UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Master Thesis

on the topic of

Modelling and optimization of ship's fuel consumption using Random Forest Regression (RFR)

Submitted to the Faculty of Engineering
of University Duisburg Essen

by

Hibatul Wafi
3021919

Betreuer:	M. T. Muhammad Fakhruriza Pradana
1. Gutachter:	Prof. Dr.-Ing. B. Noche
2. Gutachter:	Prof. Dr. Ucker
Studiengang:	ISE General Mechanical Engineering
Studiensemester:	Summer semester 2023
Datum:	04.05.2023

Contents

List of Tables

List of Figures

1 Introduction

The research on efficient ship operation is a direction that is being actively pursued by marine industry stakeholders as efficient ship operations equates to increase in profitability. One of the determining factors is the reduction of Fuel Oil Consumption (FOC). FOC takes up considerable portion in ship's operating cost. This is clearly indicated through findings made by Ronen [1] and Stopford [2]. The former mentioned that FOC consumption of a large ship potentially constitute to 75% of the total ship operating cost while the latter noted that FOC makes up to two-thirds of vessel voyage cost and over one-quarter of vessel's overall cost.

With that, maritime industry stakeholder actively searches for inexpensive approach to reduce FOC. As such, they investigate ways to optimise operational measures as technical solutions are expensive [3]. The operational measures include the inclusion of weather/environmental routing, speed optimisation, trim optimisation and virtual (just-in/time) arrival policy [3]. It is noted by Beşikçi et al. [4] that lowering ship speed will have the greatest impact in fuel economy, reducing the ship speed by $2-3knots$ could halve the operating cost of shipping company [2, 5]. Beşikçi et al. further elaborated that the main cause of this is the nonlinear relationship between ship speed and fuel consumption. Ronen [1, 6] and Wang et al. [7] approximated that fuel consumption can be derived through third order function of the ship speed.

Due to volatility and ever-increasing bunker fuel price, developing a model that could accurately predict ship speed would be beneficial to forecast the ship's FOC. The model could potentially help maritime industry stakeholder make decisions at the most opportune moment. Data driven i.e., machine learning approaches have been attempted by several authors in different literatures to model fuel consumption and reported good results in its predictive performance [4,8–11]. However, powerful machine learning models are usually unintuitive making it difficult to interpret its decisions [12]. This brings us to Random Forest, a powerful model that offers partial interpretability in their decisions. With this consideration, modelling using Random Forest will be the focus of this thesis.

1.1 Research Objective, Contributions and Boundary

This thesis aims to predict the ship's speed captured by Automatic Identification System (AIS) using random forest model. In this study, this speed shall be designated as the ship's Speed Over Ground (SOG). The modelling uses fused hourly data from AIS information of Hammershus Ro-Ro ferry and local meteorological weather data in region of travel. Subsequently, the ship actual speed, which is designated as speed through water (STW), shall be derived from the predicted SOG to enable estimation for fuel consumption over different journey periods. The modelling is performed in `Python` programming language using machine learning packages `sklearn` offered by Pedregosa et al. [13].

Using this approach, we shall raise the following research questions (RQs), namely:

- **RQ1**. Is it feasible to fuse AIS data and meteorological data to accurately predict the ship's SOG ?
- **RQ2**. During modelling, which parameters have the greatest impact in increasing the model's predictive performance ?
- **RQ3**. During evaluation, what are the performance measures that should be considered to help us gain the most information out of the model's behaviour ?

To answer the research questions, the following research boundaries are set:

- Random forest has the capability to solve both classification and regression problem. Because the target variable, SOG, is continuous, we will only adopt the regression algorithm of random forest.
- The focus of this work is a detailed study on the performance and possible optimisation configuration of random forest as predictor for the target variable. As such, we will not perform exhaustive comparison study between different machine learning models.
- The estimation for fuel consumption shall be done using simple formulation by Ronen [1, 6]. This thesis will not consider the more comprehensive method such as the method proposed by Kim et al. [14] as the focus of this work is to estimate the SOG using random forest-
- The Hammershus Ro-Ro ship sails between port of Køge, Rønne, Ystad and Sassnitz. However, we will only consider the journey between port of Køge, Rønne and Ystad as part of the data for the voyage between port of Rønne and Sassnitz are missing.

The use of AIS data provides the following contributions as indicated by Rakke [15]:

- Avoid expenses of purchasing (possibly) unaffordable ship information from online database and shipping companies.
- Independent of commercial parties, as information are available in public domain.

Additionally, this work will provide the following contribution:

- Robust modelling approach that requires minimal data pre-processing and minimal model configuration.

2 Theoretical Background

This chapter deals with the past and present research in the relevant area which include literature review. This includes the significance of precise modelling of the ship's speed and its subsequent use in forecasting the ship's operation. The theoretical background of Random Forest Regression will be discussed in this chapter.

2.1 Literature Review

The work by Yan et al. [16] provides a thorough review of the different attempts that have been made by different authors to predict different parameters of ship's operation, this includes ship's fuel consumption. Per definition by Haranen et al. [17], the modelling of ship operation is categorised into White Box Model (WBM), Black Box Model (BBM) and Grey Box Model (GBM). Machine learning approach is categorised as BBM, BBM approach is defined as purely data driven approach requiring no prior knowledge about the ship operation. The literature review by Yan et al. [16] indicated that about 42% of the research utlised BBM model based on machine learning approach.

Majority of the BBM approach based on ML is dominated by ANN [16]. However, there are works that considered decision tree-based modelling approach such as Decision Tree (DT), Random Forest (RF) and Extra Tree (ET). Soner et al. [18] implemented tree-based model, which include bagging, random forest (RF), and bootstrap, using data captured from onboard sensors of a ferry to predict speed through water and fuel consumption per hour. From the test dataset, the random forest model reported root mean square error (RMSE) of 0.34 Knots during its prediction of Speed Through Water (STW). Yan et al. [19] used random forest (RF) model to minimise fuel consumption for a voyage of a dry bulk ship. The model use ship operational data and sea and weather data from noon report and EMCWF. The prediction performance report from this literature revealed mean absolute percentage error (MAPE) of 7.91%.

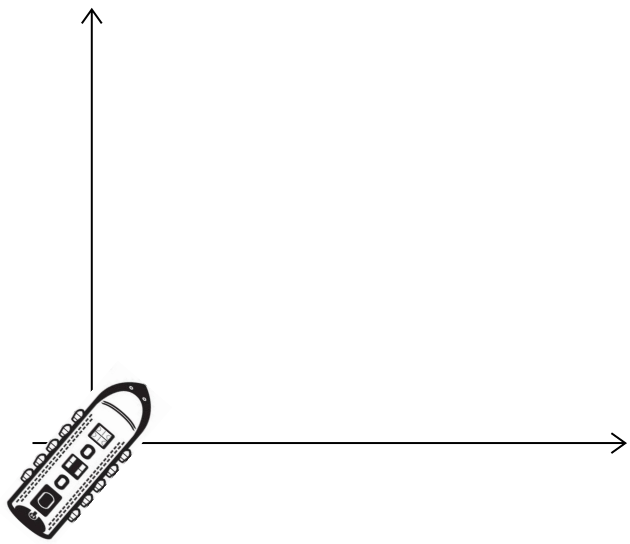
The research by Gkerekos et al. [9] highlighted the performance of different machine learning models to predict ship's fuel consumption per day using both noon data and automated data logging and monitoring (ADLM) system from a bulk carrier. This research concludes that tree based model displayed good prediction performance on both noon data and sensor based data. Using default parameters, RF model obtained R^2 score of 87.55% and 96.26% for noon data datasets and sensor-based data respectively. It is also noted that it that the data from a 3-month period in ADLM system would be sufficient to create a model with better performance than the model generated by noon data from a collection period of 2.5 year. This literature also concluded that automatic sensor-based data have the potential to increase the model accuracy score, R^2, by 5 – 7%.

Li et al. [3] performed more extensive research on the effects of data fusions between meteorological data, ship voyage data and AIS data on different machine learning models to predict the ship's FOC. This research highlighted the advantage of fusing meteorological data and ship voyage data. The evaluation on different model performance indicated that RF are among preferable model candidate that could be used in commercial scale due to its good prediction capability and robustness against different datasets. The model performance comparison in their work reported R^2 score are above 96% when deployed on the best datasets and achieved R^2 score in range between 74% – 90% over test data. The findings from their work also indicated the robustness of RF, as it displayed the lowest standard deviation at 0.015 of the R^2 score when evaluated against random splits of datasets.

Abebe et al. [10] used different approach in their research by predicting the ship's Speed Over Ground (SOG) instead of FOC. In this work, AIS data and noon-report weather data from 14 tracks and 62 ships are used for the SOG prediction. This literature reported that for random forest regressor (RFR), It achieved root mean squared error RMSE of 0.25 knots using 489 seconds for training. While decision tree achieved RMSE of 0.36 knots taking up 52 seconds for training. This shows that RFR outperforms DTR at cost of computational power.

This literature review indicated the capability of Random Forest Regressor to predict fuel consumptions and ship speed, irrespective of data source and type of data used. Promising results from different performance measures across different literatures indicated the capability random forest model as predictor. This brings us to RQ2 described in chapter 1.1: how can we extract maximum prediction performance from random forest. Due to the nonlinear, third order function estimate of fuel consumption described in chapter 1 by Ronen [6]. Accurate prediction of ship speed is necessary to ensure optimal operation and increase profitability.

2.2 Random Forest Regressor (RFR)



Suppose the following decision tree regressor

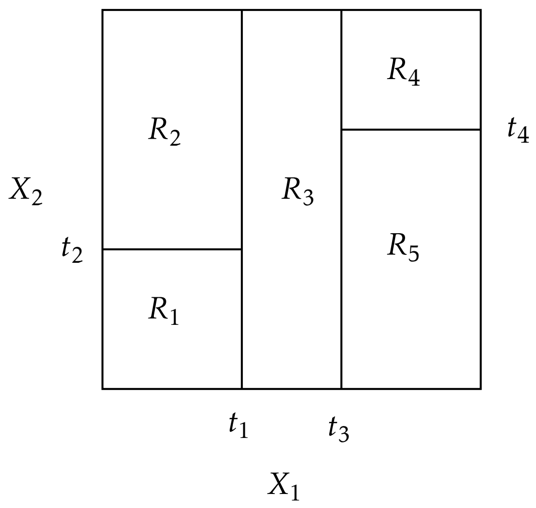


Figure 1: Example of partition space

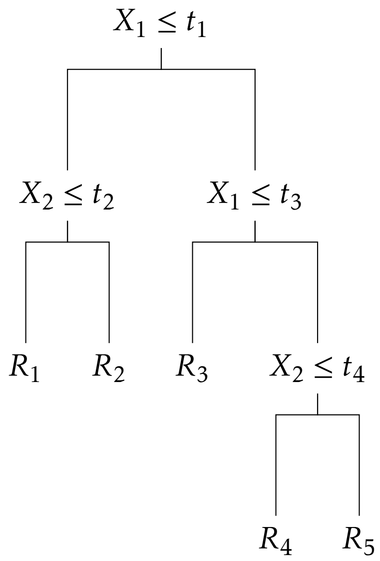


Figure 2: Example of partition tree

2.3 Ship speed

2.4 Modelling

3 Research Methodology

In this chapter the methodology used to develop the model will be discussed. The discussion on different parameters in the vessel's journey data will be discussed here. This includes the mining and merging of the features. The method used to develop the ship's speed model will be discussed in this chapter. This consists of the parameter used to develop the model. Ultimately, the model is then used to predict the ship's fuel consumption.

3.1 Data Preprocessing

- Two data sources are imported. `AIS_weather_H_ok2_copy.csv` and `AIS_weather_h_rename_copy.csv`. The information from the latter comma delimited file will be used for calculating the ship Speed Through Water (STW). The information required is the true north current direction. Which is obtained from the vector component of the Northward and Southward current.
- This dataframe will be merged with the main dataframe from the file `AIS_weather_H_ok2_copy.csv`.
- Omission of the journey data between Ronne and Sassnitz
- SOG threshold is applied to omit ship mooring and maneuvering to accurately represent the ship's steady state operation [4, 9, 10, 20]. This threshold is selected as 5 knots according to [10]
- The AIS data from June is filtered. This data will be used as validation data to check the model's performance.

3.2 Data Analysis

- The features are represented in a histogram plot. For the feature Current speed, anomaly is detected. Certain spike is detected around $0.01 - 0.03$ m/s. Reasons unknown. The data is retained, including the spike, until a definitive answer can be found.
- OPEN QUESTION : What is the necessity of feature standardization / normalization ? Normalization is required for ANN as model training requires the value between 0 and 1. But in case of RFR, there is no such requirement. Through testing, data standardization also does not seem to improve the model's performance.

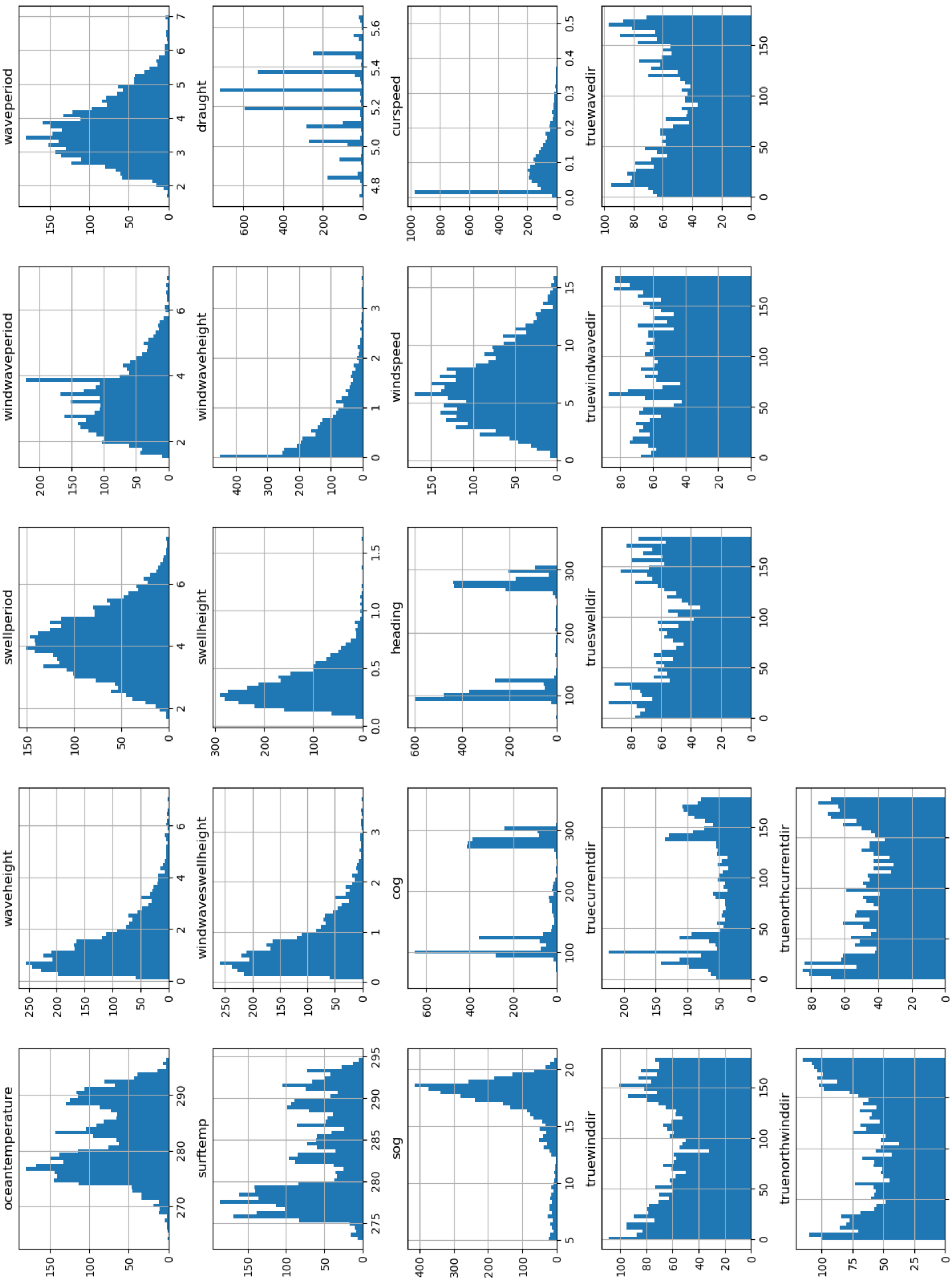


Figure 3: Histogram of the features

- The correlation of the features against SOG are determined. It is found that :
  - Draught
  - Course Over Ground (COG)
  - heading
  - Wind Speed
  - Current Speed
  - True Current direction

  Have relatively stronger correlation to SOG compared to other features, albeit the correlation is a weak one

- The correlation between the features is displayed using the following the heat map. From the heat map it can be observed that between these features:
  - Waveheight and wind wave swell height
  - Waveheight and wind wave height
  - Windwaveswellheight and wave period

  Have a strong correlation between each other.

- Open topic:
  - Feature reduction is possible, [10] suggested high feature correlation filter, the filter suggest that two features which has a high correlation ($>$ 90%) is to be combined into a single feature. But the author is unsure whether this combination is physically sensible. Hence, this filter is yet to be applied for feature reduction.
  - Some of these features can be connected through wave equations, but the author has not found an equation which could relate these features.

- The random forest regressor could not function when `NaN` values are present. With that, the missing values are filled in using the `imputer` function. The missing values are filled in by means of `KNN`.

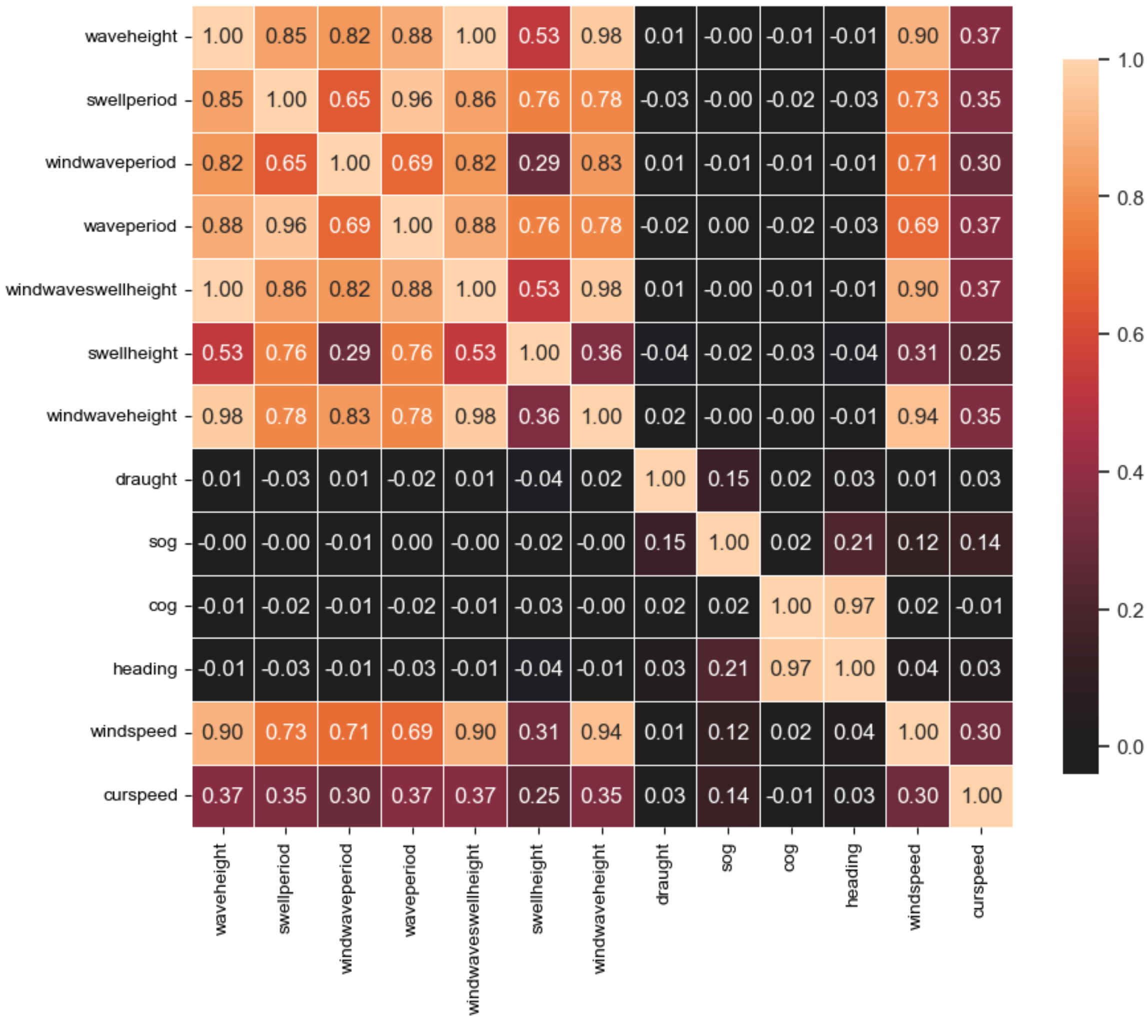


Figure 4: Correlation Heat Map

3.3 Modelling

- The data is split into 80:20 ratio. But considering the validation data, it is split into approximately 73:18:9.

- The model is then trained using Random Forest Regression (RFR). Additional training is also performed using Decision Tree Regressor (DTR). DTR model performance will be used as a benchmark as it is also a tree-based modelling method with similar methodology to RFR.

- The computational time of DTR is significantly faster than RFR Model Evaluation

3.4 Predicting STW

- The ship's Speed Through Water STW can be calculated using vector component of the SOG and current speed. The direction used will be according to True North. [20, 21]

- SOG represents the speed of the ship with reference to the ground, while the STW represent the ship's speed with reference to water.

- SOG also can be termed by the ship's speed that is captured by the GPS, and does not consider any effect of the current

- This means that the ship's STW will be greater than the ship's SOG when there is current moving against the ship's movement direction and vice versa

- The vector decomposition can be defined from the following equations, which is based on the equation by [20]:

  – The ship's SOG V_g can be decomposed into V_g^x and V_g^y, which represents the x and y components of the SOG respectively using the ship's course heading (COG) β *with respect to True North*:

  $$V_g^x = V_g \sin(\beta) \quad (1)$$

  $$V_g^y = V_g \cos(\beta) \quad (2)$$

  – To consider the effect of sea current. The current speed V_c will also be decomposed to x and y components respectively using the current direction γ *with respect to True North*:

  $$V_c^x = V_g \sin(\gamma) \quad (3)$$

  $$V_c^y = V_g \cos(\gamma) \quad (4)$$

  – from here the ship' STW V_{wx} and V_{wy} component can be found from the following equation:

  $$V_w^x = V_g^x - V_c^x \quad (5)$$

  $$V_w^y = V_g^y - V_c^y \quad (6)$$

  – The magnitude of the STW can be readily obtained from the following vector synthesis

  $$V_w = \sqrt{(V_w^x)^2 + (V_w^y)^2} \quad (7)$$

- This principle is applied to the following Python script. 3

```python

        # Convert SOG from [Knots] to [m/s]

        dfprog["vgms"] = dfprog["sog_pred"]/1.9438

        # Convert the angles from [Degrees] to [Radians]

        rad_gamma = np.deg2rad(dfprog["gamma"])
        rad_cog = np.deg2rad(dfprog["cog"])

        # Decomposition in x-component

        dfprog["vgx"] = dfprog["vgms"] * np.sin(rad_cog)
        dfprog["vcx"] = dfprog["curspeed"] * np.sin(rad_gamma)
        dfprog["stw_x"] = (dfprog["vgx"] - dfprog["vcx"])

        # Decomposition in y-component

        dfprog["vgy"] = dfprog["vgms"] * np.cos(rad_cog)
        dfprog["vcy"] = dfprog["curspeed"] * np.cos(rad_gamma)
        dfprog["stw_y"] = (dfprog["vgy"] - dfprog["vcy"])

        # Vector synthesis and reconversion to [Knots] from [m/s]

        dfprog["vwms_p"] = np.sqrt(dfprog["stw_x"]**2 + dfprog["stw_y"]**2)
        dfprog["stw_pred"] = dfprog["vwms_p"]*1.9438

```

4 Result and Discussion

The result of the research is discussed in this chapter. This comprises model validation and how different statistical metrics are used to analyze the model's performance.

4.1 Model Evaluation

The model are tested against four metrics, namely:

- R^2 : Indicate model fit. Best Score = 1
- Explained Variance `EV` : Indicate amount of variance in model. Best Score = 1
- Mean Absolute Error `MAE` : Indicate how much error a model makes in its prediction. Best Score = 0
- Root Mean Square Error `RMSE` : Same as MAE, more sensitive to outlier. Best Score = 0
- Median Absolute Error `MAD` : Check robustness against outlier. Best Score = 1

The result is summarized in the following table

Model	RFR	DTR	LR
R^2	0.9328181446941499	0.8526085810220092	1
EV	0.932872958708872	0.8526260247615258	2
MAE	0.5546347329650284	0.8108982427834758	3
RMSE	0.7095480848510665	1.5566896535262504	4
MAD	0.38484635910000087	0.5475717149999983	5

Table 1: Model performance

Model	RFR	DTR	LR
R^2	0.9328181446941499	0.8526085810220092	1
EV	0.932872958708872	0.8526260247615258	2
MAE	0.5546347329650284	0.8108982427834758	3
RMSE	0.7095480848510665	1.5566896535262504	4
MAD	0.38484635910000087	0.5475717149999983	5

Table 2: Model performance

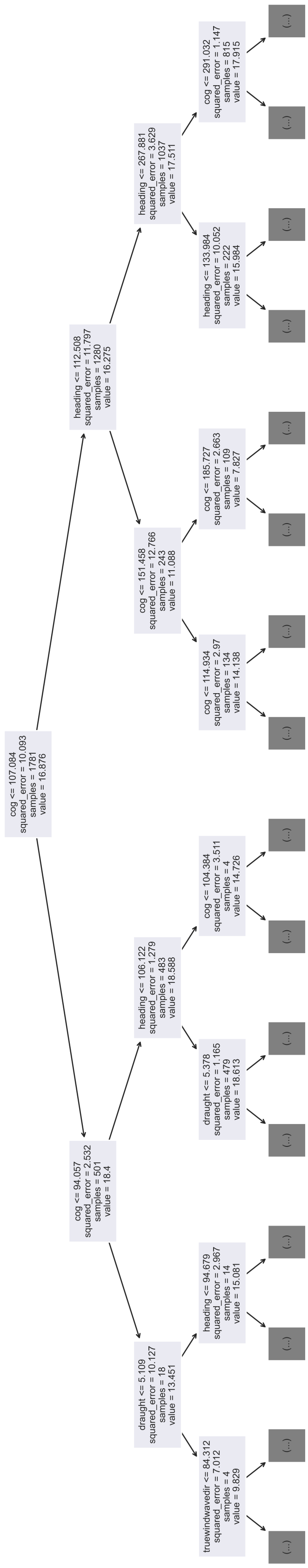


Figure 5: Correlation Heat Map

5 Summary and Outlook

In this chapter the summary of this research will be discussed. This section includes reflections of the research process and presents any possible suggestions and recommendations in this line of research. This chapter concludes this thesis.

References

[1] D. Ronen. The effect of oil price on containership speed and fleet size. *Journal of the Operational Research Society*, 62(1):211–216, 2011. doi:10.1057/jors.2009.169. 5, 6

[2] Stopford. The organization of the shipping market. page 47, 2009. 5

[3] Xiaohe Li, Yuquan Du, Yanyu Chen, Son Nguyen, Wei Zhang, Alessandro Schönborn, and Zhuo Sun. Data fusion and machine learning for ship fuel efficiency modeling: Part i – voyage report data and meteorological data. *Communications in Transportation Research*, 2:100074, 2022. doi:10.1016/j.commtr.2022.100074. 5, 7

[4] E. Bal Beşikçi, O. Arslan, O. Turan, and A. I. Ölçer. An artificial neural network based decision support system for energy efficient ship operations. *Computers & Operations Research*, 66:393–401, 2016. doi:10.1016/j.cor.2015.04.004. 5, 10

[5] N. Wijnolst, Tor Wergeland, and Kai Levander. *Shipping Innovation*. IOS Press, 2009. 5

[6] David Ronen. The effect of oil price on the optimal speed of ships. *The Journal of the Operational Research Society*, 33(11):1035, 1982. doi:10.2307/2581518. 5, 6, 8

[7] Shuaian Wang and Qiang Meng. Sailing speed optimization for container ships in a liner shipping network. *Transportation Research Part E: Logistics and Transportation Review*, 48(3):701–714, 2012. URL: https://www.sciencedirect.com/science/article/pii/S1366554511001554, doi:10.1016/j.tre.2011.12.003. 5

[8] Miyeon Jeon, Yoojeong Noh, Yongwoo Shin, O-Kaung Lim, Inwon Lee, and Daeseung Cho. Prediction of ship fuel consumption by using an artificial neural network. *Journal of Mechanical Science and Technology*, 32(12):5785–5796, 2018. URL: https://link.springer.com/article/10.1007/s12206-018-1126-4, doi:10.1007/s12206-018-1126-4. 5

[9] Christos Gkerekos, Iraklis Lazakis, and Gerasimos Theotokatos. Machine learning models for predicting ship main engine fuel oil consumption: A comparative study. *Ocean Engineering*, 188:106282, 2019. doi:10.1016/j.oceaneng.2019.106282. 5, 7, 10

[10] Misganaw Abebe, Yongwoo Shin, Yoojeong Noh, Sangbong Lee, and Inwon Lee. Machine learning approaches for ship speed prediction towards energy efficient shipping. *Applied Sciences*, 10(7):2325, 2020. doi:10.3390/app10072325. 5, 8, 10, 12

[11] Young-Rong Kim, Min Jung, and Jun-Bum Park. Development of a fuel consumption prediction model based on machine learning using ship in-service data. *Journal of Marine Science and Engineering*, 9(2):137, 2021. doi:10.3390/jmse9020137. 5

[12] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems / Aurélien Géron*. O'Reilly, Sebastopol, CA, second edition edition, 2019. 5

[13] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL: http://jmlr.org/papers/v12/pedregosa11a.html. 5

[14] Ki-Su Kim and Myung-Il Roh. Iso 15016:2015-based method for estimating the fuel oil consumption of a ship. *Journal of Marine Science and Engineering*, 8(10):791, 2020. doi:10.3390/jmse8100791. 6

[15] Stian Glomvik Rakke. *Ship emissions calculation from AIS*. PhD thesis, NTNU. URL: https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2410741. 6

[16] Ran Yan, Shuaian Wang, and Harilaos N. Psaraftis. Data analytics for fuel consumption management in maritime transportation: Status and perspectives. *Transportation Research Part E: Logistics and Transportation Review*, 155:102489, 2021. URL: https://www.sciencedirect.com/science/article/pii/S1366554521002519, doi:10.1016/j.tre.2021.102489. 7

[17] Michael Haranen, Pekka Pakkanen, Risto Kariranta, and Jouni Salo. White, grey and black-box modelling in ship performance evaluation. 2016. URL: https://www.researchgate.net/publication/301355727_White_Grey_and_Black-Box_Modelling_in_Ship_Performance_Evaluation. 7

[18] Omer Soner, Emre Akyuz, and Metin Celik. Use of tree based methods in ship performance monitoring under operating conditions. *Ocean Engineering*, 166:302–310, 2018. URL: https://www.sciencedirect.com/science/article/pii/S0029801818314446, doi:10.1016/j.oceaneng.2018.07.061. 7

[19] Ran Yan, Shuaian Wang, and Yuquan Du. Development of a two-stage ship fuel consumption prediction and reduction model for a dry bulk ship. *Transportation Research Part E: Logistics and Transportation Review*, 138:101930, 2020. doi:10.1016/j.tre.2020.101930. 7

[20] Liqian Yang, Gang Chen, Jinlou Zhao, and Niels Gorm Malý Rytter. Ship speed optimization considering ocean currents to enhance environmental sustainability in maritime shipping. *Sustainability*, 12(9):3649, 2020. doi:10.3390/su12093649. 10, 14

[21] Yang Zhou, Winnie Daamen, Tiedo Vellinga, and Serge P. Hoogendoorn. Impacts of wind and current on ship behavior in ports and waterways: A quantitative analysis based on ais data. *Ocean Engineering*, 213:107774, 2020. doi:10.1016/j.oceaneng.2020.107774. 14

Declaration in lieu of oath

I hereby solemnly declare that I have independently completed this work or, in the case of group work, the part of the work that I have marked accordingly. I have not made use of the unauthorised assistance of third parties. Furthermore, I have used only the stated sources or aids and I have referenced all statements (particularly quotations) that I have adopted from the sources I have used verbatim or in essence.

I declare that the version of the work I have submitted in digital form is identical to the printed copies submitted.

I am aware that, in the case of an examination offence, the relevant assessment will be marked as 'insufficient' (5.0). In addition, an examination offence may be punishable as an administrative offence (Ordnungswidrigkeit) with a fine of up to €50,000. In cases of multiple or otherwise serious examination offences, I may also be removed from the register of students.

I am aware that the examiner and/or the Examination Board may use relevant software or other electronic aids in order to establish an examination offence has occurred

I solemnly declare that I have made the previous statements to the best of my knowledge and belief and that these statements are true and I have not concealed anything.

I am aware of the potential punishments for a false declaration in lieu of oath and in particular of the penalties set out in Sections 156 and 161 of the German Criminal Code (Strafgesetzbuch; StGB), which I have been specifically referred to.

Section 156 False declaration in lieu of an oath

Whoever falsely makes a declaration in lieu of an oath before an authority which is competent to administer such declarations or falsely testifies whilst referring to such a declaration incurs a penalty of imprisonment for a term not exceeding three years or a fine.

Section 161 Negligent false oath; negligent false declaration in lieu of oath

(1) Whoever commits one of the offences referred to in Sections 154 to 156 by negligence incurs a penalty of imprisonment for a term not exceeding one year or a fine. (2) No penalty is incurred if the offender corrects the false statement in time. The provisions of Section 158 (2) and (3) apply accordingly.

Place,date

Signature