Predictive Maintenance

C.T.

21/03/2021

Introduction

This project will explore the data obtained from a hydraulic test rig. The test rig is equipped with a number of sensors, and four hydraulic components (cooler, valve, pump and accumulator). The failure of these hydraulic components within the rig will be modeled, and their stages of failure predicted based on sensor data.

About the dataset

The data set was obtained from kaggle.com, and as summarized by the author:

"The data set was experimentally obtained with a hydraulic test rig. This test rig consists of a primary working and a secondary cooling-filtration circuit which are connected via the oil tank. The system cyclically repeats constant load cycles (duration 60 seconds) and measures process values such as pressures, volume flows and temperatures while the condition of four hydraulic components (cooler, valve, pump and accumulator) is quantitatively varied."

The citation for the data is as follows: *Nikolai Helwig, Eliseo Pignanelli, Andreas Schütze, 'Condition Monitoring of a Complex Hydraulic System Using Multivariate Statistics', in Proc. I2MTC-2015 - 2015 IEEE International Instrumentation and Measurement Technology Conference, paper PPS1-39, Pisa, Italy, May 11-14, 2015, doi: 10.1109/I2MTC.2015.7151267.*

Goal

The goal is to correctly predict the failure of four components based on sensor data using machine learning algorithms. A number of algorithms will be trialed to determine the optimum model/technique, and mixture of features (predictors). Evaluation of the algorithms will be assessed by comparing the predicted data against the known data, using the following parameters: Precision, Accuracy, F1 Score, Recall and Model Run Time.

Note: A Validation data-set will be prepared, which will only be used at the end of the project to assess the final model. It will not be used to test the models during model development.

Key Steps

Key Steps in the project:

1. Download/extract the data.
2. Investigate the structure of the data set, generate plots to visualise data where required.
3. Pre-process the data.
    - Clean/Wrangle data.
    - Average predictors where required.
    - Remove predictors that are not useful, are highly correlated with others, have very few non-unique values, or have close to zero variation.
4. Split the Data in Model Data and Validation Data. Further split the Model data into Training and Test set in order to evaluate the algorithms.

5. Build multiple algorithms, and include additional features/biases and/or predictor pre-processing to optimise.
6. Summarise data, choosing optimal algorithm based on training/test data.
7. Apply algorithm to validation dataset.

Pre-Analysis and Data Investigation

Library

The following packages/versions have been used, this includes dependencies:

Table 1: Packages

package	packages
caret	6.0.86
data.table	1.14.0
dplyr	1.0.5
forcats	0.5.1
foreach	1.5.1
gam	1.20
ggplot2	3.3.3
kableExtra	1.3.4
knitr	1.31
lattice	0.20.41
lubridate	1.7.10
purrr	0.3.4
readr	1.4.0
stringr	1.4.0
tibble	3.1.0
tidyr	1.1.3
tidyverse	1.3.0

Initial Data Load

Data downloaded from the Github repository has multiple compressed files containing a number of text files. The initial data wrangle downloads and extracts these files, and stores them into a data-frame. This data-frame is then split into two: 'Model' and 'Validation'. The validation set is set at 10% of the Model data. The Model Data is further split into two: Train Data (used for training the algorithms) and Test Data (for verifying the models), with the Test data set at 10% of the Train data.

Cross Validation

The use of a three independent data sets: Training, Test and Validation, ensures the model can be suitably assessed, evaluated and improved without using the Validation data. This technique is know as Cross Validation, and is used to avoid over-fitting, or predictor selection bias. The split of the data 90%/10% is a common split, and ensures sufficient representative data is available to Train, Test and Validate the models.

Performance Evaluation Criteria

The following parameters have been used to evaluate the performance of the model. These parameters are described below.

- Precision: The ratio of correctly predicted positive values to the total predicted positive values.
- Accuracy: The ratio of correctly predicted values to the total values.
- F1 Score: A weighted average of Precision and Recall. Takes both false positives and false negatives into account.
- Recall: A ratio of the correctly predicted positive values to the all values in the series.
- Model Run Time: How long each model takes to run. Lower values are better.

Precision, Accuracy, F1 Score and Recall all have maximum (perfect) score of 1 (ie, 100%).

Files

There are 17 sensor files (1 file per sensors), and 1 profile file which contains the condition of each major component. All files are in a text format. The file contains rows, which represent the cycles, and columns which represent the data points within a cycle.

The files are named as follows:

Table 2: Files

Files
CE.txt
CP.txt
EPS1.txt
FS1.txt
FS2.txt
profile.txt
PS1.txt
PS2.txt
PS3.txt
PS4.txt
PS5.txt
PS6.txt
SE.txt
TS1.txt
TS2.txt
TS3.txt
TS4.txt
VS1.txt

Sensor Data

All sensor data files contain the same number of rows, which indicates the data does not have any time-line gaps (ie, missing data). In addition, sensors contain either 60, 600 or 6000 columns per row, which aligns with the expected sensor polling frequency (ie, 1 Hz, 10 Hz or 100 Hz).

Typical sensor data from one of the files is shown below. The files do not contain any headers, and only contain numeric values separated by a tab delimiter.

Relevant statistics (Mean, RSD) for each sensor cycle (row) is also calculated and shown below.

Table 3: File Sensor Row/Columns

File	Rows	Columns
CE.txt	2205	60
CP.txt	2205	60
EPS1.txt	2205	6000
FS1.txt	2205	600
FS2.txt	2205	600
PS1.txt	2205	6000
PS2.txt	2205	6000
PS3.txt	2205	6000
PS4.txt	2205	6000
PS5.txt	2205	6000
PS6.txt	2205	6000
SE.txt	2205	60
TS1.txt	2205	60
TS2.txt	2205	60
TS3.txt	2205	60
TS4.txt	2205	60
VS1.txt	2205	60

Table 4: Typical Sensor File Data (showing 10 columns only)

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
0.604	0.605	0.611	0.603	0.608	0.608	0.608	0.617	0.619	0.619
0.590	0.610	0.626	0.620	0.623	0.619	0.617	0.618	0.619	0.615
0.578	0.603	0.638	0.651	0.652	0.662	0.662	0.656	0.652	0.638
0.565	0.591	0.608	0.614	0.623	0.645	0.642	0.645	0.642	0.643
0.570	0.600	0.623	0.636	0.644	0.642	0.651	0.654	0.660	0.644
0.568	0.601	0.611	0.614	0.629	0.627	0.630	0.629	0.623	0.643

Table 5: RSD Range for each Sensor

	sensor_ce	sensor_cp	sensor_eps1	sensor_fs1	sensor_fs2	sensor_ps1	sensor_ps2	sensor_ps3	sensor_ps4	sensor_ps5	sensor_ps6	sensor_se	sensor_ts1	sensor_ts2	sensor_ts3	sensor_ts4	sensor_vs1
RSD Min	0.003	0.005	0.072	0.444	0.001	0.086	0.430	0.464	0.00	0.002	0.002	0.399	0.002	0.001	0.000	0.000	0.023
RSD Max	0.161	0.154	0.113	1.353	0.009	0.129	0.469	0.921	77.46	0.009	0.009	1.160	0.016	0.009	0.007	0.038	0.554

Table 6: Mean Range for each Sensor

	sensor_ce	sensor_cp	sensor_eps1	sensor_fs1	sensor_fs2	sensor_ps1	sensor_ps2	sensor_ps3	sensor_ps4	sensor_ps5	sensor_ps6	sensor_se	sensor_ts1	sensor_ts2	sensor_ts3	sensor_ts4	sensor_vs1
Mean Min	17.56	1.06	2361.75	2.02	8.86	155.39	104.41	0.84	0.00	8.37	8.32	18.28	35.31	40.86	38.25	30.39	0.52
Mean Max	47.90	2.84	2740.64	6.72	10.40	180.92	131.59	2.02	10.21	9.98	9.86	60.76	57.90	61.96	59.42	53.06	0.84

Profile (Component) Data

The Profile file contains the same number of rows as the sensor (2205), this confirms that the data can be joined together without causing any time-line errors. There are 5 columns: 1 per component, and 1 to indicate if the system was stable.

The unique values for each component has been summarized below, this shows there are specific distinct levels to represent the component status. From the source files these values can be interpreted as such:

Cooler condition

- 3: close to total failure
- 20: reduced efficiency
- 100: full efficiency

Valve condition

- 100: optimal switching behavior
- 90: small lag
- 80: severe lag
- 73: close to total failure

Internal pump leakage

- 0: no leakage
- 1: weak leakage
- 2: severe leakage

Hydraulic accumulator / bar

- 130: optimal pressure
- 115: slightly reduced pressure
- 100: severely reduced pressure
- 90: close to total failure

Stable Flag

- 0: conditions were stable
- 1: static conditions might not have been reached yet

Table 7: Profile File Structure

File	Rows	Columns
profile.txt	2205	5

Table 8: Profile Component Unique Values

cooler	valve condition	internal pump leakage	hydraulic accumulator	stable
3,20,100	73,80,90,100	0,1,2	90,100,115,130	0,1

Data Structure

The mean value for each Sensor cycle, and each of the Profile values have been combined into a single data-set. There are 17 columns representing each sensor, and 5 columns representing each profile value. The structure of the data is shown below.

Variable	Class	First.Values
sensor_ce	double	39.60135, 25.7864333333333, 22.2182333333333
sensor_cp	double	1.86275, 1.25555, 1.11321666666667
sensor_eps1	double	2538.92916666667, 2531.4989, 2519.928
sensor_fs1	double	6.709815, 6.715315, 6.71852166666667
sensor_fs2	double	10.3045916666667, 10.4030983333333, 10.36625
sensor_ps1	double	160.673491666667, 160.60332, 160.34772
sensor_ps2	double	109.4669135, 109.354890333333, 109.158844666667
sensor_ps3	double	1.99147533333333, 1.97623433333333, 1.972224
sensor_ps4	double	0, 0, 0
sensor_ps5	double	9.8421695, 9.63514216666667, 9.53054783333333
sensor_ps6	double	9.7280975, 9.52948783333333, 9.42794883333333
sensor_se	double	59.1571833333333, 59.3356166666667, 59.54315
sensor_ts1	double	35.6219833333333, 36.6769666666667, 37.8808
sensor_ts2	double	40.9787666666667, 41.5327666666667, 42.44245
sensor_ts3	double	38.4710166666667, 38.9789666666667, 39.63195
sensor_ts4	double	31.74525, 34.4938666666667, 35.64615
sensor_vs1	double	0.57695, 0.56585, 0.576533333333333
profile_cooler	integer	3, 3, 3
profile_valve_condition	integer	100, 100, 100
profile_internal_pump_leakage	integer	0, 0, 0
profile_hydraulic_accumulator	integer	130, 130, 130
profile_stable	integer	1, 1, 1

The data-set contains the following number of rows and columns:

Table 9: Data Set Row/Column Count

	value
rows	2205
columns	22

Pre-Processing

Predictors that don not add value should be removed from the data-set. These predictors slow down the Model training (consume CPU cycles), and/or can reduce the model performance (eg. accuracy).

Near Zero Variance (NZV)

Near Zero Variance (NZV) analysis diagnoses predictors that have one unique value or predictors that are have both of the following characteristics: they have very few unique values relative to the number of samples and the ratio of the frequency of the most common value to the frequency of the second most common value is large. The following summaries the results of this analysis, and shows that all predictors should be used based on the NZV analysis.

Definitions

- Frequency Ratio: The ratio of frequencies for the most common value over the second most common value
- Percent Unique: The percentage of unique data points out of the total number of data points
- Zero Variation/Near Zero Variation: A vector of logicals for whether the predictor has only one distinct value or whether the predictor is a near zero variance predictor.

Table 10: Predictor Near Zero Variation Analysis

	Freq. Ratio	Percent Unique	Zero Variation	Near Zero Variation
sensor_ce	1.000000	99.66311	FALSE	FALSE
sensor_cp	1.333333	95.00281	FALSE	FALSE
sensor_eps1	1.000000	100.00000	FALSE	FALSE
sensor_fs1	1.000000	99.32622	FALSE	FALSE
sensor_fs2	1.000000	99.88770	FALSE	FALSE
sensor_ps1	2.000000	99.94385	FALSE	FALSE
sensor_ps2	1.000000	100.00000	FALSE	FALSE
sensor_ps3	2.000000	99.94385	FALSE	FALSE
sensor_ps4	495.000000	44.41325	FALSE	FALSE
sensor_ps5	2.000000	99.94385	FALSE	FALSE
sensor_ps6	2.000000	99.94385	FALSE	FALSE
sensor_se	1.000000	99.77541	FALSE	FALSE
sensor_ts1	1.000000	99.60696	FALSE	FALSE
sensor_ts2	1.000000	99.60696	FALSE	FALSE
sensor_ts3	1.000000	99.71926	FALSE	FALSE
sensor_ts4	1.000000	99.49467	FALSE	FALSE
sensor_vs1	1.000000	91.01628	FALSE	FALSE

Predictor correlations

Predictor correlations must also be evaluated. Highly correlated predictors can improve a machine learning algorithm, but can also reduce the effectiveness, depending on the model used. The below summaries the correlation between each predictor. From this analysis, we can see that no predictor pairs are highly correlated (>0.999).

Table 11: Predictor Correlation

	sensor_ce	sensor_cp	sensor_eps1	sensor_fs1	sensor_fs2	sensor_ps1	sensor_ps2	sensor_ps3	sensor_ps4	sensor_ps5	sensor_ps6	sensor_se	sensor_ts1	sensor_ts2	sensor_ts3	sensor_ts4	sensor_vs1
sensor_ce	1.000	0.972	0.474	0.376	0.919	-0.044	-0.119	0.698	0.807	0.974	0.974	0.293	-0.946	-0.946	-0.942	-0.956	-0.851
sensor_cp	0.972	1.000	0.431	0.381	0.871	-0.065	-0.137	0.678	0.741	0.933	0.933	0.302	-0.908	-0.905	-0.899	-0.923	-0.816
sensor_eps1	0.474	0.431	1.000	-0.609	0.324	0.833	0.789	-0.269	0.453	0.418	0.420	-0.679	-0.372	-0.398	-0.364	-0.377	-0.073
sensor_fs1	0.376	0.381	-0.609	1.000	0.545	-0.924	-0.946	0.921	0.197	0.461	0.459	0.995	-0.505	-0.480	-0.512	-0.500	-0.713
sensor_fs2	0.919	0.871	0.324	0.545	1.000	-0.223	-0.295	0.799	0.685	0.980	0.979	0.466	-0.994	-0.993	-0.996	-0.991	-0.922
sensor_ps1	-0.044	-0.065	0.833	-0.924	-0.223	1.000	0.995	-0.719	0.043	-0.124	-0.123	-0.944	0.175	0.148	0.184	0.170	0.445
sensor_ps2	-0.119	-0.137	0.789	-0.946	-0.295	0.995	1.000	-0.767	-0.021	-0.200	-0.198	-0.958	0.249	0.222	0.258	0.244	0.506
sensor_ps3	0.698	0.678	-0.269	0.921	0.799	-0.719	-0.767	1.000	0.480	0.755	0.754	0.884	-0.776	-0.758	-0.781	-0.776	-0.895
sensor_ps4	0.807	0.741	0.453	0.197	0.685	0.043	-0.021	0.480	1.000	0.738	0.738	0.125	-0.702	-0.705	-0.700	-0.713	-0.656
sensor_ps5	0.974	0.933	0.418	0.461	0.980	-0.124	-0.200	0.755	0.738	1.000	1.000	0.377	-0.993	-0.993	-0.991	-0.995	-0.898
sensor_ps6	0.974	0.933	0.420	0.459	0.979	-0.123	-0.198	0.754	0.738	1.000	1.000	0.376	-0.993	-0.993	-0.991	-0.995	-0.897
sensor_se	0.293	0.302	-0.679	0.995	0.466	-0.944	-0.958	0.884	0.125	0.377	0.376	1.000	-0.423	-0.396	-0.431	-0.418	-0.650
sensor_ts1	-0.946	-0.908	-0.372	-0.505	-0.994	0.175	0.249	-0.776	-0.702	-0.993	-0.993	-0.423	1.000	0.999	1.000	0.999	0.913
sensor_ts2	-0.946	-0.905	-0.398	-0.480	-0.993	0.148	0.222	-0.758	-0.705	-0.993	-0.993	-0.396	0.999	1.000	0.999	0.998	0.902
sensor_ts3	-0.942	-0.899	-0.364	-0.512	-0.996	0.184	0.258	-0.781	-0.700	-0.991	-0.991	-0.431	1.000	0.999	1.000	0.998	0.914
sensor_ts4	-0.956	-0.923	-0.377	-0.500	-0.991	0.170	0.244	-0.776	-0.713	-0.995	-0.995	-0.418	0.999	0.998	0.998	1.000	0.912
sensor_vs1	-0.851	-0.816	-0.073	-0.713	-0.922	0.445	0.506	-0.895	-0.656	-0.898	-0.897	-0.650	0.913	0.902	0.914	0.912	1.000

Table 12: Predictors >0.999 Correlation

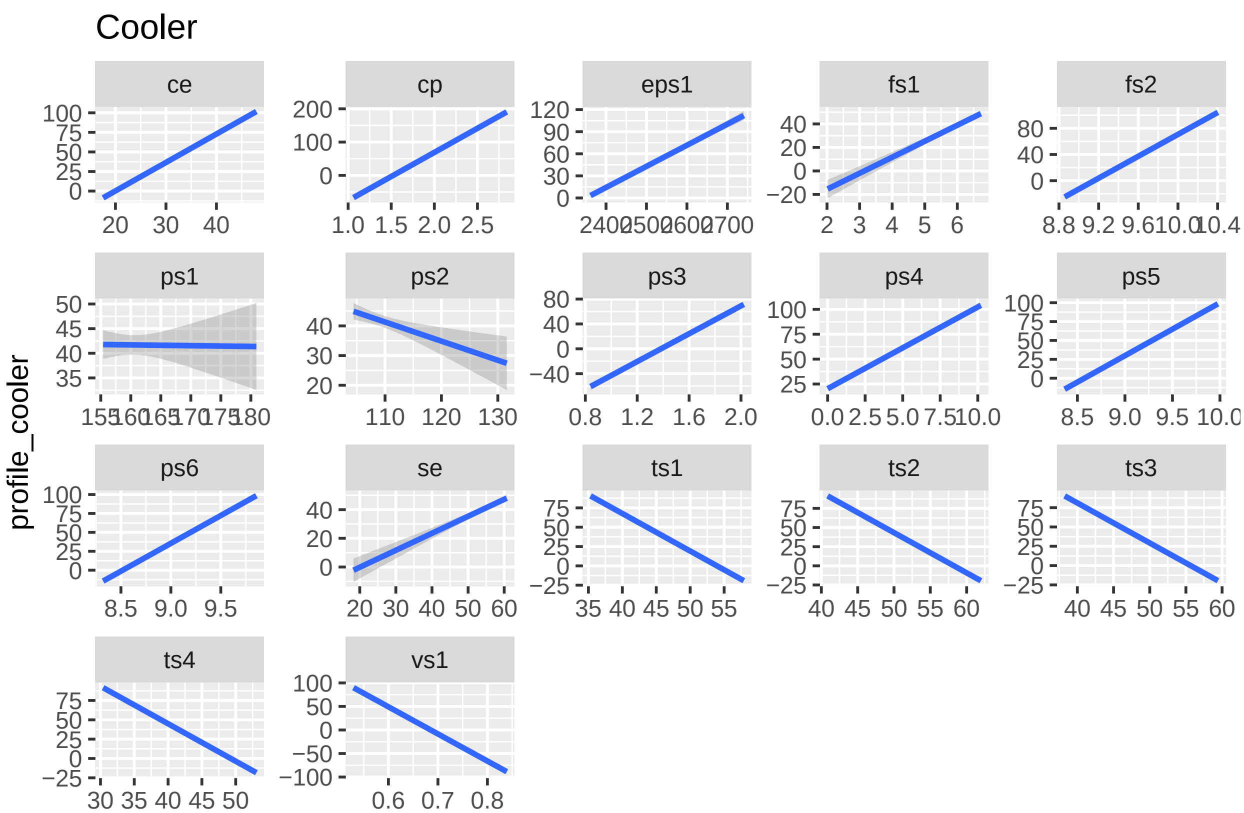
x
0

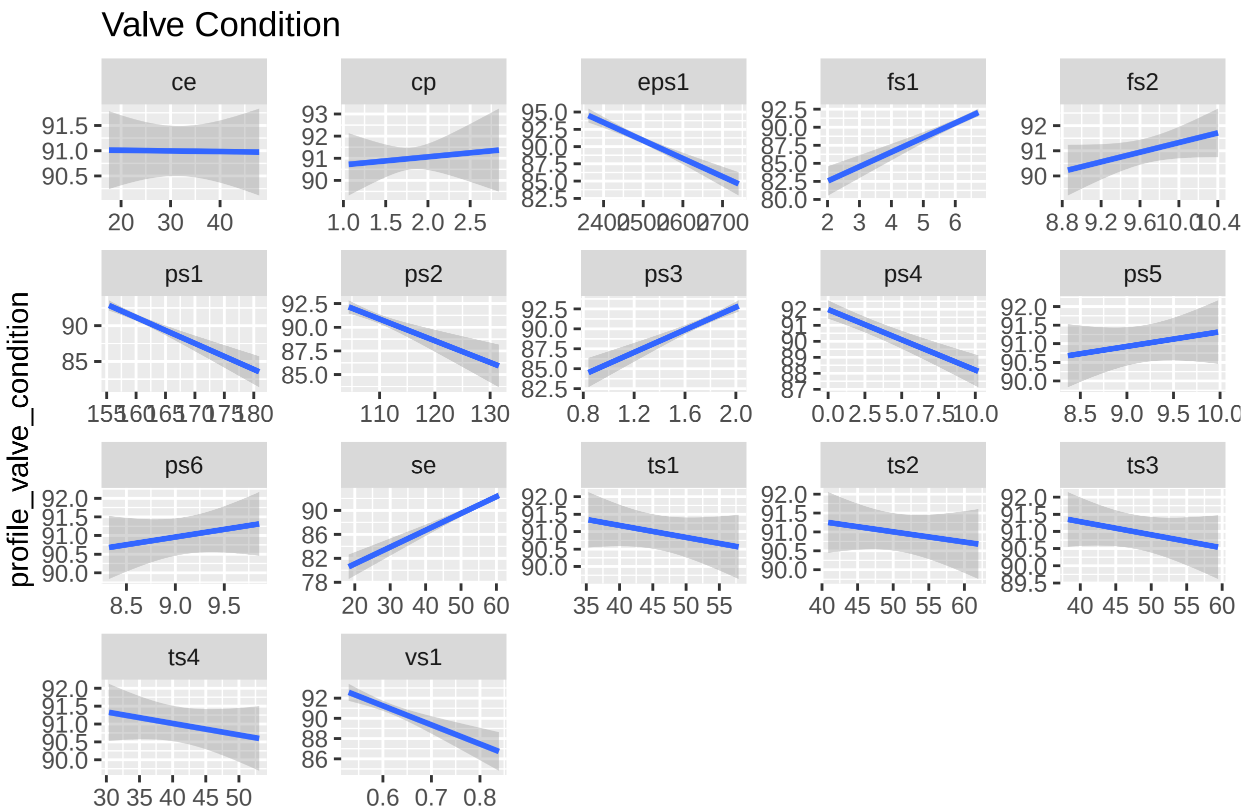
Plots

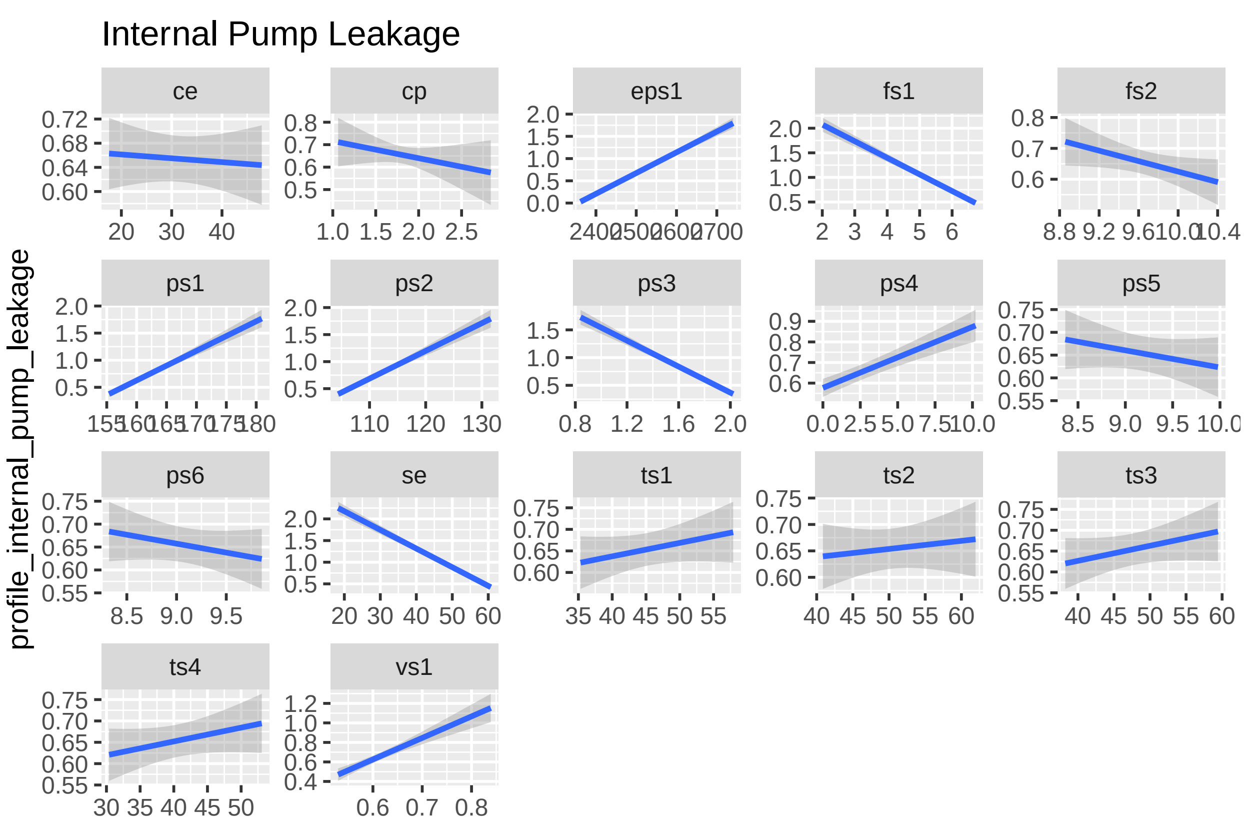
Each Predictor is plotted against the Performance criteria. Two smoothing fits are shown: Linear and Loess.

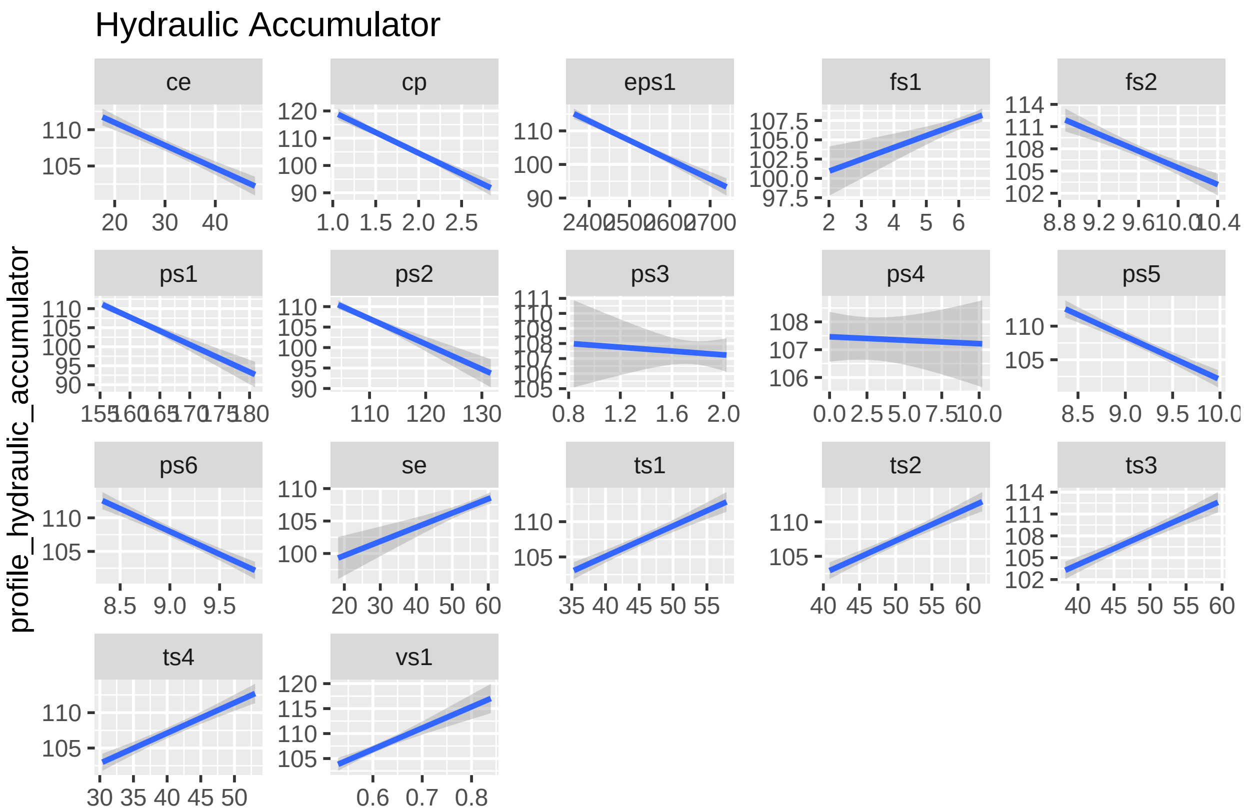
Plots Linear Fit

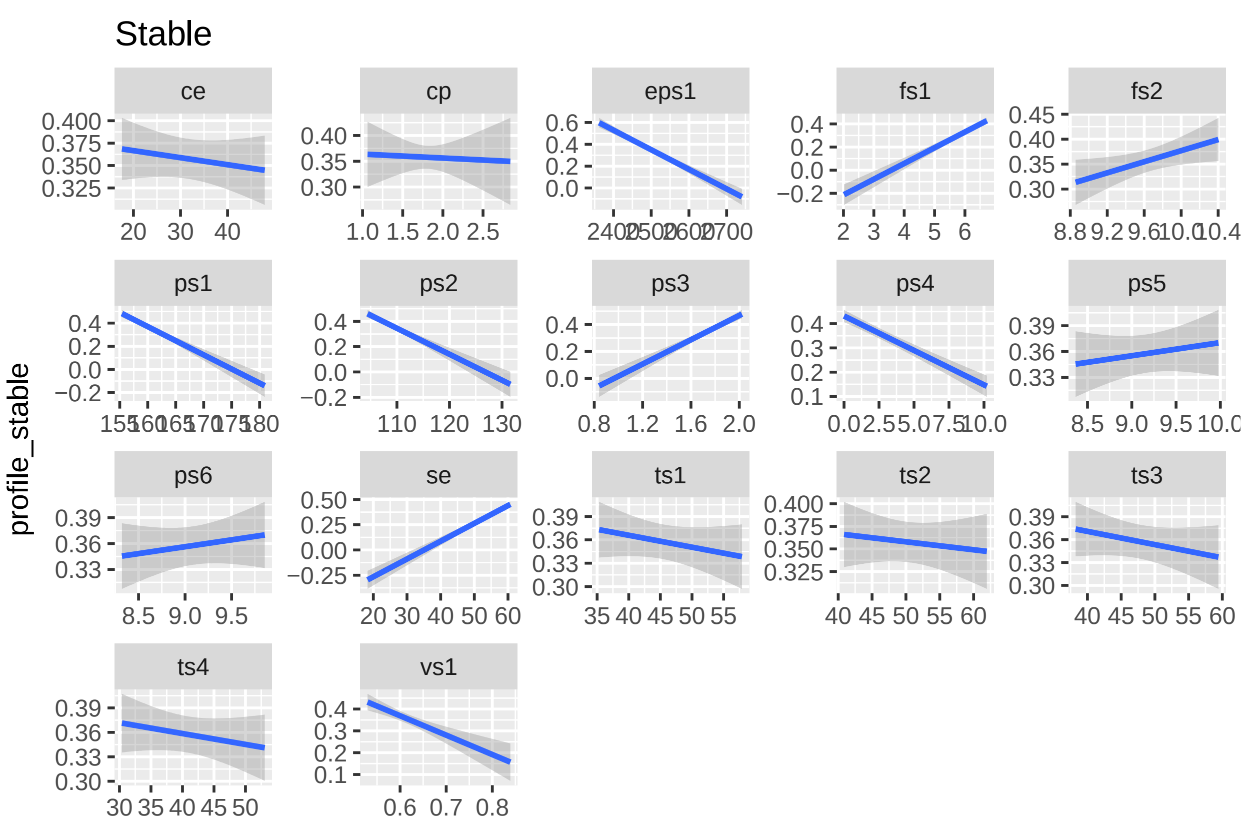
Generally the data shows a good fit for some Sensors/Components. Some sensors are positively correlated (e.g., Cooler: CE, Cooler: CP), while others are negatively correlated (e.g Cooler: TS1, Cooler: VS1). The plot data indicates that a linear fitting model may produce satisfactory results for the following components: Cooler, Hydraulic Accumulator, Internal Pump Leakage and Valve condition.




Valve Condition
ce
cp
eps1
fs1
fs2
ps1
ps2
ps3
ps4
ps5
ps6
se
ts1
ts2
ts3
ts4
vs1
profile_valve_condition

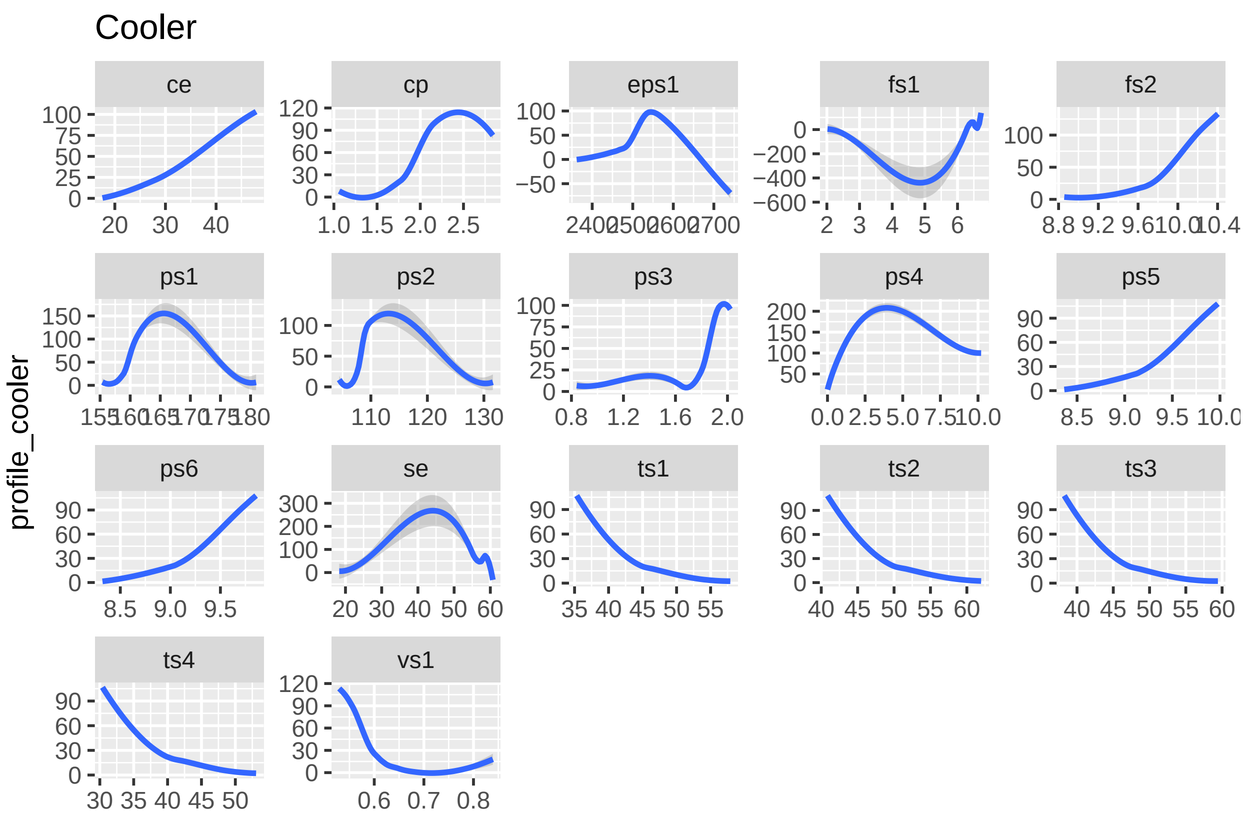

Internal Pump Leakage
ce
cp
eps1
fs1
fs2
ps1
ps2
ps3
ps4
ps5
ps6
se
ts1
ts2
ts3
ts4
vs1
profile_internal_pump_leakage


Hydraulic Accumulator
profile_hydraulic_accumulator
ce
cp
eps1
fs1
fs2
ps1
ps2
ps3
ps4
ps5
ps6
se
ts1
ts2
ts3
ts4
vs1

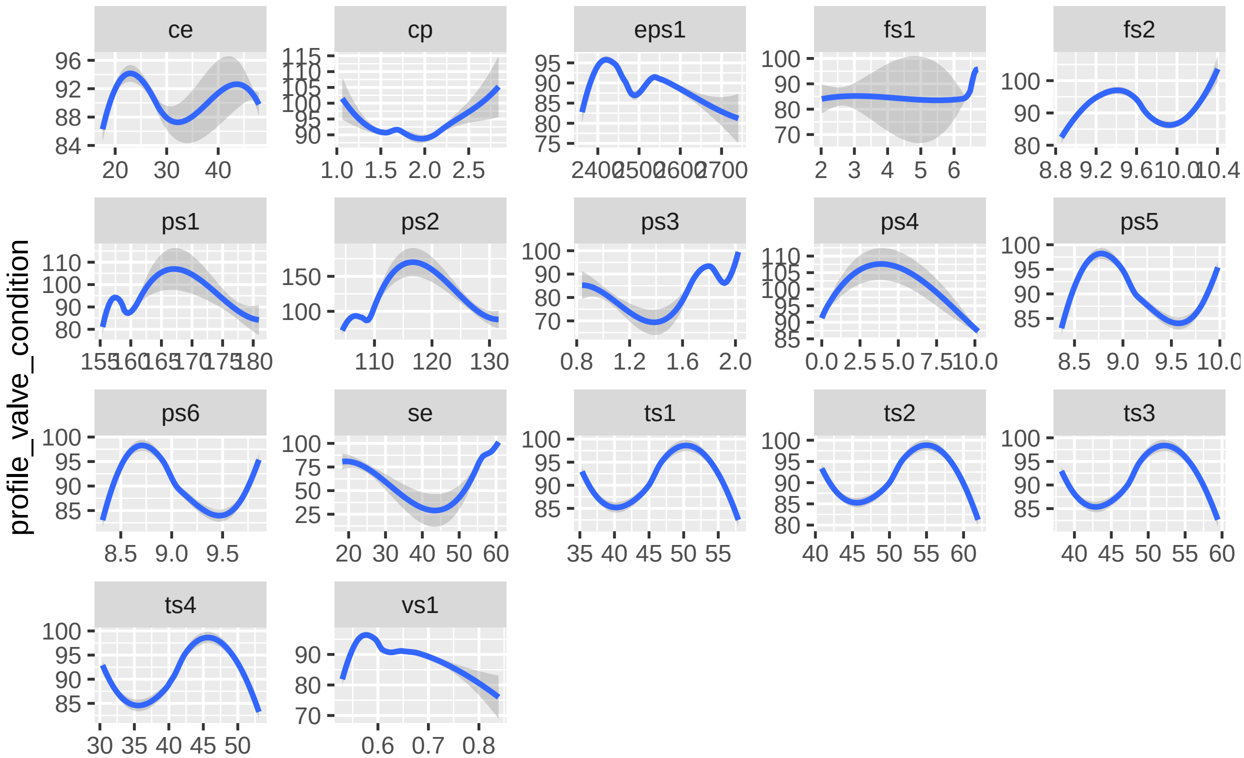

Stable
profile_stable
ce
cp
eps1
fs1
fs2
ps1
ps2
ps3
ps4
ps5
ps6
se
ts1
ts2
ts3
ts4
vs1

Plots Loess Fit

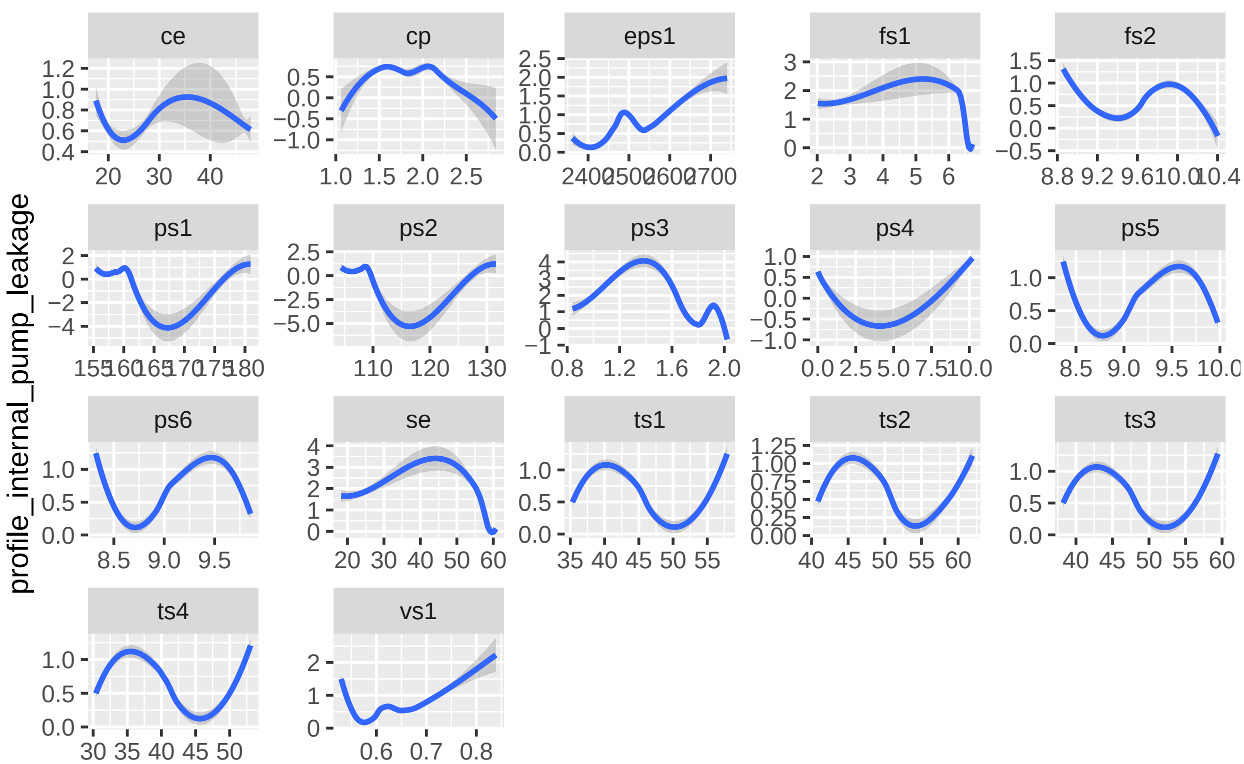
Generally the data shows a good fit for some Sensors/Components. The plot data indicates that a Loess model should produce satisfactory predictive results.

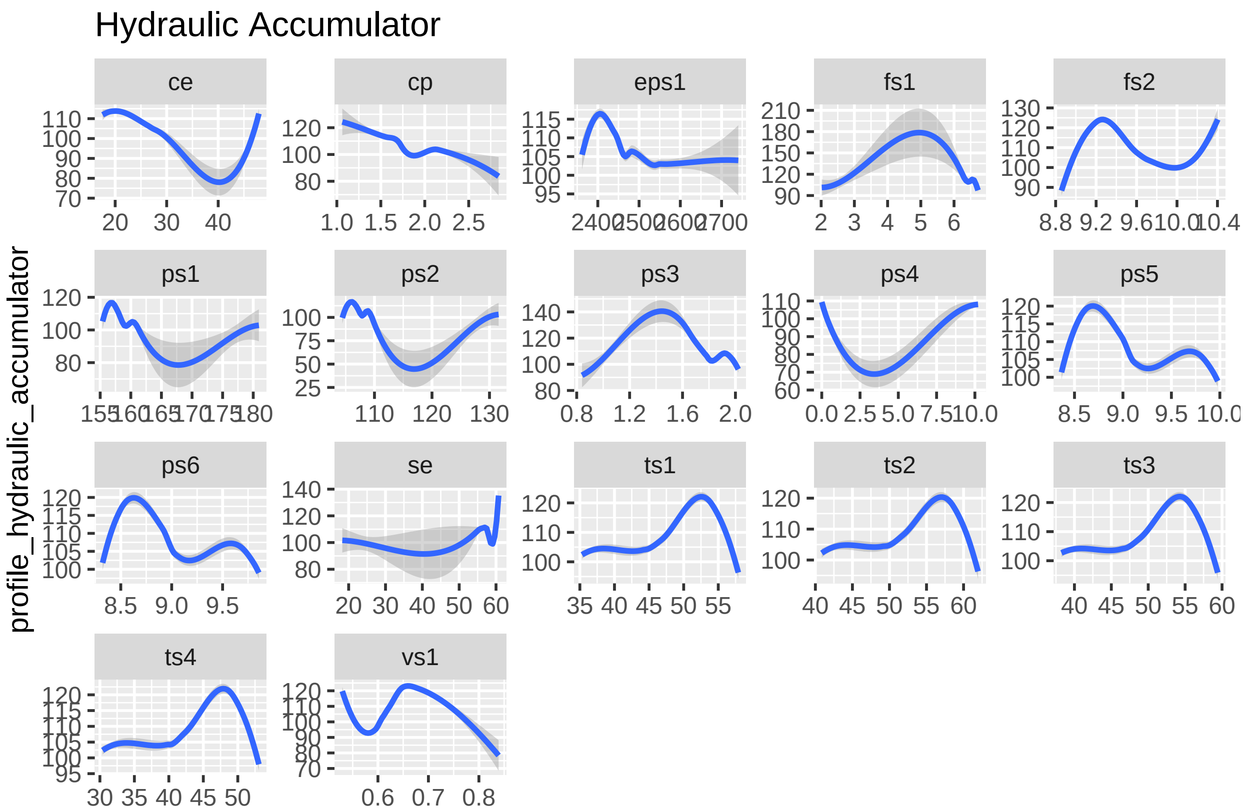


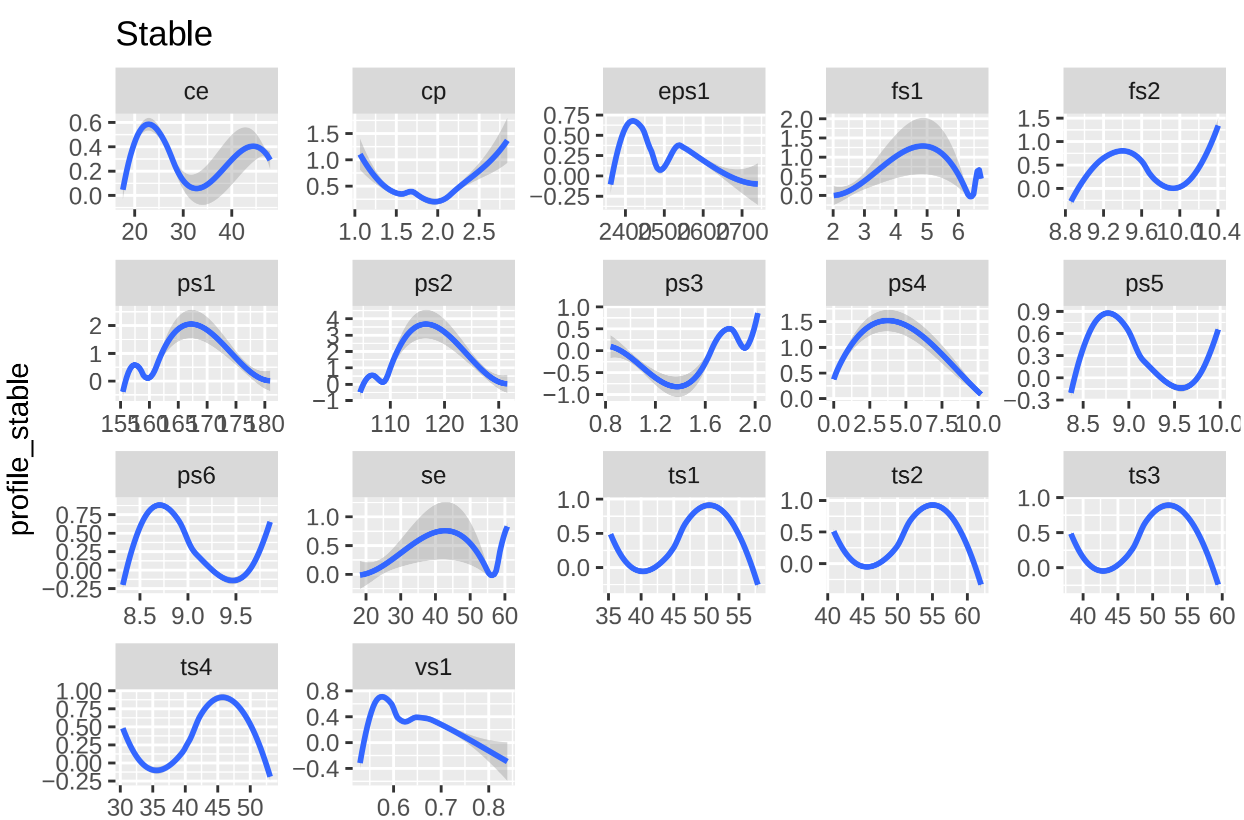
Valve Condition



Internal Pump Leakage




Hydraulic Accumulator
profile_hydraulic_accumulator
ce
cp
eps1
fs1
fs2
ps1
ps2
ps3
ps4
ps5
ps6
se
ts1
ts2
ts3
ts4
vs1


Stable
profile_stable
ce
cp
eps1
fs1
fs2
ps1
ps2
ps3
ps4
ps5
ps6
se
ts1
ts2
ts3
ts4
vs1

Data Models

Background and Model Selection

Data models have been produced for the following models:

- glm: A generalized Linear Mode. A linear regression model that allows for response variables that have error distribution models other than a normal distribution.
- gamLoess: gamLoess is a generalized additive model (GAM) and uses Loess regression. Loess regression is a nonparametric technique that uses local weighted regression to fit a smooth curve through points in a scatter plot. A generalized additive model (GAM) is a generalized linear model in which the linear response variable depends linearly on unknown smooth functions of some predictor variables, and interest focuses on inference about these smooth functions.
- knn: k-Nearest Neighbors, a non-parametric technique. KNN works by finding the distances between a predictor and all the other predictors in the data, selecting the specified number examples (K) closest to the predictor, then votes for the most frequent outcome (or average of the outcomes).
- Rborist: A Random Forest model. Uses decision trees (a series of yes/no questions) asked about our predictors eventually leading to the outcomes. Random Forest combines many of these decision trees together to increase the accuracy of the model.

For each model the following parameters were used:

- Base model
- Pre-Processing Data - Center Predictors: For each predictor, subtract mean from values.
- Pre-Processing Data - Scale Predictors: For each predictor, divide values by standard deviation.
- Tuning: TrainControl and TrainLength. trainControl generates parameters that further control how models are created. tuneLength parameter tells the algorithm to try different default values for the main parameters. Both options attempt at optimising the algorithm by trialling/testing tuneable parameters.

Performance and Outcomes

Performance for each model has been summarized below. The highest performance for each outcome has been highlighted in blue. Where multiple outcomes had the same performance value, then the performance value with the lowest run time has been highlighted.

The data shows that Pre-processing and Tuning has minimal impact on the final models performance criteria, and the most benefit to the analysis is selecting the best model for predictors/outcome.

The optimal models per component is summarized below:

- Cooler: glm
- Valve Condition: gamLoess
- Internal Pump Leakage: glm
- Hydraulic Accumulator: Rborist (with Pre-processing: Center)
- Stable: Rborist (with Pre-processing: Center)

Table 13: Precision

Model	cooler	valve condition	internal pump leakage	hydraulic accumulator	stable
gamLoess (Pre-Process: center)	1.000	0.992	1.000	0.638	0.940
gamLoess (Pre-Process: scale)	1.000	1.000	1.000	0.655	0.945
gamLoess Base	1.000	1.000	1.000	0.648	0.938
gamLoess Pre-Process: Scale, with Tune	1.000	0.992	1.000	0.638	0.940
glm (Pre-Process: center)	1.000	0.992	1.000	0.583	0.978
glm (Pre-Process: scale)	1.000	0.992	1.000	0.583	0.978
glm Base	1.000	0.992	1.000	0.583	0.978
glm Pre-Process: Scale, with Tune	1.000	0.992	1.000	0.583	0.978
knn (Pre-Process: center)	0.995	0.692	0.986	0.819	0.939
knn (Pre-Process: scale)	1.000	0.783	1.000	0.924	0.966
knn Base	0.995	0.692	0.986	0.819	0.939
knn Pre-Process: Scale, with Tune	0.995	0.692	0.986	0.819	0.939
Rborist (Pre-Process: center)	1.000	0.983	0.997	0.965	0.966
Rborist (Pre-Process: scale)	1.000	0.986	0.997	0.961	0.966
Rborist Base	1.000	0.984	0.997	0.965	0.966
Rborist Pre-Process: Scale, with Tune	1.000	0.977	1.000	0.957	0.966

Table 14: Accuracy

Model	cooler	valve condition	internal pump leakage	hydraulic accumulator	stable
gamLoess (Pre-Process: center)	1.000	0.995	1.000	0.565	0.935
gamLoess (Pre-Process: scale)	1.000	1.000	1.000	0.570	0.930
gamLoess Base	1.000	1.000	1.000	0.575	0.920
gamLoess Pre-Process: Scale, with Tune	1.000	0.995	1.000	0.565	0.935
glm (Pre-Process: center)	1.000	0.995	1.000	0.460	0.930
glm (Pre-Process: scale)	1.000	0.995	1.000	0.460	0.930
glm Base	1.000	0.995	1.000	0.460	0.930
glm Pre-Process: Scale, with Tune	1.000	0.995	1.000	0.460	0.930
knn (Pre-Process: center)	0.995	0.670	0.990	0.835	0.930
knn (Pre-Process: scale)	1.000	0.790	1.000	0.925	0.960
knn Base	0.995	0.670	0.990	0.835	0.930
knn Pre-Process: Scale, with Tune	0.995	0.670	0.990	0.835	0.930
Rborist (Pre-Process: center)	1.000	0.980	0.995	0.965	0.975
Rborist (Pre-Process: scale)	1.000	0.980	0.995	0.960	0.975
Rborist Base	1.000	0.975	0.995	0.965	0.975
Rborist Pre-Process: Scale, with Tune	1.000	0.975	1.000	0.955	0.975

Table 15: F1

Model	cooler	valve condition	internal pump leakage	hydraulic accumulator	stable
gamLoess (Pre-Process: center)	1.000	0.995	1.000	0.578	0.956
gamLoess (Pre-Process: scale)	1.000	1.000	1.000	0.580	0.952
gamLoess Base	1.000	1.000	1.000	0.583	0.945
gamLoess Pre-Process: Scale, with Tune	1.000	0.995	1.000	0.578	0.956
glm (Pre-Process: center)	1.000	0.995	1.000	0.464	0.950
glm (Pre-Process: scale)	1.000	0.995	1.000	0.464	0.950
glm Base	1.000	0.995	1.000	0.464	0.950
glm Pre-Process: Scale, with Tune	1.000	0.995	1.000	0.464	0.950
knn (Pre-Process: center)	0.995	0.627	0.988	0.820	0.952
knn (Pre-Process: scale)	1.000	0.772	1.000	0.925	0.972
knn Base	0.995	0.627	0.988	0.820	0.952
knn Pre-Process: Scale, with Tune	0.995	0.627	0.988	0.820	0.952
Rborist (Pre-Process: center)	1.000	0.978	0.995	0.967	0.983
Rborist (Pre-Process: scale)	1.000	0.979	0.995	0.962	0.983
Rborist Base	1.000	0.973	0.995	0.967	0.983
Rborist Pre-Process: Scale, with Tune	1.000	0.974	1.000	0.958	0.983

Table 16: Recall

Model	cooler	valve condition	internal pump leakage	hydraulic accumulator	stable
gamLoess (Pre-Process: center)	1.000	0.997	1.000	0.603	0.972
gamLoess (Pre-Process: scale)	1.000	1.000	1.000	0.611	0.958
gamLoess Base	1.000	1.000	1.000	0.609	0.951
gamLoess Pre-Process: Scale, with Tune	1.000	0.997	1.000	0.603	0.972
glm (Pre-Process: center)	1.000	0.997	1.000	0.491	0.924
glm (Pre-Process: scale)	1.000	0.997	1.000	0.491	0.924
glm Base	1.000	0.997	1.000	0.491	0.924
glm Pre-Process: Scale, with Tune	1.000	0.997	1.000	0.491	0.924
knn (Pre-Process: center)	0.995	0.649	0.990	0.825	0.965
knn (Pre-Process: scale)	1.000	0.789	1.000	0.926	0.979
knn Base	0.995	0.649	0.990	0.825	0.965
knn Pre-Process: Scale, with Tune	0.995	0.649	0.990	0.825	0.965
Rborist (Pre-Process: center)	1.000	0.973	0.993	0.972	1.000
Rborist (Pre-Process: scale)	1.000	0.973	0.993	0.969	1.000
Rborist Base	1.000	0.965	0.993	0.972	1.000
Rborist Pre-Process: Scale, with Tune	1.000	0.970	1.000	0.965	1.000

Table 17: Run Time (min)

Model	cooler	valve condition	internal pump leakage	hydraulic accumulator	stable
glm Base	0.03	0.02	0.02	0.02	0.02
gamLoess Base	8.44	7.49	7.64	5.47	5.22
knn Base	0.06	0.05	0.05	0.05	0.05
Rborist Base	0.59	1.88	0.90	1.79	1.27
glm (Pre-Process: center)	0.03	0.03	0.02	0.03	0.02
gamLoess (Pre-Process: center)	8.47	8.16	8.07	7.89	7.98
knn (Pre-Process: center)	0.08	0.07	0.07	0.07	0.07
Rborist (Pre-Process: center)	0.61	1.90	0.91	1.81	1.30
glm (Pre-Process: scale)	0.03	0.02	0.03	0.02	0.03
gamLoess (Pre-Process: scale)	6.76	5.99	6.45	5.04	5.64
knn (Pre-Process: scale)	0.08	0.07	0.08	0.07	0.07
Rborist (Pre-Process: scale)	0.61	1.90	0.92	1.80	1.29
glm Pre-Process: Scale, with Tune	0.02	0.02	0.02	0.02	0.02
gamLoess Pre-Process: Scale, with Tune	4.85	4.75	4.77	4.72	4.76
knn Pre-Process: Scale, with Tune	0.06	0.06	0.06	0.06	0.06
Rborist Pre-Process: Scale, with Tune	0.78	2.47	1.18	2.40	1.67

Table 18: Final Model using Test/Train Data Set

Model	cooler	valve condition	internal pump leakage	hydraulic accumulator	stable
Model	glm	gamLoess	glm	Rborist	Rborist
Pre-Processing Method	NA	NA	NA	center	center
Precision	1.000	1.000	1.000	0.969	0.966
Accuracy	1.000	1.000	1.000	0.970	0.975
F1	1.000	1.000	1.000	0.971	0.983
Recall	1.000	1.000	1.000	0.976	1.000
Time (min)	0.02	6.85	0.02	1.80	1.30

Table 19: Final Model Using Validation Data Set

Model	cooler	valve condition	internal pump leakage	hydraulic accumulator	stable
Model	glm	gamLoess	glm	Rborist	Rborist
Pre-processing	NA	NA	NA	center	center
Precision	1.000	1.000	1.000	0.979	0.988
Accuracy	1.000	1.000	1.000	0.982	0.987
F1	1.000	1.000	1.000	0.983	0.991
Recall	1.000	1.000	1.000	0.988	0.994
Time (min)	0.02	6.74	0.02	1.80	1.25

Conclusion

The final models used (GLM, gamLoess and Rborist) produced results with excellent Precision, Accuracy and Recall. The models can be used as a tool for prediciting down-time or equipment failure based on sensor data.

Final model and performance criteria is summarised below:

Table 20: Final Model Using Validation Data Set

Model	cooler	valve condition	internal pump leakage	hydraulic accumulator	stable
Model	glm	gamLoess	glm	Rborist	Rborist
Pre-processing	NA	NA	NA	center	center
Precision	1.000	1.000	1.000	0.979	0.988
Accuracy	1.000	1.000	1.000	0.982	0.987
F1	1.000	1.000	1.000	0.983	0.991
Recall	1.000	1.000	1.000	0.988	0.994
Time (min)	0.02	6.74	0.02	1.80	1.25

Analysis of additional models was limited by computer processing power, as a function of run-time. The analysis/evaluation of additional models was limited by computer processing capabilities, with total run time exceeding available computing hours.

Further work could use updated hardware to evaluate additional training models or optimisation parameters, including the use of techniques such as Ensembles (averaging of multiple models), or TuneGrids. In addition, an investigation into selectively choosing sensors as predictors for each model to improve the predicted outcomes should be performed. It would be informative to determine the minimum number and type of sensors required to produce accurate models.

Appendix 1: Total Code Run Time

Table 21: Total Code Run Time (min

x
180.2