Gergo Gyori
gegy@itu.dk

Assignment 1 for Large Scale Data Analysis

1. Data

There were two datasets provided:

- Orkney's renewable energy power generation, source: Scottish and Southern Electricity Networks(SSEN
- Weather forecasts for Orkney, source: UK MetOffice.

The renewable energy dataset contains data regarding the output energy registered in the past, the interval is 1 minute. The column 'Total' represents the energy in MegaWatts.

	Total
time	
2022-05-15 16:14:00+00:00	20.522000
2022-05-15 16:15:00+00:00	20.522000
2022-05-15 16:16:00+00:00	20.380000
2022-05-15 16:17:00+00:00	21.312998
2022-05-15 16:18:00+00:00	21.562002

Fig 1.: Tail of the dataset

The Weather forecasts dataset contains entries regarding the wind direction and speed (the other columns are deleted for the assignment), having as index the time these measurements were taken.

	Direction	Lead_hours	Source_time	Speed
time				
2022-05-15 03:00:00+00:00	SSW	1	1652572800	4.02336
2022-05-15 06:00:00+00:00	SW	1	1652583600	3.12928
2022-05-15 09:00:00+00:00	E	1	1652594400	1.78816
2022-05-15 12:00:00+00:00	SE	1	1652605200	7.15264
2022-05-15 15:00:00+00:00	ESE	1	1652616000	7.15264

Fig 2.: Tail of the dataset

2. Data preprocessing

Both tables have as an index the timestamp the measurements were recorded. While for the energy generation dataset the timestamp was 1 minute apart, the wind dataset was recorded at an interval of 3 hours. Thus, the energy data was resamped to 3 hours - mean values were used - to match with the wind sample. As for the column 'Total', the average of the energy generation was aggregated for the 3 hours interval.

The wind direction strings were transformed to degrees ("Wind Direction and Degrees", 2022) with a simple dictionary, where North as 0, then gradually increasing with 22.5 degrees for all of the 16 directions, then the two dataset were merged

time	Total	Direction	Lead_hours	Source_time	Speed	Angle
2022-05-15 03:00:00+00:00	6.105578	SSW	1	1652572800	4.02336	202.5
2022-05-15 06:00:00+00:00	5.687467	SW	1	1652583600	3.12928	225.0
2022-05-15 09:00:00+00:00	7.284089	E	1	1652594400	1.78816	90.0
2022-05-15 12:00:00+00:00	12.961211	SE	1	1652605200	7.15264	135.0
2022-05-15 15:00:00+00:00	19.858924	ESE	1	1652616000	7.15264	112.5

Fig 3.: Tail of the merged dataset

3. Constructing the pipeline

Three algorithms were used for prediction, namely Linear Regression, KNeighbors Regression and an LTSM model. Since the overall accuracy of the Linear Regression and the LTSM model were lower than the KNeighbors's these two were skipped to implement into the pipeline.
The KNeighbors gave an R2 score of 73% with 365 days of data. PolynomialFeatures was also considered to be used, but since 1 year data point generated 73% of accuracy without any preprocessing, this step is skipped from the pipeline.

```
Query is running
Dataframe cleaning is running
Wind direction tranformer is running
Grid Search is running
Best parameter (CV score=0.704):
{'KNN__algorithm': 'brute', 'KNN__n_neighbors': 16}
Retrain, predict and store the results and the model
Load latest forecasts, make a prediction of Generated Power
Wind direction tranformer is running
###### Predictions
```

time	Direction	Lead_hours	Source_time	Speed	Angle	Generated Power Prediction
2022-05-15 18:00:00+00:00	SE	3	1652619600	8.04672	135.0	18.380516
2022-05-15 21:00:00+00:00	ESE	6	1652619600	7.15264	112.5	18.303089
2022-05-16 00:00:00+00:00	ESE	9	1652619600	7.15264	112.5	18.303089
2022-05-16 03:00:00+00:00	ESE	12	1652619600	8.04672	112.5	20.246995
2022-05-16 06:00:00+00:00	E	15	1652619600	8.94080	90.0	20.369399

```
R2 score of the loaded model is 0.73
```

Fig 4.: Output of the model

During the training the different models and different predictions were saved for further analysis.

time	Direction	Lead_hours	Source_time	Speed	Angle	Generated Power Prediction	prediction_time
2022-05-15 18:00:00+00:00	SE	3	1652619600	8.04672	135	18.38051616	2022-5-15-18-18-38
2022-05-15 21:00:00+00:00	ESE	6	1652619600	7.15264	112.5	18.30308897	2022-5-15-18-18-38
2022-05-16 00:00:00+00:00	ESE	9	1652619600	7.15264	112.5	18.30308897	2022-5-15-18-18-38
2022-05-16 03:00:00+00:00	ESE	12	1652619600	8.04672	112.5	20.2469947	2022-5-15-18-18-38
2022-05-16 06:00:00+00:00	E	15	1652619600	8.9408	90	20.36939935	2022-5-15-18-18-38
2022-05-16 09:00:00+00:00	E	18	1652619600	11.176	90	23.7794033	2022-5-15-18-18-38
2022-05-16 12:00:00+00:00	E	21	1652619600	11.176	90	23.7794033	2022-5-15-18-18-38
2022-05-16 15:00:00+00:00	E	24	1652619600	11.176	90	23.7794033	2022-5-15-18-18-38

time	Direction	Lead_hours	Source_time	Speed	Angle	Generated Power Prediction	prediction_time
2022-05-15 18:00:00+00:00	SE	4	1652616000	8.04672	135	18.38051616	2022-5-15-17-59-31
2022-05-15 21:00:00+00:00	ESE	7	1652616000	7.15264	112.5	18.29223434	2022-5-15-17-59-31
2022-05-16 00:00:00+00:00	SE	10	1652616000	7.15264	135	13.75897003	2022-5-15-17-59-31
2022-05-16 03:00:00+00:00	ESE	13	1652616000	8.04672	112.5	20.23614007	2022-5-15-17-59-31
2022-05-16 06:00:00+00:00	ESE	16	1652616000	8.9408	112.5	24.35777291	2022-5-15-17-59-31
2022-05-16 09:00:00+00:00	E	19	1652616000	11.176	90	23.7794033	2022-5-15-17-59-31

Fig 5.: Snippet of different predictions in different times

4. Discussion

4.1 Pipeline, KNeighbors Regression

The KNeighbors Regression give a decent result in terms of accuracy. Designing a more complex model with more preprocessing would cause a more complex pipeline. The GridSearch was seeking through a 1-100 range for n_neighoburs and the mode is set up for 4 different method (‘auto’, ‘ball_tree’, ‘kd_tree’, ‘brute’) however the n_neighbours had a peak in terms of accuracy between 10-50 neighbours
Dealing with null and minus values was easy, deleting null values (which are caused when wind turbines need to be slowed down in case of heavy wind) was not significantly change the accuracy results.

4.2 Data

Bigger data set (means more dates) generated better accuracy for the model without any special preprocessing but using years for the training has no significant improvement. Another relevant point for the future work is to convert and handle the compass measurement in different ways, meaning for instance apply the difference between two measured points to get a more accurate values, instead of using only 16 values.

	Energy	Source_time	Speed	Angle
Energy	1.000000	0.309119	0.788888	0.129034
Source_time	0.309119	1.000000	0.276272	0.067394
Speed	0.788888	0.276272	1.000000	0.149473
Angle	0.129034	0.067394	0.149473	1.000000

Fig 6.: Correlation of 1 year data

5. Conclusion

I believe there is the trade off of the complex models and pipelines: find the limits where we use the minimum resources to maximise accuracy of predictions. Through the process of developing a code for this assignment it can be highlighted that implementing a pipeline has a strong beneficial value to process and predict, also it could generate reusable models and results which can save a lot of resources as well.

References

Wind Direction and Degrees. (2022). Retrieved 15 May 2022, from http://snowfence.umn.edu/Components/winddirectionanddegrees.htm