Norwegian School of Economics

Bergen, Spring 2023

Forecasting the Fish Pool Index

Can tree-based models produce accurate and reliable forecasts of the salmon spot price?

Cecilia Andrea Mowinckel & Simen Nordskag

Supervisor: Stein Ivar Steinshamn

Master thesis, Economics and Business Administration,
Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

Acknowledgements

This thesis was written as the final part of our Master of Science in Economics and Business Administration at the Norwegian School of Economics (NHH). In writing this thesis we have had a chance to apply the knowledge gathered throughout our majors in Business Analytics. It has been both challenging and rewarding to write a thesis about a field of such interest to the Norwegian economy, and we hope it will be as interesting to read as it was for us to write.

We would like to thank our supervisor, Professor Stein Ivar Steinshamn, for providing valuable knowledge on the studied topics and feedback on our thesis throughout the process. We would also like to thank Professor Jonas Andersson for providing great advice on the topic of predictive analytics, Hans Vanhauwaert Bjelland at Sintef for input on feature variable selection, in addition to Lars Erik Flatøy at Kontali Analyse and John Selnæs at Lusedata for providing data not publicly available. Lastly, we thank our family and friends for supporting us the whole way.

Norwegian School of Economics

Bergen, May 2023

Cecilia Andrea Mowinckel

Simen Nordskag

Abstract

Industrial salmon farming is becoming an increasingly important industry, both globally and in Norway. One of the main risk factors in salmon production is the highly volatile spot price, so access to high-quality price forecasts could prove immensely valuable throughout the value chain. In this thesis, we therefore attempt to make accurate and reliable forecasts of the salmon price 12 months ahead and assess the potential economic value of such forecasts. We chose to use tree-based models for this task, and the models applied were decision trees, random forests, and xgBoost, with- and without seasonal adjustment.

The tree-based models displayed different levels of forecast accuracy, however all models performed better than the seasonal naïve benchmark. Measured by mean absolute error and mean squared error the best performing model was random forest, followed by xgBoost and then decision tree. Overall, the seasonally adjusted random forest performed best, with a directional accuracy of 82%, implying that the model correctly predicted up- or down price movements around 8 out of 10 times. We found that the potential economic value of such forecasts to SalMar, the third largest salmon producer in Norway in 2021 with a market share of around 11%, could be 51.2 million NOK in additional earnings, corresponding to a 2% increase compared to their total 2021 earnings.

List of Contents

List of Figures

List of Tables

List of Appendixes

List of Abbreviations

ARIMA	Autoregressive Integrated Moving Average
CP	Complexity Parameter
CPI	Consumer Price Index
DA	Directional Accuracy
FPI	Fish Pool Index
GAM	Generalized Additive Models
GDP	Gross Domestic Product
ISA	Infectious Salmon Anemia
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MSE	Mean Squared Error
RF	Random Forest
RMSE	Root Mean Squared Error
S.A.	Seasonally Adjusted
SNAIVE	Seasonal Naïve Model
SSB	Statistics Norway
STL	Seasonal and Trend Decomposition using Loess
SVM	Support Vector Machines
WFE	Whole-Fish-Equivalent
WHO	World Health Organization
WTI	West Texas Intermediate crude oil
XgBoost	Extreme Gradient Boosting

1. Introduction

In this chapter we explain the motivation behind choosing to write about salmon price forecasting, before defining the research question of the thesis.

1.1 Motivation

Securing sustainable food supply is increasingly becoming a major global challenge, as the global population is expected to increase to a total of 9.8 billion (United Nations, n.d.) and the overall food demand is projected to increase by more than 50% by 2050 (Searchinger et al., 2019). Increasing demand for food implies more pressure on the world's resources, and new sustainable methods of food production becomes an integral part of the solution. Farmed Atlantic salmon is likely to be part of that solution because it is considered a sustainable and healthy alternative to other types of meat. For example, the carbon footprint of beef is about ten times higher than that of salmon. Pork's carbon footprint is more than twice as large, and even chicken has a 50% larger carbon footprint (Global Salmon Initiative, 2021).

Comparing the area used to produce 100g of protein from different meat sources also highlights the increasing importance of salmon. Farmed salmon requires 3.7 square meters to produce 100g of protein, while poultry requires roughly twice that area, pork almost three times, beef close to 30 times, and lamb 50 times the area of salmon. Farmed salmon is also the most eco-friendly type of meat measured by feed-conversion. Feed conversion measures kg in feed required to increase the animal's bodyweight by one kg. Salmon's feed conversion ratio is around 1.5, making it extremely efficient in food production. Salmon is also attractive measured by edible yield, which is the ratio of edible meat to total body weight. 68% of the salmon's body weight is edible, a much higher yield than other meat types (Global Salmon Initiative, 2021). Farmed salmon will therefore be a valuable, sustainable, and eco-friendly addition to the future global food supply.

Another trend which may cause salmon to increase in importance is changing consumer preferences. As lifestyle diseases such as obesity, heart disease, stroke and diabetes increasingly become major public health problems, the world not only needs more food, but healthier food. WHO estimates that by 2030 the proportion of deaths caused by lifestyle diseases will increase to 70% (Al-Maskari, n.d.). As consumers become more aware of the importance of a healthy diet salmon demand could increase, as salmon is a nutrient-rich food and contributes protein, healthy fats, and several essential vitamins and minerals (Global Salmon Initiative, 2021).

While increased demand is undoubtedly positive for salmon farmers, one of the major risk factors in production is the highly volatile salmon spot price. This makes planning decisions challenging and represents increased economic risk for the firms operating within the industry (Bloznelis, 2018). In fact, salmon price volatility has more than doubled over the last 10 years and is now higher than many comparable commodities (Asche et al., 2019). Gaining a more complete understanding of what causes price fluctuation, and ultimately producing more accurate and reliable forecasts of salmon spot prices could provide huge economic benefits throughout the salmon farming value chain. For example, producers could adjust short-term supply by harvesting more fish when the price is high, and less when the price is low, allowing them to capitalize on high prices.

1.2 Research Question

In this master thesis we will attempt to accurately and reliably forecast the salmon spot price represented by the Fish Pool Index in NOK per kg by applying different tree-based prediction models. We will assess the statistical accuracy as well as the potential economic value of our models in addition to investigating which explanatory variables are most important in creating price forecasts. We define the following research question:

Can tree-based prediction models produce accurate and reliable monthly forecasts of the Fish Pool Index 12 months ahead, and what may be the potential economic value of such forecasts?

2. Background and Literature

In this chapter, we will take a closer look at the global and Norwegian salmon farming industry, and then go through previous academic work within salmon price research and tree-based models in commodity price prediction. We will use the lessons from this chapter later when we build our own models.

2.1 Salmon Farming Industry

To do intelligent salmon price forecasting, we need to gain an understanding of how the salmon industry works. We will start by exploring the salmon farming industry in a global perspective, and then take an even closer look at the Norwegian industry.

2.1.1 Global Salmon Farming Industry

Traditionally the vast majority of food production has taken place on land, however this seems to be changing as both the per capita food intake and seafood consumption is steadily rising. As expressed in *figure 1,* from 1961 to 2019 the global annual seafood consumption per capita has increased from 8.9 kg to 19.8 kg, an increase of more than 120%. The average proportion of protein intake accounted for by seafood has also increased, from 4.4% to 6.7% in the same period (Ritchie & Roser, 2021).

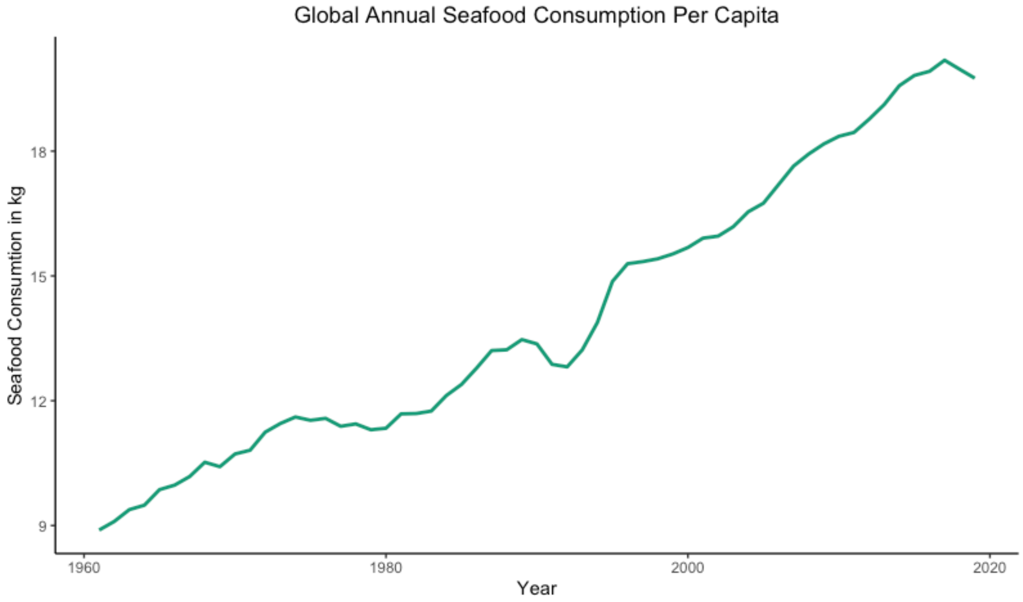


Figure 1: *Global annual seafood consumption per capita from 1961 to 2019 (Ritchie & Roser, 2021).*

With increased demand follows higher production, as evidenced by the rise of industrial aquaculture. Historically most of the seafood supply has come from wildly caught animals, but since 1970 there has been exponential growth in output from industrial aquaculture. Today aquaculture has surpassed wild fishing as the biggest supply source of seafood, as shown in *figure 2* below.

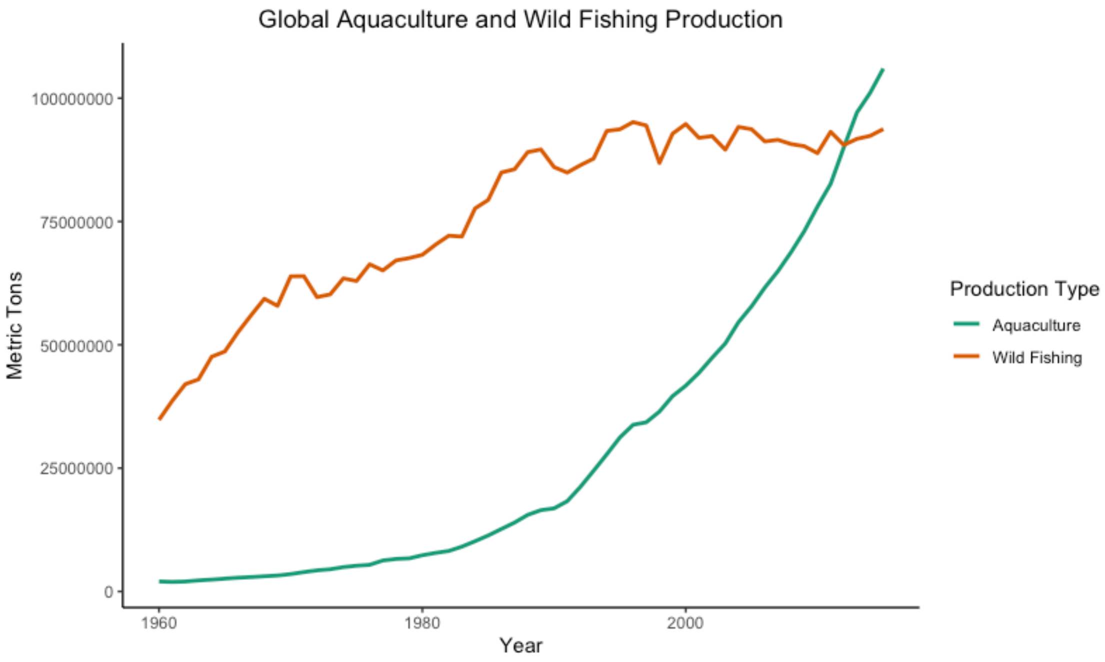


Figure 2: Global aquaculture and wild fishing production from 1960 to 2015 (Ritchie & Roser, 2021).

An increasingly important element of aquaculture production is Atlantic salmon. Since 2004 global harvest volumes of Atlantic salmon has increased by over 5% per year on average, and in 2021 production reached 2 895 000 tons WFE (whole-fish-equivalent), an all-time-high (Kontali Analyse, 2022). The salmon farming industry is expected to grow further in the future as increased population growth and changing consumer food preferences will most likely increase demand.

Because salmon requires a certain range of sea temperatures for optimal growth, the industry is dominated by a few countries with geographical locations well suited for salmon farming. The highest producing countries are Norway, Chile, the United Kingdom, Canada, and the Faroe Islands, accounting for some 93% of global production. The largest markets are the European Union and the United Kingdom making up 45% of the global market, United States with 22%, and Japan with 2.5% (Kontali Analyse, 2022).

2.1.2 Norwegian Salmon Farming Industry

Norway is by far the largest producer of Atlantic salmon. In 2022, 1 532 000 tons of Atlantic salmon was harvested in Norway, accounting for 53% of the global production. Norwegian companies also dominate the industry. Eight of the largest 15 salmon farmers in the world are Norwegian, with Mowi, Lerøy, and SalMar being the biggest both in terms of harvest volumes (Kontali Analyse, 2022), revenues, and market capitalizations (Euronext$_{1, 2, 4}$, 2023). These three companies alone account for nearly a third of the global harvest volume (Kontali Analyse, 2022).

Salmon farming is an increasingly important industry for the Norwegian economy. In 2020 salmon accounted for 69% of total Norwegian seafood exports (Albertsen et al., 2021), and the industry is highlighted as a potential growth area as Norway looks to reduce its dependence on petroleum exports. In 2021 close to 9 000 people were directly employed in the industry with many more jobs created in adjacent and supporting industries as well (Fiskeridirektoratet, 2021).

In the last decades there has been considerable consolidation in the industry. In 2000 there were 296 production companies (Asche et al., 2019), while in 2021 there were only 166 (Fiskeridirektoratet, 2021). This trend is also visible in sales. In 2010, the 10 largest companies accounted for a third of all sales, while in 2016 that proportion had doubled to two thirds (Asche et al., 2019). Harvest volumes of Norwegian farmed salmon are also quite concentrated. The three largest producers in Norwegian waters, Mowi, Lerøy, and SalMar, accounted for close to 50% of the total Norwegian harvest in 2021 (Kontali Analyse, 2022). In summary, Norwegian supply is now mostly driven by a few large companies.

2.2 Fish Pool Index

In 2006, the Fish Pool Index (FPI) was established, introducing a reference price of farmed Atlantic salmon which is widely used to settle futures salmon contracts. It is primarily owned by Oslo Børs ASA and was licensed by the Norwegian Ministry of Finance to operate as a regulated marketplace for fish and seafood derivatives. This means Fish Pool does not offer physical trading of fish, but rather financial contracts that are settled based on the spot price. The price index is published weekly and includes the current spot price illustrated in *figure 3*, in addition to forward-looking prices reflecting the expectation for the coming months.

Furthermore, the Fish Pool index is composed of two elements: the Nasdaq Salmon Index with a 95% weight, and Norwegian export prices from SSB with a 5% weight. The Fish Pool index includes the following salmon weight classes: 3-4 kg with a 30 % weight, 4-5 kg with a 40% weight and 5-6 kg with a 30% weight (Fish Pool$_1$, n.d.).

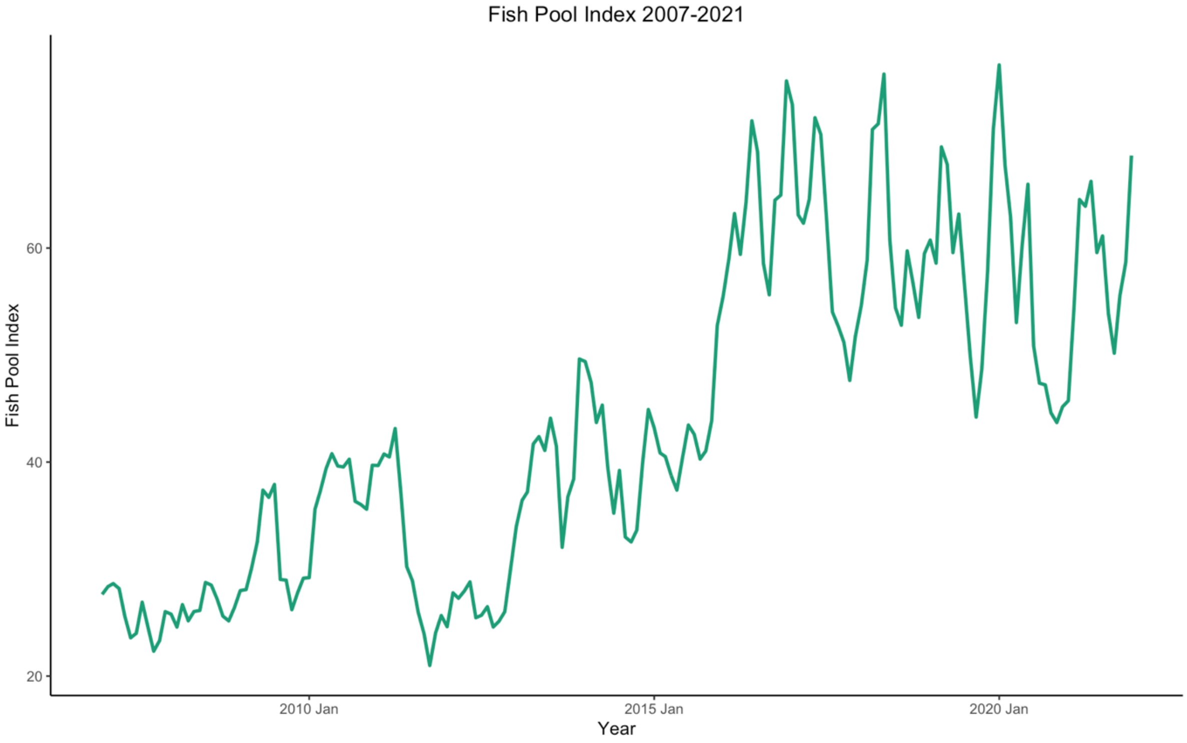


Figure 3: Development of the Fish Pool Index from 2007 to 2021 (Fish Pool$_2$, 2023).

Fish Pool financial contracts are primarily used by salmon farmers, -exporters, -importers, -processors and -retailers to hedge their salmon price risk. The total trading volume of 2021 was 72 336 tons (Fish Pool$_2$, 2023), which corresponds to about 4.7% of the total Norwegian salmon production of 1 546 000 tons (Fiskeridirektoratet, 2022). The spot price is characterized by high volatility with prices ranging from above 60 NOK per kg in January 2021 to about 45 NOK per kg in October the same year. Considering that 90% of the Norwegian Atlantic salmon production is sold at spot price as opposed to futures contracts (Ankamah-Yeboah et al., 2017), this demonstrates the importance of the spot price in determining value creation.

2.3 Literature Review

The literature on predictive models for the Atlantic Salmon spot price appears to be somewhat scarce. There are however some important contributions, and in this section we will go through existing literature. Finally, we will summarize what key lessons we derive from this, which we will incorporate in our own modelling.

There are two main branches of salmon price research that are of interest to us: predictive models to directly forecast salmon price, and research on salmon price volatility. In addition to this, we will investigate the literature on tree-based prediction models used to forecast prices of other commodities such as oil and gold.

2.3.1 Salmon Price Forecasting

Bloznelis (2018) provides the most recent and arguably most important contribution to research on salmon price forecasting. He establishes a benchmark for short-term price forecasting of one to five weeks and explores 16 alternative forecasting methods. He includes four exogenous variables in the models as predictors, including salmon export volume, share prices of salmon farming companies on the Oslo Stock Exchange, the EUR to NOK exchange rate, and salmon futures prices. The best predictions are produced by the k-nearest neighbour method for 1 week-ahead, vector error correction model for two and three weeks-ahead, and futures prices for four and five weeks-ahead. He finds that even though the nominal gains in forecast accuracy over a naïve benchmark are small, the economic value of the forecasts are significant, suggesting that implementing a trading strategy for timing sales based on price forecasts could increase the net profit of a salmon farmer by around 7%.

Dahl et al. (2021) do not predict the salmon price explicitly, but rather explore the relationship between the Fish Pool Index (FPI) and stock prices of major publicly traded salmon companies through cointegration analysis. They document that stock prices reflect salmon price information earlier than the FPI, identifying a possible source of bias in the salmon futures pricing design that relies on the index. The effect is found to be greater for large companies, which means that movements in stock prices of large salmon farmers could contain predictive power on the FPI. The authors explain that one of the reasons behind this effect may be that the FPI reflects current supply- and demand factors, while stock prices are forward-looking, discounting future supply- and demand information.

Guttormsen (1999) uses six easily applicable procedures to forecast weekly producer prices for salmon of different weight classes. This paper focuses on univariate time series-methods, and models used were classical additive decomposition, Holt-Winters exponential smoothing, autoregressive moving average, vector auto regression, and two naïve benchmarks. Evaluating forecasts by mean percentage error, mean absolute percentage error (MAPE), and ratio of accurate forecasts, he finds that classical additive decomposition performed best in forecasting the up- and down-direction of price movements, correctly predicting the price direction 70-90% of the time on 4-12 weeks-ahead forecasts. Vector auto regression performed best according to accuracy measures, producing a MAPE of 1.20%-1.78%. One key finding is the importance of detrending and deseasonalizing salmon price time series data when producing forecasts, as salmon prices exhibit yearly seasonality.

In Anderson & Gu (1995) an approach that combines seasonality removal with a multivariate, state-space, time series forecasting model is developed to provide short-term forecasts for the United States salmon market. Four versions of the state-space forecasting model are compared in terms of their performance on out-of-sample forecasts by the MAPE. Out-of-sample 3-, 6-, and 12-month-ahead directional predictions are generated to test performance in terms of direction. Empirical results indicated that deseasonalization improved the overall performance of the state-space model, and as a result, a linear, deseasonalized state-space forecasting model was selected to provide 12 months-ahead out-of-sample forecasts. The best model produced an out-of-sample forecast MAPE of 5.48%.

2.3.2 Salmon Price Volatility

The literature on salmon price volatility is comparatively more plentiful and recent than that of price forecasting. There are two articles in particular that offer insight into the mechanics of salmon price volatility.

The first is Asche et al. (2019) who find that salmon price volatility, measured as the standard deviation of log-returns, has more than doubled over the last 10 years, and is now higher than many comparable commodities. Having established that, the article then investigates possible explanations of this phenomenon and conclude that the likely cause is reduced short-run elasticity of supply. Three major developments in the salmon market supply-chain provide support for this hypothesis: consolidation of farming companies into fewer and larger units, a premium on fixed harvest schedules to satisfy retail demand for stability, and restrictions on new production capacity in conjunction with strong demand promoting “race to raise” harvest policies.

Bloznelis (2016) uses ARMA-GARCH and dynamic correlation models on weekly data from 1995 to 2013 to examine the behavior of weight-class specific prices. He identifies two periods of different volatility regimes, before and after 2006, and finds that both volatility and conditional correlations increased after 2006. Specifically, he calculates that the standard deviation of log-returns on spot price more than doubled from 3% before 2006 to 7.3% after 2006. Several possible reasons are offered to explain why this occurred specifically around 2006. First, the introduction of maximum allowable biomass constraints in Norway in 2005 put a hard constraint on supply growth. Second, the opening of Fish Pool futures and options exchange in 2006, and third, the ISA-crisis in Chile from 2007-2016 which caused demand for Norwegian salmon to increase substantially. Compared to cattle, wheat and other agricultural commodities, salmon price volatility has been exceptionally high in the latest period. The article suggests relevant factors that could help explain this, including volatility in supply, volatility in exchange rates (because most of Norwegian harvest volumes are exported, and most transactions are invoiced in foreign currencies), and prices of substitutes such as beef, pork and chicken, as well as other factors that may influence demand for salmon.

2.3.3 Tree-based Models in Commodity Price Forecasting

Tree-based models have shown promise in predictive tasks in recent years. In Chen & He (2019) the authors try to develop decision trees to predict WTI crude oil spot-prices and compare their accuracy to benchmark models such as multiple linear regression and ARIMA. They used a dataset spanning from January 1992 to December 2017, and included eight exogenous predictors, including crude oil demand and supply, monthly GDP, CPI, USD Exchange Index, and United States Federal Reserve Interest Rate. Their main finding is that the decision tree models are expected to have higher forecasting accuracy than the benchmark models. Random forest was the best performer with a mean absolute error (MAE) of 1.25 compared to MAEs of above 2.8 for both multiple linear regression and ARIMA. In addition to improved forecast accuracy, the tree-based models also gave a clear understanding of which of the predictors were the most important in price prediction, showing that the 1-month lagged WTI price provided the largest incremental reduction in MSE. The second most important predictor was the USD exchange rate index.

Fattah et al. (2021) aim to develop univariate tree-based models and compare their accuracy to ARIMA as a benchmark. They used decision trees, random forest, and gradient boosted trees to predict monthly gold prices. The dataset ranged from November 1989 to December 2019 and consisted of 362 observations. 90% of the data were used as a training set, while the rest was used to test forecast accuracy. They also find that random forest produced the best accuracy with a root mean squared error (RMSE) of 38.5 compared to ARIMA with an RMSE of 75.46. The authors explain that tree-based methods can overcome problems of forecasting non-linear and non-stationary time series data.

Baser et al. (2023) attempts to predict daily gold commodity prices specifically using tree-based models. Models used were decision trees, adaptive boosting (AdaBoost), random forest, gradient boosting, and extreme gradient boosting (xgBoost). Four metrics were used to evaluate models: RMSE, MAE, MSE, and R^2. Gradient boosting produced the best forecasts according to all performance metrics, with a MAE of 0.47, which was marginally lower than the MAE of xgBoost and random forest, but significantly lower than that of decision trees and AdaBoost. The authors conclude that tree-based models demonstrated "astounding" potential in regression problems to forecast the future price of gold.

2.3.4 Key Takeaways from Literature Review

What follows is a short summary of key lessons we have learned from the literature, and factors we will try to incorporate in our predictive models.

- Salmon prices exhibit 1-year seasonality patterns. Both Bloznelis (2018), Guttormsen (1999), and Anderson & Gu (1995), all find that deseasonalizing their price time series data generally improves forecast accuracy. We will explain how we account for this seasonality in more detail in section *3.6 Time Series Decomposition*.
- Salmon prices are highly volatile compared to other commodity prices. This is the consensus finding by Asche et al. (2019) and Bloznelis (2016) and means that our models must be flexible and able to capture non-linearities in the time series.
- Salmon farmers' stock prices may contain predictive power on the Fish Pool Index. This is the main message from Dahl et al. (2021). Consequently, we will include Oslo Stock Exchange's Seafood Index as a predictor in our models.
- Asche et al. (2019) and Bloznelis (2016) both seem to suggest that short-term supply inelasticity is one of the main causes of salmon price volatility. In other words, supply

is not able to quickly adjust to changes in demand. This motivates the inclusion of demand factors as predictors in our models. Demand factors we will include are prices of substitute protein sources, consumer purchasing power, and salmon import statistics. Other factors assumed to influence salmon prices include environmental factors, such as sea temperatures, biological factors, like sea lice, other diseases, treatments and feed consumption, and financial factors, including exchange rates and stock prices.

- There is a difference between the statistical and economic value of forecasts. As Bloznelis (2018) shows, even marginally better forecasts in terms of statistical accuracy could translate to huge economic value. Thus, we will evaluate our forecasts both based on statistical measures and potential economic value.
- Tree-based methods has demonstrated great potential in both oil and gold price predictions as suggested by Chen & He (2019), Fattah et al. (2021), and Baser et al. (2023), and to our knowledge these methods are entirely unexplored in regard to salmon spot price predictions. Therefore, our work will provide a new contribution to salmon price forecasting.

3. Methodology

In this thesis we will attempt to forecast the salmon price using tree-based prediction models. Specifically, we will use decision trees, random forest and xgBoost. We choose these models because they are flexible and able to capture non-linearities in the time series, as emphasized in section *2.3.4 Key Takeaways from Literature Review*. These models involve segmenting the predictor space into many simple regions. To make predictions we normally use the mean of the training observations in the region to which it belongs (Hastie et al., 2013). As the name suggests, tree-based models can be visualized as a tree structure. The node at the top of the tree is called the root node, while the nodes in the tree that have branches beneath them are called internal nodes or splits. The nodes at the bottom of the tree are called terminal nodes or leaves (Nielsen, 2016). One example is shown under section *3.2 Decision Tree* in *figure 4.*

In this chapter we will define a simple benchmark model, then explain in further detail how the tree-based models work and illustrate some advantages and challenges of using this type of models to do time series forecasting. We will also elucidate how we assess which variables are most important in predicting the Fish Pool Index. Finally, we describe how we aim to evaluate the quality of the models we develop.

3.1 Simple Benchmark Model (Seasonal Naïve)

Before delving into the more complex tree-based models, it may be useful to define a very simple forecasting model that will serve as a benchmark with which to compare the more complex models. If the more complex models fail to beat this simple benchmark based on the evaluation metrics that are described in section *3.9 Evaluation of Forecasts*, one would be better served by employing the simple model.

In this analysis we will use the seasonal naïve model (SNAIVE) as the simple benchmark. In the SNAIVE-model, each forecast is set to be equal to the last observed value from the same season. For instance, with monthly data the forecast for next January is set equal to the observed value from last January, the forecast for next February is set equal to the observed value from last February, and so on. This is expressed mathematically in *equation (1)* below.

(1)

$$\hat{y}_{T+h|T} = y_{T+h-m(k+1)}$$

In this equation m is the seasonal period, and k is the integer part of $\frac{(h+1)}{m}$, that is the number of complete years in the forecast period prior to time $T + h$ (Athanasopoulos & Hyndman, 2018).

3.2 Decision Tree

Decision tree is the simplest kind of tree-based model we will use in this analysis. We apply this algorithm because it is easily interpretable, able to handle complex non-linear relationships in data, and do not require feature scaling or normalization (Hastie et al., 2013). In addition, decision trees will serve as another simple model with which to compare the more complex random forest and xgBoost models. Decision trees can be used both for classification and regression purposes, but as we want to forecast a numerical variable we will limit our discussion to regression trees.

Decision trees utilize a "top-down, greedy" approach known as recursive binary splitting. Recursive binary splitting works by selecting the predictor X_j and the cutpoint *s* such that splitting the predictor space into {X | $X_j < s\}$, the region of predictor space in which X_j takes on a value less than s, and {X | $X_j \geq s\}$, the region of predictor space in which X_j takes on a value greater or equal than s, leads to the greatest possible reduction in prediction inaccuracy. As an example, one of the X-variables we will use is monthly Norwegian harvest volumes of farmed salmon. A decision tree will then split the range of harvest observations based on what split best predicts the response variable, which in our case is the Fish Pool Index.

Generally, for any predictor *j* and any cutpoint *s*, we define the partition as in *equation (2)*:

(2) $$R_1(j,s) = \{X|X_j < s\} \text{ and } R_2(j,s) = \{X|X_j \geq s\},$$

and we seek the values of *j* and *s* that minimize *equation (3)*:

(3)

$$\sum_{i:\, x_i \in R_1(j,s)} (y_i - \bar{y}_{R1})^2 + \sum_{i:\, x_i \in R_2(j,s)} (y_i - \bar{y}_{R2})^2 \,,$$

where $\hat{y}_{R_1}$is the average of the response variable for the training observations in R_1 and $\hat{y}_{R_2}$is the average of the response variable for the training observations in R_2 (Hastie et al., 2013).

In the harvest volume example, the observations range from 74 545 to 165 760 tons (see *table 4*). One could imagine a decision tree defining a split s = 122 152, which would divide the harvest volume variable in two roughly equal parts: R_1 where harvest volume < 122 152, and R_2 where harvest volume ≥ 122 152. The predicted value of the Fish Pool Index would then be the average observed FPI-values in R_1 and R_2 respectively.

Next, the process is repeated, but instead of splitting the entire predictor space, we split one of the two previously identified regions, R_1 or R_2. Now we would have three regions, and again we would split one of these three regions further with the same optimization problem as above. In this way we build a tree-like structure as illustrated in *figure 4* below, with subgroups of predictor space until a stopping criterion is reached. In our case the stopping criterion will be the so-called "complexity parameter", which is the minimum improvement in the model needed at each node.

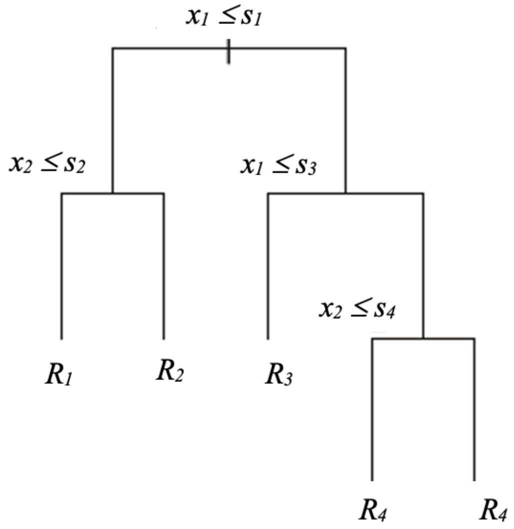


Figure 4: Illustration of how recursive binary splitting leads to a tree-like model structure (Hastie et al., 2013).

To forecast with decision trees, we use the rpart- and caret-packages in R. rpart is a package that contains the decision tree-algorithm, while the caret-package is chosen because it offers an easy way to optimize hyperparameters when developing predictive models. In this instance we conduct a grid search to find the optimal value of the complexity parameter which determines the number of nodes in the decision trees. The optimal value of this parameter will be determined by rolling-origin cross validation as explained in section *3.7 Rolling-Origin Nested Cross Validation*. This involves creating a trainControl-object that will define the cross validation-process. Finally, we use the train-function to build the decision trees.

3.3 Random Forest

While decision trees have several advantages, including simplicity and interpretability, a commonly cited disadvantage is that single decision trees generally do not have the same level of predictive accuracy as more complex models, and are prone to overfitting the training data set. However, by aggregating many decision trees the predictive performance can be substantially improved, and the risk of overfitting to training data reduced. One of the methods to do so is random forest. Compared to xgBoost, an advantage with random forest is that it is known to be more efficient with large datasets and that it requires less hyperparameter tuning (Hastie et al., 2013). These reasons explain why we employ the random forest algorithm in our analysis.

Random forest is based upon "bagged" decision trees. Bagging is the method of bootstrap aggregation, a general-purpose procedure for reducing the variance of a statistical learning method, thereby increasing prediction accuracy. This first involves bootstrapping, which means taking repeated samples from a single training data set with replacement. In our case we will generate *B* different bootstrapped training data sets. Then a different decision tree is fit on each of the *B* training data sets to obtain the best tree $\hat{f}^{*b}(x)$ given that particular dataset. Predictions are made by averaging all the predictions made by the *B* different trees as presented in *equation (4)* below.

(4)

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}\,(x)$$

In our random forests model, we apply a method very similar to bagging, but with a small tweak that decorrelates the trees. We build *B* different decision trees, and each time a split is made, a random sample of *r* predictors is chosen as split candidates from the full set of *p* predictors. Then the split is made by using one of those *r* predictors. At every split a new sample of *r* predictors is chosen as split candidates. Predictor subsampling means that each individual tree has high variance, but low bias. The average of the predictions of many trees still has low bias, but also lower variance (Hastie et al., 2013).

By forcing our model to choose between only a subset of predictors we secure that the correlation between the individual trees is lower. To understand why, imagine an example with one very strong predictor and several weak predictors. If the model was allowed to use the full set of *p* predictors at each split, it would probably choose to split based on the strong predictor very often, and we would end up with *B* regression trees that all looked very similar to each other and produced very similar predictions. In other words, the individual trees would be very highly correlated with each other, and the reduction in variance obtained by averaging predictions would be smaller than if we averaged across uncorrelated predictions. This explains why random forest may perform better than ordinary bagging (Hastie et al., 2013).

To build the random forest models in R, we again employ the caret-package, but this time in conjunction with the randomForest-package. Now it is the mtry-hyperparameter, which corresponds to the *r* parameter introduced above*,* we would like to optimize by rolling-origin nested cross validation. We do this by running a grid search where we try mtry-values from five to 50 with an interval of five. We try a fairly wide range of mtry-values, but still lower than the total number of predictors in the dataset to capture the benefits of predictor subsampling at each split in the trees. In other words, we try 10 different values and allow the train function to find the optimal value. When applying the train-function we specify the method-argument to "rf", which will build random forest-models.

3.4 xgBoost

Another approach for improving the prediction accuracy of a decision tree is called boosting*,* and this technique has spawned several types of boosted tree-models. In this analysis we will focus on one application of boosting, namely Extreme Gradient Boosting, or xgBoost as it is also commonly known. We use this model because it usually provides more accurate predictions than simple decision trees and random forests, as it applies a differentiable loss function with a regularization term. We also employ this algorithm because it can capture complex relationships and interactions between features, which means it handles non-linear patterns (Hastie et al., 2013). This is valuable to us as it seems the Fish Pool Index demonstrates a high degree of non-linearity, as shown in *figure 3.*

When we apply boosting the trees are grown sequentially, which means that each tree uses information from previously grown trees by fitting new decision trees to the residuals from previously grown trees. This means a tree is fit using the current residuals rather than the outcome Y (in our case the Fish Pool Index), as the response. Then, this new decision tree is added into the fitted function in order to update the residuals. By fitting new decision trees to the residuals, our model slowly improves in areas where it was weak originally (Hastie et al., 2013).

xgBoost utilizes boosting, and in addition weights are assigned to all the independent variables, which are then used to build sequential decision trees generating predictions. Variables that the tree is unable to predict receives increased weighting, and these variables are then fed to the next decision tree. These individual decision trees are then averaged to provide reliable predictions.

What makes xgBoost special is that it incorporates a regularized model to prevent overfitting. When we apply xgBoost we minimize the regularized objective of *equation (5)*:

(5)

$$l(\Phi) = \sum_{i} l(\hat{y}_i, y_i) + \sum_{k} \Omega(f_k)$$

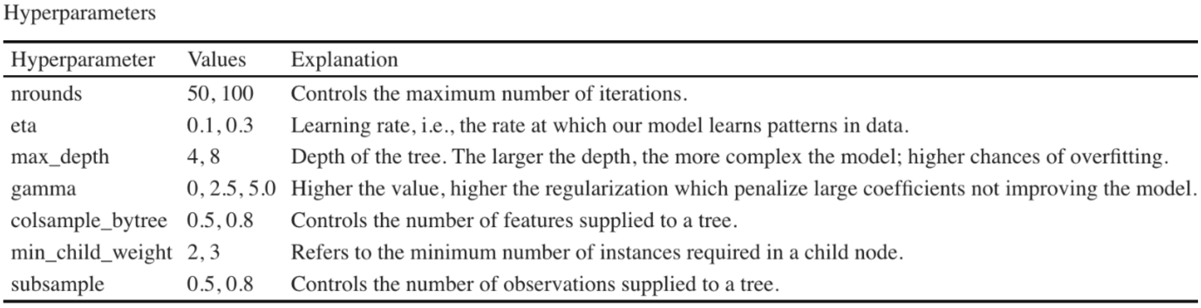
$$Where\ \Omega(f) = \sqrt{T} + \frac{1}{2}\lambda||w||^2$$

In this equation *l* is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and the target y_i. The second term Ω penalizes the complexity of the regression tree functions, and this additional regularization term helps to smooth the final learnt weights w to avoid overfitting. T is the number of leaves in each tree. Each f_k corresponds to an independent tree structure q (which represents the structure of each tree that maps an example to the corresponding leaf index) and leaf weights *w*. λ is the regularization parameter and determines how much more complex models are penalized compared to simpler ones. The regularized objective will tend to select a model employing simple and predictive functions (Chen & Guestrin, 2016).

In addition to this regularized objective, two techniques are used to further prevent overfitting. The first is shrinkage which scales newly added weights by a factor η after each step of tree boosting. Shrinkage reduces the influence of each individual tree and leaves space for future trees to improve the model. The second technique is predictor subsampling, the same technique as described above in section *3.3 Random Forest* (Chen & Guestrin, 2016).

To train the xgBoost-models, we again employ the caret-package in R, this time in combination with the xgboost-package. Our xgBoost model has several hyperparameters that need to be optimized through the grid search method. Ideally, we would prefer to run a much larger grid search, but we are constrained by computational resources. Therefore, we attempt to choose the most relevant values for all hyperparameters only. The chosen values and the explanation of each hyperparameter is presented in *table 1* below.

Table 1: Overview of hyperparameters in the xgBoost model (Kuhn, 2019).



Hyperparameters

Hyperparameter	Values	Explanation
nrounds	50, 100	Controls the maximum number of iterations.
eta	0.1, 0.3	Learning rate, i.e., the rate at which our model learns patterns in data.
max_depth	4, 8	Depth of the tree. The larger the depth, the more complex the model; higher chances of overfitting.
gamma	0, 2.5, 5.0	Higher the value, higher the regularization which penalize large coefficients not improving the model.
colsample_bytree	0.5, 0.8	Controls the number of features supplied to a tree.
min_child_weight	2, 3	Refers to the minimum number of instances required in a child node.
subsample	0.5, 0.8	Controls the number of observations supplied to a tree.

Next, we will allow the train-function to find the optimal values, and we specify the method-argument to “xgbTree” to employ the xgBoost-algorithm.

3.5 Variable Importance

Another advantage of using tree-based models is the availability of variable importance measures. Gaining a more complete understanding of what causes salmon price fluctuations could prove valuable for salmon farmers in their decision making. Variable importance attempts to measure each feature’s relative importance when developing the tree model, or in other words, how important that particular feature is for the predictive accuracy of the model. So, for each tree model, each feature will receive a variable importance score. The variable importance calculation can be done in multiple ways, and each method has advantages and drawbacks, but in this analysis we will use the varImp-function from the caret-package. The advantage of this function is that it scales the variable importance in such a way that the maximum value is 100, enabling easier comparisons across different models. Because each model contains 12 different models, one for each forecast horizon, we average the variable importance over these 12 models to obtain the final variable importance scores. Further explanation of why we employ 12 different models can be found in section *3.8 Direct Forecasting*.

Variable importance is computed differently for the different tree models. For decision trees, where we use the rpart-package in R, the reduction in the loss function attributed to each feature at each split is calculated, and the sum is returned. Since there may be candidate variables that are important but are not used in a split, the top competing variables are also included at each split. For random forest, the function calculates the prediction accuracy measured by mean squared error (MSE) on the out-of-bag portion of the data for each tree. Then the same is done after permuting each feature. The difference between the two accuracies is then averaged over all trees and normalized by the standard error. In xgBoost, variable importance is based on a “gain”-measure. Gain indicates the contribution of each feature to the model by measuring the reduction in node impurity, meaning how well the trees split the data. Each gain of each feature is summarized in each tree, and then averaged over the number of trees (Kuhn, 2019). Considering that the variable importance is computed differently for each tree model, we will be careful with direct comparisons across models. However, we are mostly interested in which variables stand out as the most important, not necessarily the numerical value of the variable importance measure.

3.6 Time Series Decomposition

We know from the literature on salmon prices that there exist 1-year seasonal patterns in the price movements, as Bloznelis (2018), Guttormsen (1999), and Anderson & Gu (1995) all find that deseasonalizing their price data generally improves forecast accuracy. We will attempt to account for seasonality using a time series decomposition method. Decomposing a time series involve splitting it into several components, each component representing a different pattern in the series. There are usually three components to a time series: trend-cycle, season, and a remainder.

The trend-cycle component is the long-term pattern of the time series, this could for instance be an up- or down-movement over time. The seasonal component is a systematic, calendar-related variation in the time series. One relevant example could be the higher harvest volumes of salmon in autumn due to higher salmon growth in summer with warmer sea temperatures. Salmon harvest volumes could thus be said to exhibit seasonality. The remainder term is whatever is left after calculating the trend-cycle and season-components. An additive time series decomposition would take the form as presented in *equation (6)*.

(6)
$$FPI_i = y_i = T_i + S_i + R_i$$

In this equation y_i is the observation at time i, T_i is the trend-cycle component, S_i is the seasonal component, and R_i is the remainder (Athanasopoulos & Hyndman, 2018).

In this analysis we will apply a method known as STL decomposition. STL is an acronym for “seasonal and trend decomposition using Loess”. The STL method has several advantages: it will handle any type of seasonality, the seasonal component is allowed to change over time, and the rate of change can be controlled. In addition, the smoothness of the trend-cycle can be controlled. Also, the method can be robust against outliers, so that a few unusual observations will not unduly affect the estimates of trend-cycle and season (Athanasopoulos & Hyndman, 2018).

To produce deseasonalized forecasts, we will first perform STL decomposition on the Fish Pool Index observations in our dataset, then remove the seasonal term from these observations, which will be S_i from the equation above. Next, we will utilize this new deseasonalized Fish Pool Index as the response variable in the tree-based prediction models. After predictions are made, we will simply add the seasonal component back to the predictions and obtain the final forecasts. This process is summarized below:

1. Perform STL decomposition on the Fish Pool Index (FPI) time series and obtain an equation of the form presented in *equation (6)*.
2. Produce a deseasonalized response variable $y_i^{ds} = y_i - S_i$.
3. Use y_i^{ds}as the response variable in the decision tree, random forest and xgBoost, and generate predictions $\hat{y}_i^{ds}$.
4. Add back the seasonal component to obtain the final forecast, $\hat{y}_i = \hat{y}_i^{ds} + S_i$.

To perform the STL decomposition in R, we use the STL-function from the feasts-package.

3.7 Rolling-Origin Nested Cross-Validation

Cross-validation is a standard technique used in predictive analytics for the purposes of testing and improving model quality. As we are dealing with time series data, traditional cross-validation techniques may become problematic. This is because of the temporal dependencies in time series, which means one must ensure that all observations in the training data set occurs chronologically before all the observations in the testing data set. To accurately simulate real-world forecasting, we cannot use information from the future to forecast said future. It also does not make sense to use information from the future to predict values from the past. At any given point in time, we must utilize information available at that point in time and produce forecasts into the future.

In this analysis we will employ a method called rolling-origin nested cross-validation. This is based upon nested cross-validation which builds an outer loop for error estimation and an inner loop for parameter tuning. The inner loop works by splitting the training set into a training subset and a validation set. The model is then trained on the training subset, and the predictive accuracy is tested against the observations in the validation set. The hyperparameter values that minimize forecast error on the validation set are chosen. There is also an outer loop, which splits the dataset into multiple different training and test sets. The error on each split is then averaged in order to compute a robust estimate of model error (Simon & Varma, 2006).

Rolling-origin nested cross-validation also involves successively updating the forecasting origin and producing forecasts from each new origin (Tashman, 2000). In other words, we successively update the test set while assigning all previous data into the training set. To make the most of our data, each split into training and validation sets is moved chronologically forward by one single observation. *Figure 5* below is an illustration of this process.

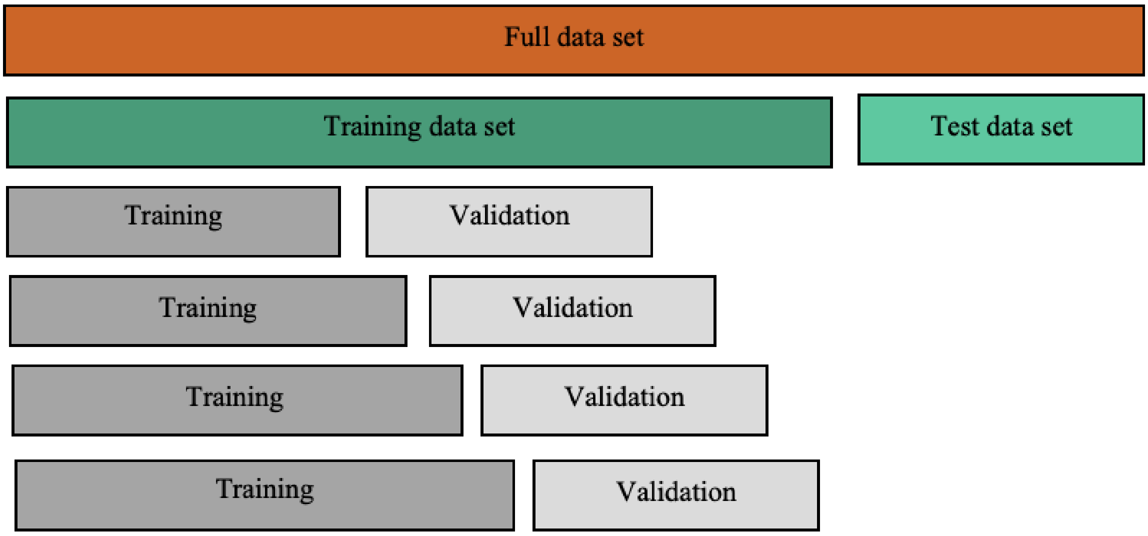


Figure 5: Illustration of rolling-origin nested cross validation (Petropoulos & Svetunkov, 2018).

To implement rolling-origin nested cross validation in R, we create a trainControl-object with the caret-package. We specify that the method should be “timeslice”, and we choose the value of the initial window-parameter to be 36. We choose 36 months, three full years, to ensure enough data when training the first model and to allow the model to capture the one-year seasonal pattern. This means that the first model will be trained on the 36 first observations in the dataset, while the next model will be trained on the 37 first observations, and so on until the whole training data set is used. We set the horizon-parameter to one, meaning that we evaluate forecasts based on one test observation. We then use this trainControl-object to specify the cross-validation procedure in the train-function when building the models.

3.8 Direct Forecasting

When making multi-period time series forecasts, we face a choice between different forecasting methods. One option is the so-called recursive method. It involves identifying and fitting an initial model to the time series and producing multi-step ahead forecasts by a sequence of one-step ahead forecasts. At each new forecast horizon, previously made forecasts are plugged in to replace unknown future values (Bhansali, 1999). In this analysis we are interested in producing monthly predictions one year ahead, and therefore we will create a 12-step ahead forecast. For example, if we were to employ the recursive method and produce a 12-step ahead forecast in December 2020, the December 2021-forecast would depend heavily on the forecasted values in January 2021, February 2021, …, November 2021.

Another option is to do what is known as direct forecasting. Direct forecasting entails building a separate model for each forecast horizon. This means that for a 12-step ahead forecast, we would build 12 different models, each one with a different response variable (Bhansali, 1999). We are interested in making monthly predictions, so the response variable in the first model would be the one-month lead value of the Fish Pool Index (y_{m+1}), the second would be the two-month lead value (y_{m+2}), and so on. The first model would then produce the one-step ahead forecast ($\hat{y}_{m+1}$), the second model would produce the two-step ahead forecast ($\hat{y}_{m+2}$), and so on until one reaches the end of the forecast horizon.

In this thesis we choose to use the direct forecasting method because of two main reasons. First, we want our forecasts to be robust against model misspecification. The biggest risk with recursive forecasts seems to be error propagation. If a mistake is made when building the one-step ahead forecasting model, this will cause a chain of errors throughout the forecasting horizon. On the other hand, a direct forecasting method builds a new model for each forecast horizon and is thus less sensitive to poor models on any specific horizon (Marcellino et al., 2006).

The second reason we choose the direct method is because it is a more practical one. We are building multivariate models, which means that to do recursive forecasting we would need to forecast not only the dependent variable (the Fish Pool Index) recursively into the future, but also all the feature variables. This seems like a complex and high-risk strategy, as our forecasts would not only be dependent on forecasted values of the Fish Pool Index, but also on forecasted values of all features as well.

3.9 Evaluation of Forecasts

To evaluate the quality of our models, we will test their forecasting accuracy against the actual observations of the Fish Pool Index in 2021. In other words, we will use data from December 2020 to forecast the FPI throughout 2021, and then evaluate how close our forecasts are to the actual observations of the FPI. It is important to emphasize that these are out-of-sample forecasts, as we do not use data from 2021 to train our models. All data from 2021 are in the test data set. We explain this further in section *4.4.3 Train and Test Split*. We will use three statistical accuracy measures, as well as evaluate the potential economic value of the forecasts to salmon farmers.

3.9.1 Evaluation of Statistical Forecast Accuracy

The first two statistical accuracy measures will evaluate the distance between observations and forecasts. The closer these measures are to zero, the more accurate predictions. The third will measure the accuracy in predicting the directional up- or down-movement from month to month, and a higher value is therefore better.

The first statistical measure we will utilize is mean absolute error (MAE). As expressed in *equation (7)*, this is quite simply the mean of the absolute values of the differences between observations and forecasts of the response variable Y, which in our case is the Fish Pool Index.

(7)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Here *n* is the number of observations, y_i is the *i*-th observation of the response variable *Y,* and $\hat{y}_i$ is the *i*-th forecast of *Y.* The advantage of this measure is that the MAE is expressed in the same unit as the observation y_i, so it is very easily interpretable.

The second measure of statistical accuracy we will employ is the mean squared error (MSE). This is the mean of the squared differences between observations and forecasts, which is expressed in *equation (8)* below.

(8)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Here n is again the number of observations, while y_i are observations of *Y*, and $\hat{y}_i$ are the corresponding forecasts of *Y*.

One difference between MSE and MAE is that MSE overweighs large differences between observation and forecast, because the difference is multiplied by itself. The MSE punishes very large forecasting errors more severely than the MAE. Both MAE and MSE measure the Euclidean distance between observation and forecast and will thus provide insight into how close forecasts are to the actual observations of the Fish Pool Index.

Another measure that will prove insightful is the directional accuracy (DA). Instead of measuring closeness to observations, it measures how often forecasts get the directional up- or down-movement of the Fish Pool Index correct. To do this, one approach is to first compute a dummy-variable U_i that takes the value 1 if the observation at time i, y_i, is larger than the observation from the previous period, y_{i-1}, and 0 if y_i is smaller than y_{i-1}. Then we create a similar $\widehat{U}_i$ that is 1 when the forecast at time i, $\hat{y}_i$, is larger than the forecast from the previous period, $\hat{y}_{i-1}$, and 0 when $\hat{y}_i$ is smaller than $\hat{y}_{i-1}$. We can then produce a so-called confusion matrix by counting how often the predicted movements correspond to the actual movements. The form is presented in *table 2* below.

Table 2: Confusion matrix.

	Actual Up-movement	**Actual Down-movement**
Predicted Up-movement	True Positives (TP) (If $\mathrm{U_i} = \widehat{\mathrm{U}}_\mathrm{i} = 1$)	False Positives (FP) (If $\mathrm{U_i} < \widehat{\mathrm{U}}_\mathrm{i}$)
Predicted Down-movement	False Negatives (FN) (If $\mathrm{U_i} > \widehat{\mathrm{U}}_\mathrm{i}$)	True negatives (TN) (If $\mathrm{U_i} = \widehat{\mathrm{U}}_\mathrm{i} = 0$)

In *equation (9)*, we obtain directional accuracy by summing the true positives and negatives, expressed as a percentage.

(9)
$$DA = \frac{1}{n}(True\ Positives + True\ Negatives)$$

Here n is the number of observations. DA then pays no attention to how big the difference between observation and forecast is, instead it just measures the probability that a forecasted up- or down-movement will correspond to the actual movement. If for instance directional accuracy equals 80% and a forecast for the next time period indicates an up-movement (i.e. $y_i < \hat{y}_{i+1}$), we should expect the direction of the forecast to be correct about 8 out of 10 times, that is, $y_i < y_{i+1}$ with 80% probability.

3.9.2 Evaluation of Economic Value of Forecasts

We will also attempt to measure the economic value of the forecasts in the perspective of a salmon farmer by utilizing a similar approach as introduced by Bloznelis (2018). Assume a particular farmer has $\bar{X}_{harvest\ volume}$ volume per month ready for harvesting, which could be done immediately or with a delay of one month at no additional cost. Knowing the direction of a change in the spot price could lead to considerable economic gain if the farmer were to time the harvest to when the price is higher.

If the salmon farmer was to make his harvest decisions at random, then one would expect him to be correct only 50% of the time and the net value added of the forecasts would be zero. This will serve as our benchmark. By comparing how much the salmon farmer could increase revenue applying our forecasts instead of the simple benchmark strategy, we can then evaluate the economic value of our forecasts. This involves first finding the mean absolute price difference $\bar{X}_{price\ difference}$ between two months in our data, which is expressed in *equation (10)* below.

(10)

$$\bar{X}_{price\ difference} = \frac{\sum_{i=1}^{i} |x_i - x_{i+1}|}{n}$$

In this equation x_i is the price of a particular month, and n is the number of observations in the data. Next, in *equation (11)* we compute the potential economic gain per month given a perfect forecast $G_{perfect\ forecast}$.

(11)

$$G_{perfect\ forecast} = \bar{X}_{price\ difference} \times 0.5\ \bar{X}_{harvest\ volume}$$

Now, in *equation (12)*, we utilize our best performing model and calculate potential economic gain per month of timely forecasting associated with this model, $G_{best\ model}$.

(12)

$$G_{best\ model} = G_{perfect\ forecast} \times DA_{best\ model} - G_{perfect\ forecast} \times (1 - DA_{best\ model})$$

Here $DA_{best\ model}$ is the directional accuracy of the best performing model of this study, stating the probability of accurately predicting up- or down-movements in price. This means our forecasts will have economic value if they have directional accuracy above 50%.

4. Data

The data of this study is a collection of variables composed from multiple public sources. In this chapter we will first describe how the dataset is structured and the selection of variables. Next, we present descriptive statistics of the feature variables, which includes correlation with the Fish Pool Index and summary statistics. Finally, we explain several pre-processing steps that was necessary to model the data.

4.1 Introduction to Dataset

The original dataset consists of one response variable and 33 explanatory variables, and spans from January 2007 to December 2021. The data occurs at a monthly frequency and entails a total of 180 observations. Originally, we envisioned more than 180 observations, but we were limited by the fact that complete data were not available further back in time. Another problem we encountered when obtaining data was low data frequency. For example, several possible explanatory variables were only available at annual frequency and could therefore not be included in the analysis.

As the data was collected from multiple public sources, some initial steps had to be performed for the variables to be comparable. For instance, some variables occurred at daily or weekly frequency and were therefore aggregated to monthly frequency. For variables expressed in monetary value such as Fish Pool forward prices, we aggregated by computing the mean, while for variables that expressed volume such as Norwegian exports, we aggregated by summarizing. Furthermore, we aggregated some variables that provided similar information. For example, dead fish, low quality fish, escaped fish and other fish loss was transformed to one variable, as we were interested in the absolute value of fish loss as opposed to the cause of fish loss.

For each variable a hypothesis about how and when it is likely to impact the salmon spot price was derived. Based on this we computed new variables that were lagged in order to capture features that would impact the spot price several months later in time. For instance, at month *m*, a variable that is lagged three months express the observation of that variable at month *m-3*. The number of lags chosen for each individual feature variable will be further described in the next section, *4.2 Variable Selection*.

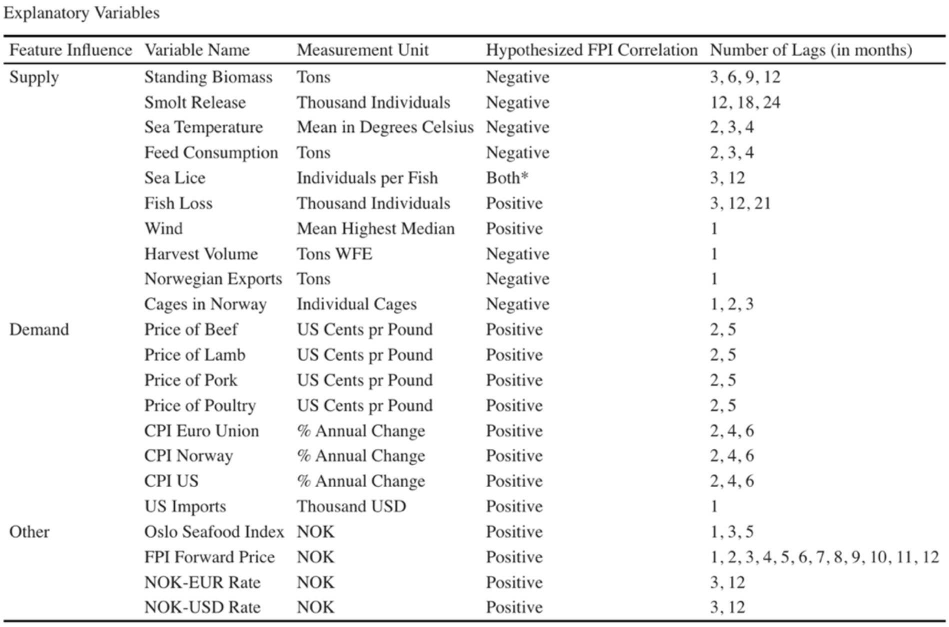
4.2 Variable Selection

The response variable of this study is the Fish Pool Index, which express the salmon spot price in NOK per kg. When predicting this response, the prediction accuracy depends upon the feature variables considered. Variable selection was therefore done through comprehensive research and input from our supervisor, Stein Ivar Steinshamn, and Hans Vanhauwaert Bjelland from Sintef.

Even though the Fish Pool Index express the global spot price, we simplified supply by only including variables based on Norwegian data. This was considered reasonable as Norway is by far the largest farmed salmon producer, accounting for over half of the global supply. With regards to demand, variables were chosen based on the largest markets being the European Union and United Kingdom making up 45% of the global market, followed by United States at 22% (Kontali Analyse, 2022).

An overview of the explanatory variables can be viewed in *table 3*. The figure is organized based upon the variable's influence on either supply or demand. This follows from the interaction of price, supply and demand, as suggested by general market mechanism theory. In addition, some other variables were included which can drive both supply and demand. Below follows an in-depth description of the selection and data-retrieval process of each variable.

Table 3: Overview of explanatory variables.



Explanatory Variables

Feature Influence	Variable Name	Measurement Unit	Hypothesized FPI Correlation	Number of Lags (in months)
Supply	Standing Biomass	Tons	Negative	3, 6, 9, 12
	Smolt Release	Thousand Individuals	Negative	12, 18, 24
	Sea Temperature	Mean in Degrees Celsius	Negative	2, 3, 4
	Feed Consumption	Tons	Negative	2, 3, 4
	Sea Lice	Individuals per Fish	Both*	3, 12
	Fish Loss	Thousand Individuals	Positive	3, 12, 21
	Wind	Mean Highest Median	Positive	1
	Harvest Volume	Tons WFE	Negative	1
	Norwegian Exports	Tons	Negative	1
	Cages in Norway	Individual Cages	Negative	1, 2, 3
Demand	Price of Beef	US Cents pr Pound	Positive	2, 5
	Price of Lamb	US Cents pr Pound	Positive	2, 5
	Price of Pork	US Cents pr Pound	Positive	2, 5
	Price of Poultry	US Cents pr Pound	Positive	2, 5
	CPI Euro Union	% Annual Change	Positive	2, 4, 6
	CPI Norway	% Annual Change	Positive	2, 4, 6
	CPI US	% Annual Change	Positive	2, 4, 6
	US Imports	Thousand USD	Positive	1
Other	Oslo Seafood Index	NOK	Positive	1, 3, 5
	FPI Forward Price	NOK	Positive	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
	NOK-EUR Rate	NOK	Positive	3, 12
	NOK-USD Rate	NOK	Positive	3, 12

* *Negative short-term- and positive long-term correlation.*

Standing Biomass

The Standing Biomass data was retrieved from Fiskeridirektoratet$_2$ (2023) and express tons of salmon in Norwegian facilities at a given month. As valuation of farmed salmon is mainly based upon size distribution and quality, biomass can provide insight into farmers' harvest volume (Mowi, 2020). Considering that salmon is fresh produce, an increase in harvest volume will likely increase short-term supply, which in turn will impact the spot price negatively. This leads to a hypothesized negative correlation between standing biomass and the Fish Pool Index. As to when the effect on the price will be apparent, large manufacturers such as Mowi (2020) and SalMar (2022) suggests that salmon spend about 12 to 24 months in sea before harvested. However, diverse age groups in sea suggests a large proportion of the biomass will impact short-term supply because the oldest fish will be larger and weigh more. We therefore operate with lags of three, six, nine and twelve months.

Smolt Release

The Smolt Release data was also retrieved from Fiskeridirektoratet$_4$ (2023) and indicate release of smolt in Norwegian facilities measured in thousands. Smolt release is the process where juvenile fish transition from an existence in freshwater to a life in the sea. There they live in cages, which are large, enclosed nets suspended in the ocean by flotation devices until the fish are ready for harvest (SalMar, 2022). This usually takes 12 to 24 months (Mowi, 2020), which suggests smolt release may be a good indicator of supply at that time. Because of this, we model with lags of 12, 18 and 24 months.

Sea Temperature

Sea temperature is an important factor for salmon farming. The data exhibits the mean temperature in Norway measured in degrees Celsius and was obtained from Lusedata (Selnæs, 2023). Lusedata graciously provided us with data from 2007 to 2011 that was not publicly available. The remaining data from 2012 to 2021 was available on the Lusedata website. Originally the data occurred at a weekly frequency, but it was later aggregated to monthly frequency by computing the mean. At the ideal sea temperature between 8 and 14°C, farmed Atlantic salmon eats well and grows quickly. However, with rising temperature, the growth rate may increase further and shorten production time. Historically this has been the case in Chile, which at 10°C has the highest mean temperature of the salmon producing regions. If the temperature becomes either too high or too low, the salmon gets stressed, eats less, experiences reduced growth and may even die (Mowi, 2020). Knowing that temperature is an important driver for production time and that warmer temperatures in summer influences harvest volumes in autumn, this suggests a negative correlation with the FPI and a short- to medium-term effect on supply. Consequently, we have introduced lags of two, three and four months.

Feed Consumption

The feed consumption data was retrieved from Fiskeridirektoratet$_3$ (2023) and indicate reported feed consumption at Norwegian facilities measured in tons. Even minor modifications in feeding may to a large extent affect the growth and quality of farmed Atlantic salmon, which in turn will affect supply. Feeding also makes up the largest share of the total cost in production (SalMar, 2022) and naturally the older, bigger fish will consume the largest amount of feed. These are almost slaughter-ready, which suggest that the effect on supply would be short-term. Thus, we operate with lags of two, three and four months. We assume the correlation with the Fish Pool index to be negative as increased supply would presumably have negative effect on the spot price.

Sea Lice

Salmon lice are natural seawater parasites that could threaten fish welfare, damage the quality of the salmon's flesh and in the worst cases, lead to disease and death. Norwegian authorities therefore have clear guidelines for handling of the lice. As the threat increase with the number of lice found in the cage, a maximum number of lice is permitted, and all Norwegian suppliers must count and report on this weekly (SalMar, 2022). The data was retrieved from Lusedata (Selnæs, 2023) and include data from 2007 to 2011 that was not publicly available. It states the average number of individual lice per fish during a month in Norway. Originally the data had a weekly frequency and consisted of three variables: moving lice, fixed lice and adult female lice per fish. These were later aggregated by computing the monthly average. High amounts of sea lice can cause premature harvest, which signal an increase in short-term supply with negative effect on price and reduction of long-term supply with positive effect on price. To capture both effects, we introduced lags of three and twelve months.

Fish Loss

Fish loss is defined as the number of fish either reported as dead, disposed of due to low quality, escaped, or lost due to other causes such as counting errors. The data was retrieved from Fiskeridirektoratet$_5$ (2023) and is expressed in thousands individuals. In Norwegian farming facilities around 15% of the production is usually discarded, which underlines the risk in relation to production volume (Hoddevik, 2023). Fish loss affects salmon of all ages and sizes, thus fish loss influence both short- and long-term supply. An increase in fish loss will affect supply negatively which in turn would increase the spot price, and the hypothesized correlation with the Fish Pool Index is therefore positive. This effect on price is expected to occur with lags of three, twelve and twenty-one months.

Wind

Waves occur when wind blows over the sea surface. This could lead to extreme weather conditions and reduce short-term harvest of farmed salmon as necessary sea transportation from the cages to slaughterhouses becomes challenging. We assume that the corresponding effect on the spot price is positive and occur with a one-month lag. We retrieved the wind data from Norsk Klimaservicesenter (2023). It is expressed as the mean of the highest median wind per month at the following Norwegian locations: Bergen Florida, Bodø Vi, Halten Fyr, Sortland, Svolvær Lufthavn, Tafjord, Vega Vallsjø, Vigra and Ørsta-Volda Lufthamn. These stations were chosen as they are located within areas where much of the Norwegian salmon farming takes place.

Harvest Volume

Harvest volume is the reported harvest of slaughtered salmon in tons whole-fish-equivalent (WFE) from Norwegian suppliers. This data was retrieved from Fiskeridirektoratet$_6$ (2023) and include all fish extracted from the cages, excluding fish that have been moved or sold alive. Harvest is an obvious indicator of very short-term supply as salmon is fresh produce. The effect on price will therefore be almost immediate, hence a one-month lag. In addition, the harvest pattern is largely influenced by seasonal fluctuation, such as warmer months when the fish grows quicker and higher demand during certain holidays (Bloznelis, 2018). Such increased supply will presumably have a negative effect on price, thus the hypothesized correlation with the Fish Pool Index is negative.

Norwegian Exports

Norwegian exports indicate tons of salmon sold internationally from Norwegian producers per month. This data was retrieved from SSB (2023). Norway exports nearly all its production and is the main supplier of salmon in Europe, the largest salmon market in the world (Bloznelis, 2018). Again, because of the short expiry associated with salmon products, exported volume is likely to follow harvested volume closely. For this reason, an increase in Norwegian export is thought to have very short-term effect on supply, thereby influencing the Fish Pool Index negatively with a one-month lag.

Cages in Norway

In Norway, aquaculture is a permit-based industry, meaning the salmon farmers must acquire a legal disposition for commercial salmon production. Because of high demand for such permits, the Norwegian government control the number of allowed cages and when permits are sold. The data express the number of cages with live salmon and rainbow trout in Norway and was acquired from Fiskeridirektoratet[1] (2023). The number of cages can be interpreted as a measure of investments in the Norwegian salmon farming industry, as the acquisition of permits can be expensive and signify plans for future salmon farming. If the number of cages increase, we expect an increase in short- to medium-term supply and therefore the correlation with the spot price is supposed to be negative. We have chosen lags of one, two and three months.

Price of Alternative Proteins

According to the Food and Agriculture Organization of the United Nations (2022), beef, lamb, poultry, and pork are the most consumed meats in the world. As alternative animal protein sources these are natural substitutes for salmon. The data of these four proteins was obtained from the Federal Reserve Bank of St. Louis[1, 2, 3, 4] (2023) and are expressed in US Cents per pound, which amount to about 0.45 kg. These products may impact demand of salmon as consumer preferences could shift in favor of these goods if the price of salmon increase. A reduction in demand of salmon will consequently have an adverse effect on the spot price, leading to a hypothesized positive correlation with the Fish Pool Index. Considering the time it takes for consumers to adapt to price changes in alternative proteins, we assume a short- to medium-term effect in demand for salmon. The effect on the Fish Pool Index is expected with lags of two and five months.

Consumer Price Index Year-over-Year Changes

Inflation results in higher production costs, which in turn means salmon farmers will require higher prices. Moreover, inflation affects consumers' purchasing power and demand negatively, again signaling change in the spot price. The most well-known indicator of inflation is the Consumer Price Index (CPI), which measures the change in the prices paid by consumers for a selection of goods over time (Bryan & Cecchetti, 1993). In this analysis we utilize data from Eurostat (2023) that measures percentage annual CPI change in the European Union, United States and Norway. This means that February 2020 will be measured against the previous year, February 2019. The European Union and the United States was chosen due to these being the largest markets (Kontali Analyse, 2022). We also include the Norwegian

CPI as this may influence production costs. Increased CPI means higher inflation and would presumably result in a higher spot price, and therefore the presumed correlation with the Fish Pool Index is positive. It is difficult to determine when consumers and producers will adjust their demand in response to a change in the CPI, but we assume a short- to medium term effect. Based on this we apply lags of two, four, and six months.

US Imports

Considering that the United States is one of the main markets for Norwegian farmed salmon, US imports should provide insight into short-term demand. The data was retrieved from the National Oceanic and Atmospheric Administration (2023), and it states the value of US imports of fresh and frozen farmed Atlantic salmon. It is expressed in thousand USD and measured as customs value, meaning the price actually paid for merchandise when sold for export to the US, excluding import duties, freight, insurance, and similar charges. Increased import value suggests increased short-term demand and will have an almost immediate positive effect on the Fish Pool Index. We therefore apply a one-month lag.

Oslo Seafood Index

The Oslo Seafood Index measures the development in the seafood industry through stock prices of large salmon farming companies listed on the Oslo Stock Exchange. Such stock prices may contain information about future supply and demand, as discovered by Dahl et al. (2021), unlike the Fish Pool Index that expresses current supply- and demand information. This data was acquired from Euronext$_3$ (2023), is expressed in NOK and consists of daily data that was aggregated by computing the monthly mean. Positive changes in stock prices means that investors expect increased revenues and earnings for salmon farmers. It is reasonable to assume that this will correspond with an increase in the salmon spot price. Therefore, we assume a positive correlation between the seafood index and the FPI in the short- to medium-term future. We utilize lags of one, three and five months.

Fish Pool Forward Prices

The forward prices reflect the price expectations for member companies of Fish Pool for the coming months, based on contracts, orders, as well as interests to buy or sell at Fish Pool. It is expressed in NOK. We include one- to twelve-months ahead futures prices because this corresponds to our forecast horizon. This means the forward price 12 in February 2020 indicates the forward price at that time twelve months ahead, for February 2021. The data was retrieved from Fish Pool$_3$ (2023) and originally occurred at daily frequency but was aggregated by computing the monthly mean. The forward price is expected to correlate positively with the Fish Pool Index, as an increase in the forward price indicates a similar development in the spot price.

Exchange Rates

Exchange rates are important because they somewhat explain foreign consumers' purchasing power relative to Norwegian cost of production (Bloznelis, 2018). Most of the raw materials required for salmon farming production in Norway are bought from Europe or the United States, in addition to the fact that these are the largest consuming markets. In other words, the vast majority of currency flows for Norwegian salmon producers are dominated by conversions of NOK to EUR and USD, both on the cost and revenue side (Mowi, 2020). We therefore include NOK to EUR and NOK to USD rates that was retrieved from Norges Bank [1, 2] (2023). Depreciation in NOK means NOK-EUR and NOK-USD exchange rates increase. This will increase the salmon export prices measured in NOK, but at the same time entail increased costs associated with imported raw materials, hence indicating an increase in the spot price. Appreciation in NOK however will have the opposite effect. This implies a positive correlation with the Fish Pool Index. The effect on supply is most likely to be long-term because harvest is planned a long time ahead and difficult to change. The impact on demand is expected to be in the medium term, as consumption could be adjusted faster. Based on these expectations, the lags on exchange rates are set to three and twelve months.

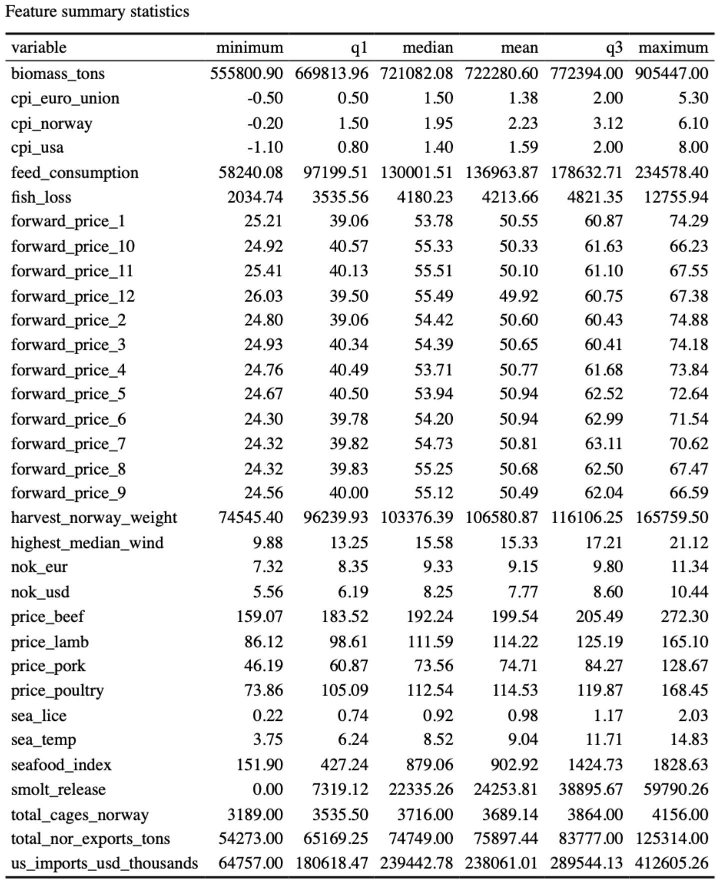
4.3 Descriptive Statistics

In this section we present descriptive statistics as a way to explore patterns in the data. This involves feature summary statistics, which provide an overview of various statistical measures for each feature variable, in addition to visualization of the variables. Finally, we calculate correlation between each feature variable and the Fish Pool Index.

4.3.1 Feature Summary Statistics

In *table 4* below, we show the minimum and maximum, first and third quantiles, median and mean values for each feature variable. This provides an overview of our data. In *appendix 1* we also present graphs of all features. From these we observe some important patterns in the data. First, we notice that the variables standing biomass, smolt release, sea temperature, feed consumption, harvest volume and cages in Norway appear to exhibit seasonality. Secondly, prices of alternative protein sources appear to be highly volatile, and lastly, we observe that the NOK has depreciated against both EUR and USD within the period.

Table 4: Feature summary statistics.



Feature summary statistics

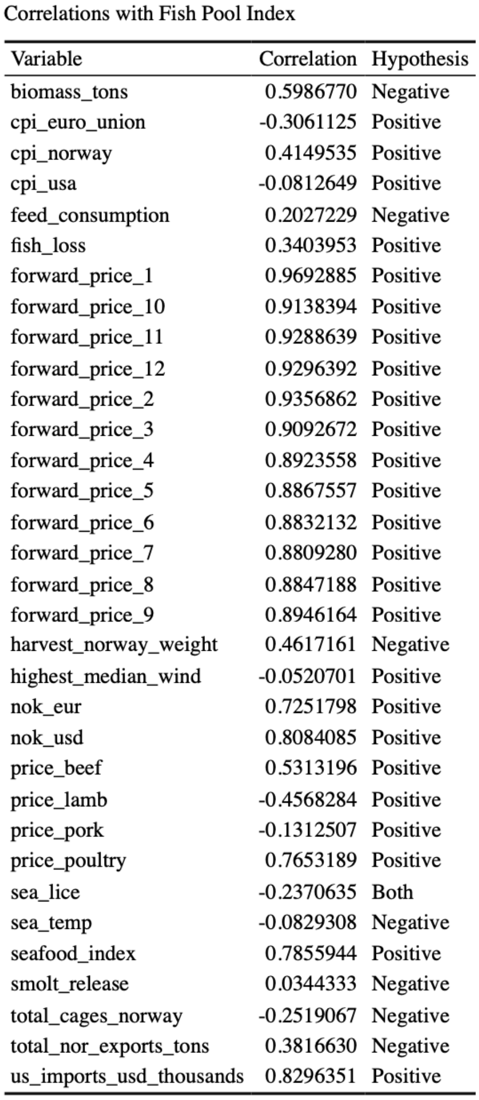
variable	minimum	q1	median	mean	q3	maximum
biomass_tons	555800.90	669813.96	721082.08	722280.60	772394.00	905447.00
cpi_euro_union	-0.50	0.50	1.50	1.38	2.00	5.30
cpi_norway	-0.20	1.50	1.95	2.23	3.12	6.10
cpi_usa	-1.10	0.80	1.40	1.59	2.00	8.00
feed_consumption	58240.08	97199.51	130001.51	136963.87	178632.71	234578.40
fish_loss	2034.74	3535.56	4180.23	4213.66	4821.35	12755.94
forward_price_1	25.21	39.06	53.78	50.55	60.87	74.29
forward_price_10	24.92	40.57	55.33	50.33	61.63	66.23
forward_price_11	25.41	40.13	55.51	50.10	61.10	67.55
forward_price_12	26.03	39.50	55.49	49.92	60.75	67.38
forward_price_2	24.80	39.06	54.42	50.60	60.43	74.88
forward_price_3	24.93	40.34	54.39	50.65	60.41	74.18
forward_price_4	24.76	40.49	53.71	50.77	61.68	73.84
forward_price_5	24.67	40.50	53.94	50.94	62.52	72.64
forward_price_6	24.30	39.78	54.20	50.94	62.99	71.54
forward_price_7	24.32	39.82	54.73	50.81	63.11	70.62
forward_price_8	24.32	39.83	55.25	50.68	62.50	67.47
forward_price_9	24.56	40.00	55.12	50.49	62.04	66.59
harvest_norway_weight	74545.40	96239.93	103376.39	106580.87	116106.25	165759.50
highest_median_wind	9.88	13.25	15.58	15.33	17.21	21.12
nok_eur	7.32	8.35	9.33	9.15	9.80	11.34
nok_usd	5.56	6.19	8.25	7.77	8.60	10.44
price_beef	159.07	183.52	192.24	199.54	205.49	272.30
price_lamb	86.12	98.61	111.59	114.22	125.19	165.10
price_pork	46.19	60.87	73.56	74.71	84.27	128.67
price_poultry	73.86	105.09	112.54	114.53	119.87	168.45
sea_lice	0.22	0.74	0.92	0.98	1.17	2.03
sea_temp	3.75	6.24	8.52	9.04	11.71	14.83
seafood_index	151.90	427.24	879.06	902.92	1424.73	1828.63
smolt_release	0.00	7319.12	22335.26	24253.81	38895.67	59790.26
total_cages_norway	3189.00	3535.50	3716.00	3689.14	3864.00	4156.00
total_nor_exports_tons	54273.00	65169.25	74749.00	75897.44	83777.00	125314.00
us_imports_usd_thousands	64757.00	180618.47	239442.78	238061.01	289544.13	412605.26

4.3.2 Correlation with Fish Pool Index

Correlation between the feature variables could provide information of patterns in our data and indicate the strength and direction of a particular variable's effect on the Fish Pool index. In *appendix 4* we have gathered scatter plots of all features plotted against the Fish Pool Index. From this we observe both linear and non-linear relationships, which again underlines the importance of flexible models.

In *table 5* below, we display the correlations between the Fish Pool Index and each feature, and how they compare to our hypotheses as first presented in *table 3*. The correlation measure is the Pearson correlation coefficient, which takes values between -1 and 1, indicating the degree of negative or positive correlations.

Table 5: Actual and hypothesized correlation between the Fish Pool Index and feature variables.



Correlations with Fish Pool Index

Variable	Correlation	Hypothesis
biomass_tons	0.5986770	Negative
cpi_euro_union	-0.3061125	Positive
cpi_norway	0.4149535	Positive
cpi_usa	-0.0812649	Positive
feed_consumption	0.2027229	Negative
fish_loss	0.3403953	Positive
forward_price_1	0.9692885	Positive
forward_price_10	0.9138394	Positive
forward_price_11	0.9288639	Positive
forward_price_12	0.9296392	Positive
forward_price_2	0.9356862	Positive
forward_price_3	0.9092672	Positive
forward_price_4	0.8923558	Positive
forward_price_5	0.8867557	Positive
forward_price_6	0.8832132	Positive
forward_price_7	0.8809280	Positive
forward_price_8	0.8847188	Positive
forward_price_9	0.8946164	Positive
harvest_norway_weight	0.4617161	Negative
highest_median_wind	-0.0520701	Positive
nok_eur	0.7251798	Positive
nok_usd	0.8084085	Positive
price_beef	0.5313196	Positive
price_lamb	-0.4568284	Positive
price_pork	-0.1312507	Positive
price_poultry	0.7653189	Positive
sea_lice	-0.2370635	Both
sea_temp	-0.0829308	Negative
seafood_index	0.7855944	Positive
smolt_release	0.0344333	Negative
total_cages_norway	-0.2519067	Negative
total_nor_exports_tons	0.3816630	Negative
us_imports_usd_thousands	0.8296351	Positive

It is worth noting the positive coefficients for supply-side variables such as Norwegian biomass, Norwegian harvest volume, and Norwegian export volume, which contradicts our original hypotheses. This seems to indicate that in our dataset, high Norwegian supply generally corresponds with high values of the Fish Pool Index. However, all twelve forward prices, as well as the exchange rates, the Oslo seafood index and US import volumes exhibit strong positive correlation in line with our expectations. Prices of alternative protein sources have varying correlations, but the price of poultry seems to be highly positively correlated.

4.4 Pre-processing of Data

Raw data is usually not ready for modelling, and therefore needs some pre-processing to make the data consistent and reliable. Pre-processing is the concept of transforming the raw data into a clean and understandable data set, which may involve removing missing values, noisy data and dealing with other inconsistencies before executing any analysis (Singh et al., 2021). In this section we will describe the pre-processing, starting with handling of missing values, before describing the several variable adjustments that was made. Finally, we explain the splitting of data into separate training and testing sets.

4.4.1 Missing Values

Missing values are values that were intended to be obtained during data collection, but due to various reasons are absent. The problem of such values is a common occurrence in all real-world data. If not treated correctly, this could reduce the statistical power of the analysis and lead to biased estimates, causing invalid conclusions (Kang, 2013). Moreover, many machine learning algorithms fail if the data contain missing values. This highlights the importance of dealing with them.

In this study, we aimed to prevent the problem of missing values by collecting data carefully, and we therefore encountered no missing values in the original data set. However, as described in section *3.8 Direct Forecasting* and section *4.1 Introduction to Dataset*, we constructed lead- and lag-variables which led to missing values as there was no data prior to January 2007 and after December 2021. The absence of these values appeared to have no pattern or be related to other variables in the data, which is referred to as Missing Completely at Random (Kang, 2013).

The advantage of such missing values is that the estimated parameters are not biased by the absence of the data. If the dataset is large enough, then listwise deletion is considered a reasonable strategy. This involves simply dropping the rows that do not have complete data for all variables and analyzing the remaining data (Kang, 2013). Utilizing this approach meant we dropped all cases containing missing values, which reduced the data from 180 to 145 observations. One drawback of this however, was that it significantly reduced our data size, which in turn could lead to a loss of statistical power. To increase the reliability of our analysis, we compare the more complex models to a seasonal naïve serving as a simple benchmark, as described in section *3.1 Simple Benchmark Model*.

4.4.2 Variable Adjustments

As mentioned in section *4.1 Introduction to Dataset*, new lagged variables were computed in order to capture the delayed effect of features that would impact the spot price several months later in time. Next, we removed all original feature variables, meaning we utilized only the lagged variables. The intention of this was to just use information that was available at the time of forecast, meaning we forecast the spot price at least one month ahead in time. Consequently, our data set now entails 61 explanatory variables as opposed to 33 in the original data set.

4.4.3 Train and Test Split

In machine learning, dividing the data into separate training and testing data sets is common practice. This approach utilizes the training data to fit the model, before applying the model to the previously unseen test set and evaluating the prediction accuracy (Hastie et al., 2013). In our case, the training set consists of data from January 2007 to December 2020, while the test set consists of observations from January 2021 to December 2021. We chose this split as we want a test set with equal number of observations as our forecast horizon. In addition, we aimed for a large training set since we have a relatively small sample and prefer more training data to develop the algorithm.

5. Analysis and Results

In this chapter, we will look at the results obtained from the models described in chapter *3 Methodology*. We will evaluate the quality of each model based on forecasting accuracy as explained in section *3.9.1 Evaluation of Statistical Forecast Accuracy,* and then assess the feature variable importance as described in section *3.5 Variable Importance*. Finally, we will compare the models to each other.

5.1 Seasonal Naïve

As explained in section *3.1 Simple Benchmark Model*, the seasonal naïve (SNAIVE) method serves as a simple benchmark against which we can compare the quality of our more complex models. From this we know that the SNAIVE forecast equals the last observed value from the corresponding season. Considering that this study consists of monthly data and employs a test data set from January 2021 to December 2021, the SNAIVE forecasts simply equal the observed Fish Pool Index values from January 2020 to December 2020. In *figure 6* below we compare the forecasts with the actual observations in a graph.

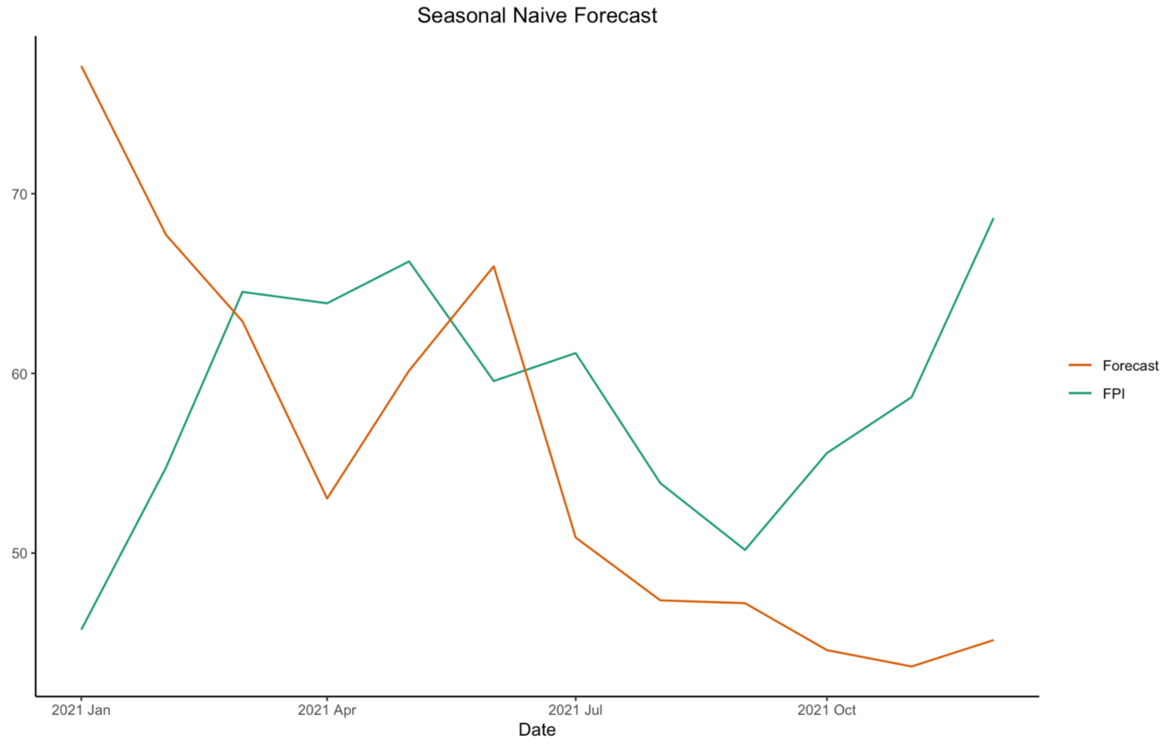


Figure 6: Seasonal naïve forecast compared to the Fish Pool Index.

From *figure 6* we observe that the seasonal naïve is quite poor in predicting the Fish Pool Index in 2021. We find that the mean absolute error (MAE) is 11.54, while the mean squared error (MSE) equals 200.35. The directional accuracy (DA) is 45%, indicating that the SNAIVE-method perform worse than a coin toss in predicting whether the FPI will go up or down in the next month. The results are summarized in *table 13* under section *5.5 Summary of Results*.

5.2 Decision Tree

In this section we will present the results of the decision trees without and with seasonal adjustment.

5.2.1 Decision Trees without Seasonal Adjustment

By implementing the procedure outlined in section *3.2 Decision Tree* in R, we obtain forecasts as expressed in *figure 7* below.

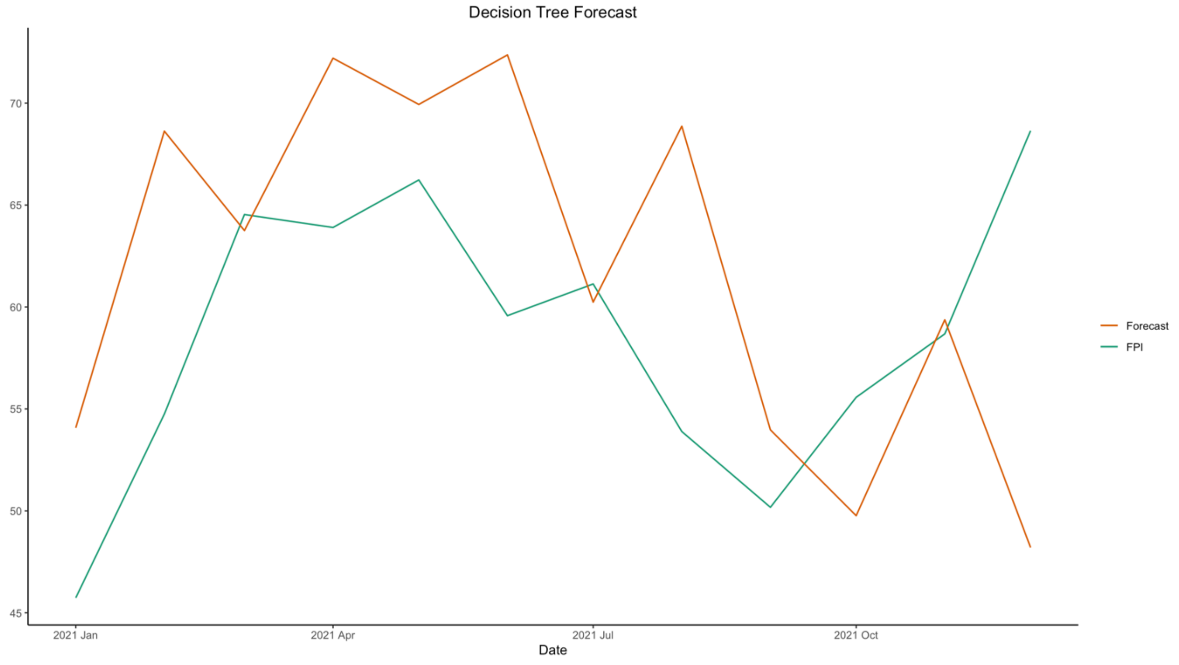
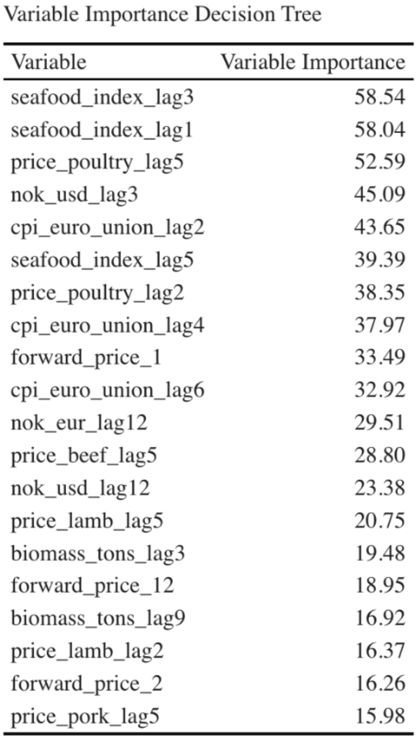


Figure 7: Decision tree forecast compared to the Fish Pool Index.

It appears that the forecasts for 2021 are quite inaccurate. The MAE is 7.87, which is a substantial improvement compared to the SNAIVE-forecast of 11.54, but still high compared to the more complex tree-models (see *table 13*). The MSE is 100.04, while the DA is 27%. While the forecasts are significantly closer to the actual observations according to MAE and

MSE, decision trees are worse than the seasonal naïve forecast in predicting the directional change of the Fish Pool Index over the next month.

Table 6: Variable importance in the decision tree models.



Variable Importance Decision Tree

Variable	Variable Importance
seafood_index_lag3	58.54
seafood_index_lag1	58.04
price_poultry_lag5	52.59
nok_usd_lag3	45.09
cpi_euro_union_lag2	43.65
seafood_index_lag5	39.39
price_poultry_lag2	38.35
cpi_euro_union_lag4	37.97
forward_price_1	33.49
cpi_euro_union_lag6	32.92
nok_eur_lag12	29.51
price_beef_lag5	28.80
nok_usd_lag12	23.38
price_lamb_lag5	20.75
biomass_tons_lag3	19.48
forward_price_12	18.95
biomass_tons_lag9	16.92
price_lamb_lag2	16.37
forward_price_2	16.26
price_pork_lag5	15.98

From *table 6* we observe that the most important explanatory variables in constructing the decision trees seems to have been the seafood index with three- and one-month lags, the price of poultry with five lags, and the NOK-USD exchange rate with three lags. These values are the average of the variable importance scores across all 12 individual decision trees. See *appendix 2* for plots of the 12 trees, where each tree provides an easy to interpret illustration of which variables are most important in predicting the Fish Pool Index.

5.2.2 Decision Trees with Seasonal Adjustment

Now, we seasonally adjust the Fish Pool Index by doing an STL decomposition and subtracting the seasonal component. Then we employ this seasonally adjusted FPI as the response variable in the decision trees and add back the seasonal component to obtain the final forecasts. This procedure is explained in detail in section *3.6 Time Series Decomposition*. In R, we use the STL-function from the feasts-package to do the seasonal decomposition. The rest of the procedure is similar as described in chapter *3.2 Decision Tree*. Using this method, we obtain forecasts as seen in *figure 8* below.

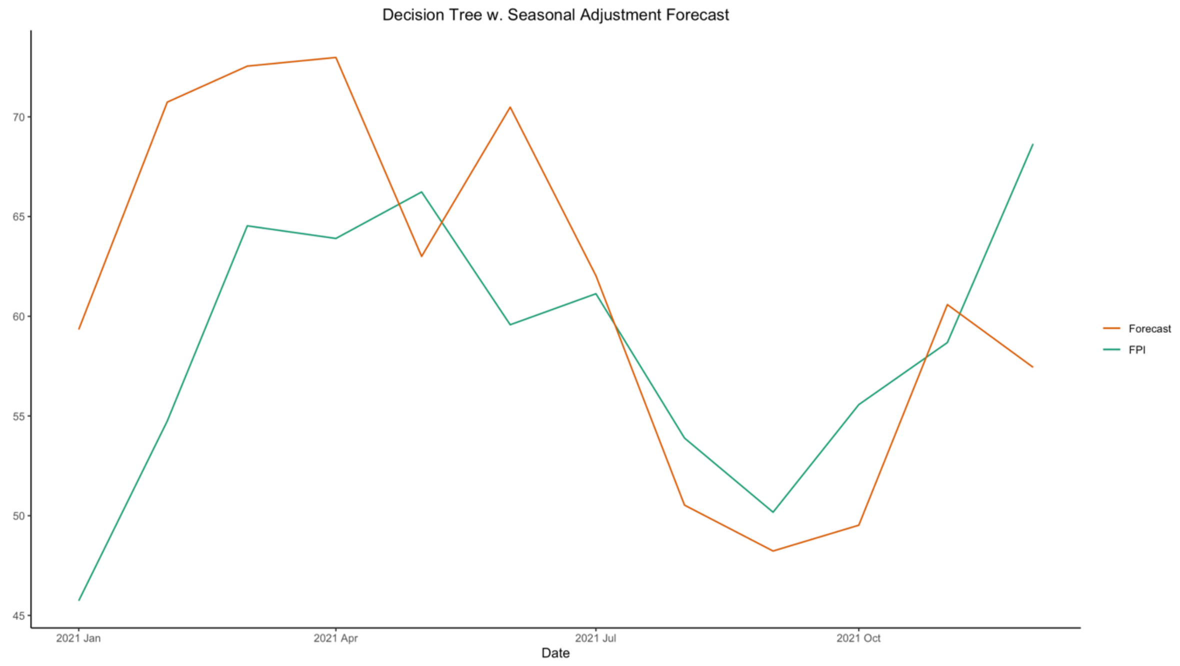
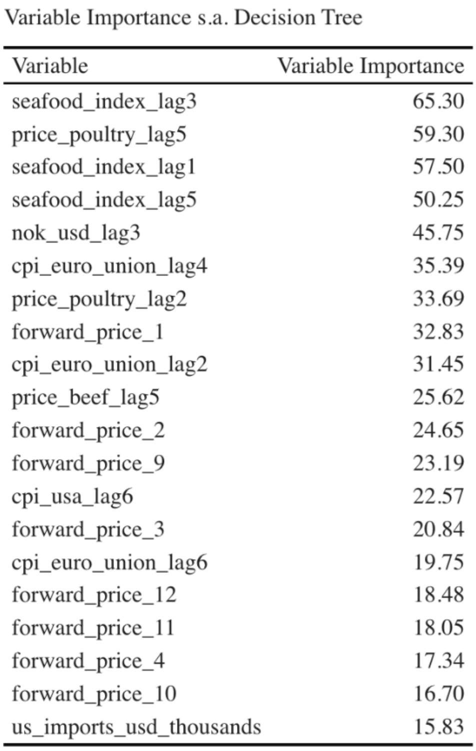


Figure 8: Decision tree with seasonal adjustment forecast compared to the Fish Pool Index.

By first glance it does appear that forecasting accuracy has improved compared to the non-seasonally adjusted decision trees. This is confirmed by the accuracy measures. MAE is down to 7.18, while MSE is 74.85. DA has also significantly improved and is now at 55% accuracy.

Table 7: Variable importance in the seasonally adjusted decision trees.



Variable Importance s.a. Decision Tree

Variable	Variable Importance
seafood_index_lag3	65.30
price_poultry_lag5	59.30
seafood_index_lag1	57.50
seafood_index_lag5	50.25
nok_usd_lag3	45.75
cpi_euro_union_lag4	35.39
price_poultry_lag2	33.69
forward_price_1	32.83
cpi_euro_union_lag2	31.45
price_beef_lag5	25.62
forward_price_2	24.65
forward_price_9	23.19
cpi_usa_lag6	22.57
forward_price_3	20.84
cpi_euro_union_lag6	19.75
forward_price_12	18.48
forward_price_11	18.05
forward_price_4	17.34
forward_price_10	16.70
us_imports_usd_thousands	15.83

In *table 7* we see that the most important feature variables in these models again are the seafood index with three-, one- and five-month lags, and the price of poultry with five lags. See *appendix 3* for visualizations of the 12 individual seasonally adjusted decision trees.

5.3 Random Forest

In this section we present the results from the random forest models.

5.3.1 Random Forest without Seasonal Adjustment

We build random forest-models in R as described in chapter *3.3 Random Forest,* and obtain forecasts as presented in *figure 9* below.

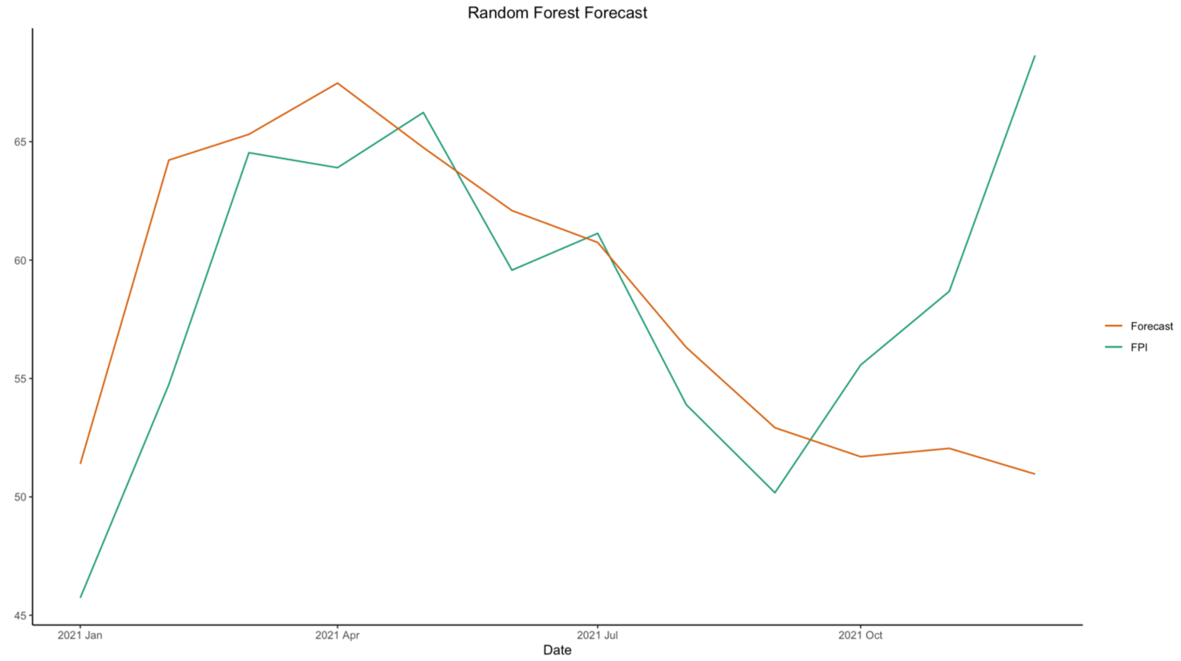
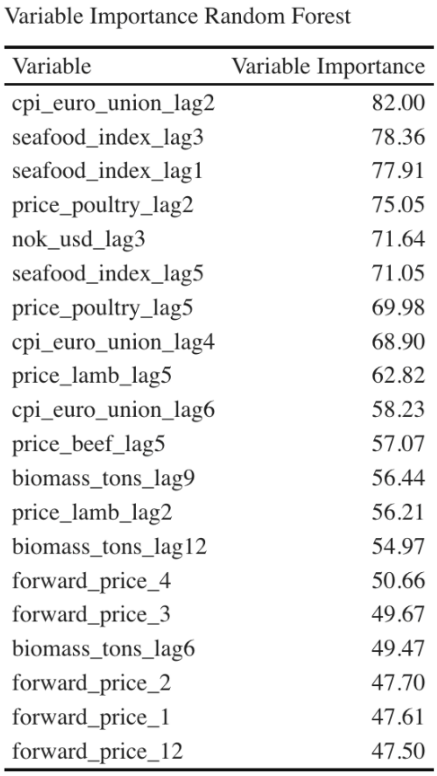


Figure 9: Random forest forecast compared to the Fish Pool Index.

This time it appears the forecasts are generally quite accurate with the notable exceptions of the November and December 2021 forecasts. MAE is 4.77, MSE is 44.07, and DA is 55%. This indicates that while the Euclidean distance between forecast and observation is quite short, the model still performs roughly equal to a random coin toss in predicting the directional movement of the Fish Pool Index one month ahead. Perhaps this means the model is not quite able to capture the seasonal price patterns.

Table 8: Variable importance in the random forest models.



Variable Importance Random Forest

Variable	Variable Importance
cpi_euro_union_lag2	82.00
seafood_index_lag3	78.36
seafood_index_lag1	77.91
price_poultry_lag2	75.05
nok_usd_lag3	71.64
seafood_index_lag5	71.05
price_poultry_lag5	69.98
cpi_euro_union_lag4	68.90
price_lamb_lag5	62.82
cpi_euro_union_lag6	58.23
price_beef_lag5	57.07
biomass_tons_lag9	56.44
price_lamb_lag2	56.21
biomass_tons_lag12	54.97
forward_price_4	50.66
forward_price_3	49.67
biomass_tons_lag6	49.47
forward_price_2	47.70
forward_price_1	47.61
forward_price_12	47.50

From *table 8*, we see that the most important feature variables are EU inflation with a two-month lag, the seafood index with three- and one-month lags, and the price of poultry with a two-month lag.

5.3.2 Random Forest with Seasonal Adjustment

We repeat the STL-decomposition and execute all the random forest models again, in the same way as described in chapter *3.3 Random Forest*, except with the seasonally adjusted Fish Pool Index as the response variable. Then we add back the seasonal component, as explained in chapter *3.6 Time Series Decomposition* to obtain the final forecasts. The forecasts are illustrated graphically in *figure 10* below.

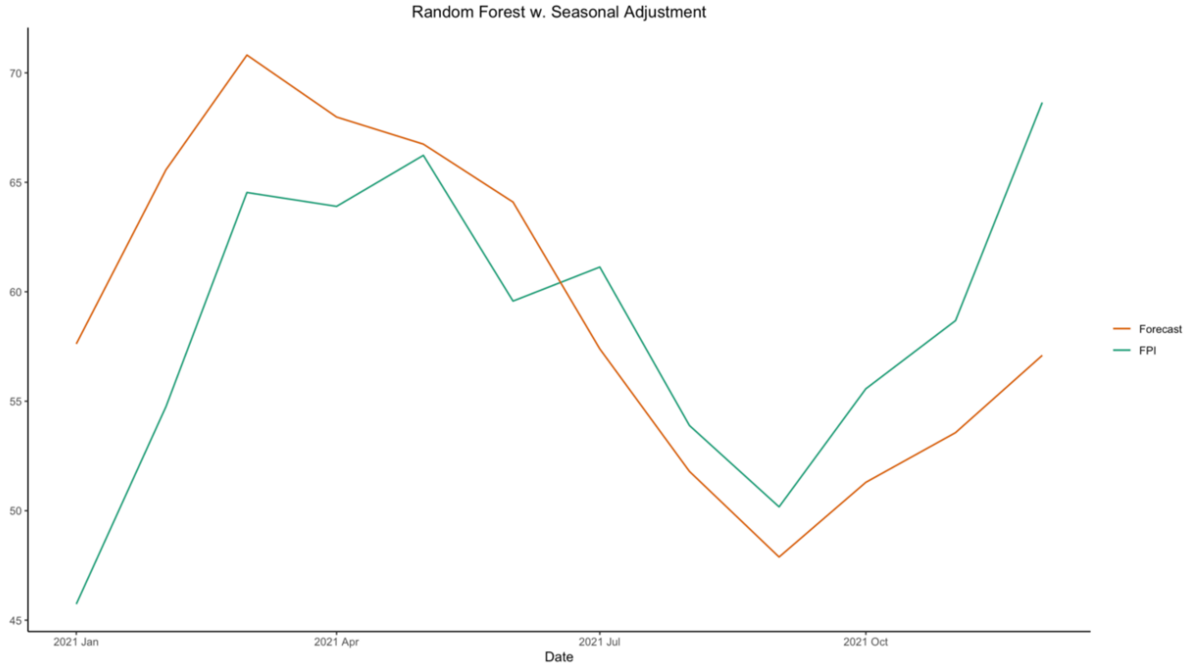
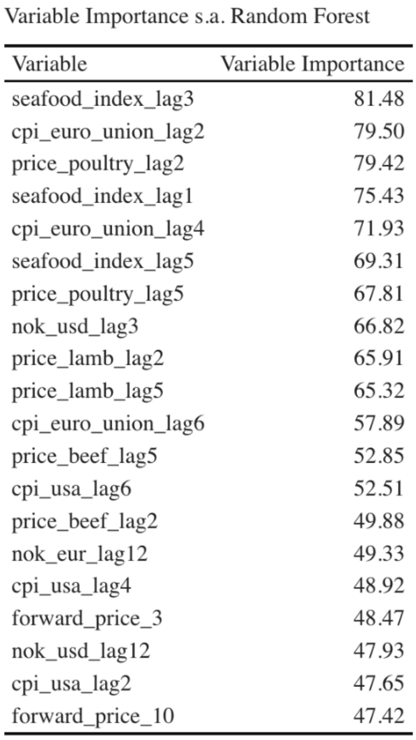


Figure 10: Random forest with seasonal adjustment forecast compared to the Fish Pool Index.

By visual inspection, it seems like the seasonally adjusted random forest models are better able to capture the seasonal pattern of price movements. MAE and MSE are slightly higher than in the non-seasonally adjusted model at 5.60 and 44.72 respectively, but DA has massively improved to 82%. While the average distance between forecast and observation is shorter in the non-seasonally adjusted random forest, the directional accuracy is much higher in the seasonally adjusted model, suggesting that the STL-decomposition is valuable in capturing the seasonal price patterns in the Fish Pool Index.

Table 9: Variable importance in the seasonally adjusted random forest models.



Variable Importance s.a. Random Forest

Variable	Variable Importance
seafood_index_lag3	81.48
cpi_euro_union_lag2	79.50
price_poultry_lag2	79.42
seafood_index_lag1	75.43
cpi_euro_union_lag4	71.93
seafood_index_lag5	69.31
price_poultry_lag5	67.81
nok_usd_lag3	66.82
price_lamb_lag2	65.91
price_lamb_lag5	65.32
cpi_euro_union_lag6	57.89
price_beef_lag5	52.85
cpi_usa_lag6	52.51
price_beef_lag2	49.88
nok_eur_lag12	49.33
cpi_usa_lag4	48.92
forward_price_3	48.47
nok_usd_lag12	47.93
cpi_usa_lag2	47.65
forward_price_10	47.42

From *table 9* we again see that the most important variables are the seafood index with three- and one-month lags, EU inflation with a two-month lag, and the price of poultry with a two-month lag.

5.4 xgBoost

In this section we will present the results of xgBoost without and with seasonal adjustment.

5.4.1 xgBoost without Seasonal Adjustment

By implementing the xgBoost-algorithm with hyperparameter tuning in R as described in section *3.4 xgBoost,* we obtain forecasts as illustrated in *figure 11* below.

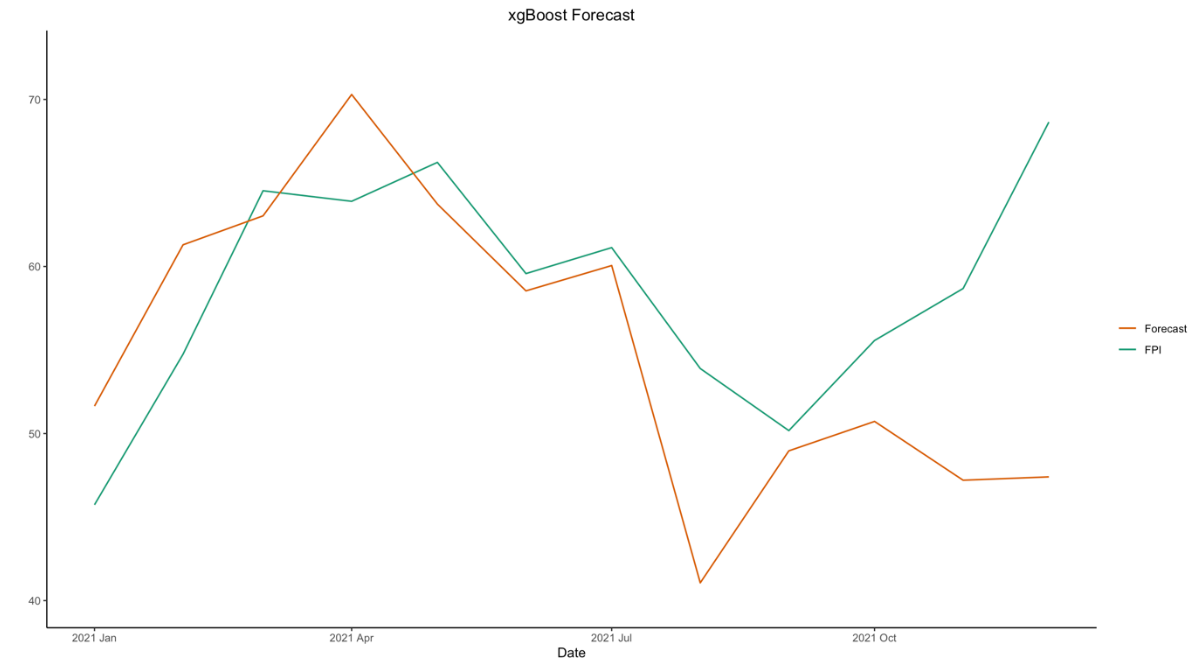
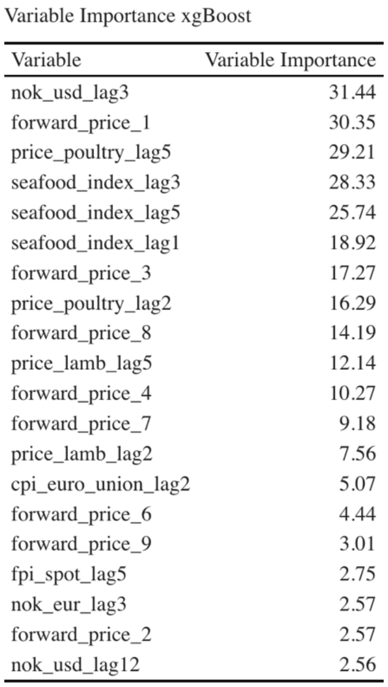


Figure 11: xgBoost forecast compared to the Fish Pool Index.

Here it seems as though the forecasts are generally quite close to the observations in the first half of 2021, and then significantly more inaccurate in the second part. The MAE is 6.38 and the MSE is 75.14, which is relatively high compared to random forests, but lower than we obtained from decision trees. The directional accuracy is 64%, which is better than a coin toss, but still worse than the seasonally adjusted random forest's DA.

Table 10: Variable importance in the xgBoost models.


Variable Importance xgBoost

Variable	Variable Importance
nok_usd_lag3	31.44
forward_price_1	30.35
price_poultry_lag5	29.21
seafood_index_lag3	28.33
seafood_index_lag5	25.74
seafood_index_lag1	18.92
forward_price_3	17.27
price_poultry_lag2	16.29
forward_price_8	14.19
price_lamb_lag5	12.14
forward_price_4	10.27
forward_price_7	9.18
price_lamb_lag2	7.56
cpi_euro_union_lag2	5.07
forward_price_6	4.44
forward_price_9	3.01
fpi_spot_lag5	2.75
nok_eur_lag3	2.57
forward_price_2	2.57
nok_usd_lag12	2.56

In *table 10*, we see that the most important explanatory variables in constructing the xgBoost models were the NOK-USD exchange rate with a three-month lag, the one-month ahead forward price, as well as the price of poultry with a five-month lag, and the seafood index with three- five- and one-month lags.

5.4.2 xgBoost with Seasonal Adjustment

Again, we performed seasonal adjustment by the method outlined in section *3.6 Time Series Decomposition,* and then repeated the xgBoost-algorithm as described in section *3.4 xgBoost* with the seasonally adjusted Fish Pool Index as the response variable. *Figure 12* below portray the forecasts.

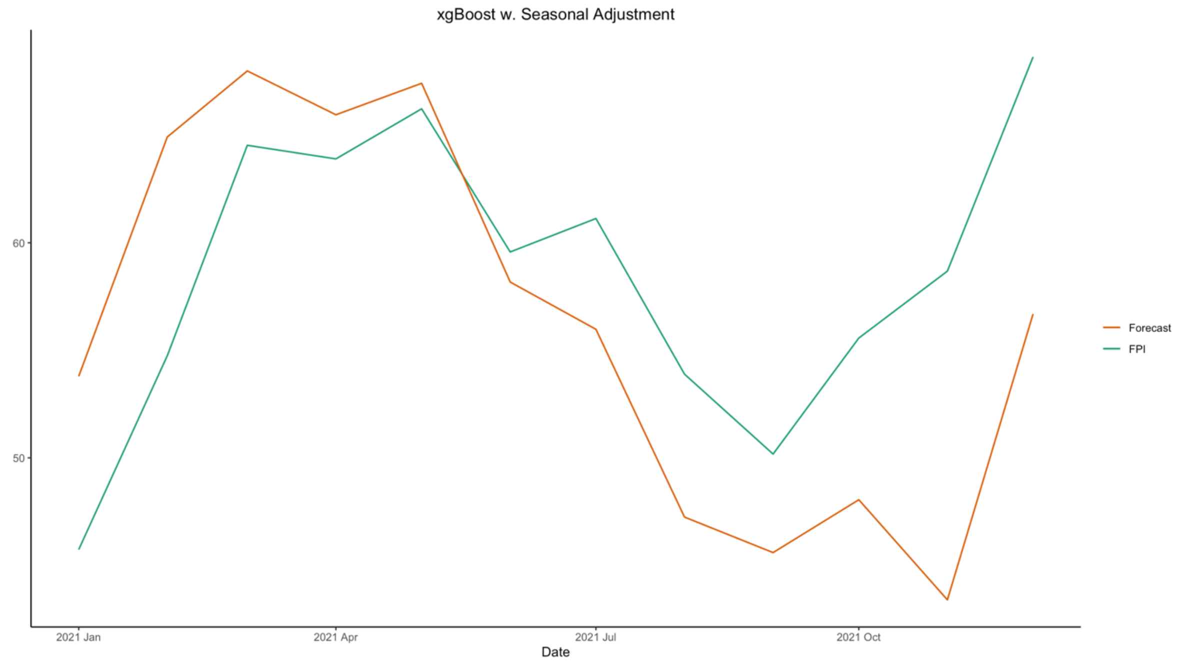
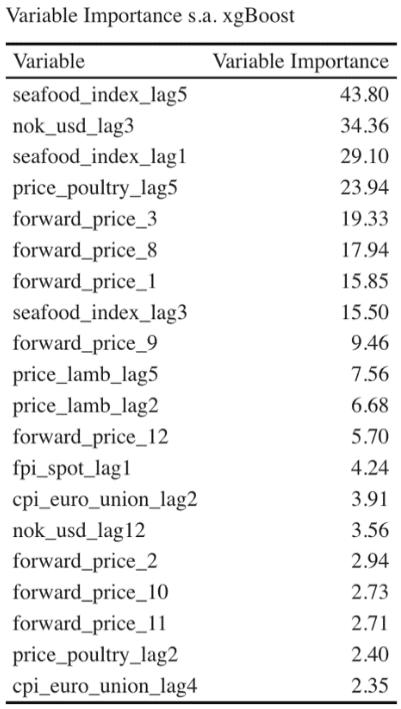


Figure 12: xgBoost with seasonal adjustment forecast compared to the Fish Pool Index.

The MAE is now slightly higher at 6.46. The MSE has decreased however, and is now 59.40. DA has improved significantly, and the seasonally adjusted xgBoost accurately predicts 82% of the monthly price directional movements in 2021, equaling seasonally adjusted random forest as the highest performing model in terms of DA.

Table 11: Variable importance in the seasonally adjusted xgBoost models.



Variable Importance s.a. xgBoost

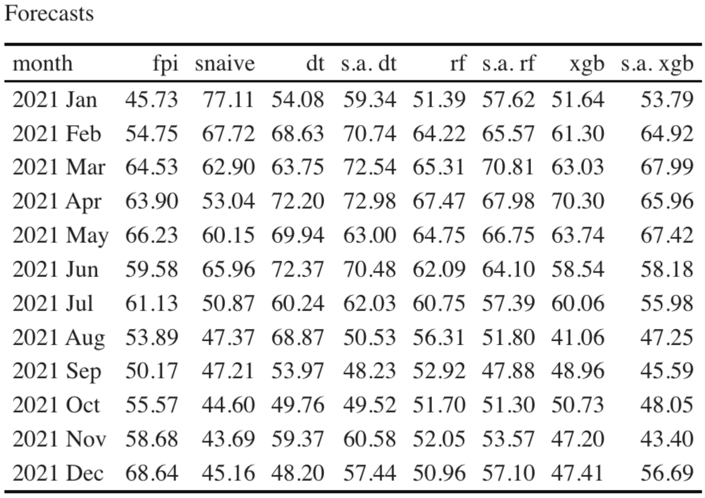
Variable	Variable Importance
seafood_index_lag5	43.80
nok_usd_lag3	34.36
seafood_index_lag1	29.10
price_poultry_lag5	23.94
forward_price_3	19.33
forward_price_8	17.94
forward_price_1	15.85
seafood_index_lag3	15.50
forward_price_9	9.46
price_lamb_lag5	7.56
price_lamb_lag2	6.68
forward_price_12	5.70
fpi_spot_lag1	4.24
cpi_euro_union_lag2	3.91
nok_usd_lag12	3.56
forward_price_2	2.94
forward_price_10	2.73
forward_price_11	2.71
price_poultry_lag2	2.40
cpi_euro_union_lag4	2.35

In *table 11* we see that the seafood index with five- and one-month lags, the NOK-USD exchange rate with a three-month lag, and the price of poultry with a five-month lag are the most important explanatory variables in the seasonally adjusted xgBoost model.

5.5 Summary of Results

To compare the performance of the models we will now present summaries of the obtained results. *Table 12* below contains the forecasts from all models.

Table 12: Summary of model forecasts.

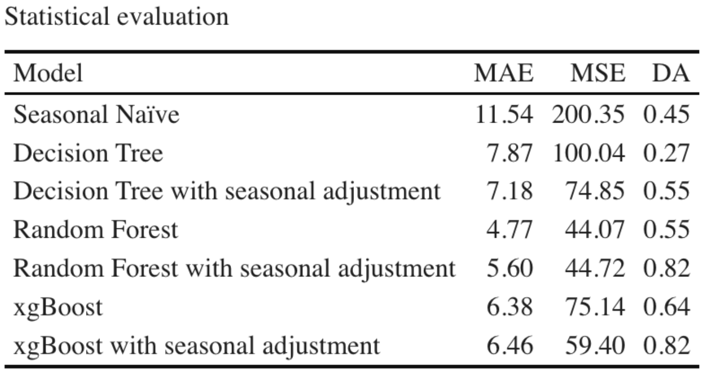


Forecasts

month	fpi	snaive	dt	s.a. dt	rf	s.a. rf	xgb	s.a. xgb
2021 Jan	45.73	77.11	54.08	59.34	51.39	57.62	51.64	53.79
2021 Feb	54.75	67.72	68.63	70.74	64.22	65.57	61.30	64.92
2021 Mar	64.53	62.90	63.75	72.54	65.31	70.81	63.03	67.99
2021 Apr	63.90	53.04	72.20	72.98	67.47	67.98	70.30	65.96
2021 May	66.23	60.15	69.94	63.00	64.75	66.75	63.74	67.42
2021 Jun	59.58	65.96	72.37	70.48	62.09	64.10	58.54	58.18
2021 Jul	61.13	50.87	60.24	62.03	60.75	57.39	60.06	55.98
2021 Aug	53.89	47.37	68.87	50.53	56.31	51.80	41.06	47.25
2021 Sep	50.17	47.21	53.97	48.23	52.92	47.88	48.96	45.59
2021 Oct	55.57	44.60	49.76	49.52	51.70	51.30	50.73	48.05
2021 Nov	58.68	43.69	59.37	60.58	52.05	53.57	47.20	43.40
2021 Dec	68.64	45.16	48.20	57.44	50.96	57.10	47.41	56.69

In *table 13* below the statistical accuracy measures from all models are summarized.

Table 13: Statistical evaluation of forecasts.



Statistical evaluation

Model	MAE	MSE	DA
Seasonal Naïve	11.54	200.35	0.45
Decision Tree	7.87	100.04	0.27
Decision Tree with seasonal adjustment	7.18	74.85	0.55
Random Forest	4.77	44.07	0.55
Random Forest with seasonal adjustment	5.60	44.72	0.82
xgBoost	6.38	75.14	0.64
xgBoost with seasonal adjustment	6.46	59.40	0.82

The model with the lowest MAE is non-seasonally adjusted random forest, with a MAE of 4.77. Then follows seasonally adjusted random forest, suggesting these models are closest to the observations of the Fish Pool Index in 2021, measured by absolute distance. Measured by MSE, we see that again, non-seasonally adjusted random forest is the best performer with 44.07, closely followed by seasonally adjusted random forest at 44.72. The third best performer according to MSE is seasonally adjusted xgBoost. In terms of directional accuracy, it is seasonally adjusted random forest and seasonally adjusted xgBoost that performs best, both with 82% accuracy. Importantly, non-seasonally adjusted random forest, the best performer both in terms of MAE and MSE, is among the worst performers in DA.

6. Discussion

In this thesis we have developed several tree-based predictive models to forecast the Fish Pool Index, a measure of the Atlantic salmon spot price. The objective was to produce accurate and reliable forecasts over a 12-month horizon, which could then be used by salmon farming companies in their decision-making to obtain economic gain. First, we developed a seasonal naïve forecast as a benchmark against which to measure the performance of the tree-based models. Then, we built decision trees, random forests, and xgBoost models without any seasonal adjustments before using an STL-decomposition to attempt to capture the seasonality of the salmon price. Finally, we evaluated the statistical accuracy of all forecasts. In the following chapter we will present and discuss the general findings in this thesis, before evaluating the potential economic value of applying the best model to salmon harvesting decisions. Finally, we will address possible limitations in our work, and suggest avenues for improvements and further research.

6.1 General Findings

All forecasts are inevitably inaccurate, and this remains true of the forecasts developed in this thesis. The relevant question, however, is what we know about the frequency and magnitude of the inaccuracies. The MAEs of the tree-based models ranged from 4.77 to 7.87, while the Fish Pool Index in 2021 ranged from 45.7 to 68.6. The average MAE was 6.38, while the average Fish Pool Index observation was 58.57, suggesting that our forecasts were inaccurate by about 11% on average. The best directional accuracy was 82%, suggesting that the models correctly predicted up- or down-movements around 8 out of 10 times. We are reasonably satisfied with the forecast accuracy of the best models, which was both random forest models with MAEs of 4.77 and 5.60 and DAs between 55% and 82%. If we compare this to the seasonal naïve benchmark's MAE of 11.54 and DA of 45%, it seems quite clear that the more complex models added some value in forecast ability.

It is also interesting to compare differences in performance within the tree-based models. The simplest models were non-seasonally adjusted and seasonally adjusted decision trees, and they achieved MAEs of 7.87 and 7.18, with DAs of 27% and 55% respectively. Comparing these results to the more complex random forest models, with MAEs of 4.77 and 5.60 with DAs of 55% and 82%, one gets the impression that there is a decent increase in forecast accuracy with increased complexity in the models. There could be many reasons behind this. We know from section *3.3 Random Forest*, that random forest is a tree ensemble model, averaging predictions made by many “bagged” decision trees that only utilize a subset of available predictors at each internal split. Given that we have somewhat limited data with only 132 observations in the training set, it is conceivable that bootstrapped sampling is a valuable tool in extracting as much information as possible from few observations. The predictor subsampling at each split could also be part of the reason why random forest generally performed better as it reduces the dominance of a few strong predictors. xgBoost also performed somewhat better than the simple decision trees. The process of boosting, as explained in chapter *3.4 Extreme Gradient Boosting*, seems to have extracted some relevant information, as the MAEs are about 15% lower. Again, it seems like averaging predictions from many tree models produced higher forecast accuracy.

Another finding worth mentioning is the apparent value of seasonality removal. Decision trees, random forest, and xgBoost all performed better when the seasonal component was removed from the Fish Pool Index. This effect was quite clear in the case of decision trees as both MAE and MSE was lower in the seasonally adjusted case, while DA was significantly higher. In random forest, the non-seasonally adjusted model produced marginally lower MAE and MSE, but the huge increase in directional accuracy in the seasonally adjusted model, of 82% vs. 55%, means we would still argue that seasonality removal caused an improvement. In xgBoost where the non-seasonally adjusted model provided marginally lower MAE, the seasonally adjusted model produced the lowest MSE, and higher DA of 82% vs. 64%. This seems to indicate that there is a clear presence of a seasonality in the Fish Pool Index, and that the STL-decomposition was better able to capture the seasonal swings than the unassisted tree models.

From the variable importance scores, we observe that the Oslo seafood index appeared among the top four most important predictors in all the decision tree-, random forest-, and xgBoost-models. This finding appears to be in line with the main finding in Dahl et al. (2021) who concluded that the stock prices of large salmon companies may contain predictive power on the Fish Pool Index. This may suggest that the stock prices of salmon companies incorporate forward-looking supply- and demand-information, while the Fish Pool Index mostly reflects current supply and demand. In addition, demand variables such as the price of poultry and EU inflation seem to be important predictors, which aligns with Asche et al. (2019) and Bloznelis (2016) who suggested that short-term supply inelasticity is one of the main causes of salmon price volatility. This may indicate that supply is not able to quickly adjust to changes in demand. We also observe that the top five most important variables in all models have lags shorter than six months, possibly indicating that more recent data is better suited for use in forecasting the Fish Pool Index 12 months ahead, than data lagged more than half a year.

6.2 Evaluation of Economic Value

To evaluate the economic value of the best forecasts we will look at the harvest volume of the third largest Norwegian salmon farmer, SalMar, and calculate what potential economic gain they could obtain by implementing a timely harvesting strategy, as explained in section *3.9.2 Evaluation of Economic Value of Forecast.* Given SalMar's annual harvest volume of 182 100 tons in 2021 (SalMar, 2022), their average monthly harvest volume was 15 175 tons. We assume that every month SalMar has 15 175 tons available for harvest and could choose to delay harvesting by one month at no extra cost. Naturally, it pays to harvest when the price is higher.

The average absolute monthly price difference, $\bar{X}_{price\ difference}$, of the Fish Pool Index in 2021 was 4.998 NOK per kg, or 4998 NOK per ton. Given a perfect forecast, SalMar could capture the whole price difference, which yields $G_{perfect\ forecast}$ of 4998 NOK × 0.5 × 15 175 tons = 37.9 MNOK per month, corresponding to 455 MNOK per year. Considering the best performing model of this study, the month with the higher price would be correctly predicted in close to 82% of the cases. Utilizing this model, we can now calculate potential economic gain per month of accurate forecasting associated with this model, $G_{best\ model}$ = 37.9 MNOK × 81.82% – 37.9 MNOK × (1 – 81.82%) = 24.1 MNOK per month or 289 MNOK per year.

SalMar had a net profit margin of 17.7% in 2021 (SalMar, 2022). If we assume the same margin on additional revenue due to timely harvesting, this would mean 289 MNOK × 17.7% = 51.2 MNOK in additional earnings after tax. Given SalMar's annual earnings of 2668 MNOK, applying our model could lead to about 2% increase in SalMar's annual earnings. If the salmon farmer was to employ the benchmark and make his harvesting decisions at random, then one would expect him to be correct 50% of the time, which would mean no value added after subtracting the loss of incorrect forecasts.

This analysis relies on a few heroic assumptions, but still illustrates that there is potentially tremendous economic value in accurate salmon price forecasts. In the real world, one would have to account for the extra cost of keeping salmon in cages for an additional month, as well as potential negative externalities. For example, holding back a large proportion of the world's salmon harvest in anticipation of higher prices could impact the price itself. It is also reasonable to assume that SalMar's competitors would adjust their behavior accordingly, possibly reducing the additional earnings from accurate forecasts. Still, this analysis does indicate that accurate forecasts could translate to significant economic gain.

6.3 Limitations

Perhaps the biggest limitation in this thesis was limited data availability, which caused the low number of observations in the dataset. By carefully examining open data sources we were able to retrieve a total of 180 monthly observations spanning from January 2007 to December 2021 of 33 variables. It proved impossible to find data going further back than this. When we built models, we divided into train and test datasets as well as introduced lags to all of the explanatory variables. This caused the number of observations in the training set, on which the models were fit, to be reduced to 132. Machine learning models such as the ones used in this thesis generally thrive on many more observations (usually in the thousands) in order to find patterns in the training data that will generalize well to out-of-sample test data. There is a risk that our training dataset does not have enough observations for the models to find these general patterns in the data.

Furthermore, with a small training dataset there is also increased risk of overfitting. This is because the model fit may be unduly influenced by random patterns in the training data that do not generalize well to out-of-sample predictions. With more training data, such random patterns would be averaged out, and the model would have less risk of overfitting. We attempted to encounter the problem of overfitting by evaluating the forecasts on out-of-sample test data. We used no information from 2021 to produce forecasts throughout the whole year. Our hypothesis was that if our models were terribly overfitted to the training data, these out-of-sample forecasts would be quite poor. In our case, the models seem able to generalize reasonably well to unseen new data, but it is still quite likely that out-of-sample forecast accuracy would improve with more training data.

Another aspect of limited data availability is inaccessible variables. As mentioned, we collected the vast majority of our data set through publicly available sources. We were able to find 33 relevant explanatory variables, but it is quite likely that we still missed variables that could be valuable in forecasting the Fish Pool Index. An example of a potentially relevant variable may be the age distribution of salmon in cages. If for instance, there were a high proportion of mature fish in the cages, one may expect short-term harvest volumes to increase, potentially causing the Fish Pool Index to decrease. Other examples of potentially relevant variables not included in this analysis may be consumption in the European market, and population data in relevant salmon markets. It is of course highly likely that our models would perform better if we were able to include all relevant variables in our data.

In this thesis we have not conducted any variable transformations except for seasonality removal of the Fish Pool Index and producing lag- and lead-variables. Traditional time series forecasting methods have stationary data as a standard assumption, and therefore stationarity tests, and potential logarithmic transformations and differencing are standard procedures in the case of non-stationary data. The tree-based methods employed in this thesis do not come with strict stationarity requirements, and we have therefore chosen not to perform tests or transformations except seasonality removal. It is however still possible that logarithmic transformations and/or differencing would lead to more stationary data and therefore a more hospitable forecasting environment.

Another limitation was computational resources. Random forest and xgBoost are quite computationally intensive, especially when one needs to optimize several hyperparameters simultaneously. This caused us to reduce the available values of the hyperparameters to ensure that the search grid in R did not become too large, as this would lead to incredibly slow computation. Naturally enough, this was particularly limiting in the xgBoost-case where the grid of hyperparameters was largest. We tried to mitigate this problem by ensuring that the hyperparameter values we did try were all likely to be relevant for the algorithm.

6.4 Possible Improvements and Further Research

One area of potential improvement in future research is to include more data. We have already explained why the dataset used in this thesis contains few observations, but as the salmon industry becomes more sophisticated, we expect the amount of data that is collected to increase. It is therefore likely that one could repeat the analysis done in this thesis with more data at some point in the future. This could potentially lead to higher forecast accuracy. Another way to increase the number of observations would be to predict the Fish Pool Index on a weekly rather than a monthly frequency. This would however also decrease the number of features as most of the features employed in this thesis is only available at a monthly frequency. In fact, only the Oslo seafood index, Fish Pool forward prices, sea lice, sea temperature, and Norwegian export volumes were available at a weekly frequency. While this would increase the number of observations it would come with the cost of excluding potentially important predictors.

This thesis has focused on using tree-based prediction models to do time series forecasting. This was because we expected tree-based models would be flexible enough to capture the non-linearities in the highly volatile Fish Pool Index, and that the literature on tree-based models showed promising results in predicting the price of commodities such as oil and gold. However, by focusing exclusively on tree-based models, we may have overlooked other models that could potentially perform better. For example, it is possible that one could obtain higher forecast accuracy by employing methods specifically designed for time series forecasting such as ARIMA models. Other predictive models that could yield good results in salmon price forecasting may include generalized additive models (GAM), support vector machines (SVM), or neural networks, as these are considered flexible and able to capture non-linear relationships in data.

We also chose to employ the direct forecasting strategy rather than the recursive method. The reasons behind this were outlined in section *3.8 Direct Forecasting.* In essence, we wanted to reduce the dependency on previously forecasted values in making new forecasts, as this could lead to error propagation. Also, we wanted to avoid creating forecasts for all features, as would be needed in a multivariate recursive forecasting model. This means however that we had to forecast 12 steps ahead directly, for example using information from December 2020 to forecast December 2021. It is possible that a recursive model could perform better as it only has to forecast one-step ahead, and then use the previous forecast for the next step. This would require a very accurate and reliable one-step ahead forecasting model, but still would be an interesting avenue for further research.

We were also constrained by computational resources in determining the optimal values of hyperparameters in the tree-based models. This was particularly limiting when developing the xgBoost models. With more computational resources one could expand the grid search and increase the probability of finding the best hyperparameter values. This would naturally lead to higher forecast accuracy and is thus an area of potential future improvement.

7. Conclusion

Industrial salmon farming is becoming an increasingly important industry, both globally and in Norway. One of the main risk factors in salmon production is the highly volatile spot price, so access to high-quality price forecasts could prove immensely valuable throughout the value chain. From academic literature we knew that tree-based models have shown promise in commodity price prediction tasks in recent years, so in this thesis we therefore tried to answer the following research question: *Can tree-based prediction models produce accurate and reliable monthly forecasts of the Fish Pool Index 12 months ahead, and what may be the potential economic value of such forecasts?*

To explain the variation in the spot price, represented by the Fish Pool Index, we included several predictors assumed to have influence on either supply or demand. First, we established a seasonal naïve model as a benchmark with which the more complex tree-based models were compared. Decision trees were chosen because they are computationally efficient and easily interpretable, while both random forest and xgBoost are more complex, and rely on averaging many individual decision trees for improved forecast accuracy. Because random forest and xgBoost are more complex, we expected them to perform better. The tree-based models are flexible, non-linear, and should be able to capture the high volatility in the Atlantic salmon price found by Asche et al. (2019) and Bloznelis (2016). In addition, we attempted to capture the one-year seasonality of the Fish Pool Index by employing STL-decomposition to break down the time series into trend-cycle, seasonal, and remainder-components, and compare the non-seasonally adjusted tree models with their seasonally adjusted counterparts.

The tree-based models displayed different levels of forecast accuracy, however, they all performed significantly better than the seasonal naïve benchmark. In general, the mean absolute errors (MAE) of the models ranged from 4.77 to 7.87, while the Fish Pool Index in the period ranged from 45.7 to 68.6. The average MAE of the tree-based models was 6.38, while the average Fish Pool Index observation was 58.57, suggesting that our forecasts are inaccurate by about 11% on average. The best directional accuracy was 82%, implying that the models correctly predicted up- or down-movements around 8 out of 10 times.

We found that both non-seasonally adjusted and seasonally adjusted random forest produced the overall best results, with significantly lower MAE and MSE than decision trees and xgBoost. The seasonally adjusted random forest also had the highest DA. Furthermore, we found that deseasonalizing the Fish Pool Index generally improved the prediction accuracy, in line with the findings of Bloznelis (2018), Guttormsen (1999), and Anderson & Gu (1995), suggesting that the spot price does exhibit seasonality. Utilizing the seasonally adjusted random forest, which in our view is the best model, we found that the third biggest Norwegian farmer, SalMar, could increase their annual earnings by 51.2 million NOK, or some 2%, by timing harvest decisions based on our forecasts.

In our analysis, the most important variables in predicting the Fish Pool Index were the Oslo seafood index, as well as demand-related factors such as the price of alternative meats (poultry in particular), and inflation in the European Union, the biggest salmon market in the world. We also observed that the top five most important variables in all models had lags shorter than six months, indicating that more recent data is more relevant in salmon price predictions. This was in line with the finding of Dahl et al. (2021), who found that the stock prices of large salmon companies may contain predictive power on the Fish Pool Index, as well as Asche et al. (2019) and Bloznelis (2016) who theorized that short-term supply inelasticity is the main cause of salmon price volatility.

References

Al-Maskari, F. (n.d.). *Lifestyle Diseases: An Economic Burden on the Health Services*. United Nations. Retrieved from: https://www.un.org/en/chronicle/article/lifestyle-diseases-economic-burden-health-services

Albertsen, M., Basso, M. N., Erraia, J., Fjose, S., Hernes, S., Jakobsen, E. (2021). *Eksportmeldingen 2021* (Menon-Publikasjon nr 58/2021). Menon Economics.

Anderson, J. L. & Gu, G. (1995). Deseasonalized state-space time series forecasting with application to the US salmon market. *Marine Resource Economics*, *10*(2), 171-185.

Ankamah-Yeboah, I., Nielsen, M., & Nielsen, R. (2017). Price formation of the salmon aquaculture futures market. *Aquaculture Economics & Management*, *21*(3), 376-399.

Asche, F., Misund, B., & Oglend, A. (2019). The case and cause of salmon price volatility. *Marine Resource Economics*, *34*(1), 23-38.

Athanasopoulos, G. & Hyndman, R. J. (2018). *Forecasting: principles and practice* (3rd ed.). OTexts.

Baser, P., Baser, N. & Saini, J. R. (2023). Gold Commodity Price Prediction Using Tree-based Prediction Models. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(1s), 90-96.

Bhansali, R. (1999). Parameter Estimation and Model Selection for Multistep Prediction of a Time Series: A Review. Ghosh, S. (Ed.). *Asymptotics, Nonparametrics, and Time series* (201-225). Marcel Dekker, Inc.

Bloznelis, D. (2018). Short-term salmon price forecasting. *Journal of Forecasting*, *37*(2), 151-169.

Bloznelis, D. (2016). Salmon price volatility: A weight-class-specific multivariate approach. *Aquaculture economics & management*, *20*(1), 24-53.

Bryan, M. F., & Cecchetti, S. G. (1993). The consumer price index as a measure of inflation. *National Bureau of Economic Research Working Paper Series* (Working Paper No. 4504).

Chen, E., & He, X. J. (2019). Crude oil price prediction with decision tree based regression approach. *Journal of International Technology and Information Management*, *27*(4), 2-16.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785-794).

Dahl, R. E., Oglend, A., & Yahya, M. (2021). Salmon stock market prices revealing salmon price information. *Marine Resource Economics*, *36*(2), 173-190.

Euronext1. (Retrieved: 01.02.2023). *Lerøy Seafood Group.* Retrieved from: https://live.euronext.com/nb/product/equities/NO0003096208-XOSL

Euronext2. (Retrieved: 01.02.2023). *MOWI.* Retrieved from: https://live.euronext.com/nb/product/equities/NO0003054108-XOSL

Euronext3. (Retrieved: 01.02.2023). *Oslo Seafood Index* [Dataset]. Retrieved from: https://live.euronext.com/nb/product/indices/NO0010580624-XOSL

Euronext4. (Retrieved: 01.02.2023). *SALMAR.* Retrieved from: https://live.euronext.com/nb/product/equities/NO0010310956-XOSL

Eurostat. (Retrieved: 01.02.2023). *HICP – monthly data (annual rate of change)* [Dataset]. Retrieved from: https://ec.europa.eu/eurostat/databrowser/view/PRC_HICP_MANR__custom_5078522/default/table?lang=en

Fattah, A. M. A., Rady, E. H. A. & Fawzy, H. (2021). Time series forecasting using tree based methods. *J. Stat. Appl. Probab*, *10*, 229-244.

Federal Reserve Bank of St. Louis1. (Retrieved: 01.02.2023). *Global Price of Beef* [Dataset]. Retrieved from: https://fred.stlouisfed.org/series/PBEEFUSDM

Federal Reserve Bank of St. Louis$_2$. (Retrieved: 01.02.2023). *Global Price of Lamb* [Dataset]. Retrieved from: https://fred.stlouisfed.org/series/PLAMBUSDM

Federal Reserve Bank of St. Louis$_3$. (Retrieved: 01.02.2023). *Global Price of Poultry* [Dataset]. Retrieved from: https://fred.stlouisfed.org/series/PPOULTUSDM

Federal Reserve Bank of St. Louis$_4$. (Retrieved: 01.02.2023). *Global Price of Swine* [Dataset]. Retrieved from: https://fred.stlouisfed.org/series/PPORKUSDM

Fish Pool$_1$. (n.d.). *Fish Pool Index™*. Retrieved from: https://fishpool.eu/fish-pool-index/

Fish Pool$_2$. (Retrieved: 21.02.2023). *FPI Weekly Details* [Dataset]. Retrieved from: https://fishpool.eu/fpi-weekly-details/

Fish Pool$_3$. (Retrieved: 01.02.2023). *Forward Price History* [Dataset]. Retrieved from: https://fishpool.eu/forward-price-history/

Fiskeridirektoratet. (2021). *Nøkkeltall fra norsk havbruksnæring 2021.* (1893-6946) Fiskeridirektoratet.

Fiskeridirektoratet. (25.05.2022). *Økt salg av oppdrettsfisk i 2021.* https://www.fiskeridir.no/Akvakultur/Nyheter/2022/okt-salg-av-oppdrettsfisk-i-2021

Fiskeridirektoratet$_1$. (26.01.2023). *Antall tillatelser 1994-2022* [Dataset]. Retrieved from: https://www.fiskeridir.no/Akvakultur/Tall-og-analyse/Akvakulturstatistikk-tidsserier/Laks-regnbueoerret-og-oerret/Matfiskproduksjon

Fiskeridirektoratet$_2$. (23.01.2023). *Beholdning ved månedsslutt fordelt på art 2005-2023 (Fylke)* [Dataset]. Retrieved from: https://www.fiskeridir.no/Akvakultur/Tall-og-analyse/Biomassestatistikk/Biomassestatistikk-etter-fylke

Fiskeridirektoratet$_3$. (19.01.2023). *Forbruk av fôr fordelt på art 2005-2023 (Fylke)* [Dataset]. Retrieved from: https://www.fiskeridir.no/Akvakultur/Tall-og-analyse/Biomassestatistikk/Biomassestatistikk-etter-fylke

Fiskeridirektoratet4. (23.01.2023). *Produksjonsoversikt fordelt på art 2005-2023 (Fylke)* [Dataset]. Retrieved from: https://www.fiskeridir.no/Akvakultur/Tall-og-analyse/Biomassestatistikk/Biomassestatistikk-etter-fylke

Fiskeridirektoratet5. (19.01.2023). *Svinn i produksjonen fordelt på art og årsak 2005-2023 (Fylke)* [Dataset]. Retrieved from: https://www.fiskeridir.no/Akvakultur/Tall-og-analyse/Biomassestatistikk/Biomassestatistikk-etter-fylke

Fiskeridirektoratet6. (19.01.2023). *Uttak av slaktet fisk fordelt på art 2005-2023 (Fylke)* [Dataset]. Retrieved from: https://www.fiskeridir.no/Akvakultur/Tall-og-analyse/Biomassestatistikk/Biomassestatistikk-etter-fylke

Food and Agriculture Organization of the United Nations. (2022). *Meat Market Review: Emerging trends and outlook 2022.* Retrieved from: https://www.fao.org/markets-and-trade/commodities/meat/en/

Global Salmon Initiative. (2021). *Farmed Salmon's Role in Sustainable Food Systems.* (2021 Sustainability Report). Retrieved from: *https://globalsalmoninitiative.org/en/sustainability-report/sustainable-food-systems/#carbon-footprint*

Guttormsen, A. G. (1999). Forecasting weekly salmon prices: Risk management in fish farming. *Aquaculture Economics & Management*, *3*(2), 159-166.

Hastie, T., James, G., Tibshirani, R. & Witten, D. (2013). *An introduction to statistical learning* (1st ed.). Springer.

Hoddevik, B. (13.02.2023). *Risikorapporten: Fortsatt høy dødelighet hos oppdrettslaks.* Havforskningsinstituttet. Retrieved from: https://www.hi.no/hi/nyheter/2023/februar/fortsatt-hoy-dodelighet-hos-oppdrettslaks

Kang, H. (2013). The prevention and handling of the missing data. *Korean journal of anesthesiology*, *64*(5), 402-406.

Kontali Analyse. (2022). *Salmon World 2022* (Yearly, 2022). Kontali Analyse.

Kuhn, M. (27.03.2019) *The caret Package,* Github, Retrieved from: https://topepo.github.io/caret/index.html

Marcellino, M., Stock, J. H., & Watson, M. W. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of econometrics*, *135*(1-2), 499-526.

Mowi. (2020). *Salmon Farming Industry Handbook 2020.* Retrieved from: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjH7cWx-pL_AhWSSPEDHRdWCS0QFnoECA8QAQ&url=https%3A%2F%2Fmowi.com%2Fit%2Fwp-content%2Fuploads%2Fsites%2F16%2F2020%2F06%2FMowi-Salmon-Farming-Industry-Handbook-2020.pdf&usg=AOvVaw1bYJczO560AF_d54qRjpu7

National Oceanic and Atmospheric Administration. (Retrieved: 01.02.2023). *Landings* [Dataset]. Retrieved from: https://www.fisheries.noaa.gov/foss/f?p=215:200:1190813404312:Mail::::

Nielsen, D. (2016). *Tree boosting with xgboost-why does xgboost win" every" machine learning competition?* [Master thesis]. Norwegian University of Science and Technology (NTNU).

Norges Bank1. (Retrieved: 01.02.2023). *Valutakurser NOK-EUR* [Dataset]. Retrieved from: https://www.norges-bank.no/tema/Statistikk/Valutakurser/?tab=currency&id=EUR

Norges Bank2. (Retrieved: 01.02.2023). *Valutakurser NOK-USD* [Dataset]. Retrieved from: https://www.norges-bank.no/tema/Statistikk/Valutakurser/?tab=currency&id=USD

Norsk Klimaservicesenter. (Retrieved: 01.02.2023) *Observasjoner og værstatistikk* [Dataset]. Retrieved from: https://seklima.met.no/observations/

Petropoulos, F. & Svetunkov, I. (2018). Old dog, new tricks: a modelling view of simple moving averages. *International Journal of Production Research*, *56*(18), 6034-6047.

Ritchie, H. & Roser, M. (2021, October). *Fish and Overfishing* [Dataset]. Our World in Data. Retrieved from: https://ourworldindata.org/fish-and-overfishing

SalMar. (18.02.2022). *Quarterly Report* (Fourth Quarter 2021). Retrieved from: https://www.salmar.no/en/quarterly-reports/

Searchinger, T., Waite, R., Hanson, C., Ranganathan, J., Dumas, P., & Matthews, E. (2019). *World resources report:* Creating *a sustainable food future.* (Final Report 2019). World Resources Institute. Retrieved from: https://research.wri.org/wrr-food

Selnæs, J. (personal communication, 26.04.2023). *Lakselus* [Dataset]. Lusedata.

Simon, R. & Varma, S. (2006). Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, *7*(1), 1-8.

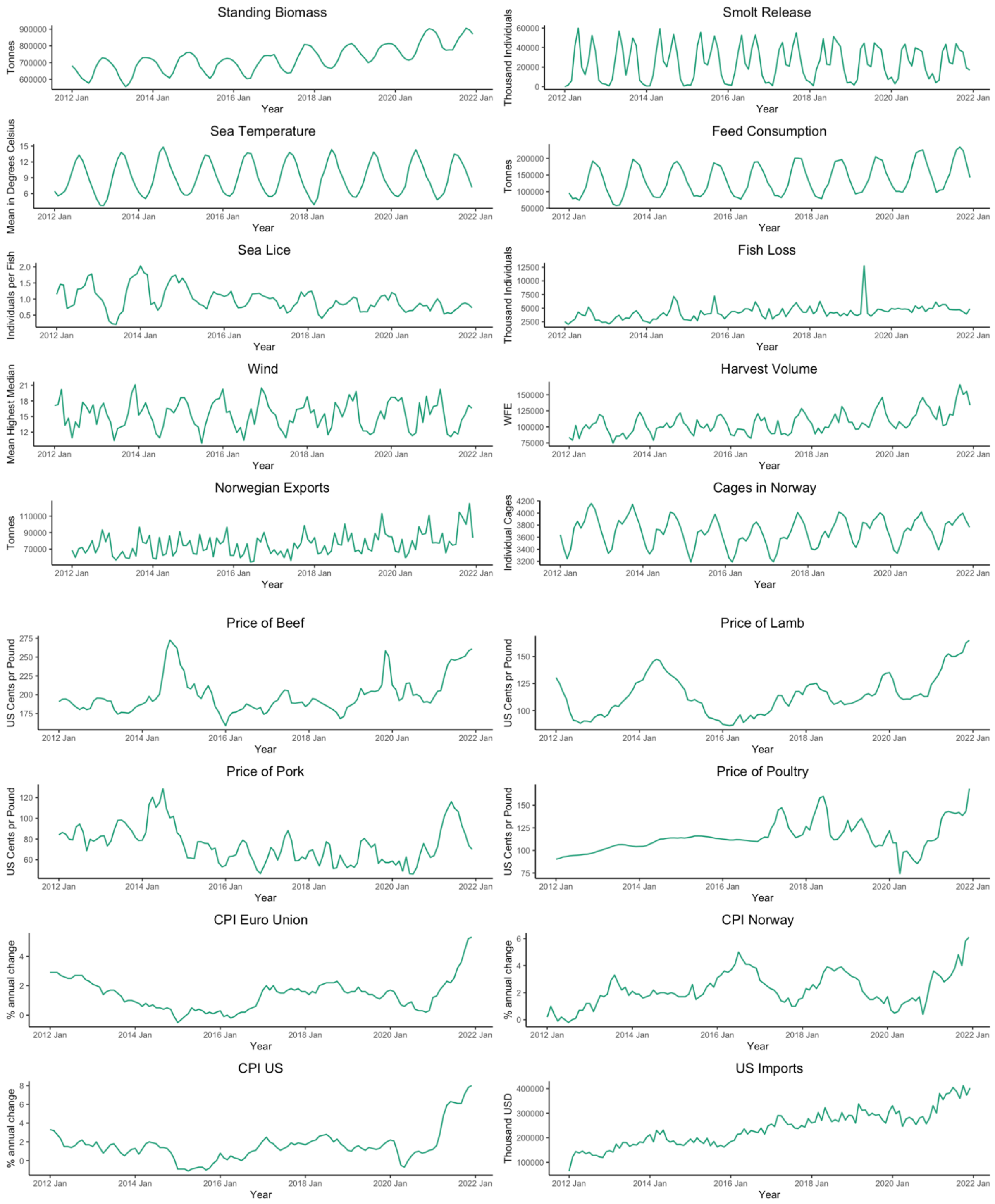
Singh, P., Singh, N., Singh, K. K., & Singh, A. (2021). Diagnosing of disease using machine learning. In Machine learning and the internet of medical things in healthcare (pp. 89-111). Academic Press.

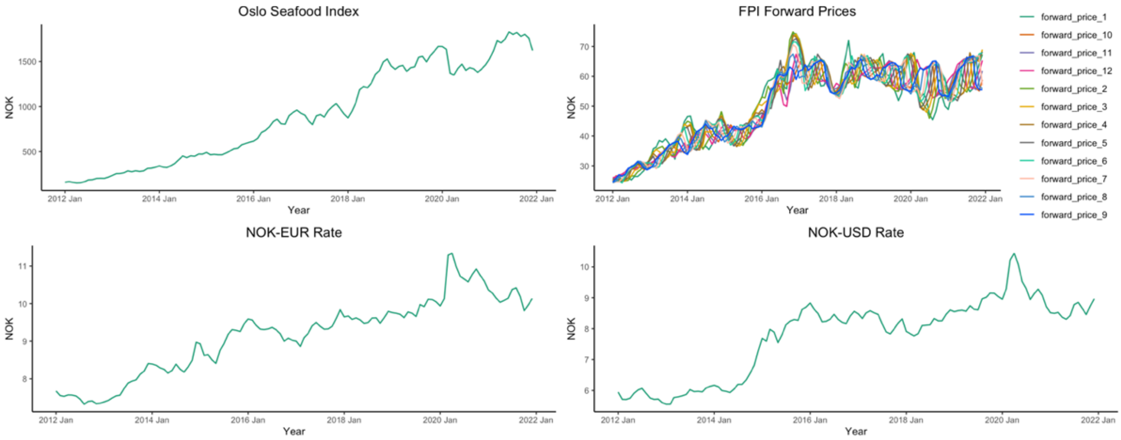
SSB. (Retrieved: 01.02.2023). *Eksport av laks* [Dataset]. Statistics Norway (Statistisk Sentralbyrå). Retrieved from: https://www.ssb.no/statbank/list/laks

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International journal of forecasting*, *16*(4), 437-450.

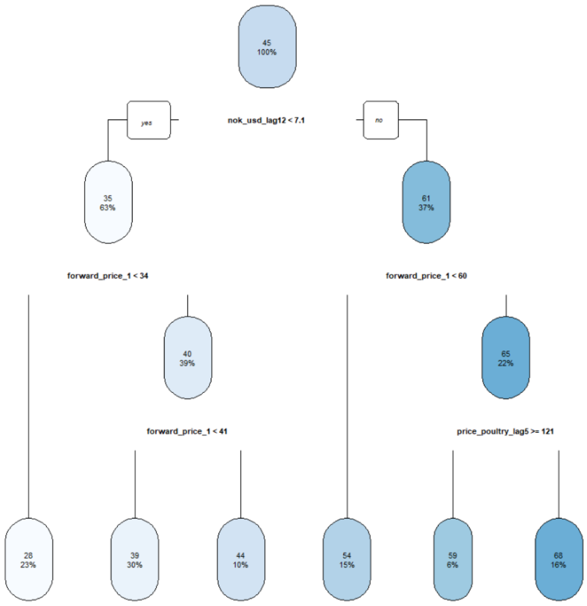
United Nations. (n.d.). *World population projected to reach 9.8 billion in 2050, and 11.2 billion in 2100.* Retrieved from: https://www.un.org/en/desa/world-population-projected-reach-98-billion-2050-and-112-billion-2100#:~:text=COVID-19 ,World%20population%20projected%20to%20reach%209.8%20billion%20in%202050%2C%20and,Nations%20report%20being%20launched%20today.

Appendix 1 – Feature Vizualisations

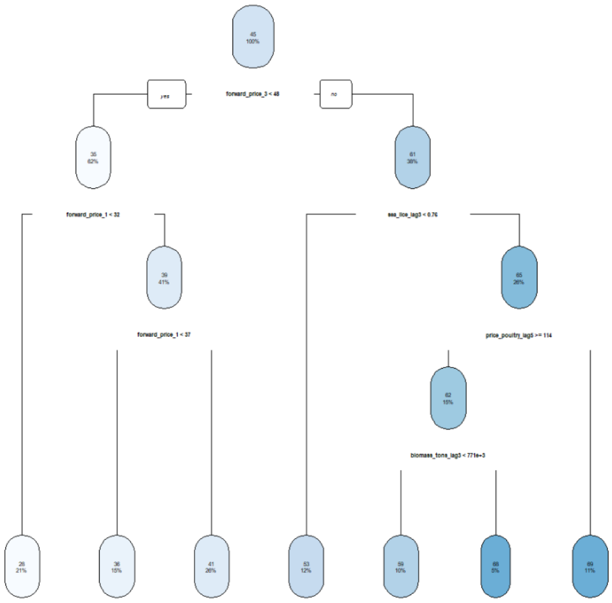



Oslo Seafood Index
NOK
1500
1000
500
2012 Jan
2014 Jan
2016 Jan
2018 Jan
2020 Jan
2022 Jan
Year
FPI Forward Prices
NOK
70
60
50
40
30
2012 Jan
2014 Jan
2016 Jan
2018 Jan
2020 Jan
2022 Jan
Year
forward_price_1
forward_price_10
forward_price_11
forward_price_12
forward_price_2
forward_price_3
forward_price_4
forward_price_5
forward_price_6
forward_price_7
forward_price_8
forward_price_9
NOK-EUR Rate
NOK
11
10
9
8
2012 Jan
2014 Jan
2016 Jan
2018 Jan
2020 Jan
2022 Jan
Year
NOK-USD Rate
NOK
10
9
8
7
6
2012 Jan
2014 Jan
2016 Jan
2018 Jan
2020 Jan
2022 Jan
Year

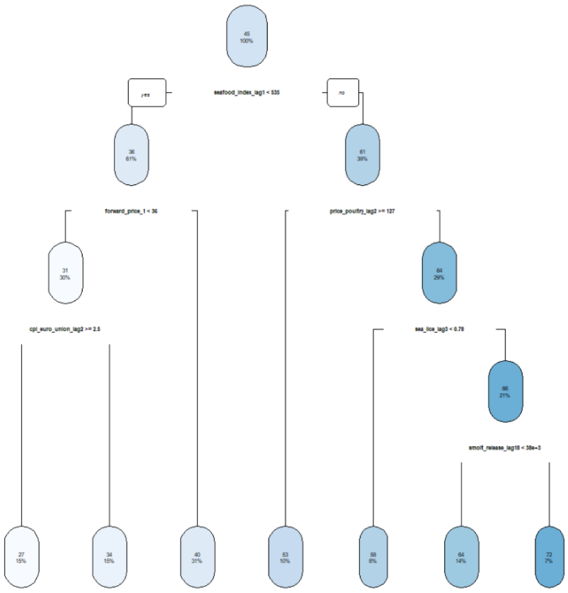
Appendix 2 – Decision Tree Plots



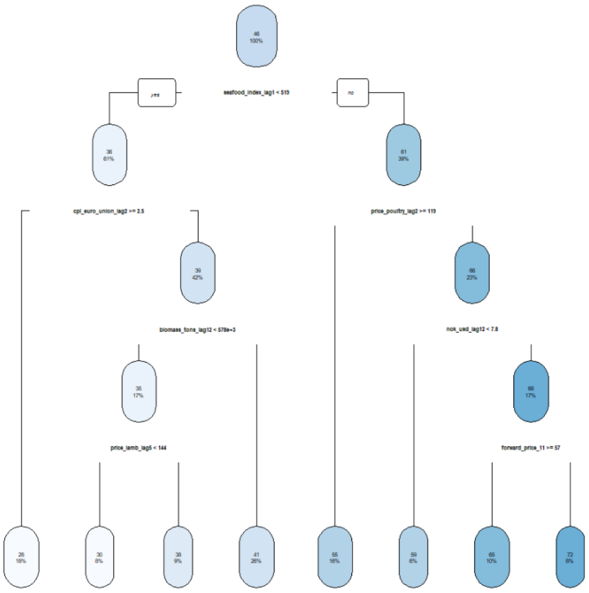
Decision Tree 1



Decision Tree 2

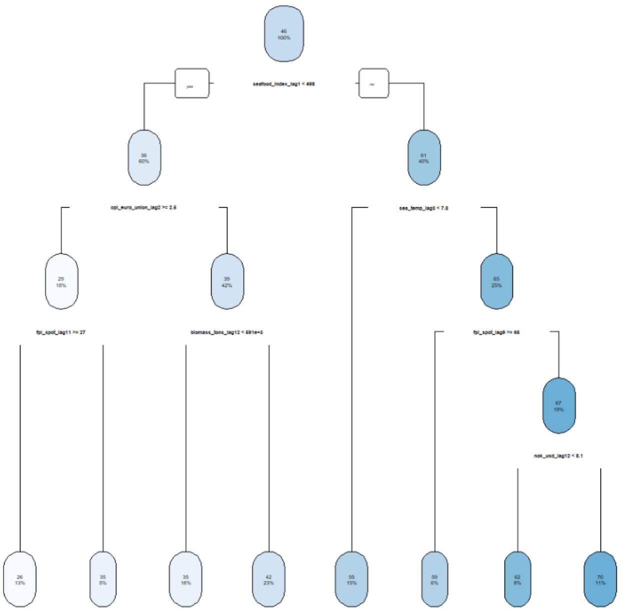


Decision Tree 3

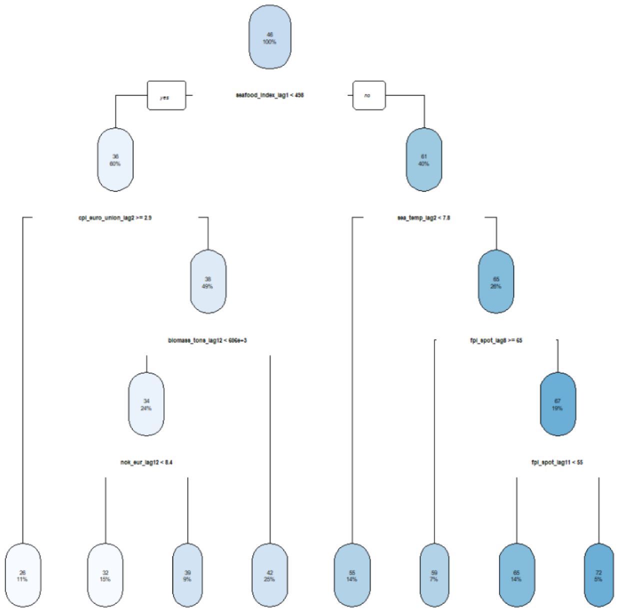


Decision Tree 4

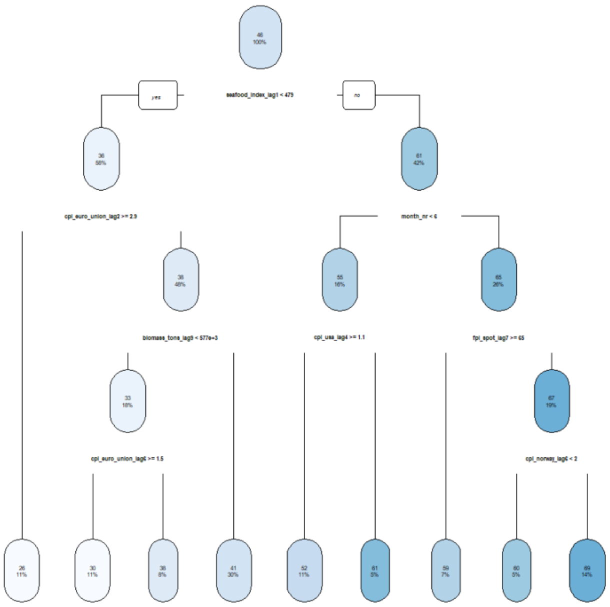
Decision Tree 5



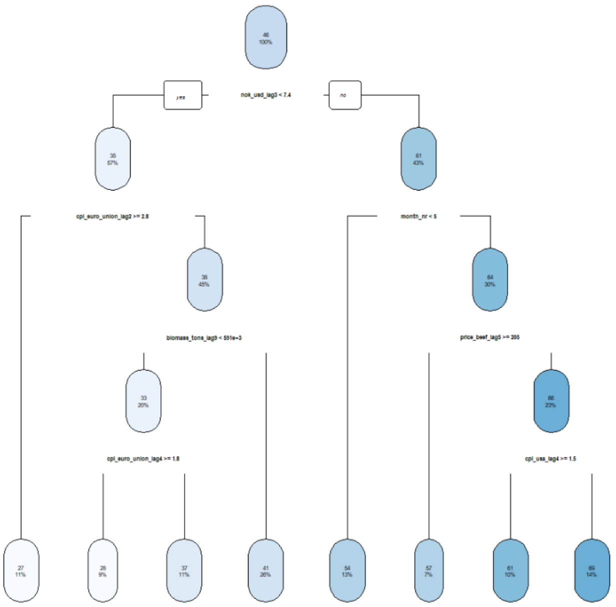
Decision Tree 6

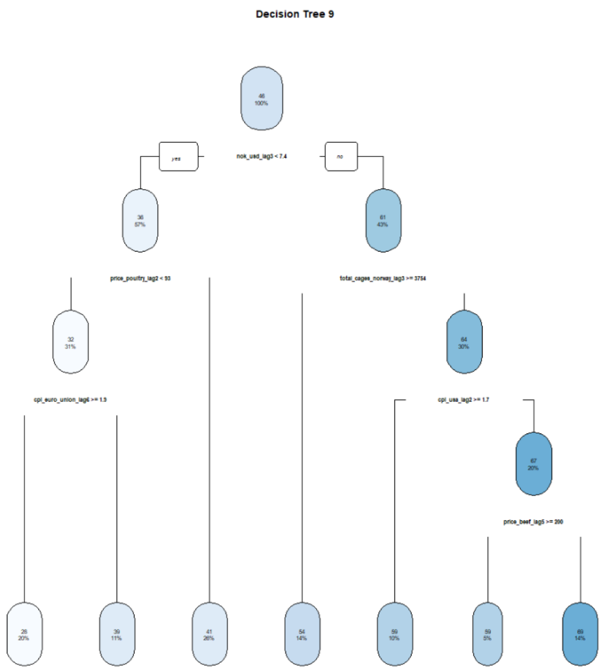


Decision Tree 7

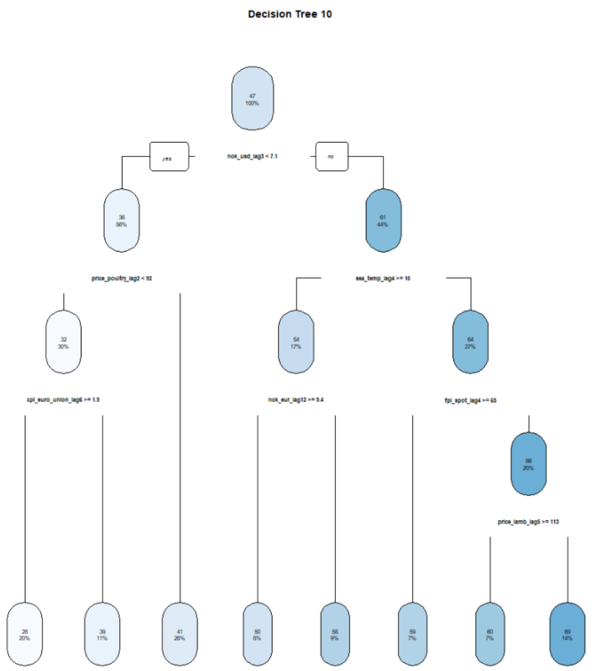


Decision Tree 8

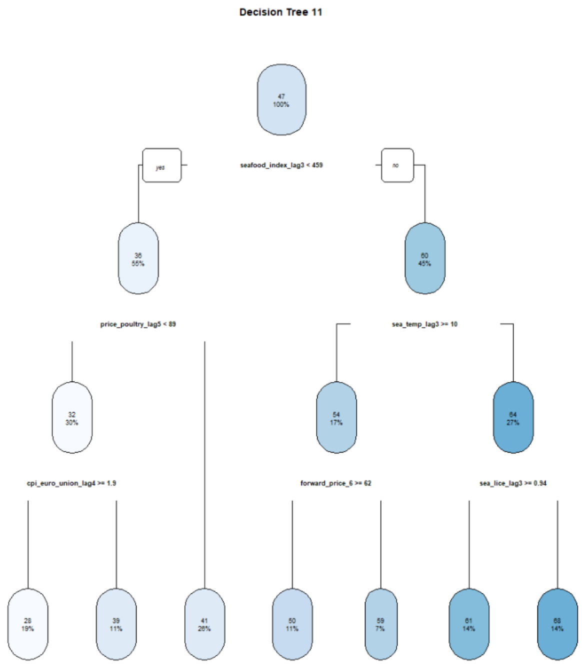




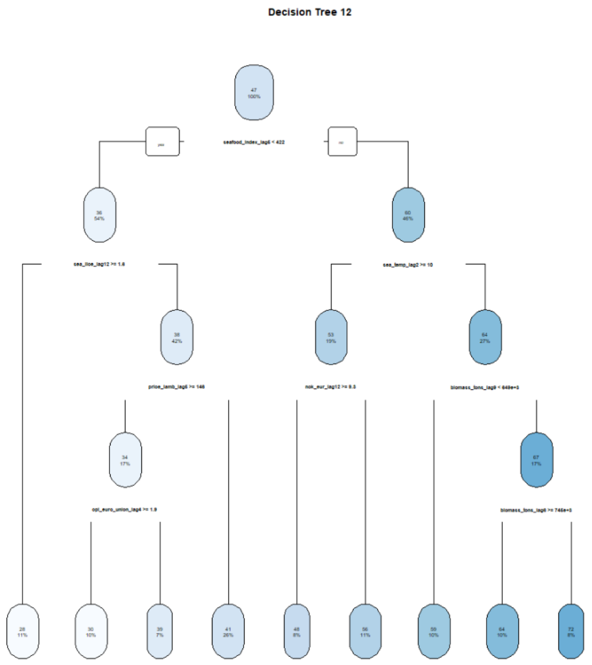
Decision Tree 9



Decision Tree 10

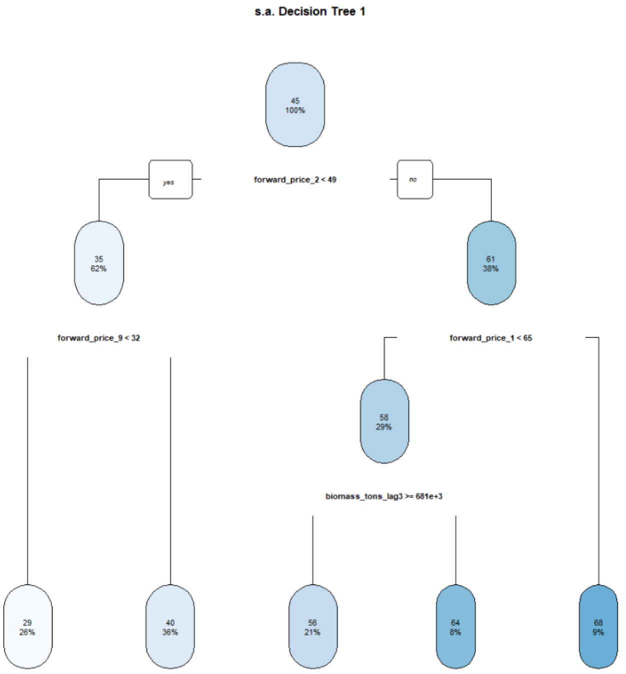


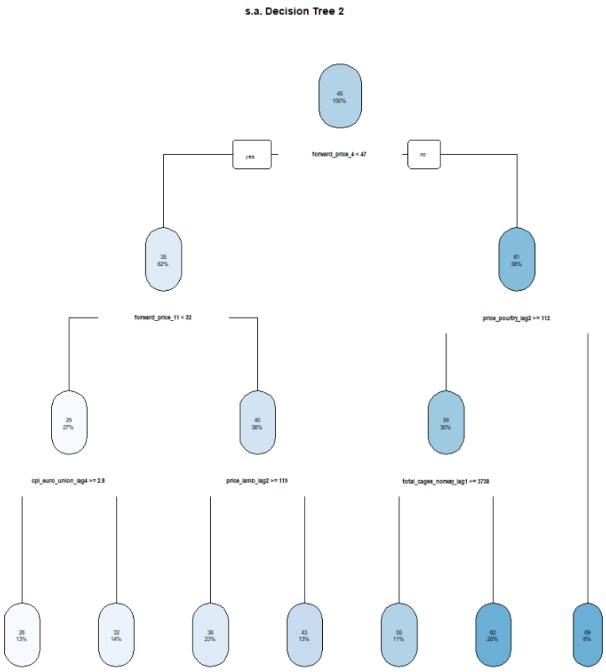
Decision Tree 11

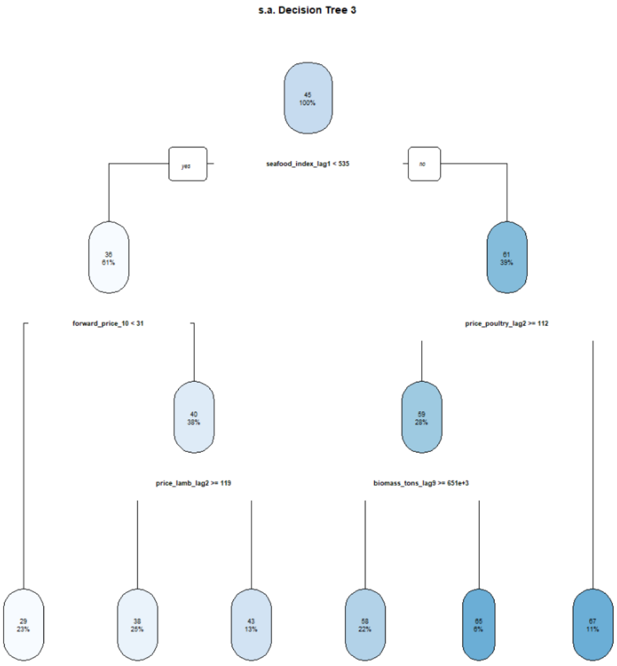


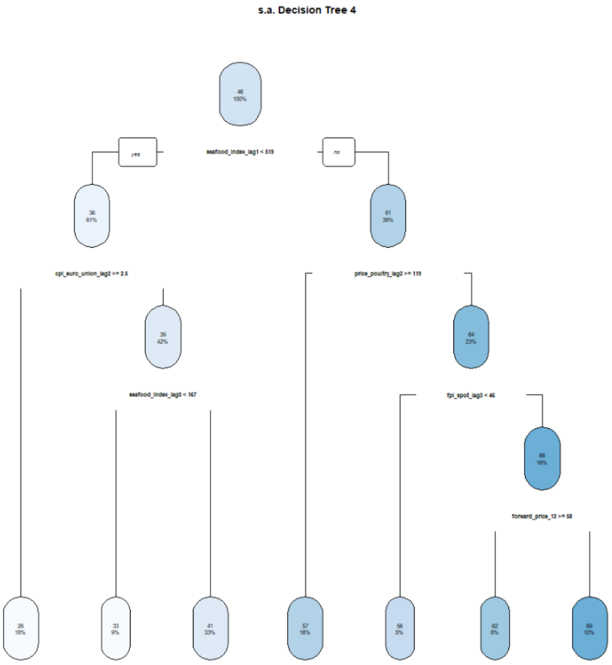
Decision Tree 12

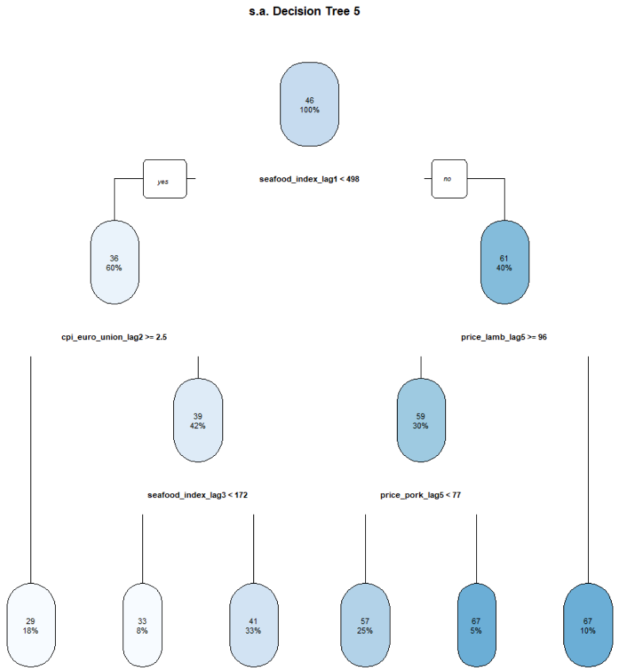
Appendix 3 – Seasonally Adjusted Decision Tree Plots

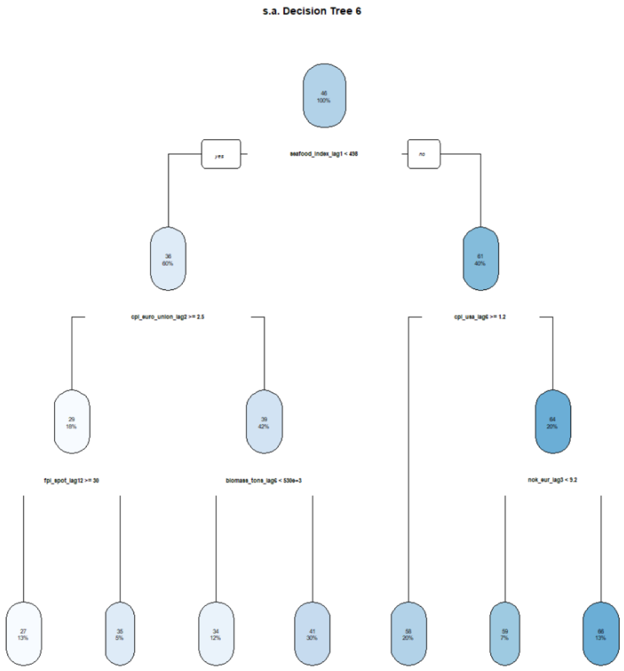

s.a. Decision Tree 1

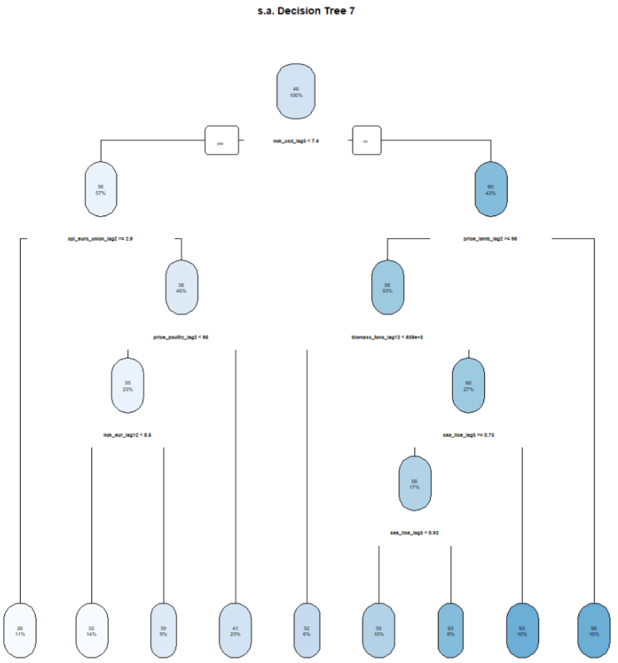

s.a. Decision Tree 2

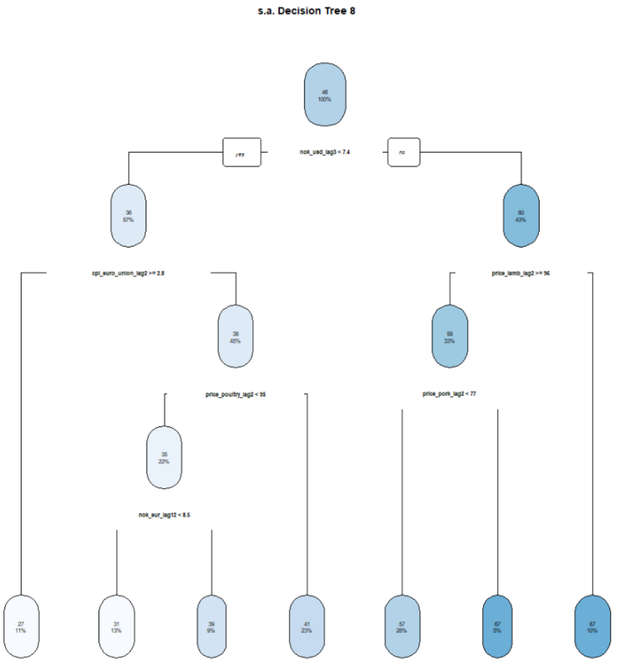

s.a. Decision Tree 3

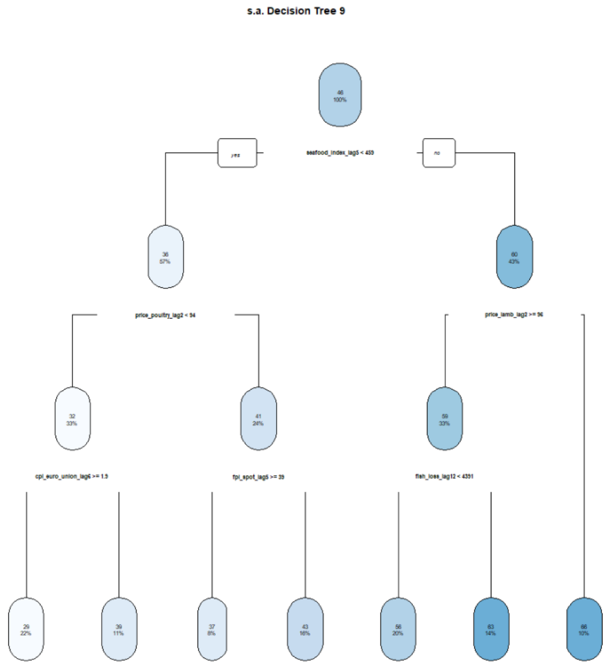

s.a. Decision Tree 4


s.a. Decision Tree 5
46
100%
yes
seafood_index_lag1 < 498
no
36
60%
61
40%
cpi_euro_union_lag2 >= 2.5
price_lamb_lag5 >= 96
39
42%
59
30%
seafood_index_lag3 < 172
price_pork_lag5 < 77
29
18%
33
8%
41
33%
57
25%
67
5%
67
10%

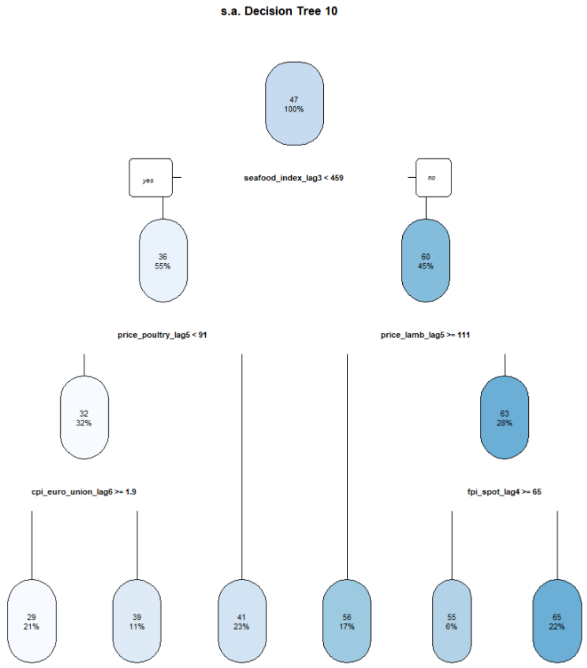

s.a. Decision Tree 6
46
100%
yes
seafood_index_lag1 < 498
no
36
60%
61
40%
cpi_euro_union_lag2 >= 2.5
cpi_usa_lag6 >= 1.2
29
18%
39
42%
64
20%
fpi_spot_lag12 >= 30
biomass_tons_lag6 < 530e+3
nok_eur_lag3 < 9.2
27
13%
35
5%
34
12%
41
30%
58
20%
59
7%
68
13%


s.a. Decision Tree 7
46
100%
yes
nok_usd_lag3 < 7.4
no
36
57%
60
43%
cpi_euro_union_lag2 >= 2.9
price_lamb_lag2 >= 96
38
46%
58
33%
price_poultry_lag2 < 96
biomass_tons_lag12 < 639e+3
35
23%
60
27%
nok_eur_lag12 < 8.5
sea_lice_lag3 >= 0.73
58
17%
sea_lice_lag3 < 0.92
26
11%
32
14%
39
9%
41
23%
52
6%
55
10%
63
8%
63
10%
66
10%

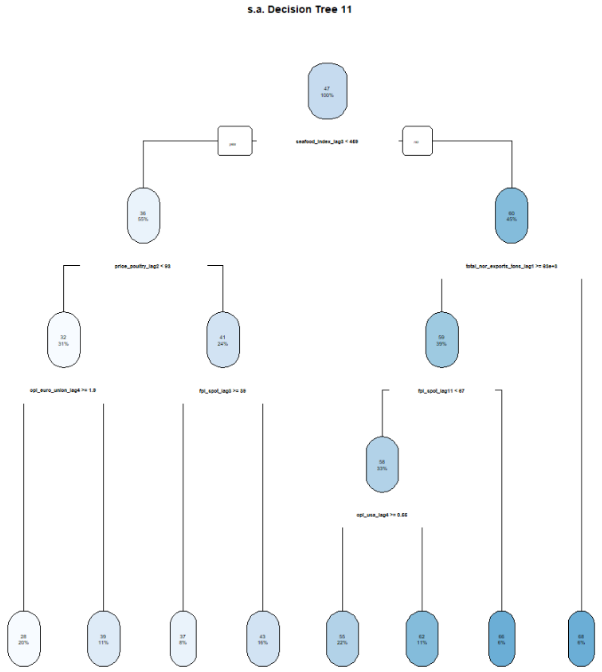

s.a. Decision Tree 8
46
100%
yes
nok_usd_lag3 < 7.4
no
36
57%
60
43%
cpi_euro_union_lag2 >= 2.8
price_lamb_lag2 >= 96
38
45%
58
33%
price_poultry_lag2 < 95
price_pork_lag2 < 77
35
22%
nok_eur_lag12 < 8.5
27
11%
31
13%
39
9%
41
23%
57
28%
67
5%
67
10%



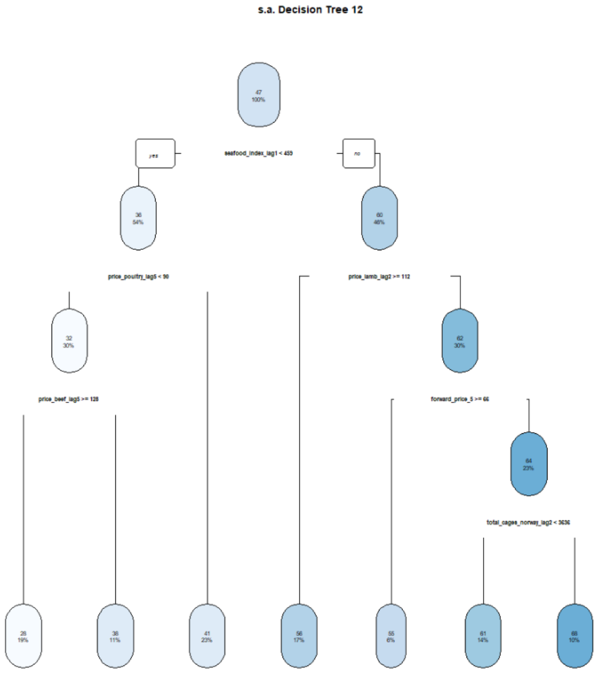
s.a. Decision Tree 9



s.a. Decision Tree 10

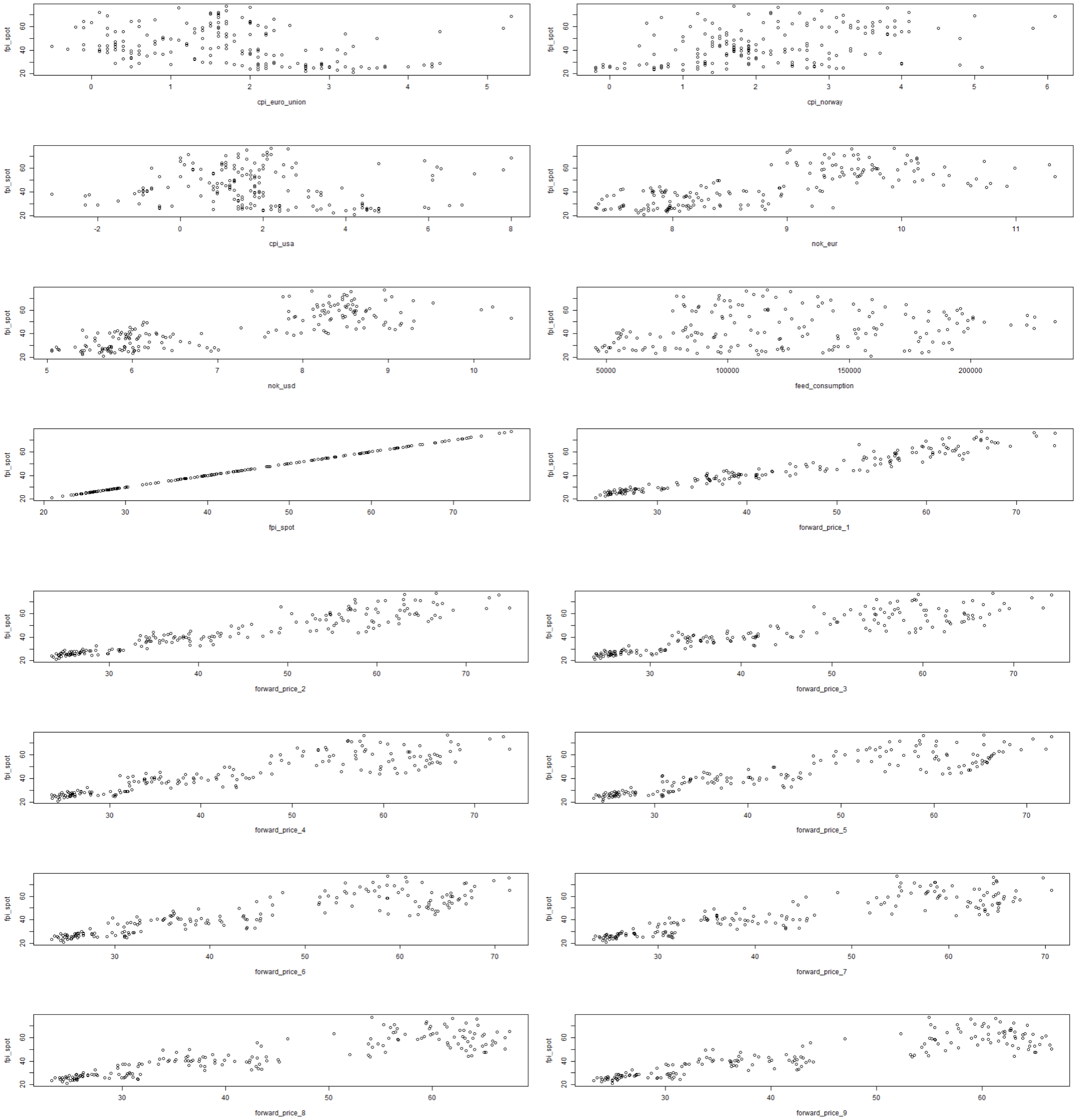


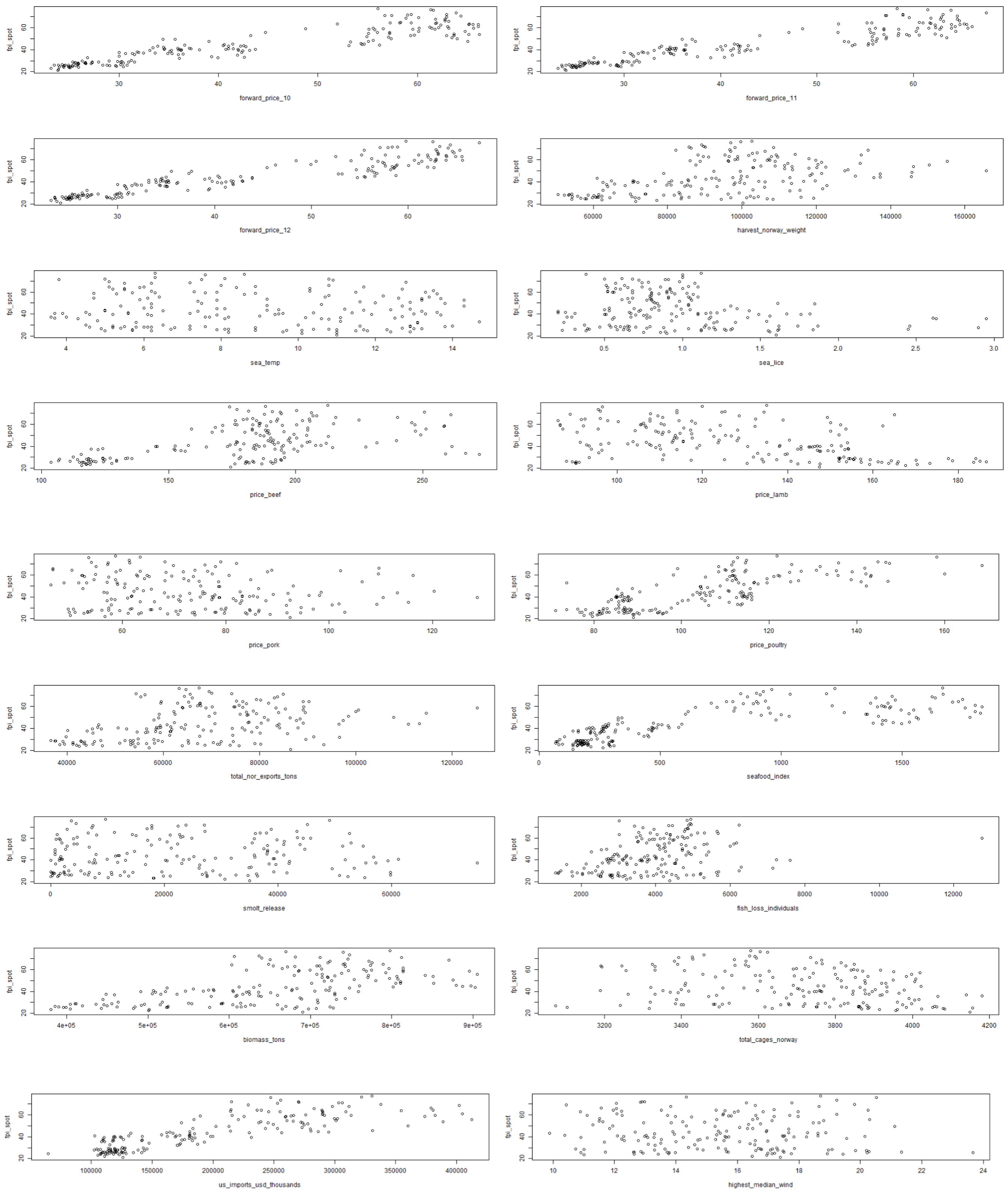
s.a. Decision Tree 11



s.a. Decision Tree 12

Appendix 4 – Fish Pool Index against Features




fpi_spot
20
40
60
30
40
50
60
forward_price_10
fpi_spot
30
40
50
60
forward_price_11
fpi_spot
30
40
50
60
forward_price_12
fpi_spot
60000
80000
100000
120000
140000
160000
harvest_norway_weight
fpi_spot
4
6
8
10
12
14
sea_temp
fpi_spot
0.5
1.0
1.5
2.0
2.5
3.0
sea_lice
fpi_spot
100
150
200
250
price_beef
fpi_spot
100
120
140
160
180
price_lamb
fpi_spot
60
80
100
120
price_pork
fpi_spot
80
100
120
140
160
price_poultry
fpi_spot
40000
60000
80000
100000
120000
total_nor_exports_tons
fpi_spot
0
500
1000
1500
seafood_index
fpi_spot
0
20000
40000
60000
smolt_release
fpi_spot
2000
4000
6000
8000
10000
12000
fish_loss_individuals
fpi_spot
4e+05
5e+05
6e+05
7e+05
8e+05
9e+05
biomass_tons
fpi_spot
3200
3400
3600
3800
4000
4200
total_cages_norway
fpi_spot
100000
150000
200000
250000
300000
350000
400000
us_imports_usd_thousands
fpi_spot
10
12
14
16
18
20
22
24
highest_median_wind