Predictive maintenance of Naval Propulsion Plants Data Set

Benno Weber

Abstract

This report describes the analysis steps conducted on the the Naval Propulsion Plants Data Set available in the UCI machine learning repository. Goal of the analysis is to predict values of the two output variables 'GT Compressor decay state coefficient' and 'GT turbine decay state coefficient' based on the values of 16 features representing measurements in a vessel and specifically the engine compartment. Basic data exploration steps were conducted, the dataset was cleaned and normalized. Dataset was split into train and test dataset, a linear regression and a decision tree regression model were applied. Both algorithms performed well on the given test dataset. Accuracy of the algorithms was measured by the mean squared error. Results of the analysis can be used to evaluate the engine state and to implement condition based or predictive maintenance on the system.

Motivation

The Naval Propulsion Plants Data Set yields interesting opportunities for machine learning applications. The application of regression algorithms like Linear regression, decision trees, GLM, GAM, quantile regression or SVM as well as a combination of several models could be the basis of a system able to assist companies in maintaining production machines. Predictive maintenance is a hot topic nowadays for any company with machines that cost a lot of money when not producing goods. Implementing a system that is able to predict a possible machine failure in time (=before it happens) can cut down repair costs and be an important part of a efficient production environment.

Dataset(s)

The Naval Propulsion Plants Data Set from [1] includes 11934 observations consisting of 18 measurements of different engine properties respectively.

- 14 features were used for prediction
- 2 features were dropped due to no or low variance
- 2 variables are our outcome variables which were predicted by the models, these are:
    - Gas turbine Compressor decay state coefficient (GT_comp_decay) and
    - Gas turbine Turbine decay state coefficient (GT_turbine_decay)

Data Preparation and Cleaning

The following data cleaning steps were conducted on the dataset:

- Drop NA values (drop those observations which had a NA value in any column)
- Drop two features, that is GT Compressor inlet air temperature and GT Compressor inlet air pressure due to no or very low variance
- Scale all 14 predictor variables using the standard scaler from sklearn due to differences in units and properties.

Research Question(s)

Based on 14 features from the Naval Propulsion Plants Data Set, is it possible to predict the value of two system-critical key indicators, GT_comp_decay and GT_turbine_decay?

If yes - how accurate are regression models able to predict the actual values of the two output variables mentioned above?

Methods

Multivariate linear regression and decision tree regression was used in order to tackle the supervised machine learning problem formulated in the research question.

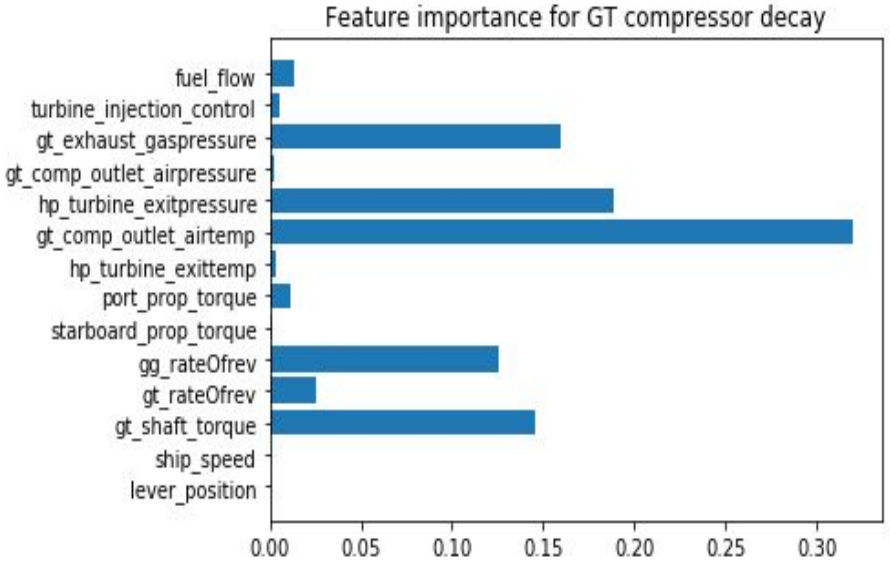
The dataset was split into training and test datasets, the models were fitted on the training set and afterwards evaluated on the test set. Accuracy of the models was evaluated by calculating the round mean squared error. RSME was calculated for the first (gt_comp_decay) and the second outcome variable (gt_turbine_decay) separately.

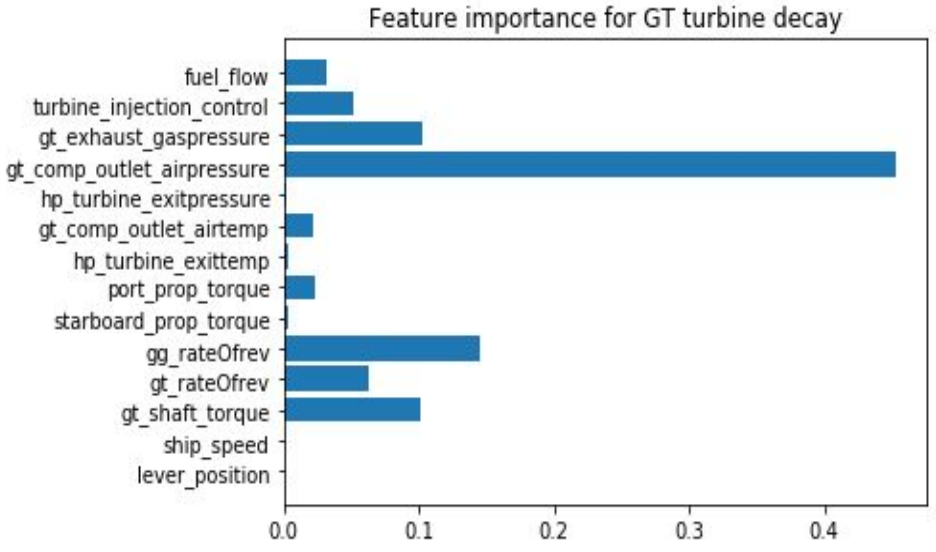
Findings

RMSE of the two models linear regression and decision tree regressor was calculated for both of the outcome variables gt_comp_decay and gt_turbine_decay. Here is the result: (rounded to 6 decimal numbers)

RSME Linear Regression GT comp decay	RSME Desc. Tree GT comp decay	RSME Linear Regression GT turbine decay	RSME Desc. Tree GT turbine decay
0.005949	0.002194	0.001527	0.001408

The difference in RSME is very small for GT turbine decay (only about 0.0002) and a little larger for GT compressor decay (0.0038).We see that decision tree regressor performs better to predict the GT compressor decay and only a little better on the turbine decay state. Next we want to investigate which variables were the most important ones for the decision tree predictor in order to predict our two output variables GT compressor decay and GT turbine decay. We do this by plotting the features importances from the predictor as barplots on the next slide.





We see in the plots above that the decision tree regressor uses features differently for predicting our two output variables. For GT compressor decay it is the variable 'gt_comp_outlet_airtemp' that is most important for the algorithm. 'gt_comp_outlet_airpressure' is of highest importance for predicting GT turbine decay.

Limitations

For the next micromasters course, machine learning, I would like to implement other regression models on the Naval Propulsion Plants Data Set like GLM,GAM, quantile regression or SVM. I would also like to do a combination of models in order to further minimize RMSE.

I know that this report lacks of elegant visualizations but I couldn’t think of any plots that yield a better understanding of the analysis steps conducted so I preferred not to include more color to this presentation just for the sake of doing it.

I am happy to receive feedback from you - maybe you can leave some ideas for plots in your review!

Conclusions

Both algorithms, linear regression and decision tree regression provide good results in predicting the values of our two output variables based on the 14 features available in the dataset used.

The difference in RSME is very small for GT turbine decay (only about 0.0002) and a little larger for GT compressor decay (0.0038).We see that decision tree regressor performs better to predict the GT compressor decay and only a little better on the turbine decay state.

Acknowledgements

The analysis this presentation is based on was conducted by me alone. I used code and information from the micromasters course ‘Python for Data Science’ from UCSD. Data was obtained from UCI machine learning repository, see references on the next slide.

References

The dataset used was obtained by:

[1] A. Coraddu, L. Oneto, A. Ghio, S. Savio, D. Anguita, M. Figari, Machine Learning Approaches for Improving Condition Based Maintenance of Naval Propulsion Plants, Journal of Engineering for the Maritime Environment, 2014, DOI: 10.1177/1475090214540874, (In Press)

Dataset was downloaded from UCI machine learning repository:

https://archive.ics.uci.edu/ml/datasets/Condition+Based+Maintenance+of+Naval+Propulsion+Plants