31761

RENEWABLES IN ELECTRICITY MARKETS

Assignment 3 - Forecast

Authors:	*Student Number:*
Sigurd Indrehus	s193028
Lorenzo Mininni	s192445
Jorge Montalvo Arvizu	s192184

Contents

List of Figures

1 Exploratory Data Analysis (EDA)

The aim of the EDA is to identify the main characteristics of the data set through initial investigations, in order to elaborate initial assumptions using visual representations and summary statistics [1]. At the first stage, the historical dataset consists of 26305 hourly observations N, with 4 features M: wind speed retrieved at two heights, 10m and 100m above sea level, given in term of their zonal and meridional components (u and v). Weather forecasts are also given and issued once a day at 00:00, with lead times between 1 and 24 hours ahead, as part of the forecast process. There are 459 missing values which were removed.

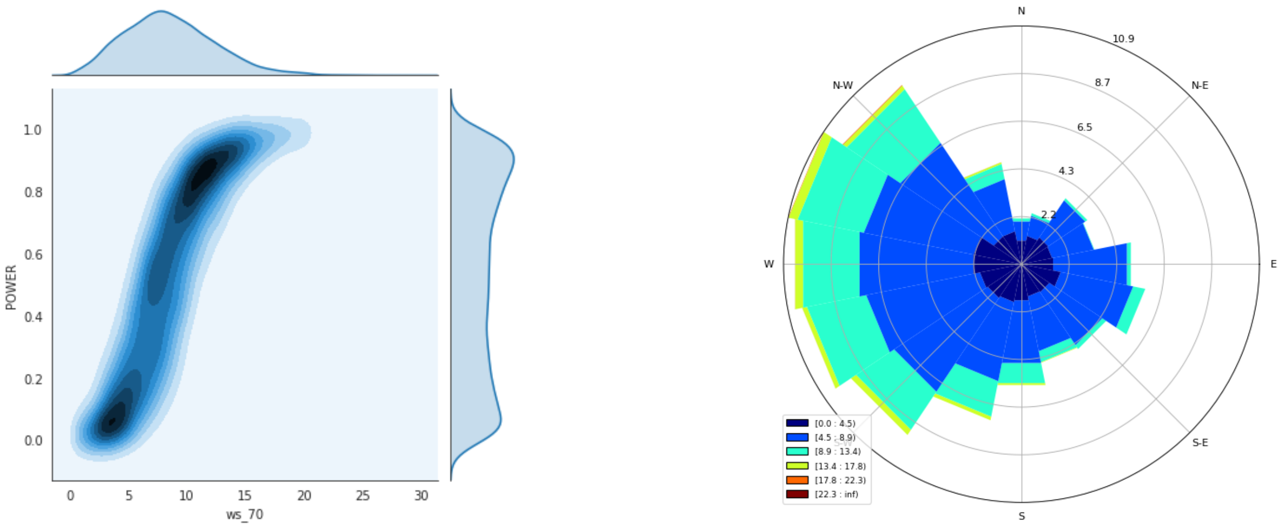
$$\mathrm{P} = \frac{1}{2}\rho A v^3 c_p \tag{1}$$

Equation 1 describes the physical relationship between power P, wind speed v (at hub height m), the turbine swept area A, the air density ρ, and the power coefficient c_v. This equation is important, since the generative story of the data that will be analyzed throughout the document is based on this specific formula. From **Figure 1a**, it can be seen that there is a certain wind speed needed before the turbines generate power, and also a plateau where the power stabilises before it reaches the cutoff speed where no power is generated due to safety reasons, the non-linear relationship between the two variables can be seen together with the fact that most of the wind speed measurements are between 0 and 15 m/s, while there are two concentrations of wind power around 0.8 and 0.2, i.e. at high and low production. The power also depends on the wind direction, normal to the blades. As seen in **Figure 1b**, the wind direction for Horns Rev 1 is mainly coming from the west which corresponds to the geographical location of the wind farm. However, as it can be verified in **Table 1**, the standard deviation of the wind direction variable indicates the presence of high variation in the measurements. The description of how the variables were transformed is explained in **Section 2**.

As seen in **Equation 1**, the density of the air also impacts the power generation. The density varies with temperature, which again depends on the time of day and season. Thus, the forecast should consider variations between i.e. day and night, and winter and summer. Furthermore, the wind speed depends on the height, i.e. the wind speed at 50 m is different from the wind speed at 250 m. Thus the hub height of the wind turbine (in this case, 70 m) has to be considered in order to properly estimate the power generated.

Table 1: Descriptive statistic parameters

	Power	**Wind speed**	**Wind direction**
Mean	0.520661	8.557326	210.380281
std	0.312567	3.865215	89.400875
min	0	0.052816	1.280727
25%	0.235322	5.793083	140.490232
50%	0.542307	8.249827	228.729195
75%	0.812662	10.972296	283.821567
max	1	29.805431	359.986437



(a) Wind speed - Power plot with KDE. The x-axis represents the wind speed at hub height 70m, while the y-axis represents the normalized wind power of the wind farm. KDE represents the density of the measurements.

(b) Windrose showing wind direction and speed (m/s) for the given location.

Figure 1: Wind speed - Power plot (a) and wind rose b).

2 Feature Engineering

In order to include the insights from the EDA, the following feature engineering steps were performed. First of all, missing observations are dropped since the models won't be able to impute the missing values or learn from them. Then, using the two components u and v, two functions are implemented to transform them into wind speed w_s and its direction w_d, using the following equations and taking into account the meteorology convention:

$$w_s = \sqrt{u^2 + v^2} \tag{2}$$

$$\text{"From" or meteorology convention: } w_d = 180° + \frac{180°}{\pi} \arctan(u, v) \tag{3}$$

Since the hub height of the turbines at Horns Rev 1 is at 70 meters, while the weather forecasts are given for 10 and 100 meters height, the wind speed and direction at 70 meters height have been estimated using the wind profile power law. The relationship shown in **Equation 4** has been used, where V_1 and V_2 is the wind speed at height Z_1 and Z_2 respectively, and α is a wind shear exponent.

$$v_2 = v_1 \cdot \left(\frac{z_2}{z_1}\right)^{\alpha} \qquad \alpha = \frac{log(\frac{v_2}{v_1})}{log(\frac{z_2}{z_1})} \tag{4}$$

It is important to underline that the α value used is not the expected value of the variable ($\mathbb{E}[\alpha] = 0.137$) rather its value at each hour, therefore every measurement is obtained with a different α. The columns of u and v at 10 and 100 meters are then dropped. Then, dummy variables are introduced to represent the month and hour of each measurement in the data set, assuming value 1 for the corresponding hour and month and zero for all the others. Therefore, 24+12 new features were added to the feature matrix **X**. These variables have the purpose of capturing the seasonality of the measurements. Therefore, the number of features M used for the forecasting is 38: 36 dummy variables, and wind speed and wind direction at 70m of height. Together with the observations N, the matrices used by the models are $\mathbf{X} = 25746 \times 38$ and $\mathbf{y} = 25746 \times 1$, where $\mathbf{y}$ is a vector with the power measurements and **X** the features matrix. Another important decision taken regarding the data is the standardization of the wind speed and wind direction, since some models are sensitive to the scaling of the variables, e.g. neural networks. The standardization is operated using the following equation, in which μ is the mean and σ the standard deviation of the feature.

$$\tilde{\mathbf{x}}_i = \frac{x_i - \mu}{\sigma} \tag{5}$$

3 Modelling

The forecasts performance is evaluated throughout the stages based on the following scores: bias, MAE, RMSE, and R2, as a function of the lead time. These scores, as well as diagnostic plots and visual inspection of the data, will help decide which models are performing better, and what to improve for future forecasts. This step is vital for any forecast improvement, as it is important to know how good a forecast performed, and how to improve it, in order to tweak the model in forecasting and learning from new data.

3.1 Stage 1

For the first stage, the following models were created and compared: random, persistence, climatology, unregularized linear regression, and ridge regression. The **random** model consists of a random guess at time t for each lead time $t + k$ with the following equation:

$$\hat{y}_{t+k|t} = u_k \quad \forall k, \text{ where } u_k \sim \mathcal{U}[0, 1] \tag{6}$$

The **persistence** approach is based on issuing a forecast at time t for all lead times $t + k$ according to the last m measurements available. Therefore, having y_{t-i} as last available date, the persistence forecast can be described as:

$$\hat{y}_{t+k|t} = \sum_{i=1}^{m} y_{t-i}, \forall k, \text{for any } m \text{ selected} \tag{7}$$

The **climatology** approach consists in providing a prediction at time t for all the $t + k$ lead times based on the average of the whole data set of measurements. The model can be described as:

$$\hat{y}_{t+k|t} = \sum_{i=1} y_{t-i}, \forall k \tag{8}$$

Linear regression aims to explain the relationship between a target variable (power generation $\mathbf{y_i}$) and explanatory variables $\mathbf{x_i}$ which are influencing $\mathbf{y_i}$. In this case, $\mathbf{X}$ includes all the 38 features defined previously. Then, the model can be described as in **Equation 9**, where β includes β_0 (the intercept) and β_i (slope), where i relates to each of the 38 features, while ε_i is the noise term. The noise term represents the forecast error which is desired to be as low as possible.

$$y_i = \boldsymbol{\beta}^\top \mathbf{x}_i + \varepsilon_i, \quad i = t-n, \ldots, t \tag{9}$$

Ridge regression is instead a regularized linear regression approach. The reason for this implementation can be found in the fact that, while performing a linear regression, the risk of overfitting the model is high. In fact, one feature might distort a lot the model under certain conditions (i.e. a very high leverage) by changing its β parameter and therefore not allowing it to perform well. Therefore, a regularization parameter is defined, which constraints the value of the weights. In this case, α is the regularization parameter (defined equal to 1 for stage 1 as a exploratory decision for stage 1), satisfying the following minimization equation:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{\sqrt{n}} \sum_i \varepsilon_i^2 + \alpha \sum_j \beta_j \quad \text{where } \alpha : \text{regularization parameter} \tag{10}$$

3.1.1 Results Stage 1

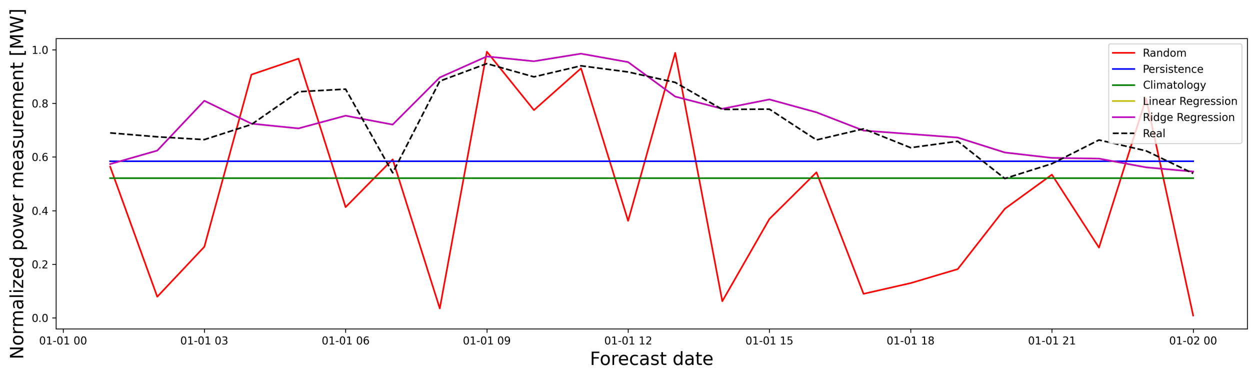


Figure 2: Actual power measurements compared to the models considered for stage 1.

The forecast and actual measurement for each hour for the whole day of 01.01.2018 can be seen in **Figure 2**. The regression models (linear and ridge) are very similar, and follow each other almost completely, as no difference can be seen in **Figure 2**. For short lead times linear models perform quite well, providing forecast similar to the actual measurements. In general, the forecast shows a smoother pattern than the actual measurement. Looking at the simplest models, the persistence is performing better than the climatology in this case, which probably can be explained by the fact that it provides a better results for shorter lead times, therefore giving a better performance for a 24 hour forecast. The random model is the one with the worst result, since its performance is unstable and unpredictable, performing quite well like during hour 11 but at the same time having a very low accuracy in hour 14. Eventually, the three baseline models all perform worse than the regression models, and will not be considered for later forecast stages. In general, the regression models fail to account for the non-linear relationship of the features against the target variable, as can be seen on the additional plots (**Figure 15** and **Figure 16**) for stage 1 in the **Appendix A**.

3.2 Stage 2

In this stage, a neural network and a regularized linear regression model was developed including two non-linear features: squared and cubed wind speed. Cross-validation was also added to each model. The objective in this stage was to better represent the non-linearity of the target variable given the features, as well as solidify the linear models with cross-validation.

For the **regularized linear regression**, in order to define the regularization parameter, 10-fold **nested** cross-validation is performed, subdividing the data set in 10 parts and considering each one of them as validation set while the others as training data. Therefore, only nine of them will be used to effectively build the model while the other one will be evaluating the quality of it. Eventually, the parameters are chosen by taking the average of the 10 cases, as shown in **Figure 3**. For each split, the training set is further divided into test and training set, to determine the value of the hyperparameter λ. This value is chosen as the one giving the lowest validation error. Then, based on the optimal hyper-parameter obtained from the sub-splits, the normal training is performed on the current split.

Figure 3: 10-fold cross validation with average [3]

In order to assess a non-linear method more directly, a **Neural Network** (NN) was developed. A NN is a set of connected neurons which operates iteratively based on feed-forward and backpropagation steps, in order to find out the relationship between values given as input against output by optimizing a cost function and adjusting the weights connecting each layer of neurons. In between, a certain number of *hidden layers* are recursively connected (in our case 2 hidden layers with M neurons), until the defined loss function (MSE) is lower than a certain tolerance or the maximum number of iterations is reached. In **Figure 4**, the general architecture of a NN is presented. In this model, the number of total features considered is 38, without taking into account the squared and cubed wind speed. Therefore, the architecture of the neural network is defined as having 38 inputs, one single output, and two hidden layers, all with ReLU as the activation function, since it provides the best overall performance, especially in terms of convergence time and simplicity [5], further avoiding the vanishing gradient problem present with the sigmoid activation function.

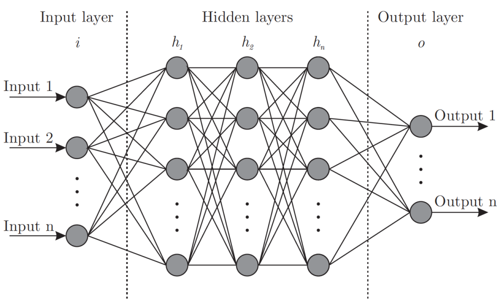


Figure 4: Neural Network - general definition [4].

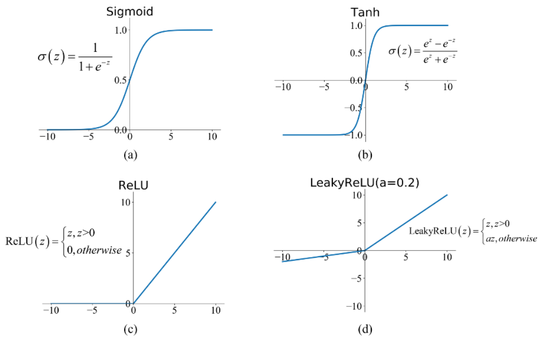


Figure 5: Main activation functions - ReLU is the chosen one [6]

3.2.1 Results Stage 2

Table 2: Scores - Stage 2

	BIAS	MAE	RMSE	R2
Linear Regression	-0.030441	0.103860	0.141952	0.807409
Regularized Linear Regression	-0.030959	0.104337	0.142444	0.806255
Ridge Regression	-0.030664	0.104055	0.142207	0.806789
Neural Network	-0.028646	0.099431	0.142560	0.800167

Figure 6 shows the forecasts compared to the actual measurement for stage two, for the whole month of January 2018 (except 1st of January). The forecast models fail to adapt to the rapid changes of the actual measurement, as can be clearly be seen between 16-20th January. Other than that, the forecasts follows the actual measurement well. Again, the regression models, also the regularized, are very similar, and little difference can be seen in **Figure 6**. The neural network perform slightly different than the regression models, although fails to beat any of them in means of RMSE as seen in **Table 2**. In fact, the linear regression model gives the lowest RMSE at this stage. The neural network gives the highest RMSE but the lowest MAE, which indicates that there are more large errors in the neural network model. All the models have a similar average bias, with the neural network being closer to zero than the others, meaning the consistent difference between the actual outcome is lower. Thus all the models tend to forecast less power than what is actually measured. **Figure 7a** shows the average forecast error for 24 hours lead time for the chosen regularized LR. The forecast error tends to increase with lead time, which is natural for forecasts as the uncertainty increases with lead time.

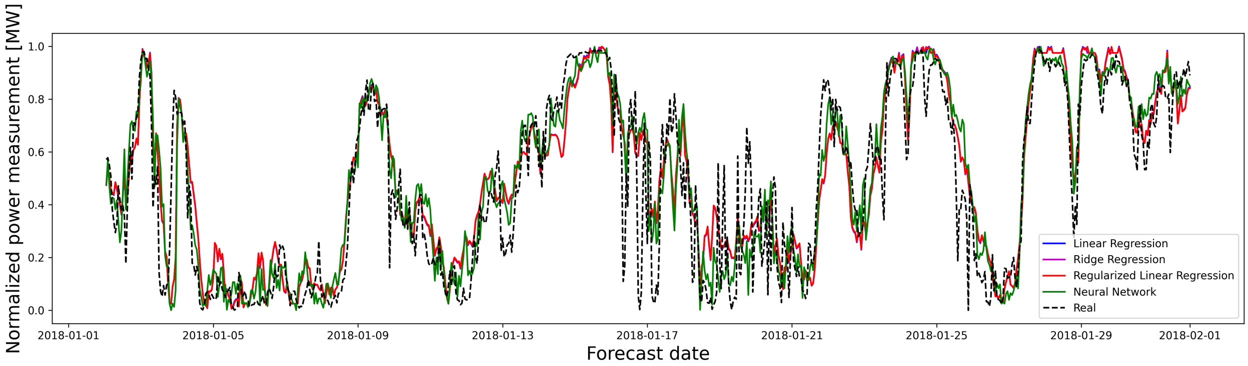
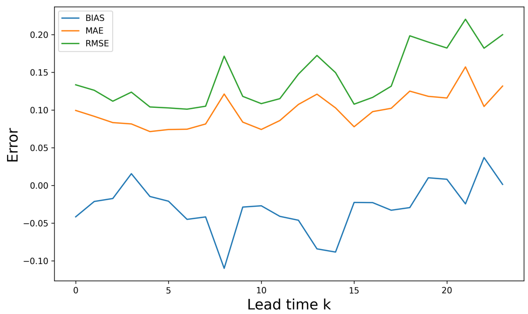
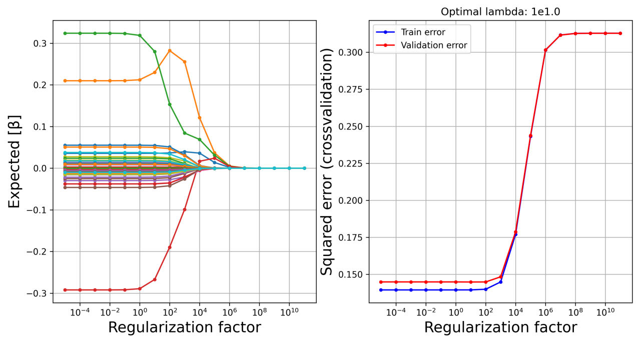


Figure 6: Actual power measurements compared to models forecasting - Stage 2.



(a) Forecast error per lead time k.



(b) Optimal regularization parameter λ.

Figure 7: Forecast error (a) and optimal regularization parameter (b).

Figure 7b shows the optimal regularization parameter λ for nested cross-validation. The optimal λ is 10, which gives the lowest validation error.

As seen by **Figure 8 and 9**, the trend is similar as for stage one, where the spread in the data is not properly accounted for. The predicted power curve is much more narrow than the actual.

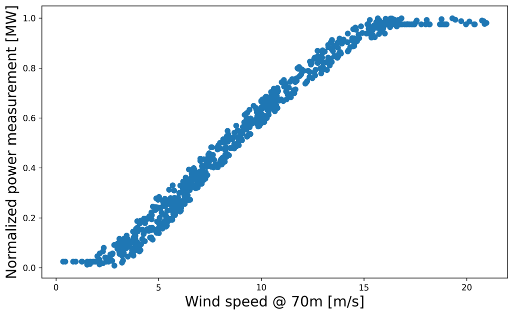


Figure 8: Predicted power curve for stage 2.

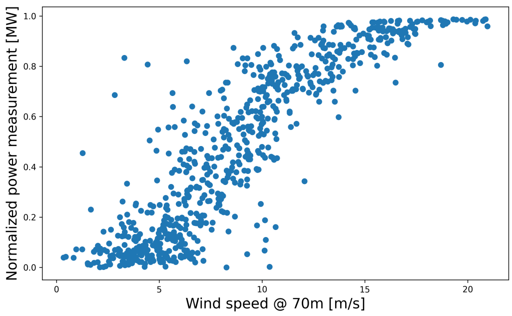


Figure 9: Actual power curve for stage 2.

3.3 Stage 3

During this stage, another two methods are implemented: a Bayesian Linear Regression and Catboost. The **Bayesian linear regression** operates in the same way of the linear regressions already implemented, but it defines a prior distribution to each β parameters based on a normal distribution instead as point estimates. Therefore, the y_i response variable will be as well derived from a normal distribution, as shown in **Equation 11**. Uncertainty is then assessed in a proper way, since all the possible values of the parameters' distributions are taken into account, instead of having just one single estimation.

$$y_n \sim \mathcal{N}\left(\alpha + \beta^T \mathbf{x}_n, \sigma^2\right) \tag{11}$$

The bayesian linear regression is based on applying Bayes' Theorem **Equation 12**, so that a value subject to uncertainty (in this case, a forecast) can be derived from a certain data set by updating its distribution given observed data.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \quad \text{with } P(x|y) \text{ probability of x knowing y is true} \tag{12}$$

Catboost, on the other hand, is a pre-defined algorithm based on gradient-boosted decision trees. In fact, it builds a consecutive set of decisions using a training dataset. The loss function of the model is then checked against the validation set. The main advantage of using Catboost is that it avoids doing the average of all y_i after cross validation and therefore reducing the risk of *prediction shift*. In fact, it performs a random permutation on all the training sets maintaining one single supporting model (the decision tree), common to all of them [7].

3.3.1 Results Stage 3

Table 3: Scores - Stage 3

	BIAS	MAE	RMSE	R2
Regularized Linear Regression	-0.030557	0.104633	0.137634	0.822211
Bayesian Linear Regression	-0.031434	0.105158	0.138106	0.821022
Neural Network	-0.035516	0.090795	0.123064	0.857716
Catboost	-0.027449	0.079767	0.107856	0.889642

As seen from **Table 3**, the catboost model performs better than the other models, in RMSE, MAE and bias. There also seems to be a slight phase shift between the catboost and the actual measurement as in day 5 and 18 in **Figure 17** on the A, where the catboost follows very similar pattern as the actual measurement, while slightly shifted to the right. For stage 3, the neural network performed better than the regression models unlike in stage 2. We account this correction in the neural network given the optimization on the cross-validation procedure.

In **Figure 10** catboost forecast captures better the error spread at medium wind speeds. The average forecast error for 24 hours lead time for the catboost model is shown in **Figure 12**. The increasing trend is not as evident as it was for the model in stage 2, since the error decreases until lead time 8, then it tends to increase before it drops heavily for the last three hours. This shows a good behaviour of the model by addressing the stationarity of the data.

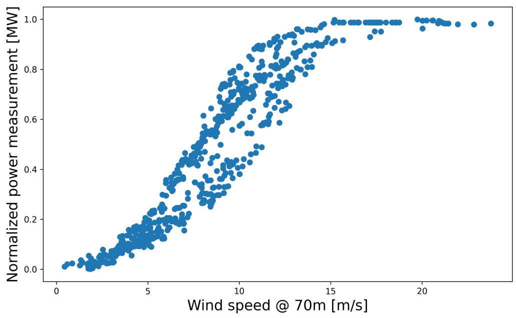


Figure 10: Predicted power curve for stage 3.

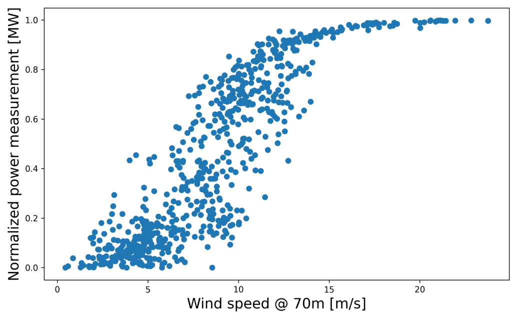


Figure 11: Actual power curve for stage 3.

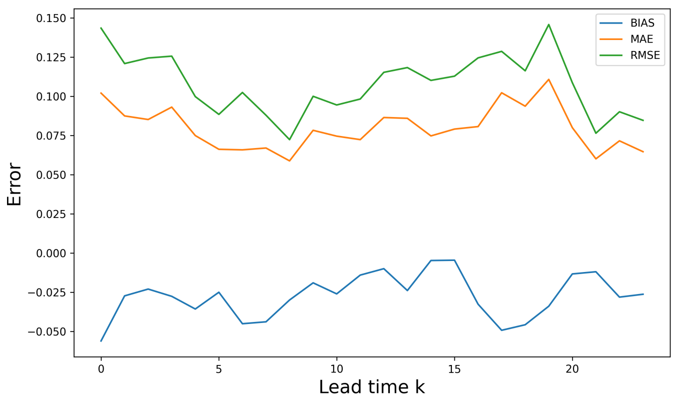
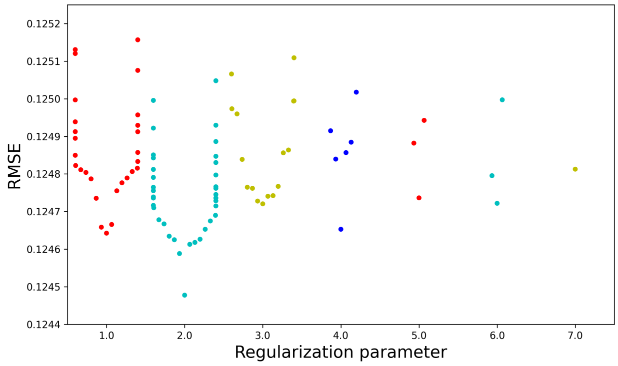


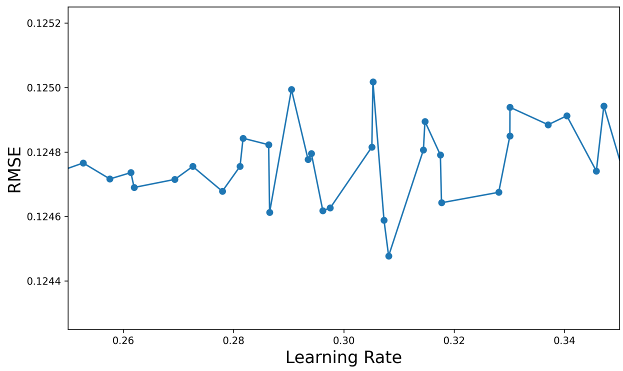
Figure 12: Forecast error per lead time stage 3.

3.4 Stage 4

For the last stage of developing the forecast, the catboost model was improved with with hyperparameters optimization. It was then tested with the same data as the catboost model from stage 3 in order to understand if it would have performed better under the same conditions. The updated catboost model gives slightly lower errors, providing then an overall better performance with a RMSE of 0.107311, MAE of 0.078654, bias of -0.025716, and R2 of 0.890836. Two hyperparameters were tuned during this stage: L2 leaf regularization and learning rate. The first one adjusting the degree of regularization at each leaf of the decision tree, while the latter adjusts the gradient step of the algorithm. The resulting optimal hyperparameters, as shown in **Figure 13**, are $\lambda = 2$, and $LR = 0.3081$.



(a) RMSE for each regularization parameter.



(b) RMSE for each learning rate.

Figure 13: RMSE for regularization parameter (a) and RMSE for learning rates b).

4 Future work and conclusion

In this report the development of wind power forecast based on given weather forecast for a off-shore wind farm in western Denmark over four stages have been documented. An explanatory data analysis was initially performed to understand the data, and to develop understanding on how to improve the forecast for later stages. As the results from previous forecasts were obtained, the forecasts were improved for the next stage. The initial regression models failed to properly account for the non-linear relationship between wind speed and wind power, and thus a Catboost model was eventually developed, showing improved bias, MAE and RMSE, while taking into account the spread in the data and the seasonality. The Catboost model was then eventually improved by including hyperparameter optimization. As future work, given that the training data was based on forecasts, a database with actual wind speed measurement might substantially improve the models developed. Also, working with uncertainty as in this assignment could be beneficial for a bayesian-based model such as a bayesian neural network, gaussian processes, or a bayesian physics-informed neural network, given the uncertainty nature of the weather variables. Hence, other variables that may account for the existing weather predictions at the wind farm might further explain away the relationship between the weather state and its effect on the wind turbine performance. Further, a combination of a physics-informed model with a data-based model might improve the forecasts, as it could take into account the actual physical laws governing the power output of the wind farm, e.g. by modelling the wake loss between turbine rows.

References

[1] Lenni pvgismine. *Solving the Problem of Overfitting - Regularized Linear Regression*. 2019. Available at: https://pvgisours.tistory.com/62?category=743773

[2] Prasad Patil. *What is Exploratory Data Analysis?*. 2018. Available at: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

[3] Ashfaque J - *Introduction to Support Vector Machines and Kernel Methods*

[4] Facundo B, Jimenez J. - *Prediction of wind pressure coefficients on building surfaces using Artificial Neural Networks*

[5] Krizhevsky A, Sutskever I, Hinton G - *ImageNet Classification with Deep Convolutional Neural Networks*

[6] Feng J, He X, Teng Q, Ren C, Chen H, Li1 Y - *Reconstruction of porous media from extremely limited information using conditional generative adversarial networks*

[7] Tal Peretz. *Mastering The New Generation of Gradient Boosting*, Towards data science. 2018 Available at: https://towardsdatascience.com/https-medium-com-talperetz24-mastering-the-new-generation-of-gradient-boo

A Appendix

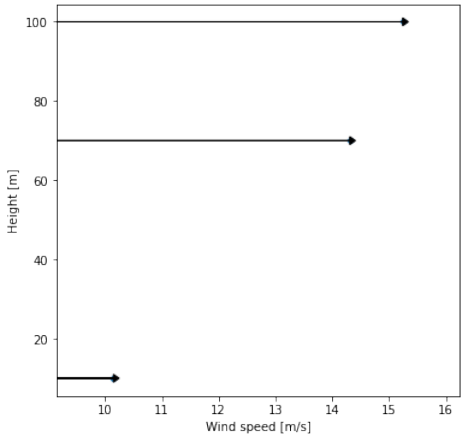


Figure 14: Wind speed at 10, 70 and 100 meters for observation 2000.

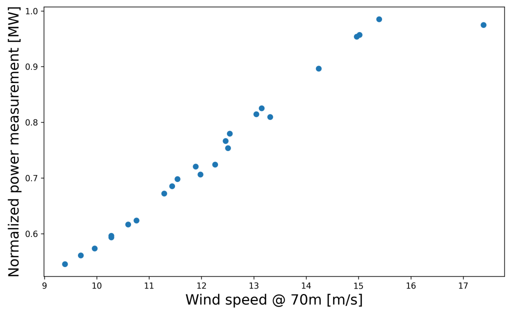


Figure 15: Predicted power curve for stage 1.

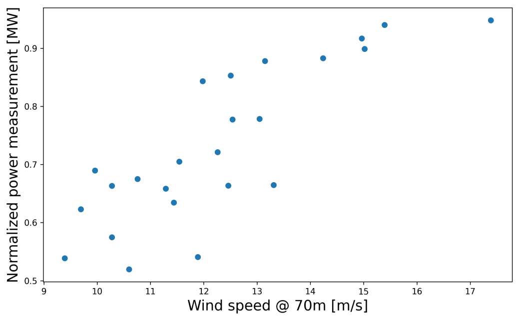


Figure 16: Actual power curve for stage 1.

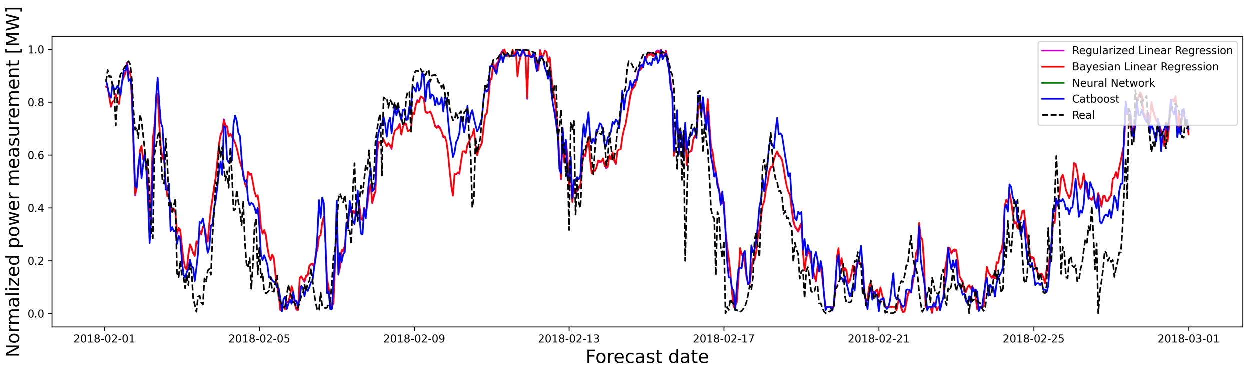


Figure 17: Actual power measurements compared to the models considered for stage 2.

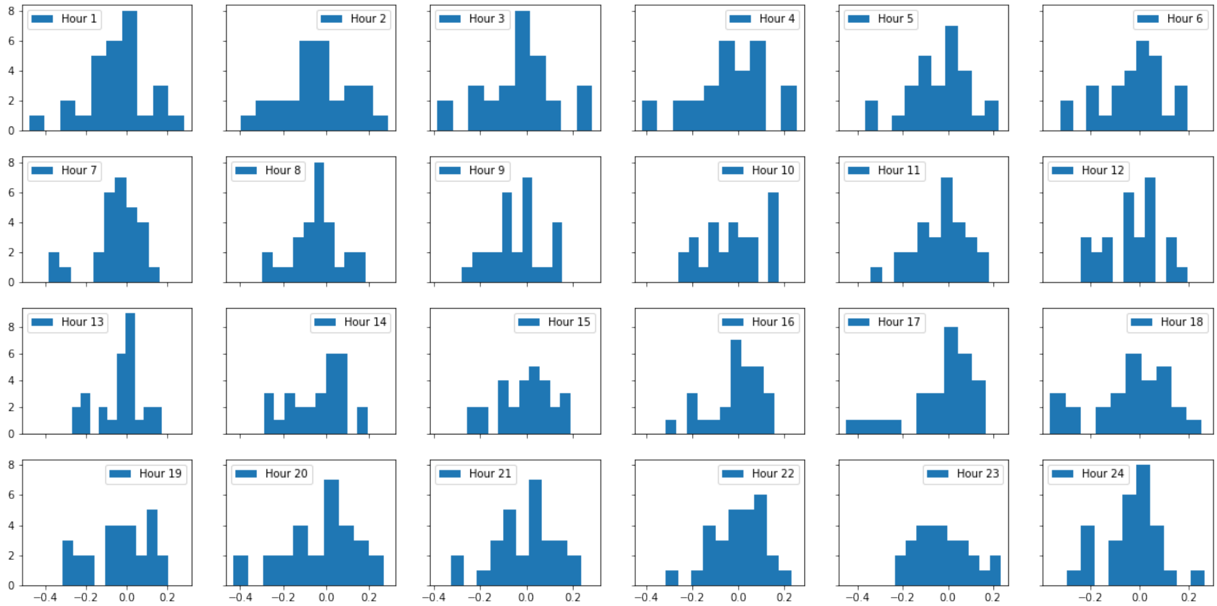


Figure 18: Distribution errors

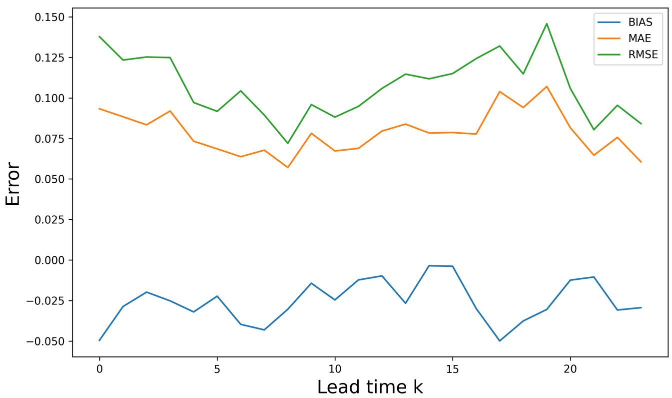


Figure 19: Forecast error per lead time stage 4.