Wind Forecasting track

**Overview**

The topic for the wind forecasting track is focused on mimicking the operation 48-hour ahead prediction of hourly power generation at 7 wind farms, based on historical measurements and additional wind forecast information (48-hour ahead predictions of wind speed and direction at the sites). The data is available from the 1st hour of 2009/7/1 to the 12th hour of 2012/6/28.

The period between 2009/7/1 and 2010/12/31 is a model identification and training period, while the remainder of the dataset, from 2011/1/1 to 2012/6/28, is there for evaluation. The training period will be used for designing and estimating models permitting predicting wind power generation at lead times from 1 to 48 hours ahead, based on past power observations and/or available meteorological wind forecasts for that period.

The evaluation part, it is aimed at mimicking real operational conditions. For that, a number of 48-hour periods with missing power observations were defined. All these power observations are to be predicted. These periods are defined as follows.

The first period with missing observations is from 2011/1/1 at 01:00 until 2011/1/3 at 00:00.

The second period with missing observations is from 2011/1/4 at 13:00 until 2011/1/6 at 12:00.

Note that to be consistent, only the meteorological forecasts for that period that would actually be available in practice are given. These two periods then repeat every 7 days until the end of the dataset. In between periods with missing data, power observations are available for updating the models.

**Data Description**

[**Link to dataset**](https://drive.google.com/file/d/1XjSCeLNWvGNY2L5m78fQ4A4U-PxJN0OK/view?usp=sharing)

"train.csv" contains the training data:

- the first column ("date") is a timestamp giving the date and time of the hourly wind power measurements in the following columns. For instance "2009070812" is for the 8th of July 2009 at 12:00;

- the following 7 columns ("wp1" to "wp7") gather the normalized wind power measurements for the 7 wind farms. They are normalized so as to take values between 0 and 1 in order for the wind farms not to be recognizable.

In parallel, files with explanatory variables (wind forecasts) are also provided for those who may want to use them. For example, the file "windforecasts\_wf1" contains the wind forecasts for the wind farm 1. In these files:

- the first column ("date") is a timestamp giving the date and time at which the forecasts are issued. For instance "2009070812" is for the 8th of July 2009 at 12:00;

- the second column ("hors") is for the lead time of the forecast. For instance if "date" = 2009070812 and "hors" = 1, the forecast is for the 8th of July 2009 at 13:00

- the following 4 columns ("u", "v", "ws" and "wd") are the forecasts themselves, the first two being the zonal and meridional wind components, while the following two are corresponding wind speed and direction.

Finally, the file "benchmark.csv" provide example forecast results from the persistence forecast method ("what you see is what you get"). This file also gives a template for the submission of results that should be strictly followed. Also, you need to include the notebook/python files for evaluation purposes as there will be bonus marks for data preprocessing and EDA.

Please notice that the first column of "benchmark.csv" is called "id" and contains a unique identifier for each row. The other 8 columns are the same as "train.csv". **When submitting the results, please make sure that your file includes total 9 columns in the same format as "benchmark.csv".**

**Evaluation metrics**

The forecasting accuracy will be evaluated with a Root Mean Square Error (RMSE) criterion. The RMSE is calculated as the square root of the average of square forecast errors. All forecasts will be equally weighted in the score calculation.

You need to score an RMSE <= 0.9 on the test data to pass this assignment.

You are required to submit the submission.csv along with your code implementation. Note that there are bonus points for data preprocessing and EDA.

The deadline for this project is May 8 2023.

**Reference:**

You can refer to [this](https://www.kaggle.com/code/mohitkhurana21/mk-wind-forecast) sample notebook for a first-hand approach