Final Project

Fundamentos matemáticos para el análisis de datos

Beatriz Torreiro Mosquera

November 17, 2019

1 Problem Statement

A dataset is provided containing information about a wind generation power station. There are 30 sensors located at different locations and altitudes of the station measuring temperature, solar radiation and wind speed and direction. The measures are taken every 15 minutes and for about a year.

The dataset contains 35136 observation and 124 variables:

- Date: date of the measurement
- Hour: hour of the day
- Min: minutes with respect to the hour
- TLXHY: temperature measured at location X for altitude Y
- GSRLXHY: solar radiation measured at location X for altitude Y
- WSLXHY: wind speed measured at location X for altitude Y
- WDLXHY: wind direction measured at location X for altitude Y
- WG: wind power generation of the station [MWh]

2 Problem 1

In this section, it will be discussed, how the power generation changes inside one hour of operation.

In order to study the power generated variations in one particular hour and throughout one day, it is necessary to know the average values of power generation throughout one day, in order to have an idea of the orders of magnitudes that are been used. The following graph shows the average power generation values during one day, Figure 1.

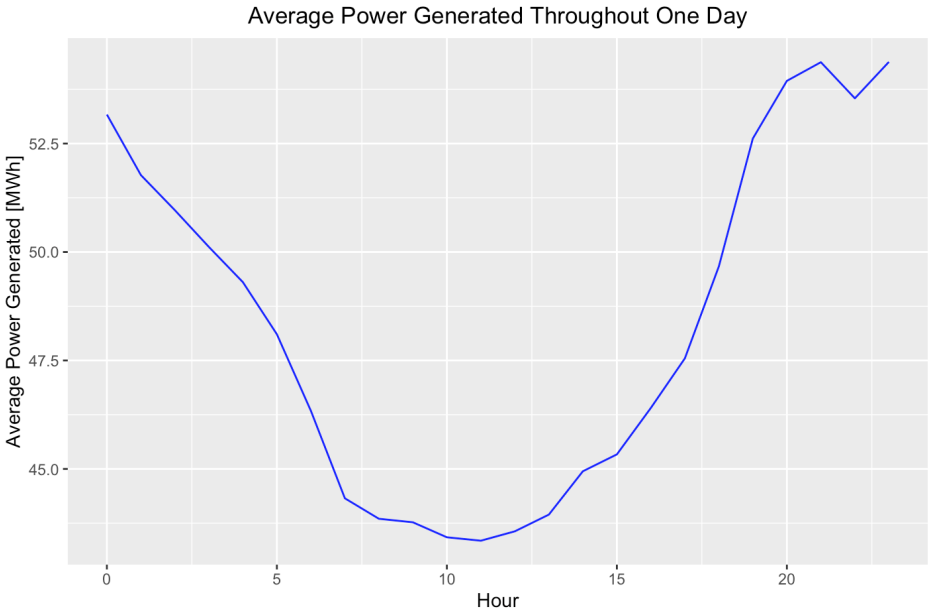


Figure 1: Average power generation throughout one day

As it can be seen in the previous image, the values of the power generated vary between 55 MWh and 40 MWh during an average day. With this information, we can better understand the power variations inside one hour.

To get a sense of the dimensions of the variation, the average range of variation and the maximum range of variation inside one hour have been calculated, Table 1. This measures have been calculated using absolute values, as in some hours the power will increase and in others it will decrease.

Range of variation of the power generated in one hour [MWh]	
Average	5.949993
Maximum	202.9
Minimum	0

Table 1: Variations of power inside one hour

In the previous table it can be seen that there is a big difference between the average and the maximum variation of power, this is due to the presence of outliers, Figure 2.

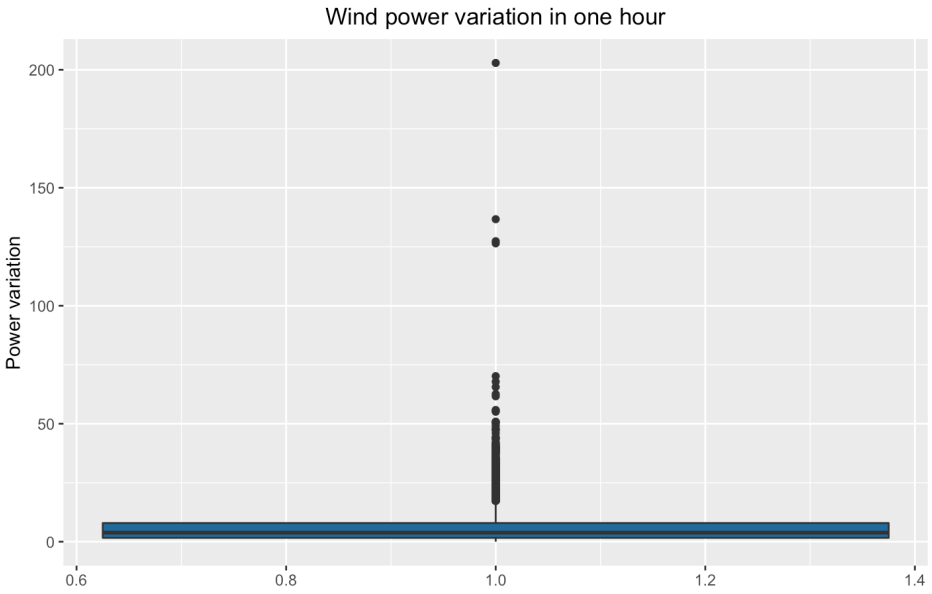


Figure 2: Boxplot of the variation of power in one hour

Studying how the power generation changes in one hour as a whole, does not give us a lot of information. That is the reason why the range of variation per hour has been studied. The following image,Figure 3, shows the boxplots of the power variation every hour of the day.

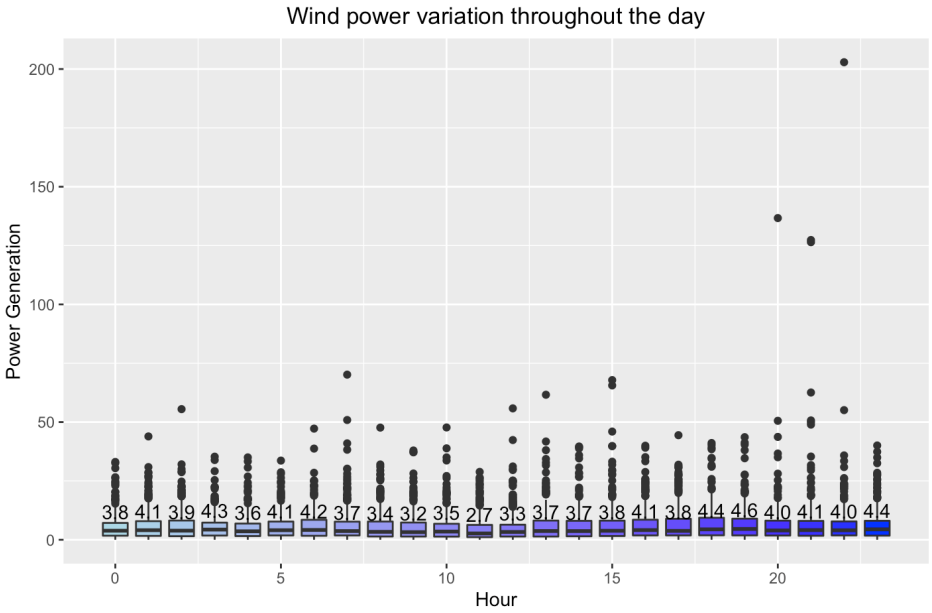


Figure 3: Boxplot of variation of power per hour of day

In the previous image, it can be seen that the presence of outlier measures is constant throughout the day. In order to be able, to better understand how the ranges of variation change throughout the day, the outliers have been removed in the following image, Figure 4.

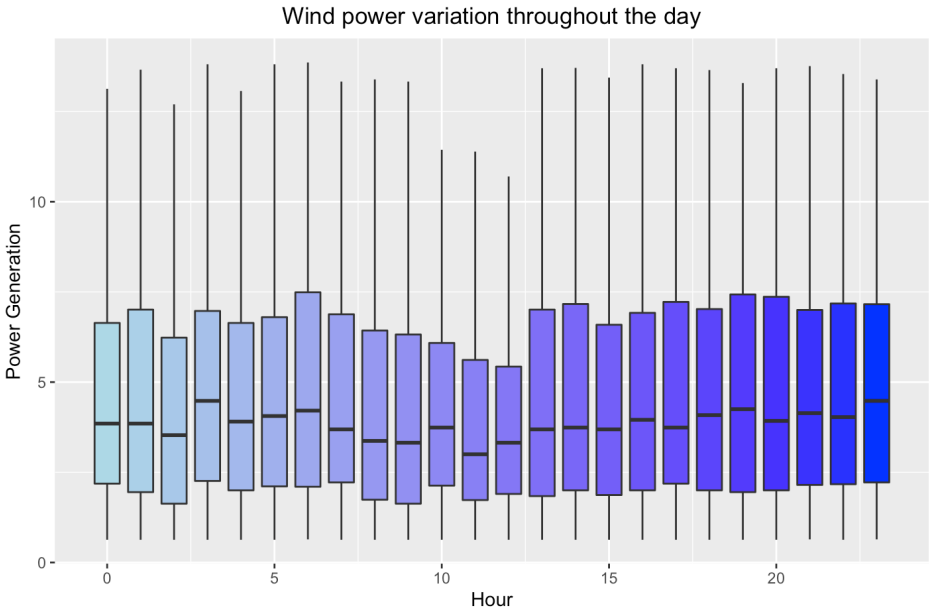


Figure 4: Boxplot of variation of power per hour of day without outliers

Moreover, the ranges of variation vary around the same levels in every hour, the median values oscillate between 2.69 MWh and 4.59 MWh.

3 Problem2

The sensor used for solar radiation in latitude 10 and altitude 100 is a new model of sensors and the supplier informed us that some of them are defective. As an indicator of these defective sensors the supplier has determined that if the difference between the measurement in latitude 9/altitude 100 and latitude 10/altitude 100 is higher than 100%, this measurement is wrong. Furthermore, the vendor says that the defective rate they have in production is between a 2% and 5%. Given the data obtained it has been decided to corroborate the reliability of the vendor.

In order to do so, the Bayes Theorem will be used, Equation 1, where:

- P(B) = prior knowledge
- P(A) = evidence of A, normalization constant
- $P(A|B)$ = likelihood of A given B
- $P(B|A)$ = posterior knowledge

$$P(B|A) = \frac{P(A|B) * P(B)}{P(A)} \tag{1}$$

First of all, using the indicator defined by the supplier, Equation 2, the number of defective measures has been obtained. There are 3219 wrong measures in the dataset. In other words, 9.16% of the measures are wrong.

$$\frac{GSRL9H100 - GSRL10H10}{GSRL9H100} \tag{2}$$

Taking this information into account the Bayes Theorem formula, Equation 1, has been filled using the following parameters:

- Likelihood: a binomial distribution, as the measure can be right or wrong, Figure 5.
- Prior knowledge: a beta distribution centered around 9.16%, Figure 6.

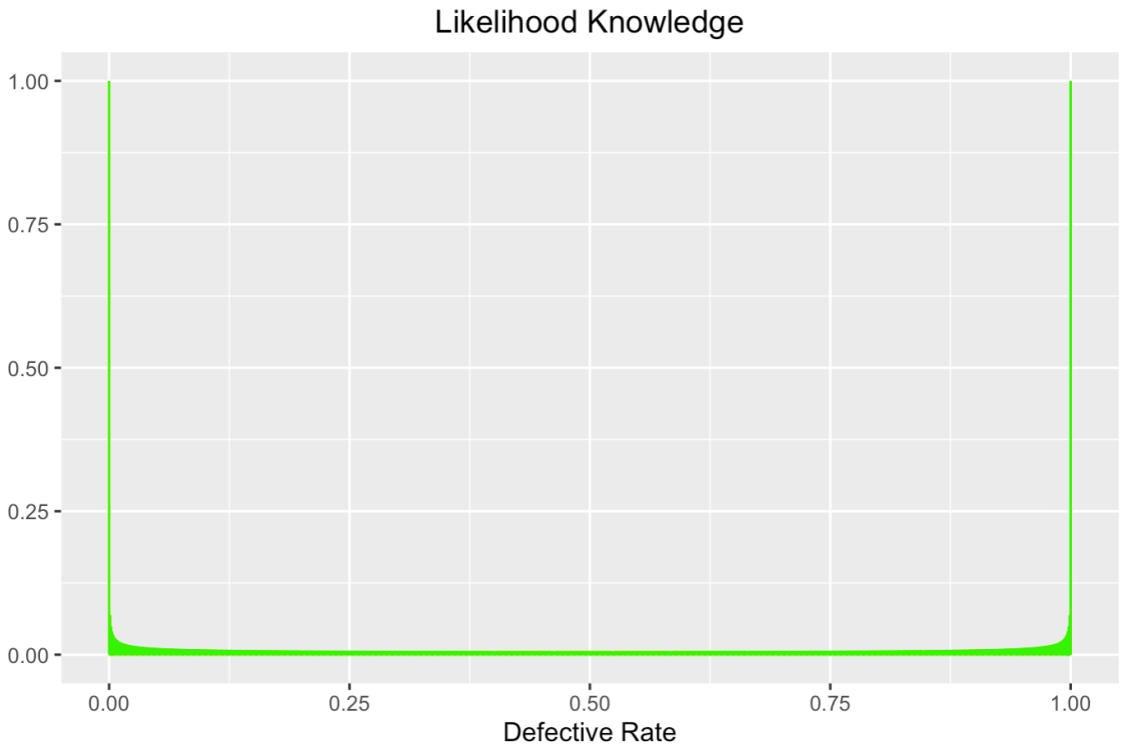


Figure 5: Likelihood knowledge

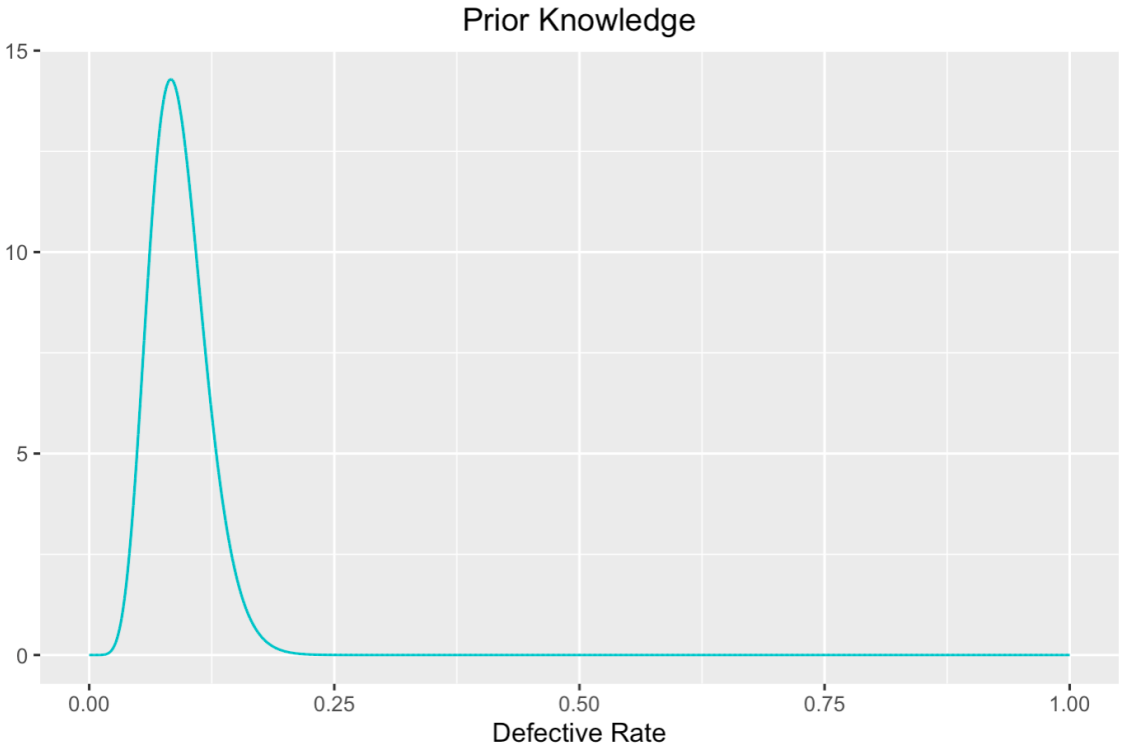


Figure 6: Prior knowledge

The Bayes inference gives the following probabilities for the defectiveness rate range of the vendor, Figure 7. As it can be seen in the image, the reliability of the vendor is, in most cases lower than 50%. It can be said with a confidence of 50% that the defectiveness rate of the measures is equal to 2%. For the other values of defectiveness rate indicated by the vendor, between 2% and 5%, the confidence is lower, being equal to 35%, approximately, at 5%. The decay in the confidences follows the exponential decay of the binomial distribution chosen as likelihood knowledge, Figure 5.

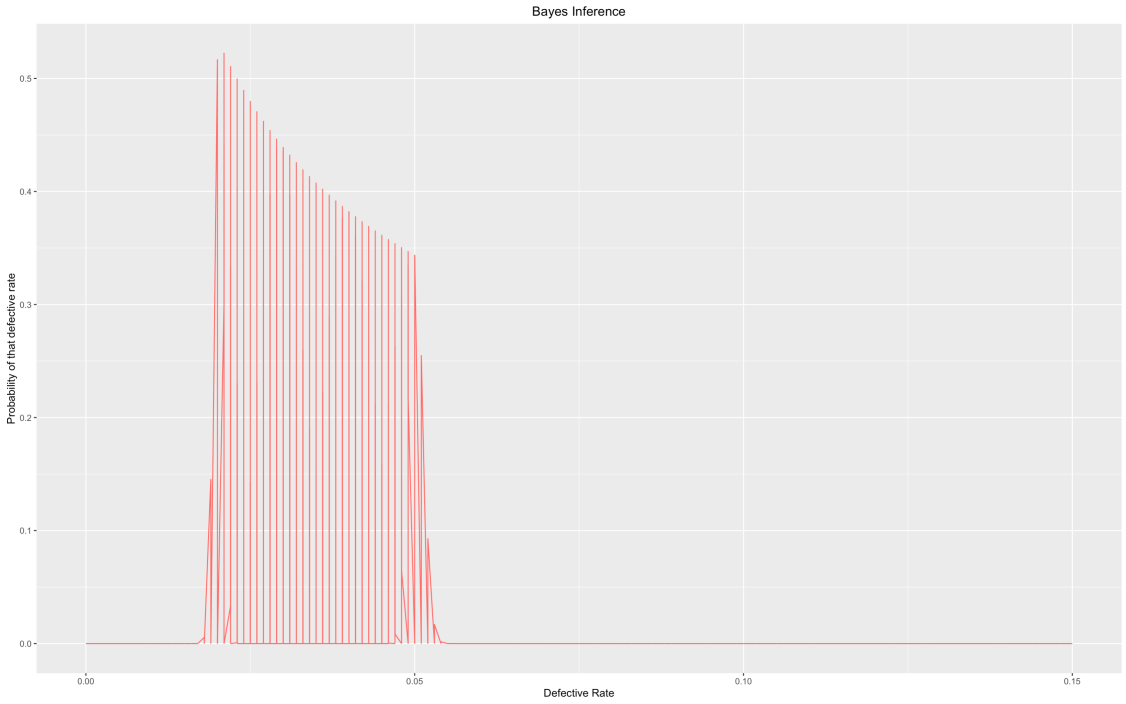


Figure 7: Bayes Inference

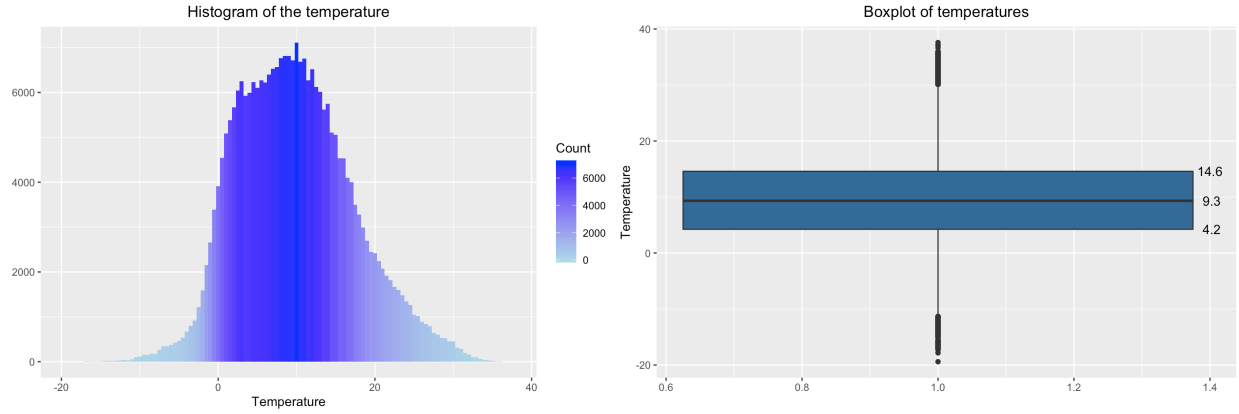
4 Problem 3

4.1 Exploratory Analysis

For this problem and the next one, the dataset used is an hourly dataset containing average values of all variables for each hour. After creating the new dataset, it is necessary to explore the data before doing the predictions requested by the client.

First, a description analysis of all the variables has been done in order to better understand the distributions they follow.

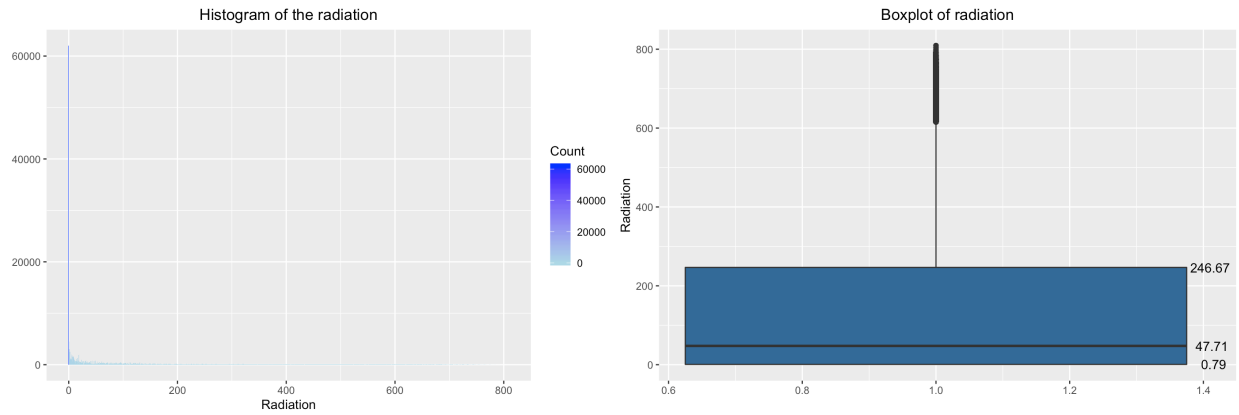
The temperature of all the sensors, presents an almost normal distribution centered around 10ºC, Figure 8. The usual values of the temperature fall between 4.2ºC and 14.6ºC. However, there are many outliers. Some sensors detect temperatures above 30ºC, while other detect temperatures below -10ºC. These outliers have been studied, and it can be said that all sensors detect temperatures above the usual values, but only half of the sensors detect temperatures below usual values. All sensor with altitude equal to 2, except the ones located at position 5 and 6, detect low temperatures. Furthermore, only the sensors located at position 4, 9 and 10, detect low temperatures in all their altitudes.



(a) Histogram (b) Boxplot

Figure 8: Temperature distribution

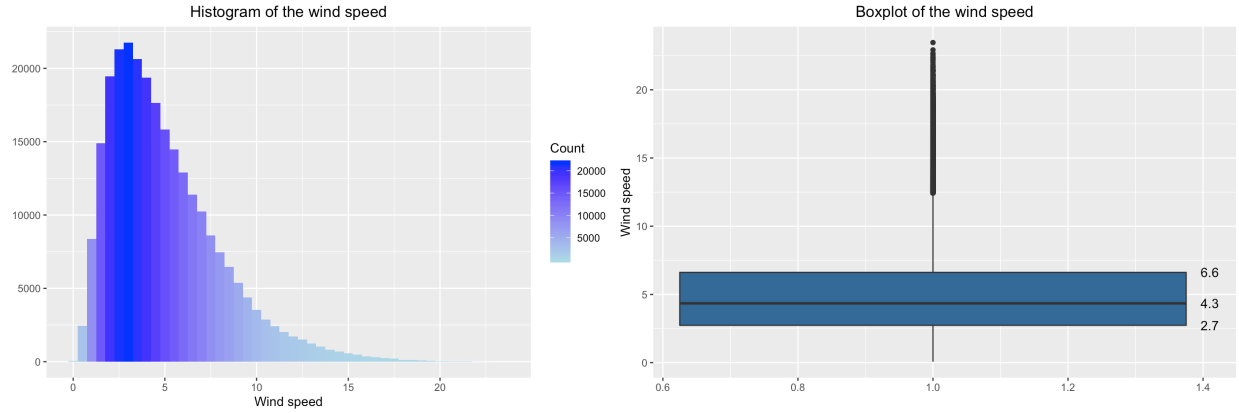
On the contrary, the solar radiation presents a positive skewed data distribution, Figure 9. The value with the highest occurrence is 0. The usual values of the solar radiation are between 0.79 and 246.67. In this case, there are also a lot of outliers, values above 600. These outliers have been studied and all the sensors detect at some point, very high values in radiation.



(a) Histogram (b) Boxplot

Figure 9: Solar radiation distribution

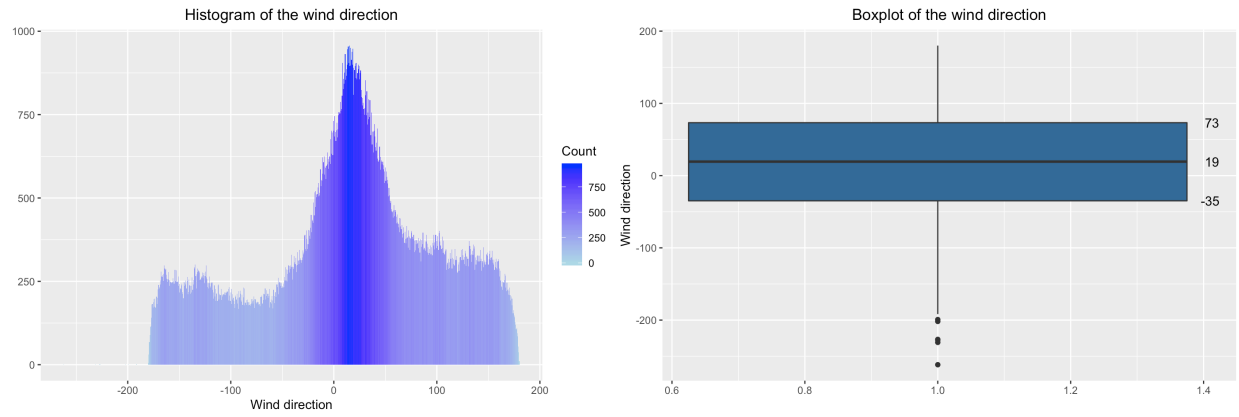
The speed of the wind also has a positive skewed data distribution, Figure 10. The value with the highest occurrence is 3. The usual values of the wind speed are between 2.7 and 6.6. In this case, there are a lot of outliers, values above 12.5. The outliers have been studied and all sensors detect high values except WSL1H2, WSL7H2 and WSL10H2.



(a) Histogram (b) Boxplot

Figure 10: Wind speed distribution

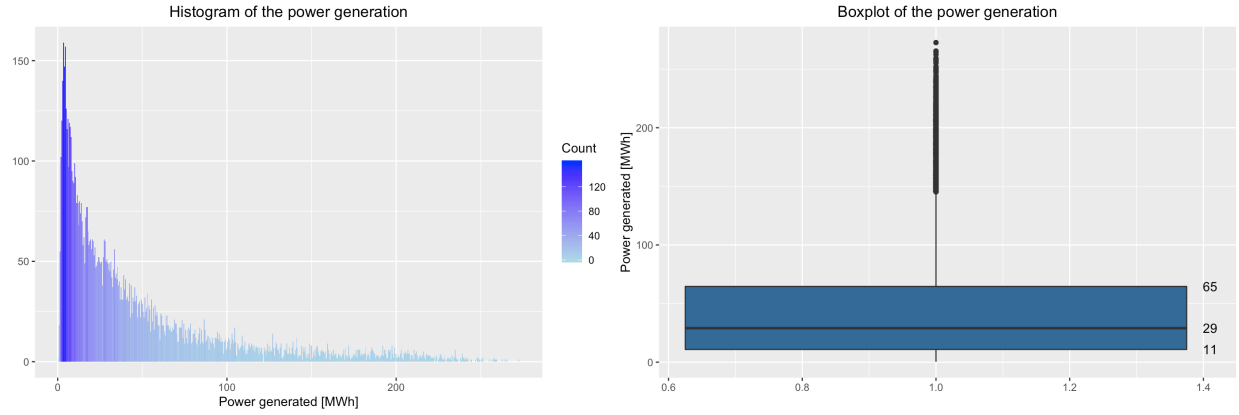
The direction of the wind detected by all the sensors, presents a rough normal distribution centered around 13.5º, Figure 11. The usual values of the wind direction fall between -35º and 73º, which makes sense, because the closer the direction of the wind is to 0º, the higher the performance of the wind mills. However, there are some values with a direction degree lower than -180º. This is impossible, because in the range [-180º,180º], all the possible directions are considered, so it does not make sense to have a measure outside that range. These measurements were taken by the sensors at location number 2, at all altitudes. When training the predictive model, this observations will be dismissed.



(a) Histogram (b) Boxplot

Figure 11: Wind direction distribution

The distribution of the power generated follows a logarithmic distribution, Figure 12. The usual values of the power generated fall between 11MWh and 65MWh. However, there are a lot of outliers, values above 145.26MWh. These outliers have been studied and they represents 7.58% of all observations.



(a) Histogram (b) Boxplot

Figure 12: Power generation distribution

Now that we have a preliminary image of how all the different variables behave on their own, it is time to find relationships between them. This is very important because the entries of the predictive model that will be developed need to be as uncorrelated as possible in order to assure the correct behavior of the model.

First, lets take a look at how the different measures (temperature, solar radiation, wind speed and wind direction) are affected by the location and the altitude of the sensors.

All the variables of temperature have a very high correlation among them, over 0.9, Figure 13. Furthermore, there is an even higher correlation between temperatures measured

at the same altitude. This observation, is also valid for measures that are taken in the same location. In other words, with only one measure of temperature, taken from any location at any altitude, we could be able to obtain the temperature at any other location and altitude.

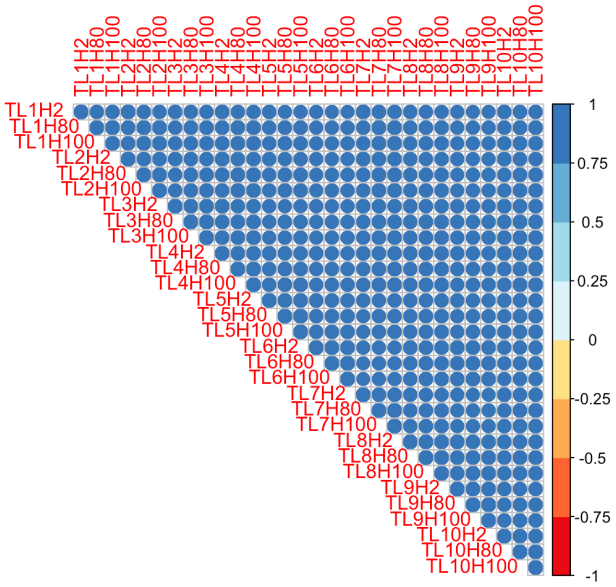


Figure 13: Correlation between temperatures

All the variables that represent values of solar radiation, have a very high correlation among them, over 0.9, Figure 14. Moreover, there is the exact same correlation with any point of one particular location at any altitude. In other words, all the measures taken in one specific moment of time in one location are the same regardless of the altitude of the sensor, therefore, the altitude of the sensor does not affect, at all, the value of the solar radiation.

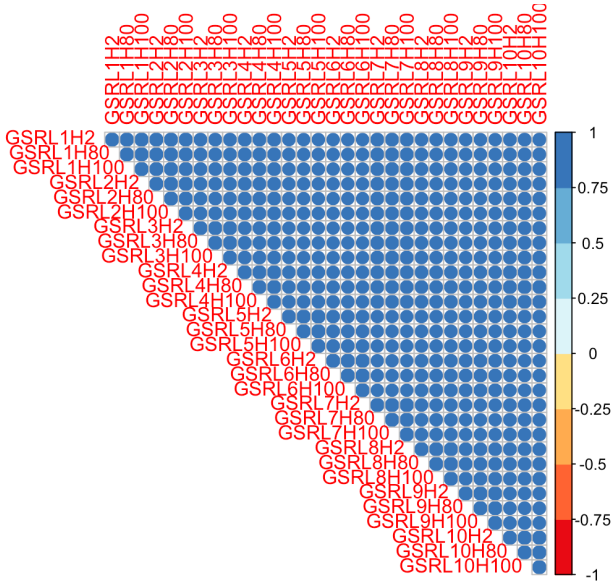


Figure 14: Correlation between solar radiation values

The speed of the wind, is very correlated among sensor at the same location, therefore, the altitude does not have a significant impact in the measure, Figure 15. However, measures taken at different location do not have a strong correlation. In other words, the speed of the wind depends on the location of the sensor, but does not depend on the altitude.

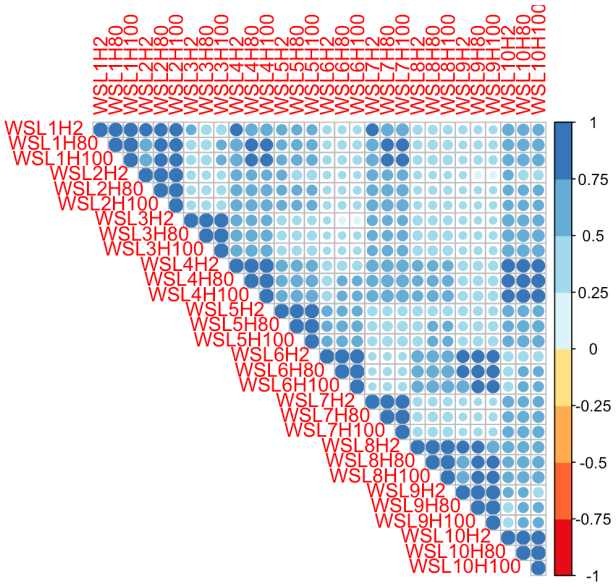


Figure 15: Correlation between wind speed values

The correlations between wind directions, behave like correlations between wind speeds. The direction of the wind depends on the location of the sensor, but does not depend on the altitude, Figure 16.

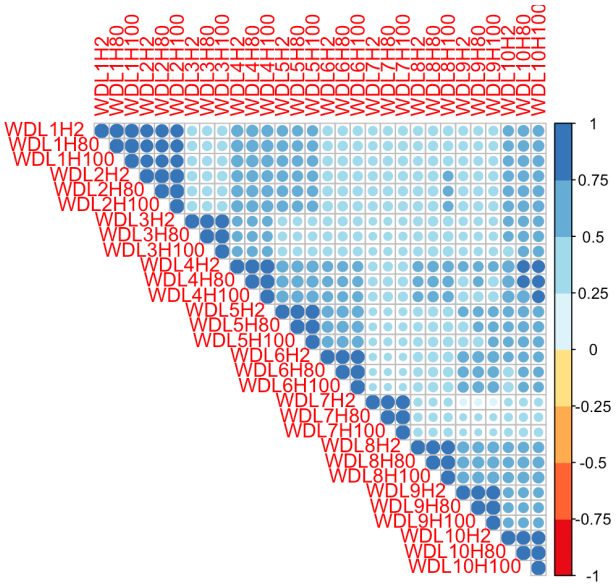


Figure 16: Correlation between wind direction values

In addition, it has been studied the correlation between the temperature, the solar radiation and the speed and the direction of the wind, Figure 17. No significant correlations have been found among variables of different nature, this is very important, as the input variables of the predictive model should be as independent as possible.

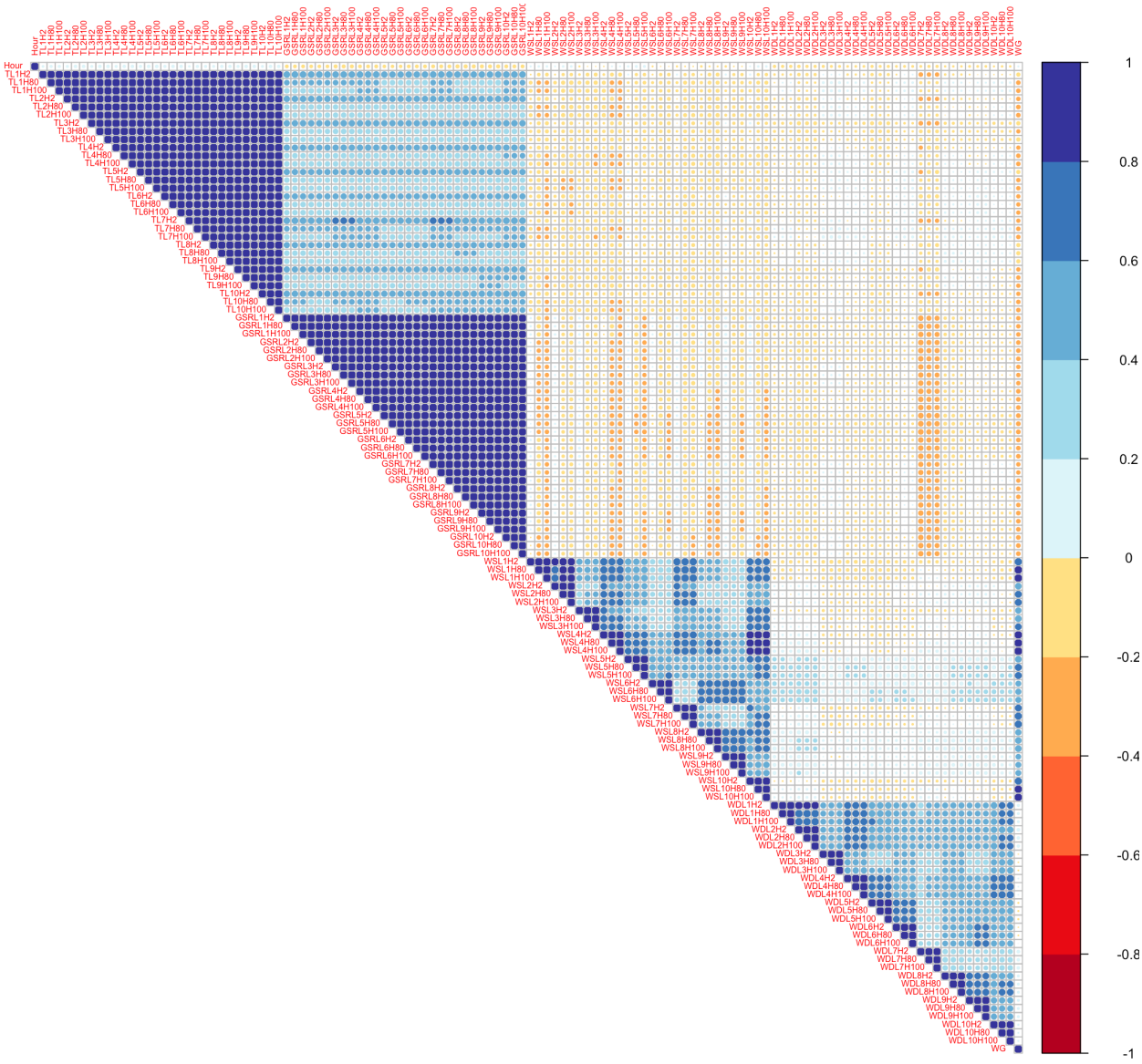


Figure 17: Correlation of all variables

In the image above, Figure 17, it can be seen that some variables have a high correlation with the output variable (WG). These values of those correlations are the following, Table 2:

Variable	Correlation
WSL4H100	0.88
WSL4H80	0.88
WSL4H2	0.83
WSL1H80	0.82
WSL1H100	0.82
WSL10H100	0.81
WSL10H80	0.81

Table 2: Highest correlationa with the ouput variable

4.2 Predictive Model

The first thing that needs to be done is cleaning the data. During the exploratory analysis, it was found that some observations had a wind direction degree lower than -180º, which is impossible taken into consideration the range of measures handled by that kind of variables ([-180º,180º]]). No more incorrect data was found, and it has been decided to keep the outlier values during the training of the model but keeping into account its presence while evaluating the model.

In this dataset, there are a multitude of variables which values differ in orders of magnitude, therefore, in order to build the model, the data has been normalized, dividing each variable by its maximum value. Afterwards, the normalized dataset has been divided into a training set and a testing set.

With all the data ready to be used as inputs of the model, it is necessary to determine which variables will be used. From the exploratory analysis, a lot of useful information was found regarding which variables to select:

- The temperature does not depend on location nor altitude, therefore, with only one measure the temperature in any other location and altitude can be inferred.
- The solar radiation does not depend on location nor altitude, therefore, with only one measure the solar radiation in any other location and altitude can be inferred.
- The speed of the wind only depends on the location, not on the altitude.
- The direction of the wind only depends on the location, not on the altitude.

Taken the previous information into account, it has been decided that the first set of variables will be composed of:

- One temperature variable

– One solar radiation variable

– One wind speed variable

– One wind direction variable

– The hour of the day

During the exploratory analysis, Table 2, it was found that WSL4H100 is the wind speed variable that has the highest correlation with WG, therefore this will be the wind speed variable used. Moreover, the altitude of the sensor is H100, as none of the variables depend on the altitude, this would be the altitude used as reference.

In order to select the temperature variable and the solar radiation variable, a method called recursive variable selection has been used. As a result, it has been determined that the temperature variable that will be used is TL3H100 and the solar radiation variable that will be used is GSRL5H100.

The first model that has been trained is a linear regression, in order know if any of these variables can de modeled using a linear expression, or they cannot, which it would mean the need for a more complex model.

$$WG \sim Hour + TL3H100 + GSRL5H100 + WSL4H100 + WDL7H100$$

The residuals plot obtained is the following, Figure 18.

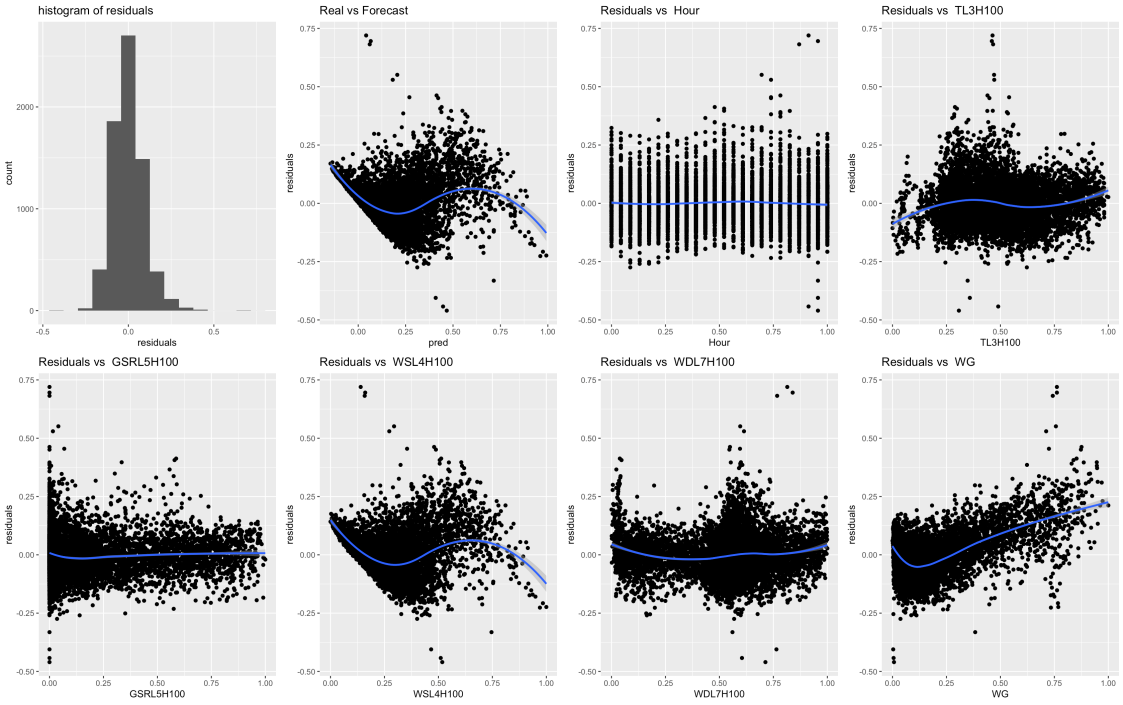


Figure 18: Residuals of the linear regression

Ideally, the residuals should be equally distributed around zero and throughout the different values taken by the variable. In the previous image, Figure 18, it can be seen that only GSRL5H100 is correctly modeled by the linear regression, therefore it is recommended to move onto more complex methods that support non-linear relationships between variables. After trying different models, it has been decided that the algorithm that best fits the dataset is neural networks.

Using the same input variables as the ones used in the linear regression, the neural network obtains the following residuals, Figure 19.

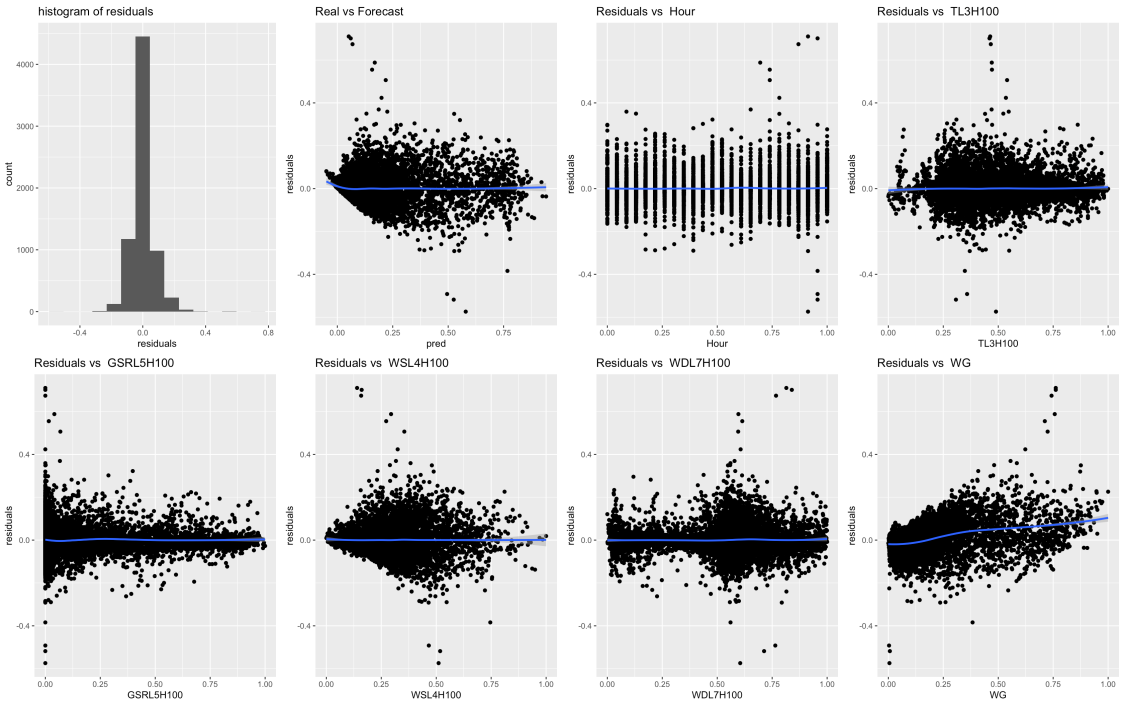


Figure 19: Residuals of MLP_1

In the image, Figure 19, it can be seen that the neural network better models all the variables, especially TL3H100 and WSL4H100. In order to determine, which are the most important variables of this model, a sensitivity analysis is done, Figure 20. The sensitivity analysis shows that the most important variable is WSL4H100, followed by GSRL5H100. The Hour and WDL7H100 have a lower importance, being TL3H100 the less important.

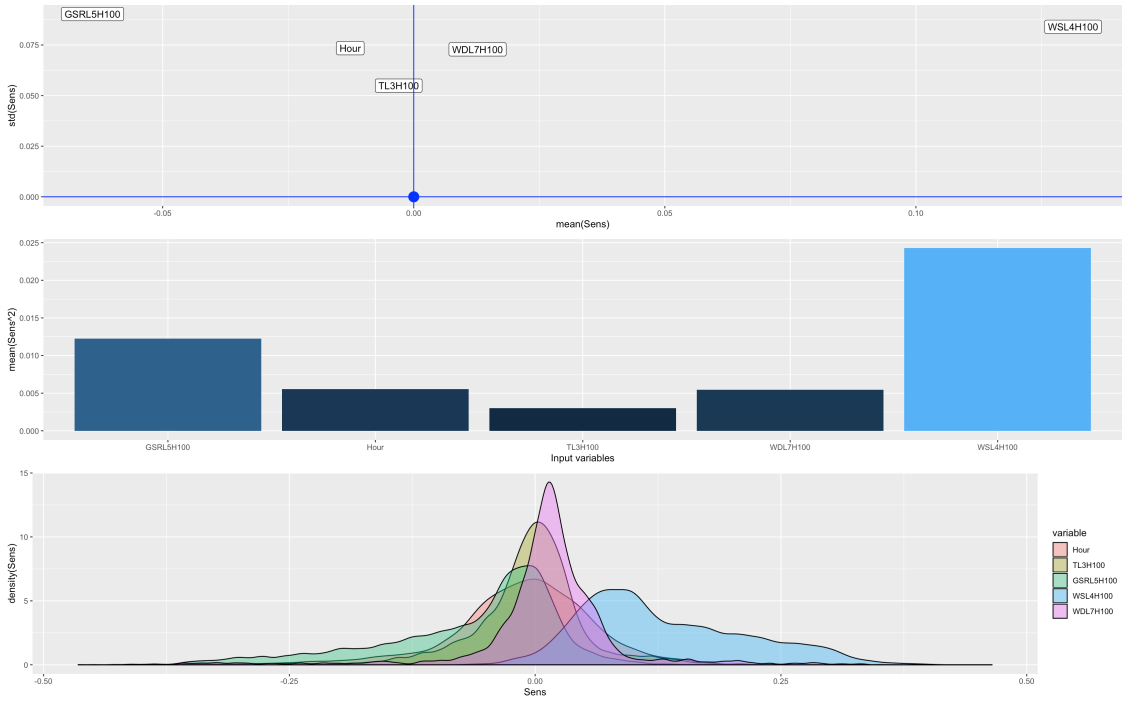


Figure 20: Sensitivity Analysis of MLP_1

Taking the previous information into account that the three most important variables represent wind speed, wind direction and solar radiation, it has been decided to create a model using GSRL5H100, all the wind speed variables and all the wind direction variables, for the location H100. All the solar radiation variables have not been used due to the high correlation of the solar radiation measures regardless of the location and the altitude.

$$WG \sim WSL4H100 + WSL1H100 + WSL2H100 + WSL3H100 + WSL5H100 + WSL6H100 + WSL7H100 + WSL8H100 + WSL9H100 + WSL10H100 + GSRL5H100 + WDL1H100 + WDL2H100 + WDL3H100 + WDL4H100 + WDL5H100 + WDL6H100 + WDL7H100 + WDL8H100 + WDL9H100 + WDL10H100$$

After building the model with the previous inputs, the sensitivity analysis shown in Figure 21, is obtained. It can be seen that the most important variables are WSL4H100, WSL1H100 and WDL2H100.

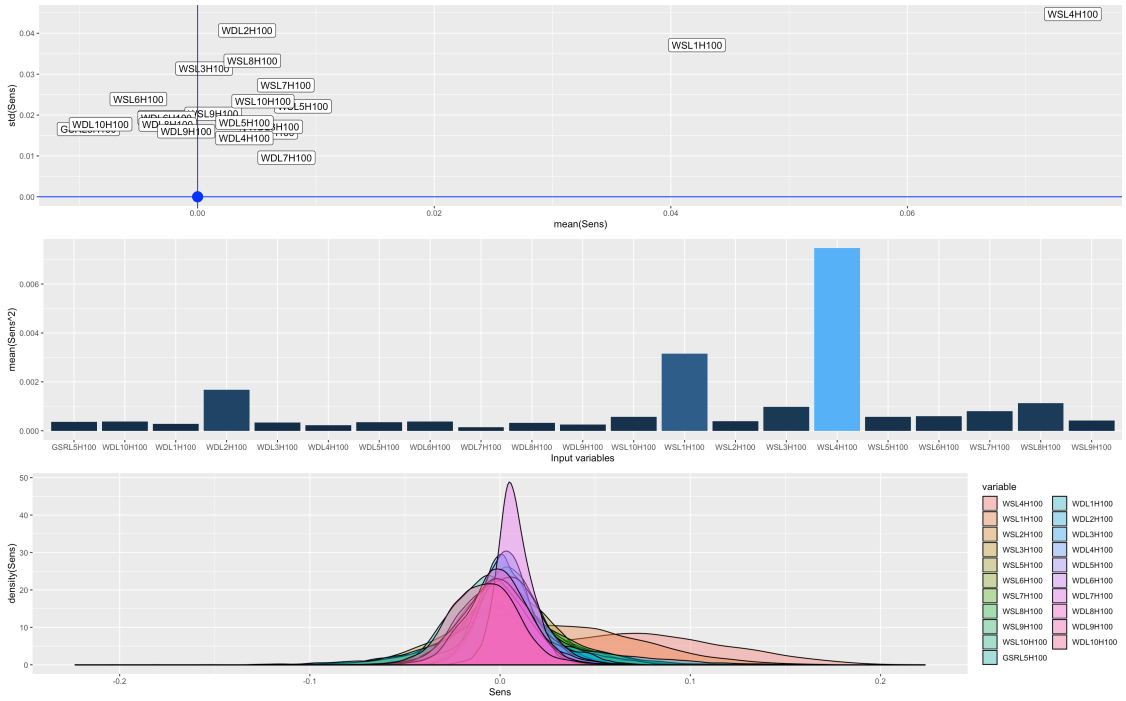


Figure 21: Sensitivity Analysis of MLP_2

Considering all the information gathered throughout the exploratory analysis and the fitting of models, the final model has been developed using the following equation:

$$WG \sim WSL4H100 + WSL1H100 + WDL2H100 + GSRL5H100$$

The following image, Figure 22, shows the sensitivity analysis of the final model. The analysis shows that the most important variable is WSL4H100. Moreover, the least important variables is GSRL5H100, therefore it can be said that in this particular model, the prediction of the power generated will be driven by wind-related characteristics.

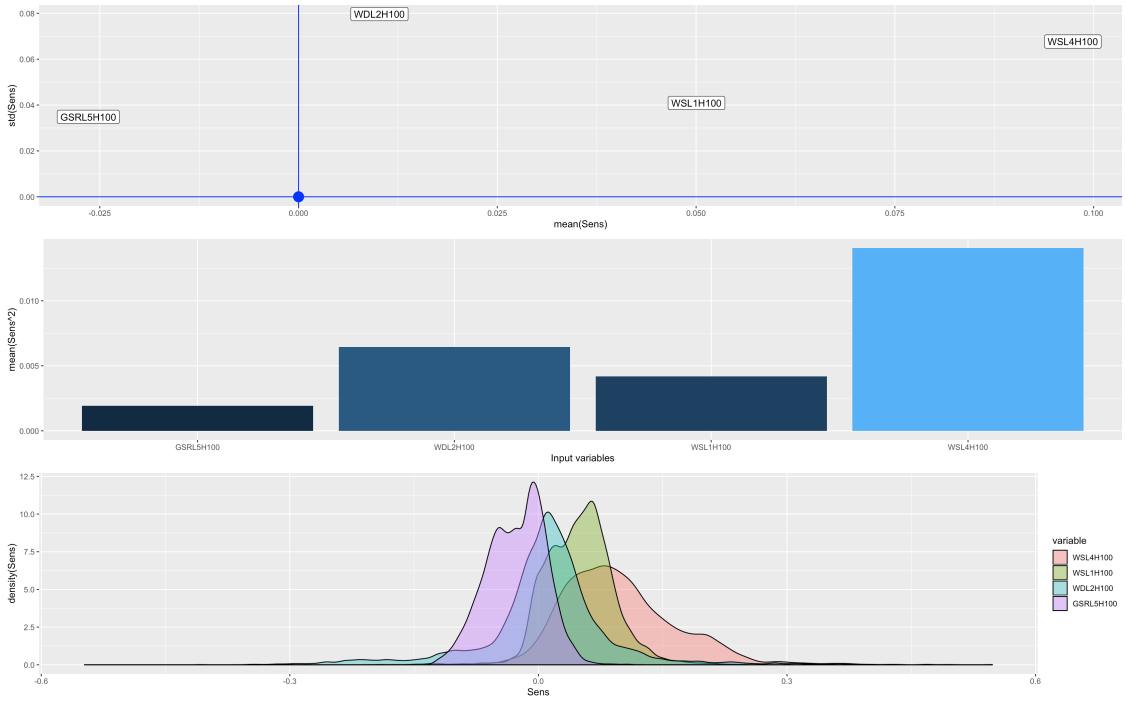


Figure 22: Statistics of the Final Model

The model created has the residuals presented in the image, Figure 23. The residuals show that all the input variables are very well modeled by the neural network, the residuals are distributed around zero and throughout all the values taken by the different variables. Some residuals are completely away from the main distribution, those points are due to the presence of outliers in the dataset. The output variable (WG), is pretty well fitted, up until 0.5, from that point onward the residuals keep growing. Ideally, the level of the residuals should be constant and equally distanced from zero for all values. This model does not show that behavior in WG. However, this is not something to be very concerned about as if 0.5 is denormalized, it converts to 136.325 MWh, which is very close to the maximum usual value, Figure 12.

Apart from the good shape of the residuals, this model has been chosen because it has one of the lowest Root Mean Squared Errors (RMSE) found while maintaining a small differences between the training and the testing values, and therefore avoiding overfitting. The following table shows the values of the error, Table 3. Taking into account the residuals, Figure 23, and the values of the error, Table 3 it can be said that the model has a good performance.

RMSE	Normalized	Denormalized
Train	0.06384	17.38009
Test	0.06816	18.55730

Table 3: Error measures of the final model

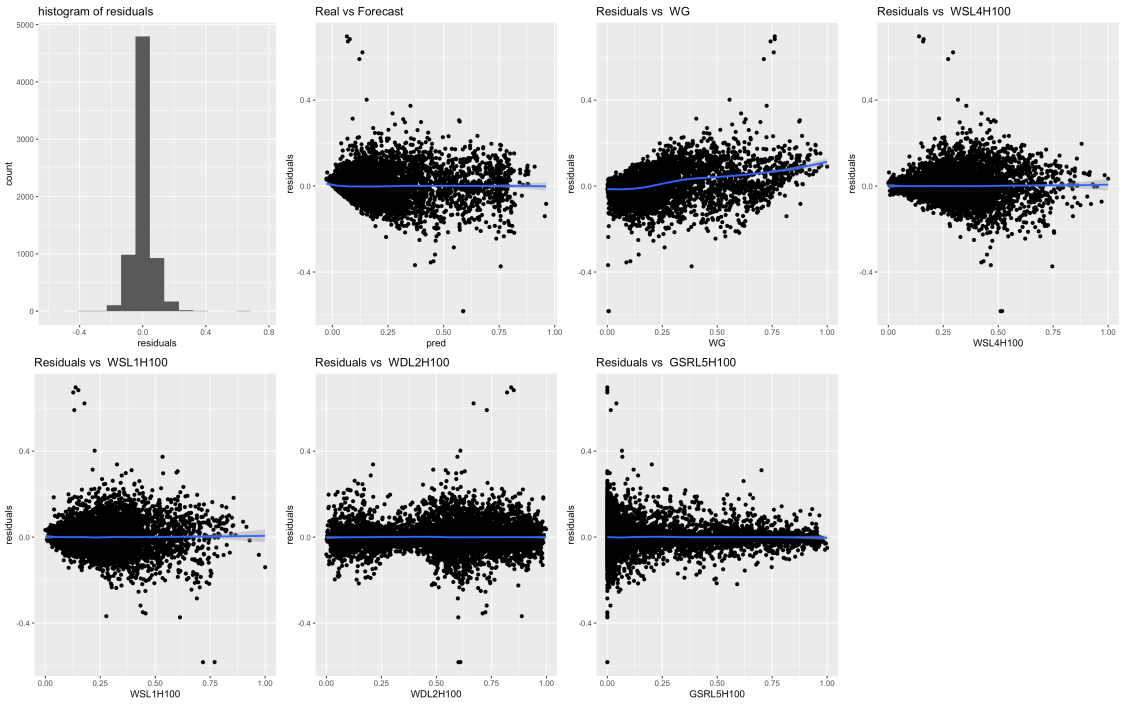


Figure 23: Residuals Final Model

4.3 Analysis of the residuals

The wind power generated at the station can be considered as a time series, therefore this problem is a dynamic regression problem, as the output also depends on explanatory variables.

The previously obtained regression model has been used to predict the power generation of the whole dataset, not just the hourly dataset. The residuals of the model have been calculated using:

$$residuals = WG - estimated \quad WG$$

When plotting the time series of the residuals the following image is obtained, Figure 24. The residuals can be considered as stationary, even though some outliers can be seen, especially at the end of the time series.

If the residuals cannot be considered as white noise, it means that they can be modeled using an ARIMA model. Therefore, the next step should be to determine whether the residuals can or cannot be considered white noise. In order to find out, the autocorrelation function (ACF) and the partial autocorrelation function (PACF) have been plotted , Figure 25. The horizontal blue lines represents the confidence interval, bot the ACF and the PACF have values that exceed the interval, therefore, the residuals cannot be considered as white noise and need to be modeled using an ARIMA model.

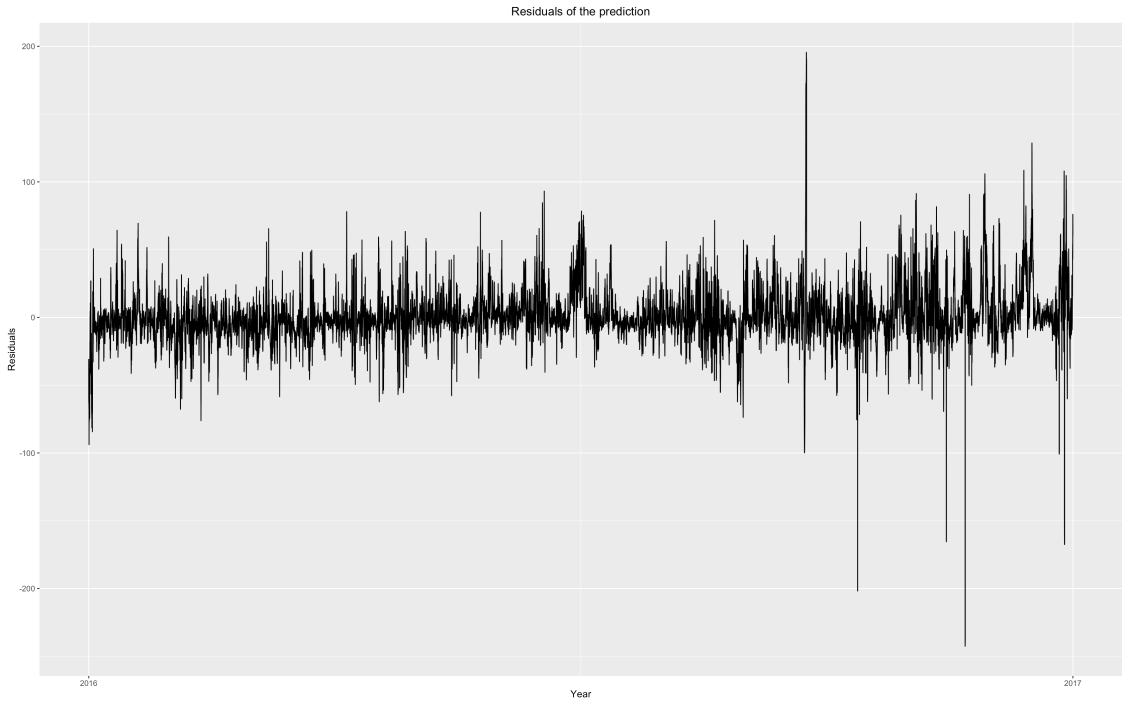


Figure 24: Residuals of the prediction

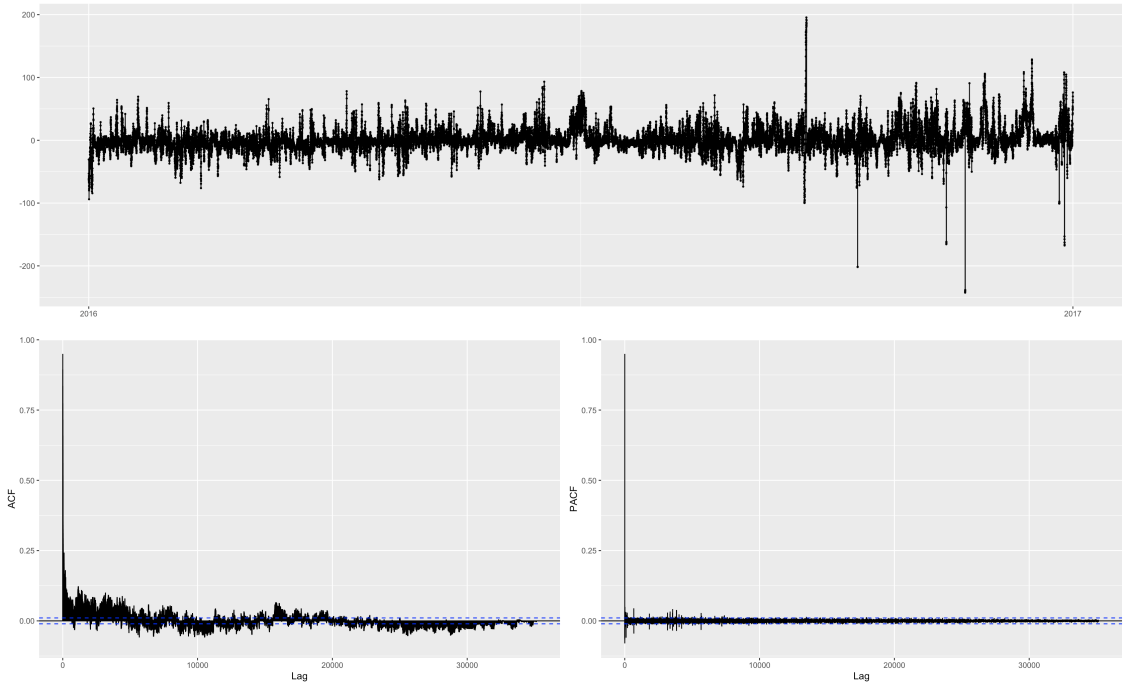


Figure 25: ACF and PACF plots of the residuals

In Figure 25, can be seen that the PACF has one predominant component, therefore the first estimation is done using an ARIMA(1,0,0), Figure 26.

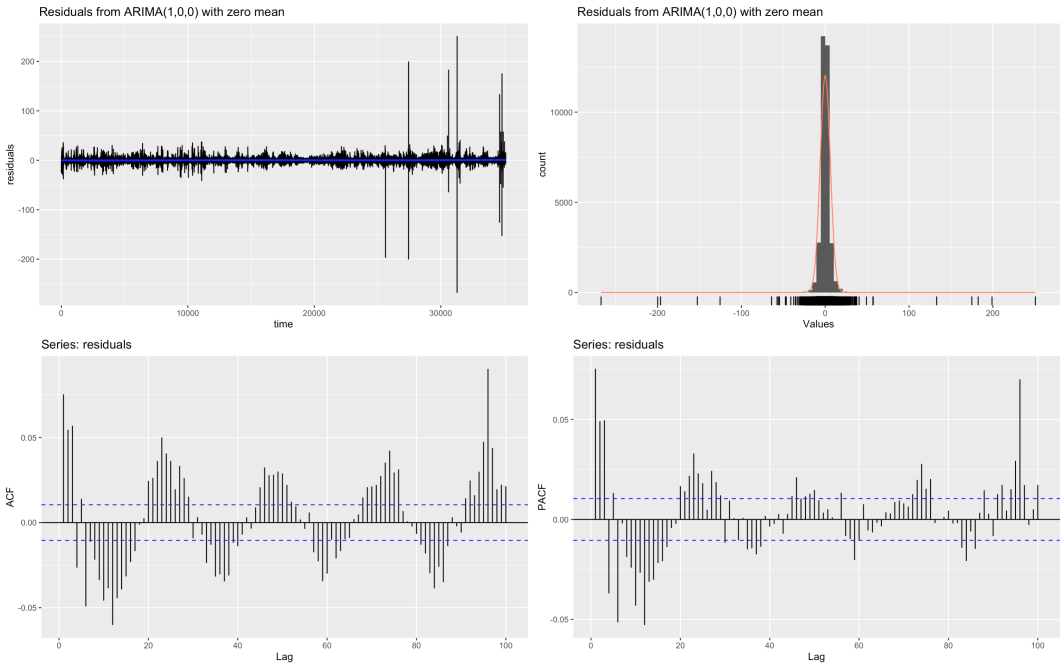


Figure 26: ARIMA(1,0,0)

Looking at the previous image, it can be seen that the residuals present a seasonal component with a period equal to 24. Taking into account only the values at positions that are multiple of 24, it can be inferred that the decay present in the ACF, as well as in the PACF is very small, therefore the residuals need differentiation in the seasonal component. The next model implemented is an ARIMA(1,0,0)(0,1,0), Figure 27.

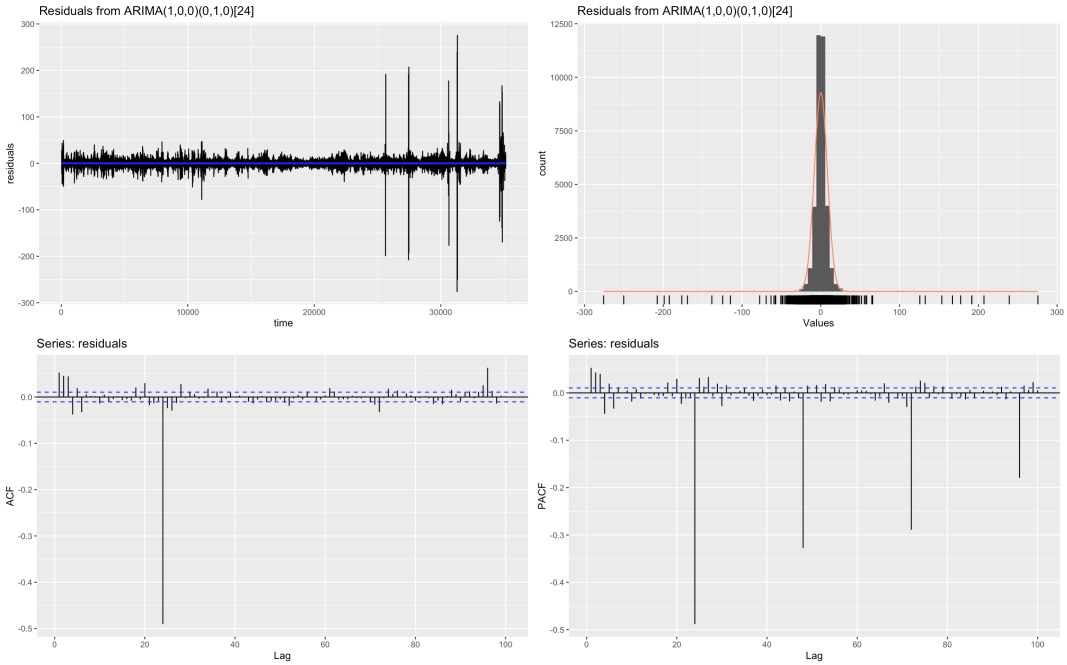


Figure 27: ARIMA(1,0,0)(0,1,0)

That model has a clear main component in ACF, so we add that information to the model, ARIMA (1,0,0)(0,10), Figure 28.

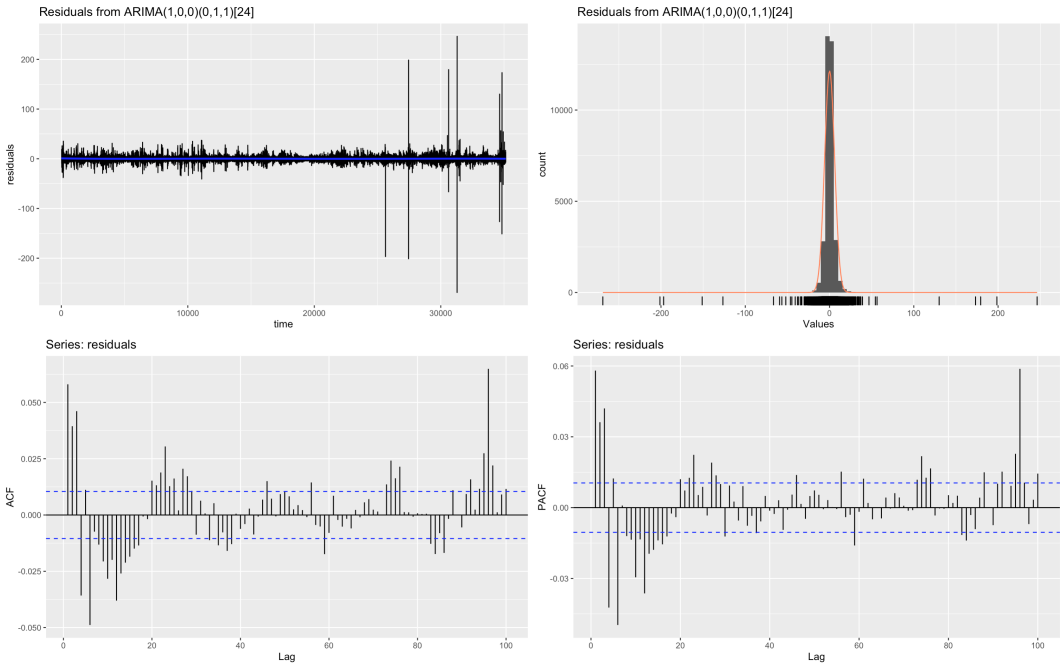


Figure 28: ARIMA(1,0,0)(0,1,1)

After various iterations of studying the time series and the seasonal component separately, the model found that best fitted the residuals is an ARIMA(3,0,1)(1,1,2)[24], Figure 29.

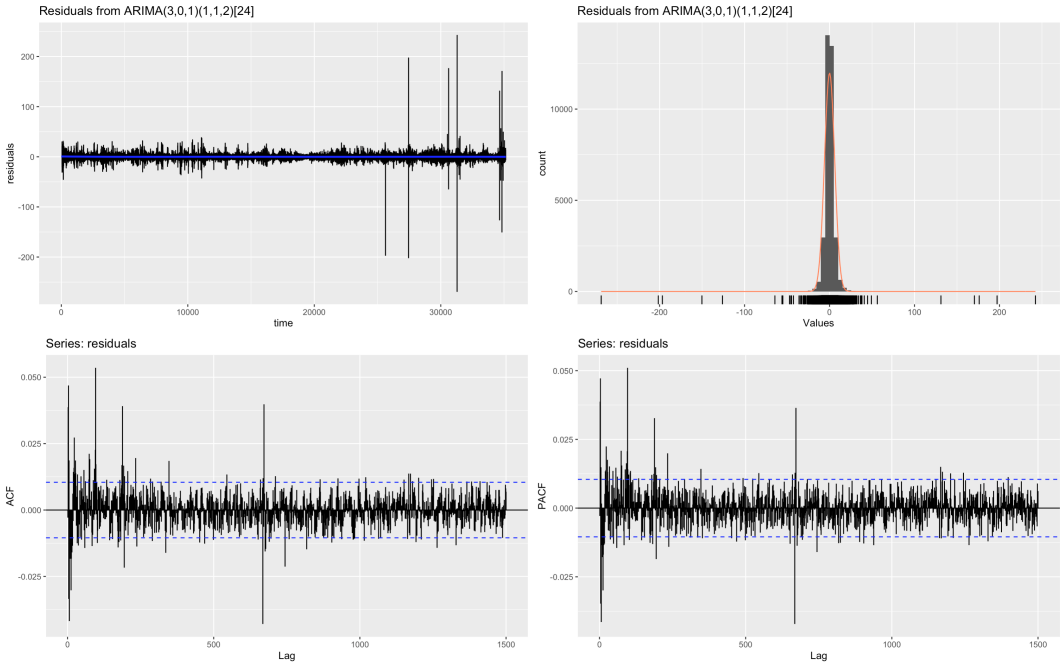


Figure 29: ARIMA(3,0,1)(1,1,2)

Most of the values are inside the white noise confidence interval, which is what it is desired. However, some of the ACF and PACF values are outside the confidence interval, those values probably correspond to the outliers presents in the dataset. Furthermore, the distribution of the residuals follows a normal distribution with a mean around zero and a small standard deviation, which is optimal.

The following image, Figure 30, shows the real values in blue versus the fitted values in red. It can be seen that the model fits the time series almost perfectly. The greatest difference between the fitted and the real values is found in the peaks.

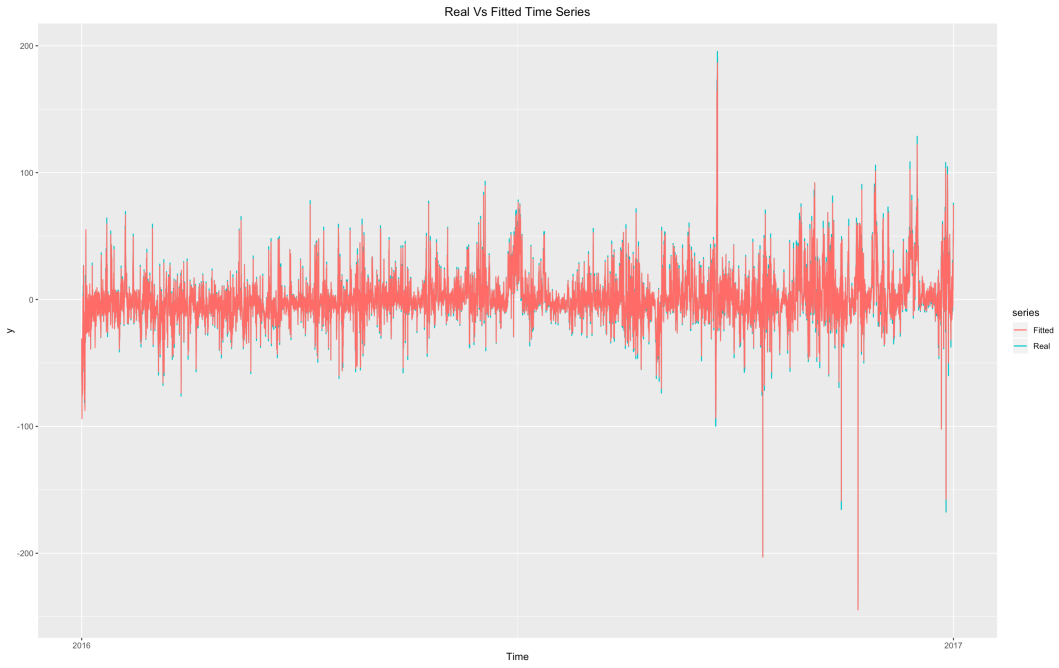


Figure 30: Real Vs Fitted Residuals

Furthermore, the prediction for next week has been calculated, Figure **??**. In that image it can be seen that the values corresponding to the first couple of days are predicted. However, after the second day, approximately, the prediction is equal for the following days and equal to the average values of the time series. This means that the prediction cannot be more exact that far ahead, so it supposes the values taken by the time series will be close to the average values.

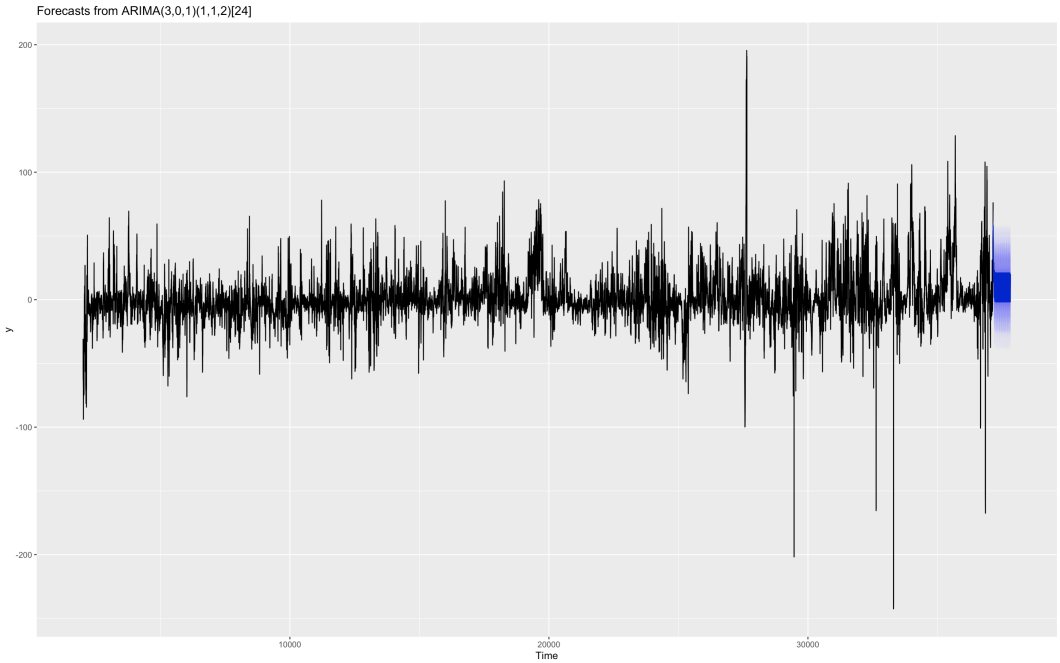


Figure 31: Next week's forecast

5 Problem 4

The company is also interested in knowing when the plant will reach a minimum generation. Specifically, the company wants to know when the wind generation take values over 10 MWh. In order to answer this question, a classification model needs to be implemented.

Knowing that the minimum power generation required is 10 MWh, a new variable has been created to determine if an observation meets the requirement or not. This new variable is called "MinimumGeneration" and will be the one predicted.

While developing the best classification model, previous information can be used to determine which variables are the most relevant. For instance, it is important to remember that all the temperature variables are very correlated, so with only one the rest can be inferred accurately, the exact same thing happens to the solar radiation. Furthermore, none of the variables depend on the height of the sensor, so the dataset has been trimmed, only the measures taken at height 100 will be taken into consideration.

When creating a classification model is very important that the input variables are uncorrelated. From the exploratory analysis, is known that variables that measure different characteristics are uncorrelated. However, the correlation between the different variables and the new output variable needs to be studied.

In the image, Figure 32, it can be seen that during the hours of the day, from 5:00 until 17:00, the number of observations that do not meet the minimum power requirement is greater than the number of observation that meet the condition. Moreover, from August until December, the requirement is most-likely met.

In Figure 33, it can bee seen that once again, the variable with the highest correlation with the output variable in WSL4H100. Furthermore, from that image it can be inferred that when the wind speed is between 0 and 10 the number of observations that do not reach the minimum power is greater.

Taking into account only the wind direction variables, Figure 34, WDL2H100 has the highest correlation with the minimum power requirement. Furthermore, the distribution of observations that meet the minimum and the number of observations that do not is very similar for every wind direction, except in WL2H100. Moreover, at location number 2 it can be seen that when the direction of the wind is between -180º and 0º, the number of observations that do not meet the minimum is significantly higher.

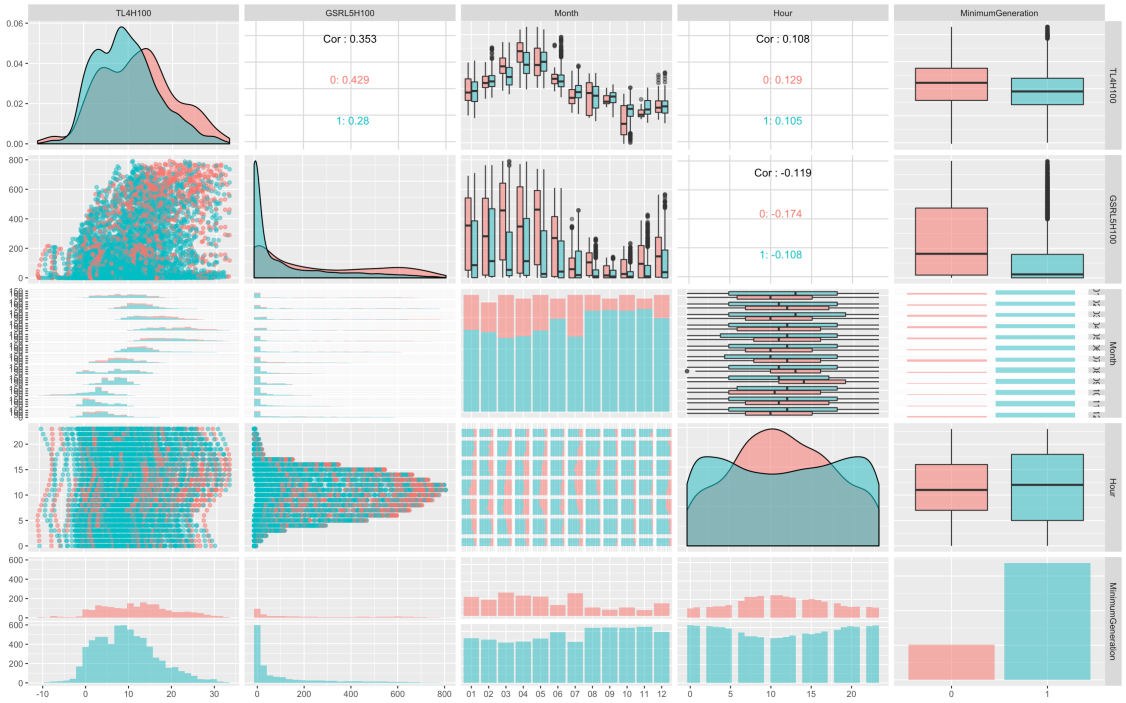


Figure 32: Correlations with Minimum Generation, part 1

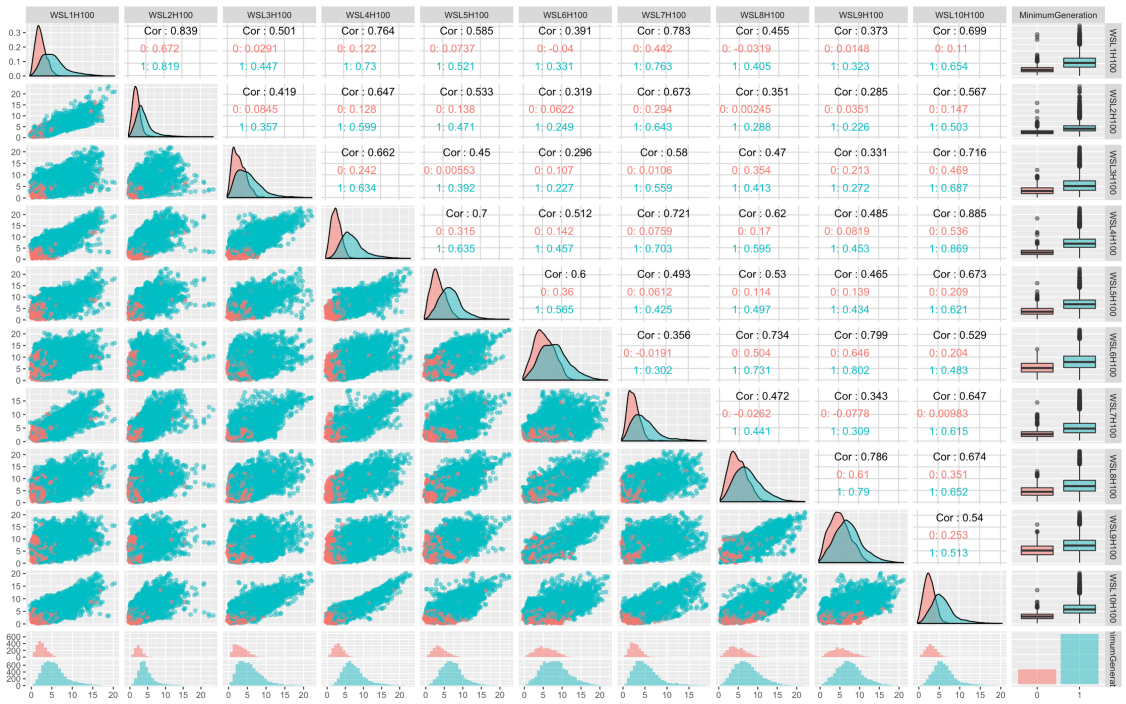


Figure 33: Correlations with Minimum Generation, part 2

Figure 34: Correlations with Minimum Generation, part 3

While developing a classification model is necessary to keep in mind the classes distribution in the dataset, as a very unbalanced dataset will bias the classification. In this case, the data in very unbalanced, only 23,10% of the observations do not meet the minimum power generation requirement. Therefore, in order to avoid biases while training the model, the training data has been upsampled, so that the proportion of each class is equal to 50%.

The first classification model tried, is a logarithmic regression using all the variables of the trimmed, as described before, dataset. The performance measures obtained using this model are the following, Table 4.

Measure	Train	Test
Accuracy	0.9049	0.8979
AUC	0.9625	0.9622

Table 4: Performance measure of the first logarithmic regression

The fact that the previous values are very similar indicates that the trained model has not incurred in overfitting. However, this model takes into account 24 variables and not all variables have the same importance. The logarithmic regression model indicates the following p-value levels for the input variables, Table 5.

Variable	P-value
TL4H100	$8.36e-10$
GSRL5H100	$1.30e-09$
WSL1H100	$<2e-16$
WSL2H100	0.82652
WSL3H100	$1.24e-15$
WSL4H100	$<2e-16$
WSL5H100	$<2e-16$
WSL6H100	0.40841
WSL7H100	$9.39e-05$
WSL8H100	0.31355
WSL9H100	0.33139
WSL10H100	0.05774
WDL1H100	0.18819
WDL2H100	0.43210
WDL3H100	$2.73e-08$
WDL4H100	0.53211
WDL5H100	0.14640
WDL6H100	0.28028
WDL7H100	0.00015
WDL8H100	0.48883
WDL9H100	0.69474
WDL10H100	0.01218
Month	$<2e-16$
Hour	$1.10e-05$

Table 5: P-values for the input variables of the first logarithmic regression

In an attempt to improve the model, a new logarithmic regression has been trained using only the most significant variables shown in Table 5, that is, the ones with a p-values lower than $9,39 * 10^{-5}$. The equation of the model is the following:

$$MinimumGeneration \sim TL4H100 + GSRL5H100 + WSL1H100 + WSL3H100 + WSL4H100 + WSL5H100 + WSL7H100 + WDL3H100 + WDL7H100 + Month + Hour$$

This model has the following performance measures, Table 6.

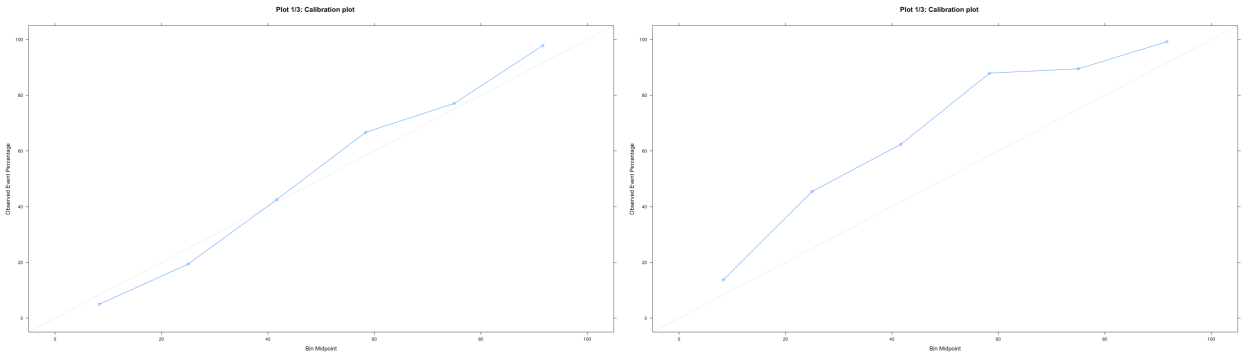
Measure	Train	Test
Accuracy	0.9036	0.9030
AUC	0.9623	0.9618

Table 6: Performance measure of the final logarithmic regression

These performance values are even more similar than the previous, the AUC diffenrence has increase in 0.002, but the accuracy difference has decreased in 0.0064. Therefore, it can

be said that this model generalizes better.

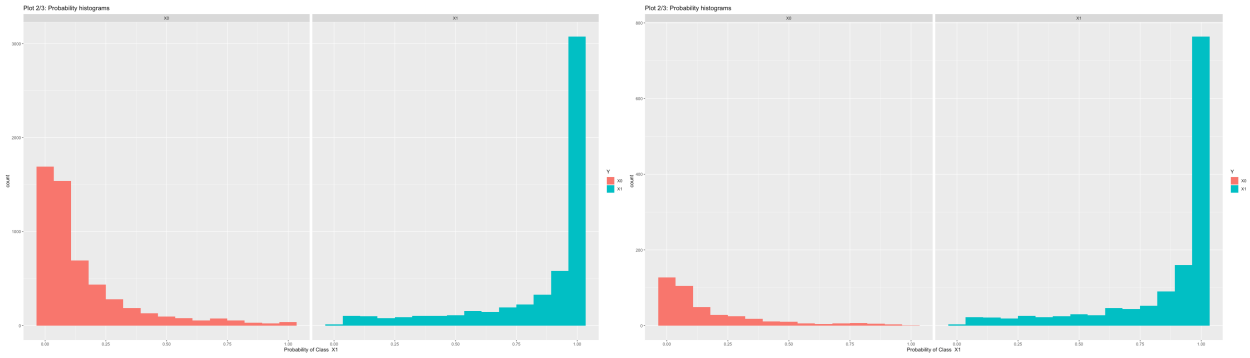
In Figure 35, it can be seen that in the calibration plots, specially in the training plot, the plot is close to y=x, which means that the model obtained fits the model pretty well. The testing calibration plot has a similar shape, but a little bit shifted to the left. This means that the model underpredicts.



(a) Training Set (b) Testing Set

Figure 35: Calibration plot

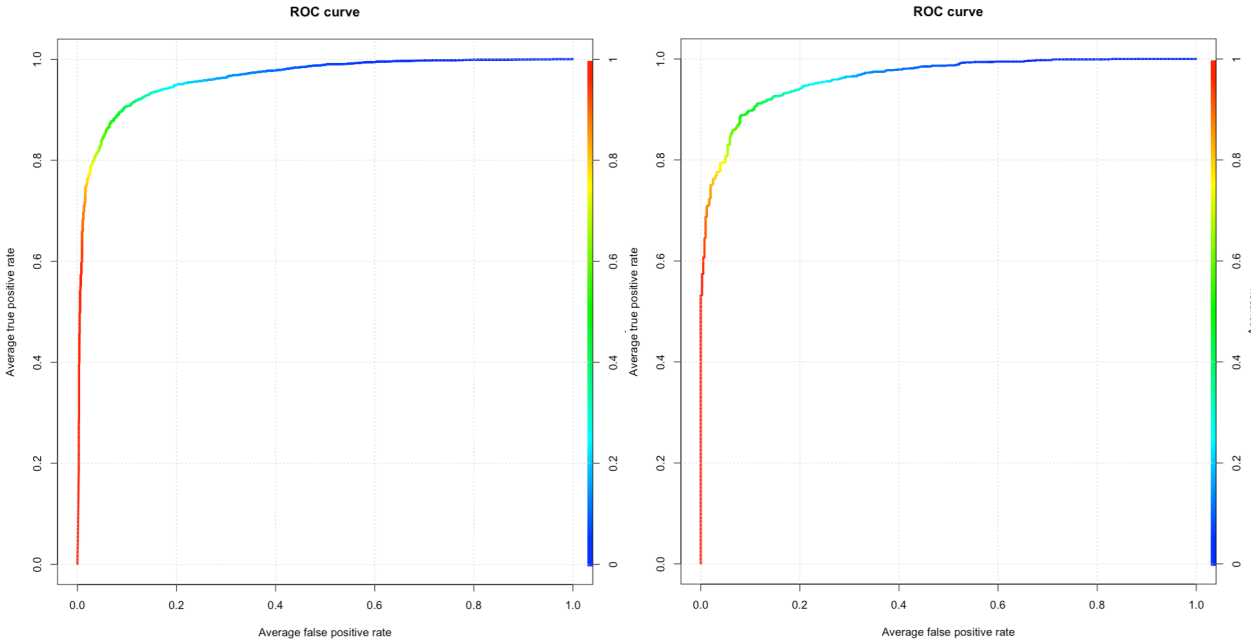
Moreover, the distribution of probability of meeting the minimum power, Figure 36, shows that in most cases the input will be classified correctly, when P(X1)=1, P(X0)=0, and vice versa.



(a) Training Set (b) Testing Set

Figure 36: Probability distribution

The AUC is 0.9623 for the training set and 0.9618 for the validation set, this means that the model is well fitted, as there is no significant difference between the values of both sets. Furthermore, the fact that the values are very close to 1, indicate that the model is working as a good predictor of the MinimumGeneration variable.The following images, Figure 37, show the practically identical ROC plots of the training ant the testing sets.



(a) Training Set (b) Testing Set

Figure 37: ROC

More classification model have been trained using different algorithms, such as K-Nearest Neighbours, Support Vector Machine or Neural Networks. However, this logarithmic regression model was the one with the highest accuracy and AUC while maintaining a small difference between training and testing sets.