Wind Energy Forecasting

Cosimo Fabbri, Daniele Hlatcu

June 4, 2023

Introduction

RenewaBolo Energy s.p.a. is a company based in Bologna and specialized in renewable energy production, with a primary focus on wind power. As a prominent player in the industry, RenewaBolo has a Data Science Department, which is entrusted with various statistical analysis, including forecasting problems.
In the following chapters, three different forecast scenarios encountered in the past years will be presented.

1 Problem statement

RenewaBolo sought to acquire a wind farm near Le Havre, France, from a smaller competitor, to establish itself as a leading renewable energy supplier in the French market. Recognizing the intermittent nature of wind energy and the need for informed decision-making about the acquisition, the company tasked its Data Science team with the development of two models: a *first model* to forecast the average energy generated, with the objective of providing valuable insights into the overall production potential; a *second model* for predicting the 10th quantile of power produced, with the aim of capturing worst-case scenarios where energy production fell considerably below the average.

1.1 Dataset

The dataset used for the analysis originally contained $6,904$ observations related to technical measurements of a wind tower, along with meteorological conditions parameters. For analytical purposes, all the observations with missing values were removed, resulting in a final dataset of $6,719$ observations with 15 variables, shown in **Table 1**.
Figure 1 showcases some important insights acquired through exploratory data analysis.

1.2 Average power generated

1.2.1 Method

The aim of the analysis was to predict the average power generated by the wind farm. Only the best models, obtained after variable selection (performed with the *best subset*

Name	Description	Type	Outcome
`hub.temp`	Hub temperature (°C)	Numeric	
`pitch.angle`	Pitch angle	Numeric	0-360°
`gen.speed`	Generator speed (m/s)	Numeric	
`gen.bear.temp`	Generator bearing temperature (°C)	Numeric	
`gen.stat.temp`	generator stator temperature (°C)	Numeric	
`gear.bear.temp`	Gear bearing temperature (°C)	Numeric	
`gear.oilsump.temp`	Gear oilsump temperature (°C)	Numeric	
`nacelle.angle`	Angle of nacelle position	Numeric	0-360°
`nacelle.temp`	Nacelle temperature (°C)	Numeric	
`wind.speed`	Wind speed (m/s)	Numeric	
`wind.direction`	Wind direction	Numeric	0-360°
`temperature`	Temperature (°C)	Numeric	
`rotor.speed`	Rotor speed (m/s)	Numeric	
`rotor.bear.temp`	Rotor bearing temperature (°C)	Numeric	
`power`	Power generated (kW)	Numeric	

Table 1: Variables in the dataset

selection method), are presented in the report. Let Y be the power produced and X the matrix of covariates chosen for the models:

$$X = \begin{bmatrix} \texttt{pitch.angle} & \texttt{hub.temp} & \texttt{gen.speed} & \texttt{gen.speed}^2 \\ \texttt{gen.speed}^3 & \texttt{gen.bear.temp} & \texttt{gen.stat.temp} & \texttt{wind.speed} \\ \texttt{wind.speed}^2 & \texttt{wind.speed}^3 & \texttt{rotor.bear.temp} & \end{bmatrix}$$

The first model examined was a *Normal Linear Regression*, defined as:

$$Y = \alpha + X^T\beta + \epsilon, \quad \epsilon \sim N(\mu, \sigma^2) \tag{1}$$

Next, based on model residuals' distribution, depicted in **Figure 2**, a *Student's t Regression* was fitted[1]:

$$Y = \alpha + X^{*T}\beta + \sigma\epsilon, \quad \epsilon \sim t(\nu) \tag{2}$$

Lastly, due to the presence of numerous observations where energy was not produced, as shown in the histogram of `power` **(Figure 1)**, a *Tweedie-Based Regression* was employed. The Tweedie distribution is particularly suitable for handling variables characterised by a mixture of zero and not-negative data points (Tweedie et al., 1984). Formally, the model was defined as:

$$Y \sim Twd(\mu, \phi, p), \quad \log(\mu) = X^T\beta + \epsilon; \tag{3}$$

with $E(Y) = \mu$ and $VAR(Y) = \phi\mu^p$, $\phi > 0$.[2]

[1]For technical reasons, the quadratic and cubic terms have been removed from the *Student's t Regression*, resulting in a modified covariates matrix X*.

[2]By varying the *power parameter* p, the distribution of Y changes and can degenerate in other known distributions. Some important cases include: Normal Distribution ($p = 0$), Poisson Distribution ($p = 1$) and Gamma Distribution ($p = 2$). When $1 < p < 2$, the Tweedie distribution assumes the shape of a Compound Poisson-Gamma Distribution (Hasan and Dunn, 2012), which was the distribution the Data Science team was looking for.

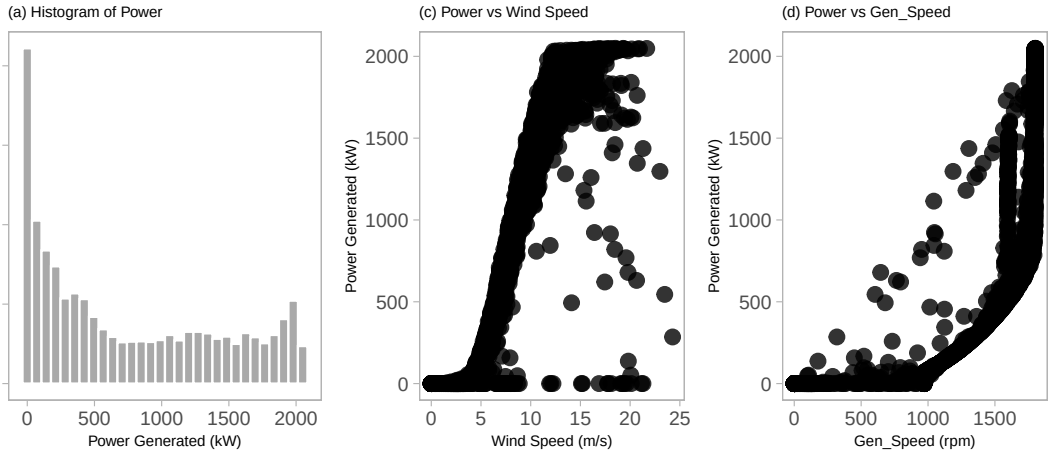

Figure 1: (a) Histogram of `power`, highlighting the abundance of observations without power generated. (b) Scatterplot of `wind.speed` against `power`. (c) Scatterplot of `gen.speed` against `power`. Both scatterplots suggest a potential cubic relationship between the independent and dependent variable.

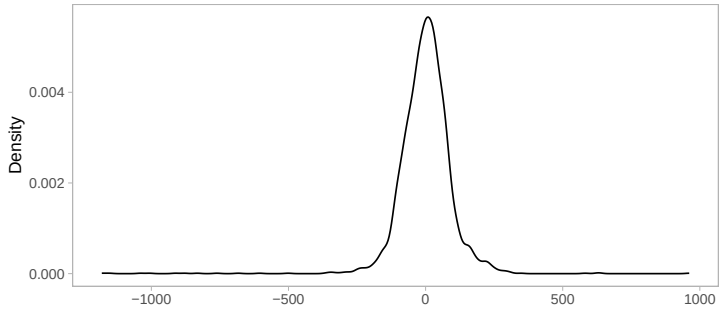

Figure 2: Residuals' Distribution of the Normal Regression.

1.2.2 Analysis

The dataset was randomly split into two parts, with half of the observations $(3,360)$ allocated to both the *train set* (used for model estimation) and the *test set* (used for evaluating prediction accuracy).
By examining **Figure 3**, which displays the predicted values of the regression models against the actual values, it is evident that the *Tweedie regression* outperforms the other models: it does not exhibit any distinct pattern around zero, and the forecasted average values are always non-negative, as expected. Consequently, the Mean Squared Error (MSE) of `Model` $\texttt{III}_a$ is significantly lower than the MSE values of the other model specifications, as shown in **Table 2**.

Model	**Specification**	**MSE**
$\texttt{I}_a$	`Normal Regression`	$9,173$
$\texttt{II}_a$	`Student's t Regression`	$26,405$
$\texttt{III}_a$	`Tweedie Regression`	$2,926$

Table 2: MSE of the different regression models

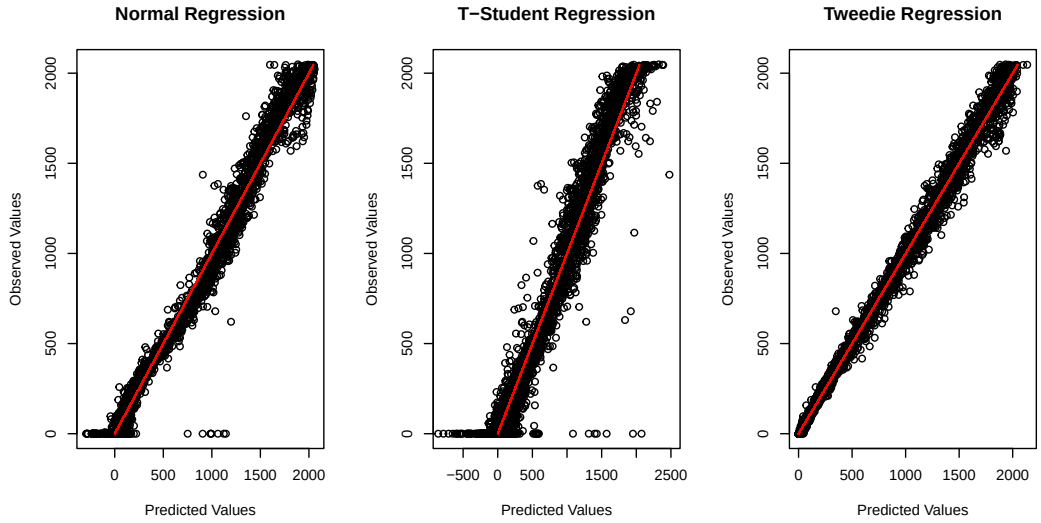

Figure 3: Predicted values vs Observed values. From left to right: *Normal Regression*, *Student's t Regression* and *Tweedie Regression*. Note that in *Tweedie Regression*, no prediction is below zero and there are no unusual patterns at the boundaries of the plot.

1.3 Quantiles estimation

1.3.1 Method

For the second task, the team pursued two different approaches: computing the quantiles using a Regression Model and fitting a *Quantile Regression*.
`Model` $\mathtt{I}_a$ was chosen as the starting point. Residuals' distribution of the model, shown in **Figure 2**, resembled a Student's t, so the decision was to explore both location scale and skewed distributions.
Firstly, the 10^{th} quantile for each observation was computed using:

$$\hat{q}_{\tau,i} = x_i^T \hat{\beta} + \hat{\sigma}_i F_t^{-1}(\tau) \tag{4}$$

where $F_t^{-1}(\tau)$ is the τ^{th} quantile of a Student's t distribution ($\tau = 10$ for this analysis)[3].
Subsequently, the quantile regression was specified as:

$$Y = X^T \beta + \epsilon, \quad Q_{10}(\epsilon|X) = 0 \tag{5}$$

1.3.2 Analysis

The same *training set* and *test set* from the previous problem were used to conduct the analysis and to evaluate the performances of the models. Each model's specification included the covariates specified in X.
To compare the quantile predictions of the models, an indicator variable[4] I was created, defined as:

$$I_i = 1(yi < q_{\tau,i}). \tag{6}$$

The three alternatives defined above were evaluated using a likelihood-ratio (LR) test, commonly employed to compare the goodness of fit of two models based on the ratio of

[3] $F_t^{-1}(10)$ depends on the specification for the residuals' distribution: location scale or skewed t.
[4] I compares each test observation y_i, with its corresponding quantile estimation, $\hat{q}_{\tau,i}$. If the observed value is smaller than the estimated quantile, $I_i = 1$.

their likelihoods[5].
Table 3 displays the p values of the LR test for each model, as well as the percentages of times where the test observations were smaller than the quantile estimations[6].

Model	**Method**	**E(I)**	**p-value**
$\texttt{I}_b$	`Location Scale Student t`	10.414%	0.4235
$\texttt{II}_b$	`Skewed Student t`	13.00%	0.0000
$\texttt{III}_b$	`Quantile Regression`	9.46%	0.2967

Table 3: Log-Likelihood test for the models

For both `Model` $\texttt{I}_b$ and `Model` $\texttt{III}_b$ there was not enough evidence for rejecting H0. In the end, `Model` $\texttt{I}_b$ was chosen as the best one, because it provided a higher value of E(I) than `Model` $\texttt{III}_b$ and, thus, demonstrated a more conservative approach. On the other hand, `Model` $\texttt{II}_b$ rejects H0, and so its estimate $\hat{q}$ is significantly different from the true quantile.

2 Problem statement

RenewaBolo faced height restrictions when expanding in Brasil: their existing 80-meters tall turbines exceeded the legal limit of 30 meters. However, a new deal allowed the firm to acquire 25-meters tall turbines, that could generate significant amounts of energy. The drawback is that they are unable to produce energy when the wind speed falls below 5 km/h. Consequently, RenewaBolo tasked its Data Science Department with the development of a model able to predict the production frequency.

2.1 Dataset

The dataset on which the analysis was performed consisted of $403,042$ observations across 27 variables, collected from several wind towers in the Mato Grosso region. After removing all the observations with at least a missing value (the majority of which had missing for all variables), the dataset was reduced to $161,570$ observations. For the purpose stated, only 10 variables were selected, as listed in **Table 4**. Important plots related to the the analysis are displayed in **Figure 4**.

2.2 Method

The aim was to develop a model that estimates the proportion of time in which the wind speed exceeds 5 km/h (`energy` = 1). Regardless of the model used, it was decided to penalize the *false positive* predictions (i.e. cases where the model predicts `energy` = 1, but the actual value is 0). This decision was based on the fact that it was more serious for the company to predict that energy was produced when actually it was not, as it incurred costs for preparing the wind turbines in advance, which involved both time and human resources.

[5]In this case, the hypotheses are H0: $E(I) = \tau$; H1: $E(I) \neq \tau$
[6]Ideally, this percentage should be 10%.

Name	Description	Type	Outcome
`precipitation`	Total Precipitation (mm)	Numeric	
`sl.pressure`	Air Pressure at the station level (mB)	Numeric	
`max.pressure`	Maximum Air Pressure (mB)	Numeric	
`min.pressure`	Minimum Air Pressure (mB)	Numeric	
`solar.radiation`	Solar Radiation (kj/m^2)	Numeric	
`dew.point.temp`	Dew Point Temperature (°C)	Numeric	
`max.temp`	Maximum Temperature (°C)	Numeric	
`min.temp`	Minimum Temperature (°C)	Numeric	
`relative.humid`	Relative Humidity (%)	Percentage	
`energy`	Energy produced	Binary	0/1

Table 4: Variables in the dataset

Consider all the covariates in the following matrix X:

$$X = \begin{bmatrix} \texttt{precipitation} & \texttt{sl.pressure} & \texttt{max.pressure} \\ \texttt{min.pressure} & \texttt{solar.radiation} & \texttt{dew.point.temp} \\ \texttt{max.temp} & \texttt{min.temp} & \texttt{relative.humidity} \end{bmatrix}$$

The first model analysed was a *Logistic Regression*, that defines the relationship between `energy` and the covariates as:

$$p = Pr(\texttt{energy} = 1|X) = \frac{1}{1 + \exp(-X'\beta)} \tag{7}$$

where p is a *Bernoulli* random variable representing the proportion of times in which `energy` was produced (and thus the wind speed exceeded 5 km/h). The parameters of the model were estimated by maximum likelihood, adding weights that doubled the likelihood value of observations with `energy`=0, in order to match the management request.

2.3 Analysis

The dataset was randomly split into two parts, allocating half of the observations (80, 785) to both the training set (used for estimation) and the test set (used for evaluation).
Two metrics were considered to evaluate the models: *accuracy* and *precision*. Accuracy assesses the overall correctness of the predictions, while precision focuses on correctly identifying positive instances.
Table 5 depicts the results for the three models presented. `Model I` comprised all the variables mentioned above, but performed poorly compared to `Model II`, which was obtained by applying variable selection using LASSO[7]. `Model II` excluded the covariates `dew.point.temp` and `max.temp` and lost 0.4 points in accuracy in respect to `Model I`, but it achieved a precision of 0.92, significantly better than the other models. Lastly, a Random Forest was considered in `Model III`. It performed well in terms of accuracy, but its precision was lower than `Model II`. Hence, `Model II` was identified as the best model for the purpose of the analysis, considering the preference for an higher precision and a lower False Positive rate.
An attempt was made to use the same model specifications for probit regression. The results were similar, thus the decision was to report only the logistic regression.

[7]LASSO: Least Absolute Shrinkage and Selection Operator

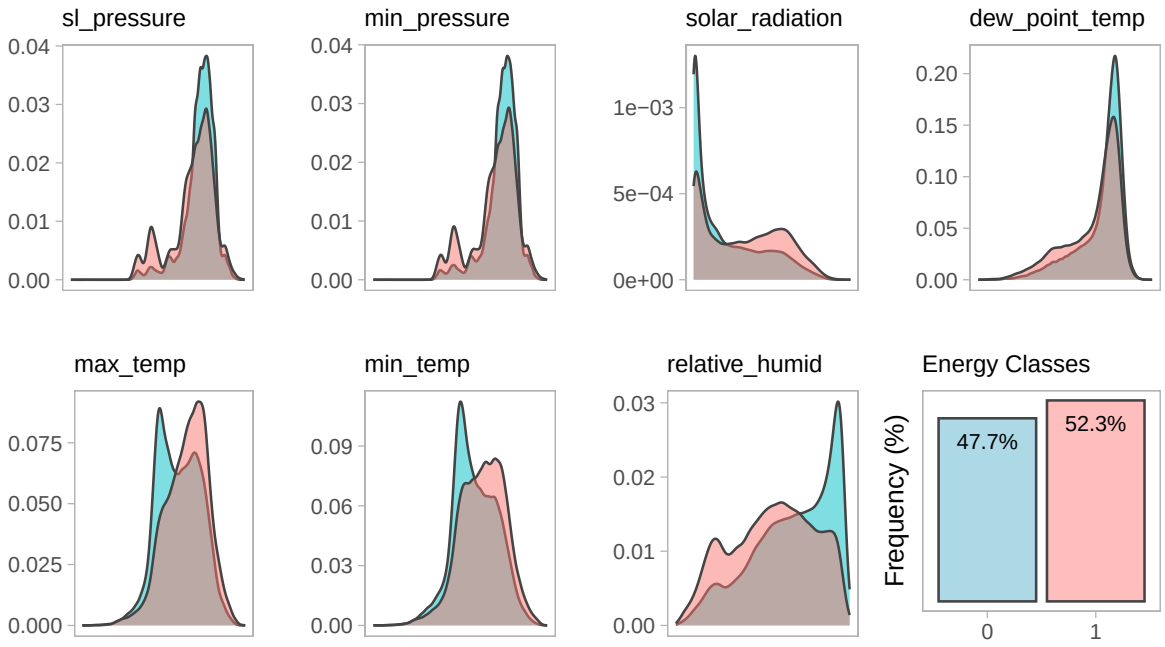

Figure 4: From top left to right: Graphical representation of the distribution of some important variables conditioning on whether `energy`=0 [light blue] or `energy`=1 [pink]; all of them display a different pattern depending on `energy` and were deemed appropriate to include in the models. Bottom right: barplot of energy, which shows balanced classes.

Model	Specification	Accuracy	Precision
I	Logit model	0.6492	0.6610
II	Logit model	0.6064	0.9232
III	Random Forest	0.7123	0.7425

Table 5: Accuracy and Precision of the models

3 Problem statement

Renewable energy suppliers can achieve substantial profitability by maintaining lower costs compared to the general gas-based energy price (Boer and Stet, 2022). Moreover, they often leverage tax break programs established by governments to support the advancement of renewable technologies (Regueiro-Ferreira and Cadaval Sampedro, 2022). However, the Spanish Government recently announced plans to eliminate tax breaks for renewable energy producers. As a result, selling energy at low costs may no longer be financially viable, as post-tax profits might not justify the investments. This has presented RenewaBolo, which entered the Spanish Market to benefit from tax breaks, with a weekly dilemma on whether to engage in energy production or deactivate the turbines. Their decision relies on predictions of the weekly price level. To address this challenge, the Data Science Team was tasked with developing a model to accurately forecast the weekly energy prices in Spain, enabling RenewaBolo to make informed decisions.

3.1 Dataset

The dataset used was collected by the Spanish Transmission Service Operator (TSO) Red Electrica Espana (REE). It includes 1460 observations of daily price data, spanning from January 1, 2015 to December 31, 2018. The variables are presented in **Table 6**, and the behaviour of the price over the 4 years is displayed in **Figure 5**.

Name	Description	Type
`Date`	Daily date	Date
`Price`	Energy Price €	Numeric

Table 6: Variables in the dataset

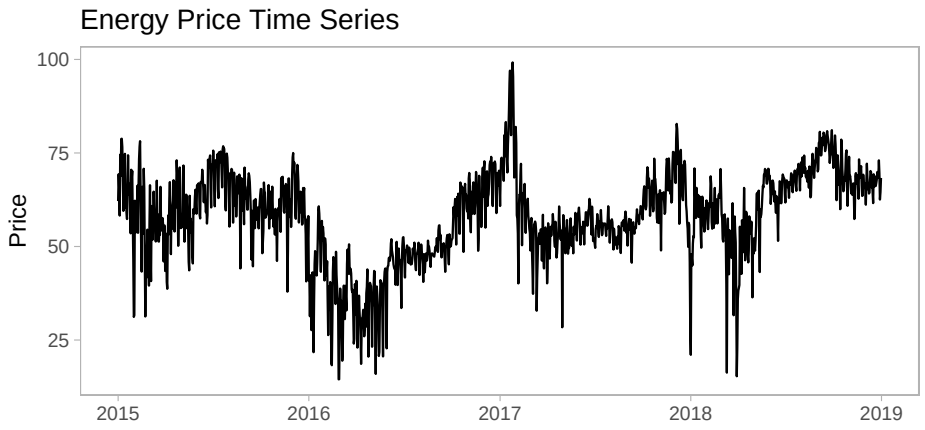

Figure 5: Time Series of the energy price in Spain between 2015 and 2018. Analyzing the time series, it is clear that there was a consistent behaviour throughout 2015, followed by a decrease at the beginning of 2016, which led the price to its four-year-minimum of €14.48. Subsequently, from 2016 to 2017, there was a rapid growth of price, which reached its peak value of €99.20 in January 2017. After returning to average levels, the price remained relatively stable throughout 2017 and the first half of 2018, with occasional negative fluctuations. The value of the time series exhibited a slight increase in the last six months. The average price over the period was €57.90.

Using time series decomposition tools, the team observed that there was no trend, but identified the presence of seasonality. In order to capture this pattern, various specifications (semester, quarters, months and daily) were tried and ultimately the team decided that price levels fluctuated based on the day of the week (**Figure 6**). Consequently, daily dummy variables were added in the modelling steps.

3.2 Method

The aim of the analysis was to predict future energy prices. The time series of 1460 observations was split into a *training* and *test set*, containing respectively 1095 observations (3 years of data, from 2015 to 2017) and 365 observations (2018).
Initially, the stationarity of the price series was studied. A Dickey-Fueller test was conducted to test the null hypothesis of non-stationarity ($H0 : \beta_1 = 1$) against the alternative hypothesis of stationarity ($H1 : \beta_1 < 1$). The model used for the hypothesis testing was:

$$\texttt{Price}_t = \beta_0 + \beta_1 \texttt{Price}_{t-1} + \sum_{j=1}^{6} \gamma_j \texttt{Day}_j + \epsilon_t \tag{8}$$

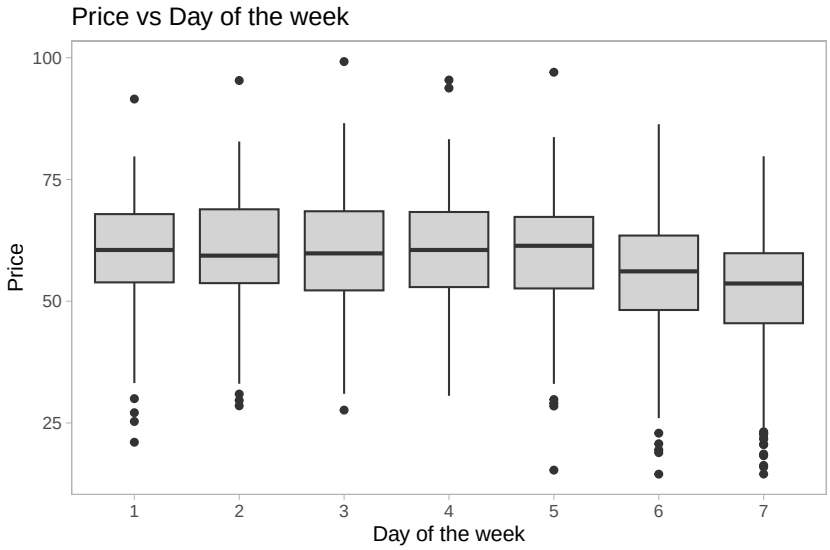

Figure 6: Boxplot of price depending on the day of the week. It can be observed that there was a significant decrease in price during the weekends, with a difference on average, in Saturday and Sunday, of respectively €5 and €8 less compared to the overall mean, accompanied by an increase in variance. This pattern may be due to week-end closure of companies. On the other hand, in the remaining five days, there were more subtle differences in median values. Despite this, a daily seasonality was chosen because in the model specified by Eq (8) all the dummies' coefficients γ_j were significantly different from 0, in respect to other seasonality specifications.

where the dummy variables $\texttt{D}_j$ are equal to 1 on the respective day they represent and 0 otherwise.[8] The result of the Dickey-Fueller test led to the rejection of H0, indicating that modelling should be performed on the stationary price series with the daily dummies.
After conducting several trials, the best model found based on BIC was the `ARIMAX(5,0,0)`, specified by the equation:

$$\texttt{Price}_t = \beta_0 + \sum_{i=1}^{5} \beta_i \texttt{Price}_{t-i} + \sum_{j=1}^{6} \gamma_j \texttt{Day}_j + \epsilon_t \tag{9}$$

In order to account for potential heteroskedasticity in the residuals, a Breusch-Pagan test was performed. The test resulted in a p-value of 0.051, slightly above the significance level of 0.05.[9] Consequently, a `GARCH(1,1)-ARIMAX(5,0,0)` model was estimated, following the formulation in Eq (9). However, the variance of the residuals is not assumed constant, but modeled as:

$$\sigma_t^2 = a_0 + a_1 \epsilon_{t-1}^2 + b_1 \sigma_{i-1}^2 \tag{10}$$

The choice of the order of the residuals variance was made by comparing the BIC values of different specifications, and the the `GARCH(1,1)-ARIMAX(5,0,0)` was selected because it had the smallest value of the metric.

[8] Note that there are only six dummy variables, as the effect of one of them is captured by the intercept β_0.

[9] In theory, this result does not bring to a rejection of the null hypothesis H0, indicating insufficient evidence to state that there is heteroskedasticity. However, due to the p-value being very close to the significance level of 5%, it is desirable to consider models that account for heteroskedasticity in the residuals.

3.3 Analysis

The forecasting was conducted in a short-term scenario, where the train set was updated with the respective test values after predicting 7 steps ahead (one week, as requested by management). The results for the two models are reported in **Table 7**.

Model	Specification	BIC	MSE
I	ARIMAX(5,0,0)	6,500	39.39
II	GARCH(1,1)-ARIMAX(5,0,0)	5,797	45.83

Table 7: MSE and BIC of time series models.

Model II has a smaller BIC value, indicating a better fit to the data. However, it exhibits less predictive power, with a MSE 16% higher than Model I. Based on this result, the final model chosen was ARIMAX(5,0,0), given the higher interest in the forecasting purpose. A visual comparison of the predictions for the two models is shown in **Figure 7**.

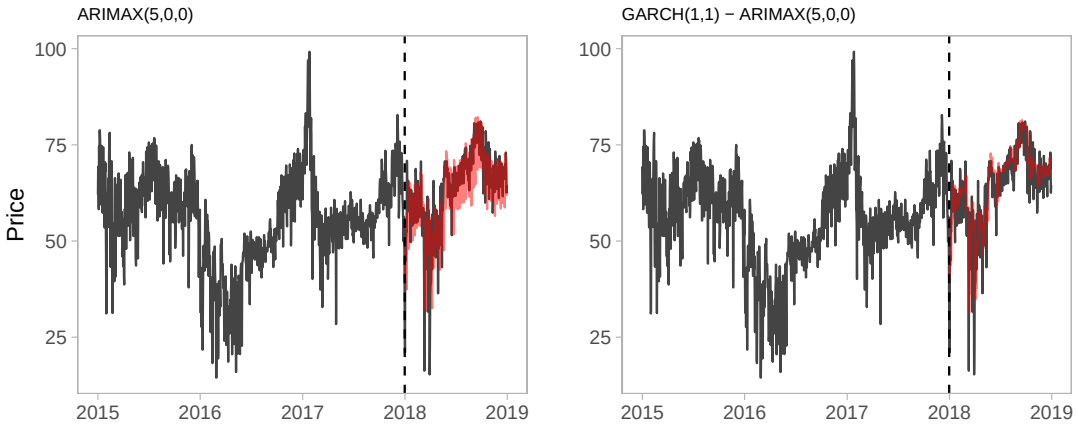

Figure 7: Visualization of 7-steps ahead forecasts for the two models. In red the predicted values. It is noticeable that Model I performed better overall, although in some dates it predicted a lower prices compared to the test values.

References

Boer, S. D. and Stet, C. The basics of electricity price formation. 2022.

Hasan, M. M. and Dunn, P. K. Understanding the effect of climatology on monthly rainfall amounts in australia using tweedie glms. *International Journal of Climatology*, 32(7):1006–1017, 2012.

Regueiro-Ferreira, R. M. and Cadaval Sampedro, M. Renewable energy taxes and environmental impacts: A critical reflection from the wind tax in spain. *Energy & Environment*, page 0958305X221083249, 2022.

Tweedie, M. C. et al. An index which distinguishes between some important exponential families. In *Statistics: Applications and new directions: Proc. Indian statistical institute golden Jubilee International conference*, volume 579, pages 579–604, 1984.