Linear Regression Analysis – Energy Production in European Countries

Object: To use linear regression algorithms to predict the renewable energy (solar + wind) production of key countries in the EU.

Approach: A dataset comprising energy values expressed in the form of MWh, corresponding to wind and solar generation, is analyzed using time-series analysis, while tested with 3x different linear regression algorithms.

Out of the energy and weather data corresponding to the 20+ countries contained in the Open Power System Data, our analysis was limited to the following countries:

- United Kingdom
- Sweden
- Denmark
- Germany

Our analysis was conducted using the python platform and common libraries such as Numpy, Pandas, Matplotlib and Seaborn.

The most complete dataset corresponded to Germany, whereas the other countries present significantly less weather data points.

Initially, we tried evaluating energy values starting in 2010 until the present date. But the evaluation period needed to be drastically reduced by the lack of energy data corresponding to the other countries.

Preliminary Graphical Results:

The time series plots for each country identify a seasonal variation in energy production, both wind energy and photovoltaic energy.

For Germany, wind energy more than doubled between 2010 and 2019, whereas the solar energy production was first reported in 2012 and has slightly increased over the years.

Similarly, Denmark observed a doubling of its wind energy production over the same period. And, solar energy output is first reported in 2015 increasing slightly over the years to reach ca. 15% of its wind energy counterpart.

Energy datasets corresponding to Britain indicate an increase of 500% in wind energy production as well as an increase of 400% in solar energy for the same period.

On the contrary, wind energy production has slowly increased over the year but to a lesser extend to all other countries. Further, solar energy data is only reported for 2018 and on.

As far as weather data,

Variables such as Temperature, Radiation Direct Horizontal and Radiation Diffuse Horizontal are plotted over time, showing a direct correlation to the energy production levels (peak values) observed on a year-to-year basis.

In addition, Seaborn was used to identify in graphical form the interdependence of the energy production levels with temperature, radiation direct horizontal and radiation diffuse horizontal, by plotting pairplots against combined weather data variables.

Although large data clusters were obtained, there is some an indication of a direct correlation of values, which points towards the feasibility of linear regression analysis.

Due to the large of Not Available datapoints, we limited the size of the datasets and proceeded to apply linear regression algorithms obtaining the following results:

Linear Regression Algorithm:

Country						Average
Germany	-0.92309326	-0.10978097	0.44140219	0.57089763	0.57172048	0.11022921
Sweden	-1.463e+03	-4.5283e+01	-9.9482e-01	5.37557e-02	-1.8826e-01	-301.90415
Denmark	-0.030949	0.00931171	0.01579143	0.02611322	-0.0042439	0.00320466
Britain	-2.25405	-0.786215	0.0027278	0.02359805	-0.1691703	-0.63663376

Ridge Linear Regression Algorithm, w. alpha = 5:

Country						Average
Germany	-0.9230931	-0.1097811	0.441402	0.570897	0.5717207	0.00320466
Sweden	-1.4631e+03	-4.5283e+01	-9.948e-01	5.3755e-02	-1.882e-01	-301.90342
Denmark	-0.030949	0.0093117	0.015791	0.0261132	-0.004243	0.003204
Britain	-2.25404	-0.786214	0.00272816	0.0235978	-0.169170	-0.6366199

Lasso Linear Regression Algorithm, w. alpha = 0.1:

Country						Average
Germany	-0.923091	-0.10978	0.441402	0.570897	0.5717207	0.1102292
Sweden	-1.4631e+03	-4.52835e+01	-9.948e-01	5.3755e-02	-1.8826e-01	-301.903423
Denmark	-0.030949	0.009311	0.015791	0.0261132	-0.004243	0.0032046
Britain	-2.25404	-0.786214	0.002728	0.0235978	-0.1691706	-0.6366193

Conclusions:

A large data set was analyzed using simple graphical tools that helped determine direct correlations between energy production levels and key weather parameters.

Unfortunately, a myriad of missing values limited the extend to which historical data could be analyzed. For this reason, python libraries such as NUMPY and PANDAS were used to wrangle and clean the data sets.

When evaluating the energy production levels, for both types of energy, Eolic and Solar, over one year period, it is hard to distinguish any trends as far as growth. But when plotted over a period of about 5 years, a clear seasonal behavior and overall trend to growth can be observed, which corresponds to the overall penetration rates of renewable energy observed in Europe.

The statistical analysis approach was adequately chosen to be linear regression and approached as time-series analysis.

As indicated, the plotting of energy production levels against weather parameters shows a linear dependency over time, suggesting that the output is a linear function of the input. This observation indicates that linear algorithms can be used to predict the response of test datasets.

The linear regression algorithm was adopted from sklearn.linear_model (least square linear regression), using the cross-validation procedure. In our case, we set the number of folds to be cv=5. In here, the dataset is divided into k-smaller subsets or folds, and the model is trained in k-1 of the defined folds.

The performance of the algorithm is determined by the coefficient of determination, R^2. An average value close to 1 will indicate that the regression makes predictions close to the true values. As covered in previous task, a desired value would for R^2 average would be rather between 0.6 to 0.8.

But our actual results put the average close to 0 or even assigns it negative values, indicating that the linear regression model is weak in predicting future energy production levels.

At this moment, the low average R^2 values are low due to the fact that a long period of time, containing various weather cycles, have a deleterious effect on the linear dependency of the data.

Future work should focus in applying the models to only one cycle of data, perhaps evaluating datasets corresponding only to one-year cycle (including all seasonal fluctuations).