IT UNIVERSITY OF COPENHAGEN

LSDA: Assignment 1

Author	**Email address**
Maciej Jalocha	macja@itu.dk

1 Introduction to the problem

The goal is to predict the total power generated by Orkney's renewable power generation plant as reported by SSEN using weather forecasts for Orkney from UK MetOffice.

2 Datasets

In both datasets, rows are indexed by time.

SSEN provides power generation at a given time, every minute. From the SSEN data, we pick only the 'Total ' key because we are interested in the generation of electricity regardless of the specific source:

Key	Type	Desc
time	datetime	index
ANM	float	Source
Non-ANM	float	Source
Total	float	ANM+'Non-ANM'

Table 1: Desc. of SSEN data (target variable)

UK MetOffice provides **point-in-time** forecasts. We have access only to the forecasts made exactly one hour earlier, every 3 hours. Therefore, from the data we pick 'Speed' and 'Direction' only - there is no reason to pick Lead_hours which is constant for all rows nor 'Source_time' which would be 'time'-1h/Lead_hours. If the 'Lead_hours' varied, perhaps we could incorporate this into our predictions as a quality measure of a forecast.

Key	Type	Desc.
time	datetime	forecast target time
Speed	float	Wind speed, [m/s]
Direction	string	like "S" or "NW"
Source_time	integer	of a forecast
Lead_hours	string	horizon, hours

Table 2: UK MetOffice data (feature data)

Both datasets have missing data. Upon inspection for up to two years back from 03-03-2025, we saw that both datasets contained **few** and **isolated** missing data points. We consider them negligible and droppable. However, target labels contain two consecutive periods of a few dozens of days of missing data. (April/May 2024 and Feb 2025). We drop them as explained in EDA. We also noticed negative values. We set these to zero.

3 EDA

- Determine if, given the current features, the problem could be solved at all.
- Determine which metric is optimised
- If so, identify which features incl. transformed could be used.
- Decide which models would be promising
- Consider merging datasets and handling missing data*.
- Consider scaling data*.

*These aspects are considered in the context of the proposed models.

3.1 Metric (Goal)

RMSE is used, which is suitable for regression problems.

3.2 Features

3.2.1 Speed

is necessary and sufficient to produce energy in wind turbines. We see that its relationship with the target variable follows a known 'power curve'. A polynomial of degree 2 or 3 or 4 could be used for training. We also see that 'Total' in some way follows 'Speed' It also gives a good idea that this problem could be solved at all. Following figures are using 2 years of data.

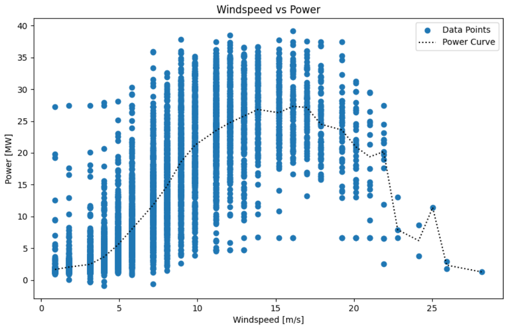

Figure 1: Wind Speed vs Power Relationship

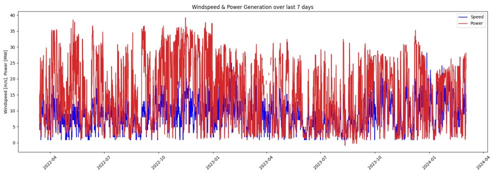

Figure 2: Speed and Power Generation vs Time

From The Power Curve of a Wind Turbine the authors read that ‘power curve‘ on its own is not the best predictor, and past variance of the speed should be incorporated. The model could learn that the variance of ‘Speed‘ is important either by being provided ‘Speed‘ variance or std from the previous n entries (rolling window) or by being directly provided n ’lagged’ ‘Speed‘ entries.

‘Speed‘ n-Lag	Corr.	-=-	-=-
1	0.67	4	0.40
2	0.57	5	0.34
3	0.48	6	0.29

Table 3: Corr. between lag of ‘Speed‘ and ‘Total‘

Unfortunately despite promising results in the end due to time constraints we didn’t manage to incorporate that. We also think that the lags of ‘Total‘ itself could be incorporated.

‘Total‘ n-lag	Corr.	-=-	-=-
1	0.88	4	0.55
2	0.76	5	0.48
3	0.64	6	0.42

Table 4: n-lagged ‘Total‘ corr. to ‘Total‘

3.2.2 Direction

is neither necessary nor sufficient to produce energy. However, changing direction introduces the loss of energy. Therefore, an idea for the feature would be a rolling window of (some measure of) recent wind variance or lagged wind direction. Secondly, due to the way the plant’s wind turbines are positioned, some directions might be favoured.

Wind at the current time We perform a few checks:

- There is no collinearity between ‘Direction‘ and ‘Speed‘. This is a good sign.

Direction	Value	Direction	Value
E	-0.04	S	-0.05
ENE	-0.08	SE	0.01
ESE	0.02	SSE	-0.02
N	-0.02	SSW	-0.07
NE	-0.06	SW	0.01
NNE	-0.06	W	0.09
NNW	0.03	WNW	0.05
NW	0.02	WSW	0.11

Table 5: Corr. with Speed (Rounded)

- No correlation between one-hot encoded ‘Direction‘ and ‘Total‘. This is perhaps not the best sign, though it does not mean that together with Speed it will not help. Correlation between ‘Total‘ and 8 lagged one-hot encoded ‘Direction‘ similarly with no single one trespassing +/- 0.20 (In this case it’s not interesting nor useful: should I write about it at all?).

Direction	Corr.	Direction	Corr.
E	-0.05	S	0.05
ENE	-0.09	SE	0.07
ESE	0.03	SSE	0.05
N	-0.06	SSW	-0.02
NE	-0.11	SW	0.01
NNE	-0.11	W	0.05
NNW	-0.03	WNW	0.01
NW	-0.03	WSW	0.10

Table 6: Corr. with target variable (Rounded)

We also noticed that the ratios of the means of ‘Total‘ and ‘Speed‘ within each direction seem almost the same, suggesting that direction indeed does not have an effect. (Actually a question to reviewers: does it make sense to make this plot?)

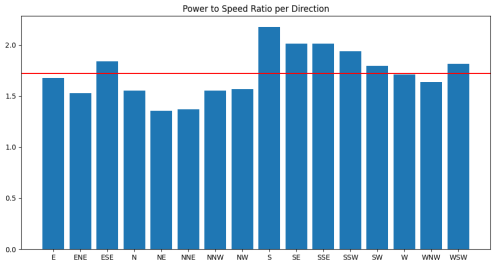

Figure 3: Power-Speed Ratio by Direction

Capturing change of the wind In order to compute variance of the wind change, the wind direction was encoded as a unit vector represented in the data as magnitudes of the OX and OY vectors, i.e. $\sqrt{W_x^2 + W_y^2} = 1$, and then the standard deviation of the previous N (to be determined) rows was calculated for both W_x and W_y and these two values were summed. We loop over the range from N: 0 to 10 to assess which yields the highest correlation with ‘Total‘. N = 3 produced a -.3 correlation. It is worthwhile to note that this negative correlation followed expectations. (Is it a good idea to pick the one with the highest negative correlatin? again: other might be better together).

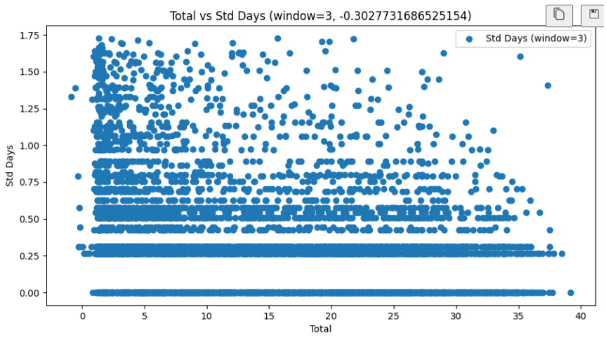

Figure 4: Correlation with ‘Total‘

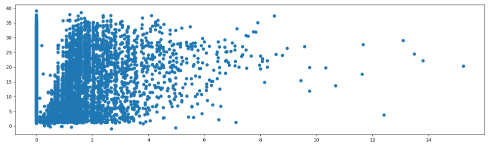

Figure 5: N=3 Rolling Window of variance of the wind * ‘Speed‘ against Total, with correlation = 0.14

We inspected a lagged direction, both one-hot encoded and vector encoded - the latter offered a higher correlation of -0.2 across 8 lags, but we didn't end up testing it. We would like to utilise periodograms for that.

3.3 Models & Scaling

. Since speed is clearly the most important feature and could be modelled as a polynomial, the authors decided to use Linear Regression (pure + with L1 i.e. Lasso + with L2 i.e. Ridge). Though Linear Regression does not require scaling, to simplify the code, the data is scaled for all of them (Ridge and Lasso require it).

3.4 Merging and Missing Data

3.4.1 External loss

Authors decided to drop every row with missing data after performing the inner join on two tables. These were dropped because:

- the chosen models depend only on the ‘Speed‘ in the current moment and the wind variance drawing *only* from the last three previous entries. Authors acknowledge that it means that there are cases where variance is based on irrelevant rows, but since these are so few cases in relation to the whole dataset, authors consider it negligible and don't handle it.
- concerning these two large regions of empty labels - a smart imputation of these would be difficult. The authors attempted imputations with different interpolation algorithms with little success.
- However, at the end ¿ 90% i.e. ¿5000 datapoints remains, so the authors consider it enough data.

3.4.2 Loss caused by inner join

The authors realise that a lot of data from SSEN dataset was actually lost at the moment of the inner join. Clearly, imputation on a 3h interval would be more feasible than imputation of a 20 day interval, but on the other hand, it would mean that 179 out of 180 datapoints (every minute per 3h) did not come from the original data. In the end, the authors did not check it (due to time constraints).

3.5 Best training timeframe

No initial guess was found by the team. We realise that the bigger samples are, the more data for the model there is to learn from, but on the other hand, the older data is, the more expired it is. We answer this question using grid search.

3.6 Model Selection and Validation

To find the best model and validate ideas, a grid search of 27 combinations was performed. The search considers the following:

- **Model Choice** – One of (3) models
- **Polynomial Degree** – Chosen from $\{2, 3, 4\}$
- **Handling of Wind Direction** (3):
  - Not using direction at all
  - Using direction only as an interaction term (multiplying it by all polynomial terms of speed)
  - Computing polynomial powers of direction variance as well

For each combination, we perform time series cross-validation and report:

- The **average RMSE** across top ten splits.
- The **average number of training datapoints** for the top 10 splits.

There are a total of **104 splits**, each corresponding to **one week** in two years of past data. So the tests are done with week forward.

RMSE	Data Points	Handling	Degree	Model
2.95	361D	Drop	4	LinRe
2.97	349D	PolyInt	4	LinRe
3.06	349D	PolyInt	3	LinRe
3.07	363D	Drop	3	LinRe

Table 7: Results of model evaluation

3.6.1 Results

3.6.2 Analysis of the results

All results are on: dagshub Opposite to expectations, Lasso and Ridge did not help. Best Ridge is 5th with 3.08 top-10-RMSE and Lasso further away with 4.03 top-10-RMSE. Secondly, the rolling window of the wind variance did not help. A model which we would consider a baseline, that is, a Linear Regression with the polynomial transformation of ‘Speed‘ is the best. More statistical checks would be required, but the authors recognise that the baseline was not beaten and remains the same. The authors settled to use the 4th model as the final model, as it is the least complex one without much loss in the top-10-RMSE. The reason could be that simply the wind is not changing much or that Orkney has technology, which prevents energy loss and the effect is negligible. The authors did not perform more analysis due to time constraints.

3.7 Deployment

3.7.1 Access

The model can be accessed by:

```
curl -X POST \
51.120.241.52:5000/invocations \
     -H ’Content-Type: application/json’ \
     -d ’{
           "dataframe_split": {
             "columns": ["Speed"],
             "data": [[5], [24], [4]]
           }
         }’
```

Note: On Windows, replace single quotes (’) with double quotes (") and double quotes (") with two double quotes ("") to avoid errors.

3.7.2 Retraining

Every minute a retraining is run on a new data (‘crontab‘). If the new model is better than the latest best one, then it’s registered as the latest best one.

Note: Apparently ITU i.e. provider disabled access to Influx and this fails.

3.8 Code

: Available on: DAGsHub and ITU’s Github Experiments can be seen either via DAGsHub or MLFlow

3.9 Reproducibility

Environment for development and production has been saved and is available in the repo.

3.10 Limitations, authors’ mistakes and ideas

The authors acknowledge that:

- It would be beneficial to arrange a test dataset and provide a score for it.
- In the current setup, the whole dataset was used for EDA, training and validation. The authors are seeing that in such a setup, there is a data leakage between training and validation datasets via EDA (done on all data). (?How to resolve that??)
- Obviously whole time-series engineering could be deployed here i.e. lag or trend features, or hybrid models. **Note:** The authors carried out a vastly unsuccessful attempt with ARIMA model which was too embarrassing to include in the report. The authors learnt that they should have started simple with new models to them, like ARIMA, before deploying whole pipeline of extreme feature engineering only to see that the model was different and did not provide results.
- More rigid analysis could be deployed, i.e. checking if the difference between the scores was statistically significant and shuffling the target variable to see how much the model picks on the noise.
- More metrics could be used, especially to better understand why ‘Direction‘ did not help despite an educated attempt. We also think that the wind change did not help due to a too long span (std of wind direction=3, computes std based on current wind direction, 3h and 6h earlier which does not capture the actual change in the last 5 minutes)
- Above we performed a check that ‘Total‘ is uncorrelated with the ‘Direction‘ (One Hot encoded). Given that it is known that even though not all variables are independently highly correlated, together these might give the best score, is computing individual correlations in the EDA useful at all? **This is important** As you see, I’m doing my initial

preselection based on the fact which variables correlates the most with the target variable - but is there any reason at all to do this if we know, that in the end another combination might be the best? I think no and felt a little pressured to just write something. Should I do any preselection of features in EDA then or leave it to experimentation part?

- Since we had only two initial features, i.e. Speed and Rolling window perhaps plotting an interactive 3D plot would greatly help assess what is going on (or just 'Total' power being colour in a 2D plot.
- Insight for us: it's interesting to see that depending on feature representation correlation might be different (correlation of Wx with target variable is twice as big as the highest correlating one-hot encoded direction.)
- Now I'm thinking I would an infographic explaining what features were actually used.