parameter (Conductor Temperature)

The building and running time of the models are measured and stored **manually** in the error metric matrix.

**Predictive models for historical observations**

The steps taken in this analysis can be seen in Figure 1. After the discussion with Vattenfall, it was concluded that Current is to some extent dependent on the weather and temporal parameters. Therefore, the Current models are built using the historical observations of the Current, weather and temporal parameters. Then, the Conductor Temperature models are built on the already predicted Current observations and on the measured weather parameters as well as on temporal variables. Lastly, the NormalDLR predictive models are built using the already predicted Current and Conductor Temperature observations and also on weather and temporal variables.

The training dataset represents the time period: Sep., 15th 2014 to Mar., 30th 2015. The test dataset is the month of April, 2015. The predictive models are trained on the training dataset and tested on the test dataset.

Figure 1. The flowchart of predicting the NormalDLR using the historical observations

**Predictive models for forecasted observations**

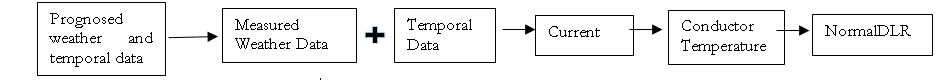
The steps taken in this analysis can be seen in Figure 2. The novelty in this approach, compared to the one used above, is the usage of the prognosed weather observations obtained from Swedish Meteorological and Hydrological Institute (SMHI). Only two input variables (Ambient Temperature and Wind Speed) were used to build the predictive models for. SMHI has communicated at a later stage that the Solar Radiation observations were erroneous. Also, the Wind Direction parameter could not be modeled and therefore, predicted. After discussing this problem with SMHI and Vattenfall, it was concluded that the Wind Direction varies significantly in three dimensions: time and space (vertically and horizontally). The weather stations from where the weather observations were downloaded, were located far from the line sensor.

Figure 2. The flowchart of predicting the NormalDLR using the forecasted observations

Ambient Temperature and Wind Speed prognosed data were available from SMHI only until March, 31st.

The columns in the +24H and +48H time-horizons dataframes were built such as the prognosed DateTime column would be the first one – “PDateTime”, where “P” stands for “prognosed” (also in the case of “PAirTemp” and “PWindSpeed”).

There were built predictive models for 24H and 48H time-horizons. The predictive models were built for Ambient Temperature and Wind Speed parameters. The Ambient Temperature model represents a simple linear regression model. The models used for predicting the Wind Speed involved the temporal parameters. Then, the Current models were built given the temporal variables and the already predicted Wind Speed and Ambient Temperature. The Conductor Temperature models were built using the temporal variables and the already predicted Current, Wind Speed and Ambient Temperature. The NormalDLR models were built using the temporal variables and the already predicted Conductor Temperature, Current, Wind Speed and Ambient Temperature.

For every parameter there were chosen different predictive models- this selection was done after analyzing each model’s performances on the test dataset. Again, the train dataset consisted of observations: Sep., 15th 2014 to Mar., 30th 2015. The test dataset is the month of April, 2015. The predictive models are trained on the training dataset and tested on the test dataset.

The “24H.R” and “48H.R” scripts combine the dataset received from SMHI (“AirTempP.csv” and “WindP.csv”; “P” stands for “prognosed”) with the one processed from Vattenfall (“mean.Hours.csv”), build, train and test the models for each predicted parameter. Later, the “24H+48H.R” combines the “24H.R” and “48H.R” scripts and analyzes the performance of the predictive models for each predicted parameter. My findings (figures) can be found in the appropriate files in the folder “ForecastedObservations”.

The models have found to give slightly different results than initially.