Predicting Fuel Consumption of MS Smyril with Speed, Pitch, and Rudder data

Satu Laukkanen

1 Introduction

In this research paper, we will explore the data set collected from a ship in the time period of February to April 2010.

This paper explores the prediction of the estimated fuel consumption in tons/day based on sensor readings from the different sensors on MS Smyril. The goal is to train models to predict the estimated fuel consumption in terms of tons/day at a given time.

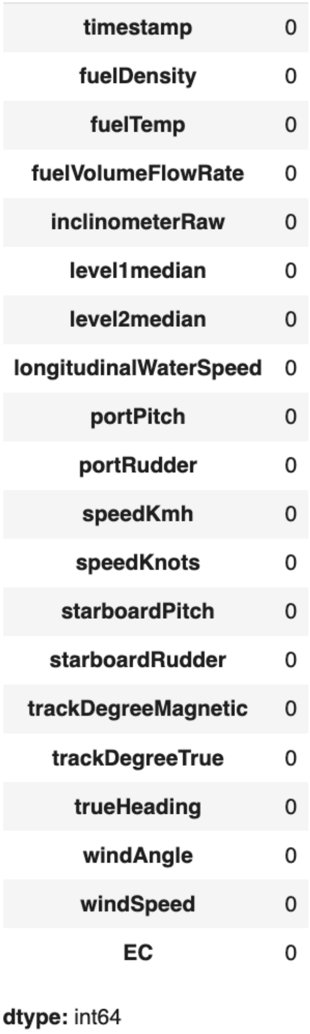
Fuel consumption is one of the largest consumption costs for ships, and has a substantial environmental impact. Thus, optimizing fuel consumption is essential for economic and environmental reasons.

2 Data processing

The data consists of sensor readings from the ship's sensor systems from 16th February 2010 to 12th April 2010, with a total of 246 voyages and 1,627,324 data records.

The calculation interval chosen for this task is seconds. The data is initially presented in a .NET timestamp format, and timestamps vary greatly between different sensors. A data frame for fuel-related sensors was created initially. If the fuelVolumeFlowRate was 0, the row was dropped, resulting in a data frame with 1,627,251 data records.

The tables were merged based on proximity to the nearest timestamp to handle the different timestamps in the sensor logs.



```
              timestamp  0
            fuelDensity  0
               fuelTemp  0
     fuelVolumeFlowRate  0
         inclinometerRaw 0
            level1median 0
            level2median 0
longitudinalWaterSpeed   0
               portPitch 0
              portRudder 0
                speedKmh 0
              speedKnots 0
          starboardPitch 0
         starboardRudder 0
    trackDegreeMagnetic  0
         trackDegreeTrue 0
             trueHeading 0
               windAngle 0
               windSpeed 0
                      EC 0

dtype: int64
```

Figure 1: NaN-values in the initial data frame

As the data frame was merged in a manner that found the closest match to each row in the fuel-related sensor data, there are no NaN-values.

2.1 Estimated Fuel Consumption

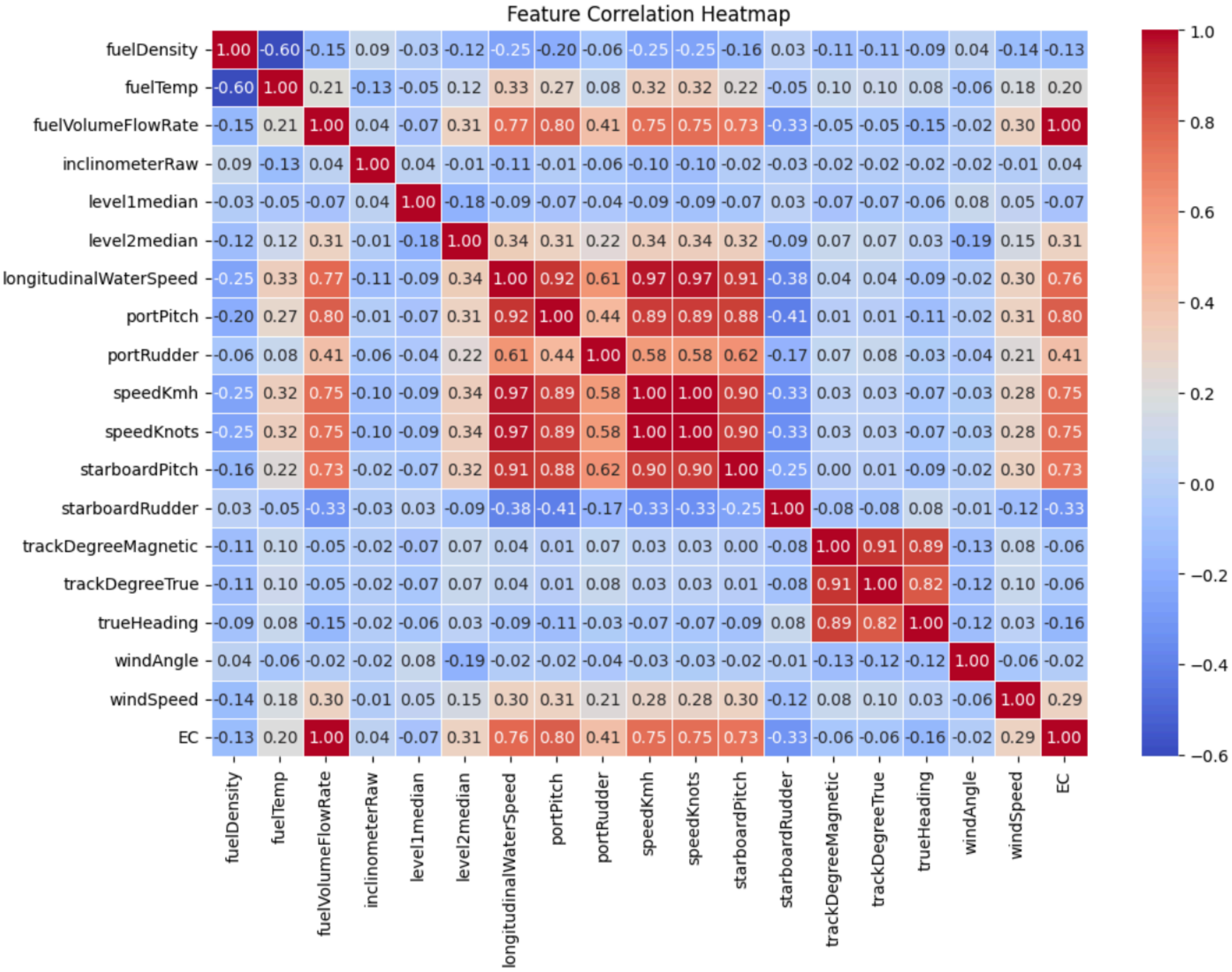
In this analysis, the target value is the estimated consumption of fuel, measured in tons in a day. The formula gives the estimated fuel consumption on a given day. This target value was added to the data frame on each row.

2.2 Correlation Analysis

A correlation analysis was conducted on the whole data frame to explore the correlation between different features and the target variable EC. As seen in the correlation matrix, fuelVolumeFlowRate has the highest correlation with the target variable.

Among other highly correlated variables, we can find longitudinalWaterSpeed, portPitch, speedKmh, speedKnots, and starboardPitch. None of the features have a high negative correlation with the target EC.

As seen, many of the other features have a high correlation between each other, such as speedKnots and speedKmh.



	fuelDensity	fuelTemp	fuelVolumeFlowRate	inclinometerRaw	level1median	level2median	longitudinalWaterSpeed	portPitch	portRudder	speedKmh	speedKnots	starboardPitch	starboardRudder	trackDegreeMagnetic	trackDegreeTrue	trueHeading	windAngle	windSpeed	EC
fuelDensity	1.00	-0.60	-0.15	0.09	-0.03	-0.12	-0.25	-0.20	-0.06	-0.25	-0.25	-0.16	0.03	-0.11	-0.11	-0.09	0.04	-0.14	-0.13
fuelTemp	-0.60	1.00	0.21	-0.13	-0.05	0.12	0.33	0.27	0.08	0.32	0.32	0.22	-0.05	0.10	0.10	0.08	-0.06	0.18	0.20
fuelVolumeFlowRate	-0.15	0.21	1.00	0.04	-0.07	0.31	0.77	0.80	0.41	0.75	0.75	0.73	-0.33	-0.05	-0.05	-0.15	-0.02	0.30	1.00
inclinometerRaw	0.09	-0.13	0.04	1.00	0.04	-0.01	-0.11	-0.01	-0.06	-0.10	-0.10	-0.02	-0.03	-0.02	-0.02	-0.02	-0.02	-0.01	0.04
level1median	-0.03	-0.05	-0.07	0.04	1.00	-0.18	-0.09	-0.07	-0.04	-0.09	-0.09	-0.07	0.03	-0.07	-0.07	-0.06	0.08	0.05	-0.07
level2median	-0.12	0.12	0.31	-0.01	-0.18	1.00	0.34	0.31	0.22	0.34	0.34	0.32	-0.09	0.07	0.07	0.03	-0.19	0.15	0.31
longitudinalWaterSpeed	-0.25	0.33	0.77	-0.11	-0.09	0.34	1.00	0.92	0.61	0.97	0.97	0.91	-0.38	0.04	0.04	-0.09	-0.02	0.30	0.76
portPitch	-0.20	0.27	0.80	-0.01	-0.07	0.31	0.92	1.00	0.44	0.89	0.89	0.88	-0.41	0.01	0.01	-0.11	-0.02	0.31	0.80
portRudder	-0.06	0.08	0.41	-0.06	-0.04	0.22	0.61	0.44	1.00	0.58	0.58	0.62	-0.17	0.07	0.08	-0.03	-0.04	0.21	0.41
speedKmh	-0.25	0.32	0.75	-0.10	-0.09	0.34	0.97	0.89	0.58	1.00	1.00	0.90	-0.33	0.03	0.03	-0.07	-0.03	0.28	0.75
speedKnots	-0.25	0.32	0.75	-0.10	-0.09	0.34	0.97	0.89	0.58	1.00	1.00	0.90	-0.33	0.03	0.03	-0.07	-0.03	0.28	0.75
starboardPitch	-0.16	0.22	0.73	-0.02	-0.07	0.32	0.91	0.88	0.62	0.90	0.90	1.00	-0.25	0.00	0.01	-0.09	-0.02	0.30	0.73
starboardRudder	0.03	-0.05	-0.33	-0.03	0.03	-0.09	-0.38	-0.41	-0.17	-0.33	-0.33	-0.25	1.00	-0.08	-0.08	0.08	-0.01	-0.12	-0.33
trackDegreeMagnetic	-0.11	0.10	-0.05	-0.02	-0.07	0.07	0.04	0.01	0.07	0.03	0.03	0.00	-0.08	1.00	0.91	0.89	-0.13	0.08	-0.06
trackDegreeTrue	-0.11	0.10	-0.05	-0.02	-0.07	0.07	0.04	0.01	0.08	0.03	0.03	0.01	-0.08	0.91	1.00	0.82	-0.12	0.10	-0.06
trueHeading	-0.09	0.08	-0.15	-0.02	-0.06	0.03	-0.09	-0.11	-0.03	-0.07	-0.07	-0.09	0.08	0.89	0.82	1.00	-0.12	0.03	-0.16
windAngle	0.04	-0.06	-0.02	-0.02	0.08	-0.19	-0.02	-0.02	-0.04	-0.03	-0.03	-0.02	-0.01	-0.13	-0.12	-0.12	1.00	-0.06	-0.02
windSpeed	-0.14	0.18	0.30	-0.01	0.05	0.15	0.30	0.31	0.21	0.28	0.28	0.30	-0.12	0.08	0.10	0.03	-0.06	1.00	0.29
EC	-0.13	0.20	1.00	0.04	-0.07	0.31	0.76	0.80	0.41	0.75	0.75	0.73	-0.33	-0.06	-0.06	-0.16	-0.02	0.29	1.00

Figure 2: Correlation Matrix of Initial Features

2.3 Visual Data Analysis

A brief visual data analysis was conducted. A daily mean calculation over the estimated fuel consumption EC was plotted in Figure 3, showing that MS Smyril was docked on the 2nd and 4th of April. The daily mean for estimated fuel consumption is 46,9771 tons/day.

Figure 4 depicts the rate of estimated fuel consumption during the hours of a day in the given time interval.

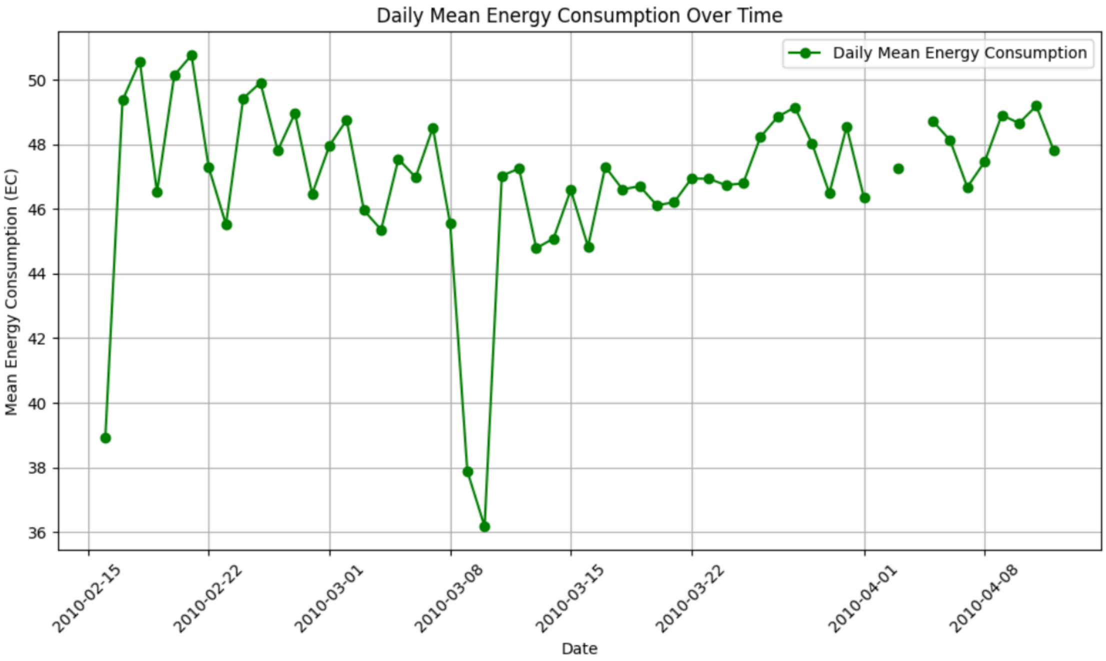


Figure 3: Estimated fuel consumption mean for each day of the measuring period

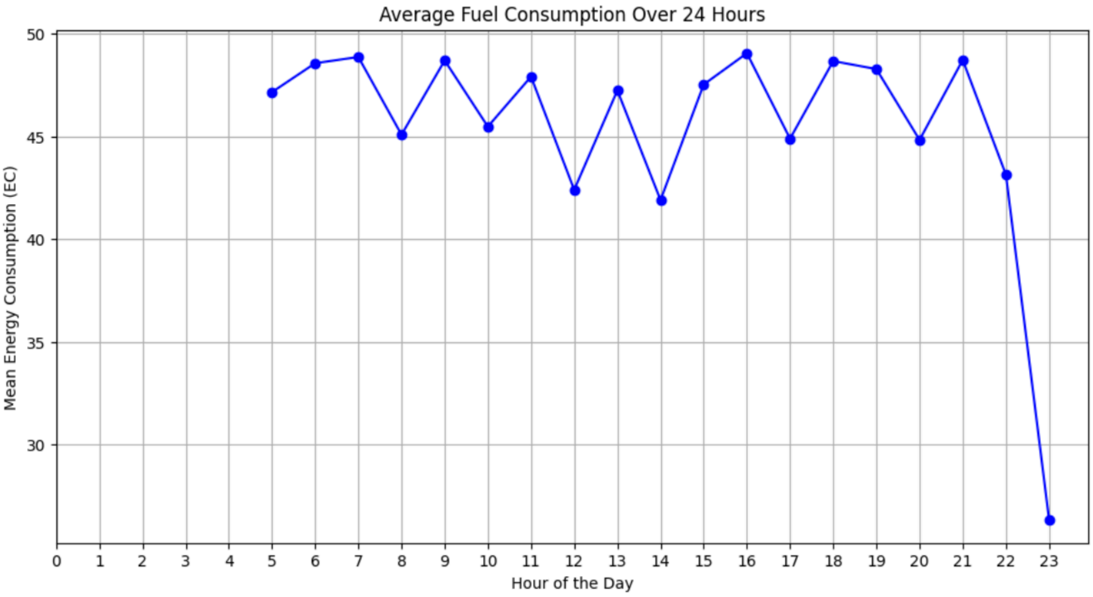


Figure 4: Estimated fuel consumption mean during the hours of a day

2.4 Selection of Features for Training

In this analysis, the effects of speed, pitch, and rudder data are explored to predict the target value EC. The fuel-based data is not selected, as they are used to compute the target value EC. The final selection of features:

- longitudinalWaterSpeed
- portPitch
- portRudder
- speedKnots
- starboardPitch
- starboardRudder

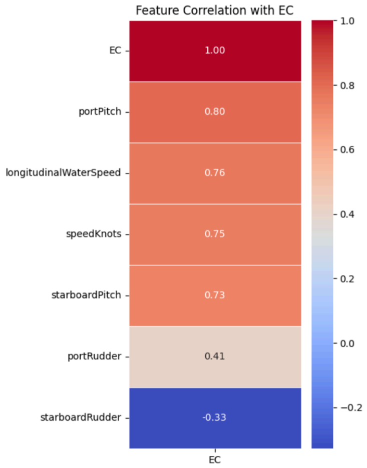


Figure 5: Correlation between chosen features and the target value EC

3 Splitting the Data for Testing and Training

When creating training, validation, and test sets, a choice between using a randomized split and a date-based split was made. Instead of splitting the data set into different portions based on never-before-seen dates in the data set, randomization was chosen.

As the training, testing, and validation sets do not include the timestamps or datetime features, but predict based on sensor data in an isolated moment, randomization was chosen.

Set sizes:

- Training: 70%, 1139075
- Validation: 15%, 244088
- Testing: 15%, 244088

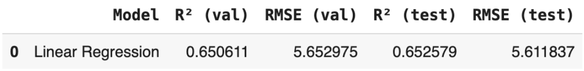
4 Model Training

For this paper, three models were chosen for comparison: Linear Regression, Polynomial Regression, and Random Forest Regressor. All models were trained with the seed 42.

4.1 Linear Regression

A linear regression analysis was conducted. Linear Regression is a popular machine learning algorithm for regression problems.

At the beginning of the analysis, the sets were scaled for better model predictions.



	Model	R² (val)	RMSE (val)	R² (test)	RMSE (test)
0	Linear Regression	0.650611	5.652975	0.652579	5.611837

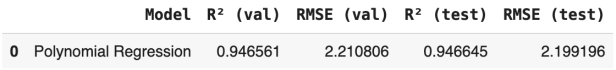
Figure 6: Model performance of Linear Regression

As we can see, the linear model performs only moderately, suggesting there might be non-linear components in the data.

4.2 Polynomial Regression

Polynomial regression was chosen as the second model to address the suspected non-linearity of the data. Polynomial Regression differs from Linear Regression by adding a degree parameter, which in Linear Regression can be seen as a degree of 1. For this analysis, the degree used was 2.

The features were scaled at the beginning of the analysis for better model performance.



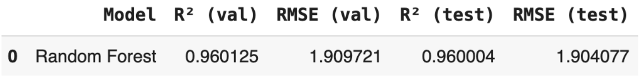
	Model	R^2 (val)	RMSE (val)	R^2 (test)	RMSE (test)
0	Polynomial Regression	0.946561	2.210806	0.946645	2.199196

Figure 7: Model performance of Polynomial Regression

Polynomial Regression performed well on both the validation and test sets, so no adjustments were made between them. The accuracy score R-square is over 0.94 for both the validation and test sets.

4.3 Random Forest Regressor

Random Forest is a decision tree-based model well suited for regression and classification tasks. For this task, the maximum tree depth was 10. Restricting the tree to a maximum depth reduces the risk of overfitting.



	Model	R^2 (val)	RMSE (val)	R^2 (test)	RMSE (test)
0	Random Forest	0.960125	1.909721	0.960004	1.904077

Figure 7: Model performance of Random Forest Regression

Of the three models, Random Forest performed the best, with an average score of 0.96. The polynomial regression performed so well on the validation test set that no additional tweaks were made before running it.

5 Conclusions

This paper explored the prediction of the estimated fuel consumption in tons/day based on sensor readings from the different sensors on MS Smyril. The goal was to train models to predict the estimated fuel consumption in tons/day at a given time.

At the beginning of the analysis, the data sets were collected from various sensor systems of MS Smyril. As all the sensor data was not measured at the same timestamps, the data was merged based on the timestamps in the fuel-related data sets. This ensured that no timestamp had NaN values, as the merge function always picked the closest match for each row in the fuel-based time logs. Longitudinal and latitudinal data sets were discarded in the beginning, as they are not numerical or easily categorified.

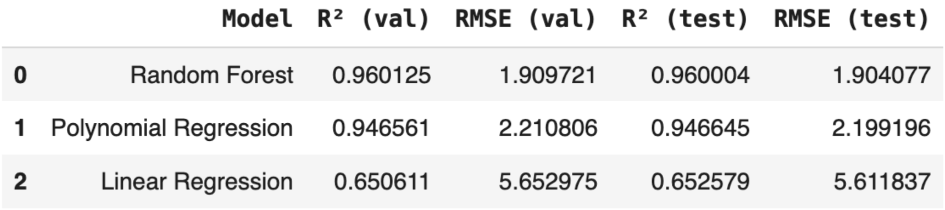
The timestamps were converted to datetime format to perform a visual data analysis of the ship’s voyages and median estimated daily fuel consumption.

The following sensor logs were chosen for the analysis:

- longitudinalWaterSpeed
- portPitch
- portRudder
- speedKnots
- starboardPitch
- starboardRudder

As the fuel-related data sets were used to calculate the target value, no fuel data sets were used as features.

Test sets were split randomly with 70% training data, 15% validation data, and 15% testing data.



	Model	R² (val)	RMSE (val)	R² (test)	RMSE (test)
0	Random Forest	0.960125	1.909721	0.960004	1.904077
1	Polynomial Regression	0.946561	2.210806	0.946645	2.199196
2	Linear Regression	0.650611	5.652975	0.652579	5.611837

Figure 8: Model Performance Comparisons

Linear Regression, Polynomial Regression, and Random Forest Regression were compared to assess their effectiveness in modeling the dataset. As anticipated, Linear Regression did not adequately capture the complexity of the data due to its inherent assumption of linearity. Consequently, Polynomial Regression was selected as it is better suited for modeling non-linear relationships.

Polynomial Regression and Random Forest Regression demonstrated strong predictive performance at the validation stage, eliminating the need for extensive hyperparameter tuning. Each model achieved an R-squared value of approximately 0.95, indicating a high degree of explanatory power and suggesting that both approaches effectively captured the underlying patterns in the data.

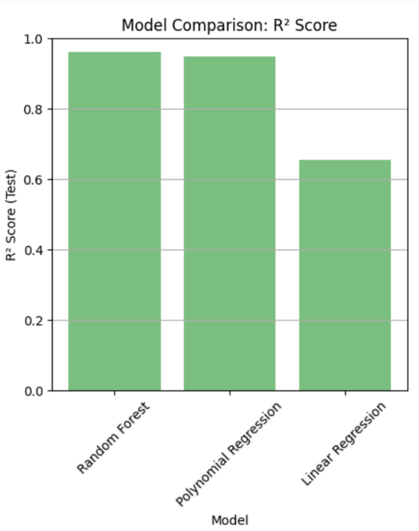


Figure 9: Model Performance Comparisons

The data set that consists of sensor data from MS Smyril is versatile and offers a lot of different angles to explore. Furthermore, the effects of weather sensors on fuel consumption could be explored further to estimate best practices to minimize fuel consumption to make an even greater economic and environmental impact.