Using Machine Learning to Model Yacht Performance

MARINE 2021

C. Byrne[1,2,*], T. Dickson[2], M. Lauber[2], C. Cairoli[1] and G. Weymouth [2,3]

[1] BAR Technologies, The Camber, East Street, PO1 2JJ Portsmouth, United Kingdom. E-mail: cian.byrne@bartechnologies.uk, web page: http://www.bartechnologies.uk

[2] University of Southampton, University Rd, SO17 1BJ Southampton, United Kingdom. E-mail: enquiries@southampton.ac.uk , web: http://www.southampton.ac.uk

[3] Alan Turing Institute, Euston Rd, NW1 2DB London, United Kingdom. E-mail: research@turing.ac.uk, web: http://www.turing.ac.uk

[*] Corresponding author: Cian Byrne, cian.byrne@bartechnologies.uk

ABSTRACT

Accurate modelling of the performance of a yacht in varying environmental conditions can significantly improve a yachts performance. However, a racing yacht is a highly complex multi-physics system meaning that real-time performance prediction tools are always semi-empirical, leaving significant room for improvement. In this paper we first use unsupervised machine learning to analyse full-scale yacht performance data. The widely documented ORC VPP (ORC, 2015) and the commercial Windesign VPP are compared to the data across a range of wind conditions. The data is then used to train machine learning models. A number of machine learning regression algorithms are explored including Neural Networks, Random Forests and Support Vector Machines and improvements of 82% are obtained compared to the commercial tools. The use of physics based learning models (Weymouth and Yue, 2013) is explored in order to reduce the amount of data required to achieve accurate predictions. It is found that machine learning models can outperform empirical models even when predicting performance in environmental conditions that have not been supplied to the model as part of the training dataset.

Keywords: Machine Learning; Unsupervised Learning; Neural Network; Random Forest.

NOMENCLATURE

Bsp	Boat speed [knots]
Twa	True wind angle [deg]
Tws	True wind speed [knots]
NN	Neural Network
CFD	Computational Fluid Dynamics
VPP	Velocity Prediction Progrmam
RF	Random Forest
IM	Intermediate model
GLM	Generic Learning Model
ORC	Offshore Racing Congress

1 INTRODUCTION

Accurate predictions of the performance of a racing yacht is extremely beneficial to both amateur and professional crews alike. When a yacht is performing worse than expected changes can be made to the sail settings in order to increase performance. Similarly if a yacht is performing better than expected the setup can be recorded and predictions updated. It becomes increasingly hard to improve a crews performance if there is a low level of accuracy in the performance prediction tools available.

Currently there are a number of different methods which are used to predict and simulate the performance of a sailing yacht. Classic Velocity Prediction Programs (VPP) use a mixture of theoretical analysis and empirical data to estimate the motion of a yacht. This type of method is now fast to implement but suffers from inaccuracies. More modern VPP's incorporate Computational Fluid Dynamics (CFD) to provide high fidelity data for a specific yacht that a VPP can then use to predict performance. This method is useful but performing sufficient CFD calculations to obtain accurate results from a VPP is extremely computationally expensive (Böhm, 2014) and thus is beyond the capabilities of most amateur racing teams.

There are two main approaches in the applications of VPP's. A quasi-static approach views a yacht at a snapshot in time. The forces acting on the yacht are estimated in order to establish if the yacht is in equilibrium, if not, the underlying assumptions such as boat speed or heel angle are changed until an equilibrium is reached. A dynamic approach allows the yacht to accelerate in all directions. This approach involves using time dependent equations to model the yachts behaviour.

New methods which utilise data are needed in order to improve on these issues associated with the current methods. Advances in machine learning and data gathering capabilities provide a unique opportunity to analyse and improve on existing models. Existing models can be extremely complex in nature as a by-product of introducing many different parameters in order to fit the model to what is observed in reality. The use of data should allow for a reduction in the number of parameters required in a model. A comparison will be undertaken between physics based models, fully data based models and a mixture of each.

2 DATA CLASSIFICATION

Data that is typically available from real world racing yachts contains a lot of noise. Typical data files provide no metadata which would identify when a yacht is racing, motoring to/from the race course or simply sailing with non optimal sail trim. It is obvious that comparing a VPP result to data describing a yacht motoring is pointless. Moreover, it is essential that the data which we use for analysis accurately represents the performance of the yacht in question by excluding data when the boat is not in "race-mode". If sails are not trimmed optimally and/or crew positioning is not optimal then the comparison of physics based models to the real world data is not valid. It is assumed that when the yacht is in "race-mode" the yacht will be raced optimally i.e. the crew extracts maximum performance from the yacht. This assumption can only be made as the available data set is from a yacht competing in a professional level regatta. T skill level amongst a crew competing at such a high level regatta is extremely high. If the crew were not of a high level then it would be entirely possible that the skill factor would contribute to a higher performance prediction from the physics based models when compared to data of a yacht being sailed sub-optimally.

2.1 Data

The data that forms the basis of this paper was gathered on a TP52 class yacht *"Spookie3"* during the 2015 TP52 Super Series Regatta in Cascais Portugal. This data is freely available from the Sailing Yacht Research Foundation online library (Benjamin, 2015).

The data consists of log files from 8 days of competitive sailing containing a total of 193,727 data points sampled at a frequency of 1Hz. The files consist of a total of 42 data fields including Boat Speed (Bsp), True Wind Angle (Twa), True Wind Speed (Tws), etc.

In order to extract racing data a number of methods will be discussed. The available data was labeled manually after rigorous analysis by the author. The data was classified as either Upwind (UW), Downwind (DW) or Not Racing (NR). This labelled data will be used in a supervised learning approach to both train a supervised learning model and test the accuracy of such a model. The labelled data can also be used to assess the ability of unsupervised learning models to separate the data into correct categories.

2.2 Supervised Classification

Traditionally for supervised classification tasks a fully labelled data set is split into two different sets of data, namely training data and test data. The model will be given both the features and labels of the training set in order to learn the appropriate classification model. The feature values in the test data set are then provided to this learned model which in turn predicts the labels that should be associated with each data point. These predictions can then be compared with the actual test labels that were not provided to the model in order to establish the accuracy of the learned model.

In typical machine learning applications test data usually consists of 20-30% of the full dataset. The train-test split is performed by randomly selecting points to be included in the test data set. However, in the case of classifying yacht racing data it does not make practical sense to apply this exact approach. This is because in practice data will be manually labelled on a day to day basis. In other words we would have access to labelled data from say, three days of data files and we would wish to use this data to classify a fourth day of data files. Therefore the accuracy of this type of classification method will be highly dependant on which days are used as training data and which days are used as test data. Test days that have similar conditions to the conditions experienced in the corresponding training data sets will perform better than test days in different conditions to those supplied in the training data. In order to avoid this dependence we will look at using every possible combination of test/training days for varying amounts of training data (i.e. 1 day training data/7 days test data, 2 days training data/5 days test data).

Two main supervised classification models are explored namely Random Forests (Breiman, 2001) and Support Vector Machine classifiers (Cortes and Vapnik, 1995).

The feature sets that were used in all the supervised classification algorithms were:

- Standard Features - [Bsp, abs(Awa), abs(Heel), Tws, abs(Leeway)]
- Enhanced Features - [Bsp, abs(Awa), abs(Heel), Tws, abs(Leeway), Forestay, Bsp/Aws, abs(Awa)*Bsp, Tws/abs(Heel)]

2.3 Unsupervised Classification

The labeling process required in order to perform supervised classification can be very time consuming. Unsupervised learning may be used in order to avoid this task. Unlike supervised learning models the desired output labels are not be supplied to an unsupervised model. In the case of yacht data classification this means that no manual labelling of the data is required and that an unsupervised model can learn to classify the data from only inputs.

Two of the most widely used unsupervised classification methods are K-means clustering (Forgy, 1965) and hierarchical clustering (Rokach and Maimon, 2005).

Both the k-means and hierarchical clustering models are most often used to cluster individual data points into respective clusters based on their similarity to other points within their cluster. However, classifying each data-point separately fails to take advantage of the temporal nature of the yacht data. In yacht racing a yacht tends to stay on an upwind course for a significant period of time before turning onto another leg and remaining on that leg for another period of time. Thus, classifying a window of data may be more beneficial than classifying individual points. This makes intuitive sense as two windows of data each describing a yacht sailing on an upwind leg will have similar values for features such as Twa, Heel, Bsp, etc. and should then be classified in the same cluster by the clustering algorithm. This process of windowing the data also helps to filter the effect of noise in the data collection process.

In reality more than one cluster will describe upwind sailing. This is the case as when the wind speed is significantly different between two windows of data, each describing a yacht sailing on an upwind course, then values for features within these windows will be far enough apart that the algorithm will classify them as separate clusters.

In order to apply windowing of the data for either the k-means or hierarchical clustering algorithms the data is stretched into a single high dimensional point. Consider a matrix A that contains a single window of the data of length n with m different features being used for clustering. This matrix is then mapped to a single point in $\mathbb{R}^{n \times m}$ as shown:

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,m} \end{pmatrix} \longmapsto (a_{1,1}, a_{1,2}, \cdots a_{1,m}, a_{2,1}, a_{2,2}, \cdots a_{n,1}, a_{n,2}, \cdots a_{n,m})$$

This single high dimensional point will then be used by the k-means or hierarchical clustering algorithms to group with similar high dimensional points which each represent a window of the data.

As both K-Means clustering and hierarchical clustering are fundamentally geometrically based methods of clustering it is important to scale the features supplied to these algorithms Mohamad and Usman (2013). This is due to the fact that, when classifying a data point, a 10 knot change in boat speed should not be regarded as of equal importance to a 10 degree change in wind angle. There are a number of different methods of standardization such as: min-max, Robust scaling and z-score standardization methods. In line with the findings in Mohamad and Usman (2013) it was found that z-score standardization performed the best on this data-set.

The features that were used in all the unsupervised classification algorithms (K-means, Hierarchical) were:

- Standard Features - [Bsp, abs(Awa), abs(Heel), Tws]
- Enhanced Features - [Bsp^2, abs(Awa), abs(Heel), Tws, abs(Leeway), Forestay, Bsp/Tws, abs(Awa)*Bsp, Tws/abs(Heel)]

The application of feature engineering in the enhanced features helps the unsupervised algorithm separate the data and allows important relationships such as the ratio of Bsp to Tws to be used as an important measure in separating the data into groups.

The output of such clustering models is that each datapoint is assigned to one of k distinct groups. Without any further information the model does not know what label to assign to each of these clusters. Each cluster could be inspected manually at this point and a appropriate label applied. This method can still significantly reduce workload of labeling data as typically number of clusters required to separate data are about 10-50 clusters. This is a significant reduction than manually labeling thousands of data points.

A more efficient way of completing this process is to supply the algorithm with some simple parameters. These parameters can be used by the model to decide which label ('UW', 'DW' or 'NR') to give each of the k different clusters. This process is known as semi-supervised learning.

3 PHYSICS BASED MODELS

The ORC VPP is a widely used VPP in the yacht racing industry. The main purpose of VPPs is to provide a handicap to yachts racing in fleets containing boats of varying sizes, designs and speeds. The goal of a VPP is to accurately predict the performance of a yacht in a wide range of wind conditions. Simple VPPs will typically use basic boat parameters such as length, breadth, sail area, etc. as well as the true wind speed (TWS) and true wind angle (TWA) as inputs and output predictions for a yachts speed (Bsp), heel angle and leeway angle. For conciseness the Bsp output is the main focus of this paper, however, the same methods and models presented for predicting Bsp can be applied to model both heel and leeway angles. The documentation for the ORC VPP is freely available ORC (2015). This VPP uses empirical data gathered in model testing as a basis for its underlying physical models.

In order to compare the VPP to the data the VPP was run for every data point classified as racing, taking as input the Twa and Tws from the data point and outputting predicted Bsp from the VPP.

4 DATA BASED MODELS

There are a number of available machine learning models that are suitable for this type of regression problem. Neural Networks (NN) are a popular form of nonlinear machine learning model that is capable of learning a nonlinear function from a set of training data. Random Forests (RF) are an ensemble learning model that consists of a number of different decision trees that are constructed from the training data with each tree then contributing a vote towards the desired output, in the case of regression the average of the tree predictions is then taken to be the predicted output Breiman (2001). The ensemble nature of random forests makes them less likely to overfit to training data than NN's or other forms of learning models.

In order to assses accuracy of a machine learning model it is common practice to split the available data into a *train* and *test* set. The model is then trained on the *train* data-set and the *test* data-set containing data points that have not been supplied to the model is then used to determine the models ability to generalise to data which it has not seen. The splitting of the data is done by choosing random points with the *train* set containing 80% of the available data and the test set containing the remaining 20%.

Added resistance due to waves is one of the most difficult aspects of a sailing yachts performance to capture. VPPs usually assume that the yacht is sailing in calm water. For real-time use models need to have some information of what sea conditions a yacht is experiencing in order to give informed predictions. Sensors detecting wave height and period do exist but are extremely rare onboard sailing yachts. However, it is common place for a racing yacht to have an accelerometer and gyroscope sensor fitted that can measure and record the pitch and heel of a yacht. By utilising the process of feature engineering it is possible to transform the pitch and heel data to new features namely heel amplitude, pitch amplitude, heel frequency and pitch frequency. This was done by identifying "*peaks*" and "*troughs*" in the time-series data. The frequency and amplitude of these peaks and troughs were then calculated and then added as respective features to the data set.

5 PHYSICS BASED LEARNING MODELS

Combining physical models with data should reduce the amount of data needed in order to achieve good model predictions. Weymouth and Yue (2013) have shown how combining a simple physical model with a small number of data points can improve the performance of a model in data sparse cases. This type of model is known as a physics-based learning model (PBLM). PBLM's utilize physics based insights of the problem to improve the accuracy and also reduce the data dependence of a General Learning Model (GLM).

In this work the ORC based VPP is used an an IM and both simple regression techniques as well as a random forest are explored as potential GLMs. The features supplied to the GLM will be Twa, Tws, pitch amplitude, heel ammplitude, pitch frequency and heel frequency. The IM takes Tws and Twa as inputs along with boat shape parameters.

6 RESULTS

6.1 Data Classification

6.1.1 Supervised Classification

The resulting accuracy score of the supervised random forest model can be seen in Figure 1. The accuracy score relates to simply the percentage of points in the test data that had predicted labels equal to the actual manually assigned labels. The points in the plot relate to a type of "mean of means", this is due to the fact that to reduce the stochastic influence of the RF model the model was run 10 times for each given training data with the mean accuracy taken as the score for that training data. In order to reduce analyse the influence of using different days as training data for a fixed n number of days each possible combination of n days was fit (10 times each) as training data with the remaining data being used as test data. Therefore for a given n the accuracy score represents the mean accuracy score over each possible combination of training data. The error bars represent the standard deviation of these accuracy scores.

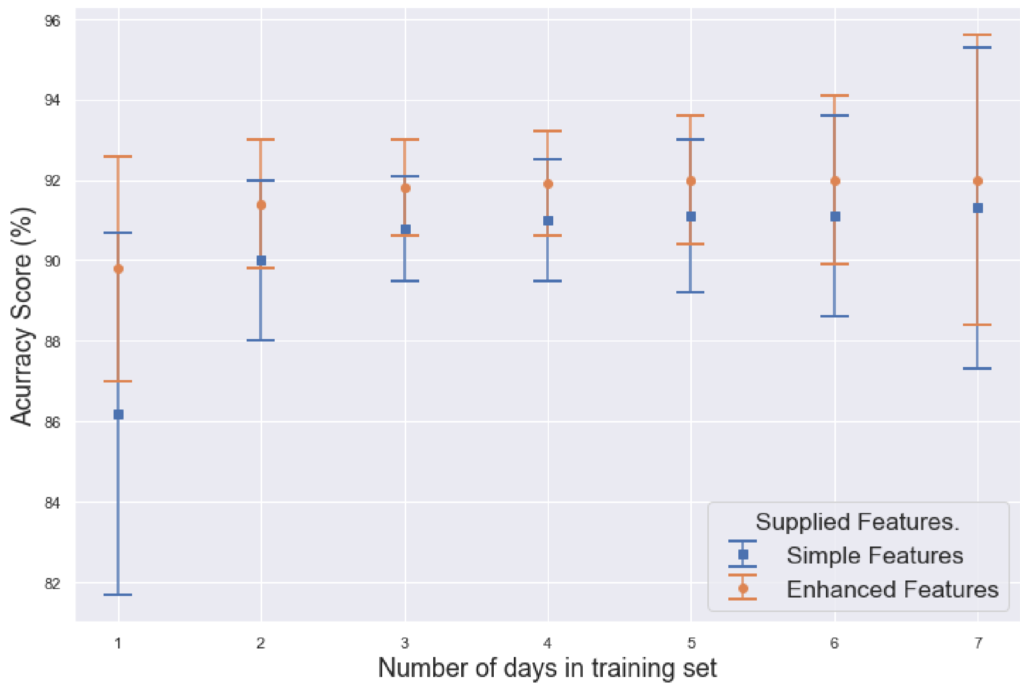

Figure 1: Accuracy of supervised Random Forest classification. The RF model was fit to each possible combination of n training days , with the remaining number of days of data being used for testing. The model was fit 10 times to each given training data to reduce the influence of the stochastic nature of RF model. Presented is the mean and standard deviation of the model over the different training data provided for a given number of days of training data.

6.1.2 Unsupervised Classification

There are two main user defined parameters in the version of both k-means clustering and hierarchical clustering used: window length and number of clusters (k). It was found that the model performance is optimal using a window length of 40s, 30 clusters for the k-means model and 12 clusters for the hierarchical model.

In Table 1 the results of the data classification process are presented. The unsupervised classification models outperform the supervised classification models. For both the unsupervised models the appropriate window length was found to be about 40 seconds.

Table 1: Summary of Classification model accuracy.

Type of Model	Model	Parameters	Accuracy (%)
Supervised	RF	Enhanced Features $n_{days} = 7$ $n_{trees} = 100$	92.0
Supervised	SVM	Simple Features n days = 7	92.7
Unsupervised	K-Means	Enhanced Features k = 30 window len = 40	93.06
Unsupervised	Hierarchical	Enhanced Features k = 12 window len = 40	93.5

6.2 Physics Based Performance Models

Table 2 summarises the accuracy scores of the ORC based VPP. When considering Bsp the ORC based VPP performs significantly better when compared with UW data rather than DW data. However, when looking at the accuracy of Heel predictions the VPP is significantly more accurate at predicting the DW Heel than UW Heel. VPP's in general tend to overestimate both the UW speed and UW Heel experienced by a yacht while also failing to capture when a yacht will experience planing thus underestimating DW speed in certain conditions.

Table 2: Summary of ORC based VPP accuracy metrics. VPP outputs compared to observed steady state values at a minimum Twa of 25 degrees.

Output	Data Range	RMSE	MAE	R^2	Max error
Bsp	Racing	1.61	1.264	0.361	7.14
Bsp	UW	1.23	0.989	-0.79	5.99
Bsp	DW	2.079	1.706	0.044	7.14
Heel	Racing	10.775	8.729	-1.629	38.698
Heel	UW	13.373	12.324	-8.66	38.698
Heel	DW	3.93	2.962	-0.243	30.07

6.3 Data Based Performance Models

Fitting data based models to the data shows a significant increase in the accuracy when compared with the ORC based VPP. Figure 2 shows the comparison of Bsp models based on the R^2 score of a given model on the steady-state racing data (consisting of both UW and DW data points). Even when the machine learning models are only trained using basic features such as just Tws and Twa all models significantly outperform the physics based VPP in predicting both Bsp and Heel. Using just Tws and Twa as input parameters the models achieves a test set accuracy of $R^2 = 0.856$ for Bsp predictions (compared to 0.361 for the VPP model).

The inclusion of parameters relating to the sea conditions that the yacht is experiencing into the data based models sees a large increase in model accuracy. The test set score for predicting Bsp of the random forest model increases from $R^2 = 0.839$ to $R^2 = 0.949$ with the addition of these motion based features.

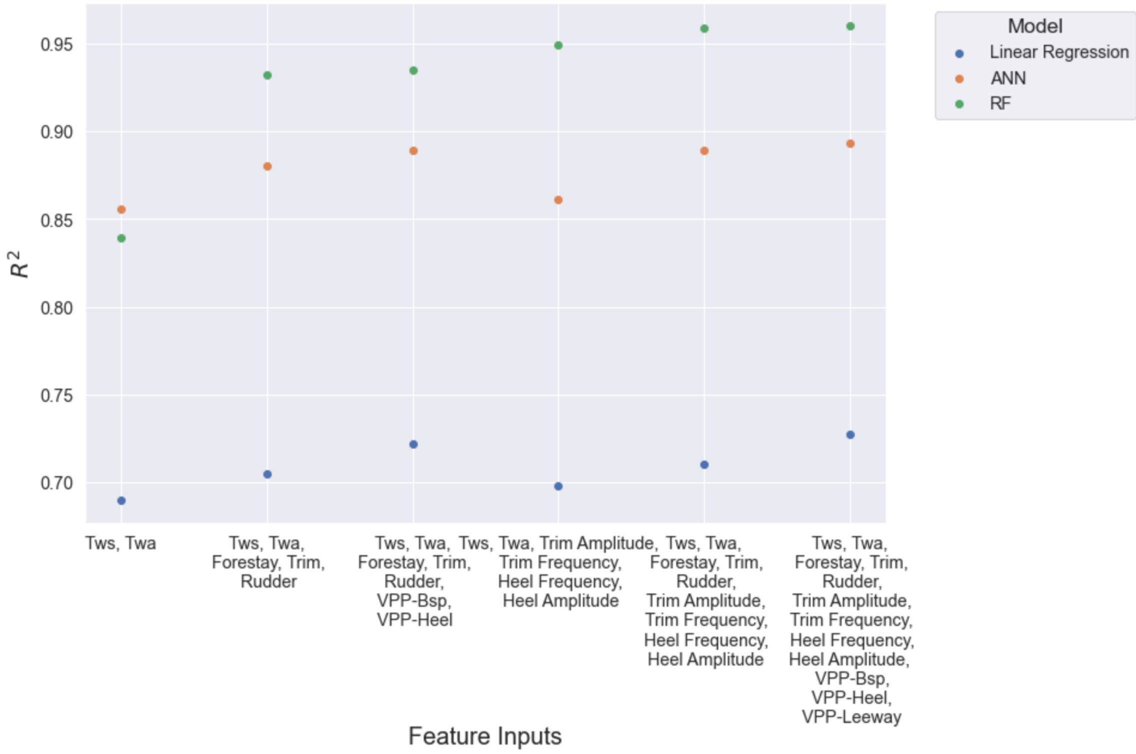

Figure 2: Data based models test set accuracy score for predicting boat speed. The VPP accuracy score on the same data is $R^2 = 0.361$

The ability of the data based models to extrapolate beyond the type of data which it was trained on was investigated. Training data here consisted of steady state data with a Tws $<$ 15 knots. The test data consisted of data points with Tws $>$ 15 knots and consisted of approximately 30% of the total dataset. Figure 4 shows how the accuracy of the data based models for extrapolation is significantly worse than when the model learned on data similar to the data it was tested on. This is a drawback to using data based models which may be improved upon using physics based learning models in the future. However, the support vector machine model is seen to generalise much better than the other data based models and performs almost as well as the Windesign VPP and still significantly better than the ORC VPP on this data. It should be noted that the ORC VPP is also significantly less accurate on this set of high wind speed data compare to lower wind speeds. The RMSE of the ORC VPP is 2.19 knots for high wind speed data compared to 1.34 knots for the data point with Tws $<$ 15 knots.

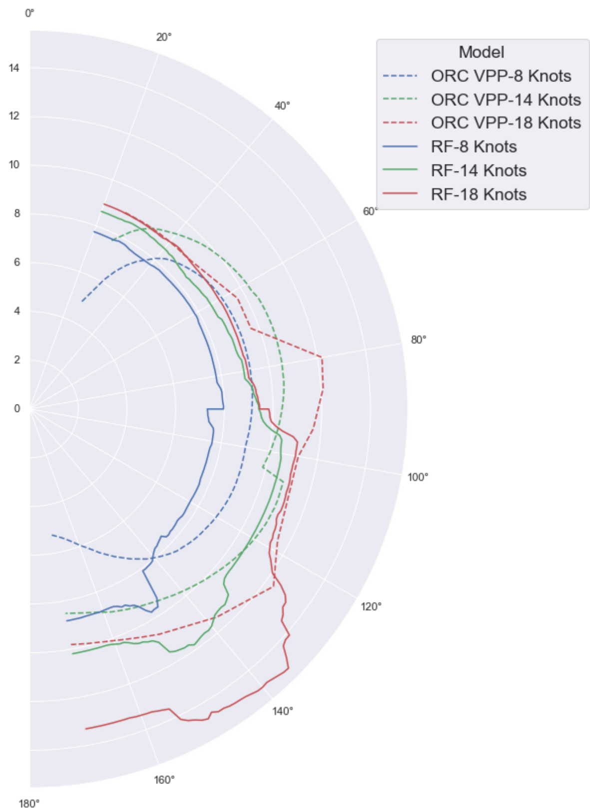

Figure 3: Comparison of Polar plot generated from Random Forest model to the polar plot generated from the ORC VPP.

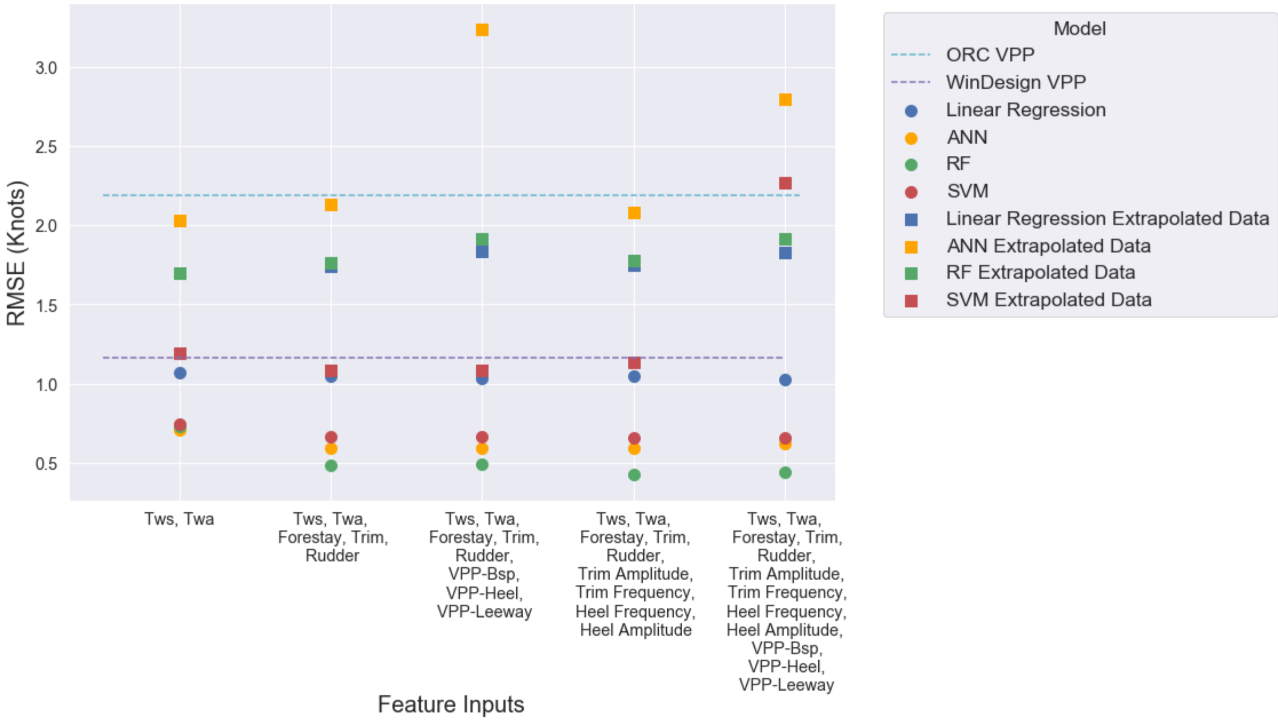

Figure 4: Data based Bsp prediction accuracy tested on data outside the learning range. Test data for all models consists of data points wih Tws > 15 knots.

6.4 Physics-Based Learning Models

In Figure 5 the use of a PBLM model can be seen to greatly increase the accuracy of a model using a basic GLM such as ridge regression when the number of data points is small. In the case of using the

RF as a GLM, there is only a slight increase in model accuracy in using a PBLM model even when the number of data points is small.

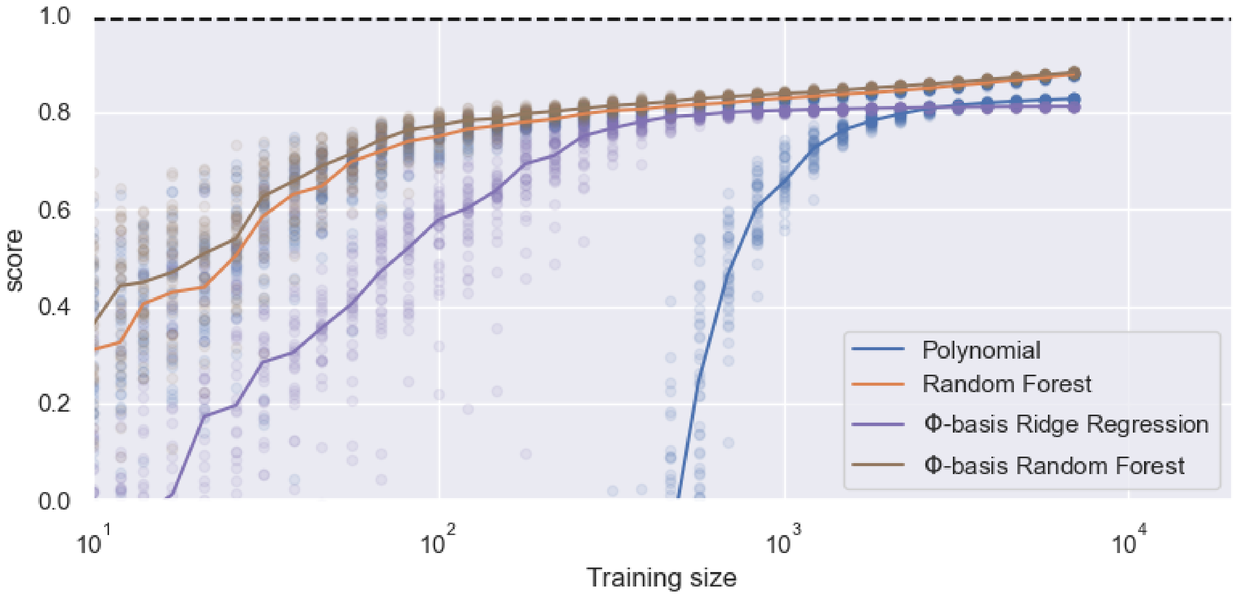

Figure 5: Explained variance score of PBLM models for varying number of available data points. At each training size the available data is randomly split into train and test sets (with the required number of points in training set). This is done at random 35 times for each training size. The scatter points represent the individual model score for each of the random training sets. The solid line represents the median of the model scores for a given training set size.

7 CONCLUSIONS

This work has shown the advantages of using data based models for modeling a yachts performance when compared to semi-empirical methods. Simple data based methods such as linear regression were shown to outperform the VPPs, however, more advanced methods of regression such as random forests were shown to perform significantly better than using linear regression on this type of data.

The power of feature engineering was shown by the ability to transform the pitch and heel data of the yacht into features that the machine learning models can use as replacements for wave data. This has the potential to greatly increase the performance of using such a model for real-time predictions.

A brief overview of using unsupervised machine learning in the pre-processing of yacht racing data is presented and is shown to be a very useful tool in reducing the amount of time it takes to group a data set into distinct groups. This method of classification can be generalised to many other forms of real world data that is captured in the form of a time series.

The use of PBLMs is explored in use with real world sailing data. It is shown however that a physical basis does not greatly improve the accuracy of the model. A more accurate IM or less noisy data points may be needed for PBLM to be utilised more effectively.

References

S. Benjamin. Tp52 performance data - spookie3 - superseries cascais, 2015. http://library.sailyachtresearch.org//images/library/Spookie3/Spookie%20TP52%20Cascais%20Dat

C. Böhm. A velocity prediction procedure for sailing yachts with a hydrodynamic model based on integrated fully coupled ranse-free-surface simulations. 2014.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

E. W. Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, 21:768–769, 1965.

I. B. Mohamad and D. Usman. Standardization and its effects on k-means clustering algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, 6(17):3299–3303, 2013.

ORC. Orc vpp documentation 2015. Technical report, Technical report. 118, 120, 2015.

L. Rokach and O. Maimon. Clustering methods. In *Data mining and knowledge discovery handbook*, pages 321–352. Springer, 2005.

G. D. Weymouth and D. K. Yue. Physics-based learning models for ship hydrodynamics. *Journal of Ship Research*, 57(1):1–12, 2013.