# 3.1 Methodology

## 3.1. Data

**[Description of the data being used in the analysis.]**

Prior to feature selection, an initial assessment of the wind turbine’s power curve was undertaken [**add short description of what the power curve explains**]. In doing so allows the researcher to determine whether (i) there are any outliers within the data and (ii) whether there was any lost production.

Morrison et al (2022)[[1]](#footnote-1)[[2]](#footnote-2) highlight that fairly basic anomaly filtering can occur initially based on several parameters, including whether the SCADA data indicates there were any operational issues with the wind turbine during the period or whether the blade pitch angle feature is greater than a certain degree. Fortunately, the SCADA dataset used in this analysis provides an overall amount of lost production resulting from factors including curtailment and performance. Removing observations where there was lost production results in a cleaner power curve which should improve the predictive accuracy of the model.

The initial power curve chart (top of Figure 1) shows a significant amount of anomalies, i.e. points outside the ‘traditional’ power curve, and observations where there was lost production. Comparatively, the bottom half chart shows fewer anomalies.

**Figure 1: Penmanshiel wind turbine power curve before and after data cleaning**

Chart, scatter chart

Description automatically generatedChart

Description automatically generated with medium confidence

## 3.1. Feature description and selection

Based on the available literature, a number of meteorological variables were deemed relevant to the modelling of wind turbine power production. These included wind speed, wind direction, nacelle position, nacelle ambient temperature (selected as a proxy for outdoor temperature), and blade angle pitch position for each blade. Given the time series nature of the data, time of year information was extracted from the date and time observation; sin and cosine transformations performed on the month, day, week and hour. [**add a bit more detail about what these sin and cosine transformations help to do i.e. they’re relational transformations**].

Additionally, lagged values of the target variable were explored to determine whether there was any partial autocorrelation between the current value and the its previous values. [**add a bit more detail about auto-correlation**]. Figure 2 shows the partial autocorrelation between the lagged time series points and the present value of the target variable. Here we see that the last six lagged values sit outside of the area depicting the 95% confidence interval which indicates the lagged values hold some significance in explaining the present value. New variables for each of these lagged values were created and further explored in the next stage of the exploratory analysis.

**Figure 2: PACF of wind turbine power output**

Chart

Description automatically generated

The aforementioned variables were explored to determine their correlation with and statistical significance in relation to the target variable. Figure 3 shows the upper diagonal of a correlation matrix filled with ellipses that have their eccentricity parametrically scaled to the correlation value. It can be seen that several variables are highly correlated with the target variable: namely wind speed and lagged values of the target variable. Variables that are lightly shaded show low correlation and should be considered as non-important and therefore potentially dropped from the final model. These include some of the transformed hour, day, week and month variables.

**Figure 3: Correlation matrix of potential model features**

Diagram

Description automatically generated

Additional checks on the importance of each of these variables were completed prior to their removal from the model. Initially, a simple linear regression was performed on the data. [**needs a short description of what a linear regression is**]. Variables that were not statistically significant, as determined by their p-value, were considered potential candidates to drop from the final model. These results were tested further through the use of lasso regression. [**needs a short description of what lasso regression is and how it shrinks non-important variable coefficients to zero**]. Feature coefficients which shrank to zero indicate non-important relationships and therefore should be dropped from the final analysis.

**[Visualisation**: Create chart which shows in descending order the relative importance of each variable and colour code those which are significant at the various significance levels from the linear regression. Find a way to also show in the visualisation which features shrank to zero in the lasso regression something like this – please ignore the male-female stereotypes included in this chart.[[3]](#footnote-3)]

Chart

Description automatically generated

Some early results from random forest (super basic only trained on 50 trees with no hyperparameter tuning) – don’t need to write these up now, just wanted to show you guys.

Chart

Description automatically generated Chart

Description automatically generated

Graphical user interface, chart

Description automatically generated

**LINEAR REGRESSION OUTPUT – NOT TO BE INCLUDED, JUST MAKING A NOTE [RED = POTENTIALLY DROP – NOT SIGNIFICANT]**

Call:

lm(formula = Energy.Export..kWh. ~ ., data = explore1)

Residuals:

Min 1Q Median 3Q Max

-464.26 -13.39 -3.51 9.52 235.85

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.463e+01 9.928e-01 -34.877 < 2e-16 \*\*\*

Wind.speed..m.s. 1.218e+01 1.255e-01 97.060 < 2e-16 \*\*\*

Wind.direction.... -1.276e-04 3.960e-03 -0.032 0.974291

Nacelle.position.... -1.072e-02 3.910e-03 -2.742 0.006103 \*\*

Nacelle.ambient.temperature...C. -2.845e-01 7.730e-02 -3.680 0.000233 \*\*\*

Blade.angle..pitch.position..A.... -1.594e-01 5.632e+00 -0.028 0.977426

Blade.angle..pitch.position..B.... 1.375e-01 3.170e+00 0.043 0.965397

Blade.angle..pitch.position..C.... 1.740e-01 2.484e+00 0.070 0.944145

sinmonth -8.913e+00 2.653e+00 -3.360 0.000780 \*\*\*

cosmonth 8.055e+00 2.658e+00 3.030 0.002444 \*\*

sinday -2.160e-01 2.930e-01 -0.737 0.461037

cosday -2.935e-01 2.944e-01 -0.997 0.318665

cosweek -4.933e+00 2.693e+00 -1.832 0.067018 .

sinweek 8.954e+00 2.667e+00 3.357 0.000788 \*\*\*

sinhour 5.015e-01 2.937e-01 1.707 0.087762 .

coshour -4.349e-01 3.171e-01 -1.372 0.170182

Energy.Export..kWh.1 2.664e-01 5.414e-03 49.204 < 2e-16 \*\*\*

Energy.Export..kWh.2 1.393e-01 5.599e-03 24.882 < 2e-16 \*\*\*

Energy.Export..kWh.3 2.133e-01 5.640e-03 37.821 < 2e-16 \*\*\*

Energy.Export..kWh.4 -8.425e-02 5.621e-03 -14.988 < 2e-16 \*\*\*

Energy.Export..kWh.5 -5.161e-03 5.523e-03 -0.935 0.350020

Energy.Export..kWh.6 4.450e-02 4.999e-03 8.902 < 2e-16 \*\*\*

**LINEAR REGRESSION – RELATIVE IMPORTANCE**

|  |  |  |
| --- | --- | --- |
| Wind.speed..m.s. | Energy.Export..kWh.1 | Energy.Export..kWh.2 |
| 0.169 | 0.15 | 0.135 |
| Energy.Export..kWh.3 | Energy.Export..kWh.4 | Energy.Export..kWh.5 |
| 0.134 | 0.114 | 0.11 |
| Energy.Export..kWh.6 | Blade.angle..pitch.position..A.... | Blade.angle..pitch.position..B.... |
| 0.108 | 0.016 | 0.016 |
| Blade.angle..pitch.position..C.... | cosweek | cosmonth |
| 0.016 | 0.011 | 0.01 |
| Nacelle.ambient.temperature...C. | Wind.direction.... | Nacelle.position.... |
| 0.005 | 0.002 | 0.002 |
| sinmonth | sinweek | sinday |
| 0.002 | 0.001 | 0 |
| cosday | sinhour | coshour |
| 0 | 0 | 0 |

LASSO REGRESSION OUTPUT – SAME AS ABOVE

Wind.speed..m.s. 0.4241853085

Wind.direction.... .

Nacelle.position.... -0.0070303372

Nacelle.ambient.temperature...C. -0.0075769854

Blade.angle..pitch.position..A.... 0.0222782776

Blade.angle..pitch.position..B.... .

Blade.angle..pitch.position..C.... 0.0004515191

sinmonth -0.0063357964

cosmonth 0.0058024955

sinday -0.0005205013

cosday -0.0006945822

cosweek .

sinweek .

sinhour 0.0023563492

coshour -0.0012387466

Energy.Export..kWh.1 0.2641087833

Energy.Export..kWh.2 0.1333201597

Energy.Export..kWh.3 0.2067052070

Energy.Export..kWh.4 -0.0674173342

Energy.Export..kWh.5 .

Energy.Export..kWh.6 0.0376127924

1. https://www.sciencedirect.com/science/article/pii/S0960148121017134?via%3Dihub#sec2.4 [↑](#footnote-ref-1)
2. <https://ieeexplore.ieee.org/document/7384769> - also potentially useful [↑](#footnote-ref-2)
3. http://www.sthda.com/english/articles/32-r-graphics-essentials/129-visualizing-multivariate-categorical-data/ [↑](#footnote-ref-3)