Norwegian School of Economics

Bergen, Autumn 2020

The Value of Interpretable Machine Learning in Wind Power Prediction

An emperical study using Shapley Addidative Explanations to interpret a complex wind power prediction model

Ulrik Slinning Tenfjord and Thomas Vågø Strand

Supervisor: Johannes Mauritzen

Master thesis, Economics and Business Administration, Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

Abstract

The main objective of this thesis is to evaluate if interpretable machine learning provides valuable insight into TrønderEnergi’s wind power prediction models. As we will see, interpretable machine learning provides explanations at different levels. The main objective is therefore answered by dividing the analysis into three different sections based on the scope of explanations. The sections are global, local, and grouped explanations. Global explanations seek to interpret the whole model at once, local explanations aim to explain individual observations and the grouped explanations aims to uncover observations with similar explanation structure. To quantify these explanations, we use Shapley Additive Explanations (SHAP). This approach takes a complex machine learning model and estimates a separate explanation model from which each feature´s marginal contribution to the predicted output is estimated.

The global analysis shows that wind speed is the biggest contributor to the prediction, while wind direction contributes to a lower degree. However, wind direction SHAP-dependence plot shows why wind direction is an important feature in wind power predictions. When including wind direction as a feature, random forest seems to take speed-up effects and wake effects into account.

In the local explanations we examine the observation with the highest prediction error and the one with highest imbalance cost. Inaccurate wind speed forecasts seem to be the cause of the observation´s large prediction error. An underestimation of the real production and a large spread between the spot price and RK-price seems to be the main contributor to the observation with highest imbalance cost.

In the cluster analysis, we see that when Numerical Weather Prediction (NWP) models predict different wind speeds for the same observation, the model tends to perform worse in terms of RMSE. Observations where NWP-models all predict either high or low wind speeds for the same observation, performs significantly better, with less than half as low RMSE.

We also discuss how these three explanation frameworks can be used to gain business benefits. We find that there are many potential benefits but some of the more prominent are legal, control and trust.

Acknowledgements

First, we like to thank our supervisor Johannes Mauritzen for excellent guiding and constructive feedback throughout the writing process. We would also like to thank TrønderEnergi AS and especially their AI department for the collaboration, suggestions, and providing us with the necessary data. Finally, we want to thank the Norwegian Computing Centre, and Anders Løland, for providing valuable insight regarding the methodology used in this thesis.

Key Words: Interpretable Machine Learning, Machine Learning, Wind Power Predictions, Random Forest, Clustering, Hierarchical Clustering, SHapley Additive exPlanations, SHAP-values, Business Benefits.

Contents

Figures and Tables

1. Introduction

Machine learning models are steadily increasing their presence as a part of the decision making process. As they aim to become even more “human“ in their appearance, we cannot simply measure their performance in terms of predictive accuracy. Humane properties like discrimination, safety and the right to an explanation will become important parts of any machine learning algorithm in the future. Assuming we are on a path, where machine learning adapts human behavior, then a question arises: How can we quantify these attributes? How can we measure an algorithm’s ability to discriminate an outcome or specific feature? How can we supply an explanation to those affected by an algorithm?

Enter Interpretable machine learning. Interpretable machine learning is all about understanding the reasoning of complex machine learning models. Only when we understand the reasoning of a model, can we evaluate the outcome in any other measure than predictive accuracy. In recent years, interpretable machine learning has experienced increased attention. Traditionally, the focus has been on increasing the predictive accuracy of machine learning models with little regard to interpretability. However, we now see a surge in scientific papers trying to explain what happens behind the curtain of machine learning models. In this thesis we will review some of the more established theories in the field and apply them in a case study of wind power predictions.

1.1 Motivation

In the context of wind power predictions, the focus has been on increasing predictive accuracy. For TrønderEnergi AS, the main objective has been to build a model that provides wind power predictions as close as possible to real production levels. This has led to increased complexity of their prediction model. As we know, accuracy tends to increase with complexity but when a model gets too complex for a human to easily interpret, it can be categorized as a black box model. We were tasked with unboxing this model to make it more interpretable.

Focusing on interpretability is important for a number of reasons such as operative and financial decision making and the reduction of prediction errors. TrønderEnergi´s prediction models occasionally suffer from large prediction errors, which could have an impact on the company´s financial performance. In particular, prediction errors can significantly increase their production imbalance costs, and as a result, decrease their net income. Hence, the company has a desire to know if there are any methods that enables humans to understand the decisions made by complex machine learning models, and if it discriminates on certain features.

TrønderEnergi supplied all data for this work. The data mainly consists of weather predictions and historical production at Ytre Vikna, a wind power plant just west of Rørvik (Northern Trøndelag).

1.2 Literature Review

Some of the more notable contributions to interpretable machine learning, have come in the form of local interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016a) and Shapley additive explanations (Lundberg & Lee, 2017). These articles put forward general methodologies that can be applied to explain any type of machine learning model. The field is experiencing rapid contributions and constant improvement. Recently, Aas et al. (2019) released an article refining Lundeberg’s approach in the case of feature dependency and there are many more examples which we will refer to as we go along. Christopher Molnar does a good job of collecting and presenting different approaches to interpretable machine learning, and its benefits in his guide to make black box models explainable (Molnar, 2019).

In recent years, some sectors have focused more on the usage of interpretable machine learning, than others. Typically, these are sectors or areas with a decision-making process that affect human-life, and also, areas that may have significant legal or ethical ramifications as a consequence of relying on AI. Medicine is an example of this. Here, interpretable machine learning has been used to explain a random forest classifier that predicts stroke (Prentzas et al., 2019). Another example is in financial risk-management, and typically credit-risk modelling. Bussmann et al. (2020) use TreeSHAP to estimate Shapley values, and to explain a tree-based prediction model that predicts the probability of whether a company will default or not. Lin (2018) also uses TreeSHAP to explain a multiclassification model that predicts whether a transaction may be fraud or not. Other fields where interpretability may be crucial include transport and self-driving cars or military operations using AI (Adadi & Berrada, 2018).

Very recently, a literature has started to emerge on the importance of interpretable machine learning within the energy sector. Vega García & Aznarte (2020) use SHAP-values to explain a deep neural network used to forecast the air quality or nitrogen dioxide concentration in Madrid. Other examples include Carlsson et al. (2020) who applies interpretable machine learning to interpret the consumption side of the energy market. The authors apply the original KernelSHAP method to estimate Shapley values, and use them to explain an artificial neural network predicting the energy consumption of an electric arc furnace. The authors pointed out a problem, namely that their explanations rely on estimates of SHAP-values that assume feature independency. As we have come to know, explanations with correlated features might be inaccurate. Interpretable machine learning has also been used for the supply side of the energy market. Kuzlu et al. (2020) use SHAP-values and LIME to explain solar PV power generation forecasts, provided by a random forest model. The authors objective was to create acceptable explanations, so interpretable machine learning can be applied in smart grid applications.

To our knowledge, interpretable machine learning has not yet been thoroughly explored in relation to wind power prediction, especially if we consider the new and improved SHAP estimates from the Norwegian Computing Centre (Aas et al., 2019). We think that we bring something new to the table and that we demonstrate how interpretable machine learning can be used, and the benefits of using it for wind power suppliers.

1.3 Problem Statement

The main objective of this thesis will be to evaluate if interpretable machine learning gives TrønderEnergi valuable insight into their wind power prediction models. As we will see, interpretable machine learning provides explanations at different levels. The main objective is therefore answered by dividing the analysis into three different sections based on the scope of explanations.

Global explanations:

How does the wind power prediction model behave globally and what features are most important in making predictions?

In the first section, we try to create a holistic overview of the prediction model. Global explanations seek to interpret the whole model at once and provide valuable information regarding the estimated relationships between the response variable and features.

Local explanations:

Why did the wind power prediction model make a certain prediction for a specific observation?

In the second section, we aim to analyze and establish explanations for single observations. By singling out specific observations we can get an understanding of why this exact prediction, turned out the way it did. The analysis will look at two particular observations, which are the one having the highest production imbalance cost and the one with highest prediction error. The method applied can be used to analyze any given observation.

Group-based explanations:

Why did the wind power prediction model make certain predictions for a group of observations?

In the third section, explanations for multiple predictions are aggregated into group levels. Groups are found by using hierarchical clustering. Our goal is to analyze whether any groups of observations can be distinguished from one another in terms of higher production imbalance cost and prediction error.

Business benefits with interpretable machine learning:

The main objective is answered by integrating the three explanation frameworks with a discussion of the importance of interpretable machine learning, and the possible business benefits for TrønderEnergi.

1.4 Results

The global, local, and cluster analysis yielded a variety of explanations. The global analysis showed that the three wind speed features contributes the most to the predicted output. Wind direction has smaller contribution in general. The SHAP-dependence plots gave especially two valuable insights. First, they verified that the trained relationships between wind speed and production are in accordance with wind turbines fundamentals or the power curve. Second, they exhibit the importance of wind direction as a feature in wind power predictions. This includes random forest's ability to consider the complex terrain at Ytre Vikna. When using wind direction as a feature, random forest seems to take speed-up effects and wake effects into account. These insights could potentially increase the trust of those domain experts or wind power engineers that doubt on using machine learning in wind power predictions.

In the section with local explanations, we chose to focus on explaining the observation with the highest prediction error and the observation with the highest production imbalance cost. In the case of the largest prediction error, this seems to be caused by bad estimates of wind speed. The large imbalance cost is mainly caused by the huge spread between the spot and RK-price and the fact that the prediction resulted in an underestimation of actual production.

The cluster analysis found groups of observations that had different patterns when considering prediction error and production imbalance cost. The main findings were the following: When NWP-models differentiate in their estimate of wind speed for any given hour, the prediction model tends to predict worse on an average. When predicted wind speed was relatively high for all NWP-models, the predictions where sufficiently better in terms of predictive accuracy. Low predicted wind speeds for all NWP-models, results in the predictions closest to actual production, likely because of low production elasticity when wind speeds are low.

In our discussion we establish why interpretable machine learning grants valuable insights into wind power predictions. Three areas we focus on are to comply with *regulation*, retaining *control*, and maintaining *trust*. Regulations may be put forward, that demands a sound reasoning for individual predictions. Local explanations can be used to see how features contributed to the prediction and uncover any irregularities.

Control is an area of focus because with interpretable machine learning you maintain a certain control over developments regarding the learning process. This control can be exploited to improve future versions of machine learning models, and subsequently improve the model´s predictive performance.

Trust can be important in an operative manner. The people working on the production central or in general internal stakeholders have little experience with machine learning. By giving them simple explanations as to why things are like they are, they might improve their understanding of the system as a whole. Increasing trust is beneficial since it potentially eases the implementation of new machine learning models in the future with less internal obstacles.

1.5 Thesis Outline

We will begin our thesis with an introduction to the Norwegian/Nordic power market. Here we will try to explain some features of the power market that has repercussions for our approach to creating explanations for the model. We will also define some economic variables used when analyzing different segments or clusters of observations.

After exploring the power market, we will take a deeper dive into the theory behind interpretable machine learning to give the reader an introduction to the subject. In addition, we aim to explain some of the choices that had to be made during the process of creating the explanations such as setting tuning parameters and distributions etc.

Then we explore the data supplied to us by TrønderEnergi. We examine the statistical properties of the different variables, take a look at the wind roses for Ytre Vikna, and lastly, perform a correlation analysis.

The next section covers the methodologies that are used in our analysis. First, we cover how we built the prediction model on which the explanations are based. Second, we cover the explanation model used to interpret the prediction model´s individual predictions. Lastly, we cover the clustering method used to find similar groups of explanations.

When estimating our wind power production model some considerations had to be taken. Since we cannot present TrønderEnergi's original model for competitive reasons, we made our own random forest model. Our model consists of many of the same features and returns similar predictions. The sources of the features have also been masked to avoid the need for secrecy.

When analyzing the results, we have focused on global and local explanations before doing a cluster analysis. Local and global explanations focus on getting an understanding of how the features contribute to the outcome as well as discovering interaction effects. The cluster analysis groups observations based on their SHAP-values. After clustering we can analyze attributes within each cluster such as economic properties and prediction errors.

1.6 Experimental Design

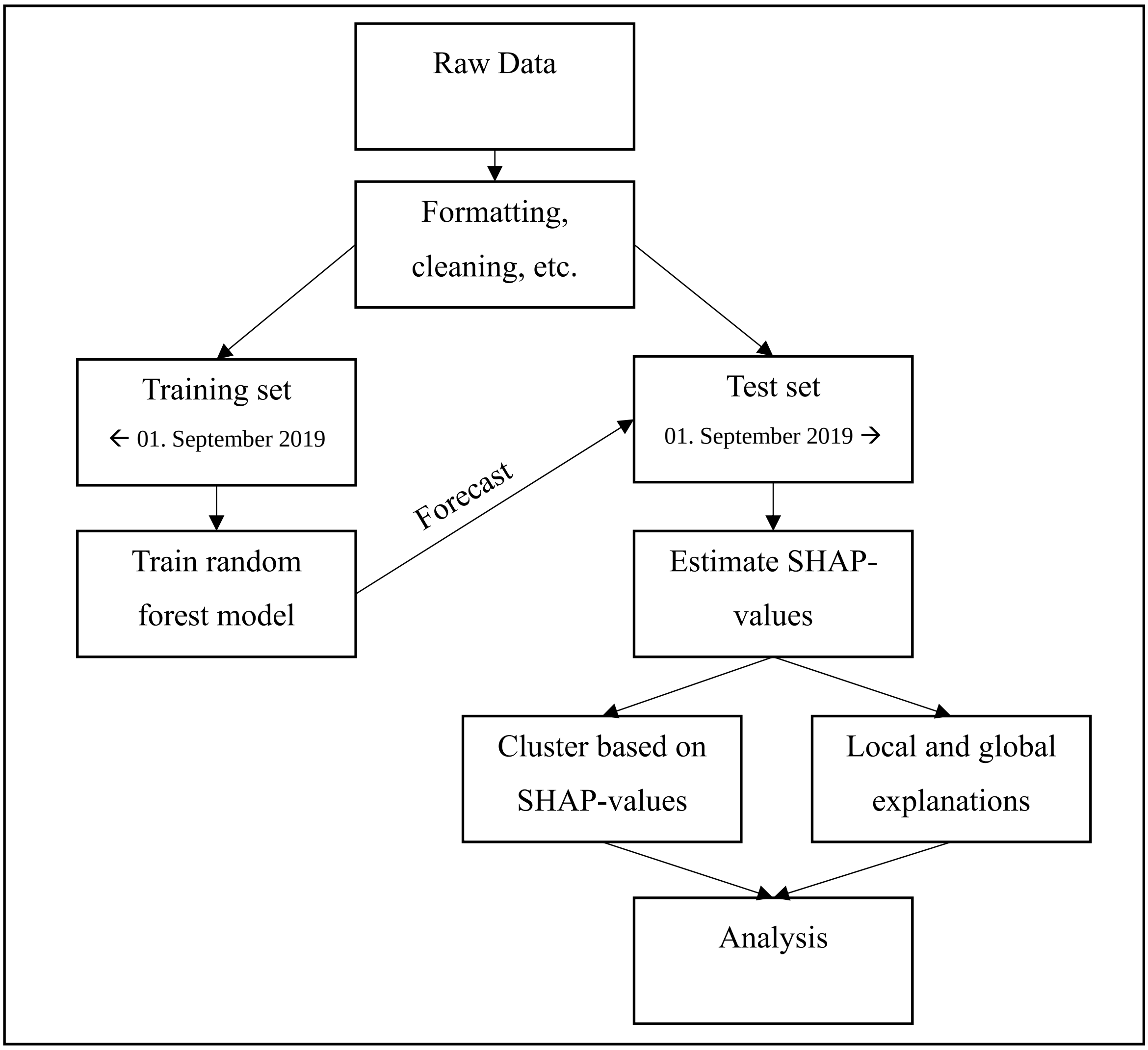

Raw Data
Formatting, cleaning, etc.
Training set
← 01. September 2019
Test set
01. September 2019 →
Forecast
Train random forest model
Estimate SHAP-values
Cluster based on SHAP-values
Local and global explanations
Analysis

Figure 1: Experimental Design

This is a visualization of our workflow or experimental design if you will. From raw data to complete analysis. After receiving the data, it had to be formatted, cleaned, and merged before we could split it into a training- and test set. We then trained our model on the training set and make forecasts for the test set. Based on this forecast, we estimate the SHAP- values. These values are then used to create local and global explanations, as well as being the features of our cluster analysis.

2. The Power market

The power market is a market in constant need of balancing as electricity must be consumed and produced at the same time. In Norway, Statnett is the responsible party for maintaining this balance between production and consumption. The participants of the power market do most of the job themselves through bidding at the day-ahead or intraday market driven by Nord Pool, but where there are discrepancies between supply and demand, Statnett offers solutions through their balance market.

In the wholesale market, suppliers and consumers make hourly bids based on their projections for production and consumption the following day or hours (Intraday). In the day-ahead market bids are delivered to the transmission system operator (Statnett) no later than 12:00 the day before. The market is split into price areas based on transmission capacity and bottlenecks. When a supplier/consumer delivers a bid, it is implicitly also a bid for some of the capacity of the transmission system. The price might vary from area to area given bottlenecks and lower transmission capacity. The intraday market closes one hour before the operating hour and gives market participants some room for adjusting their bids.

Still, there are uncertainties and challenges in planning power production as there are many factors involved, this is especially true for wind power since production cannot be planned to the same extent as for example hydro power. Production of power must also take the projected demand into consideration which is subject to rapid change given changes in the temperature and other unforeseen events such as malfunctions.

Challenges like this is what creates the discrepancies between supply and demand. In these events Statnett is tasked with creating balance through their reserve market which aims to adjust the consumption/production up or down, securing a momentarily balance.

2.1 Balance Markets

When the market participants cannot ensure a balance through the wholesale markets, Statnett makes use of the balance market. At a frequency of 50 Hz the market/power grid is in balance and no action is needed. When the frequency drops the consumption is increasing relative to supply and vice versa.

When an imbalance occurs, it is first regulated by primary reserves known as frequency containment reserves. An automated market exists to ensure there is enough reserves to respond to imbalances. This reserve market is divided in a weekly market, concluded before the elspot market, and a D-1 market concluded after trading in the elspot market has closed to cover any residuals.

Should the imbalance continue the secondary reserves are activated. The transmission system operator sends a signal to the suppliers which automatically changes the production at the facility.

In the case of further imbalances, the tertiary reserves are activated. These reserves are known as regulating power and are manually operated with an activation time of fifteen minutes. Here, producers and consumers of power can meet to offer their latest estimates delivered as close as 45 minutes before the next operating hour. The regulating power price is determined in this market, hereby referred to as the RK (regulerkraft) price. The market is often used as way to manage the deviations between planned and real production. The regulating power market and the wholesale markets organized by Nord Pool are the physical markets for electrical power in Norway.

2.1.1 How Trønder Energi Manages Production at Ytre Vikna

The balance market and how it is operated has multiple implications for TrønderEnergi. First of all, they must maintain their own power balance given by the balance equation:

$$Production + Import = Export + Consumption \tag{1}$$

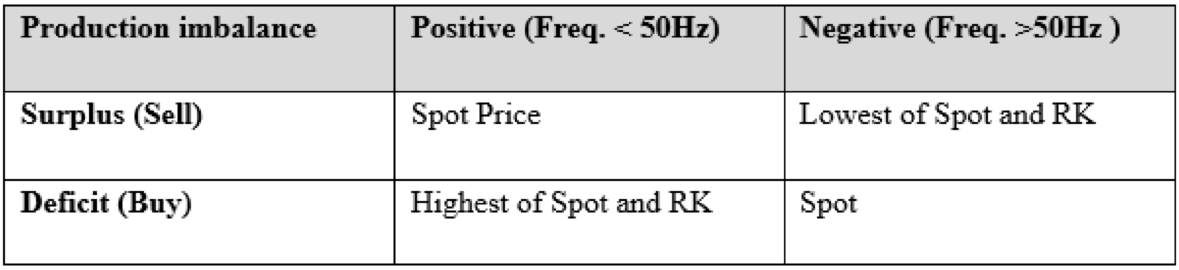
All power producers are tied to a company responsible for their balance and are legally obligated to maintain their own power balance. The companies plan their balance as good as possible and use the RK market for corrections (Ministry of Petroleum and Energy, 1999). There are two types of imbalances, consumption- and production balance (eSett, n.d.).

2.2 Production Imbalance

The production imbalance is reported hourly and can be found by comparing planned production to actual production and adding any activated corrections (use of balance markets).

$$Prod. Imbalance = Production - Planned\ Production\ \pm Activated\ Corrections \tag{2}$$

Imbalance is priced differently based on whether there is a surplus or a deficit in production and depends on the overall situation in the price area. A surplus of power (Frequency greater than 50Hz) in any given area demands a negative system correction to decrease the load and so the price for selling surplus production is the lowest of the spot price and the RK price. Should you need to buy during a negative correction you could get away with paying the spot price. If there is a deficit (frequency less than 50Hz) the situation is opposite. You would receive the spot price when selling during a positive correction and you would face the highest of the spot price and RK price when buying. This system is depicted in the table 1.



Production imbalance	Positive (Freq. < 50Hz)	Negative (Freq. >50Hz)
Surplus (Sell)	Spot Price	Lowest of Spot and RK
Deficit (Buy)	Highest of Spot and RK	Spot

Table 1: Price overview

2.3 Consumption Imbalance

Consumption imbalance is calculated as the deviation between consumption, planned production, trades, MGA imbalance and imbalance adjustment (eSett, n.d.).

$$\begin{aligned} &Consumption\ Imbalance \\ &\quad = consumption + planned\ production + trades - adjustments \end{aligned} \tag{3}$$

Consumption represents all consumption in the price area, planned production is the corrected production balance (replan – day-ahead), trades is the producers recorded trade of power before the operating hour and the adjustments is considered to be any manually activated regulations. The consumption imbalance is priced at the RK price, and do not follow the two-price system.

2.3.1 Deriving Income From a Production Facility

In the production process, TrønderEnergi relies on two forecasts of production; The day-ahead forecast which is reported the day before production, and a replan forecast which takes updated variable information into account and is reported one hour before production. The volume reported by the day-ahead forecast is sold at the spot price, while the difference between day-ahead and replan volume is bought/sold at the RK price.

The difference between the replan volume and actual production volume is bought/sold in accordance with table 1 but without regard to the overall situation in the price area. If there is a production surplus, this is sold at the lowest price, and any deficit is bought at the highest price. In other words, we take the worst-case scenario into account.

$$Income = Dayahead * Spot\ Price + RK\ Price(Replan - Dayahead) + \qquad (4)$$

$$(Production - Replan) * \begin{cases} Max(RK, Spot)\ if\ Production < Replan(Buy) \\ Min(RK, Spot)\ if\ Production > Replan(Sell) \end{cases}$$

The preferred result is to sell the whole volume at the spot price.

$$"Perfect\ Income" = Production * Spot\ Price \qquad (5)$$

By comparing the real income with the “perfect” income, we can get some understanding of how the forecasting models behavior impact the income from the production facility. This difference is referred to as “monetary loss” and is defined as:

$$Monetary\ loss = Perfect\ income - \ Income \qquad (6)$$

The income is a result of the complex relationship between the dayahead- and replan prognosis, real production, and the prices in the different markets. Analyzing this relationship has not been the primary focus in this thesis. We have focused on creating interpretations for the replan prognosis and because of this, as well as the limited data we had access to in regard to the dayahead prognosis, we have simply used Trønder Energi's dayahead prognosis in our analysis. In other words, we have estimated the replan prognosis on our own, and borrowed the historically recorded dayahead forecasts to compute the monetary loss.

One interesting scenario occurs when the RK price is greater than the spot price. In this scenario it would be beneficial to have a low day-ahead prognosis relative to the replan prognosis. Although power suppliers are bound by law (Ministry of Petroleum and Energy, 1999) to ensure their prognosis is the best estimate of actual production, it can be challenging to prove this if the model is in fact a black box model. Regulations may come into play that forces producers of power to account for their models in a way that makes interpretable machine learning obligatory for all suppliers.

2.4 Production Imbalance Cost

In this thesis we will focus on replicating TrønderEnergi´s replan prediction model, and subsequently explain this model´s individual predictions. We will also analyze the potential economic consequences that arises as a result of prediction errors from the replan prediction model. The economic consequences will mainly be from the last part of the income function. This is where TrønderEnergi needs to sell surplus power to the lowest of replan – and spot price, and they need to buy deficit power to the highest of replan – and spot price.
Surplus power means that the replan prediction model underestimates the power produced at Ytre Vikna, and deficit power means that the replan prediction model overestimates the power produced. There are two situations where potential loss of income can occur as a result of replan prediction errors. We will refer to them as "*production imbalance cost 1"* and *"production imbalance cost 2"*.

Production Imbalance cost 1

Production imbalance cost 1 arises when the replan model ***underestimate*** produced power at Ytre Vikna, and at the same time the RK-price ***is higher*** than the spot price. In these situations, the surplus power is sold to the spot price since this is the lower one. With a perfect replan prediction model, or a model that simply predict the power produced perfectly, the amount of surplus power would have been sold to the RK-price. This is a higher price than what they sold it for, and it is therefore a calculative cost, which is a result of the replan model's prediction error. In these situations, there is a loss of income due to prediction errors and the requirement to sell surplus power for a lower price than what they could have got by decreasing/eliminating these prediction errors. Mathematically we express this calculative cost as:

$$Prod. imb. cost\ 1 =$$
$$replan_{nok} * surplus\ power\ (MW) - spot_{nok} * surplus\ power\ (MW)$$

$Surplus\ power$:

$$Underestimated\ power\ they\ need\ to\ sell\ in\ order\ to\ achive\ balance$$

There are also situations where the replan model ***underestimates*** produced power at Ytre-Vikna, while at the same time, the RK-price is ***lower*** than the spot price. In these situations, the surplus power is sold to the RK-price since that is the lower price. With a perfect replan prediction model, the amount of surplus power would also have been sold to the RK-price. Therefore, there is no loss of income in these situations. This is because TrønderEnergi gets the same price for the surplus power in both scenarios where they predict poorly and perfectly. Thus, there is no production imbalance cost in these cases.

Production Imbalance cost 2

Production imbalance cost 2 arises when the replan model **overestimates** produced power at Ytre Vikna, and at the same time the RK-price **is lower** than the spot price. In these situations, the deficit power needs to be purchased to the spot price, since this is the higher price. With a perfect replan prediction model, there would be no need to purchase power, since there is no deficit power. However, when we overestimate the power produced, we need to purchase the amount of deficit power to a higher price than what we sold it for at the balancing market. Consequently, there is second situation with loss of income and production imbalance cost, that arises as a result of replan prediction errors. Mathematically we express this calculative cost as:

$$Prod.imb.cost\ 2 =$$

$$spot_{nok} * deficit\ power\ (MW) - \ replan_{nok} * deficit\ power\ (MW)$$

$Deficit\ power$:

$$Overestimated\ power\ they\ need\ to\ purchase\ in\ order\ to\ achive\ balance$$

There are also situations where the replan model ***underestimates*** produced power at Ytre-Vikna, and at the same time the RK-price is ***higher*** than the spot price. The deficit power is purchased at the RK-price in these situations. The amount of deficit power was sold at the RK-price as well, and hence there is no loss of income in such situations. Consequently, in these situations there is no production imbalance cost, even though we see large prediction errors.

By adding production imbalance cost 1 and 2 we get the total production imbalance cost for each hour, which is the economic measure we apply for production imbalance cost in this thesis.

3. Interpretable Machine Learning

The wide range of papers and different applications for interpretable machine learning have led to confusion about the concept of interpretability. Lipton (2018) remark this by saying the term of interpretability is ill-defined, and that goals and methods used within this field are very different from each other. Lipton also point out that different ideas within interpretability machine learning need to be extricated from each other before any implementations can be done. In this section, there will first be a proper definition of interpretable machine learning, and then a classification of the different methods contained within the term.

Murdoch et al. (2019) formulated a definition that clearly express what interpretability means in terms of machine learning: *"Interpretable machine learning is the extraction of relevant knowledge from a machine-learning model concerning relationships either contained in data or learned by the model"*

Another definition comes from Molnar (2019): *"Interpretable Machine Learning refers to methods and models that make the behavior and predictions of machine learning systems understandable to humans".*

3.1 Classification of Interpretable Machine Learning methods

DARPA came up with a proposal that categories explanation models into three different classes. These classes are *Interpretable Models, Model Induction* and *Deep Explanation* (DARPA, 2016). There has since been an increasing development within interpretable machine learning. As a result, their groups could potentially be outdated or miss some important aspects that distinguish explanation models. Molnar (2019) propose an updated overview of classes that helps distinguish different interpretability models. We will use Molnar's framework to classify and distinguish the different methods used in interpretable machine learning (IML).

3.1.1 Molnar´s Classification of Interpretable Machine Learning Methods

Intrinsic Versus Post Hoc

Intrinsic: Interpretability is realized by restricting the complexity of a machine learning model. This means that prediction models are interpretable through their transparent and simple structure, like simple linear regression.

According to Lipton (2018), transparency is evaluated on the three following criteria: *simulatability, decomposability, and algorithmic transparency*. Simulatability looks at the entire model when evaluating transparency. A model is transparent if an individual can explain the entire model at once. Decomposability looks at the individual components of a model to evaluate transparency. A model is transparent if every part of the model (inputs, parameters, calculation) exhibits an intuitive explanation. Lastly, algorithmic transparency looks at the training algorithm in order to evaluate transparency. The model is transparent if there is a possibility to understand the optimization process that produces predictions.

Post hoc: Interpretability is achieved by using methods that analyze a machine learning model after it is trained. These interpretations do not clarify exactly how a model works (transparency), but they do convey valuable interpretations to stakeholders using these machine learning models. Post hoc analysis can handle all types of machine learning models, including intrinsic models, but they are often applied to explain the output from complex and less transparent machine learning models.

Model-Specific Versus Model-Agnostic

Model-Specific:

This refers to model-based interpretability. It is machine learning models that are interpretable from their simple structure. Model-specific interpretation tools are only applicable for the particular machine learning method used. The coefficients or weights from a linear regression model are typical model-specific interpretations. The tree structure from a decision tree also provides good interpretations for why a particular prediction were made. For instance, the way a tree is split gives mode-specific interpretations of the feature importance. The abovementioned interpretation tools are only applicable for the specific method, which mean that they are indeed model-specific interpretations. Model-specific interpretations are always intrinsic.

Model-Agnostic:

Model-agnostic interpretation tools are applicable for all types of machine learning models. They are used after a machine learning model is trained. With other words, model-agnostic interpretations are always post-hoc. The original model is handled as a black-box model, and the interpretations are based on a separate explanation model. In the separate explanation model, feature values are often permutated. Then we look at how the black-box model respond to these changes. Hence, for model-agnostic models, there is no need for transparency or knowledge about the inner workings of the original model in order to create interpretations. Ribeiro et al. (2016a) specifies three benefits by using model-agnostic methods:

1. *Model flexibility:* Model-agnostic interpretation methods are applicable for any type of ML-models. Even for complex models such as random forest, gradient boosting, and neural networks.
2. *Explanation flexibility:* these methods are able to produce different types of explanations, or with other words, they are not restricted to a specific type of explanation.
3. *Representation flexibility:* these methods produce different feature representations.

Local Versus Global

Local - and global explanations are two categories that methods with post-hoc interpretations lie within (Murdoch et al., 2019).

Local:

Interpretation methods that explain why an individual prediction were made from a machine learning model, are classified as local explanations. These models try to explain a machine learning models local behavior with an individual observation in mind. Two examples of local interpretation methods are LIME and SHAP.

Global:

Interpretation methods providing global explanations have all observations from a dataset in mind. They aim to explain the entire machine learning model´s behavior at once. Two examples of global interpretation methods are *partial dependence plots* and *global surrogate* *models*. It is worth noting that explanations from local interpretation methods like SHAP, can be aggregated and hence provide global explanations as well.

3.2 Tradeoff-between interpretability and performance

In this thesis we have a supervised learning problem that assumes there is a relationship between our quantitative dependent variable Y and features $X = (X_1, X_2, X_3, \dots, Xp)$. In general, this relationship can be expressed as the following equation:

$$Y = f(X) + \epsilon \tag{7}$$

f: Fixed but unknown function of the feature values X. Reflects the systematic information contained within X that can explain the behavior of Y.

ϵ: Unsystematic error term with mean equal to zero and independent of X.

f is unknown, therefore it needs to be estimated by methods from statistical learning. According to James et al. (2013), users of statistical learning have two different motives to estimate f: *prediction* and *inference*. A prediction's objective is to estimate f in order to get the best prediction of the response variable as possible. The goal of inference is to objectively understand how the response variable is affected when the independent variables changes.

Estimation methods from statistical learning are quite different when we consider a model´s structure. Linear regression has a simple structure that produces only linear functions. In other words, it is an inflexible method. On another side, random forest is very flexible. This method has a complex structure with non-linear shapes to estimate f. Theory claims that there is trade-off between interpretability and flexibility. According to this theory, it is best to apply inflexible and simply structured models when *inference* is the main objective (James et al., 2013). Flexible models are often able to find more complex relationships between the response variable and the feature variables which often leads to better predictive accuracy. Thus, when *prediction* is the main objective, we often seek more complex models in order to get superior predictive accuracy.

As a general rule of thumb, we often say that a model´s increased complexity leads to increased predictive accuracy, but at the same time decreases a model´s interpretability (DARPA, 2016; Oxborough et al., 2018).

TrønderEnergi's AI department constantly seek to find the method that estimate f best in terms of predictive accuracy. We aim to create quality explanations from these complex methods. The goal is to provide a model with high predictive accuracy and good interpretations, which breaks with the trade-off above. Interpretable machine learning consists of techniques to avoid a trade-off by applying post-hoc analysis and model-agnostic methods (section 3.3).

3.3 Model-Agnostics Methods

3.3.1 Additive Feature Attribution Methods

Explanations from model-agnostic methods handles the original prediction model f as a black box, while the interpretations are based on a separate explanation model g that tries to approximate the original model. LIME and SHAP are two different explanation frameworks used to explain a prediction model based on a single input or feature vector x. They are both *additive feature attribution models* that provide local explanations.

In an additive feature attribution model, a prediction is simply explained as the sum of the effect values of all feature's attributions. It is a linear function of binary variables:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j \, z' \tag{8}$$

Molnar (2019) describes $z' \in \{0,1\}^M$ as a coalition vector, where M is the maximum coalition size. 0 means that the corresponding feature is "absent" in the coalition, and 1 means that the corresponding feature is "present" in the coalition. ϕ_j is described as the effect that feature *j* attributes with to the coalition's output value.

3.3.2 Properties with Additive Feature Attribution Models

A property with additive feature attribution models is that there exists one unique solution that satisfies three desirable abilities (Lundberg & Lee, 2017). These three properties are *local accuracy, missingness* and *consistency*.

1. With *local accuracy* we mean that approximating the original model *f(x)* for any input of *x*, the explanation model is bound to at least match the output of *x* for any simplified input z'.

$$f(x) = g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j \, z_j^{'}$$

2. *Missingness* tells us that in the lack of presence of a certain feature, that feature should have no impact. In other words, if a feature is not present in a subset of features, its impact is constrained to zero.

$$z_j' = 0 \; (indicating\ that\ feature\ \boldsymbol{j}\ is\ missing) \rightarrow \phi_j = 0$$

3. *Consistency* is the fact that if the marginal contribution of a feature increases or stays the same, the estimated contribution follows the marginal contribution of the feature. Proof of this is provided in the appendix to Lundberg & Lee (2017).

Lundberg proves that only methods that are based on Shapley values can be used to satisfy these properties at the same time, which has great implications for our choice of methods.

3.3.3 Shapley Values

Shapley values is a concept that originally stems from cooperative game theory. It was developed of the Nobel prize winner Lloyd S. Shapley (Shapley, 1953) and is best illustrated through an example.

Let us picture a cooperative game with N players, and the main objective is to maximize every players payoff. A subset or coalition of S players has the following notation $S \subseteq N = \{1,2,3 \ldots, N\}$. The characteristic function, $v(s)$, maps every subset of players to the expected payoff they receive by collaborating. In other words, it simply describes the value of a coalition. Under the assumption that every player collaborates, the Shapley Values “fairly” distributes the total payoff between the participating players. The distribution is based on each player’s contribution to the total payout. The distribution is fair since it is the only set of values which satisfy the four properties *efficiency, symmetry, dummy player* and *linearity* (Shapley, 1953; Young, 1985).

The Shapley value for player i is computed as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(N-|S|-1)!}{N!} \left(v(S \cup \{i\}) - v(S)\right), \;\; i = 1, \dots, N \tag{9}$$

$v(S \cup \{i\})$: *Payoff received with player* i *included in the coalition*

$v(S)$: *Payoff received with player* i ***not*** *included in the coalition*

Shapley value ϕ_i is the average marginal contribution for player i across all potential subsets, and it will be player $i\acute{}s$ distributed payoff from the grand coalition with all the players. An empty coalition, $S = \emptyset$, is also considered.

To give the reader a more intuitive understanding of the application of Shapley Values, we think of a game with three players $N = \{1,2,3\}$. There are eight possible coalitions: $\{\emptyset\}, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}$. We assume the coalitions have the following payouts: $v\{1\} = 4$, $v\{2\} = 8$, $\mathrm{v}\,\{3\} = 6$, $v\{1,2\} = 12$, $v\{1,3\} = 16$, $v\{2,3\} = 14$, $v\{1,2,3\}$ = 36, $v\{\emptyset\}$ = 0.

Each player´s average marginal contribution can be computed with equation 9:

$$\phi_1 = \frac{1}{3}(36-14) + \frac{1}{6}(12-8) + \frac{1}{6}(16-6) + \frac{1}{3}(4-0) = 11$$

$$\phi_2 = \frac{1}{3}(36-16) + \frac{1}{6}(12-4) + \frac{1}{6}(14-6) + \frac{1}{3}(8-0) = 12$$

$$\phi_3 = \frac{1}{3}(36-12) + \frac{1}{6}(16-4) + \frac{1}{6}(14-8) + \frac{1}{3}(6-0) = 13$$

Note: ϕ_0 = $v\{\emptyset\}$ is the fixed payoff when no players are included, and it is usual zero in cooperative games (Aas et al., 2019). However, this is not the case when implementing Shapley values to explain a predictive model.

3.3.4 Local interpretable model-agnostic explanations (LIME)

The LIME framework proposed by Ribeiro et al. (2016b) aims to explain predictions through the use of local surrogate models. These surrogate models are trained to approximate the predictions of any underlying black box model. As the name implies, this methodology focuses on training local models to explain individual predictions as opposed to global explanations.

LIME works by sampling a new dataset from the old along with the corresponding prediction from the black-box model. We can then train any interpretable model and weight the model according to the proximity of the sampled observations to the individual observation of interest. The goal of the trained model is to be a good approximation locally, we are not too concerned with the global accuracy.

Mathematically we can write this as:

$$explanation(x) = argmin_{g \in G} \quad \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{10}$$

Where $g \in G$ is a set of interpretable models like linear regression models, decision trees etc. Since not all interpretable models have the same complexity, number of interpretable components or the same ease of interpretation, we let $\Omega(g)$ be a complexity measure of the model $g \in G$. For instance, the depth of a decision tree. f denotes the prediction model to be explained and π_x is a proximity measure, measuring the size of the neighborhood around x that we consider for the explanation.

The explanation of observation x is given by minimizing $\mathcal{L}(f, g, \pi_x)$ which measures how close the surrogate prediction is to the original model in the neighborhood π_x. The complexity measure is user defined but should be low enough so that the surrogate model is easily interpreted.

3.3.5 SHAP (Shapley Additive explanations)

To explain the prediction of a feature vector x by using cooperative game theory and Shapley values, Lundberg & Lee (2017) proposes the SHAP methodology. This method tries to explain individual predictions by using Shapley Values from cooperative game theory.

SHAP consider features as *“the players”*, combinations of different features as *“the coalitions”*, and the prediction as *“the total payout”*. The Shapley value ϕ_i is the average marginal contribution for feature i throughout all potential coalitions. Thus, it explains each feature´s contribution to a prediction.

In a prediction setting, we apply the training data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ where $\boldsymbol{x_i} = \left(x_{i1}, x_{i2}, \ldots, x_{ip}\right)^T$ to estimate a model $\hat{f}(x)$. We try to explain a prediction from $\hat{f}(x)$ for a particular feature vector $\boldsymbol{x} = \boldsymbol{x}^*$. It is here SHAP comes in to play.

Lundberg & Lee (2017) defines the contribution function $v(S)$ as follow:

$$v(s) = \mathrm{E}\,[f(\boldsymbol{x})|\boldsymbol{x_s} = \boldsymbol{x_s^*}] \tag{11}$$

$$x_s : Features\ in\ subset\ S$$

$$x_s^* : Subset\ S. Input\ values\ that\ we\ try\ to\ explain$$

$v(S)$ is the expected output from a prediction model, given that coalition S have value $\boldsymbol{x_s^*}$**.** In our case, the contribution function $v(S)$ represent the predicted production of wind power for a specified coalition. The Shapley value for a given feature x_i is then computed by substituting $v(S)$ in equation 9 by its conditional expectation (Redelmeier et al., 2020).

The SHAP framework is consistent with *additive feature attribution*. An individual prediction is decomposed by equation 8 where all features are “present” in the coalition vector. In this setting, empty coalitions have a payoff of $\phi_0 = E[f(x)]$ instead of zero (Aas et al., 2019). This expected value is the average of all production and what we refer to as the *baseline*. As a result, the sum of Shapley values in an additive attribution model is equal to the difference between the predicted value and the global average prediction. We can say that: “Feature *i´s value in a feature vector $\boldsymbol{x} = \boldsymbol{x}^*$, contributed ϕ_i to the prediction compared to the average prediction (baseline) in the training data* (Molnar 2020). In other words, the Shapley value is simply a feature´s positive or negative contribution to the prediction relative to a baseline.

There is a dimensional problem in computing Shapley values. The number of coalitions increases exponentially (2^N) as the number of features increases, and when a model contains a certain level of features the computation becomes infeasible (Molnar, 2019; Redelmeier et al., 2020). Thus, an approximation is often required in order to compute Shapley values in a practical setting.

3.3.6 KernelSHAP

KernelSHAP approximates the real Shapley values to explain individual predictions. It does so by combining LIME with Shapley values. Shapley values could be found by using equation 10, but this is dependent on the selection of kernel weight $\pi_{x'(z')}$, the regularization term Ω and the loss function L.

Under the definition that additive feature methods have an explanation model that is a linear function of binary variables, the only choice of $\pi_{k'}$, L , Ω that make the output from equation 10 consistent with the three properties *local accuracy*, *missingness* and *consistency* are:

$$\Omega = 0,$$

$$\pi_{x'(z')} = \frac{(M-1)}{\binom{M}{|z'|}|z'|(M-|z'|)},$$

$$L(f, g, \pi_{x'}) = \sum_{z' \in Z} \left[f\left(h_x(z')\right) - g(z')\right]^2 \pi_{x'}(z'),$$

where $|z'|$ is the number of non-zero components in a coalition vector z'

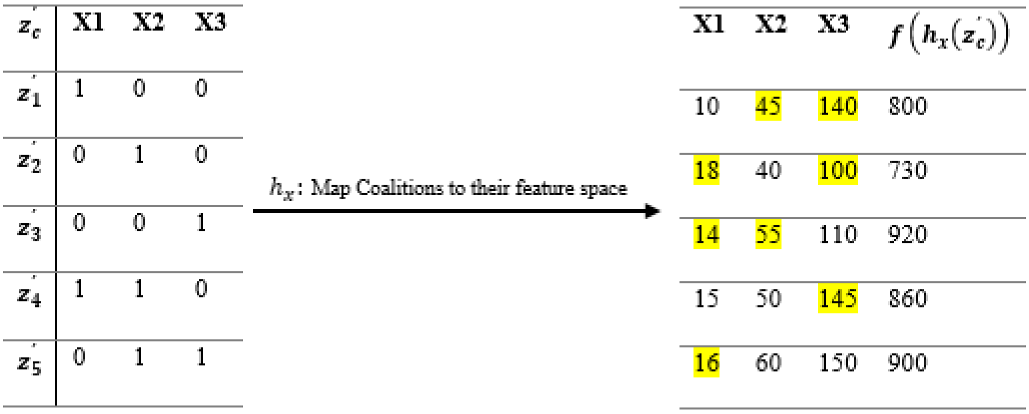
Due to the fact that $g(z')$ is a linear function and the loss function is a squared loss, Equation 10 is solved, and the Shapley values are calculated by using weighted linear regression. As a result, features coefficient in the weighted linear regression model will correspond to their estimated Shapley value or their attribution in a prediction of a specific instance.

Molnar (2020) divides the computation of Shapley values using KernelSHAP in five parts:

1. Sample coalitions $z'_c \in \{0,1\}^M,\ c \in (1,2,\dots,C)$,
   $where\ \mathbf{0} =$ "absent" feature, $\mathbf{1} =$ "$present$" $feature$

2. Compute predictions for all coalitions z'_c by using model $f\left(h_x(z'_c)\right)$.

   Here, all C coalitions are mapped to their original feature space by using the mapping function h_x. 1´s are mapped to their corresponding value from feature vector $\boldsymbol{x} = \boldsymbol{x}^*$ that we try to explain. 0´s are mapped to a random sampled value from the training data. The original KernelSHAP method assumes feature independency. Mapped values for “absent” features will therefore be sampled from their marginal distribution. Mapped values are used as input values in the original prediction model *f* and it subsequently gives each coalition a predicted value. The first two steps are illustrated in the figure 2, where five coalitions are sampled. Yellow values illustrate the mapped values for “absent features”, which are sampled values from the feature’s marginal distribution.



z'_c	X1	X2	X3
z'_1	1	0	0
z'_2	0	1	0
z'_3	0	0	1
z'_4	1	1	0
z'_5	0	1	1

h_x: Map Coalitions to their feature space

X1	X2	X3	$f\left(h_x(z'_c)\right)$
10	45	140	800
18	40	100	730
14	55	110	920
15	50	145	860
16	60	150	900

Figure 2: Example of mapped coalitions

3. Calculate the weight for each coalition z'_c by using the kernel weight function $\pi_{x'(z')}$. From the kernel weight function, we see that coalitions with few or many “present” features (1´s) are weighted most heavily.
4. Fit a weighted linear regression model, using the mapped features values as input values, and the predicted values for each coalition as the target value, and each coalition´s kernel weight. In other words, minimize the loss function L.
5. Coefficients from the weighted linear regression model are returned as features approximated Shapley values or their marginal attribution in a prediction for a particular feature vector $\boldsymbol{x} = \boldsymbol{x}^*$.

Proof of why explanations from the KernelSHAP method are consistent with properties 1-3 are found in Lundberg & Lee (2017) supplementary material. For a more extensive mathematical explanation of the KernelSHAP estimation, we recommend looking closer at Aas et al. (2019) paper.

3.3.7 New version of KernelSHAP

A disadvantage in using the original KernelSHAP in order to estimate SHAP values is the method´s assumption of feature independency. This can potentially lead to an estimation that weight unlikely instances in a too large degree. As a result, it might yield unreliable explanations. If there is a high correlation or dependency between a model´s features, explanations from KernelSHAP could potentially be wrong.

Aas et al. (2019) solves this problem by relaxing the independency assumption. They developed a method that takes dependency between features into account. In their approach, a conditional distribution between features $p(x_{\bar{S}}|\boldsymbol{x_S} = \boldsymbol{x_S^*})$ is estimated using the training data. The notation $x_{\bar{S}}$ is the part of x not in x_S, or simply absent features in a coalition. In step two of Molnar's overview of how to estimate Shapley Values using the KernelSHAP, absent features are now sampled using the estimated conditional distribution, instead of the marginal distribution. Aas et al. (2019) propose four different approaches to estimate the conditional distributions:

1. **Multivariate Gaussian distribution**
   Absent values are sampled from a multivariate gaussian distribution, where the expectation vector $\boldsymbol{\mu}$ and full covariance matrix $\sum$ are estimated using the training data.

2. **Gaussian copula**
   If the features are nowhere near being multivariate Gaussian distributed, this approach is better to use in order to estimate the features conditional distribution. Here, features marginals are characterized by their empirical distributions, and then the dependence structure is modelled by a Gaussian copula.

3. **Empirical conditional distribution**
   If features dependency structure and multivariate distributions are nowhere near being Gaussian, this method is proposed by the authors. It is a non-parametric estimation method, motivated by the Nadaraya-Watson estimator.

4. **Combined approach using the empirical – and the Gaussian or Gaussian copula**
   This approach combines the empirical distribution with either the Gaussian or Gaussian copula distribution. An example is to use the empirical approach when we are conditional on 1-2 features, and Gaussian distribution when we are conditional on more than two features.

All approaches for estimating the conditional distribution are thoroughly explained in Aas et al. (2019) paper.

4. Data

This section aims to give the reader better insight into the data on which the analysis is built. The raw data is supplied by TrønderEnergi. It has a time interval from 09. April 2017 to 09. September 2020 and the observations are of an hourly frequency which results in 29 976 observations.

4.1 Data preprocessing

Two of the features used in predicting wind power contains 365 missing values. According to Pratama et al. (2016) there are three different conventional methods to handle missing values in time series data. They are *ignoring, deletion* and *mean/mode imputation.* Ignoring the missing values is not an option, due to the characteristics of our prediction model. It is not possible to predict an outcome using a random forest model and a feature vector containing missing values. This problem is often solved in random forest algorithms by imputing missing values (Gupta, 2015). *Mean imputation* replaces missing values with the mean value for a particular feature. Wind speed is one of the features containing missing values. Wind speed has a random and fluctuating pattern throughout the time series. In our opinion, it will be a mistake to use *mean imputation* because there is a probability that a replaced mean value is not representable for a particular observation. Replacing values using *mean imputation* could lead to misinformation about the relationship between a feature vector and the response variable. Hence, we are left with *deletion*. By deleting these observations, we lose some information, however it should not be a significant problem, since it only accounts for roughly one percent of the total data set.

The dataset is split into a training and test set. The training data contain all observations before 01. September 2019, and the test data contain all observations after 01. September 2019. The test and training data make up about 30 % and 70% of the total data, respectively. The main reason for this split is a desire to have explanations on out-of-sample predictions throughout a whole year.

4.2 Numerical weather predictions

The features used in our prediction model are forecasted values from different numerical weather prediction (NWP) models. NWP models use current weather observations and combine them with a mathematical model to provide a forecast of future weather (NCEI, 2020). In developing a wind power prediction model, the choice of NWP model(s) is a crucial step. According to Foley et al. (2010) a model developer should consider the three following criteria when choosing between NWP models: *The geographical area*, *resolution* and *forecast horizon*. In their article, there was a solid inter-dependence between a wind power prediction model´s accuracy and a numerical weather prediction model´s accuracy.

Nielsen et al. (2007) conducted an analysis with multiple wind power forecasts, which were all based on different numerical weather prediction models. In general, the models used in forecasting wind power, all had an approximately equivalent performance or accuracy. The authors displayed, by combining wind power forecasts based on different NWP models, the forecast error decreased compared with the single model’s accuracy level. The concept of combining multiple NWP models is used in building our prediction model as well, only now within a machine learning setting.

Our features are based on three different NWP models. The weighting of the different NWP models changes for every single prediction. In fact, these are the weights we will be able to see in the separate explanation model.

We aim to replicate TrønderEnergi’s prediction model, and then explain the predictions from this model. Therefore, we do not face the challenge of selecting the best NWP models for our data and wind park. That is a challenge already solved by TrønderEnergi and we use the same.

4.3 Features

In our wind power prediction model, we apply three different NWP models, and that is NWP100, NWP200 and NWP500. Each of these models provide a one-hour forecast of the wind speed and wind direction at Ytre Vikna. As a result, we have six features, *three wind speed forecasts* and *three wind direction forecasts*.

4.3.1 Wind Speed forecasts

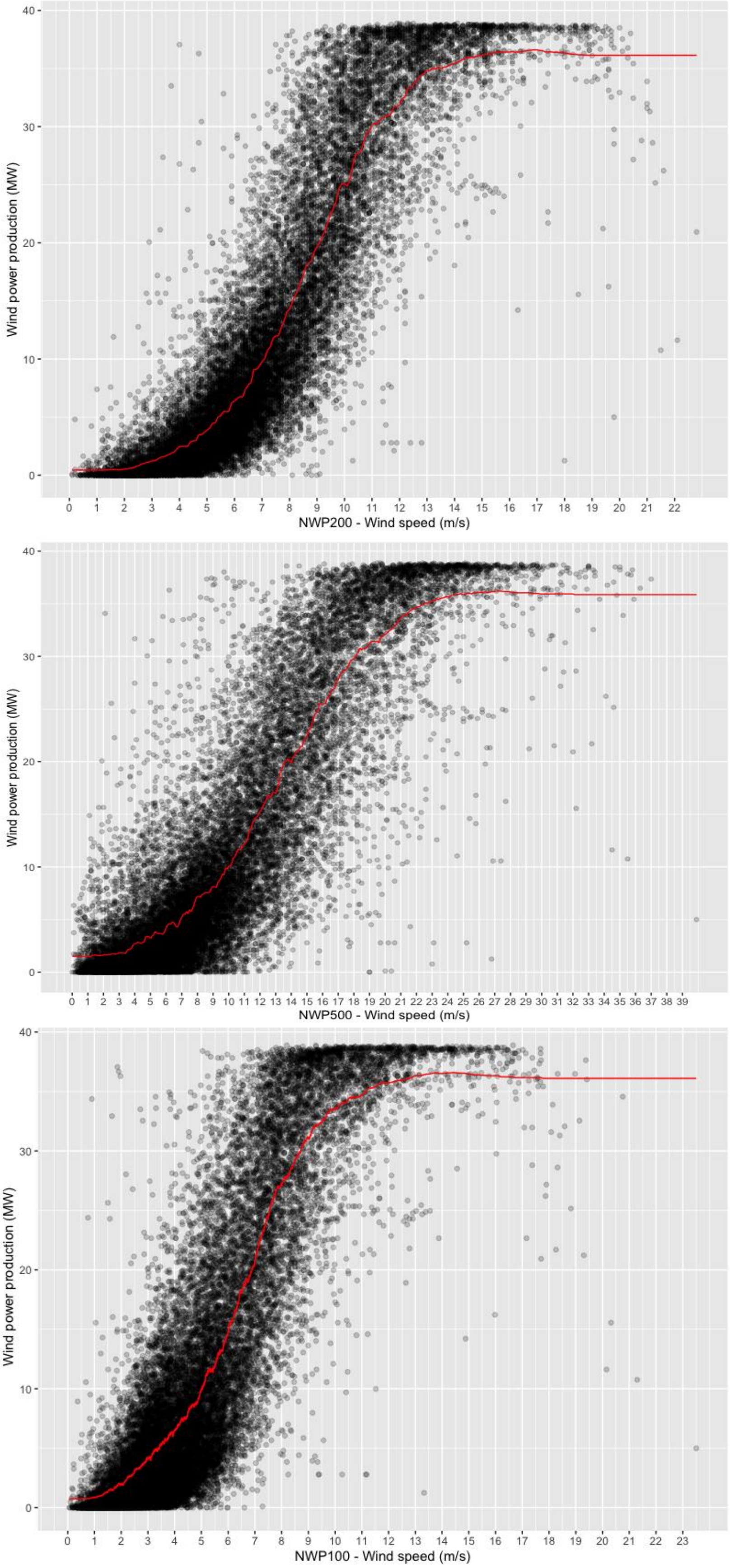


Figure 3: Power curves

The three forecasts of wind speed, all have continuous scales, and their unit is *meter per second (m/s)*. The figures illustrate the relationship between the power production and wind speed forecasts given by the different NWP models.

All of the 17 wind turbines at Ytre Vikna are 2,3 MW wind turbines of the type E70, and they are delivered by Enercon GmbH (Akram, 2014). The power curve for this type of wind turbine is provided by the manufacturer, and it displays the relationship between wind speed and power produced at a constant air density level. This is shown in appendix (see Appendix H). From this power curve, we see that wind turbines of this type produce wind power within a wind speed range of 3 m/s and 30 m/s. The production tends to be highest, and most stable, in the range of 15 m/s to 30 m/s. Wind speed within this range is categorized as the *rated speed*, since within this range, the turbines produce their rated/maximum power (Lydia et al., 2014). 3 m/s is the turbines cut-in speed, which means there is no production if the wind speed is below this boundary. The same is true for wind speeds above 30 m/s, but this is the cut-off speed, which means that there is no production if the wind speed is above 30 m/s (Akram, 2014).

The figures show the empirical power curve for each of the NWP models, and they are modelled using the K-nearest neighbors (KNN). The power curves all exhibit the same pattern: A flat and constant production level when the wind speed is very low, which is a result of the cut-in speed. When the wind speed increases, the production increases in linear like fashion until the wind turbines capacity is reached. When the capacity is reached, the production level starts to flatten out until it reaches the cut-off speed. K is set to 500 in the modeling process, and therefore, the empirical power curves are not able to perfectly envision the cut-in/off points.

The NWP models vary somewhat in regard to the empirical power curve. *NWP500* and *NWP200* tend to have the same cut-in speed. *NWP100* starts increasing between 1 to 2 m/s. The rated speed or maximum capacity seems to be reached around 12 m/s, 10 m/s and 20 m/s for NWP200, NWP100 and NWP500, respectively. The cut-off speed is not displayed at the estimated power curves, but observations tend to decrease at 21-22 m/s, 18-19 m/s and 34 m/s for NWP200, NWP100 and NWP500. This may be an indication that the cut-off speed is reached. NWP-models differences occur mainly as a result of different height-profiles. The wind park is located around 100 meter above sea level, and the turbines are between 64-74,5 meter above ground. NWP200, NWP500 and NWP100´s predictions are approximately at 200, 500 and 100 meter above sea levels respectively (Source: TrønderEnergi).

4.3.2 Wind Direction forecasts

The three predictions of wind direction are all circular variables. This means, their values are represented as a point on the circumference of the unit circle. Wind direction´s unit is *degrees,* and so the values are within the range of 0-360 degrees. 0 and 360 degree is far from each other on a linear line but is nearly identical when measured in degrees. According to Pewsey et al. (2013) there is a need to abandon standard statistical techniques developed for linear data, and instead apply statistical techniques developed for circular data.

When building a prediction model f, there are two different approaches to handle circular variables. One option is to make a linear transformation, by using cosines and sinus. The other option is to keep the data as is. We executed a comparable analysis by training a model with and without transformed wind direction variables. We then compare the predictions of the two models. This comparison can be seen in the appendix (see Appendix A), but the difference is minimal. Because of the small difference we went in the most interpretable direction and chose to keep the direction features in their original format as degrees.

4.4 Response

The response variable is *hourly wind power produced* at Ytre Vikna. There are 17 wind turbines at this wind park, and the response variable´s value correspond to the aggregated production level from all these turbines within an hour (Rosvold, 2019). The response variable´s unit is *megawatt (MW)*.

4.5 Descriptive statistic

In the descriptive analysis, there is a mix of linear- and circular variables. Circular statistical approaches are required for the wind direction features. Trigonometric functions can be used when calculating the descriptive statistic for wind direction. The NWP-models average wind direction is calculated using *the resultant vector average wind direction*. This approach is explained by Grange (2014). Wind direction´s standard deviation is approximated using the

Yamartino method (Yamartino, 1984). These are both methods for circular variables and developed specifically for wind direction.

	Mean	Median	Max	Min	SD
Response					
Production[1]	12.51	7.68	38.88	0.00	12.24
Features					
NWP100 Wind Speed[1]	5.19	4.60	23.51	0.06	3.01
NWP500 Wind Speed[1]	10.20	9.00	39.90	0.00	6.17
NWP200 Wind Speed[1]	7.16	6.70	23.90	0.10	3.44
NWP100 Wind Direction	213.23[2]	-	-	-	92.42[3]
NWP500 Wind Direction	217.56[2]	-	-	-	87.33[3]
NWP200 Wind Direction	202.69[2]	-	-	-	94.51[3]

Table 2: Descriptive statistics

[1] Traditional statistical techniques for linear variables

[2] Resultant vector average wind direction (Grange, 2014)

[3] Yamartino method to approximate wind direction´s standard deviation (Yamartino, 1984)

4.6 Wind Rose

A wind rose is a visualization tool often applied by meteorologists to outline the wind direction and wind speed probability distribution (Lepore et al., 2020). Here, a wind roses provide a good overview of the distribution of both wind direction and wind speed. The size of each pillar represents the frequency of a wind direction bin in the dataset. The colors represent different bins of wind speed, and their size within a pillar display their frequency for a particular wind direction bin.

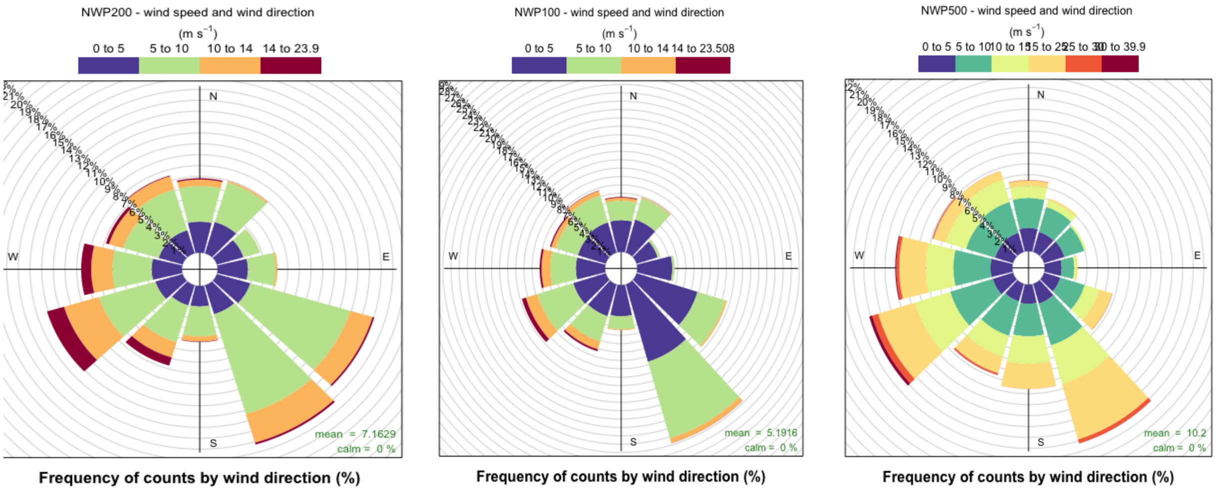


Figure 4: Wind Roses

NWP200: The two largest pillars are between east and south, which indicate that it blows most frequently towards these directions. The speed in this wind direction tend to be relatively low, since the largest frequency within these pillars are green (5-10 m/s.) The wind speed tends to be highest when the wind direction is between south-west (210) and west (270). The wind blows less frequently towards the north (0) and east (90), and the wind speed also tend to be relatively low in this direction.

NWP100: Explanations from NWP100´s wind rose is more or less the same as the explanations provided for NWP200.

NWP500: The wind direction tends to have highest frequency towards 150 and 270 degrees. The bins with highest wind speed tend to be in the same direction as the other NWP-models, between 210 - and 270 degrees. The wind blows less frequently towards 0 and 90 degrees, and the wind speed tend to be very low in this direction.

4.7 Correlation analysis

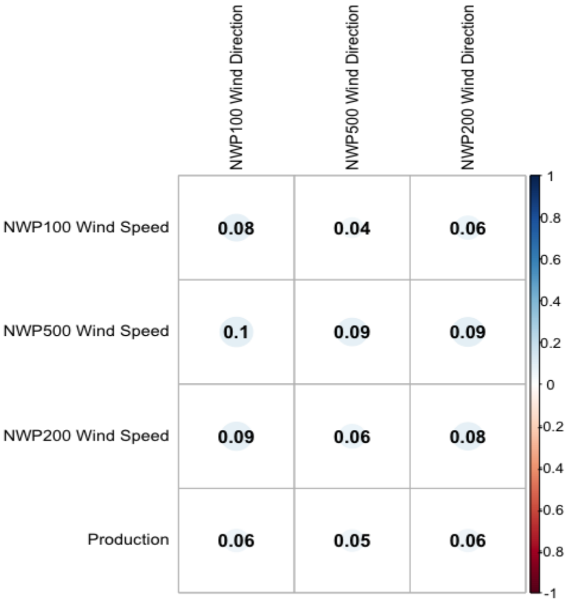
A correlation analysis is executed to support our choice of the SHAP estimation method. From 3.5.5, we know that explanations from the original KernelSHAP method are potentially inaccurate if features are dependent on each other. The correlation analysis needs to be adjusted compared to a traditional analysis using the typical *Pearson-correlation*. The reason for the adjustment is the circular variables used in our model. The correlation analysis is split into three different parts, where each part depends on the linear or circular characteristic of the variables measured. The three parts use three different methods to calculate the correlation coefficient, and they are the following:

Correlation between two linear variables: Traditional statistical correlation analysis, where *Pearson correlation is* used as the correlation coefficients. The correlation coefficient has a range from -1 to 1.

Correlation between two circular/angular variables: To measure the correlation between two angular variables we use a method developed by Jammalamadaka & Sarma (1988). In the computation, $(x_i - \bar{x})\ and\ (y_i - \bar{y})$, from traditional correlation, are substituted with $sin(x_i - \bar{x})\ and\ sin(y_i - \bar{y})$. The correlation coefficient ranges from -1 to 1.

Correlation between a linear- and circular variable: The correlation between wind direction and wind speed or production is found by a parametric method developed by Mardia (1976). The correlation coefficient ranges from 0 to 1. The mathematical computation of this method is explained by Lototzis et al. (2018).

4.7.1 Correlation between a linear and circular variable



	NWP100 Wind Direction	NWP500 Wind Direction	NWP200 Wind Direction
NWP100 Wind Speed	0.08	0.04	0.06
NWP500 Wind Speed	0.1	0.09	0.09
NWP200 Wind Speed	0.09	0.06	0.08
Production	0.06	0.05	0.06

Figure 5: Correlation between linear and circular features

We see that there is a very weak or no correlation between wind direction – and wind speed forecasts. There is also a very weak or no correlation between wind direction forecasts and the wind power produced at Ytre Vikna.

4.7.2 Correlation between two linear variables

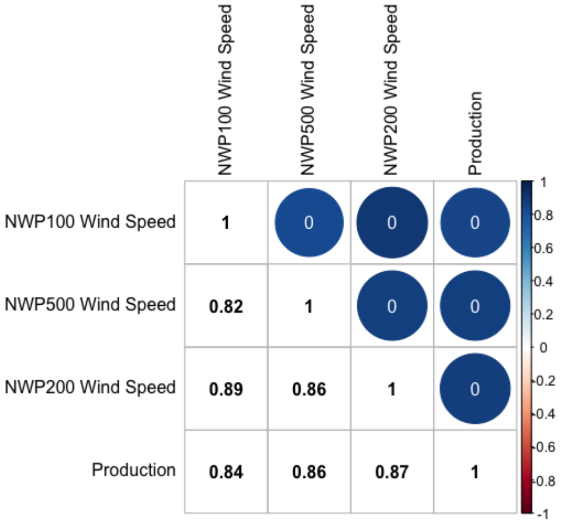


Figure 6: Correlation between linear features

From these numbers, we can conclude that all linear relationships in our model have a strong positive relationship. White numbers are P-values and low P-values signify statistically significant correlations.

4.7.3 Correlation between two circular variables

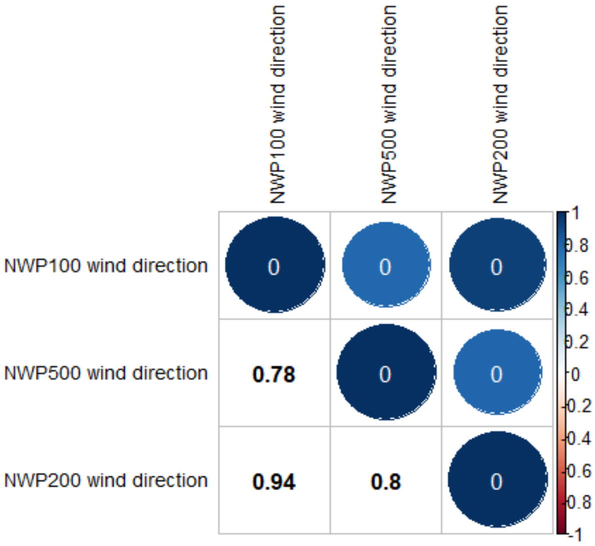


Figure 7: Correlation between circular features

There is a positive relationship between the wind direction forecasts. Wind direction forecasts tend to positively change with each other.

5. Methodology

TrønderEnergi needs to train two separate prediction models, one for the day-ahead market and another for the regulation-market which we refer to as the replan model. In order to make this thesis in a more general form, and without revealing corporate secrets we are not using their exact models. The objective of this thesis is not to produce the best model in terms of predictive accuracy, but instead to develop an explanation framework that works for any complex model. That includes the models used by TrønderEnergi. Hence, we will try to replicate TrønderEnergi´s model by using many of the same features, but with another algorithm. After training the model and using it to predict wind power based on the observations in the test data, we try to explain the individual predictions using our explanation model.

Most of the features used in predicting day-ahead production had missing values throughout the dataset which made it hard to train a model anywhere near TrønderEnergi's original. Because of this we focus on predicting wind power one hour before real production, or in particular for the regulating/replan market.

We refer to the appendix if you wish to see the R-implementations and libraries needed to make use of the different methodologies (see Appendix E).

5.1 Random Forest

When building our prediction model, we decided to use random forest. Random forest is a tree based method, that seeks to improve upon the basic decision tree by fitting multiple trees through bootstrapping and then averaging the results. This process of fitting multiple trees on bootstrapped datasets is known as bagging. Bagging and random forest are set apart from the fact that in the random forest approach, the trees are to some extent uncorrelated. The benefit of this is that the trained trees look less similar, which can happen if say one predictor is dominant. Since only a subset of the predictors is considered at each split, the dominant feature may not be considered at all in some cases and the trees will be less correlated and have a greater learning curve (James et al., 2013).

Bagged trees have high variance and low bias due to letting the trees grow deep without “pruning”, but the variance is reduced severely by averaging across the trees. However, this leads to very low interpretability which is what this thesis is all about.

To determine the degree of “decorrelation” we use the tuning parameter m to define the size of the predictor subset. The smaller m is relative to the number of predictors p, the less correlated the trees will be. If m equals p it will simply amount to bagging. A small m is typically helpful if the features are correlated, which they are in this instance. Typically, $m \approx \sqrt{p}$ gives a small edge compared to regular bagging (James et al., 2013), and that is the tuning parameter we choose to use. One must also decide on the number of trees to grow, but this is simply a computational choice as the number of trees grown will not lead to overfitting.

5.2 Shapley Addidative Explanations (SHAP)

In our analysis, the model-agnostic method used to explain individual predictions is the KernelSHAP approach. As mentioned in the theory section, models not based on Shapley values violates the properties of *local accuracy and consistency*. This is typical for the LIME approach, and it consequently leads to unintuitive behavior in certain situations (Lundberg & Lee, 2017). KernelSHAP provides a unique solution in the class of additive attribution methods.

In the correlation analysis, we displayed a strong correlation between multiple features. Wind speed predictions from different NWP models are highly correlated, and the same can be said about the different wind direction predictions. Dependency between features can potentially lead to incorrect explanations if the original KernelSHAP approach is used. Consequently, we choose to use the new version of KernelSHAP explained in section 3.3.7, which take dependency between features into account.

5.2.1 Method used to estimate features conditional distribution

Aas et. al (2019) conducted an experiment to analyze which method provides the most accurate approximation of the real Shapley values. The original KernelSHAP and the four different approximation methods proposed by Aas et. al (2019) were compared and measured against the true Shapley values. When features distributions were asymmetric, or skewed and heavy-tailed, the combined approach with *empirical conditional distribution (with bandwidth equal* *to 0.1)* and *multivariate gaussian distribution* outperformed the other approaches. By looking at the distributions to our features, we can say that they are asymmetric and skewed (see Appendix G). Hence, we assume that the above-mentioned combination performs best with our data as well. We therefore use the new version of KernelSHAP with a combined approach, applying the empirical approach for subset of dimension equal to 1 and the Gaussian approach otherwise.

This implies that when we are conditioned on one feature, we use the empirical approach to estimate feature´s conditional distributions, and to sample values for the absent features. When we are conditioned on 2-6 features, we use the gaussian distribution. A bandwidth needs to be set by the user when applying the empirical approach to estimate the conditional distribution. This choice is often shown as a bias-variance trade-off. Empirical analysis shows that bandwidth equal to 0.1 typically delivers sufficient results, and therefore we apply the same bandwidth in our estimation of Shapley values (Aas et al., 2019).

5.3 Clustering

After estimating the SHAP values we have an nxp matrix, where $n = number\ of\ observations\ in\ the\ test\ set$ and $p = number\ of\ features$. This matrix consists of each feature´s marginal attribution to its respective prediction for every observation in the test data:

	$\boldsymbol{\phi_1}$	$\boldsymbol{\phi_2}$	$\cdots$	$\boldsymbol{\phi_p}$
1	ϕ_{11}	ϕ_{12}	$\cdots$	ϕ_{1p}
2	ϕ_{21}	ϕ_{22}	$\cdots$	ϕ_{2p}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\boldsymbol{n}$	ϕ_{n1}	ϕ_{n2}	$\cdots$	ϕ_{np}

Figure 8: SHAP matrix

Based on this matrix we will identify groups with high prediction error and high production imbalance cost. In order to find groups with similar explanation structure, we use a technique within unsupervised learning called *agglomerative clustering*. This technique gathers observations from the matrix into separate clusters based on their similarity (James et al., 2013).

5.4 Hierarchical Clustering

To find distinct groups with similar characteristics we use agglomerative clustering, which is a common method within hierarchical clustering. This is a bottom-up approach where all observations start at the bottom of the hierarchy. The final hierarchy is illustrated as a dendrogram. From the bottom, where all observations are their own cluster, similar observations/clusters are fused as we move upwards. At the highest level of the hierarchy there is simply one cluster, including every observation (James et al., 2013):

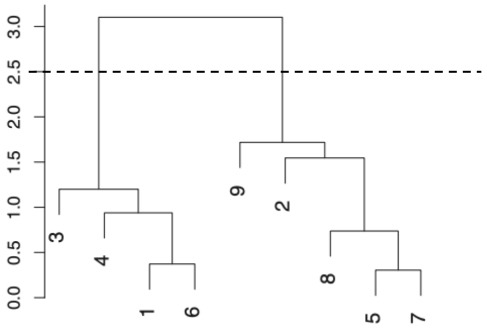


Figure 9: Example of dendrogram

In the example, there are nine different observations which are clustered together. A prespecified similarity measure decides which branches are combined. The height gives us information about the similarity of clusters. The higher the observations are fused, the less similar they tend to be. Final clusters are identified by cutting the tree with a horizontal line. Connected observations under this cut are grouped as single clusters. In the example, a cut at a height of 2.5 produces two distinct clusters.

5.4.1 Agglomerative clustering algorithm

Our dendrogram is given by the following algorithm (Hastie et al., 2009; James et al., 2013):

1. Compute the pairwise "distance" between observations in the dataset. For this purpose, we use the Euclidean distance:

$$d(x_i, x_{i'}) = \sum_{j=1}^{p} |x_{ij} - x_{i'j}| \tag{12}$$

The equation gives us the distance between observation i and $i^{'}$ in Euclidean space. This is computed pairwise for all n observations, producing a nxn distance matrix.

2. For $i = n,\ \ n-1,\ \ n-2, \dots, 2$

   I. Consider the dissimilarity between i clusters and find the pair of clusters having the lowest distance between them. This pair is then fused. The distance between the two combined clusters decides the height at which they are combined. In the first iteration, there are n clusters, and the distance matrix is used to find the pair of clusters with the lowest distance.

   II. Calculate the new pairwise inter-clustering distance between the $i - 1$ outstanding clusters. A new measure is needed at this point since we consider the dissimilarity between a cluster with more than one observation and a cluster with either one or multiple observations. We use the *average linkage approach*:

$$d(C1, C2) = \frac{1}{N_{C1}N_{C2}} \sum_{i \in C1} \sum_{i \in C2} d\left(x_i, x_{i^{'}}\right) \tag{13}$$

$$where\ N_{C1}\ and\ N_{C2}\ are\ number\ of\ observations\ in\ these\ clusters$$

After using the average linkage approach on every pair of clusters we have a $(i-1)x(i-1)$ distance matrix. We then fuse and repeat the process.

After clustering we need to choose the height at which we should cut the dendrogram. This cutting process decides the number of clusters obtained. In practice we often simply look at the dendrogram and choose a rational number of clusters. A rational choice means that the user should analyze the height on the dendrogram at which clusters are combined, and also the number of clusters preferred (James et al., 2013). In our situation, with almost 9000 observations, the dendrogram will be large and difficult to interpret. A possible way to cut it is proposed by Joseph Larmarange. He proposes an algorithm where the superior partition to cut a dendrogram is decided by the higher relative loss of inertia criteria (Larmarange, 2020).

6. Analysis

Our analysis is sectioned into global explanations, local explanations and grouped explanations. The basis for all the different scopes of explanations is the estimated SHAP values. To recap, the SHAP value is a feature´s marginal contribution to any given prediction, relative to the baseline. The baseline being the expected value of production.

6.1 Global Explanations

In this section we have aggregated the SHAP values for every observation in the test set and we aim to interpret them "at once" or in an a more global view. This provides especially two valuable insights. First, we are able to see what features that is most important in predicting wind power. Second, we are able to see the trained relationships between the response variable and features. From these explanations we are able to see if the trained relationships between features and the response variable are consistent with our prior beliefs. Specifically, we can see if they are consistent with the fundamental characteristics of wind turbines, such as their power curves.

6.1.1 SHAP Summary

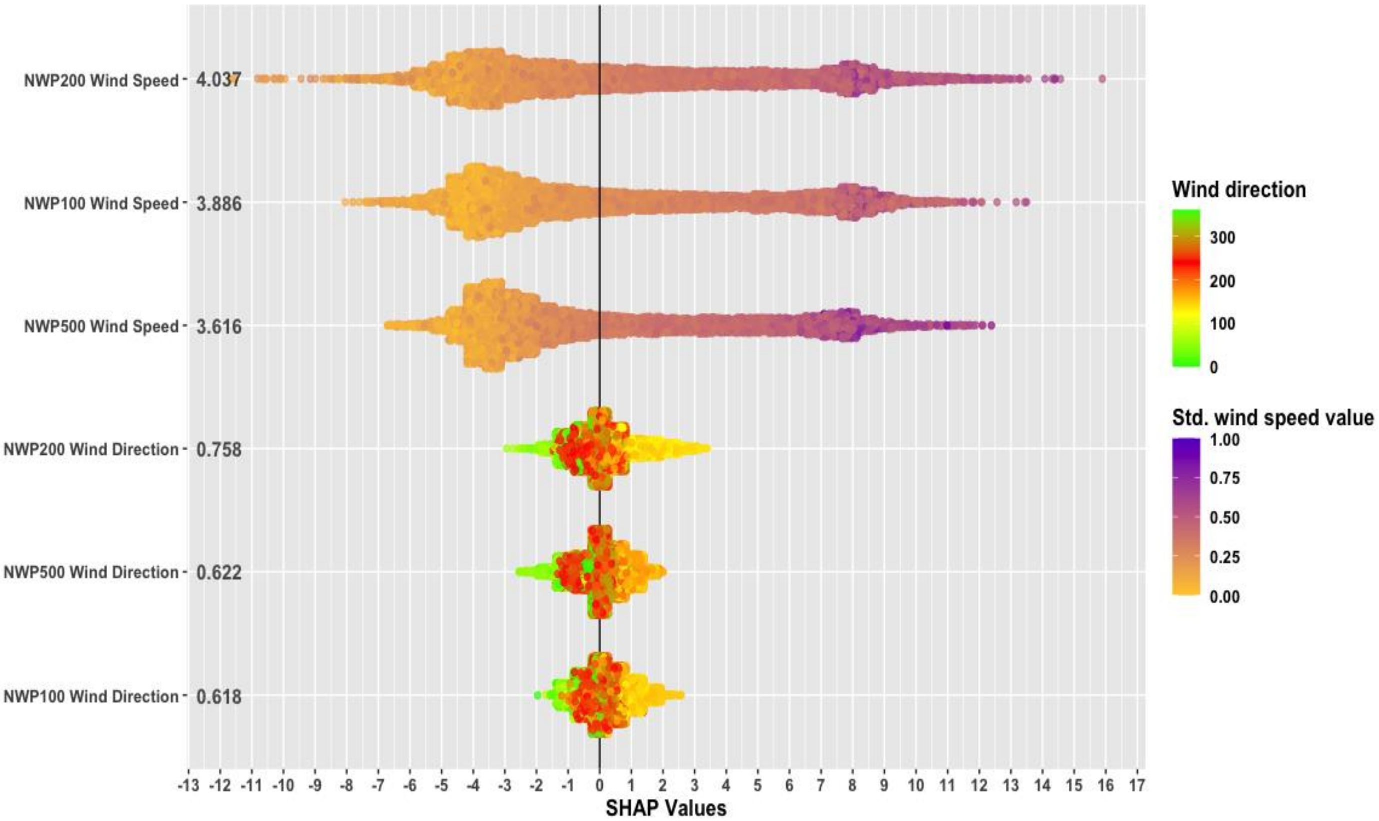


Figure 10: SHAP summary plot

The SHAP summary plot shows the replan prediction model´s feature importance combined with each feature effects. The x-axis measures the SHAP-value, and the y-axis the different features. The black number beside a feature name is the mean absolute Shapley value for the corresponding feature. Each dot represents the SHAP-value for an individual observation and specific feature. For wind speed features, the color represents high or low wind speed values. This value is standardized (0 to 1). For the wind direction features, the color represents the wind direction in degrees. Green is represented at both 360 and 0 degree since this is the same wind direction.

The SHAP summary plot grants a good overview of our model´s overall behavior. From the mean absolute SHAP-values we can tell that wind speed features have the greatest impact on the predictions. NWP200 wind speed is the most impactful feature. This means that NWP200 wind speed has the largest marginal contribution to the predicted output on average. On the other side, NWP100 wind direction has the smallest marginal contribution to the predicted output on average.

High wind speed values are associated with high and positive contribution to the predicted output as high wind speeds tends to lead to higher SHAP- values. Low wind speed values are the opposite, typically seen at the lower spectrum of SHAP values.

The yellow color for wind direction represents a range of approximately 100-200 degrees which roughly equates to a range from the east to south-west. When an observation is in this range, it is associated with a positive, but weak contribution to the prediction. The green color represents the range of approximately 300 to 100 degrees which roughly covers the northern sphere. When an observation is in this range, wind direction is often associated with negative and weak contribution to the prediction. The wind roses in section 4.6 showed that in this range it tends to blow less frequent and at low wind speeds. The low wind speeds observed within this range is a possible explanation for why observations in this range tends to be followed by negative contribution from wind direction on the predicted output. The red color represents a range of 200 to 300 degrees. When an observation is in this range, wind direction often has no or a very weak negative contribution to the prediction. The wind roses showed that the fastest wind speeds tend to be observed most frequently when it blows in this direction. Therefore, it might be strange that this wind direction range, globally contributes less to wind power predictions. The potential explanation for this is discussed in the SHAP-dependence section. The outlined relationship for wind direction is consistent for all the different NWP-models.

6.1.2 SHAP Dependence – Wind Speed Predictions

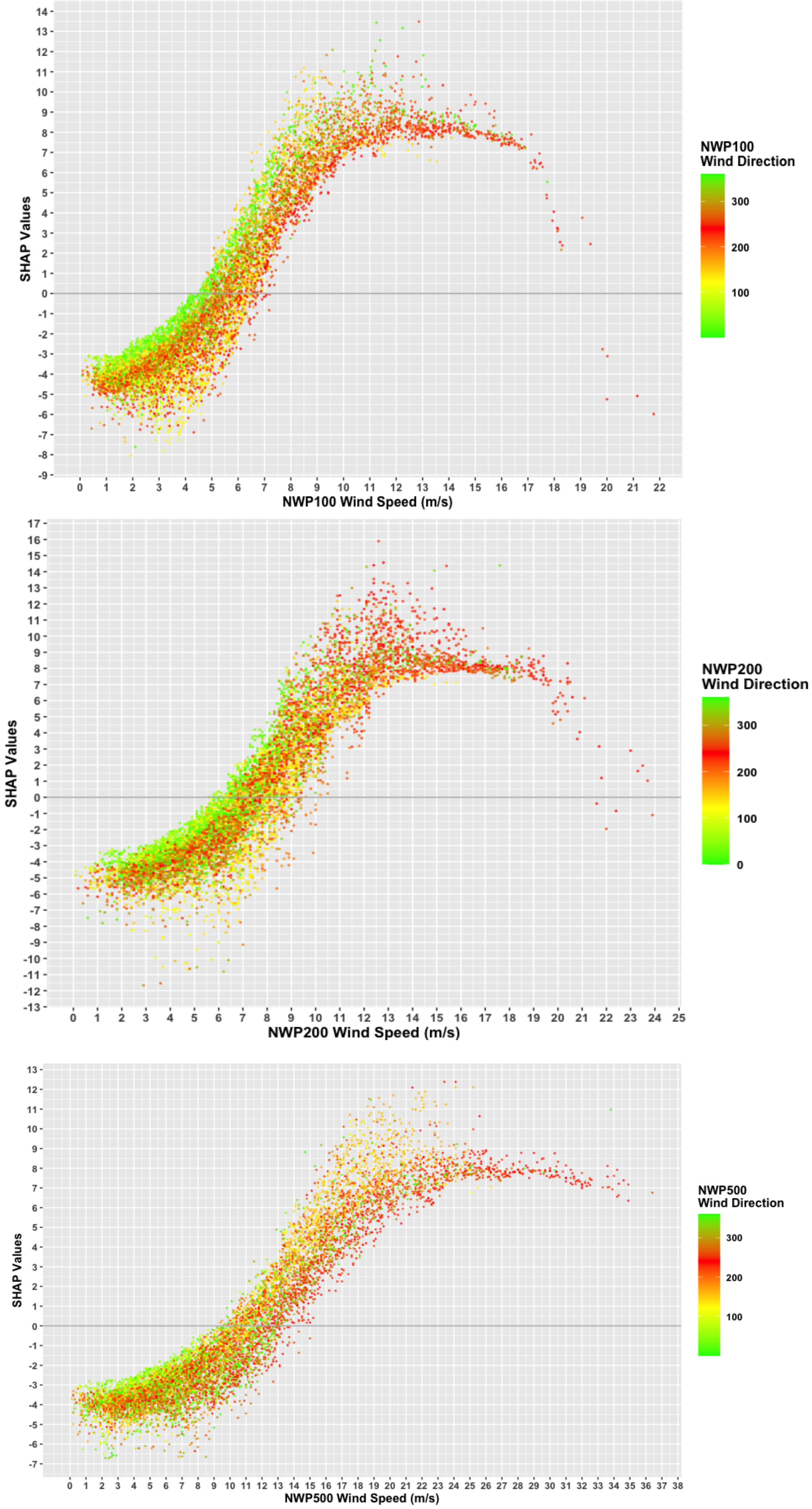


Figure 11: SHAP dependence plots Wind Speed

The SHAP dependence plots display NWP-models wind speed forecast plotted against their SHAP value. SHAP-values on the y-axis, the wind speed values (m/s) on the x-axis. The colors correspond to the same observation’s observed wind direction value. The overall pattern is similar for all three NWP-models. By first ignoring the colors, we can explain wind speeds effect on the predicted output. The following interpretation considers NWP200´s dependence plot, figuratively split into four “regions” based on the level of wind speed:

The first region is from 0 m/s towards 3 m/s approximately. Here, wind speed´s effect on the predicted output tend to be negative and stable at a SHAP value of -5. This means, that low wind speed values tend to contribute by reducing predicted production by 5 MW from the expected production at 12.307 MW. This region is heavily tied to the wind turbines cut-in speed.

The second region is from 3 m/s to 12 m/s. Here, wind speed´s contribution to the predicted output tend to increase in a positive and linear fashion. At around 7 m/s the SHAP-values crosses zero, signifying the limit at which wind speeds starts to positively contribute to the predicted output, compared to the baseline. Given the steep trend in this region, we can see that small changes in the wind speed leads to big changes in the effect on predicted output, implying a high elasticity. Fluctuations in the NWP-models wind speed forecast in this region will likely have a greater effect on predicted production. Hence, inaccurate wind speed forecasts in this region will have a great impact on the wind power prediction model and increase the chances of large prediction errors.

The third region is from approximately 12 m/s towards 20 m/s. Wind speed´s contribution to the predicted output tend to be positive and stable, with a SHAP-value around 8. This means, high wind speeds usually contribute with 8 MW relative to the baseline. The SHAP-values stop increasing as the turbines maximum capacity is getting closer.

The fourth and last region is from 20 m/s and beyond. Wind speed’s contribution to the predicted output drops off steeply. This is probably caused by the wind turbines cut-off speed.

The other NWP’s dependence plots can be viewed in the same light. However, NWP100 and NWP500 start increasing from the cut-on speed at an earlier and later wind speed values, respectively. They also reach the maximal capacity at an earlier and later wind speed values. This is probably due to the NWP-models different height profile, discussed in section 4.3.1. Wind speed´s SHAP dependence plots, and the different regions, shows our prediction model ability to detect and estimate the different relationships and patterns during the power curve. In other words, the plots confirm that the random forest model is able to estimate the correct relationship between wind speed and wind power produced.

Interaction effects arises when the effect of one feature depends on the value of another feature (James et al., 2013). When the wind direction is in the range of 0-100 degrees and 300-360 degrees, the effect of wind speed seems to be more positive than other wind direction ranges. This is especially clear in the dependence plots of NWP200 and NWP100, and particularly when the SHAP values are below zero as illustrated by the horizontal line.

6.1.3 SHAP Dependence – Wind Direction Predictions

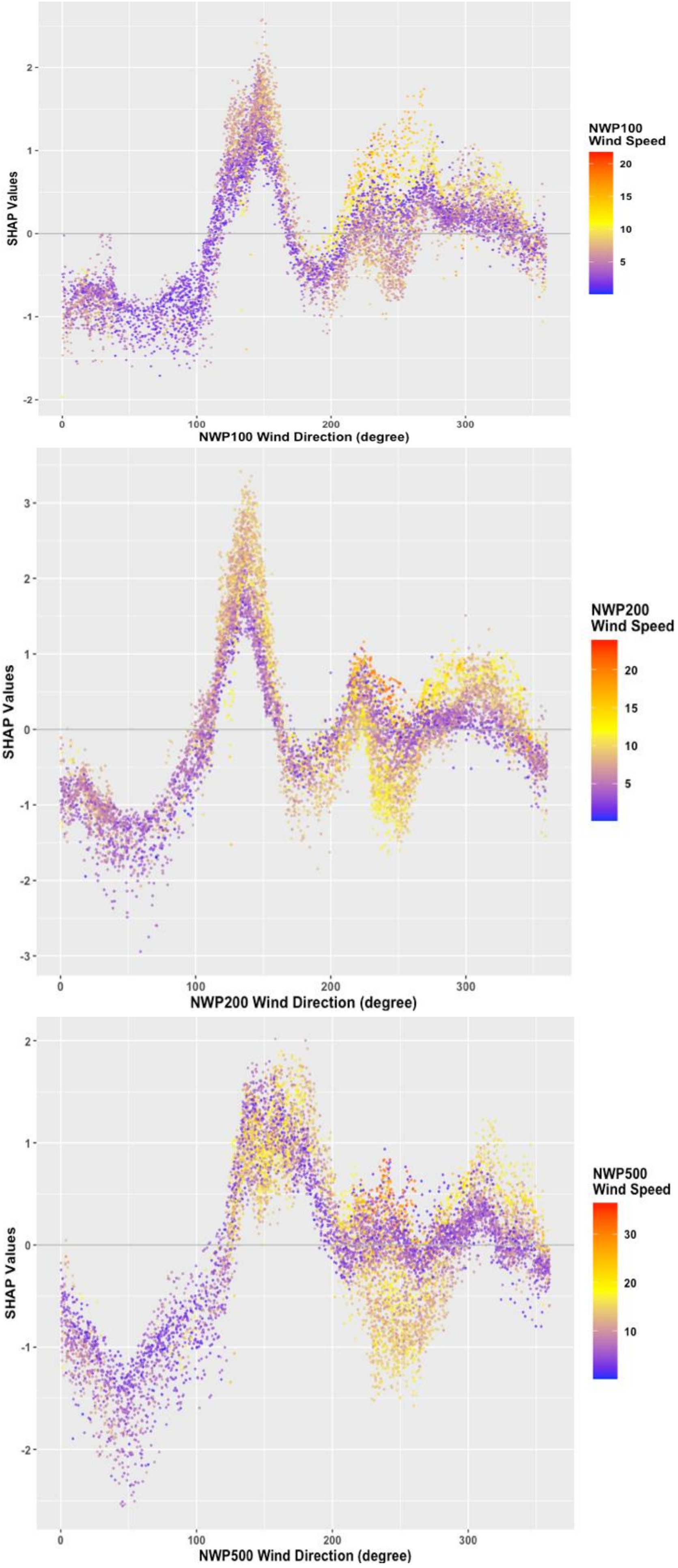


Figure 12: SHAP dependence plots Wind Direction

These SHAP Dependence plots display the SHAP values of wind direction features plotted against the wind direction. The colors correspond to the same observations wind speed values, this time not standardized.

From north to east (0-100 degrees), wind direction tends to contribute negatively to the predicted output. Between south-east and south (100-180 degrees), the SHAP-values spike upwards, so that the wind direction has a relatively high and positive contribution to the predicted output. From south to west (180-270), the SHAP-values drop off again. Within this range there is a great variability in the SHAP Values. Wind direction most often have a small, negative effect on the predicted output. However, there is a large number of observations that seemingly has no contribution and a great deal that also exhibits a low and positive SHAP-values. A possible explanation for the large variance is interaction effects with the wind speed features. From west to north (270-360 degrees), wind direction tends to either have no effect or a slightly positive impact on the predicted output and relative to the baseline.

Interaction effects are in place if the SHAP-values for wind direction depends on the wind speeds values. Potentially, we see two interaction effects in the dependence plots. The first can be seen when wind direction ranges from 100 to 180 degrees and 270 to 360 degrees. In these ranges, higher wind speeds, tend to coincide with high SHAP values for wind direction. This interaction effect occurs where yellow observations (high wind speed values) tend to systematically be above the purple observations (low wind speed values). It is particularly strong in the NWP200 and NWP100 dependence plot.

The second interaction effect can be viewed in the range from 180 to 270 degrees. In this range, higher wind speed values tend to decrease the SHAP-values of wind direction. This occurs where yellow observations tend to lay below the purple observations. At first, the second interaction-effect seems odd, but there may be a good explanation for this. Akram (2014) analyzed the deviations between predicted and actual production levels at Ytre-Vikna. One of Akram’s findings was the reason for why production tends to be lower when it blows towards the south-west compared to the south-east, given equal wind speed values. The author concluded that a potential reason for this is the terrain and the placement of turbines. From the south-west, most of the turbines are placed in a straight line with short distances in between, typically 250-350 meter. This causes wake effects and turbulence. Wake effects are the decreased wind speed that occur as a result of the wind hitting/passing a previous wind turbine. Wake effects lead to lower production of wind power, even with high wind speeds in the area.

From the dependence plots of wind speed, we know that high wind speed levels are associated with a large and positive SHAP-value for wind speed. However, wind direction tends to counteract this positive effect and decrease the overall prediction when the wind direction is between 200-270 degrees and the wind speeds tend to be relatively high. This interaction effect seems to take the wake effect into account when it blows towards the south-west (200-270 degrees), and subsequently lower the predicted production.

When the wind blows in the range between south-east and north-east, the terrain is not as homogeneous or flat as it is for the south-west. There is larger variation of heights when the wind blows in this range. The heights at the wind park can be very steep when the wind blows in the south-east direction (Akram, 2014). Wind turbines placed on the top of heights tend to experience higher power output and lower turbulence than turbines placed on flatter ground. This is due to the speed-up effect (Eilenberg, 2012). That is, when the wind moves upwards steep heights, the wind speed increases when it reaches the top of the height. This might be the explanation of why wind direction’s effect on the prediction increases when the wind direction is between 100-180 degrees.

6.2 Local Explanations

In this section, we use the SHAP-values to explain individual observations. The focus of this section will be to explain the observation in the test data with highest prediction error, and the observation with highest production imbalance cost. The force plot used to explain these individual observations, is applicable for any observation a user wants to get more insight into, here we simply examine two observations of interest. In the plots below, and the cluster analysis, WS and WD are abbreviations for wind speed and wind direction, respectively.

6.2.1 Explaining the individual observation with highest prediction error

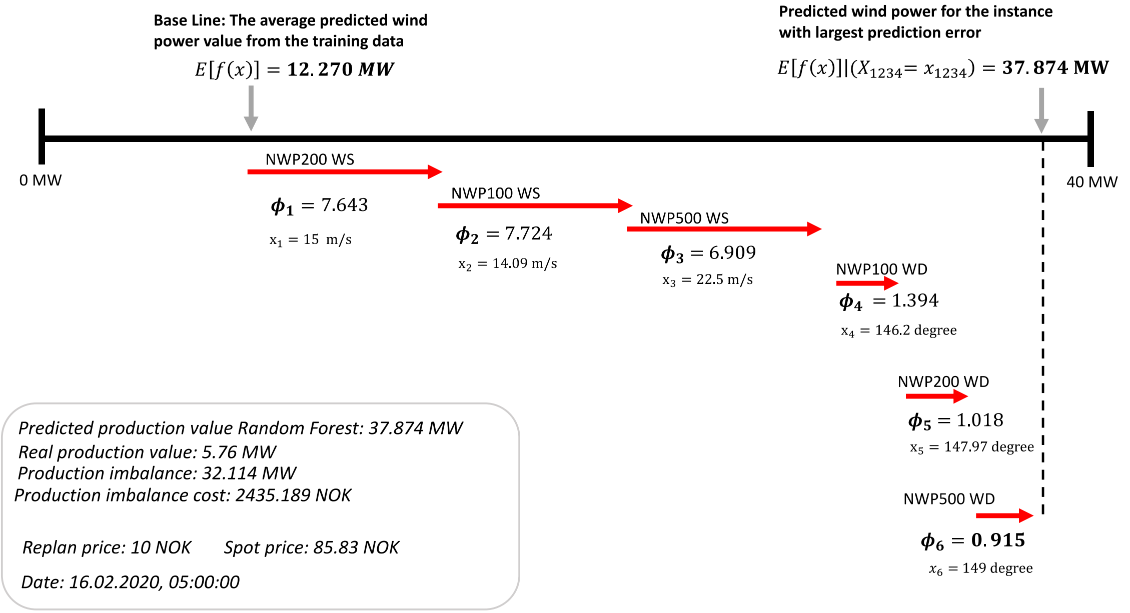

Base Line: The average predicted wind power value from the training data
$E[f(x)] = 12.270\ MW$
Predicted wind power for the instance with largest prediction error
$E[f(x)]|(X_{1234} = x_{1234}) = 37.874$ MW
0 MW
40 MW
NWP200 WS
$\phi_1 = 7.643$
$x_1 = 15$ m/s
NWP100 WS
$\phi_2 = 7.724$
$x_2 = 14.09$ m/s
NWP500 WS
$\phi_3 = 6.909$
$x_3 = 22.5$ m/s
NWP100 WD
$\phi_4 = 1.394$
$x_4 = 146.2$ degree
NWP200 WD
$\phi_5 = 1.018$
$x_5 = 147.97$ degree
NWP500 WD
$\phi_6 = 0.915$
$x_6 = 149$ degree
Predicted production value Random Forest: 37.874 MW
Real production value: 5.76 MW
Production imbalance: 32.114 MW
Production imbalance cost: 2435.189 NOK
Replan price: 10 NOK
Spot price: 85.83 NOK
Date: 16.02.2020, 05:00:00

Figure 13: Explanation of the individual observation with highest prediction error

The plot shows all features marginal contribution to the individual prediction with the highest prediction error. ϕ correspond to the features´ SHAP values, and x correspond to the features´ actual values. $E[f(x)]$ is the expected value when no features are contributing to the prediction. This is simply the average of all predicted values in the training data and referred to as the baseline. Each feature´s arrow is red and points to the right, which indicates that they all have a positive attribution to predicted production, increasing it from the baseline. All three wind speed features have a large, positive effect on the predicted output, and together, these features increase the predicted output with 22.276 MW relative to the baseline. All three wind direction features also have a positive effect. They increase the predicted production with 3.327 MW from the baseline. This leads to the final prediction of 37.874 MW.

The prediction of 37.874 MW is very far from the real production of only 5.76 MW. This leads to a production imbalance of 32.114 MW. TrønderEnergi is required to purchase this difference at regulating markets. The spot price is higher than the RK-price, which leads to a production imbalance cost. The production imbalance cost is 2435.19 NOK, which is significantly larger than the average production imbalance cost of 53.95 NOK. The cost is magnified by a large spread between the spot- and RK-price. TrønderEnergi sells the production imbalance of 32.114 MW to a significant lower price (10 NOK) than they are required to purchase it for (85.83 NOK) in order to maintain balance.

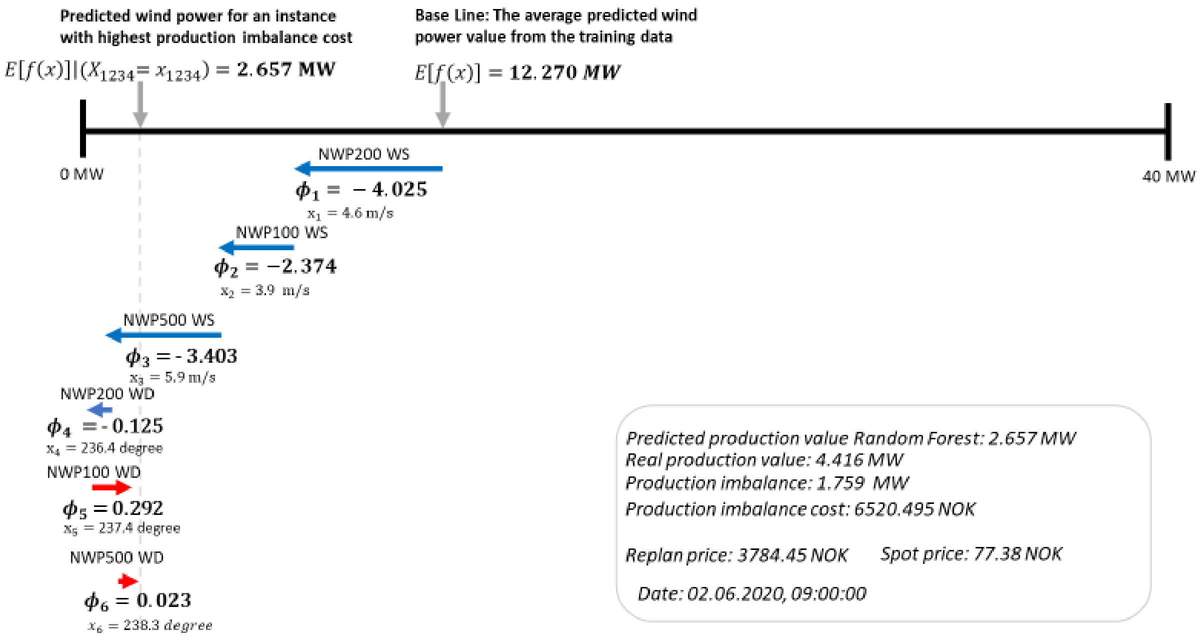
In general, three reasons can explain why large production imbalances or prediction errors occurs. Firstly, NWP- model's predicted wind speed and wind direction values can deviate heavily from their real values. Secondly, weaknesses in the training data which subsequently leads to poor models. Lastly, the local power grid can be out of service or there is unscheduled maintenance on one or multiple wind turbines (Akram, 2014).

The force plot shows that the three wind speed features have the largest contributions to the prediction. 86.7 % of the total positive increase from the baseline stems from wind speed features. If they deviate heavily from the real wind speed values that hour, it subsequently leads to a poor prediction. The three NWP-models' predictions are measured at different heights, and it was therefore difficult to collect representable real wind speed data for the time period. Consequently, it is difficult to measure the predicted wind speed values accuracy, and if they are the main reason for the large prediction error. According to Akram (2014), predicted wind speed values deviates heavily from their real values when the wind direction points towards the south-east direction at Ytre Vikna. In this case, all three NWPs tell us that the predicted wind direction is in the interval 146.2-149 degrees, or in the south-east direction. If Akram´s analysis and the wind direction predictions are consistent, there is great chance that the predicted wind speeds deviate from their real values. This may partially explain the large prediction error.

The SHAP dependence plots indicates that the relationships estimated by the random forest model, are in accordance with the prior understating of how wind speed and wind direction impact power production. Hence, weaknesses in the training data seems to not be a factor contributing to the large prediction error. According to data received from TrønderEnergi, the available power from the turbines at the observation´s time, 16.02.20 05:00:00, were at maximum capacity. This means, there were no unscheduled maintenance at this hour. As a result, the main contributor to the large prediction error seems to be overestimated and wrongly predicted wind speed values.

NWP200, NWP100 and NWP500 predicted wind speeds of 15 m/s, 14.09 m/s and 22.5 m/s, respectively. In the SHAP dependence plots of wind speed, these are all in region three, were the contribution is large, but stable. Wind speed features contribution to predicted production seems to be too large when looking at the hour´s real production level. In other words, NWP-models´ predicted wind speeds seems to be overestimated, and they will therefore have a large positive effect on the predicted production which leads to an overestimation.

6.2.2 Explaining the individual observation with highest production imbalance cost



Predicted production value Random Forest: 2.657 MW
Real production value: 4.416 MW
Production imbalance: 1.759 MW
Production imbalance cost: 6520.495 NOK

Replan price: 3784.45 NOK *Spot price: 77.38 NOK*

Date: 02.06.2020, 09:00:00

Figure 14: Explanation of the individual observation with highest prod. imbalance cost

All three wind speed features have a negative contribution to the predicted production, and together, these features decrease the predicted output with 9.802 MW relative to the baseline. NWP200 wind direction has a negative contribution and the two other wind direction features have a positive contribution to the predicted output. Together, they increase the predicted output with 0.19 MW.

Compared to the real production level of 4.416 MW, the predicted output of 2.657 MW is not too far away. NWPs wind speeds and wind directions seems to be somewhat accurate for this observation. However, it is here we find the largest production imbalance cost out of all observations in the test set. The main reason for this is the wide gap between the RK price and spot price. The prediction is an underestimate of the real production level, and the RK-price is larger than the spot price. From section 2.4, we know that this is what creates *production imbalance cost 1*. TrønderEnergi can sell their surplus power of 1.759 MW to the spot price 77.38 NOK. However, if the replan prediction model had been better and predicted perfectly, they could have sold this amount at the RK-price of 3784.75 NOK. This is why the production imbalance cost is so large for this observation, although the prediction error is relatively small.

In general, the spot price and RK-price tend to follow each other, and they have a correlation coefficient[4] equal to 0.87. However, the RK-price has more significant positive and negative spikes throughout the time period, and also a higher standard deviation (see Appendix B). This implies that the RK-price fluctuations from its mean are greater than the spot price fluctuations, and therefore will wide gaps occasionally occur.

[4] Pearson Correlation

6.3 Cluster analysis – Identifying groups with high prediction error and high production imbalance cost

In this section we use the SHAP values to create clusters and see if they can identify explanation structures with higher prediction errors and production imbalance costs, than others. As explained in section 5.3, the clusters are based on the test data and each observation´s SHAP-values. This implies that we cluster observations based on their explanation similarity (Molnar, 2019). An issue that often comes into play when dealing with clustering, is the question if you should standardize your features in order to get them on an equal scale before the clustering process (James et al., 2013). SHAP-values are all using the same scale or unit. Therefore, we do not face the issue of standardizing. This is one the benefits of using SHAP-values in a clustering analysis (Molnar, 2019; Niemerg, 2020).

6.3.1 Three clusters - Results

According to the Larmarange's (2020) algorithm discussed in section 5.4.1 and the higher relative loss of inertia criteria, the best partition to cut our dendrogram was by creating three distinct clusters.

Cluster	**Mean ϕ NWP100 WD**	**Mean ϕ NWP500 WD**	**Mean ϕ NWP200 WD**	**Mean ϕ NWP100 WS**	**Mean ϕ NWP500 WS**	**Mean ϕ NWP200 WS**	**RMSE Prediction Error Random Forest**	**RMSE Prediction Error TE**	**Mean Production**
1	0,067	-0,020	0,040	-1,983	-1,576	-1,216	3,827	3,938	7,090
2	0,442	0,179	0,415	6,220	4,360	5,081	5,681	5,706	28,044
3	0,277	0,138	0,027	10,793	5,541	-7,917	10,117	10,654	22,776

Table 3: Mean SHAP Values (ϕ), Prediction Errors and Production

In table 3 we see the average SHAP-value for each feature and each cluster. Additionally, the table shows an overview of the clusters forecast accuracy. The measure used to compute each cluster´s forecast accuracy, is the root-mean-square of error (RMSE). The predictions for the observations within each cluster are all based on the same data, and RMSE is therefore the preferable measure since it is also in the same scale as our predictions (Hyndman & Koehler, 2006). In this context, RMSE is the predictions average deviation from production in terms of MW. RMSE is computed for our model (Random Forest) and TrønderEnergi´s model (TE).

Cluster	Mean NWP100 WD	Mean NWP500 WD	Mean NWP200 WD	Mean NWP100 WS	Mean NWP500 WS	Mean NWP200 WS
1	206,187	217,814	197,142	3,876	7,821	5,910
2	234,619	235,947	232,001	9,519	17,604	11,578
3	212,270	235,493	193,675	10,263	17,445	6,560

Table 4: Mean Numerical Weather Predictions – m/s and degrees

In table 4 we have the mean real values of each feature. For wind speed features, this means the average value in terms of *m/s*, and for wind direction it means the average value in terms of *degrees*. The average wind direction values are calculated using the *resultant vector average wind direction method*, presented by Grange (2014).

Cluster	Mean RK-price	Mean spot price	Mean total production imbalance cost	Max total production imbalance cost	Mean monetary loss	Number of observations	Number of under-estimations	Number of over-estimations
1	180,412	184,560	43,554	6520,495	29,326	6405	2285	4120
2	170,143	188,516	79,385	4309,147	76,444	2539	1341	1198
3	80,451	108,464	156,052	963,422	65,817	20	10	10

Table 5: Mean Economical Variables – Prices, Prod. Imbalance Cost and Monetary Loss

Table 5 shows economic variables that impacts TrønderEnergi´s net income. First, we see the average RK-price and spot price. Then comes the production imbalance cost presented in section 2.4, we include the mean production imbalance cost, max production imbalance cost, and average monetary loss. At the end we have the number of observations in each cluster, the number of underestimations (surplus) and the number of overestimations (deficits).

6.3.2 Three clusters - Analysis

From the SHAP summary plot in section 6.1.1, we know that the wind speed features typically have the largest contribution to the prediction. This is also seen in table 3, for all of the three clusters. Predictions with a negative contribution on the predicted output from NWP200 wind speed, and a positive contribution from NWP100 and NWP500 wind speed, tend to have inaccurate predictions. This is seen for *cluster three* in table 3, where the RMSE equals 10.117

MW. When wind speed features effect on the predicted output pulling in the same direction, predictions tend to be better when measured in terms of RMSE. In *cluster two,* all wind speed features have a positive contribution to the predicted output on average. The RMSE for *cluster two* is 5.681 MW. This is almost half of that of *cluster three*. In *cluster one,* all wind speed features have a negative contribution, and RMSE equals 3.827 MW. The predictive accuracy therefore seems superior for observations where the NWP-models pull in the same direction. In other words, the model is more accurate when there is a low degree of variability in the weather forecasts. Further, it may imply that the model is more accurate when it is less windy, so that the contributions from wind speed features are negative.

Table 4 gives a more intuitive interpretation about the cluster's predictive accuracy. In order to not create any confusion in the discussion, when referring to high or low wind speeds, we need to consider NWP-models different height profile, as discussed in section 4.3.1. This means, a high estimated wind speed value for NWP100 does not necessary imply a high estimated wind speed value for NWP500, where the wind speed is estimated at a higher height profile. In *cluster three,* NWP200´s average wind speed value is low, while NWP100 and NWP500 average wind speeds are relatively high. We can say that there is a "disagreement" between the three NWP models in terms of whether the wind speed is high or low. In these situations, the model tends to predict worse compared to cluster one and two, where the NWPs are closer to each other. In *cluster two,* the wind speed features have less spread on average. In these cases, the RMSE value is half of that seen in cluster three where the spread is greater. In *cluster one,* there is a much smaller spread between the NWPs. When this is the case, the model´s predictions tend to be better in terms of RMSE.

A potential reason for why *cluster one* is superior to *cluster two* in terms of RMSE is that the NWP-models tend to exhibit low wind speeds in *cluster one*. These values are at end of the flat section at the beginning of the power curve (section 4.3.1). Under- or overestimated wind speeds are not affecting the prediction too much at this stage of the power curve. As a result, NWP-models with a low predictive accuracy will not affect our prediction model´s accuracy significantly, as long as the wind speeds are low. On another side, the NWP-models have a higher wind speed value in *cluster two* on average. These values are clustered around the region of the power curve with much greater production elasticity. Under– and overestimations of wind speeds changes the predicted production level to a much greater extent in this region of the power curve. This leads to a larger deviation between predicted and real production than seen in *cluster one,* and subsequently a larger RMSE value. This conclusion is consistent with

Akram (2014) analysis of TrønderEnergi´s prediction models´, and their errors at Ytre Vikna in 2014. The same explanations could be achieved by looking at wind speed features SHAP dependence plot. By looking at the SHAP dependence plots in section 6.1.2, we see much of the same. The elasticity is much smaller in region one compared to region two, and so changes in wind speed leads to smaller changes in the predicted output for region one. The high elasticity in region two leads to a higher likelihood of great prediction errors when the wind speed is in region two.

Table 5 shows that the average production imbalance cost for the observations within *cluster one, two* and *three* are 43.554 NOK, 79.385 NOK and 156.052 NOK, respectively. *Cluster three* has the largest production imbalance cost on average. This implies that observations with a similar explanation structure to *cluster three,* tend to produce larger production imbalance costs. In general, the size of the production imbalance cost depends on three factors. First, whether the prediction is an over- or underestimation relative to the real production level. Second, the relationship between the spot– and RK-price. Third, the size of production imbalance/ prediction error.

The average spot price is greater than the average RK-price for all three clusters. From section 2.4, we know that when the spot price is greater than the RK-price, it is not beneficial to overestimate production levels. The proportion of overestimated observations is approximately 50 % for *cluster three and two*, and roughly 60 % for *cluster one*. The spread between the spot – and RK-price is larger for *cluster two* and *three* on average, which means it will be extra costly to overestimate production for these clusters. From table one, we know that *cluster three* has the largest prediction error on average, *cluster two* has the second largest, *and cluster one* has the smallest.

A greater RK–price relative to the spot price, a great number of overestimated observations, and a large prediction error are all reasons for why *cluster three* has largest average production imbalance costs. When NWPs wind speed predictions “disagree” about the wind speeds, the prediction tend to be inaccurate, and consequently lead to higher average production imbalance costs. Luckily, there is only twenty observations throughout the test data that exhibits this explanation structure.

Low spread between the RK-price and the spot price and a small prediction error are the main reasons for why *cluster one* has the smallest average production imbalance cost. This implies, that when NWPs wind speed predictions all have a negative effect on the predicted output, or they simply predict low wind speeds, the prediction accuracy tends to be better, and the production imbalance cost is low. However, this is only based on the average of all the observations in the cluster. From table three, we see that *cluster one* actually has the observation with the largest production imbalance cost. This is the observation explained in section 6.2.2. We know that the prediction error was low for this observation. However, the RK-price was at the greatest, the prediction underestimated the real production level, and the spread between the replan – and spot price was extremely wide. Due to the difference between the prices, the production imbalance cost became very large, even though the prediction error was low. This implies, that observations with similar explanation structure might vary greatly in terms of production imbalance costs. This mainly occurs when the difference between the spot- and RK-price is particularly great. This spread is controlled through supply and demand in the power markets and is therefore outside of TrønderEnergi´s control. Extreme situations such as this will occur now and then and should be expected as long as it is not possible to perfectly predict production levels.

We would also like to point out that the choice of where to cut the dendrogram is based on a simple rule in this cluster analysis. However, there are different suggestions to decide where to cut a dendrogram. Instances of such suggestions are the Elbow, Silhouette and Gap statistical method. There is no common census what the right cut-off is, and the choice is often situation-based. The objective is to find clusters that provide us with useful information about observations typical explanation structure. Hence, for our objective, it might be useful to analyze a larger number of clusters as well. In the appendix we show the results of using six clusters, however, the general findings are very much the same.

7. Discussion

7.1 The Importance of Interpretable Machine Learning for Trønder-Energi

According to Molnar (2019), interpretability is essential when a prediction model has a significant impact on the stakeholders. TrønderEnergi has multiple prediction models that are trained to forecast the upcoming wind power production at their wind parks around Norway. As we have seen, large prediction errors tend to have a significant impact and may lead to poor outcomes for TrønderEnergi. Poor in the sense that they increase the company´s production imbalance cost, which subsequently leads to a negative impact on TrønderEnergi´s bottom line. Interpretability and understanding why a model behaved in a certain way, is important due to the models´ economic impact. Stakeholders might require explanations for why prediction errors and imbalances occur, and why they fluctuate over the course of a year. Interpretable machine learning and model-agnostics methods can help to answer these questions and explain why the prediction model behaved in a certain why.

In a recent report from PwC, Oxborough et al. (2018) discusses the main drivers of whether interpretability is crucial to implement or not. They are:

1. The potential economic impact of a single prediction.
2. The economic utility of understanding why the prediction was made with respect to the choice of actions that may be taken as the result of the prediction.
3. The economic utility of the information gleaned from understanding trends and patterns across multiple predictions.

TrønderEnergi´s energy management and bidding process in the day-ahead market and balancing market are based on their prediction models. In section 6.2, we saw that a single prediction can create a significant production imbalance cost. The total production imbalance cost for TrønderEnergi throughout the one year of test data is 483 644 NOK when estimates are based on our model. This amount could potentially be reduced with better insight into the model´s behavior. The cluster analysis in section 6.3 presents itself as an economic tool to understand trends and patterns across groups of predictions. A pattern seen in the cluster analysis, with both three – and six clusters, is that when NWP-models differentiate in their estimate of wind speed for any given hour, the prediction model tends to predict poorly on an average basis. For instance, in section 6.3.1, the same pattern is seen for *cluster three*, while cluster *one and two* have NWP-models that have smaller differences in their estimates of wind speed. *Cluster three* has twice as big RMSE as that of *cluster one* and *two*. Observations with a similar explanation structure to *cluster three* therefore has the least accurate predictions on average. A possible solution is to develop prediction models that only applies features that exhibit roughly the same estimates of wind speeds across the different height profiles. For instance, if two of the NWPs are considerably different from the third, the model will only use the two most similar. By using alternative prediction models in periods where NWP-models differentiates greatly in regard to wind speeds, the replan predictions´ accuracy can be improved, and subsequently increase TrønderEnergi´s income.

TrønderEnergi´s prediction models have a significant impact on the stakeholders. Furthermore, we can say that the predictions generated by these models are in line with PWCs critical factors outlining why interpretability is a crucial aspect in a practical setting.

7.2 Business Benefits with Interperatble Machine Learning

Oxborough et al. (2018) discusses how interpretable machine learning can be applied as a competitive advantage for businesses. In order to optimize a model´s performance, they suggest retaining *control*, maintain *trust* and comply with *accountability*. To optimize the decision making they suggest retaining *safety*, maintain *ethics* and comply with *regulation*. In the sense of this thesis, and in general for wind power predictions, we see retaining *control*, maintaining *trust*, and complying with *regulations* as the most important factors. In the sections to come, we will discuss these three factors to see how interpretable machine learning benefits TrønderEnergi.

7.2.1 Regulation

The usage of AI and machine learning have increased rapidly over the last years. The rapid growth partially explains why it is so lightly regulated. However, as the usage of AI increases and starts impacting individuals´ life to a larger extent, regulations will increase. Standards for accuracy, transparency and reasoning will be focus areas for governments when implementing new regulations (Oxborough et al., 2018).

A regulation from EU that came into effect in 2018, requires a “right to an explanation” for automated individual decision-making processes. It requires a business to explain the decisions of their algorithm to individuals, and especially when it impacts an individual legally and financially. Denmark was the first country to implement a compulsory regulation for AI and Data Ethics, and it came into effect in July of 2020. The regulation gives an individual the right to an explanation. Specifically, companies need to deliver information regarding their algorithms, and demonstrate they are transparent and fair (Research, 2020; Sønnersgaard, 2020). The Norwegian Board of Technology´s main task is to give the Norwegian Government recommendations about technological regulations, and the usage of new technology. In a report written in 2018, they proposed a strategy for using AI in Norway. One of the fourteen proposal in this strategy was “the right to an explanation”. It says that the Norwegian Government should legislate a right to explanation and avoid applying decision-systems without good enough explanations (Teknologirådet, 2018). From the new regulations and the strategy proposal, we see an increasing focus on the “right to an explanation” aspect. Consequently, there is a great probability that such regulations will be implemented in Norway as well, covering all types of industries and markets. In the future, it should come as no surprise if TrønderEnergi are required to explain their decisions based on machine learning models.

TrønderEnergi is bound by law to ensure their prognosis is their best estimate of actual production. This is to create a balance in the market, and also to maximize the social welfare (NordPool, 2020). Bids from participants in the day-ahead marked must be based on its fundamental conditions and marginal-cost. The Norwegian Energy Regulatory Authority (NVE – RME) monitors the power suppliers’ bids in terms of volume and prices. If they notice significant deviations in a supplier´s bid compared to other participants, they might require an explanation. However, a phone call with the Norwegian Energy Regulatory Authority made it clear that a prediction model is categorized as a fundamental condition. Therefore, it is often enough to explain significant deviations with the model´s inaccuracy in itself, without exploring why this prediction error actually occurred.

As discussed in section 2.3.1, in periods where the RK-price is systematically higher than the day-ahead price, there is an economic incentive to underestimate the prognosis on the day-ahead market. We observe particular months with a systematically larger RK-price than the spot price throughout the data´s time interval. For instance, August is a month where this tends to be the case (see Appendix C). Potentially, a power supplier can train a separate model that systematically underestimate the predicted production level at the day-ahead market.

Therefore, in our opinion, if the Norwegian Energy Regulatory Authority wishes to maximize the social welfare in the market and ensure the law is followed, then they need to require more detailed information about the algorithms used to predict power production, and also require individual and global explanations of these predictions.

In conclusion, there is an increasing focus on regulations that require a “right to an explanation” in decision that uses machine learning processes. When the RK-price is higher than the spot price, there is an economic incentive to violate the law, without the fear of getting caught. If the Norwegian Energy Regulatory Authority does not possess full insight into power suppliers´ prognosis, it might be difficult to uncover violations. These two aspects might require TrønderEnergi to explain their predictions in more detail in the upcoming future. By developing an interpretable machine learning framework like ours, TrønderEnergi might gain competitive advantages or business benefit due to their proactive behavior. If regulations regarding the right to an explanation starts to become popular in Norway, SHAP-values is potentially the only legally compliant method (Molnar, 2019). This is mainly because it is based on a solid theory that distributes the features contribution to the predicted output fairly and consistently. The framework in this thesis and the analysis therefore seems to be a good choice when considering legal explanations.

7.2.2 Control

Interpretable machine learning can help TrønderEnergi to retain control of their machine learning models. Control can be accomplished by monitoring the prediction model´s behavior, pointing out its errors, and it enable us to shut down the system or use alternative models if required. That is exactly what the explanations in this analysis helped us to do. First of all, they gave us better insight about the features used to predict wind power. Local explanations helped us to better explain potential reasons why an individual prediction failed in terms of prediction errors, and the economic issues that arises as a result of production imbalances. Having a better understanding of these issues helps TrønderEnergi to better maintain a certain level of control. The cluster analysis in section 6.3 is especially useful. It helps us to point out specific situations where the model tends to systematically fail in terms of prediction errors and production imbalance costs. As mentioned earlier, a potential solution is to use alternative machine learning models for observations where the original model historically has failed to predict wind power accurately. For instance, an option is to use alternative machine learning models when your observation has a similar explanation structure to that of cluster three in section 6.3.1. Potentially, this could improve the replan predictions overall, and help TrønderEnergi to get a stable performance.

7.2.3 Trust

67 % of business leaders thinks AI has an adverse impact on stakeholders' level of trust (Oxborough et al., 2018). Regarding TrønderEnergi and machine learning, trust is built by providing adequate information to their stakeholders about the prediction models. Adequate information in the sense that the models are making the correct decision for the right reasons. Global explanations in section 6.1 covers this aspect well, especially when considering domain experts and operators in wind power production. These types of explanations help to understand the prediction model´s global behaviour and help to verify if the relationships learned are consistent with the fundamental characteristics at Ytre Vikna. Specifically, they help domain experts within wind power production to trust the model and verify the reliability of the prediction model applied. For instance, the SHAP dependence plots for wind speed, shows random forests capability to model the wind turbines power curve with NWP-models wind speed prediction measured at different heights.

Explicability ability helps to detect errors in the model or if there is any bias in the training data used to train the model. A deeper insight into the model helps to understand why it occasionally fails, and therefore makes it simpler to enhance the predictions (Oxborough et al., 2018). By using interpretable machine learning, TrønderEnergi can improve their predictions through a deeper understanding of their models. As explained in section 7.2.2, one option to improve the predictions is by using the clusters obtained in section 6.3. Hopefully, this will improve the predictions accuracy and make them more stable over time. Stable results build trust and makes it easier for stakeholders to accept TrønderEnergi´s machine learning models.

The production central is a department within TrønderEnergi that monitors their power plants 24/7. They have the short-term responsibility of planning the operation at TrønderEnergi´s power plants. This includes the responsibility of bids in the day-ahead - and balancing market (Trønder Energi, 2020). The volumes for these bids are based on the machine learning models developed for each power plant. The individuals employed at the production central are seldom experts of machine learning, and in periods where the models tend to predict wrongly, it is easy to lose trust in the machine. By providing reasonable explanations to the people working at the production central, their knowledge about these models increases, and their trust levels are also likely to increase. One type of such reasonable explanations is shown in section 6.2 (local explanations). If the people working at the production central, or in general internal stakeholders, have a higher confidence and trust in machine learning processes, then it will be much easier for TrønderEnergi to implement new machine learning models/processes in the coming future. Hence, in our opinion, it is beneficial for TrønderEnergi to implement model-agnostic models in order to increase internal stakeholders trust levels. Subsequently, it will be easier to implement new innovative solutions with fewer obstacles.

7.3 Limitations

To answer this thesis´s research question, an new version of KernelSHAP from Aas et al. (2019) was used to explain individual wind power predictions. This method takes feature dependency into account. Features conditional distributions are estimated to handle dependency. As explained in section 5.2.1, we use a combined approach to estimate conditional distributions. Specifically, the empirical distribution for subset of dimension equal to 1 and Gaussian distribution otherwise.

A simulated experiment in Aas et al. (2019) paper shows that their methods provide more accurate estimates of Shapley values, compared to the original KernelSHAP (Lundberg & Lee, 2017) and TreeSHAP (Lundberg et al., 2019). Although the improved KernelSHAP used in this thesis provides more accurate explanations, the same experiment also displays a certain uncertainty. This means, our estimates are not perfect approximation of Shapley values. This needs to be considered when interpreting this analysis.

Real Shapley values are difficult to compute due to the dimensional problem that arises in computing them. Therefore, explanations have an unquantified uncertainty. However, when summarizing features explanation as mean attributions, like the cluster analysis, there is a possibility to quantify the uncertainty. Merrick & Taly (2020) propose a solution using the bootstrap approach to quantify mean attributions uncertainty. In section 6.3.1 cluster analysis, uncertainty is quantified for mean Shapley values using the bootstrap approach to compute their confidence interval (see Appendix D). The confidence intervals (CI) for *cluster one, two* and *three* shows in general a low uncertainty since the CI spans very close to the mean value. This strengthens the validity of the results obtained in section 6.3.

8. Conclusion

In this thesis we have evaluated whether interpretable machine learning gives TrønderEnergi valuable insight into their wind power prediction models. For this, we needed a separate interpretation method that is applicable to explain any type of machine learning model. The interpretation model needed to yield explanations for single observations as well as the model as a whole. SHapley Additive exPlanations (Lundberg & Lee, 2017) and specifically a new version of the KernelSHAP approach (Aas et al., 2019) presented us with a suitable framework. It allowed us to examine which features that contribute the most to the final prediction of any given observation, and it let us aggregate these observations to interpret the global effects of features. When we had computed explanations for every observation in the test set, we could also use clustering to group explanations with similar explanation structure.

The global analysis showed that the three wind speed features contributes the most to the predicted output. Wind direction has smaller contribution in general. The SHAP-dependence plots gave especially two valuable insights. First, they verified that the trained relationships between wind speed and production are in accordance with wind turbines fundamentals or the power curve. Second, they exhibit the importance of wind direction as a feature in wind power predictions. This includes random forest's ability to consider the complex terrain at Ytre Vikna. When using wind direction as a feature, random forest seems to take speed-up effects and wake effects into account. These insights could potentially increase the trust of those domain experts or wind power engineers that doubt on using machine learning in wind power predictions.

In the section with local explanations, we chose to focus on explaining the observation with the highest prediction error and the observation with the highest production imbalance cost. However, the methodology can be applied to any observation, so these are just interesting examples. In the case of the largest prediction error, this seems to be caused by bad estimates of wind speed. The large imbalance cost is mainly caused by the huge spread between the spot and RK-price and the fact that the prediction resulted in an underestimation of actual production.

When clustering the observations based on their explanations, the goal was to find groups of observations that had the same basic explanation structure. Considering the prediction model´s performance, the most valuable insights were found in this section. The cluster analysis found groups of observations that had different patterns when considering prediction error and production imbalance cost. The main findings were the following: When NWP-models differentiate in their estimate of wind speed for any given hour, the prediction model tends to predict worse on an average. When predicted wind speed was relatively high for all NWP-models, the predictions where sufficiently better in terms of predictive accuracy. Low predicted wind speeds for all NWP-models, results in the predictions closest to actual production, likely because of low production elasticity when wind speeds are low.

In our discussion we establish why interpretable machine learning grants valuable insights into wind power predictions. Three areas we focus on are to comply with *regulation*, retaining *control*, and maintaining *trust*. Regulations may be put forward, that demands a sound reasoning for individual predictions. Local explanations can be used to see how features contributed to the prediction and uncover any irregularities.

Control is an area of focus because with interpretable machine learning you maintain a certain control over developments regarding the learning process. This control can be exploited to improve future versions of machine learning models, and subsequently improve the model´s predictive performance.

Trust can be important in an operative manner. The people working on the production central or in general internal stakeholders have little experience with machine learning. By giving them simple explanations as to why things are like they are, they might improve their understanding of the system as a whole. Increasing trust is beneficial since it potentially eases the implementation of new machine learning models in the future with less internal obstacles.

In conclusion, the three explanation frameworks proposed in this thesis, are all providing valuable insight into our complex prediction model. Combining this with the fact that interpretable machine learning is a potential competitive advantage through trust, control, and regulations, we definitively see it as a valuable asset for TrønderEnergi. Going forwards we think the next step would be to expand upon this thesis by including both the day-ahead and replan prognosis and digging deeper into the spread between spot prices and RK-prices. This could lead to a better understanding of interactions which we were not able to explore.

References

Aas, K., Jullum, M., & Løland, A. (2019). Explaining individual predictions when features are dependent: More accurate approximations to shapley values. In *arXiv*.

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*. https://doi.org/10.1109/ACCESS.2018.2870052

Akram, M. T. (2014). *An analysis of errors in prediction of wind power production - A study based on Ytre Vikna part 1* [Norges miljø- og biovitenskapelige universitet]. https://nmbu.brage.unit.no/nmbu-xmlui/bitstream/handle/11250/221539/Combine Result.pdf?sequence=1&isAllowed=y

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable Machine Learning in Credit Risk Management. *Computational Economics*. https://doi.org/10.1007/s10614-020-10042-0

Carlsson, L. S., Samuelsson, P. B., & Jönsson, P. G. (2020). Interpretable Machine Learning—Tools to Interpret the Predictions of a Machine Learning Model Predicting the Electrical Energy Consumption of an Electric Arc Furnace. *Steel Research International*. https://doi.org/10.1002/srin.202000053

DARPA. (2016). *Broad Agency Announcement Explainable Artificial Intelligence (XAI).*

Eilenberg, R. S. (2012). Wind Farm Arrangement: Considering the Influence of Complex Terrain and Turbine Wake Interactions . *Development of an Automated Fabric Deformation System for Composite Wind Turbine Blade Manufacturing*, 82–95. http://home.engineering.iastate.edu/~jdm/wind/Binder1.pdf#page=82

eSett. (n.d.). *eSett – Handbook*. Retrieved November 17, 2020, from https://www.esett.com/handbook/

Foley, A. M., Leahy, P. G., & McKeogh, E. J. (2010). Wind power forecasting & prediction methods. *2010 9th Conference on Environment and Electrical Engineering, EEEIC 2010*. https://doi.org/10.1109/EEEIC.2010.5490016

Grange, S. K. (2014). *Technical note: Averaging wind speeds and directions*. https://www.researchgate.net/publication/262766424_Technical_note_Averaging_wind_speeds_and_directions?channel=doi&linkId=54f6184f0cf27d8ed71d5bd4&showFulltext=true

Gupta, A. (2015). *Overcoming Missing Values In A Random Forest Classifier*. https://medium.com/airbnb-engineering/overcoming-missing-values-in-a-random-forest-classifier-7b1fc1fc03ba

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Elements of Statistical Learning 2nd ed. In *Elements*.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2006.03.001

James, G., Witten, D., Hastie, T., & Tibishirani, R. (2013). An Introduction to Statistical Learning with Applications in R (older version). In *Springer Texts in Statistics*.

Jammalamadaka, R., & Sarma, Y. (1988). A correlation coefficient for angular variables . *Statistical Theory and Data Analysis 2*. https://www.researchgate.net/publication/244954553

Kuzlu, M., Cali, U., Sharma, V., & Guler, O. (2020). Gaining Insight Into Solar Photovoltaic Power Generation Forecasting Utilizing Explainable Artificial Intelligence Tools. *IEEE Access*. https://doi.org/10.1109/access.2020.3031477

Larmarange, J. (2020). *best.cutree: Optimal partition based on the higher relative loss of criteria* . https://rdrr.io/github/larmarange/JLutils/man/best.cutree.html

Lepore, A., Palumbo, B., & Pievatolo, A. (2020). A Bayesian approach for site-specific wind rose prediction. *Renewable Energy*. https://doi.org/10.1016/j.renene.2019.12.137

Lin, C. (2018). *Application-grounded evaluation of predictive model explanation methods*.

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*. https://doi.org/10.1145/3233231

Lototzis, M., Papadopoulos, G. K., Droulia, F., Tseliou, A., & Tsiros, I. X. (2018). A note on the correlation between circular and linear variables with an application to wind direction and air temperature data in a Mediterranean climate. *Meteorology and Atmospheric Physics*. https://doi.org/10.1007/s00703-017-0508-y

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2019). Explainable AI for trees: From local explanations to global understanding. In *arXiv*.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*.

Lydia, M., Kumar, S. S., Selvakumar, A. I., & Prem Kumar, G. E. (2014). A comprehensive review on wind turbine power curve modeling techniques. In *Renewable and Sustainable Energy Reviews*. https://doi.org/10.1016/j.rser.2013.10.030

Mardia, K. V. (1976). Linear circular correlation coefficients and rhythmometry. *Biometrika*. https://doi.org/10.2307/2335637

Merrick, L., & Taly, A. (2020). The Explanation Game: Explaining Machine Learning Models Using Shapley Values. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-030-57321-8_2

Ministry of Petroleum and Energy. (1999). *Forskrift om måling, avregning, fakturering av nettjenester og elektrisk energi, nettselskapets nøytralitet mv - Lovdata*. Forskrift Om Kraftomsetning Og Nettjenester. https://lovdata.no/dokument/SF/forskrift/1999-03-11-301?q=avregningsforskriften

Molnar, C. (2019). Interpretable Machine Learning. A Guide for Making Black Box Models

Explainable. *Book*.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1900654116

NCEI. (2020). *Numerical Weather Prediction*. National Centers For Environmental Information. https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/numerical-weather-prediction

Nielsen, H. A., Nielsen, T. S., Madsen, H., San Isidro Pindado, M. J., & Marti, I. (2007). Optimal combination of wind power forecasts. *Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology*, *10*(5), 471–482. https://doi.org/10.1002/we.237

Niemerg, M. (2020). *Like Peas in a Pod: Ideas in Cluster Analysis*.

NordPool. (2020). *Day-ahead market Nord Pool*. https://www.nordpoolgroup.com/the-power-market/Day-ahead-market/

Oxborough, C., Cameron, E., Rao, A., & Westermann, C. (2018). Explainable AI - Driving business value through greater understanding. In *PricewaterhouseCoopers report*.

Pewsey, A., Neuhauser, M., & Ruxton, G. D. (2013). Circular Statistics in R. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. https://doi.org/10.1111/rssa.12222

Pratama, I., Permanasari, A. E., Ardiyanto, I., & Indrayani, R. (2016). A review of missing values handling methods on time-series data. *2016 International Conference on Information Technology Systems and Innovation, ICITSI 2016 - Proceedings*. https://doi.org/10.1109/ICITSI.2016.7858189

Prentzas, N., Nicolaides, A., Kyriacou, E., Kakas, A., & Pattichis, C. (2019). Integrating machine learning with symbolic reasoning to build an explainable ai model for stroke prediction. *Proceedings - 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering, BIBE 2019*. https://doi.org/10.1109/BIBE.2019.00152

Redelmeier, A., Jullum, M., & Aas, K. (2020). Explaining Predictive Models with Mixed Features Using Shapley Values and Conditional Inference Trees. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-030-57321-8_7

Research. (2020). *Mandatory Legislation for AI & Data Ethics Introduced In Denmark | Industry News*. https://www.explainx.ai/resources/mandatory-legislation-for-ai-data-ethics-introduced-in-denmark

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). *Model-Agnostic Interpretability of Machine Learning*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). “Why should i trust you?” Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference* *on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939778

Rosvold, K. A. (2019). *Ytre Vikna vindpark i Store norske leksikon*. Store Norske Leksikon. https://snl.no/Ytre_Vikna_vindpark

Shapley, L. S. (1953). A Value for n person Games. Contributions to the Theory of Games. *Annals of Mathematics Studies*.

Sønnersgaard, M. (2020). *Nu skal virksomheder redegøre for dataetik i årsrapporten | Læs nyheden*. https://kammeradvokaten.dk/nyheder-viden/nyheder/2020/06/nu-skal-virksomheder-redegoere-for-dataetik-i-aarsrapporten

Teknologirådet. (2018). *Kunstig Intelligens. Muligheter, utfordringer og en plan for Norge*.

Trønder Energi. (2020). *Energy Management - Energy Management*. https://energymanagement.tronderenergi.no/

Vega García, M., & Aznarte, J. L. (2020). Shapley additive explanations for NO2 forecasting. *Ecological Informatics*. https://doi.org/10.1016/j.ecoinf.2019.101039

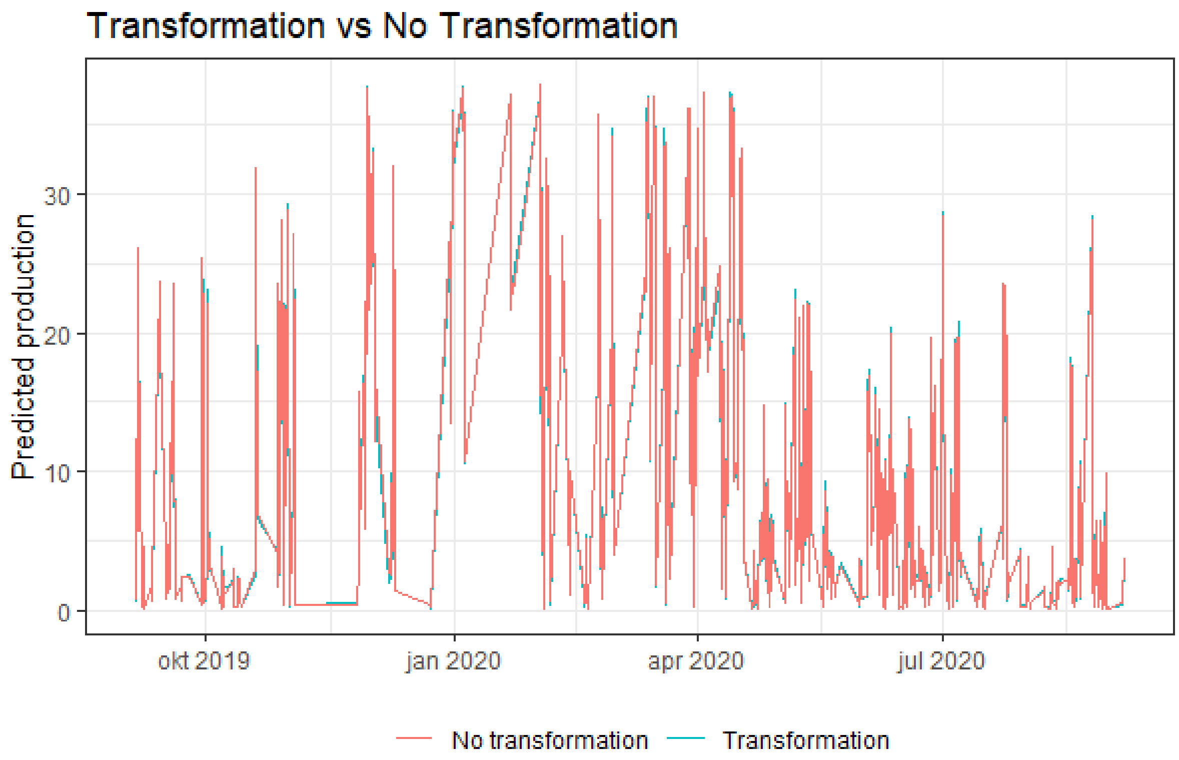
Wright, M. N., & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v077.i01

Yamartino, R. J. (1984). A comparison of several ' single-pass' estimators of the standard deviation of wind direction. *Journal of Climate & Applied Meteorology*. https://doi.org/10.1175/1520-0450(1984)023<1362:ACOSPE>2.0.CO;2

Young, H. P. (1985). Monotonic solutions of cooperative games. *International Journal of Game Theory*. https://doi.org/10.1007/BF01769885

Appendix

Appendix A

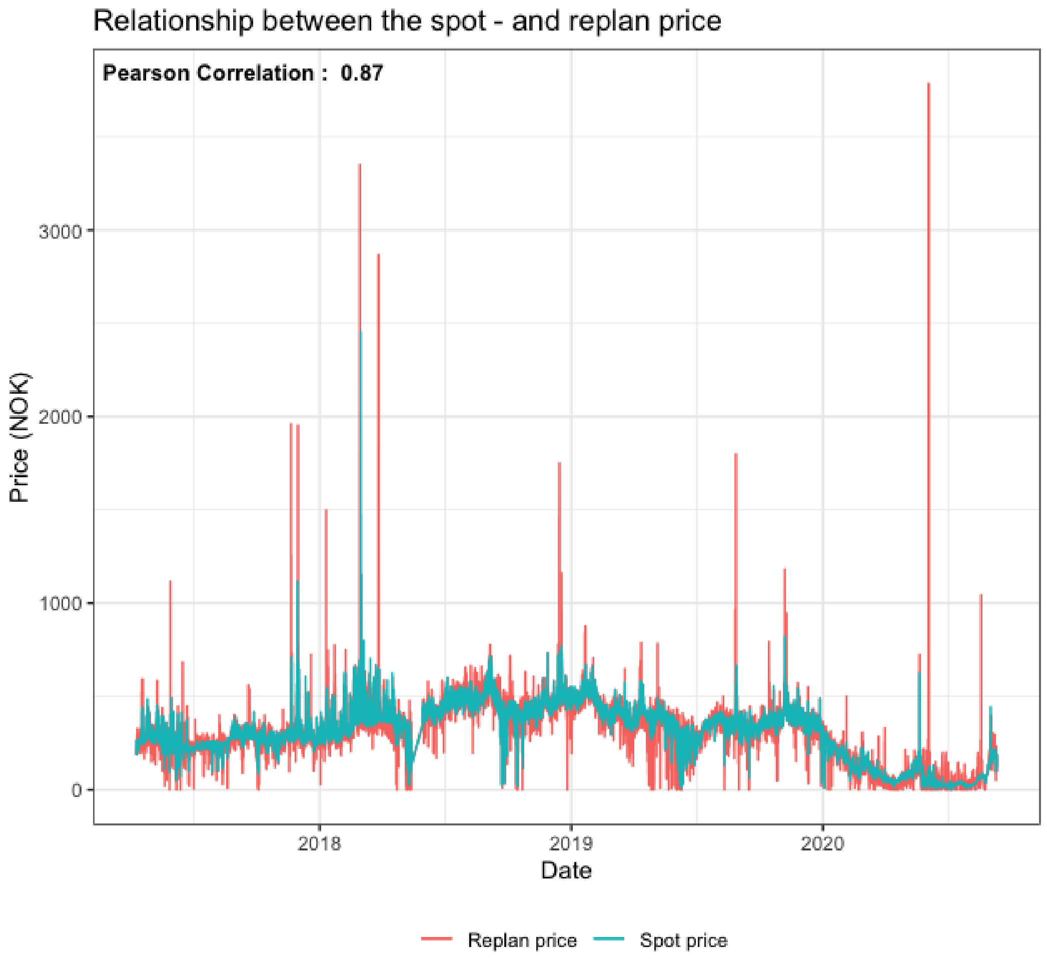


RMSE with Transformation	RMSE without Transformation
4.4509	4.4498

Appendix B

Relationship between the spot- and RK-price in the period

from 09. September 2017 to 09. September 2020



	Mean	Median	Max	Min	Standard Deviation
Spot Price	312.3 NOK	330.2 NOK	2454.3 NOK	0	145.9 NOK
RK-price	301.3 NOK	315.1 NOK	3784.4 NOK	0	161.5 NOK

Appendix C

Overview of months with an average RK-price larger than the average spot price in the period from 09. September 2017 to 09. September 2020

Year	**Month**	**Mean RK-price**	**Mean Spot Price**
2017	November	313,18 NOK	308,22 NOK
2018	August	497,96 NOK	494,63 NOK
2019	August	365,56 NOK	364,60 NOK
2020	June	63,56 NOK	33,78 NOK
2020	August	75,44 NOK	67,87 NOK
2020	September	184,59 NOK	177,14 NOK

Appendix D

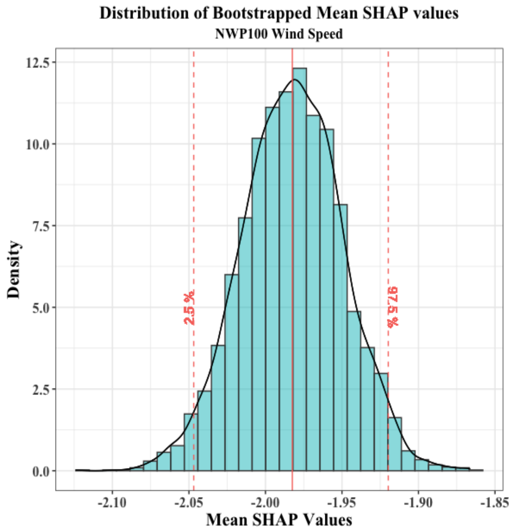
Three Clusters – Mean Shapley Values with confidence intervals, computed using the bootstrap approach.

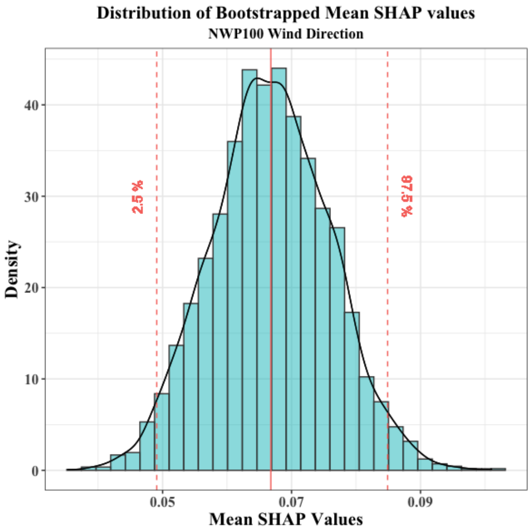
Cluster 1

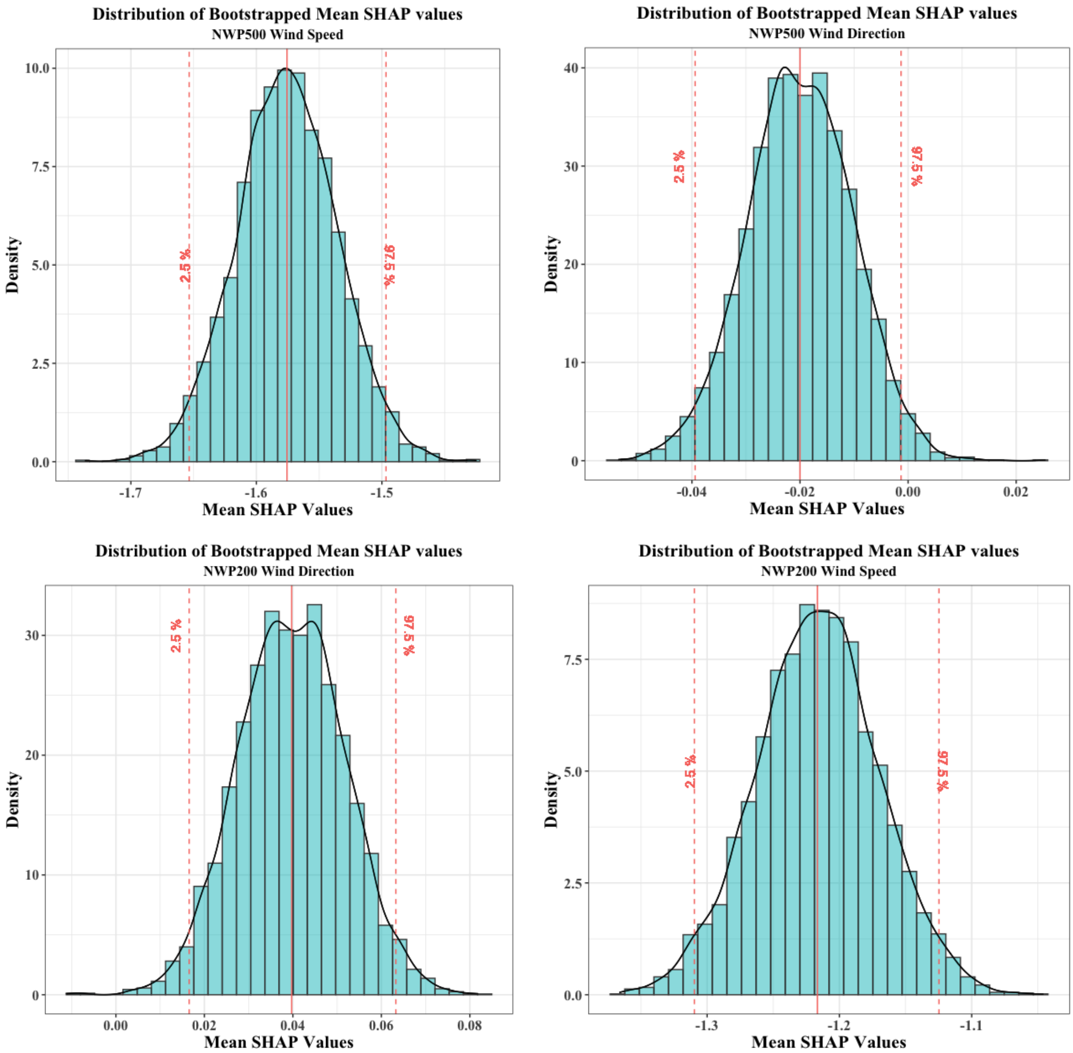
The table displays an overview of features average SHAP-value, and their respective confidence intervals for cluster number one in section 6.3.1. Mean ϕ values are the same as the values shown in table 3, and the confidence intervals illustrate these values uncertainty.

Feature	**Mean ϕ**	**2.5%**	**97.5%**
NWP100 Wind Direction	0.067	0.049	0.085
NWP100 Wind Speed	-1.982	-2.049	-1.917
NWP500 Wind Direction	-0.02	-0.039	-0.001
NWP500 Wind Speed	-1.577	-1.659	-1.495
NWP200 Wind Direction	0.04	0.017	0.064
NWP200 Wind Speed	-1.216	-1.306	-1.123

The histograms show the distribution of sampled mean SHAP-values computed with the bootstrap approach, and for cluster one. It also displays the overall mean, and 2.5 % - and 97.5 % percentiles or the confidence intervals.






Distribution of Bootstrapped Mean SHAP values
NWP500 Wind Speed
Density
Mean SHAP Values
-1.7
-1.6
-1.5
0.0
2.5
5.0
7.5
10.0
2.5 %
97.5 %
Distribution of Bootstrapped Mean SHAP values
NWP500 Wind Direction
Density
Mean SHAP Values
-0.04
-0.02
0.00
0.02
0
10
20
30
40
2.5 %
97.5 %
Distribution of Bootstrapped Mean SHAP values
NWP200 Wind Direction
Density
Mean SHAP Values
0.00
0.02
0.04
0.06
0.08
0
10
20
30
2.5 %
97.5 %
Distribution of Bootstrapped Mean SHAP values
NWP200 Wind Speed
Density
Mean SHAP Values
-1.3
-1.2
-1.1
0.0
2.5
5.0
7.5
2.5 %
97.5 %

Cluster 2

The table displays an overview of features average SHAP-value, and their respective confidence intervals for cluster number two in section 6.3.1. Mean ϕ values are the same as the values shown in table 3, and the confidence intervals illustrate these values uncertainty.

Feature	**Mean ϕ**	**2.5%**	**97.5%**
NWP100 Wind Direction	0.442	0.414	0.47
NWP100 Wind Speed	6.22	6.122	6.32
NWP500 Wind direction	0.179	0.149	0.209
NWP500 Wind speed	4.36	4.225	4.494
NWP200 Wind Direction	0.415	0.376	0.455
NWP200 Wind Speed	5.082	4.921	5.241

Cluster 3

The table displays an overview of features average SHAP-value, and their respective confidence intervals for cluster number three in section 6.3.1. Mean ϕ values are the same as the values shown in table 3, and the confidence intervals illustrate these values uncertainty.

Feature	**Mean ϕ**	**2.5%**	**97.5%**
NWP100 Wind Direction	0.281	-0.037	0.588
NWP100 Wind Speed	10.796	9.636	11.632
NWP500 Wind direction	0.139	-0.077	0.355
NWP500 Wind speed	5.539	3.263	7.705
NWP200 Wind Direction	0.032	-0.409	0.534
NWP200 Wind Speed	-7.911	-9.291	-6.504

Appendix E

R-implementations and libraries used in our methodologies

Library	Methodology	What it does
“Ranger”	Random Forest	Used to train our prediction model. Optimized for high dimensions, but also runs faster at low dimensions due to pre-ordering and memory efficiency (Wright & Ziegler, 2017).
“Shapr”	Shapley additive explanations	Used to estimate Shapley values. Applies the improved version of KernelSHAP (Aas et al., 2019).
“Dist”	Step 1. Agglomerative cluster algorithm	Calculates the Euclidean distance between each observation. Used as input in hclust.
“Hclust”	Step 2. Agglomerative cluster algorithm	Calculates the dendogram.
“Best.cuttree”	Agglomerative clustering	Used to find the superior partition to cut the dendrogram.

Appendix F

Six Clusters – Results

Mean Shapley Values (ϕ), Prediction Errors and Production

Cluster	Mean ϕ NWP100 WD	Mean ϕ NWP500 WD	Mean ϕ NWP200 WD	Mean ϕ NWP100 WS	Mean ϕ NWP500 WS	Mean ϕ NWP200 WS	RMSE Prediction Error Random Forest	RMSE Prediction Error TE	Mean Production
1	0,299	0,201	0,545	-2,427	4,577	3,651	6,192	6,311	18,123
2	0,456	0,202	0,433	6,286	4,680	4,956	5,642	5,686	28,349
3	0,025	-0,060	-0,052	-1,902	-2,696	-2,102	3,214	3,329	5,083
4	0,098	-0,405	-0,028	4,562	-3,689	8,247	6,590	6,184	20,351
5	0,374	0,176	0,018	11,264	5,229	-7,720	8,711	9,859	22,050
6	-1,565	-0,583	0,192	1,848	11,473	-11,668	24,610	20,576	36,576

Mean Numerical Weather Predictions – m/s and degrees

Cluster	Mean NWP100 WD	Mean NWP500 WD	Mean NWP200 WD	Mean NWP100 WS	Mean NWP500 WS	Mean NWP200 WS
1	174,159	187,100	175,866	5,206	15,078	9,107
2	233,989	235,531	231,239	9,598	17,941	11,627
3	261,925	252,287	246,115	3,634	6,501	5,328
4	284,342	265,172	271,595	7,524	9,132	10,351
5	211,043	230,784	191,653	10,483	17,374	6,753
6	11,295	355,300	288,200	6,078	18,800	2,900

Mean Economical Variables – Prices, Loss of Income and Monetary Loss

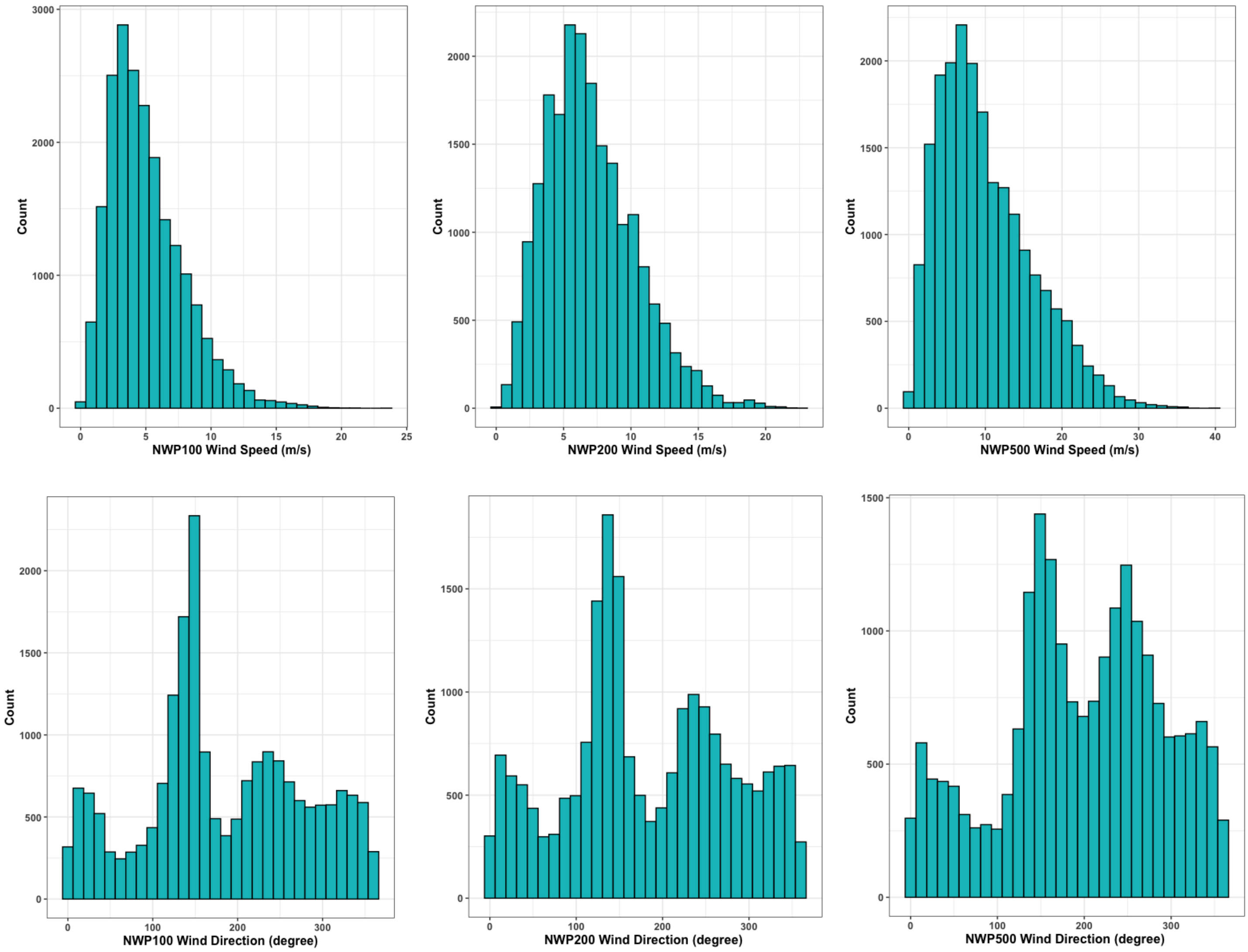
Cluster	Mean RK-price	Mean spot price	Mean total production imbalance cost	Max total production imbalance cost	Mean monetary loss	Number of observations	Number of under-estimations	Number of over-estimations
1	218,941	232,752	81,356	3841,946	58,799	986	421	565
2	168,939	188,817	79,174	4309,147	74,009	2442	1300	1142
3	173,402	175,792	36,676	6520,495	23,963	5419	1864	3555
4	200,433	180,956	84,718	907,289	137,753	97	41	56
5	81,905	111,392	164,265	963,422	69,277	19	9	10
6	52,820	52,822	0,000	0,000	0,061	1	1	0

Six Clusters - Analysis

The conclusions of the tables above are very equal to they in section 6.3.2. Cluster *one*, *four* and *five* all exhibit differences between the NWP-models wind speed feature effect on predicted output. In these clusters two NWP-models always pull in a different direction than the third. Two NWP-models wind speed features have a positive effect on the predicted output, and the last NWP-model wind speed feature has a negative effect on the predicted output. Looking away from *cluster six* which consists of only one observation, cluster *one*, *four* and *five* provides the largest prediction errors in terms of RMSE. Cluster *five* has the clearly largest RMSE value among them. This may imply that when NWP200 wind speed is the NWP-model that pulls in the other direction, the random forest model tends to perform the worst. This is consistent with the conclusion from section 6.3.1. *Cluster two* and *three* support the statement in section 6.3.1, that observations where the NWP-models wind speed predictions are in harmony tend to have the best accuracy in terms of RMSE. When all have a negative effect on the predicted output, implying low wind speeds, the predictive accuracy improves as elasticity of wind speeds effect on the prediction is low.

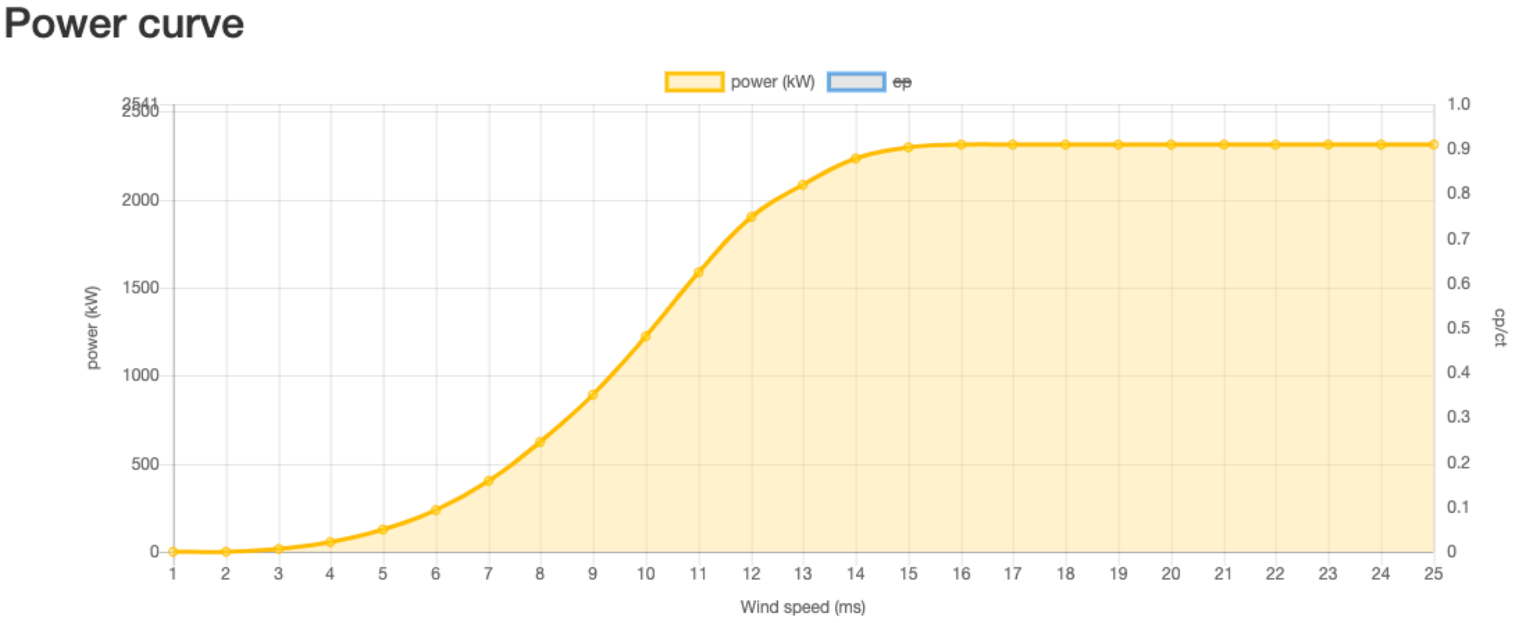
Appendix G

Feature distributions



Appendix H

Enercon E70 2,3MW power curve provided by the manufacturer


Power curve
power (kW)
~~cp~~
power (kW)
cp/ct
Wind speed (ms)