

MASTER

Predicting the number of truck repairs using logged vehicle data

Lemmen, R.M.D.

Award date:
2019

[Link to publication](#)

Disclaimer

This document contains a student thesis (bachelor's or master's), as authored by a student at Eindhoven University of Technology. Student theses are made available in the TU/e repository upon obtaining the required degree. The grade received is not published on the document as presented in the repository. The required complexity or quality of research of student theses may vary by program, and the required minimum study period may vary in duration.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain



TECHNICAL UNIVERSITY EINDHOVEN
FACULTY OF INDUSTRIAL ENGINEERING AND INNOVATION SCIENCES
IN COLLABORATION WITH DAF TRUCKS N.V. COSTING & ANALYSIS
DEPARTMENT

Predicting the number of truck repairs using logged vehicle data

Supervisors:

Prof.dr.ir. U. Kaymak	Eindhoven University of Technology
Dr. Y. Zhang	Eindhoven University of Technology
Prof.dr.ir. I.J.B.F. Adan	Eindhoven University of Technology
M. Maaskant MSc, MSc	DAF Trucks N.V., Costing & Analysis
Ing. E. Koreman	DAF Trucks N.V., Costing & Analysis

Author:

R.M.D. Lemmen (Remco) BSc ID: 0811245

January 21, 2019

Note that due to confidentiality reasons, the axes details of the figures containing claim rates have been left blank or have been scaled and the repair types in the figures containing the top most common repairs have been generalized.

Executive summary

This graduation project is executed at DAF trucks N.V. (DAF), in collaboration with DAF's costing & analysis department. DAF is a leading truck manufacturer whose core activity is the development and production of light-, medium- and heavy-duty commercial vehicles.

Problem statement

Along with the sale of the trucks, DAF provides its buyers with multiple warranty, maintenance and repair contract options. In order to offer their customers competitive prices for these contracts, accurate prediction of repair and maintenance costs of the trucks is of vital importance for DAF. Currently, the repair and maintenance (R&M) costs for DAF's trucks are predicted based on generalized averages over their whole fleet of trucks and rules of thumb for certain truck features (e.g. region where it is sold and cargo that it will carry). It is unclear how much these factors actually contribute to the total R&M costs. Furthermore, the current calculation model is a black box for DAF. There is little knowledge about the model design choices and the reasoning behind the underlying relations and calculations. As a result, reasoning behind the cost formulations themselves is not known. This results in suboptimal cost calculations, which are notable through the difference between predicted costs and actual costs of their contracts. As DAF wants to stay competitive and increase its market-share where possible, it is of importance that they can offer their customers optimal contract prices while maintaining their profit margins. This is why DAF is motivated to improve the prediction of the number of repairs and corresponding costs for their trucks. Better predictions will help them to improve the cost allocations per R&M contract that they sell to their customers, subsequently enabling them to provide better contract prices and thus improve their market position. Recently, DAF has installed a new feature on their trucks, named *DAF Connect*. Their trucks have been fitted with all kind of sensors that collect real-time data on the condition, usage and state of operation of the trucks. The collected sensory data is transmitted to DAF in 5-minutes intervals. Examples of measurements are the recording of engine temperature, intake air pressure and engine rpm. Besides the interval data, aggregated trip information is sent to DAF each time that a truck has completed a trip. DAF wants to use these data to predict the number of repairs and corresponding costs for their trucks, based on the characteristics (e.g. type, engine and age) and usage of the trucks (e.g. driving style and truck condition). At the time of research it was still unclear how these data could be used for their predictions and thus, the following research question is answered:

How can the number of truck repairs be predicted based on telemetry truck data and truck usage information?

Research approach

To answer the research question, four prediction methods have been developed. Random Forests and Neural Networks have been used as literature research showed that they are often used to predict future state, Remaining Useful Life and machine operating conditions. Although often providing good modeling performance, this comes at the cost of limited decision rule interpretability (Negnevitsky, 2005). To compare their performance to some relatively simple models which can be interpreted easily, logistic regression models and Decision Trees have been constructed as well.

Together with experts at DAF, there has been decided to construct binary classification models. The reason for this is that only a limited set of trucks is available for analysis (as DAF Connect is rather new) while many different types of repairs exist. As a result, there is not enough information available to predict the exact number of repairs or the associated costs (modeling tests showed very poor performance). Instead, a binary decision is made which predicts if a truck has more or less than the average number of repairs over the time horizon under consideration. Trucks with less than the average number of repairs are labeled 0, while trucks having more than the average number of repairs are labeled 1. With this classification task, there can be derived which trucks require more intensive repair and maintenance and which trucks do not. This can subsequently be used to substantiate R&M contract costs and identify risk vehicles regarding R&M costs.

The models have been developed according to the six phases of the CRISP-DM framework, which are Business understanding, Data understanding, Data preparation, Modeling, Evaluation and Deployment respectively. During the Business understanding phase, a literature study was executed to gain insights into the available models and methods for the problem at hand. Furthermore, expert knowledge was used to gain insights in the business objectives for DAF and the determination of the data mining goals. In the data understanding phase, the data has been collected from the various data-sources at DAF after which an initial exploration has been made and the quality of the data has been verified and reported. This is done using both visual (e.g. boxplots and barcharts) and quantitative methods (e.g. numerical summaries). Subsequently the data preparation has been performed. First, the correct data had been selected by excluding data of all trucks that fell out of the scope of this research. Secondly, various feature extraction operations have been performed in order to retrieve useful features from the available telemetry data. Lastly, various data cleaning, construction and formatting tasks have been executed to provide a clean dataset with the appropriate formatting that could be used for modeling. In the modeling phase, the prediction models have been created and optimized (e.g. hyper-parameter optimization and feature selection have been applied). In the evaluation phase, the results of the final models have been evaluated and compared on performance, usefulness for DAF and potential for future research. Finally, in the deployment phase this report was delivered and the findings were presented at (and reviewed with) both the company and the university.

Conclusions and recommendations

Conclusions

The prediction models showed that there was some predictive power in the available data. However, the performance of the models was limited. This was mainly caused by the fact that the available data only covered the early life of DAF's trucks, in which not many repairs actually occurred. Furthermore, the repairs that did occur were of many different origins, preventing the models from deriving robust patterns for specific or common repairs. Nonetheless, a combination of features on driving behaviour, truck status and some passive features showed that the telemetric data had potential regarding the prediction of repairs. The random forest showed the most potential for repair predictions on DAF's trucks, it had the most consistent results of the compared models. It reached an accuracy of 64%, precision score of 67% and recall score of 64%. Lastly, it provided some insights into the features that it selected based on the derived feature importances from the model.

Each of the models selected their own set of features for the prediction of expected repairs. However, some features were used by multiple models and therefore, a form of feature importance validation has been done by counting the occurrence frequency of the variables that have been used by the best performing models. This resulted in a top 10 of most important features. With the exception of one truck configuration feature (*Asset type*), the top 10 of most important features is comprised of features that have been derived from DAF Connect, indicating its potential for the prediction of expected repairs on DAF's trucks. In short, these features can be divided into three categories, being driving related features, truck status features and indirect features. From the top 10 of most important features, the features *Harsh braking duration*, *Tachograph speed*, *DPA anticipation scores* and *Max throttle duration* are directly influenced by the driver of the truck and thus categorized as driving related features. The features *Engine oil pressure* and *Engine intake air-pressure* give information about the truck's status and are categorized as truck status features. Lastly, the features *Asset type*, *Ambient air temperature* and *Trip distance* are features that are not directly influenced by the truck or the driver (assuming that a driver doesn't choose the trip distance, but receives this information from an external entity/company) and labeled as passive features. In conclusion these features provide the best information about the expected number of truck repairs.

Recommendations

There has been shown that the features that have been derived from DAF Connect show a better potential in the prediction of the numbers of repairs than the currently used truck specifications and contract details. Analysis of the constructed models showed that the Connect features were consistently favored

above the currently used features. However, the current amount of available Connect data is too limited to make useful predictions for individual trucks. Therefore it is recommended for DAF to wait until more data is available before revisiting the problem of predicting the exact number of repairs or its costs (i.e. at least two more years, as during that period most of the repairs occur).

It is recommended to DAF to focus on more narrowed down problems than the prediction of the total number of repairs over a given time horizon (at least for individual trucks). Instead, it is recommended for DAF to focus on the prediction of specific, expensive repairs before they actually happen. Extensive literature is available on the prediction of Remaining Useful Life, Future state and Health status of specific components or machines. For DAF, a prime example are the turbo and battery. They tend to fail rather regularly and are costly (at least the turbo) to replace. With the right sensors in place, these failures could be detected before they happen, allowing for the appropriate preventive actions to take place. It is of vital importance that DAF thinks about the implementation of the right sensors to do so now, at this point in time, because information that can be derived with these sensors only becomes useful after months or even years of data collection. This is because currently available prediction methods need to learn from the past (or at least have reference values) in order to come up with meaningful predictions.

Furthermore, it is recommended to DAF to investigate the possibility to incorporate usage information as derived in this research into their current cost calculation methods when more data has become available. This information could be used to correct cost expectations 'on the go'. I.e. the driving behavior of the connected trucks can be monitored while its operating and based on this, the expected costs and/or number of repairs for specific trucks can be adjusted in the forecasts. Subsequently, specific discounts could be offered to customers that prove to have a beneficial usage profile, either when they buy additional trucks or during the agreed contract period itself. This would require a change in the current business processes and thus should be investigated timely.

Lastly, some general recommendations regarding the data have been given. Much information that was needed for this research was hidden in different databases and different sources. In order to obtain it, many different and specific queries/scripts had to be written. Afterwards, they had to merged requiring another set of manual operations. This made the collection of data time consuming and prevents possibilities for automated analyses and data collection. It is therefore recommended for DAF to standardize their data and documents in a single (cloud) location (which is currently explored by DAF), allowing for much faster analyses and automation of data collection and processing.

Contents

Executive Summary	ii
List of Figures	ix
List of Tables	xi
1 Introduction	1
1.1 DAF trucks N.V.	1
1.2 Problem statement	1
1.2.1 Relevance	1
1.3 Research question	2
1.3.1 Research sub-questions	2
1.4 Research methodology	3
1.4.1 Available data	3
1.4.2 Methodology	3
1.4.3 CRISP-DM methodology	4
1.5 Data mining goals and deliverables	5
1.6 Scope	5
1.7 Thesis outline	6
2 Literature review	7
2.1 Predictive maintenance	7
2.1.1 Remaining useful life prediction	7
2.1.2 Future state prediction	8
2.1.3 Condition prediction	9
2.2 Time series analysis	10
2.2.1 Traditional (time series) warranty and repair analysis	10
2.2.2 Feature based time series analysis	10
2.2.3 Pattern based time series analysis	11
2.3 Modeling techniques	12
2.3.1 Logistic Regression	13
2.3.2 Decision Trees	13
2.3.3 Random Forest	14
2.3.4 MLP Neural Network	14
2.4 Previous work	15
2.5 Discussion	15
2.6 Conclusion	16
3 The company and the available data	17
3.1 The company	17
3.2 Products	17
3.2.1 Truck type	17
3.2.2 Chassis type	18
3.3 Service and aftersales	18
3.3.1 Warranty and R&M contracts	18
3.3.2 MultiSupport Calculator	19
3.3.3 Costing & analysis	19
3.4 Data sources	21
3.4.1 CCM truck data	21
3.4.2 Mi claim database	21
3.4.3 Connect database	21
3.5 Descriptive analytics of the CCM Database	22
3.5.1 Graphical analysis	28

3.6	Descriptive analytics of the Mi Database	30
3.6.1	Graphical analysis	35
3.6.2	Repair types	36
3.6.3	Assumptions on the repair data	38
3.7	Descriptive analytics of the connect data.	38
3.7.1	Trip data	38
3.7.2	Snapshot data	41
3.8	Chapter Summary	44
4	Data preparation	45
4.1	Selecting the relevant trucks and data	45
4.2	Classifying the trucks based on repairs	45
4.3	Data cleaning	46
4.3.1	Outliers	46
4.3.2	Missing values	46
4.3.3	Inconsistencies and noise	47
4.4	Feature derivation and extraction	48
4.4.1	Deriving truck usage features from the datasets	48
4.4.2	Fuzzy histograms and feature extraction	48
4.4.3	Data integration and formatting	51
4.5	Final dataset	51
4.6	Chapter summary	52
5	Repair prediction models	53
5.1	Modeling setup	53
5.1.1	Scoring metrics	53
5.1.2	Feature selection	54
5.1.3	Filtering methods	55
5.1.4	Wrapping methods	56
5.1.5	K-Fold cross validation and parameter optimization	57
5.1.6	Experimental setup	58
6	Modeling results	60
6.1	Feature selection	60
6.2	Logistic Regression	61
6.2.1	Base models	62
6.2.2	Extended models	63
6.3	Decision Tree	65
6.3.1	Base models	65
6.3.2	Extended models	67
6.4	Random Forests	68
6.4.1	Base models	69
6.4.2	Extended models	70
6.5	MLP-NN results	71
6.5.1	Base models	71
6.5.2	Extended models	72
6.6	Summary of results and relevant features	73
6.6.1	The most suitable models to predict repairs	73
6.6.2	Relevant features for repair predictions	73
6.7	Comparison to the work of Goudsmit (2018)	75
7	Conclusions and recommendations	77
7.1	Research conclusions	77
7.2	Recommendations	79
7.3	Limitations	80
7.4	Contribution to literature	81
7.5	Future research	81

8	Appendices	82
A	Overview of the available CCM truck data at DAF.	82
B	Overview of the available Mi claim data at DAF.	85
C	Overview of the available connect interval data at DAF.	87
D	Numerical summary of the snapshot data	95
E	Overview of the tasks that belong to each phase of the CRISP-DM model.	98
F	Distribution of contracts for the connected trucks	99
G	Fleet information	100
H	Overview of box-plots for the Connect Data.	101
	H.1 Box-plots of the trip database features.	109
I	The 1.5 IQR for the Connect Data	112
J	Fuzzy bins experiment results	117
K	Similarly performing Logistic Regression model beta coefficients	118
L	Average accuracy over 10 different modeling runs	119
M	T-test on model improvements using Fuzzy Bins	121
N	The derived truck usage features over the first month of operation.	124
O	The features over which fuzzy bins have been used for feature extraction.	125
P	An example of an instance in the final datasets.	126
Q	The searched hyperparameter space for each of the models	135
R	The optimal hyperparameters for each of the models.	136
S	List of abbreviations	137

List of Figures

1	Phases of the CRISP-DM process model for data mining (Wirth, 2000)	5
2	An example of a decision tree	13
3	Simple schema of a MLP-NN with 1 input layer, 1 hidden layer and 1 output layer.	15
4	DAF's newest LF, CF and XF series respectively.	17
5	DAF's available axle configurations.	18
6	DAF's available Multi-support packages (DAF, 2018c).	19
7	The distribution of trucks types for connected trucks (21-08-2018).	28
8	The distribution of chassis types for connected trucks (21-08-2018).	28
9	The number of months that the connected trucks are in service (21-08-2018).	29
10	The distribution of connected trucks in Europe (21-08-2018).	29
11	The distribution of total number of repair claims per truck (21-08-2018).	35
12	Boxplots of the R&M claims (a), warranty claims (b) and total number of claims (c) per months in service (2018).	36
13	Distribution of contracts on which repair claims are made (21-08-2018).	36
14	The top 10 most occurring repairs on both the driveline (a) and non-driveline (b).	37
15	The top 10 most occurring repairs on both the driveline (a) and non-driveline (b) (21-08-2018).	37
16	The top 10 of most expensive repairs from 01-04-2017 up to and including 21-08-2018.	38
17	Example of temperature outlier analysis using predefined baseline boundaries.	46
18	A crisp histogram (left), compared to a fuzzy histogram (right) with overlapping membership functions (van den Berg et al., 2004).	49
19	The fuzzy histogram for the distribution of truck speed measurements.	50
20	Confusion matrix for a binary classification problem.	53
21	The SFS algorithm.	57
22	The RFE algorithm.	57
23	10-fold cross validation where E is the performance metric score.	58
24	The general experimental setup for the different prediction models.	59
25	The performance of MI (a), RFE(b) and SFS (c) on one of the used random forest models.	61
26	Cross-validated 8 months ahead Logistic Regression accuracy scores plotted against the number of features selected by the SFS.	63
27	The beta coefficients for te best performing base Logistic Regression model.	64
28	The beta coefficients for te best performing extended Logistic Regression model.	65
29	Results of the Decision Tree model performance using SFS and 11 months of data.	66
30	Feature importance for the best 11 months ahead decision tree with SFS.	67
31	Feature importance for the best performing extended decision tree.	68
32	Results of the Random Forest validation performance using SFS and 11 months of data.	69
33	Feature importance for the best performing base random forest.	70
34	Feature importance for the best performing extended random forest.	71
35	SFS applied to one of the MLP-NN's.	72
36	The learning curves for the 11 months ahead predictions	79
37	Overview of the tasks that belong to each phase of the CRISP-DM model (Wirth, 2000).	98
38	Distribution of the type of warranty contracts that are sold with the connect trucks.	99
39	Distribution of the type of R&M contracts that are sold with the connect trucks.	99
40	The number of months that connect trucks are in service divided per truck class.	100
41	The production of connected trucks per month, divided over product type (model).	100
42	Snapshot data histograms (1)	101
43	Snapshot data histograms (2)	102
44	Snapshot data histograms (3)	103
45	Snapshot data histograms (4)	104
46	Snapshot data histograms (5)	105
47	Snapshot data histograms (6)	106
48	Snapshot data histograms (7)	107
49	Snapshot data histograms (8)	108
50	Trip data histograms (1)	109
51	Trip data histograms (2)	110

52	Trip data histograms (3)	110
53	Trip data histograms (4)	111
54	Trip data histograms (5)	111
55	The beta coefficients for one of the found Logistic Regression models.	118

List of Tables

1	Overview of the deliverables per CRISP-DM phase.	6
2	The features that are currently used for the price calculations of R&M contracts.	20
3	Overview of the CCM truck data groups and the corresponding features.	23
4	Quantitative summary of all numerical variables in the CCM dataset.	24
5	Overview of the CCM truck data variables containing missing values.	25
6	Overview of categorical variables in the CCM truck data that contain mostly identical values.	26
7	Overview of the irrelevant contract data features from the CCM database.	27
8	Overview of the redundant features and their duplicate in the CCM database.	27
9	Overview of the Mi data groups and the corresponding features.	31
10	Quantitative summary of all numerical variables in the Mi dataset.	32
11	Overview of the Mi truck data variables containing missing values.	33
12	Overview of the irrelevant and redundant features from the Mi database.	34
13	Overview of the trip data.	39
14	Quantitative summary of the numerical trip data.	40
15	Overview of the snapshot data.	42
16	Overview of the variables with missing values in the Snapshot data.	43
17	The number of rows (instances) per subset for both datasets	59
18	Logistic Regression base model results	62
19	Modeling results for the base Logistic Regression model with SFS.	63
20	Modeling results for the extended Logistic Regression model without SFS.	64
21	Modeling results for the extended Logistic Regression model with SFS.	64
22	The results for the base models without SFS.	65
23	The results for the base models with SFS.	66
24	The results for the extended models without SFS.	67
25	The results for the extended models with SFS.	68
26	The results for the base models without SFS.	69
27	The results for the base models with SFS.	69
28	The results for the extended models without SFS.	70
29	The results for the extended models with SFS.	70
30	The results for the base models without SFS.	71
31	The results for the base models with SFS.	72
32	The results for the extended models without SFS.	72
33	The results for the extended models with SFS.	73
34	The 10 most used features across the best performing models	74
35	The claim rate per country, where the claim rate is the average number of claims per truck.	75
36	Comparison of the model performance (accuracy) of the work of Goudsmits (2018) and this research.	76
37	Overview and explanations of the available CCM truck data.	82
38	Overview and explanations of the available Mi claim data.	85
39	Overview and explanation of the snapshot data.	87
40	Variable explanations of the trip data.	91
41	Variable explanations of the trigger data.	93
42	Quantitative summary of the numerical snapshot data.	95
43	The 1.5 IQR boundaries for the snapshot data.	112
44	The cross-validated experimental model performance using different sizes of bins	117
45	The average accuracy and it's standard deviation for 10 runs of the Logistic Regression modeling process.	119
46	The average accuracy and it's standard deviation for 10 runs of the Decision Tree modeling process.	119
47	The average accuracy and it's standard deviation for 10 runs of the Random Forest modeling process.	120
48	The average accuracy and it's standard deviation for 10 runs of the MLP-NN modeling process.	120

49	Results on one sided t-test for increase of accuracy for the Logistic Regression models based on 10 different runs/datasplits	122
50	Results on one sided t-test for increase of accuracy for the Decision Tree models based on 10 different runs/datasplits	123
51	Overview of the derived truck usage features over the first month of operation.	124
52	The features over which fuzzy bins have been used for feature extraction.	125
53	Example of an instance of the final datasets.	126
54	The optimal parameters for the decision tree models.	136
55	The optimal parameters for the random forest models.	136
56	The optimal parameters for the MLP-Neural Network models.	136
57	The list of abbreviations.	137

1 Introduction

1.1 DAF trucks N.V.

This graduation project is executed at DAF trucks N.V. (DAF), in collaboration with DAF's costing & analysis department. DAF is a leading truck manufacturer whose core activity is the development and production of light-, medium- and heavy-duty commercial vehicles.

1.2 Problem statement

Along with the sale of the trucks, DAF provides its buyers with multiple warranty, maintenance and repair contract options. In order to offer their customers competitive prices for these contracts, accurate prediction of repair and maintenance costs of the trucks is of vital importance for DAF. Currently, the repair and maintenance (R&M) costs for DAF's trucks are predicted based on generalized averages over their whole fleet of trucks and rules of thumb for certain truck features (e.g. region where it is sold and cargo that it will carry). It is unclear how much these factors actually contribute to the total R&M costs. Furthermore, the current calculation model is a black box for DAF. There is little knowledge about the model design choices and the reasoning behind the underlying relations and calculations. As a result, reasoning behind the cost formulations themselves is not known. This results in suboptimal cost calculations, which are notable through the difference between predicted costs and actual costs of their contracts. As DAF wants to stay competitive and increase its market-share where possible, it is of importance that they can offer their customers optimal contract prices while maintaining their profit margins. This is why DAF is motivated to improve the prediction of the number of repairs and corresponding costs for their trucks. Better predictions will help them to improve the cost allocations per R&M contract that they sell to their customers, subsequently enabling them to provide better contract prices and thus improve their market position.

Recently, DAF has installed a new feature on their trucks, named *DAF Connect*. In April of 2017 this application was considered to be correctly implemented (matured). Since then, trucks have been fitted with all kind of sensors that collect real-time data on the condition, usage and state of operation of the trucks. The collected sensory data is transmitted to DAF in 5-minutes intervals. Examples of measurements are the recording of engine temperature, intake air pressure and engine rpm. Besides the interval data, aggregated trip information is sent to DAF each time that a truck has completed a trip. A full list of available data from both sources can be found in Appendix C. DAF wants to use these data to predict the number of repairs and corresponding costs for their trucks, based on the characteristics (e.g. type, engine and age) and usage of the trucks (e.g. driver style, road type and cargo). This can then be used to define the costs of R&M contracts more accurately. At the time of research, the cost predictions of a maintenance or service contract were based on expert knowledge and general averages only (as explained above). Therefore, the connect data has a lot of potential and can help DAF to identify and accurately analyze the maintenance and repair costs per truck.

1.2.1 Relevance

The manufacturing industry has to deal with an increasing global competition on product quality and production costs (Brettel et al., 2014). To handle this, manufacturing companies transform into integrated networks where virtualization provides real-time access to relevant product and production information for all participating entities (Brettel et al., 2014). Furthermore, "Cyber-Physical System-based manufacturing and service innovations are two inevitable trends and challenges for manufacturing industries" (Lee et al., 2014). Big-data and connectivity are becoming an increasingly important part of modern manufacturing industry. Nonetheless, many manufacturing systems are not yet ready to manage big data as they lack smart analytic tools and knowledge (Lee et al., 2013). Now, DAF is making its first steps towards industry 4.0 with the implementation of *Connect data* in their trucks. The knowledge that is obtained by this system can be used for Condition Based Maintenance, which is a trending topic among manufacturing companies (Lee et al., 2014).

However, DAF does not only want to apply Condition Based Maintenance (CBM), but wants to make better predictions of individual truck maintenance costs as well. For such problems, probability theory, reliability theory and renewal processes are widely studied and put into practice (Hartzell et al., 2011). In these studies, failure distributions and hazard rates are linked to parts, which are subsequently used to predict the number of necessary preventive and corrective repairs that occur over certain maintenance intervals (e.g. Weibull distributions are widely used as they provide a good fit with data in many applications) (Arts, 2017). These kind of studies have been applied at DAF as well. The use of big-data and machine learning for these predictions however, is a new field of application that has not yet been implemented at DAF. R&M cost calculations are part of Product Life-cycle Management (PLM) (Lee et al., 2008) and although big data is used for PLM to some degree, its research is "still far from enough because many promising 'Big Data' applications remained undeveloped yet" (Li et al., 2015). There is a lack of academic studies of Big-Data in PLM while being vital for fast Time-To-Market, quality, cost reduction, flexibility and service (Li et al., 2015).

Machine learning prediction methods have been applied to vehicle time-series data (Frisk et al., 2014) and warranty data (Prytz et al., 2015) before. Examples are the calculation of a vehicle's Remaining Useful Lifetime (RUL) and pattern detections in machine equipment using Random Forests (Prytz et al., 2015). However, they all focus on the detection of failures shortly before they occur, while this research focuses on the prediction of the number of failures that will occur over a predefined, longer time period. Research on (lifetime) maintenance contract costs have been done as well, but often rely on statistical models, discrete event simulations and failure rate analysis (Jackson and Pascual, 2008) (Wu and Akbarov, 2011) (Wu, 2012). Machine Learning techniques such as Artificial Neural Networks (Rohani et al., 2011) and Neuro-Fuzzy approaches (Chinnam and Baruah, 2004) have also been used to predict maintenance costs, but just as in traditional reliability engineering, they focus on wear and failure rates instead of detailed usage data such as available at DAF. The research proposed in this document tries to provide a basis for DAF in their development of trucks towards an industry 4.0 automotive product. It aims to deliver a prediction tool for the maintenance and repair of individual DAF trucks based on the usage data that is retrieved from said trucks, providing DAF with a new approach to R&M predictions.

1.3 Research question

DAF has stated that, in order to stay competitive, maximize profits and simultaneously provide customers with attractive contract prices, it is essential to predict truck repairs as accurately as possible. This drives the desire of DAF to continuously improve the understanding of factors that contribute to the number of repairs on their trucks. The arrival of DAF Connect (telemetry truck data) presented DAF with new business opportunities regarding the prediction of truck repairs. It sparked their interest in the development of prediction models using the DAF Connect data in order to gain insights in the effects of truck usage data on the prediction of repairs. Therefore, the following research question is formulated:

How can the number of truck repairs be predicted based on telemetry truck data and truck usage information?

1.3.1 Research sub-questions

To help answer the research question above, a set of deliverables and associated sub questions has been formulated:

1. A definition of current repair predictions and cost calculations.
 - What cost/repair data is available?
 - Which factors are currently used to predict repairs?
2. An overview of available data and variables.
 - Which variables can be extracted from the available data?
3. A model for the prediction of truck failures over a given period.

- Which prediction method is most suitable for the problem at hand?
4. Evaluation of the models' prediction power.
 - What is the performance of the most suitable models?
 - What variables provide information about the expected number of truck repairs?
 5. Final document and presentation.

1.4 Research methodology

1.4.1 Available data

To get a better understanding of the project, a short overview of the available data is given. The data consists of general truck information and sensory data that DAF's trucks send to the DAF headquarters wirelessly in regular time intervals of five minutes. This raw data is stored in an off-site data warehouse where the data is separated and stored into the following datasets:

- General truck information data
- Trip data
- Trigger data
- Snapshot data

The *General truck information data* is stored in the so called Customer Contract Management (CCM) database. For every produced truck it contains specific truck setup information as well as sales and contract details. In fact, current contract cost predictions are based on the information in this dataset. An overview of the variables in the dataset is found in table 2.

The *Trip data* collects data for each trip that a truck makes (where a trip is considered to be the time between the start and shutdown of the engine). It aggregates the data from the trip and sends it to the data-warehouse as a single instance. Valuable information such as the total brake duration, harsh brake duration, max throttle duration and fuel consumption per trip are found in this dataset. In total, 46 variables are recorded per trip. A week of trip data constitutes to a size of 0.05GB of hard drive memory.

The *Trigger data* records and stores sensory data from the truck, each time that a message is triggered inside the truck. This allows for data analysis at the exact time of fault occurrence on DAF's trucks. Note that triggers are not just fault occurrences. For example, starting and stopping the engine is also considered to be a trigger. The trigger data set records 56 variables per entry and contains more detailed information compared to the trip data. I.e. the current gear, engine load, coolant temperatures and many more are recorded at the time of an event occurrence. A week of trigger data constitutes to a size of approximately 3.00 GB of hard drive memory.

The last set contains the *Snapshot data* (also called 5-minute data). It is the most comprehensive set of data, as it contains sensory data for each truck in five minute intervals. Data is sent at each interval, regardless of the condition and state of the truck. 108 variables are recorded per truck every five minutes. The total data size for a week's worth of data constitutes to a size of approximately 2.00 GB of hard drive memory.

1.4.2 Methodology

The research that is proposed in this document concerns a data mining task. The cross-industry standard process for data mining (CRISP-DM) methodology is often used for these type of problems (Azevedo and Santos, 2008). Although other methods exist as well (e.g. Sample, Explore, Modify, Model, and Assess (SEMMA) and Knowledge Discovery in Databases (KDD)), they do not differ that much. Slightly different terms and names for the research stages are used, but in the end they boil down to (more or less) the same methodology (Azevedo and Santos, 2008). The CRISP-DM methodology is used as it

”provides a framework for carrying out data mining projects” (Wirth, 2000). Furthermore, it is more complete than SEMMA and KDD as it specifically starts with the business understanding phase while the others do not incorporate this explicitly (Azevedo and Santos, 2008). Due to the vast amount of data available and the complexity of DAF’s products and maintenance contracts, business understanding is of importance for this research. Lastly, CRISP-DM incorporates a deployment phase, which is less emphasized by the SEMMA and KDD methodology. Successful deployment is of significance for DAF as it allows them to use the model for future cost predictions.

1.4.3 CRISP-DM methodology

The CRISP-DM model contains the phases, their respective tasks and outputs for data mining projects. The model is divided into six phases, which are depicted in figure 1 (Wirth, 2000).

A short description per phase is given below:

- The *Business understanding phase* focuses on the formulation of business objectives and success criteria, which are subsequently transformed into a data mining problem definition, goals and project plan.
- In the *Data understanding phase*, initial data is collected, described (e.g. volume, attributes, key relationships), explored and verified on quality (e.g. coverage, missing attributes, missing data and deviations).
- Afterwards, in the *Data preparation phase* the data is selected, cleaned, constructed (e.g. transformations of month into season), integrated and formatted (e.g. normalization).
- In the *Modeling phase*, the modeling technique along with its assumptions is selected and the test design is generated. Subsequently the model is iteratively built and assessed (after which parameter setting revision often takes place). In this phase additional data preparation often takes place.
- When the model is finalized it is evaluated in the *Evaluation phase*. The data mining results are assessed with respect to the business success criteria, the process is reviewed and next steps are determined (e.g. possible actions or decisions based on the findings).
- Lastly, in the *Deployment phase*, the model deployment is planned along with monitoring and maintenance needs. A final report is created and the project is reviewed.

Important to note is that the CRISP-DM methodology is an iterative process. Moving back and forth between phases is a common occurrence throughout the data mining project. The arrows in the process diagram indicate the most important dependencies and iterations between phases (e.g. Business understanding and Data understanding) (Figure 1). An overview of all tasks that belong to the model phases is given in Appendix E.

During the Business understanding phase, a literature study was executed to gain insights into the available models and methods for the problem at hand. Furthermore, expert knowledge was used to gain insights in the business objectives for DAF and the determination of the data mining goals. In the data understanding phase, the data has been collected from the various data-sources at DAF after which an initial exploration has been made and the quality of the data has been verified and reported. This is done using both visual (e.g. boxplots and barcharts) as well as quantitative methods (e.g. numerical summaries). Subsequently the data preparation has been performed. First, the correct data had been selected by excluding data of all trucks that fell outside of the scope of this research. Secondly, various feature extraction operations have been performed in order to retrieve usefull features from the available telemetry data. Lastly, various data cleaning, construction and formatting tasks have been executed to provide a clean dataset with the appropriate formatting that could be used for modeling. In the modeling phase the prediction models have been created and optimized (e.g. hyper-parameter optimization and feature selection). Different methods were used to compare their performance and evaluate if there is one best technique for the problem at hand. In the evaluation phase, the results of the finals models have been evaluated on performance, usefulness for DAF and potential for future research. Finally, in the deployment phase a report was delivered and the findings were presented at (and reviewed with) both the company and the university.

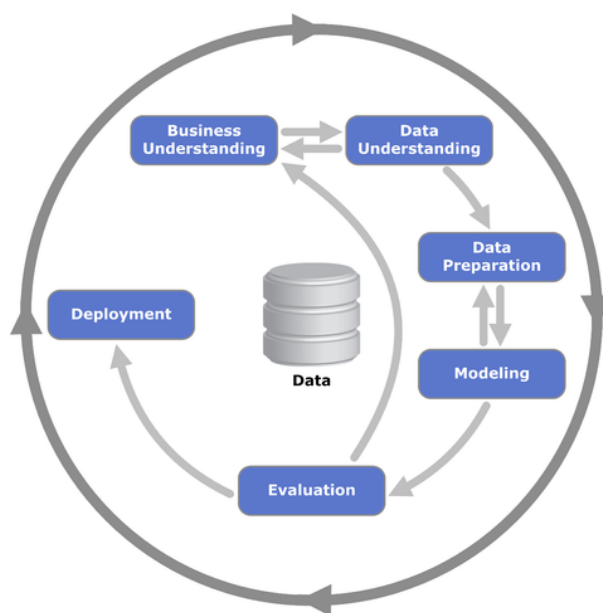


Figure 1: Phases of the CRISP-DM process model for data mining (Wirth, 2000)

1.5 Data mining goals and deliverables

Now, to provide an overview of the project, for each phase of the CRISP-DM model the deliverables are defined. The deliverables are incorporated in the report and are used to answer the (sub) research question. First, in order to obtain a proper business understanding it is important to map the current methods that are used to predict repairs and define the business objective for the project. With help of experts at DAF, the project is defined and data mining goals are established. Subsequently, the data is retrieved from the different sources at DAF after which they are analyzed on their contents and quality. The findings are summarized, visualized and presented in the data description section of this report. This is the deliverable corresponding to the data understanding phase. Then, with a proper understanding of the data, data preparation can be executed by selection, cleaning, construction, integration and formatting of the data. This is done in the data preparation phase. No specific deliverable for DAF is presented in this phase. Now, with the correct data in place, the test design and prediction models can be constructed and delivered. This is done in the modeling phase. Subsequently the results of the models are analyzed and assessed with respect to the business success criteria and data mining goals. This is done in the evaluation phase and as a deliverable, the findings are summarized and visualized. Lastly, in the deployment phase, the final report containing all findings and deliverables is presented and handed over to DAF.

The above described data mining goals are established with the help of DAF experts and are as follows:

- Predict the expected number of repairs for trucks over the available time period.
- Evaluate and determine the prediction power of model features and analyze their usefulness in predicting repairs.

The deliverables for each of the CRISP-DM phases as described above are summarized in Table 1.

1.6 Scope

To avoid loss of research quality, pre-determined scope boundaries are set.

The introduction of *DAF connect* is still very recent in terms of truck lifetime durations. At the start of this project, roughly half a year of truck data was available. DAF has started with the implementation of DAF connect before that time, but until April of 2017 this data has been inconsistent and unreliable. The available useful data is therefore from trucks of an age from one to half a year old. However, according

Table 1: Overview of the deliverables per CRISP-DM phase.

Phase	Deliverable
Business understanding	1. A definition of current repair predictions and objectives of the research
Data understanding:	2. An overview of available data and variables.
Data preparation:	No deliverable.
Modeling:	3. A model for the prediction of truck failures over a given period.
Evaluation:	4. Evaluation of the models prediction power.
Deployment:	5. Final document, presentation and recommendations/advice.

to DAF, peak maintenance and repair actions occur at an age of around four years. As more data is simply not available, the research is focused on the early life of the truck, being the first half year to a year of its life cycle.

As a consequence (of this lack of data), the research is of a more exploratory nature and thus, choices have had to be made. Not all repairs over a trucks lifetime can be predicted. Instead, predictions are made for early repairs. In future research, when more data is available, the findings and model from this research can be extended towards predictions over longer periods of time.

Only the unexpected repairs are targeted in this research. This means that maintenance actions, services, inspections and other, non-repair related adjustments to the trucks have not been taken into account.

1.7 Thesis outline

The remaining chapters of the report are structured according to the CRISP-DM phases. First, the findings of the literature review are given in Chapter 2. Then, the background information of the company and the available data are described in Chapter 3 (business understanding & data understanding phases). Chapter 4 goes into the details of the data preparation that has been executed (data preparation phase). Chapter 5 describes the modeling setup used to perform the repair predictions (modeling phase). Chapter 6 provides the results of the models and methods used (evaluation phase) and finally, Chapter 7 provides the conclusions and recommendations that followed from the research project.

2 Literature review

The research in this thesis aims to predict the number of failures of trucks based on their usage. Relevant findings on this topic as found in existing literature are described here. At first, an overview of current failure prediction methods is given. Afterwards, failure prediction with the help of multi-variate time series is elaborated in more detail. Then, the different machine learning techniques as used in this research are explained. Lastly a discussion and conclusion regarding the findings in the literature are given.

2.1 Predictive maintenance

In literature, the topic of failure prediction is often associated with Predictive maintenance and Condition based maintenance (CBM) (Peng et al., 2010). Due to the extensive attention that these topics have received and their close relation to failure prediction, it is included in this section. It provides an overview of the current data mining techniques that are used in machine and vehicle prognostics and predictive maintenance. In general, three types of predictive maintenance methods are applied, which are *Remaining useful life prediction (RUL)*, *Future state prediction* and *Condition prediction* (Peng et al., 2010).

2.1.1 Remaining useful life prediction

One way to apply condition based maintenance is by the prediction of RUL. Expected times until failure are calculated using statistical methods or data mining techniques. With these predictions, stakeholders are able to decide if and when a component or machine needs replacement.

Frisk et al. (2014) predicted the RUL of starter batteries in heavy-duty trucks. To do this, they used logged vehicle data that was retrieved from the trucks on a periodic basis. For a total of 33.306 vehicles, 291 variables were logged. They contained both numerical (e.g. temperatures and pressure) and categorical (e.g. model build and battery mount point) data. Important to note is that no time-series data was collected, only aggregates such as total distance traveled and time in service. A key problem was that out of all the available variables, it was not clear which best captured the battery degradation characteristics. The dataset was reduced to 30 variables which were selected based on area under the curve (AUC) of the receiver operating characteristic curves (ROC-curves) for single variable analysis and tree error rate analysis for multivariate analysis. These variables were used to create random survival forest, allowing for an accurate prediction of failures for a short time into the future (actual times were confidential), using a probability of failure threshold of 0.9.

Chinnam and Baruah (2004) analyzed drilling operations to predict the RUL of drill bits. In total, 12 drill-bits were tested until failure. Thrust force and torque data were collected at 250 Hz and later concatenated to 24 datapoints per drill operation. This time series data was subsequently fed to a focused time-lagged feedforward network (TLFN) to predict thrust force and torque signals. These predictions were subsequently used as two-dimensional input space for a fuzzy inference system (FIS) model to estimate the conditional reliability. Although providing a framework and showing the feasibility of the proposed method, no definitive performance measures were given.

RUL predictions have also been done by the use of vibration data. Wu et al. (2007) used accelerated testing to predict the RUL of thrust ball bearings. A set of 23 bearings was subjected to a constant load and rotational speed during which vibration signals were measured continuously. From this data a degradation signal was derived which was used to predict the bearing's residual life. A 3-3-1 Feedforward Backpropagation Neural Network (FFBPNN) was proposed which used the degradation signal of the bearings to predict the residual life percentile. Results showed an accuracy of 81.78% for a 10% prediction error threshold and an accuracy of 97.52% for a 20% prediction error threshold.

More recently, Mathew et al. (2017) have made a comparative study of machine learning (ML) techniques to predict the RUL of aircraft turbofan engines. For 250 engines with an unspecified sampling frequency, run-to-failure measurements from 21 sensors were collected. Feature extraction methods were not specified. After feature extraction, 10 different ML techniques were evaluated on their root mean square

error (RMSE) scores for RUL prediction of the engines. They found that the random forest algorithm generated the smallest error which was 29.73 time units (which are not further specified).

In general, remaining useful life predictions often concern some sort of degradation signals that have been collected from specific machine components. Although providing reliable results for predictive maintenance, RUL predictions are often performed for short times into the future only.

2.1.2 Future state prediction

Instead of RUL prediction, one could also use classification methods to predict the future state of machinery. Using historic data as input, models can be constructed that predict the future state of equipment for a given amount of time in advance (e.g. "failed" or "not failed").

Yang and Létourneau (2005) used decision trees and Naive Bayes classifiers to classify train axles as either 'going to fail' or 'not going to fail' within a certain time period. Data was extracted from so called *Wheel Impact Load Detectors* (WILD) which measures the impact force of each wheel passing the system along with additional information such as train speed, train direction, nominal weight of the car, name of the WILD site, and the time of the measurement. The data was collected over a period of 17 months for a fleet of 804 train cars with 12 axles each. The resulting dataset contained 200.808 observations per axle. Expert knowledge, linear regression and Fast Fourier Transformation were used to extract relevant features. Subsequently, different sets of these features were used to construct decision trees and Naive Bayes classifiers. The outputs of the four best scoring models were subsequently used to construct a meta-model which was used to classify the instances. To calculate the model, recall, false positive rate and an own performance indicator that took into account the time between 'going to fail' classification and actual failure were used. The final model had a recall of 0.97 and false positive rate of 0.08. Due to the lack of research in train axle failure prognostics, benchmarking with other research was not possible.

Last et al. (2011) applied multi-target information fuzzy network (M-IFN) classification to classify car batteries into 'broken' or 'not broken' as well as classification of the number of months until failure (i.e. 0.0, 0.5, 1.0, 1.5, 2.0 and 2+ months). The dataset contained 46.418 periodical battery sensor records from 21.814 vehicles. Each record contained 12 individual sensor measurements. The AUC was used as a performance measure, scoring 0.6165. Although the model was compact and interpretable (14 rules to classify both targets) and outperformed regular Weibull reliability analysis, its results were of limited accuracy.

In the domain of predictive maintenance, wind turbines are of high interest (Tchakoua et al., 2014). Large amounts of research have been executed regarding the prediction of their failures. An example is the work of Canizo et al. (2017). They used two year of sensory data from a wind farm of 17 turbines. A total of 104 operational parameters and 448 different alarm types were collected in 10 minute intervals. This data was subsequently used to predict the future state of each turbine for $t+10$ to $t+60$ minutes in advance. Principal component analysis (PCA) was used to reduce the set of parameters to 14 variables which represented 99% of the covariance. A random forest algorithm was used to generate the predictive models. They learned that the number of trees did not really affect the model's precision. The effect of maximum tree depth however, was significant. On average, the model predicted failures with an accuracy of 82.04%, sensitivity of 92.32% and specificity of 60.58%. The results are promising, although a relatively large amount of false positives was registered (due to low specificity performance).

Another popular field of application for predictive maintenance is found in the aircraft industry. Due to the potentially high impact of aircraft engine failures, early detection of failures is of high interest (Hong and Meeker, 2010). Byington et al. (2004) for example, have predicted the RUL for aircraft actuator components. Test data that was made available by Boeing was used to construct Fuzzy logic classification models, where the classification itself was based on RUL threshold levels. Feature extraction was done by signal processing and neural network modeling. Models ranged from 8 to 20 classifications, providing an overall error rate of 4% over 106 classifications and a maximum error of 10%.

On a first glance, future state prediction has a promising resemblance to our problem. It is often performed with the help of multi-variate data such as done by Yang and Létourneau (2005), Last et al. (2011) and Canizo et al. (2017). A recurring processing step involves the reduction of number of features

when one is presented with a large dataset. Different techniques such as PCA (Canizo et al., 2017), Neural network modeling (Byington et al., 2004) and expert knowledge (Yang and Létourneau, 2005) can be used for feature selection. Furthermore, the future state is predicted using a range of different machine learning techniques. Now, although these techniques show promising results for the classification and prediction of the future state of a vehicle, they often predict the future state for a relatively short time in advance (i.e. hours or days.) No research was found on the prediction of total breakdowns or repairs over a long period of time such as presented in this research.

2.1.3 Condition prediction

Lastly, machine learning is used to predict the future condition of machines. It differs from future state prediction with respect to the output of the models. Where future state prediction applies labeling for classification, condition prediction uses regression to predict values such as temperature, vibration, fuel level and so on.

For a set of 24 wind turbines Kusiak and Verma (2012) collected four months of sensory data that was recorded over 100 parameters in 10 second intervals. This data was used to predict bearing failures inside the turbines. In order to reduce the number of variables, domain knowledge was applied to discard 50 parameters. Subsequently, three different wrapper algorithms were used to select the most relevant parameters, being wrapper with genetic search (WGS), wrapper with best first search (WBFS), and a boosting tree algorithm (BTA). The final dataset contained the 18 most relevant variables. Multiple multilayer perceptrons neural networks (MLP) with different configurations were constructed to predict the turbines' future state based on bearing temperature. Model performance was evaluated on absolute error (AE), mean absolute error (MAE), relative error (RE), mean relative error (MRE) and coefficient of determination (R^2). A three layered 18-17-1 MLP proved to give the best performance with a MAE between 0.765-0.860°C, MRE between 1.65-1.88%, and (R^2) coefficient between 0.998-0.994 for the training, testing, and validation set. Although a good performance was realized, the model only predicted failures 1.5 hours in advance.

In a similar fashion as Kusiak and Verma (2012), Chaochao Chen et al. (2011) have applied condition prediction of bearings in helicopter gearboxes. For a period of 1000 ground-air-ground cycles of the helicopters, vibration features were extracted and used to develop a Neuro-Fuzzy system combined with a Bayesian estimation technique. Its performance was compared to a recurrent neural network (RNN), adaptive neuro fuzzy inference system (ANFIS) and adaptive recurrent based neuro fuzzy inference system (ARNFIS) for the RMSE on predictions from r to $5r$ time units in advance (where r is a confidential unit of time). The results demonstrated that the proposed method's prediction accuracy was higher than those of the three classical predictors. A RMSE from 0.0611 (r steps ahead on some weighted frequency measurement) to 0.0822 ($5r$ steps ahead) was realized for the bearings' vibration prediction.

Marinelli et al. (2014) predicted the future condition of earthmoving trucks in Greece. For a set of 124 vehicles their capacity, age, kilometers traveled and maintenance level were acquired. These were subsequently divided into a training, validation and test set and fed to a multilayer feedforward neural network (MLFN) with one hidden layer to predict the condition level of the trucks (divided over four classes). The neural network had an overall accuracy of 94.7% but no specification was made on the amount of time that the model predicted ahead.

Although condition prediction is useful for the planning of maintenance or the prediction of failures for a short time in advance, the prediction of condition often involves specific components or the use of basic parameters such as age and distance traveled (e.g. (Marinelli et al., 2014)) or univariate measurements such as bearing vibrations (e.g. (Chaochao Chen et al., 2011)). The resemblance with our research regarding the type of data available and the predictions desired are low and therefore not directly applicable in this research.

In general, existing research on predictive maintenance and failure prediction are focused on upcoming failures. They provide accurate results but none of them focus on the number of failures over time. They do however, show interesting insights about previous use of sensory data for the prediction of failures and breakdowns. Furthermore, insights on feature extraction and selection from multi-variate datasets have been discovered. However, the research as described above does not focus on multi-variate time series

data, such as available at DAF. This is why the next section is introduced. It describes the available literature on time series analysis on failure prediction.

2.2 Time series analysis

Reviewing the current methods of failure predictions revealed that the use of multi-variate (real time) time series data to predict the number of failures on individual trucks has (to the best of our knowledge) not been applied until this date, already pointing out the scientific relevance of this research. Nonetheless, in this section an effort has been made to couple the existing literature on time series analysis to the problem of truck failure predictions.

2.2.1 Traditional (time series) warranty and repair analysis

At first, the most common and simple form of time-series analysis for machine failure prediction is described. It is based on univariate or bivariate usage data such as age, mileage and historic warranty claims. Such predictions are common in the manufacturing industry and have been widely studied in literature (Murthy and Djameludin, 2002).

Wu and Akbarov (2012) for example, used historical warranty claim information to predict the number of warranty claims of a product in the next k months starting from the current month. They fitted an inverse Weibull distribution to the number of claims for a total of 20.000 products consisting out of 30 product types. Subsequently, they used a non-homogeneous Poisson process and constrained maximum likelihood estimations to build the forecasting model. The performance was measured using the normalized root mean squared error (NRMSE) resulting in an average error between 0.13 and 0.24 depending on the amount of months planned ahead.

Hong and Meeker (2010) used the warranty data in a different manner. They extracted use-rate and cycles-to-failure (nr. of uses until failure) information from high end copying machines together with three forms of occurred failure modes. Again, distribution functions were fitted to the data which were subsequently used to calculate the remaining life and number of failures over a given period.

In similar fashion, many other papers that use univariate or bivariate data for their research are available. Common examples for machine and truck warranty analysis are age based (Lawless, 1998), mileage based (Ye and Murthy, 2016) and failure/hazard rate based (Jackson and Pascual, 2008) predictions. Although providing a well established base for repair predictions, these methods are not able to take full advantage of the specific (and multivariate) truck usage data such as available at DAF. They are accurate for the prediction of fleet failure behavior but generalize their predictions too much to make accurate predictions on individual machines or vehicles (Meeker and Hong, 2014). Instead, a more comprehensive analysis in the form of multivariate time series analysis is desired, which is described in the following two paragraphs.

2.2.2 Feature based time series analysis

One way to classify time-series data is by the extraction of relevant features from the data that describe the time series. These can subsequently be used in classification and prediction methods such as decision trees and neural networks.

Extensive literature on feature selection from time-series data is available. They are most often used when comprehensibility of the variables and methods is desired. The extracted features can be used in white box models such as regression models, decision trees and random forest (Rodríguez and Alonso, 2004).

Rodríguez and Alonso (2004) extracted averages and deviations to construct interval-based decision trees. The method has been applied to different multivariate time series datasets. It proved to be competitive with other methods of decision tree construction. However, comprehensibility comes at a cost. Methods such as boosting outperform the interval-based decision tree on accuracy performance but reduce comprehensibility of rules.

Khaleghi et al. (2016) used estimators of distributional distance between time-series sequences to construct a clustering algorithm. Data was taken from the MOCAP database, representing human locomotion. They evaluated the model on entropy score and accuracy, showing that their model outperforms current methods as used on the MOCAP dataset. They also showed that their methods can have worse results for non-ergodic data.

Another method of feature extraction is to sum the time series data over a certain period and use these aggregations as features for the prediction models. Prytz et al. (2013) used aggregated vehicle log data from eighty trucks as input for their models. Hundreds of variables were reduced to twelve parameters based on expert knowledge and subsequently used to classify a compressor component into healthy or faulty. Random Forests and K-nearest neighbors algorithms were constructed. Although the amount of data available was limited and feature selection was based on expert knowledge only, the method proved to outperform the company's scheduled maintenance plans.

Later Prytz et al. (2015) revisited the problem of air compressor failure predictions, but applied different methods for feature selection. From approximately five-hundred available variables, the 14 most relevant ones were selected. Guyon and Elisseeff (2003) and Bolón-Canedo et al. (2013) provided an extensive overview of feature extraction methods. From these methods Prytz et al. (2015) adopted a wrapper and filter approach to search for optimal feature sets. Subsequently these feature sets were compared to a feature set as comprised by domain experts. They showed that the wrapper and filter methods outperformed the expert's feature set using the random forest method.

Additionally, clustering of multivariate time-series (MTS) datasets can be performed based on similarity factors. Singhal and Seborg (2006) classified the operating condition for a batch fermentation process in a production plant. The dataset consisted out of 12 temporal values that were measured in 1 minute intervals for thirty hours. A total of 100 process batches was clustered. They used a Principal Component Analysis similarity factor, Euclidean distance measure and Gaussian probability distribution measure as similarity factors. They showed that their clustering method was more accurate than the clustering of unfolded data and outperformed current methods as applied on that dataset.

Baydogan et al. (2013) argue that, for long time series, it is more appropriate to measure similarity from higher level structures instead of local comparisons by similarity factors such as Dynamic Time warping (DTW) and Euclidean distance measures. They extracted global properties from the time-series after which they tried to improve the classification by the addition of local properties. The time-series were divided into intervals after which features such as slopes, means and variances from these intervals are extracted and used in the prediction models. The classifier was trained with support vector machines (SVM) and random forests. Subsequently, its performance was compared to classifiers with DTW and global features using 45 of the publicly available UCR time series database. The results were promising, showing that it outperformed the other methods on most of these datasets. Baydogan et al. (2013)'s research was based on univariate time-series. The work from Wang et al. (2016) provides a method to extend these kind of feature derivations for MTS's. In a similar way as Baydogan et al. (2013) they derive features from the univariate series of a dataset. Subsequently, the most powerful features are selected after which they are concatenated to form a vector of features that represent the whole MTS dataset. These are then used to classify the MTS's. The method was tested on a set of human motion capturing data which contained 10 different movements (classes) measured on 25 variables. After feature extraction and selection, 10 features remained which resulted in a classification accuracy of 89%.

In general, the available literature on MTS analysis is relatively scarce (Fulcher and Jones, 2014). Nonetheless, global feature extraction (e.g. mean, weighted average, variance, skewness, , min, max etc.) in combination with local feature extraction have shown promising results with regard to MTS classification and clustering. To a lesser extent, similarity factors (e.g. dynamic time warping, euclidean distance measures etc.) have been used for MTS classification as well. However, they can be outperformed by feature extraction methods and result in a less interpretable analysis of the classifier rules (Fulcher and Jones, 2014). The latter methods are explained in the following paragraph.

2.2.3 Pattern based time series analysis

Besides the feature extraction methods as described above, other methods of time series analysis have been researched as well, focusing less on the comprehensibility and feature extraction from the data and

more on accuracy of results.

Wang et al. (2016) proposed the use of deep neural network learning for MTS dataset classification. They argued that their methods are pure end-to-end without the need for heavy pre-processing steps. Deep multilayer perceptrons (DMLP), fully convolutional networks (FCN) and residual networks have been constructed and evaluated on 44 univariate UCR time series datasets as a benchmark test. The results were comparable to other methods, but as stressed, required much less pre-processing. Furthermore, the proposed methods tend to over-fit due to the large number of layers and model interpretability is poor.

In a similar manner, Zheng et al. (2014) have used deep learning to classify MTS datasets. They proposed a Multi-Channels Deep Network to learn features from the individual dimensions of the time-series and subsequently concatenates them and feeds them to an MLP to perform classification. They evaluated their method on two real-world datasets. The first dataset contained 19 classes of physical activities performed by 9 subjects which were measured on 52 variables. This resulted in a dataset of 3.850.505 instances. The second dataset is comprised of 53, 8-minute recordings of ECG, PPG, and impedance pneumography signals (with a sampling frequency of 125 Hz.). They showed that their model outperformed state of the art DTM nearest neighbor algorithms on both datasets, obtaining an accuracy of over 90% on both sets. However, algorithm and classification rules are not visible and thus the model's underlying methods are a black box. Furthermore, they estimated that it would take a nearest neighbor algorithm with Dynamic Time Warping one month to perform its classification, due to the computationally expensive DTW similarity measure.

2.3 Modeling techniques

Learning method In order to select the appropriate modeling technique, the learning method has to be defined at first. In general, there are two types of inductive-learning methods, being *supervised* and *unsupervised* learning. Supervised learning is used when output values (or classes) for the training samples of the machine learning model are known and the model response can be evaluated directly. With unsupervised learning, the input values are given to the learning system while no output value (classes) are known during the learning process. The goal of unsupervised learning techniques is to discover natural structure in the data while supervised methods are used to predict labels or numbers (Kantardzic, 2011). For the study at DAF, the number of repairs on trucks and their associated costs are stored in local (and cloud) databases and thus output values are known upfront. Therefore, the learning method is of the type supervised learning.

Machine learning techniques With the learning method established, a decision had to be made on the machine learning techniques to be used for the study at DAF. Many different methods exist and there is no distinct preference for each of the models as this is dependent on multiple, case specific factors. Some models such as decision trees and logistic regression models are easily interpretable due to the ability to graphically visualize the decision rules (when their size is kept small) and feature importances can be derived. Other models such as neural networks are more of a 'black box' model and often require a relatively large amount of data but are highly capable in modeling non-linear input/output relationships (Byington et al., 2004).

Now, the goal of the project is two-sided. On the one hand there is a desire to achieve a high as possible model performance, as this is used to evaluate the potential of DAF Connect regarding repair (cost) predictions. On the other hand, clarification of decision rules and feature importance interpretability is desired as these can be used by DAF to derive business rules regarding contract costs and maintenance schemes for trucks in the future.

Random Forests and MLP's are used as modeling techniques in this research because literature research (Chapter 2) showed that they are often used to predict future state, RUL and machine operating conditions. Although often providing good modeling performance, this comes at the cost of limited decision rule interpretability (Negnevitsky, 2005). To compare their performance to some relatively simple models which can be interpreted easily, logistic regression models and Decision Trees are constructed as well. Their working is explained in short below.

2.3.1 Logistic Regression

The first modeling technique is logistic regression. It is a relatively simple method that is used to predict a binary dependent variable ($Y=1$ or $Y=0$) based on a set of predictor variables X . For the study at DAF, a truck having 'many repairs' is classified as 1 and a truck having 'few repairs' is classified as 0. Using this logic, the regression coefficients (also called beta-coefficients) of the features in the model can be used to derive the relation of the features with respect to the output class. A positive beta-coefficient indicates a contribution to output class 1 while a negative regression-coefficient does this for class 0.

In short, the Logistic Regression model with k different independent variables is given by:

$$P(Y = 1) = \frac{1}{1 + e^{(\beta_0 + \beta_1 * n_1 + \beta_2 * n_2 + \dots + \beta_k * n_k)}} \quad (1)$$

where $P(Y=1)$ is the probability of a truck having many repairs and $\beta_0 + \beta_1 * n_1 + \beta_2 * n_2 + \dots + \beta_k * n_k$ are the corresponding regression coefficients (Kurt et al., 2008).

Formally, equation 1 is the following function solved for p :

$$\log \frac{p(y = 1)}{1 - p(y = 1)} = \beta_0 + \beta_1 * n_1 + \beta_2 * n_2 + \dots + \beta_k * n_k \quad (2)$$

This transformation of the outcome is called the logistic (or logit) transformation. Fitting the data to this logit function is known as logistic regression.

2.3.2 Decision Trees

A decision tree describes a data set by a tree-like structure. It starts with a so called root node from which it develops new nodes, branches and leaves. The root node includes all data which is split over different nodes as the tree grows deeper. The goal of these splits is to separate the data into subsets of increasing purity with regards to the dependent variable (classification label in our case). Thus a split in the decision tree corresponds to a predictor (variable) with the maximum separation power over the (sub)set under consideration. For a better understanding, an example of a small decision tree is given in Figure 2.

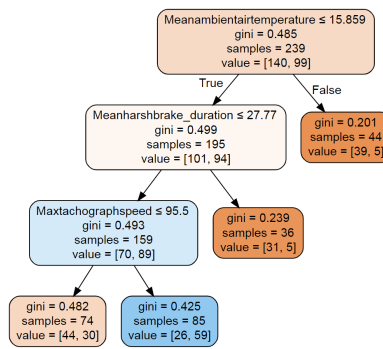


Figure 2: An example of a decision tree .

In this case, the root node is represented by the mean ambient air temperature as measured by the truck. The importance of the feature is calculated as the decrease in node impurity weighted by the probability of reaching the node. The probability of reaching the node is easily calculated by dividing the number of samples that have reached the node by the total number of samples. The node impurity is represented by the Gini scores of the nodes:

$$GiniScore = 1 - \sum_{i=1}^C (P_i)^2 \quad (3)$$

where C is the number of unique labels (or categories, which are 2 in our binary classification case) and P_i is the chance that an instance in the node under consideration is of class i . Thus, the Gini score becomes zero when all cases fall into a single category and 0.5 when the cases are equally split over both categories.

As an example, the Feature importance of the 'Maxtachographspeed' in Figure 2 is calculated as follows:

1. Impurity decrease to the left: $(0.493 - 0.482 = 0.011)$.
2. Impurity decrease to the right: $(0.493 - 0.425 = 0.068)$
3. The probability of reaching the node to the left: $74/239 = 0.309$
4. The probability of reaching the node to the right: $75/239 = 0.314$
5. Finally, the Feature importance is then calculated by: $0.011 * 0.309 + 0.068 * 0.314 = 0.025$

Note that a feature can appear in a tree multiple times. In this case, their values are summed. For a more detailed explanation of the Gini index we refer the reader to the work of Rutkowski et al. (2014).

Furthermore, for the remainder of this report, the feature importances are given as normalized scores (i.e. summing them to 1) in order to provide a clear view of the relative feature importances in the models.

2.3.3 Random Forest

The random forest is a so called ensemble learning method. Basically, it is an ensemble of Decision Trees as explained above. A multitude of different decision trees is built after which majority voting is used to perform classification. These trees are somewhat random due to the fact that random sets of observations and features are used to construct the individual trees. In short, the algorithm works as follows (Liaw and Wiener, 2002):

1. Draw (bootstrap) samples from the dataset.
2. For each sample, grow a classification tree where at each node, a random sample (instead of all) of the predictors is evaluated to choose the best split.
3. New data is subsequently classified by aggregating the predictions of the individual trees that have been constructed in the forest. To do this, majority voting is used.

The random forest returns a matrix where the rows represent the test instances and the columns represent the scores for each classification label of the data. This score is given by the fraction of trees in the model that classified (or voted) the instance to be of the label represented by the column. The model then chooses the label with the highest fraction (most votes) as the final label for the test instance under consideration. This is called majority voting. In general, the randomness and majority voting in the model prevents over-fitting and allows for good generalization over the data (Breiman, 2001).

2.3.4 MLP Neural Network

The last modeling technique that is considered is the MLP-Neural network. It is a feed-forward artificial neural network, consisting of an input layer, 1 or more hidden layers and finally an output layer. Each layer (except the input layer) consists of a set of nodes, which are neurons that utilize a nonlinear activation function. For clarification an example MLP is given in Figure 3. It is an MLP with n input nodes (features) and 1 hidden layer with k neurons. The output is a non-linear function approximation which can be used for both classification and regression.

Now, in every hidden layer, the neurons transform the values of the previous layer with a weighted sum of the inputs and the specified activation function (e.g. step functions, logistic functions or tangens hyperbolicus functions). The model is trained using back-propagation, where the weights of the connections are adjusted based on the strength of each node's contribution to the final prediction made by the

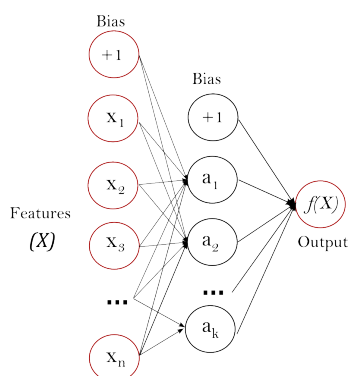


Figure 3: Simple schema of a MLP-NN with 1 input layer, 1 hidden layer and 1 output layer.

network. The mathematical optimization algorithm to do this is called gradient descent (Witten, I. H. , Frank, E., 2016). In short, the steps to train the model are as follows:

1. Initialization: Weights and biases are initialized based on a normal distribution with a mean and variance of zero.
2. Sample presentation: Provide the training samples to perform forward and backward passes.
3. Forward pass: Propagate a training sample from the input, through all layers until the output and calculate the error signal.
4. Backward pass: Recursively compute the local gradients from the output layer, through all layers until the input layer and adapt the weights according to the error gradient.
5. Iterate steps 2 till 4 until a stopping criterion is met (e.g. gradient vector has reached a small enough gradient, output error size is sufficiently small, the generalization performance has peaked or the maximum number of iterations has simply been reached).

The advantage of neural networks is that they are capable to learn complex non-linear relationships in the data. However, this comes at the cost of model interpretability as neural networks are often described as black boxes (Witten, I. H. , Frank, E., 2016). The model has been incorporated in this research in order to see if a higher classification performance could be reached by sacrificing understandability and feature importance evaluation.

2.4 Previous work

Recently, another study regarding the prediction of the number of truck repairs has been executed at DAF. The research executed by Goudsmits (2018) aimed to predict the number of truck repairs over time based on truck specifications and repair and maintenance contract details (i.e. intended truck usage, cargo, area of operation and so on). He found that *inspection interval*, *homecountry*, *body type* and *estimated yearly mileage* were important factors for the prediction of the total number of repairs over time. The difference in his work and this research is the fact that DAF Connect was not used for the repair predictions, as it was not available at the time of his research. In addition to the use of DAF connect, the truck specification data is used as well. As this has been derived from a similar dataset as used by Goudsmits (2018) and the goal of the research is similar (predicting the number of truck repairs), the results from this research are compared to those of Goudsmits (2018).

2.5 Discussion

Traditional repair and maintenance cost analyses using warranty claim data and basic time-series data such as age, failure rates and mileage do not take full advantage of the multivariate usage data at DAF and are therefore not suitable for this research. Instead, methods associated to multivariate time series analysis appear to provide more promising results.

There are two ways to approach MTS analysis. On the one hand, one can use methods such as (deep) neural networks to achieve maximum accuracy prediction at the cost of model interpretability (Wang et al., 2016) (Zheng et al., 2014). On the other hand, one can use more interpretable methods such as decision trees and random forest at the cost of accuracy performance (Baydogan et al., 2013) (Bastos et al., 2014) (Prytz et al., 2015). Which of the two approaches is best depends on the goals of the research.

Important to take into account when analyzing MTS data is the required computational power. MTS distance similarity measures such as DTW and Euclidean distance calculation show promising classification results (Singhal and Seborg, 2006) but Zheng et al. (2014) showed that unmodified similarity factor analysis of a multivariate time-series dataset with 'just' 3.850.000 instances would already take at least a month of calculation time on an average computer.

2.6 Conclusion

The goal of this research is to predict the total number of failures over a given period. This is subsequently used to explain, based on truck usage, which customers are likely to have many repairs and what the explaining variables/causes are. Literature review showed that traditional predictive and condition based maintenance methods, although providing accurate results for failure predictions over short time periods, are not used to predict the total number of failures over long periods of time. Furthermore, they make no use of multi-variate time series telemetry data to do these predictions.

More resemblance to our research can be found in literature concerning multi-variate time series analysis. Although lacking research about aggregate number of failure predictions as well, it provided useful insights about approaches to multi-variate time series analysis and its application to failure predictions.

Overall, comparing the reviewed repair and maintenance prediction methods in combination with the available data showed that global feature extraction in combination with local feature extraction such as applied by Baydogan et al. (2013) provided the most promising results.

Lastly, four different suitable modeling techniques have been identified during literature review. These are logistic regression, decision trees, random forests and neural networks respectively. Random Forests and neural networks are often found in literature regarding predictive maintenance and thus used in this research. To compare their performance to some relatively simple models which can be interpreted easily, logistic regression models and Decision Trees are constructed as well.

3 The company and the available data

In this chapter, the background of the company is given. Their products are explained and the department in which this research project is executed is described. Secondly, the available data is described and data exploration findings are given.

3.1 The company

DAF's headquarter is based in Eindhoven, the Netherlands and recently celebrated its 90th anniversary. In 1928, the company was founded by Huub van Doorne as a simple engineering and blacksmith workshop. Over time, it quickly developed into the leading truck manufacturer that it is today. The DAF Euro 6 is the latest truck model which has received an additional overhaul last year, meeting the latest environmental requirements and providing the best comfort, quality and efficiency to date.

Besides Eindhoven, DAF has production facilities in Westerlo, Leyland (England) and Ponta Grossa (Brazil). The total production area covers approximately $2.200.000m^2$, allowing for a total workforce of 9240 FTE (fulltime-equivalent). DAF produces trucks according to the 'built-to-order' principle, meaning that trucks can be built according to specific customer needs. In 1996 DAF has been taken over by PACCAR inc, being a subsidiary ever since (DAF, 2018a).

Currently, DAF produces no less than 240 trucks per day. Allowing for a European market share in the heavy and light segment of 15.5% and 10.1% respectively. DAF is the market leader for the heavy segment in the Netherlands, Great Britain, Poland, Hungary and Bulgaria. Furthermore, it has a growing presence outside of Europe. Market positions in e.g. Ecuador, Peru, Chili an Colombia are strengthening due to the expansion of dealer networks in these countries and the recently opened manufacturing plant in Brazil (DAF, 2018a).

3.2 Products

3.2.1 Truck type

DAF distinguishes three main types of trucks, being the LF, CF and XF (Figure 4.) (DAF, 2018b). The XF is DAF's long haul truck of choice. Offering optimal transport efficiency, reliability and the lowest fuel costs. The CF is DAF's medium sized all-round truck, excelling in its versatility. This makes the CF most suitable for all-round transport and non-standard applications such as garbage collection and construction work. Lastly, the LF is most suitable for short-haul (distribution) transport. Its size and engine types ensure ideal efficiency and fast delivery (DAF, 2018b).

Each customer and transport application has different requirements. This is why DAF allows its customers to configure trucks according to their own unique wishes. Choices range from axle configuration and cabin type to safety features (e.g. night-lock) and the selection of the steering wheel material (DAF, 2018b).

In addition, DAF supplies its engines and axles to coach, bus, off-road and agriculture vehicle manufacturers and provides financial lease for its own trucks through PACCAR financial inc.




Figure 4: DAF's newest LF, CF and XF series respectively.

3.2.2 Chassis type

Besides a truck type, different chassis and corresponding axle configurations are available as well. For the chassis there is an option for either a rigid (FA) or tractor (FT) chassis. Rigid trucks have their container directly attached to the chassis while tractor trucks carry detachable trailers (figure ??).

Lastly, the different kind of axle configurations (each for its own cargo and transport purposes) can be found in Figure 5. Of these configurations, the standard FA and FT configuration are sold the most. The actual most sold configuration differs per truck type. For example, for rigid XF trucks the FAR configuration is in higher demand than the FA configuration.



Bakwagenchassis

	LF city	LF R-12L	LF R-10L	LF 19L	CF PK-7	CF MX-11	CF MX-13	XF
FA 4x2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
FAR 6x2					<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
FAS 6x2					<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
FAG 6x2					<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
FAN 6x2					<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
FAT 6x4					<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
FAK 8x2					<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
FAQ 8x2					<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
FAC 8x2					<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
FAX 8x2					<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
FAD 8x4					<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>


ASCONFIGURATIE

Trekkerchassis

	LF 01L	CF MX-11	CF MX-13	XF
FT 4x2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
FTP 6x2		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
FTR 6x2		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
FTS 6x2		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
FTG 6x2		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
FTN 6x2		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
FTT 6x4		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
FTM 8x4				<input checked="" type="checkbox"/>


DAF Trucks Belux
Luxemburgstraat 17
9140 Ternes
Tel: 03/710.14.11
Fax: 03/710.14.02
www.daf.be

Naloopas



Ook verkrijgbaar in Construction-uitvoering

Aangedreven as



Ook verkrijgbaar in Low Deck-uitvoering

© DAF Trucks N.V., Eindhoven, Nederland.
Aan deze publicatie kunnen geen rechten worden ontleend. DAF Trucks N.V. behoudt zich het recht voor om zonder voorafgaande kennisgeving productspecificaties te wijzigen. Producten en diensten kunnen per land verschillen. Neem contact op met uw erkende DAF-dealer voor actuele informatie.

BE-AL-9917

DRIVEN BY QUALITY

Figure 5: DAF's available axle configurations.

3.3 Service and aftersales

3.3.1 Warranty and R&M contracts

For each produced and sold truck, DAF offers multiple services and aftersale options. A standard one year full warranty and two year driveline warranty is included in each truck sale, after which customers can add additional services such as *Repair and Maintenance contracts (R&M contracts)*, *Road side assistance*, *Financial services* and *Driver training* (DAF, 2018c). As the research concerns the repair, maintenance and failures of trucks, the R&M contracts (also called DAF MultiSupport packages) are elaborated below.

Customers can choose from a range of six pre-defined maintenance packages at DAF. Each package has the option for some additional services as well. The packages including optional services are shown in Figure 6.

The *Warranty plus - Driveline* and *Warranty plus - Vehicle* packages provide additional warranty on top of the standard warranty period, but do not include any of DAF's additional support services. They

Predicting the number of truck repairs using logged vehicle

Page 18

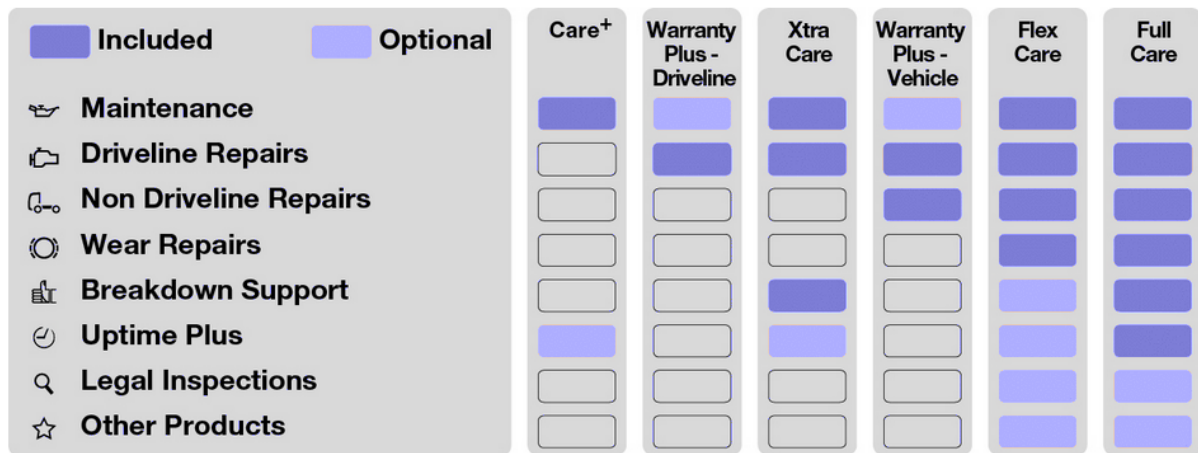


Figure 6: DAF's available Multi-support packages (DAF, 2018c).

focus solely on the repair of parts that do not experience wear. The *Care+*, *Xtra Care*, *Flex Care* and *Full Care* packages do offer these additional support services (DAF, 2018c). R&M contracts (called packages at DAF) can be closed for up to ten years, while warranty contracts run for a maximum of three years. For an overview of the distribution of contracts per truck, the reader is referred to Appendix F.

3.3.2 MultiSupport Calculator

For the price calculations of a MultiSupport package, DAF has developed an online price calculator that is used to negotiate and agree on the total package price. The costs are calculated not only on the chosen package and options, but includes specific driver/truck information as well. For example, the country in which a truck operates and the payload that it is expected to carry also influence the total price. The features as used in the current R&M contract price calculations are given in Table 2. Based on the input, contract costs are derived and one out of three maintenance schedules is selected. The initial price of the contract is based on the selected truck type. Depending on the other inputs (i.e. the variables from Table 2), additional costs are added to the initial price, based on cost curves and expected costs as assigned to the selected variables.

Customers that have bought a MultiSupport package can go to official DAF dealers and franchisers for necessary repair and maintenance actions. The dealers can subsequently make a claim for the costs of the executed repair or maintenance actions after which DAF checks the validity of the action and reimburses the dealer when approved. It is possible that dealers sell a truck without one of DAF's own support packages. In that case, the dealer often provides its own contracts or uses other third party insurances, meaning that no claims can be made at DAF (except for the standard first-year warranty repairs).

3.3.3 Costing & analysis

The department at which the research is conducted is the *Costing & Analysis department*. It is a relatively new department that is founded in 2014. Before that time they were part of the Commercial services department. The main tasks of the department are *analysis and reporting*, *cost curve management*, *data quality management* and *predictive tasks*. Among others, the department is responsible for the prediction of total R&M contract costs.

Table 2: The features that are currently used for the price calculations of R&M contracts.

Feature	Description
Contract package	MultiSupport care package of choice.
Options	Additional options to the pre-fixed care package.
Contract length	Length of the contract in years.
Start kms	The total mileage on the truck at the time of contract closing.
Vehicle age	The age of the vehicle at the time of contract closing.
Yearly mileage	The expected amount of driven miles per year.
Oil type gearbox	The type of oil in the truck's gearbox (as different grades of oil are available).
Oil type rearaxle	The type of oil in the truck's rearaxle.
Oil type engine	The type of oil in the truck's engine.
Truck model	XF, CF or LF.
Truck series	Model version.
Body type	The type of body that the truck carries (e.g. rigid, tractor, box or demounting).
Chassis type	Type of chassis on the truck.
Axle configuration	Axle setup on the truck.
Engine type	The specifications of the engine inside the truck.
Emission	The emission label of the truck (e.g. EURO 3,4,5 or 6).
ADR	Certified to transport dangerous goods (yes/no).
S&M inspection interval	The agreed time between maintenance inspections (for english trucks only).
Service interval engine	The agreed duration for engine oil replacement intervals.
Power Take Off	The average hours per day that the Power Take Off on the truck engine is used.
Gearbox	Type of gearbox inside the truck.
Tractor coupling type	Tells whether a coupling is fitted or not.
Retarder System	Type of retarder system fitted inside the truck (if any).
Nr. of drops per day	The expected amount of cargo drops that the truck will make per day.
Area of operation	The geographical area in which the truck will operate.
Type of operation	The type of trips that the truck will make (e.g. long haul or regional).
Road type	What share of it's trips the truck will drive off-road.
Application	The type of cargo that the truck will carry (e.g. sand & gravel, pallets or waste).

The data that is used in this research is extracted from a range of different data sources. In order to provide the reader with a clear overview of the available data, the contents of the datasets and the information that is extracted from them are described next.

3.4 Data sources

The data that is used in this project has been extracted from multiple sources, which on their turn, exist of one or more subsets of data. An overview of the different sources is given in this section. The data from the different datasets are linked to the trucks by truck identification number which is present in each of the datasets, called the *vehicle identification number (VIN)*.

3.4.1 CCM truck data

The CCM database contains the relevant information regarding the warranty and R&M contracts that are sold with the trucks. It contains the truck's specification information (e.g. model, engine type and production date) and customer contract information (e.g. home country, warranty package, R&M contract type and delivery date). A complete overview of the variables is given in Appendix A.

3.4.2 Mi claim database

Repairs are executed and claimed by official DAF dealers and workshops. Each claim is either entered into the so called *Dealer Claim Entry system (DCE)* or Service Claim Handling (SCH) system after which it is stored in the claim database, called the *Mi claim database*. It contains information about the truck and corresponding customer, details about the repair and all relevant costs associated to it. When a repair is claimed at DAF, it is not automatically reimbursed. Each individual claim has to pass a claim review first. It is not uncommon that claims are declined, for example when unreasonable prices are charged or excessive labor hours are registered. In this case, the dealers can opt to pay the unaccepted claim amount by themselves. The paid amount is then stored in the data base as a *local policy payout*. When dealers don't reimburse the customer either, the claim is definitively declined. The reviewed claims and the outcomes are all stored in the Mi claim database. An overview of all variables in the dataset is given in appendix B.

3.4.3 Connect database

The connect database holds all the truck usage data that is collected from the trucks that are operating in the field. Operational data is collected from the trucks by a multitude of sensors, after which it is sent to DAF by the use of telemetry and stored in off-site servers. This data is then used to analyze truck information such as fuel consumption, trip route information, combined load weight and so on. The data that is sent through connect is stored in three subsets, being the trip data, trigger data and snapshot data respectively. A more detailed explanation of these three sub databases is given in section 1.4.1. Furthermore, a complete overview of the variables in the three databases is given in appendix C. The trigger data is not analyzed in details as its relevant features are found in the snapshot data as well and the trigger messages are not related to specific repairs anyway.

3.5 Descriptive analytics of the CCM Database

In this section, the available CCM data is described and the most important findings that resulted from the data exploration are given for the data up to and including 21-08-2018. The features in the CCM database can be categorized into three groups, based on their relation to the truck. They are the following:

- Contract information
- Truck specifications
- Truck usage

The different groups are explained below, and the corresponding variables are listed in Table 3. For an explanation of each of the variables, we refer the reader to Appendix A.

Contract information (38 features) concerns all features that hold information about the R&M contracts that have been sold with the trucks. Upon closing of the contracts, the contract type, duration and other specifics are agreed upon and stored in the CCM database.

The *Truck specifications* (30 features) are also found in the CCM database. Before the closing of an R&M contract, truck specifications such as the model, axle configuration and engine fitted are defined. Furthermore, details such as the oil type in the axles and the presence of vehicle safety features are determined as they all affect the pricing of the R&M contracts.

The features in the *Truck usage* group (8 features) are part of the contract details as well. They are determined together with the customer based on the expected use of the truck and the environment in which it will operate. Among others, they concern the delivery country, type of cargo that is hauled and the number of cargo drops that the trucks will make per day. Contract fees are adjusted based on assumptions of the effects of these feature values on the total truck repair costs. Until this point in time, DAF had no method to check if truck owners actually used the truck as specified. This is why they believe that the actual usage may differ from the specified usage characteristics. Of course, with the introduction of *DAF Connect*, usage could be more closely monitored in the future.

Table 3: Overview of the CCM truck data groups and the corresponding features.

Contract Information	Contract information (2)	Truck specifications	Truck usage
Forecasting Run Date	First Registration Date	Model	Service Interval Engine
Forecasting Report Date	Country	Series	Number of Drops per day
Subsidiary	Default Service dealer Location code	Sub series	Area of Operation
Contract Number	Delivery date	Chassis number	Type of Operation
Contract Version	Delivery Country	Brand	Power Take Off (PTO)
Contract Group	Selling dealer	Engine power	Static PTO Hours per day
Contract Name	Selling dealer Location code	Axle configuration	Road Type
current Contract (version) Status	Default Service Dealer	Emission	Application
snapshot Contract (version) Status		Asset Description	
Contract Birthdate		Asset Type-info	
Contract (1st) activation date		Vehicle Park Number	
Contract (version) Start date		Vehicle Safety Features	
Contract (version) activation date		Soot Filter	
Contract (version) End date (original)		Retarder System	
Contract end-date (actual)		Fuel Specification	
Contract end-year (actual)		Factory External Camera System	
Contract closing date		Body Specification	
Contract closing year		Taillift Fitted	
Contract (version) duration in months (original)		ADR Specification	
Contract (version) duration in months (actual)		(Semi-) Trailer Coupling	
Contract (overall) duration in months (actual)		Rear Axle Oil	
FinVehAge		Gearbox Oil	
Contract (overall) Age in months (actual)		Engine Oil	
Contract (version) start kms		Driven Axle Suspension	
Contract contracted yearly mileage		Body Type	
Contract Origin		Axle configuration	
Contract package		Engine Type	
Currency		Engine	
Claim delay Date		Gearbox	
Last date invoiced		Rear Axle Type	

To provide the reader with a better understanding of the data in the CCM database, the minimum, maximum, mean, median and standard deviation of the numerical features are calculated, which provide the reader with a quick overview of the feature values and their characteristics (Nelson et al., 2003).

Table 4: Quantitative summary of all numerical variables in the CCM dataset.

Variable	min	max	mean	median	std
Contract (version) duration in months (original)	11	96	40	36	15,13
Contract (version) duration in months (actual)	11	96	40	36	15,13
Contract (overall) duration in months (actual)	11	96	40	36	15,13
FinVehAge	1	16	9	10	3,82
Contract (overall) Age in months (actual)	1	16	8	8	3,73
Contract (version) start kms	0	81.500	692	0	5754,77
Contract contracted yearly mileage	25.000	290.000	133.976	120.000	34.443
Static PTO Hours per day	0	5	0,17	0	0,27
Month_in_service	0	15	7	7	3,72

Furthermore, an initial quality check has been performed on the CCM truck data. This revealed that several of the variables contained missing values, indicated by "NULL" in the dataset. Missing values in a dataset often negatively affect the performance of prediction models (Triebel et al., 2008). They are discussed in more detail in the next chapter. The missing values are quantified and listed in Table 5. As can be seen, there are quite some variables with many missing values. This is for a large part explained by the fact that, when a truck is not fitted with a certain component, a NULL value is filed in the database. The *Power Take Off* for example, is either fitted (value: 'gearbox mounted') or not (value: 'NULL'). Consultation of CCM domain experts revealed that this data is missing as they were only gathered for older truck models, and no longer stored for the newer models as analyzed in this research. However, this doesn't mean that the features are not fitted on the trucks. When possible, the connect data has been used to derive their presence (e.g. counting PTO duration to derive if a PTO has been fitted). The features with missing values are further addressed in Chapter 4.

Table 5: Overview of the CCM truck data variables containing missing values.

Variable	DataType	Nr. Null	% Null
Vehicle Safety Features	Cat.	2864	99,3%
Soot Filter	Cat.	2864	99,3%
Factory External Camera System	Cat.	2864	99,3%
Body Specification	Cat.	2864	99,3%
Taillift Fitted	Cat.	2864	99,3%
Driven Axle Suspension	Cat.	2864	99,3%
Power Take Off (PTO)	Cat.	2864	99,3%
Contract closing date	Num.	2812	97,5%
Contract closing year	Num.	2812	97,5%
S&M Inspection Interval ('O' licence)	Cat.	2804	97,2%
Vehicle Park Number	Cat.	2147	74,4%
First Registration Date	Num.	11	0,4%
FinVehAge	Num.	3	0,1%
Delivery date	Num.	3	0,1%

Besides missing values, the dataset contains categorical features that have a single value for most (or all) of its instances. For example, all of the trucks have been fitted with the same axle oil (synthetic) and engine type (EURO-6). Furthermore, almost all trucks have the same *type of operation* (99,3%: *long distance*) and *number of drops per day* (99,3%: *1 to 6* drops). The features that have the same value for each instance are removed from the dataset as they then become a constant (i.e. zero variability predictor) and thus do not contain any information. The features that contain predominantly (but not completely) one value should be handled with caution. They could add unnecessary complexity to the prediction models as they might have limited prediction power (low variability). However, simply deleting all features with a low variability is dangerous as the few 'outliers' might contain valuable information about the target variables. An overview of all variables that have many identical values (> 65%) is given in Table 6. They will be further addressed in Chapter 4.

Table 6: Overview of categorical variables in the CCM truck data that contain mostly identical values.

Variable	DataType	% Identical	Value
Brand	Cat.	100,0%	DAF
Emission	Cat.	100,0%	EURO-6
Fuel Specification	Cat.	100,0%	Diesel EN590
Rear Axle Oil	Cat.	100,0%	Synthetic (ext)
Engine Type	Cat.	100,0%	EURO 6
Gearbox Oil	Cat.	99,9%	Synthetic (ext)
Road Type	Cat.	99,5%	On Road only
Area of Operation	Cat.	99,4%	W.-Europe (excl. Scandinavia)
Number of Drops per day	Cat.	99,3%	1 to 6
Contract Origin	Cat.	98,2%	New
Contract (version) start kms	Num.	95,4%	0
Contract Version	Cat.	95,0%	1
ADR Specification	Cat.	94,3%	No
Engine Oil	Cat.	93,9%	Synthetic (ext)
Type of Operation	Cat.	93,3%	Long Distance
(Semi-) Trailer Coupling	Cat.	92,5%	Fifth wheel
Body Type	Cat.	92,5%	Tractor Not Applicable
Axle configuration	Cat.	92,1%	4x2
Rear Axle Type	Cat.	84,2%	SR 1344
Application	Cat.	76,3%	General (dry freight, pallet loads)
Sub series	Cat.	70,5%	7
Gearbox	Cat.	69,8%	TraXon 12 speeds
Series	Cat.	69,1%	XF_F7_BH
Retarder System	Cat.	66,8%	ZF Intarder
Model	Cat.	66,5%	FT XF_F7_BH
Contract package	Cat.	65,6%	DAF MultiSupport Full Care

The CCM dataset consists of 76 features in total. However, not all of them are useful for modeling. Some features are duplicates of each-other or have no relation to the truck specification and usage whatsoever.

In collaboration with DAF knowledge experts, 26 features have been identified as irrelevant. They concern contract details such as the contract name and subsidiary. They do not provide information about the behaviour of the truck or it's configuration. in Table 7, the full list of irrelevant features is given.

Table 7: Overview of the irrelevant contract data features from the CCM database.

Feature	Feature
Forecasting Run Date	Contract end-year (actual)
Forecasting Report Date	Contract closing date
Subsidiary	Contract closing year
Contract Number	Contract (version) duration in months (original)
Contract Version	Contract (version) duration in months (actual)
Countract Group	Contract (overall) duration in months (actual)
Contract Name	Contract (overall) Age in months (actual)
Contract Birthdate	Currency
Contract (1st) activation date	Claim delay Date
Contract (version) Start date	Last date invoiced
Contract (version) activation date	First Registration Date
Contract (version) End date (original)	Vehicle Park Number
Contract end-date (actual)	Snapshot Contract (version) Status
Service dealer	Selling dealer
Brand	Emission

Furthermore, 6 redundant features have been identified. The truck model for example, is duplicated under different formats as being multiple individual features. In Table 8, the redundant features and their duplicates are listed.

Table 8: Overview of the redundant features and their duplicate in the CCM database.

Feature	Duplicate Feature
Selling dealer location code	Selling dealer
Default Service dealer location code	Service dealer
Series	Asset Type-info
Model	Asset Type-info
Country	Delivery country

In summary, the CCM dataset contains truck specification data of 2884 trucks. Of the 76 available features, many are redundant or insignificant for the purpose of this research. Collaborating with knowledge experts at DAF, 32 features have been identified as irrelevant, leaving 44 useful features. The remaining dataset is of reasonable quality. No outliers or incorrect values have been found. However, the dataset does contain a lot of missing values and some categorical variables contain predominantly

one value.

3.5.1 Graphical analysis

At first, an overview of the truck specifications is given for the connected trucks with CCM contracts. Since the introduction of DAF Connect, 2884 trucks with these contracts have been fitted with Connect (up to and including 21-08-2018). As can be seen in Figure 7, the vast majority of connected trucks is made up of DAF's XF truck, which is its largest available model. The CF and LF models comprise the remainder of sold trucks. The dataset contains 2728 XF's, 83 CF's and 73 LF's.

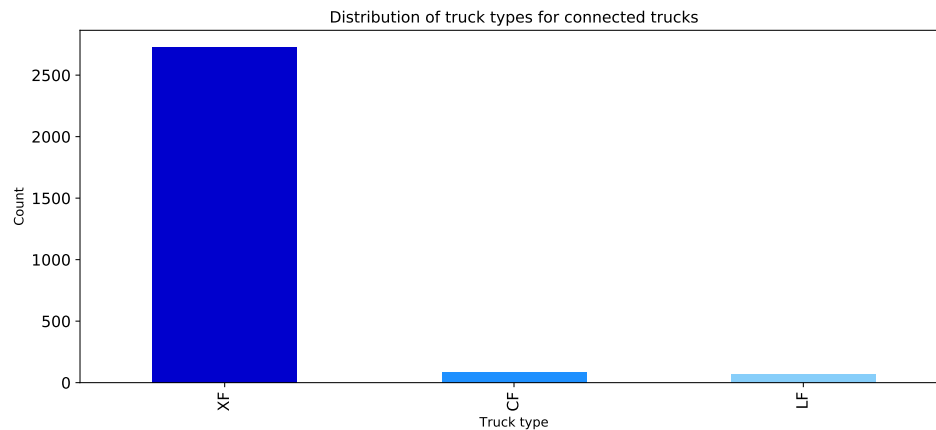


Figure 7: The distribution of trucks types for connected trucks (21-08-2018).

Looking at the distribution of chassis types for the connected trucks, there can be seen that a similar phenomenon occurs. From Figure 8, it shows that the FT is the most sold setup by far. The FT is DAF's standard 4x2 tractor truck, which is most often used for general truck and trailer activities. Other, less common chassis setups can be indicators for other types of truck usage.

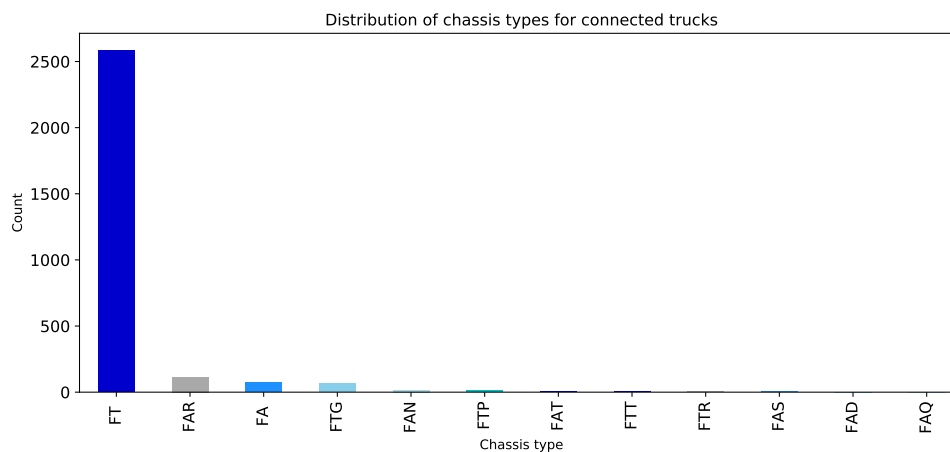


Figure 8: The distribution of chassis types for connected trucks (21-08-2018).

Another important feature is the time in service for each of the trucks. As explained in section 1.6, the time period over which the connect data is collected is limited. From the CCM database, there is derived that the majority of trucks that have been fitted with DAF connect have been in service for less than two years, which can be seen in Figure 9. On average, the connect trucks have been in service for 11 months (at 2018-08-21). As a result, the analysis and predictions in this research are limited to the early life of the trucks, as more data is simply not available.

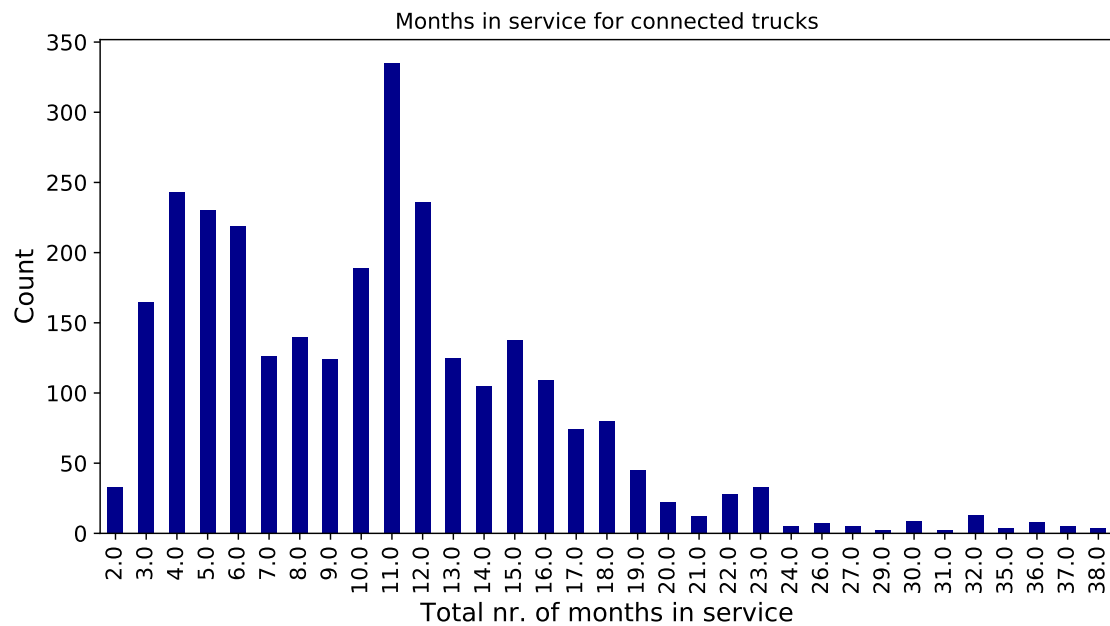


Figure 9: The number of months that the connected trucks are in service (21-08-2018).

The country in which the trucks operate can be of significance as well. Variables such as road conditions and fuel quality are, to a degree, country dependent. At the moment, as DAF connect is a new development, it has been deployed in Europe only. To give the reader a better view of the countries in which connected trucks operate, a density plot is made for the map of Europe, which is given in Figure 10. The color density of the countries indicates their number of active trucks. The most popular countries are Germany, France, Lithuania, Spain, Hungary and the Netherlands respectively.

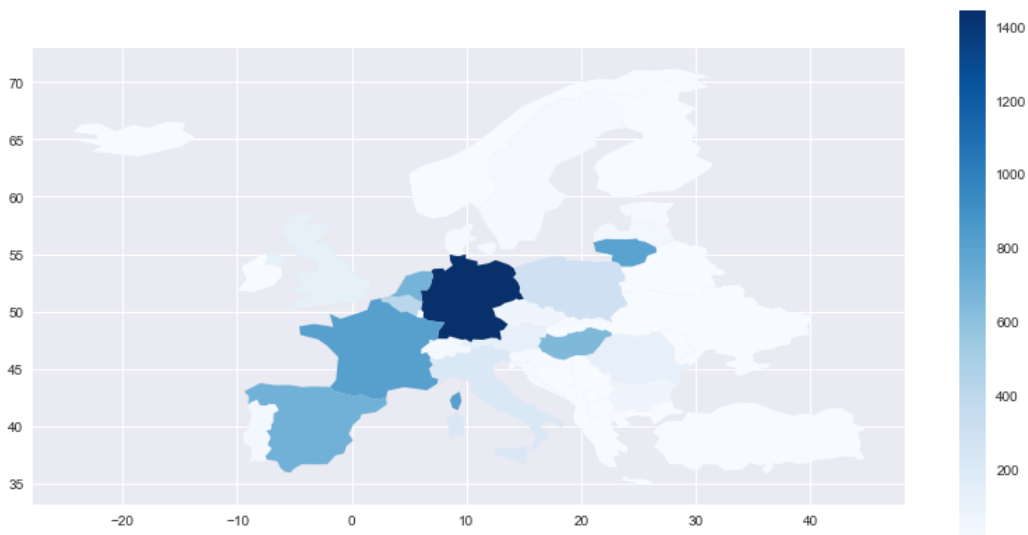


Figure 10: The distribution of connected trucks in Europe (21-08-2018).

For the interested reader, a range of other informative plots such as production rates, class distributions and the distribution of closed CCM contract types are given in Appendix F and G.

3.6 Descriptive analytics of the Mi Database

As explained in section 3.4, the Mi claim database provides information about the claimed truck repairs. An overview of the available data is given in this section. Again, it concerns all data up to and including 21-08-2018. As is the case for the CCM truck data, the features can be categorized into different groups, which are:

- Truck information
- Repair details
- Cost specification

The *truck information* is used to link the repairs to the corresponding CCM contract. Besides the VIN and chassis number, basic details such as the truck type, production date and the delivery country are given. For more specific truck information, the corresponding CCM data is consulted. More interesting are the *repair details* and *cost specifications*. They contain all necessary details about the repairs that are performed on the trucks. Besides the component that has been repaired and the dealer which made the claim, specifics such as labor hours, components used and the causes of the defects are stated. Furthermore, the date and time of the defect, repair and the claim itself are registered. Together with the repair details, the cost specifics are given. Labour hours, material prices and other costs are all reported. Now, DAF only reimburses claims when they satisfy all R&M conditions. This is why, at first, the claimed costs are registered. After the claim analysis, the paid amounts are added. The claim analysis is not binary. i.e. claims can also be denied and accepted partially. It is perfectly normal that half of the material costs are reimbursed while the other half is deemed to be outside of contract conditions. Similarly, it can happen that the material costs are paid while the labour hours are not. For the cost analyses in this project, only the paid repairs are taken under consideration as it is assumed that all unpaid claims have been rightfully rejected. To give the reader a complete overview of the available data in the Mi database, the features are listed in Table 9. Note that the features are separated per group, as specified above. For a description of the features, the reader is referred to Appendix B.

Table 9: Overview of the Mi data groups and the corresponding features.

Truck information	Repair details (1)	Repair details (2)	Cost specification
ChassisNr	ClaimCountry	Artnr1	MATT_CLAIMED
TypeName	ClaimDealer	Artnr2	LABOUR_CLAIMED
ProductRange	Claimnr	Artnr3	MISC_CLAIMED
productionsite	ClaimSort	Artnr4	TOTAL_CLAIMED
prodDate	FieldReportYN	Artnr5	MATT_PAID_DTNV
prodMonth	warrantycategory	labourcode1	LABOUR_PAID_DTNV
DeliveryDate	LastClaimStatus	labourcode2	MISC_PAID_DTNV
deliveryCountry	Laststatusstartdate	labourcode3	PAID_MIN_LANDED
deliveryDealer	DefectCode	labourcode4	LANDED_COSTS_DTNV
	defectcodedescription	labourcode5	TOTAL_PAID_DTNV
	defectcause	Misc1	ArticlePaidDN_LocalPolicy
	CausallPart	Misc2	LabourPaidDN_LocalPolicy
	DefectDate	Misc3	MiscPaidDN_LocalPolicy
	defectmonth	Misc4	HandlingPaidDN_LocalPolicy
	ClaimReceiveDate	Misc5	TotalPaidLocalPolicy
	Claimfinalised	ITSCode	HoursClaimed
	KmChassis	MonthInService	rejectioncode
	KmPart	DriveLineYN	acceptedclaimwarranty
	KindofPart	serviceproduct	AcceptedClaimLocalPolicy

Again, the numerical variables are summarized by the calculation of data characteristics such as mean, standard deviation and minimum and maximum values. The full numerical data summary is given in Table 10. A striking finding is the fact that claims sometimes contain negative repair costs. These negative costs seem to contradict the concept of repair (reimbursement) claims. However, consulting a claim assessment expert learned that they are actually debit costs resulting from wrongfully approved repair claims. Most often the corresponding (wrongfully) approved claim is found in the database as well. Do note however, that this is not always the case due to the fact that repairs can be debited only partially as well. Furthermore, some repairs have been performed before the truck has been in service altogether. This is due to the fact that a truck is considered to be in service, not when it leaves the production line, but when it has been prepared for customer use.

Table 10: Quantitative summary of all numerical variables in the Mi dataset.

Variable	min	max	mean	median	std
prodMonth	2014-01	2018-07	-	2017-07	-
defectmonth	2015-01	2018-08	-	2018-03	-
Claimfinalised	0	2018-08	-	2018-03	-
KmChassis	0	925	72	54	72
KmPart	0	925	7	0	27
MATT_CLAIMED	0	25713	267	54	965
LABOUR_CLAIMED	0	5431	133	76	236
MISC_CLAIMED	0	39533	72	4	601
TOTAL_CLAIMED	0	39533	472	220	1276
MATT_PAID_DTNV	0	25713	237	47	902
LABOUR_PAID_DTNV	0	5431	107	59	204
MISC_PAID_DTNV	0	6395	41	4	202
PAID_MIN_LANDED	0	25735	386	168	1067
LANDED_COSTS_DTNV	0	3034	30	1	107
TOTAL_PAID_DTNV	0	27053	416	176	1154
ArticlePaidDN_LocalPolicy	-1626	0	-2	0	33
LabourPaidDN_LocalPolicy	0	368	6	0	23
MiscPaidDN_LocalPolicy	0	2074	7	0	62
HandlingPaidDN_LocalPolicy	0	0	0	0	0
TotalPaidLocalPolicy	-1626	2074	11	0	77
HoursClaimed	0	57	2	1	3
MonthInService	-9	36	7	6	6
acceptedclaimwarranty	0	1	1	1	0
AcceptedClaimLocalPolicy	0	0	0	0	0
TOTAL_PAID_COMBINED	0	27053	427	189	1153

To perform a preliminary data quality analysis, the number of missing values has been evaluated for the Mi data as well. There are some variables that have a large number of missing values. However, they are mainly optional fields, providing a logical explanation of their number of missing values. The *artnr2*, *artnr3*, *artnr4* and *artnr5* features for example, are only filled when more than one component is replaced (i.e. more 'articles' are used). The same holds for the *misc* features, as they are only filled out when miscellaneous parts are used that are not in DAF's register. An exception is the *causal Part* feature, which should indicate the part that caused the defect. In practice this is most often left blank, while it could provide useful information. Furthermore, for a single claim the truck's delivery date and corresponding months in service are missing, which is an negligible amount. The full list of Mi features with missing values is given in Table 11.

Table 11: Overview of the Mi truck data variables containing missing values.

Feature	DataType	Nr. Null	% Null
Misc5	object	14,430	99.81%
labourcode5	float	14,414	99.70%
labourcode4	float	14,328	99.10%
Misc4	object	14,298	98.89%
Artnr5	object	14,012	96.92%
labourcode3	float	13,963	96.58%
CausalPart	object	13,956	96.53%
ITSCode	object	13,878	95.99%
Artnr4	object	13,604	94.09%
Misc3	object	13,031	90.13%
labourcode2	object	12,656	87.54%
Artnr3	object	12,119	83.82%
Artnr2	object	10,517	72.74%
Misc2	object	10,278	71.09%
Artnr1	object	7,052	48.78%
Misc1	object	6,686	46.24%
labourcode1	object	2,532	17.51%
DeliveryDate	datetime	1	0.01%
MonthInService	float	1	0.01%

Furthermore, of the 66 features in the Mi dataset, there are a few irrelevant features. They are either outside of the scope of this research, or already present in the CCM truck data (i.e. truck information). For example, the *claim dealers* and *defect code* are not related to the truck itself and therefore not of interest when predicting repairs. Furthermore, features such as the truck type and product range are already present in the CCM truck data. The full list of redundant and irrelevant features is given in Table 12.

Table 12: Overview of the irrelevant and redundant features from the Mi database.

Feature	Why	Explanation
TypeName	Redundant	Present in CCM data
ProductRange	Redundant	Present in CCM data
prodDate	Redundant	Present in CCM data
prodMonth	Redundant	Given in 'prodDate'
DeliveryDate	Redundant	Present in CCM data
deliveryCountry	Redundant	Present in CCM data
deliveryDealer	Irrelevant	Not truck related
ClaimCountry	Irrelevant	Not truck related
ClaimDealer	Irrelevant	Not truck related
Claimnr	Irrelevant	Not truck related
ClaimSort	Irrelevant	Not truck related
FieldReportYN	Irrelevant	Not truck related
LastClaimStatus	Irrelevant	Not truck related
Laststatusstartdate	Irrelevant	Not truck related
DefectCode	Irrelevant	Not truck related
defectmonth	Redundant	Given in 'DefectDate'
Claimfinalised	Irrelevant	Not truck related
KmPart	Irrelevant	Not truck related
KindofPart	Irrelevant	Not truck related
Artnr1	Irrelevant	Not truck related
Artnr2	Irrelevant	Not truck related
Artnr3	Irrelevant	Not truck related
Artnr4	Irrelevant	Not truck related
Artnr5	Irrelevant	Not truck related
labourcode1	Irrelevant	Not truck related
labourcode2	Irrelevant	Not truck related
labourcode3	Irrelevant	Not truck related
labourcode4	Irrelevant	Not truck related
labourcode5	Irrelevant	Not truck related
Misc1	Irrelevant	Not truck related
Misc2	Irrelevant	Not truck related
Misc3	Irrelevant	Not truck related
Misc4	Irrelevant	Not truck related
Misc5	Irrelevant	Not truck related
ITSCode	Irrelevant	Not truck related
rejectioncode	Irrelevant	Not truck related
acceptedclaimwarranty	Irrelevant	Not truck related
AcceptedClaimLocalPolicy	Irrelevant	Not truck related
prodCountry	Redundant	Present in 'Productionsite'

In summary, 39 out of the 66 available features in the Mi dataset are either redundant or irrelevant to this research. This is due to the fact that they do not hold information about the truck specifications or usage. Furthermore, there are multiple features with either missing values or predominantly one feature value. The remainder of the data is of decent quality, where especially the most important features such as repair costs, claim values and the defect/causal parts are accurately filed.

3.6.1 Graphical analysis

In total, 14,458 claims have been registered for the 2884 DAF Connect trucks with an R&M contract. Of these claims, 3170 claims are for repairs and 11,288 are maintenance, service, courtesy or field action claims. Courtesy claims are reimbursements that are not officially covered by the customer's contract(s) but are still reimbursed by DAF out of courtesy. Field action claims are claims that have been issued by DAF itself when, for example, a certain production batch has a faulty component installed and they are collectively replaced by DAF 'in the field'. Lastly, service and maintenance actions are planned and not part of unsuspected, unplanned repairs. For this research, we focus on the repairs claims only. Now, when looking at the 3170 repair claims, an average of 0.43 R&M claims and 0.67 warranty claims have been filed per truck. The actual distribution of claims per truck can be found in Figure 11. As can be seen from the calculations above, the number of repairs per truck is limited. This is logically explained by the fact that DAF's trucks are built according to a much higher life expectancy than the time that Connect has been available. In fact, DAF expects its trucks to have their repair and maintenance peak after no less than 48 months of service (which is halfway their expected lifetime of 96 months).

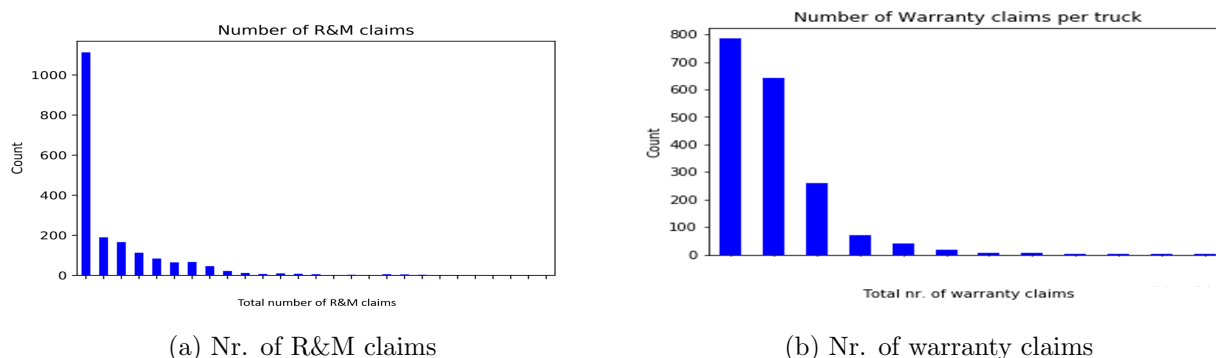
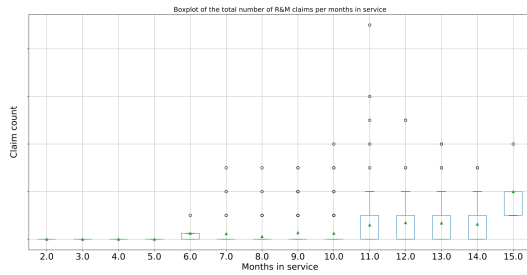


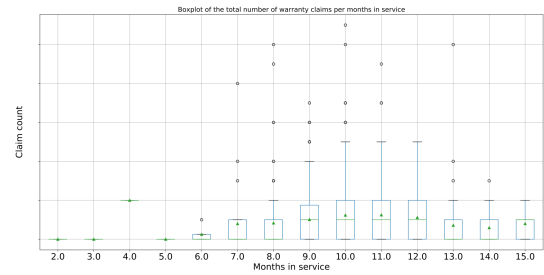
Figure 11: The distribution of total number of repair claims per truck (21-08-2018).

The numbers above include all trucks in the dataset, without any scope limitations. However, as we are only looking at the trucks that have been in service for 8 months or more and are delivered after 01-04-2017, it is more informative to look at the number of claims per months in service. During the time that a customer's warranty contract is active, repairs on the truck are handled as warranty claims. However, warranty contracts do not include wear parts. This means that repairs on components such as brake pads and clutches are not covered by warranty. Instead, they are covered by the R&M contracts (see section 3.3). This is why, when looking at the total repairs (claims) per truck, the sum of both claim types has to be considered. This sum is given as the *Total number of claims*.

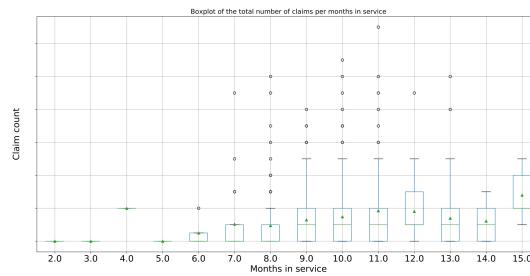
To provide a visual representation of the number of claims per months in service, boxplots are given in Figure 12. Multiple outliers are detected for the number of repairs per truck. As there can be numerous different reasons that could explain these excess number of repairs, they have been investigated in more detail before any actions were taken. Consultation with DAF experts resulted in the decision to not exclude any truck with many repairs (outliers) as they consisted out of legitimate claims from DAF's Mi claim database.



(a) R&M claims



(b) Warranty claims



(c) Total number of claims

Figure 12: Boxplots of the R&M claims (a), warranty claims (b) and total number of claims (c) per months in service (2018).

3.6.2 Repair types

Of course, the number of repairs only provides information on a high level. On a lower level, the type of repair and its corresponding costs provide a more substantial overview of repairs per truck. As we are looking at trucks that have been in service for 8 months or more, two contracts can be active on the trucks simultaneously, being the warranty contract as well as the R&M contract. Repairs on the driveline are automatically claimed through the warranty contracts while any other repairs are claimed through the R&M contracts. An overview of the distribution of claims on these contracts is given in Figure 13.



Figure 13: Distribution of contracts on which repair claims are made (21-08-2018).

Beside the contract on which a repair is claimed, further distinctions are made based on the nature of the repairs. Repairs can either be on the driveline or on the non-driveline. Driveline repairs encapsulate repairs that are performed on the engine, gearbox, differential and other parts which are directly involved with the power train of the trucks. Any other repairs are labeled as non-driveline repairs, which can range from the repair of light bulbs to the repair of electrical systems and chassis. Note that, for example, the engine and gearbox also contain parts that are not directly related to the driveline and are thus labeled as non-driveline components. Examples are the water pump system and oil cooling system. To provide the reader with an idea of the nature and quantity of claimed repairs, an overview of the 10 most common repairs on both the driveline and non-driveline are given in Figure 14.

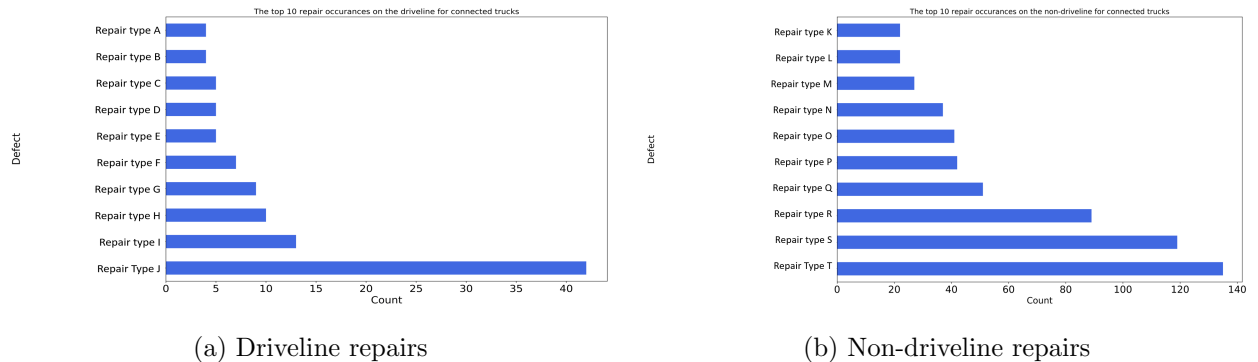


Figure 14: The top 10 most occurring repairs on both the driveline (a) and non-driveline (b).

As the name suggests, R&M contracts also include maintenance services. These services are planned on pre-defined time intervals and include standard maintenance actions such as oil top ups and vehicle condition checkups. Just as for repairs, these actions are recorded in the Mi database. However, they are not included in the repair overview above as they concern pre-defined actions, thus not adding value to the predictions in this research.

Now, to accurately predict the repair costs of trucks, the corresponding costs of these repairs are of key importance. The setup of claim entries is such that there are no pre-defined costs assigned to the different repairs. The number of labour hours and component costs differ per case and country. Instead, the average costs per repair are derived and used as indicators for the cost of the different repairs. In figure 15 an overview of the average costs for the most common repairs is given.

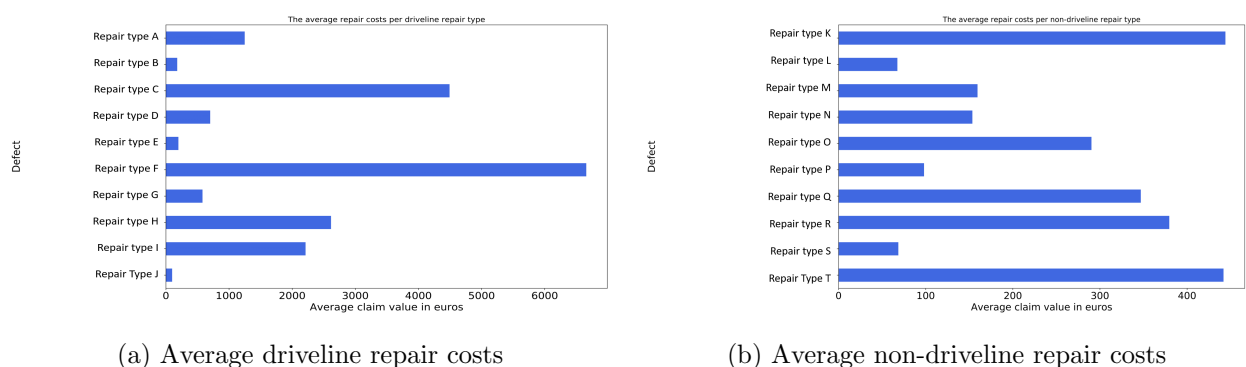


Figure 15: The top 10 most occurring repairs on both the driveline (a) and non-driveline (b) (21-08-2018).

Lastly, not only the number of repairs, but the total value of these repairs is of significance as well. In the end, they are of high interest for DAF as they represent the highest expense on the R&M and maintenance contracts. The 10 repairs with the highest combined repair costs are given in Figure 16. Note that the 'service' claims are not directly related to specific repairs and such have been excluded from this research.

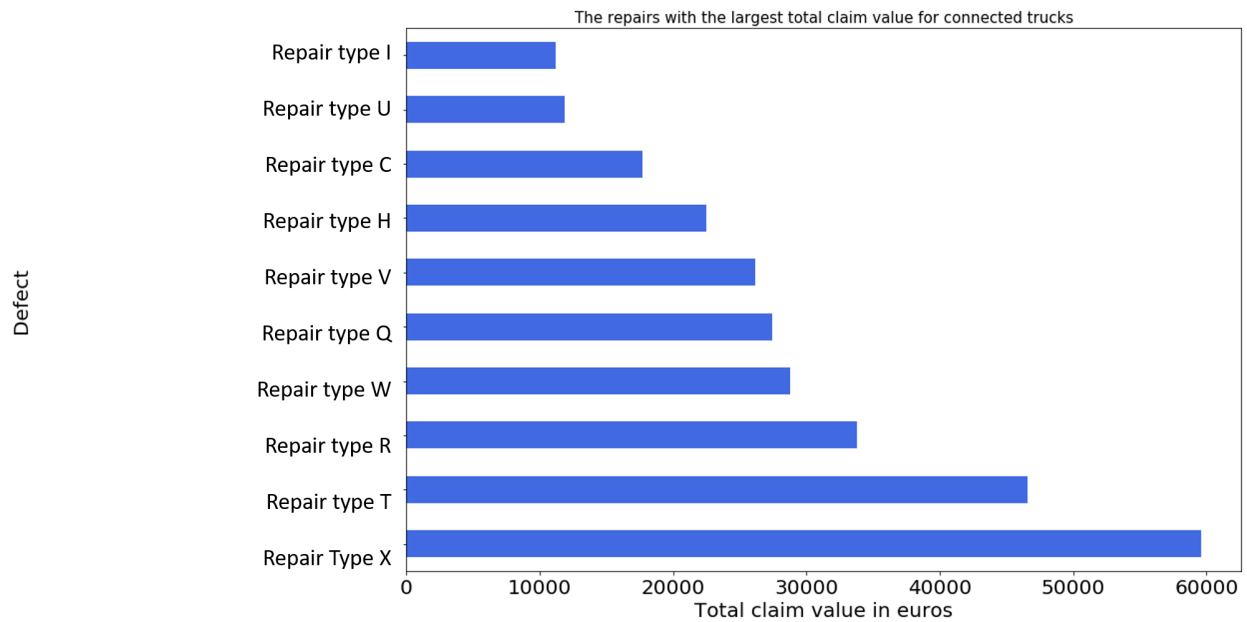


Figure 16: The top 10 of most expensive repairs from 01-04-2017 up to and including 21-08-2018.

3.6.3 Assumptions on the repair data

The sections above have provided an overview of the repair and claim information available at DAF. In collaboration with DAF experts, a few assumptions have been made on the data. They are listed below.

1. There is assumed that the claimed and reimbursed amounts accurately represent the costs of the repairs. Although fraud on the claimer's side can occur, it is assumed that DAF's claim review system is foolproof and thus the claimed values accurately represent the repair costs on the trucks.
2. Only the claims that have been reimbursed by DAF are taken under consideration as it is assumed that all unpaid claims have been rightfully rejected.

These assumptions are made as there is no detailed information available on why certain claims have been accepted or rejected. They are often individual cases which have been reviewed by claim analysts before being stored in the claim database.

3.7 Descriptive analytics of the connect data.

The last set to analyze is the connect data. It contains all truck usage information that has been collected from the operating trucks through DAF's connect data telemetry system (as explained in chapter 3.4). The two relevant subsets (trip and snapshot data), are analyzed in this section.

3.7.1 Trip data

The *Trip data* is a collection of data for each trip that a truck makes (where a trip is considered to be the time between the start and shutdown of the engine). It aggregates the data from the trip and sends it to the data-warehouse as a single instance. Valuable information such as the total brake duration, harsh brake duration, max throttle duration and fuel consumption per trip are found in this dataset. In total, 46 variables are recorded per trip. They are given in Table 13. For variable details, the reader is referred to Appendix C.

Table 13: Overview of the trip data.

Aggregated trip data measurements		
inputfiledate	brake_duration	maxthrottlepaddle_duration
datetime.begin	cruisecontrol_distance	dpabrakingscore_sum
unixtimestamp_begin	harshbrake_duration	dpaanticipationscore_sum
datetime.end	idling_duration	dpabrakingevent_count
unixtimestamp_end	gps_elevationloss	dpaanticipationevent_count
gpsdatetime_end	gps_elevationgain	cruisecontrol_fuelconsumption
gpslatitude_end	pto_count	gpsspeed
gpslongitude_end	pto_distance	cruisecontrol_distanceclass_1
gpsdatetime_begin	pto_duration	cruisecontrol_distanceclass_2
gpslatitude_begin	totalfuelconsumption_begin	cruisecontrol_distanceclass_3
gpslongitude_begin	fuellevel_begin	cruisecontrol_distanceclass_4
totaldistance_begin	totalfuelconsumption_end	cruisecontrol_distanceclass_5
totaldistance_end	fuellevel_end	vin
tripkey	gps_distance	month
dcmserialno	idling_fuelconsumption	
dcmversion	acceleration_duration	

Most of the features in the trip data are automatically aggregated before being stored in the connect database. Brake durations, idling durations and other similar features are all given as a total duration in seconds. Exceptions are the fuel level, trip duration and total fuel consumption, which are not aggregated automatically. The total fuel consumption for example, is calculated by subtracting the fuel level at the end of the trip from the fuel level at the beginning of the trip. In a similar manner the driven distance and trip duration are calculated. The VIN is used to link the trips to the corresponding CCM and Mi data of the trucks. To get a better understanding of the data, a quantitative summary including the minimum, maximum, mean, median and standard deviation of each of the numerical variables is given in Table 14.

Table 14: Quantitative summary of the numerical trip data.

Variable	min	max	mean	median	std
gpslatitude_end	33	255	51	49	25
gpslongitude_end	-10	255	10	7	31
gpslatitude_begin	33	255	51	49	25
gpslongitude_begin	-10	255	10	7	31
totaldistance_begin	0	318,201,310	62,156,352	55,179,350	45,053,079
totaldistance_end	0	318,315,425	62,200,438	55,224,953	45,061,477
brake_duration	-7	282,441,707	180	19	161,813
cruisecontrol_distance	0	808,520	21,705	0	55,434
harshbrake_duration	-3	1303	15	0	37
idling_duration	-6,896	576,173,923	1,113	190	541,719
gps.elevationloss	0	69,680	184	10	397
gps.elevationgain	0	69,772	198	12	407
pto_count	0	64	0	0	0
pto_distance	0	6410	0	0	5
pto_duration	0	21,153	13	0	182
totalfuelconsumption_begin	1	99,518,908	18,471,680	16,318,219	13,532,694
fuellevel_begin	0	100	70	71	26
totalfuelconsumption_end	7	99,522,303	18,484,618	16,330,960	13,535,013
fuellevel_end	0	100	67	67	27
gps_distance	0	29,167,437	42,837	541	87,060
idling_fuelconsumption	-179,999	202,297	294	123	2,187
acceleration_duration	-67,122,714	67,139,152	793	76	54,415
maxthrottlepaddle_duration	-56	67,133,010	81	0	38,461
dpabrakingscore_sum	0	17,805	260	0	564
dpaanticipationscore_sum	0	16,933	283	0	550
dpabrakingevent_count	0	196	4	0	8
dpaanticipationevent_count	0	314	5	0	9
cruisecontrol_fuelconsumption	-35,668	317,598	5,753	0	15,160
cruisecontrol_distanceclass_1	0	184,935	1	0	107
cruisecontrol_distanceclass_2	0	56,980	63	0	495
cruisecontrol_distanceclass_3	0	195,280	1,365	0	5,594
cruisecontrol_distanceclass_4	0	807,990	20,275	0	52,914
cruisecontrol_distanceclass_5	0	7,195	1	0	34

The quantitative summary already reveals some possible outliers and erroneous data. For example, *idling_fuelconsumption*, *brake_duration* and *maxthrottlepaddle_duration* have a negative minimum value, which is impossible in practice. Simultaneously, these features also have an unrealistically high maximum value. This becomes clear when (using Table 14) their mean and median values are compared with the corresponding minimum and maximum values. For example, The maximum for the full acceleration time duration during a trip (*maxthrottlepaddle_duration*) is found to be 67,139,152 seconds. This amounts to an unlikely long trip of 777 consecutive days without turning off the engine and applying maximum acceleration throughout. The mean value of 81 seconds however, is much more promising. Now, the same logic applies to the other features that contain such high feature values. The mean and median are often of a much lower value than the maximum, indicating a high likelihood of the presence of outliers, which are further analyzed in the data preparation chapter.

Furthermore, as is the case for the Mi and CCM datasets, the trip data contains irrelevant and redundant features as well. The date and time are captured in different formats and through different sources for each instance. The *inputfiledate*, *unixtimestamp_begin*, *unixtimestamp_end*, *gpsdatetime_begin*, *month* and *gpsdatetime_end* are redundant as they are already given by *datetime_begin* and *datetime_end* for each trip instance. Furthermore, the *fuellevel_begin*, *fuellevel_end*, *tripkey*, *dcmserialno*, *dcmversion* and *gpsspeed* are irrelevant for the repair predictions. The *tripkey* is merely used as a identifier, which also holds for the dcm serial number. The *dcmversion* holds information about the current connect software version in the truck, which is of no influence to the truck's behavior or setup. Lastly, the gps speed as recorded for each trip is of unknown origin. It doesn't in any way represent the average speed or other relevant information about the truck's speed during a trip. Thus, in total, 13 out of the 46 features are regarded to be redundant or irrelevant for further use.

Lastly, analysis showed that there are close to none missing values in the dataset. The only exception are the *gpsdatetime_end*, *gpsdatetime_begin* and *gpsspeed* where 0.50 to 1.50 percent of data is missing. They appear to be missing randomly and on random trucks without an evident logical explanation. however, as explained above, the gps datetimes and speed are not of direct relevance and thus, their missing values can be ignored altogether.

3.7.2 Snapshot data

The *snapshot data* collects sensory data from the operational trucks in five minute intervals, starting from the moment that the engine is turned on until the moment that it is turned off. The data consists out of actual sensor recordings at the time of measurement, giving information about the real-time state of the truck at that point in time. Each snapshot instance contains 108 features, ranging from fuel-, oil- and coolant levels to ambient air pressure and engine rpm's. A full list of variables is given in Table 15. For a detailed explanation of the variables, the reader is referred to Appendix C.

Table 15: Overview of the snapshot data.

Snapshot data variables		
snapshotkey	barometricpressure 1	engineload 1
inputfiledate	barometricpressure 2	engineload 2
datetime	barometricpressure 3	engineload 3
unixtimestamp	barometricpressure 4	engineload 4
gpsdatetime	barometricpressure 5	engineload 5
gpslatitude	fuellevel 1	enginespeed 1
gpslongitude	fuellevel 2	enginespeed 2
totaldistance	fuellevel 3	enginespeed 3
eventid	fuellevel 4	enginespeed 4
totalfuelconsumption	fuellevel 5	enginespeed 5
idle duration	fueltemperature 1	engineintakeairpressure 1
dcmserialno	fueltemperature 2	engineintakeairpressure 2
dcmversion	fueltemperature 3	engineintakeairpressure 3
gpsaltitude	fueltemperature 4	engineintakeairpressure 4
gpsheading	fueltemperature 5	engineintakeairpressure 5
fuellevel	engineoiltemperature 1	tachographspeed 1
aftertreatmentlevel	engineoiltemperature 2	tachographspeed 2
grosscombinationweight	engineoiltemperature 3	tachographspeed 3
wheelbasedspeed	engineoiltemperature 4	tachographspeed 4
tachographspeed	engineoiltemperature 5	tachographspeed 5
gps distance	engineoilpressure 1	totaldistance 1
enginecoolantlevel 1	engineoilpressure 2	totaldistance 2
enginecoolantlevel 2	engineoilpressure 3	totaldistance 3
enginecoolantlevel 3	engineoilpressure 4	totaldistance 4
enginecoolantlevel 4	engineoilpressure 5	totaldistance 5
enginecoolantlevel 5	enginecoolanttemperature 1	gpsspeed
engineoillevel 1	enginecoolanttemperature 2	ambientairtemperature
engineoillevel 2	enginecoolanttemperature 3	distanceuntilservice
engineoillevel 3	enginecoolanttemperature 4	enginecoolanttemperature
engineoillevel 4	enginecoolanttemperature 5	enginotalhours
engineoillevel 5	servicebrakeairpressure 1	servicebrakeairpressure1
aftertreatmentlevel 1	servicebrakeairpressure 2	servicebrakeairpressure2
aftertreatmentlevel 2	servicebrakeairpressure 3	gpshdop
aftertreatmentlevel 3	servicebrakeairpressure 4	eventname
aftertreatmentlevel 4	servicebrakeairpressure 5	tripkey
aftertreatmentlevel 5	month	vin

Almost all data in the snapshot dataset is of numerical origin (with the exception of some datetimes and the VIN). Again, a quantitative summary of the data is made, including the minimum, maximum, mean, median and standard deviation for each of the numerical variables. For readability purposes, the results have been moved to Appendix D.

The numerical summary revealed some errors in the data, which are described here. The features *engineoiltemperature*, *enginecoolanttemperature*, *fueltemperature*, *ambientairtemperature*, *gpsaltitude*, *ser-*

vicebrakeairpressure, *grosscombinationweight*, *engineintakeairpressure* and *idle_duration* show exceptionally high maximum values when comparing them to their mean and median. This indicates that they are likely to be outliers. Furthermore, fluid level measurements for *aftertreatmentlevel* and *engineoillevel* can reach levels of over 100%, which might be caused by incorrect calibration or swaying of the fluids during driving. Lastly, gps sensors record a value of 255 (which is not a coordinate) when no gps signal is available at the time of measurement. In general, many possible outliers are present in the dataset, which are analyzed in more detail in the next chapter.

The presence of missing values has been evaluated. All features that have been measured in one minute intervals contain an ascending number of missing values, where the first minute measurement contains roughly ten percent of missing data and the fifth (and last) minute measurement contains roughly thirty percent of missing data. The resulting correlation between the measurement minute and the percentage of missing data was investigated which revealed that most of the missing data was caused by early stopping when a truck was turned of. If a truck is turned of down in the middle of a five minute measurement period, the incomplete data sample is sent to DAF anyway, leaving missing values for all measurements for the remainder of the time-frame. Excluding these incomplete samples resulted in a dataset with only a few arbitrarily missing variables ($< 0.2\%$).

Also, most of the *enginototalhours* and *distanceuntilservice* measurements are missing (82%). Lastly, the features *ambientairtemperature*, *enginecoolanttemperature*, *gpsspeed*, *servicebrakeairpressure1* and *servicebrakeairpressure2* contain a limited number of missing values (0.2%). The full overview of missing values per variable is given in Table 16.

Table 16: Overview of the variables with missing values in the Snapshot data.

Variable	Nr. missing	% Missing	Variable (2)	Nr. missing (2)	% Missing (2)
distanceuntilservice	2,251,315.00	82.4	enginecoolanttemperature_3	483,018.00	17.7
enginototalhours	2,251,315.00	82.4	barometricpressure_3	483,007.00	17.7
fuellevel_5	884,399.00	32.4	enginecoolantlevel_3	482,948.00	17.7
engineoiltemperature_5	881,006.00	32.2	engineoilpressure_3	482,902.00	17.7
fueltemperature_5	881,001.00	32.2	engineload_3	482,815.00	17.7
engineoillevel_5	880,970.00	32.2	enginespeed_3	482,813.00	17.7
engineintakeairpressure_5	880,961.00	32.2	fuellevel_2	435,267.00	15.9
totaldistance_5	879,597.00	32.2	engineoiltemperature_2	431,275.00	15.8
tachographspeed_5	879,541.00	32.2	fueltemperature_2	431,235.00	15.8
servicebrakeairpressure_5	878,316.00	32.1	engineoillevel_2	431,161.00	15.8
aftertreatmentlevel_5	878,293.00	32.1	engineintakeairpressure_2	431,099.00	15.8
barometricpressure_5	878,269.00	32.1	totaldistance_2	429,343.00	15.7
enginecoolanttemperature_5	878,268.00	32.1	tachographspeed_2	429,094.00	15.7
enginecoolantlevel_5	878,234.00	32.1	servicebrakeairpressure_2	427,621.00	15.6
engineoilpressure_5	878,233.00	32.1	aftertreatmentlevel_2	427,559.00	15.6
engineload_5	878,209.00	32.1	enginecoolanttemperature_2	427,528.00	15.6
enginespeed_5	878,205.00	32.1	barometricpressure_2	427,516.00	15.6
fuellevel_4	539,632.00	19.7	enginecoolantlevel_2	427,471.00	15.6
engineoiltemperature_4	535,800.00	19.6	engineoilpressure_2	427,407.00	15.6
fueltemperature_4	535,788.00	19.6	engineload_2	427,282.00	15.6
engineoillevel_4	535,729.00	19.6	enginespeed_2	427,268.00	15.6
engineintakeairpressure_4	535,721.00	19.6	fuellevel_1	284,807.00	10.4
totaldistance_4	534,069.00	19.5	engineoiltemperature_1	280,567.00	10.3
tachographspeed_4	533,971.00	19.5	fueltemperature_1	280,554.00	10.3
servicebrakeairpressure_4	532,409.00	19.5	engineoillevel_1	280,444.00	10.3
aftertreatmentlevel_4	532,407.00	19.5	engineintakeairpressure_1	280,421.00	10.3

barometricpressure_4	532,371.00	19.5	totaldistance_1	278,125.00	10.2
enginecoolanttemperature_4	532,366.00	19.5	tachographspeed_1	277,940.00	10.2
enginecoolantlevel_4	532,319.00	19.5	aftertreatmentlevel_1	276,476.00	10.1
engineoilpressure_4	532,306.00	19.5	servicebrakeairpressure_1	276,449.00	10.1
engineload_4	532,260.00	19.5	barometricpressure_1	276,404.00	10.1
enginespeed_4	532,255.00	19.5	enginecoolanttemperature_1	276,403.00	10.1
fuellevel_3	490,472.00	17.9	enginecoolantlevel_1	276,301.00	10.1
engineoiltemperature_3	486,615.00	17.8	engineoilpressure_1	276,278.00	10.1
fueltemperature_3	486,581.00	17.8	engineload_1	276,203.00	10.1
engineoillevel_3	486,482.00	17.8	enginespeed_1	276,199.00	10.1
engineintakeairpressure_3	486,453.00	17.8	ambientairtemperature	5,597.00	0.2
totaldistance_3	484,799.00	17.7	enginecoolanttemperature	5,597.00	0.2
tachographspeed_3	484,593.00	17.7	gpsspeed	5,597.00	0.2
servicebrakeairpressure_3	483,096.00	17.7	servicebrakeairpressure1	5,597.00	0.2
aftertreatmentlevel_3	483,055.00	17.7	servicebrakeairpressure2	5,597.00	0.2

The features *servicebreakairpressure1* and *servicebreakairpressure2*, besides containing many faulty measurements, are redundant as they are measured by the other service airbreak sensors as well. The same redundancy holds for the features *gps distance*, *month*, *enginecoolanttemperature*, *fuellevel*, *aftertreatmentlevel*, *wheelbasedspeed*, *tachographspeed* and *gpsspeed*. The *distanceuntilservice* feature holds no information about the truck's condition and is therefore regarded irrelevant. The same holds for the features *snapshotkey*, *inputfiledate*, *unixtimestamp*, *gpsdatetime*, *dcmserialno*, *dcmswversion*, *gpsheading*, *eventid*, *eventname*, *datetime*, *gpshdop*, *totaldistance_1*, *totaldistance_2*, *totaldistance_3*, *totaldistance_4*, *totaldistance_5* and *tripkey*. lastly, the *VIN* is used as an identification variable only. Thus, 29 out of 108 features are removed from the snapshot dataset.

3.8 Chapter Summary

The full dataset contains data for 2884 trucks. Together, these trucks have had 14.458 claims of which 3170 were actual unplanned repairs on truck components, which are all stored in the Mi database. Both the number and type of claims as well as the associated costs for these claims have been recorded by DAF. Furthermore, truck specification data and truck usage data has been collected from three different datasources at DAF. the CCM data is used to retrieve truck setup and specifications and is of categorical origin (engine type, axle configuration, truck model etc.) Furthermore it contains some information regarding the expected truck operations (type of cargo carried, number of drops per day etc.). The Snapshot data and Trip data has been used to derive truck usage based on actual operational data of the trucks (through telemetry) and are of numerical origin. In total, 230 features are available per truck. However, the data contains quite a few variables that are redundant, irrelevant or of poor quality regarding the number of missing and erroneous values. Furthermore, some variables are of low cardinality, up to a set of just 1 for some categorical variables, depriving them of any predictive power. These findings have been elaborated in this chapter and are further addressed in Chapter 4.

4 Data preparation

In this chapter, the data pre-processing and cleaning steps that have been taken are elaborated. They follow the steps as found in the 'data cleaning' phase of the CRISP-DM model (Wirth, 2000). The data cleaning steps for this framework are given in Appendix E. Cleaning of the data is a crucial step as "(1) real-world data is impure; (2) high-performance mining systems require quality data; and (3) quality data yields concentrative patterns" (Zhang et al., 2003). In general, low quality data leads to low quality results (Zhang et al., 2003).

4.1 Selecting the relevant trucks and data

The first step in the data selection process was to exclude all of the data that falls outside of the scope of the project. The total list of connected trucks with R&M contracts consisted out of 2884 trucks at 21-08-2018. However, many of these trucks did not fit the scope. They were either too young to be able to validate the number of failures over the specified time-frame, or that old that they did not contain the correct software for data collection (based on expert knowledge at DAF). The data from DAF's connect system only became reliable for trucks that had been produced from 01-04-2017 or later, while the scope of the project required the trucks to be in service no later than 21-12-2017. This resulted in a set of 1099 trucks that were available for analysis.

Furthermore, data exploration revealed numerous irrelevant, redundant or poor quality features which have been removed with the help of knowledge experts at DAF. From the CCM database, 32 out of 76 features have been removed as they contained irrelevant contract details or redundant information. From the trip database, 13 out of 46 features have been removed from the trip database and 29 out of 108 features have been removed from the snapshot dataset. They were irrelevant or redundant as well. A detailed explanation of the removal of each of these features is given in Chapter 3,

Lastly, the data from the Mi database is only used to retrieve the type, amount and costs of repair claims that have been made on the trucks. This is due to the fact that any truck information present in this database has been derived from the CCM or Connect databases. The features *TotalPaidLocalPolicy* and *TotalPaidDTNV* contain the total of repair costs per claim which have been reimbursed by DAF. The type of repair is given by the information in *DefectCode*, *defectcodedescription* and *warrantycategory*, where the defect code defines the component (group) that has been repaired and the warranty category specifies the category of the repair (e.g. warranty or R&M contract).

In conclusion, the initial data selection resulted in a set of 1099 trucks that were available for analysis and a reduction of the number of features from 230 to 156.

4.2 Classifying the trucks based on repairs

Together with experts at DAF, there has been decided to construct binary classification models. The reason for this is that only a limited set of trucks is available for analysis (as DAF Connect is rather new) while many different types of repairs exist. Furthermore, the majority of trucks has either 0 or 1 repair during the first year of service (for more details on the number and type of repairs we refer the reader back to Chapter 3.6). As a result, there is not enough information available to predict the exact number of repairs or the associated costs. Instead a binary decision is made which predicts if a truck has more or less than the average number of repairs over the time horizon under consideration. Trucks with less than the average number of repairs are labeled 0, while trucks having more than the average number of repairs are labeled 1.

With this classification task, there can be derived which trucks require more intensive repair and maintenance and which trucks do not. This can subsequently be used to substantiate R&M contract costs and identify risk vehicles regarding R&M costs.

4.3 Data cleaning

This section describes the outliers, missing values, inconsistencies and noise that has been found in the data, together with the methods that are used to deal with them.

4.3.1 Outliers

Data exploration revealed numerous errors in the data measurements regarding pressures, weight, fuel levels, temperatures and so on. However, the physical limitations of the trucks and material/fluids inside are well known at DAF. This information is leveraged to detect and remove the most obvious outliers. Together with the DAF experts, baseline values have been established for each of the features in the connect database. The minimum and maximum of acceptable values have been determined and any measurements that fall outside of those boundaries have been removed. To clarify, an example is given for the detection of outliers in the fuel temperature that is measured for a specific truck. In Figure 17, the fuel temperature measurements during a trip are given in five minute intervals. The outlier thresholds have been set to a minimum of 0 degrees celsius and maximum of 120 degrees celsius. For the given trip, one outlier is found that surpasses the maximum threshold. The measured temperature of 217 degrees celsius would have spontaneously ignited the fuel and is therefore classified as an outlier. The same logic has been applied for the other connect data features. When outlier boundaries could not be established with the help of DAF experts, the 1.5 interquartile rule for boxplots are used to detect outliers. They provide a interpretable outlier detection rule and do not require the data to be normally distributed as the boxplot depends on the median instead of mean value of the data (Walfish, 2006). The boxplots for the connect data features are given in Appendix H. Together with the numerical analysis in Section 3.7 they provide an intuitive overview of the measurements that are found in the Connect database. Subsequently, a list of established outlier boundaries using the 1.5 IQR Rule is given in Appendix I

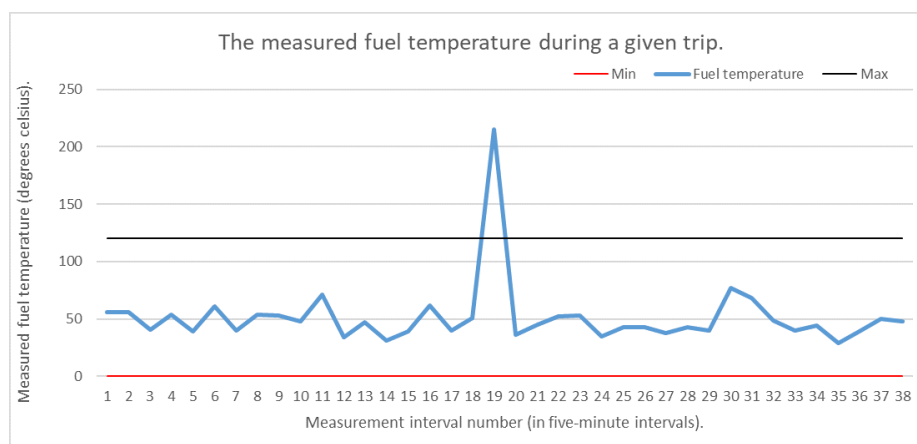


Figure 17: Example of temperature outlier analysis using predefined baseline boundaries.

The CCM dataset did not contain any outliers as each entry consisted of a set of categorical features that have been selected from a predetermined range of options upon R&M contract closing with the customer.

4.3.2 Missing values

Beside outliers, data exploration revealed numerous missing values as well. For the CCM database it has been established that it is a result of the fact that the value Nan (Not a number) is filed when a component is not fitted to the truck (e.g. the soot filter or tail-lift are almost never fitted). Therefore, there is assumed that a missing value for truck specification and truck usage features in the CCM database correspond to 'component not fitted'. Furthermore, for three trucks, the delivery date is missing. There has been decided to ignore those trucks as essential information about their age is missing.

For the snapshot and trip datasets, the data is of numerical origin and collected from the operating trucks in real time. As a result it is highly likely that missing values are a result of faulty or non-active sensors on the trucks. During the data exploration phase, it became clear that there were numerous missing variables for each of the features in the snapshot dataset (Table 16). The first step when encountering missing values is to determine if the data is missing at random or that an underlying cause can be identified (Fallis, 2013). Analyzing some sample datasets and consultation with experts revealed that many of the instances from the snapshot data with missing values had an underlying cause, as they were heavily dependent on the *eventname* of the snapshot message. Snapshots are sent in five minute intervals but also when the truck is started and stopped. Of course, at this point the truck is not yet driving and thus, no information except basics such as the datetime and current location are transmitted. For this reason, all snapshot messages with an *eventname* feature value other than 'TIME TRIGGER', and thus not recorded during operation, are removed from the dataset.

After removal of the incomplete snapshot messages, some missing values still remained. However, they were missing at random and thus, measures to deal with randomly missing values had to be taken. According to (Han et al., 2011) missing values can either be ignored, filled in manually, replaced by a measure of central tendency or changed to a global constant. Now, as the dataset under consideration is of considerable size, and removal of instances with missing values would result in loss of other information as well, there has been decided to fill each missing value with the mean value of the attribute under consideration, where the mean of the attributes is derived for each truck individually.

4.3.3 Inconsistencies and noise

As elaborated, the snapshot data is captured in five minutes intervals. However, selected measurements are executed in one minute intervals instead of the regular five. They are stored within the truck and sent as individual features in the snapshot message. This inconsistency is actually beneficial as the extra measurements can be leveraged to retrieve more detailed truck operation information. They are used to derive approximations for probability distributions of measurements as explained in Section 4.4.2.

The snapshot data also contained noisy information caused by faulty sensor measurements. When a sensor is not able to do an accurate measurement at the time of the snapshot creation, it registers a default value (e.g. gps latitude: 255 degrees or service break air-pressure: 65535 Kpa). Fortunately, these default values are easily recognizable as outliers due to their extreme values and thus dealt with appropriately during outlier removal using expert knowledge and the box-plot interquartile range.

Furthermore, the trip dataset contained many trip measurements with a total duration of zero minutes or without any distance driven. This is simply explained by the fact that truck drivers can switch the key on their trucks without driving. To avoid these noisy data inputs, only trips with a total distance of more than two kilometers have been taken into account. This number has been chosen based on expert knowledge. Local transport (e.g. city distribution) can include many short trips, but two kilometers of driving is enough to assume that a trip to a new location has been made.

Another inconsistency in the trip data is the aggregation of trip information. Although most variables contain aggregated data which can be used for analysis directly, a few exceptions are present. The trip duration, distance driven and fuel consumed are not directly present in the dataset but have to be derived from the data. The trip duration for example is derived by subtracting the *datetime.begin* from the *datetime.end* feature. In a similar fashion, the trip distance is derived from the features *totaldistance.begin* and *totaldistance.end*. Lastly, the features *totalfuelconsumption.begin* and *totalfuelconsumption.end* have been used to derive the fuel consumption used. As the derived features replace the features that have been used for their derivation, the total number of features in the trip dataset is reduced from 33 to 30 features.

The Mi claim dataset contained noise in the form of rejected claims. A repair claim could be filed and paid, after which it has been rejected later in the verification process. In these cases, a rejection code is added to the claim file and the paid amount is recovered by DAF. To avoid these invalid claims, any claim that has its paid repair costs reimbursed in full is removed from the dataset. Furthermore, all claims that are not related to repairs (i.e. maintenance claims, service actions and inspections) have been removed from the dataset, such that only relevant repairs remained.

Lastly, a generalization has been made for the one minute data in the snapshot dataset (see Section 3.7.2). As they contained arbitrarily missing values and were measured in a different time interval, the median of the measurements during each five minute interval has been derived and used as the replacement value for that specific measurement interval.

4.4 Feature derivation and extraction

Several operations had to be executed on the data in order to create meaningful variables for classification. They include feature extraction, attribute derivation and data aggregation, which are described in this section.

4.4.1 Deriving truck usage features from the datasets

The trip data has both timestamps and aggregated feature values based on the duration of a trip. Therefore, to make measured trip variables time independent, some derivations have been made. The trip start time-stamp and the trip end time-stamp are used to derive the trip duration for each entry in the trip dataset by subtracting the start time from the end time. In a similar fashion, the total distance at end and the beginning of each trip are used to determine the driven distance per trip. These trip durations and distances are consecutively used to transform the aggregated measurements (i.e. brake duration, fuel consumption etc.) into averaged measurements per kilometer. By doing this, the measurements become independent of the trip duration or length, thus providing a generalized/uniform measurement of driving behavior for each truck. The purpose of this generalization is to provide DAF with possible business insights regarding driving behavior and repairs. If they prove to be valuable repair predictors, they can be used to relate maintenance requirements and expected repairs to driving behavior. The first month of operation for each truck has been used to derive the truck usage features. It could be the case that multiple drivers are linked to a single truck. However, there is assumed that the first month of operation allows for enough time to find the usage profile based on these multiple drivers. This has been assumed because there is no data available to verify when a truck changes from one driver to the other (i.e. driver information is not stored).

To clarify, a simple example is given. Imagine that we want to score two drivers on their braking behavior. Driver one made a trip of 1000 km, while driver two drove only 100 km. The measured brake_duration of driver one is 500 seconds, while driver two's brake_duration is 55 seconds. The average score is then derived by dividing the brake durations by the driven distance. Driver one's score then becomes 0.5 seconds per kilometer, while driver two gets a score of 0.55 seconds of braking per kilometer. Their behavior is now compared in a generalized way. The same logic has been applied to the other aggregated trip measurements in the trip dataset.

In addition, the trip measurements such as the above described breaking duration have been summed over the entire month of truck operation measurement such that the total score is given. Also, the mean, standard deviation and skewness of the trip measurements has been derived. They contain information on the driving behavior dependent on distance and time, which could hold additional information about the expected number of repairs. Lastly, the number of trips during the first month of operation has been calculated by counting of the number of trip measurements for each truck. This provides information about the expected number of trips that a truck will make over the prediction period.

In summary, for every feature from the trip data (except for the redundant features as elaborated in Section 3.7.1), the value per driven kilometer has been derived, the sum of the measurements over the first month of operation has been calculated and the mean, standard deviation and skewness of the measurements has been calculated for each truck.

4.4.2 Fuzzy histograms and feature extraction

The snapshot and trip data by themselves can't be used as model input directly. In order to use the time series data for classification tasks, relevant information has to be extracted from them (Rodríguez

and Alonso, 2004). Global characteristics are retrieved for each truck's (snapshot) time-series measurements. The mean, maximum, minimum standard deviation and skewness values are derived as they are commonly used metrics to describe the global characteristics of time-series data (Yang and Létourneau, 2005)(Baydogan et al., 2013)(Nanopoulos et al., 2001). Furthermore, the number of snapshots has been calculated and added as a feature as it provides information about the expected hours of operation during the period under consideration (a snapshot counts for five minutes of operation time).

Furthermore, a more advanced method for feature extraction has been applied in the form of fuzzy histograms. Fuzzy histograms are a fuzzy generalization of crisp histograms and used to describe the probability distribution function properties of the time-series measurements of the trucks. They contain a much higher level of statistical efficiency compared to the regular crisp histograms, while maintaining a high level of computational efficiency Waltman et al. (2005). Due to the overlap in the fuzzy sets, they better approximate probability distribution functions than the regular crisp histograms (van den Berg et al., 2004).

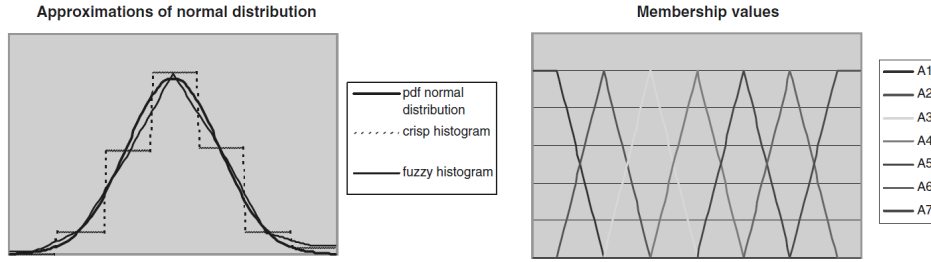


Figure 18: A crisp histogram (left), compared to a fuzzy histogram (right) with overlapping membership functions (van den Berg et al., 2004).

Based on (Waltman et al., 2005), the fuzzy histograms are explained as follows:

For samples of $x(1), \dots, x(n)$ with a random sample size n , The corresponding probability distribution function $\hat{f}(x)$ is estimated by a fuzzy histogram as follows:

$$\hat{f}(x) = \sum_i \frac{p_i \mu_i(x)}{\int \mu_i(x) dx'} \quad (4)$$

where p_i is given by:

$$p_i = \frac{1}{n} \sum_{j=1}^n \mu_i(x_j) \quad (5)$$

and μ_1, \dots, μ_i are the membership functions that describe the fuzzy partitioning. Therefore, the sum of their membership to the membership functions is equal to 1 for all x :

$$\sum_i \mu_i(x) = 1 \forall x \in \mathbb{R} \quad (6)$$

For a more detailed discussion on fuzzy histograms, the reader is referred to (Kaymak et al., 2003) and (van den Berg et al., 2004).

The membership functions that are used are triangular as they are easy to understand, commonly used in practice and capture the desired properties well (Barua and Kosheleva, 2014). Each triangular membership function $\mu_i(x)$ is defined by a lower limit a , an upper limit b and value m :

$$\mu_i(x) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{m-a}, & a < x \leq m \\ \frac{b-x}{b-m}, & m < x < b \\ 0, & x \geq b \end{cases} \quad (7)$$

This method has been tested with 5,7 and 9 membership functions (bins) respectively in order to evaluate their effect on the model performance. This was done to compare the effect of different numbers of fuzzy bins on the model performance as there is no unambiguous definition for the optimal number of bins. The tests were performed on several trip data and snapshot data measurements. The list of features over which the fuzzy bins have been used to extract features is given in Appendix O. These features have been selected as they were found to have a high feature importance during modeling of the base models. Experiments on the modeling performance showed that the best performance was reached using 5 fuzzy bins and therefore, the models with fuzzy bins in this research are made using this number of bins. An overview of the experiment results is given in Appendix J. Easy to understand linguistic terms such as 'very low', 'low', 'average', 'high' and 'very high' have been used for the membership functions in order to retain feature interpretability. An example of a fuzzy histogram with five bins for the distribution of truck speed measurements is given in Figure 19.

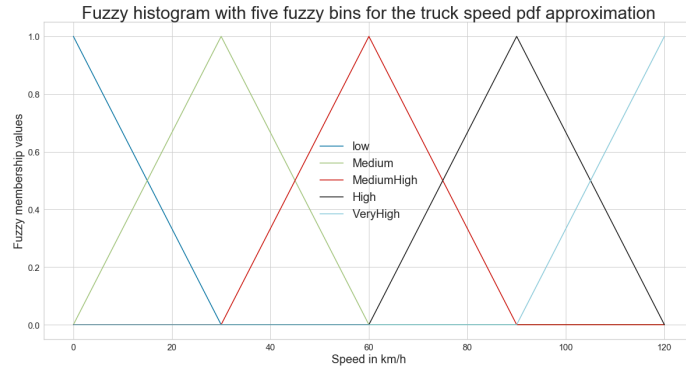


Figure 19: The fuzzy histogram for the distribution of truck speed measurements.

The membership degrees are counted for all the measurements of the time-series feature and truck under consideration, aggregated per bin and subsequently normalized, resulting in a feature vector that describes the PDF of the measurements.

To clarify, a short example for three speed measurements of an arbitrary truck is given based on the bins in Figure 19 (for this example, we assume that the three observed measurements are the complete set of measured truck speeds):

Step 1. Find the values of the measurements:

Measurement 1: 80 kmph
Measurement 2: 30 kmph
Measurement 3: 50 kmph

Step 2. Calculate the membership degrees to the bins:

Measurement 1: Membership High = 0.65, Membership MediumHigh = 0.35
Measurement 2: Membership Medium = 1
Measurement 3: Membership Medium = 0.5, Membership MediumHigh = 0.50

Step 3. Aggregate the membership values for all measurements:

Total membership to Medium: $1 + 0.50 = 1.50$
Total membership to MediumHigh: $0.35 + 0.50 = 0.85$
Total membership to High: 0.65

Step 4. Normalize the membership values to 1:

Normalized membership to Medium: 0.50
Normalized membership to MediumHigh: 0.28
Normalized membership to High: 0.22

These normalized membership functions have subsequently been used as additional input features for modeling.

4.4.3 Data integration and formatting

The data from the trip database, snapshot database, R&M contract database and MI (repair) database are easily integrated into a single dataset as the VIN (identification number) of the truck is present in each of the datasets. Merging them on the VIN results in the formation of the final dataset that is used for analysis. When formatting of the data has been required for modeling, it is explained in the next chapter together with each of the models that have been used for analysis.

4.5 Final dataset

Data has been selected based on the scope of the project and data availability. The scope has been determined to be the trucks that have been produced from 01-04-2017 or later, while they had to be in service no later than 21-12-2017. This resulted in a set of 1099 trucks that have been used for analysis. For each of these trucks, a set of 156 features remained after data cleaning.

Now, as DAF is interested in the potential of the available data in the future, two datasets have been derived for these trucks. One set contains all trucks that have been in service for 8 months or more such that predictions could be made over a time horizon on 8 months. This resulted in a set of 793 trucks. Subsequently, a set that contains all trucks that have been in service for 11 months or more has been derived, such that the repair predictions could also be made over a time horizon of 11 months. The difference in obtained results was then used to analyze the effects of a longer prediction time horizon on the modeling performance. This resulted in a set of 342 trucks.

The average number of repairs in the 8 months ahead dataset was found to be 1.06. Therefore, trucks that had one or more repairs were classified as having 'many repairs', while trucks with less than 1 repair were classified as having 'few (no) repairs'. This resulted in a nicely balanced dataset that could be used for analysis.

The average number of repairs in the 11 months ahead dataset was found to be 1.70. Therefore, trucks that had 2 or more repairs were classified as having 'many repairs' and trucks with less than 2 repairs were classified as having 'few repairs'. This resulted in a more or less (40%-60%) balanced dataset. Although some imbalance is present, the minority class is not substantially underrepresented. Furthermore, under-sampling and over-sampling have limitations of their own. Under-sampling reduces the already small dataset even further, and might remove important samples from the set, while oversampling can cause the models to over-fit on the duplicated instances. Lastly, Estabrooks et al. (2004) showed that the optimal ratio between the two classes doesn't have to be 50-50 and different sample ratios could even improve results. Nonetheless, caution has to be paid when analyzing the accuracy of the models formed with this dataset as some bias towards the majority class could be formed. Therefore, the precision, recall, f1 and Kappa score are also analyzed, as together, they provide a good indication of the model's ability to classify both classes accurately.

4.6 Chapter summary

Data has been selected based on the scope of the project and data availability. The scope has been determined to be the trucks that have been produced from 01-04-2017 or later, while they had to be in service no later than 21-12-2017. Two datasets have been created. The first set contains 793 trucks (8 months time horizon for prediction). The second set contains 342 trucks (11 months time horizon).

For these trucks, irrelevant, redundant and poor quality features have been removed, resulting in a reduction of 230 features to 156 features per truck. For these features, expert knowledge and box-plot interquartile ranges have been used to determine and remove outliers. Missing values have been imputed by the mean for each individual truck and feature. Noise in the form of faulty measurements, trips less than 2 km, and rejected repair claims has been removed from the dataset.

From the clean data and final dataset, new attributes have been derived which were more representative for truck usage. From the trip dataset, scores per kilometer of driving have been derived from the trip data in order to compare driving behavior in a generalized way. The trip measurements have also been aggregated (their values have been summed over the measurement period of one month) and the mean, standard deviation and skewness of the measurements has been calculated in order to create features that describe the cumulative usage profile over the first month of operation. Lastly, different features have been extracted from the snapshot time-series data. The min, max, mean, standard deviation and skewness have been derived for the feature measurements during the first month of truck operation. Besides these global features, fuzzy histograms have been used to create feature vectors that approximate the probability distribution functions of the measured features for each individual truck.

This resulted in a set of 324 features for each truck. For clarity purpose, an example of a complete feature set is given in Appendix P

5 Repair prediction models

Before constructing the final models, several modeling decisions had to be made. The modeling techniques had to be chosen. the scoring metric had to be defined, features had to be selected and hyper-parameters had to be tuned. In this chapter the modeling setup is elaborated, modeling choices are justified and finally the final test design is explained.

5.1 Modeling setup

5.1.1 Scoring metrics

The problem under consideration is binary. Often, a confusion matrix is used to derive the model performance for such classification tasks (Kantardzic, 2011). The confusion matrix visualizes the model performance by classifying and showing the *true positive* (TP), *true negative* (TN), *false positive* (FP) and *false negative* (FN) predictions. The TP and TN are correctly classified predictions. The FP predictions are classified as positive while actually being negative and the FN predictions are classified as negative while being positive. The confusion matrix is visualized in Figure 20.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 20: Confusion matrix for a binary classification problem.

The confusion matrix is used to derive several performance metrics. Most common are the *Accuracy*, *Precision*, *Recall*, *Kappa* and the *F1 score*.

The accuracy is the most intuitive scoring metric which gives us the proportion of correctly classified samples:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

The precision metric tells us which proportion of samples classified as positive are actually positive. It is often used when there are large consequences for a positive classification. E.g. for the prediction of a patient having cancer or not, a high precision is desired:

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

The recall metric is used to determine how many of the actual positive samples have been identified by the model. E.g. when for a sample of 100 patients, 5 patients have cancer it is desirable that the model finds all of these five instances and thus has a high recall score:

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

The Kappa score compares the expected results from random prediction with predictions in the same proportion as the predictions made by the classifier being evaluated. I.e. it compares the observed

accuracy (model accuracy) against the expected accuracy (random chance). This is especially useful for (highly) imbalanced datasets. For example, for the mentioned case of a sample of 100 patients of which 5 have cancer an accuracy of 95% is reached when all patients are classified as healthy. The Kappa score however is low as this accuracy is (almost) reachable by chance as well. According to Anthony J Viera and Joanne M. Garrett, 2005 the scores and agreements are as follows:

- 0: chance agreement.
- 0.01 - 0.2: Slight agreement.
- 0.21 - 40: Fair agreement.
- 0.41 - 0.60: Moderate agreement.
- 0.61 - 0.80: Substantial agreement
- 0.81 - 0.99: Almost perfect agreement.

The kappa equation is given by:

$$Kappa = \frac{Accuracy - p}{1 - p} \quad (11)$$

Where p is the probability of predicting the correct class due to chance.

Lastly, the F1 score represents both the Precision and Recall. Although not as being the average of the two. For example, when the model has a Recall of 1 as it classifies all patients as having cancer, the precision is only 5% (for our 100 patients example). This is why the F1 score takes the harmonic mean of both the Precision and Recall:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (12)$$

For DAF, the most interesting scoring metric is that of accuracy. It is equally important to identify trucks that will have many repairs as identifying trucks that will have few repairs. This is because R&M contract prices are based on the expectations of the total repairs (and costs) of the individual trucks. Therefore, a wrong classification of the truck would directly result in incorrect R&M expectations of the truck under consideration. Besides the accuracy, the precision, recall, F1 score and Kappa score have to be taken into account as well in order to identify possible bias towards one of the labels in the dataset. Now, the dataset for 8 months ahead predictions contains an almost perfectly balanced class distribution and thus, the additional scoring metrics are not particularly relevant. However, the dataset for 11 months ahead predictions does contain a little imbalance (60%-40%, as elaborated in Section 4.5) and thus the additional scoring metrics can be used to evaluate possible bias towards the majority class.

5.1.2 Feature selection

When a scoring metric has been selected, in theory a machine learning model can be run with some basic parameter settings using all the features that are present in the dataset. However, this often results in suboptimal results as no features have been selected and the hyperparameters have not been tuned. We'll discuss the feature selection first.

Less relevant or highly correlated features often result in a decreased classification accuracy of machine learning models. Feature selection has proven to be an effective measure to increase the predictive accuracy, learning efficiency and reduce the complexity of the used models (Vieira et al., 2012). As stated by Kantardzic (2011). Feature selection is useful for a threefold of reasons :

- It often improves the performance of data-mining models. Especially when the number of features is large and/or many noisy features are present.
- The learning process of the models becomes faster and more memory efficient when the number of features is reduced.

- Understanding of the underlying process that leads up to the results becomes easier when the number of features is reduced.

In general, three types of feature selection can be distinguished (where hybrid methods are a combination of these methods), being feature ranking (also called *filtering methods*), subset selection (also called *wrapping methods*) and embedded feature selection respectively (Vieira et al., 2012). Feature ranking is used to rank the features based on some sort of metric and discards all features that do not meet the required threshold. Where metrics are (for example) the accuracy of available data, consistency, information content and statistical dependencies (e.g. mutual information between the feature and target or a chi-squared test for dependency between feature and target). Subset selection on the other hand, searches all available features (although often greedy and not exhaustive) for the best performing subset without ranking the features within the chosen subset. Lastly, embedded methods execute variable selection as a part of the training phase and are usually specific to the learning machine under consideration. In this study, both filtering and wrapping are applied. At first, a preliminary feature selection is made based on data quality and consistency. Afterwards, further filtering based on statistical dependencies and wrapping methods have been used to further reduce the dataset. The methods that are used in this study are described below.

5.1.3 Filtering methods

A few filtering methods have been applied to the final dataset in order to try and reduce model complexity and noise. Although many, relatively easy methods to do so are available, one has to be careful as seemingly redundant features don't truly have to be. They can still increase performance when taken with additional features (Guyon and Elisseeff, 2003). Initial filtering steps are only used to identify truly useless features.

Variance threshold The first filter that has been applied is that of the variance threshold. When a feature had only one specific value for all data instances (e.g. all trucks were of the brand DAF) the feature has been removed from the subset as they didn't possess any predictive value. In other words, a baseline variance of 100% had been set. Although one could argue that a more lenient variance threshold of, for example, 95% could further reduce noise in the dataset there has been decided to not lower the threshold level. It could be possible that some valuable information is stored in the 5% of the minority values. Besides that, the subsequent use of wrapping methods that has been performed in this study further reduces the subset of features and can also discard those 'redundant' features with a low variance (when present).

Mutual information Another common filtering method for feature selection are correlation analysis and mutual information (MI) calculation. Although correlation analysis is frequently used for basic statistic analyses it is only able to capture linear dependencies between features (Chandrashekar and Sahin, 2014). Mutual information however, is capable of measuring any kind of relation between variables, both linear and nonlinear (Chen et al., 2018). Therefore, we prefer the MI method over correlation analysis as an additional filtering method for feature selection.

MI calculates the 'amount of information' that can be obtained about a random variable, through another random variable. It does this by quantifying the amount of mutual information in units such as bits. Given two discrete variables X and Y, the mutual information between them is defined in terms of their probabilistic density functions $p(x)$, $p(y)$ and $p(x, y)$ as follows:

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)}\right) \quad (13)$$

The mutual information between the continuous features and labels in our dataset is estimated based on entropy estimates from k-nearest neighbor distances as described by Kraskov et al. (2004), which is more simply explained by Ross (2014). In short, from a continuous feature vector N , for every corresponding class label data-point i , a number I_i is computed. This is based on i 's nearest-neighbors having the same label y . First, for all data-points, the k_{th} closest neighbor with the same label is identified, where the set of samples in N having the same label is defined as N_{y_i} . The distance between the two points is then defined as d where the distance is derived from the values of feature vector N corresponding to the two

points. The number of neighbors, regardless of their class label, that fall within distance d (basically, a bin of size $2d$) is calculated, which is defined as m_i . I_i can then be calculated based on N and m_i by:

$$I_i = \psi(N) - \psi(N_{y_i}) + \psi(k) - \psi(m_i) \quad (14)$$

where $\psi(\cdot)$ is the digamma function (logarithmic derivative of the gamma function). The MI of the continuous feature X and discrete class label Y is then calculated by averaging I_i over all datapoints:

$$I(X, Y) = \langle I_i \rangle \quad (15)$$

For a more detailed explanation we refer the reader to (Kraskov et al., 2004)

Besides being able to capture non-linear relationships, another advantage of the MI filter method as proposed by Chen et al. (2018) is that the method can be used to determine the optimal number of features whereas other filtering methods often rely on a (heuristically), user defined threshold for inclusion. In short, the proposed method operates as follows:

1. For featureset X , calculate the MI between each available feature x and target variable Y and rank the features on their MI score.
2. Build/train the models with the first k features of the ranked list and calculate the performance of the model (e.g. accuracy or AUC), where k is run from 1 to X in increments of 1.
3. Plot the model performance versus the number of features used and identify the optimal number of features by identifying the number of features associated with the peak of the plot.

For a more comprehensive explanation of the MI feature selection method, we refer the reader to the work of Chen et al. (2018).

5.1.4 Wrapping methods

Two different wrapping methods have been compared in this study, being Recursive Feature Elimination (RFE) and Sequential Feature Selection (SFS). Although exhaustive methods exist as well, they become computationally expensive very fast for large datasets and a large feature space (Chandrashekar and Sahin, 2014). The RFE and SFS are greedy wrapping methods which provide adequate results and have an acceptable computation time.

Sequential Feature Selection The SFS comes in two 'flavors', being forward and backward selection. The forward selection algorithm starts with an empty set of features and as a first step adds the one feature that gives the highest objective function score (e.g. classification accuracy or precision). This feature is then permanently included in the feature subset. For every next step, the model adds another feature to the subset by iterating over the remaining features and selecting the one that provides the highest objective function together with the previously included features (current feature subset). This process is repeated until the required number of features is selected (i.e. when the objective function score no longer increases) (Chandrashekar and Sahin, 2014). On the other hand, the backward selection algorithm follows a similar procedure but reversed. It starts with the full set of features and iteratively removes the feature which exclusion results in the largest increase of the (user defined) performance metric. As another backward feature selection method is already presented later on, there has been decided to use the forward selection method during the modeling phase. The forward SFS algorithm is given in Figure 21, where the accuracy (acc) is used as the scoring metric.

Recursive Feature Elimination The RFE algorithm is similar to the backwards SFS apart from the performance metric used for evaluation (Figure 22). The RFE algorithm removes features based on the feature importance rather than a user defined performance metric. It uses either the feature coefficients (linear models) or feature importance (tree based models) to rank features and iteratively removes the worst ranked feature from the subset of features until one's left with the desired number of features, making it an instance of backward feature elimination (Guyon et al., 2002).

```

Input:  Data with feature set F and class label C;
Output: S; //selected feature subset
1  acc = 0;
2  S = null;
3  while  $\sim$ isempty(F) do
4      flag = 0;
5      for i = 1 to length(F) do
6          Snew = add(copy(S); Fi);
7          accnew = evaluate(classifier, DataSnew ∪ {C});
8          if accnew > acc then
9              ind = i; acc = accnew; flag = 1;
10         if flag then
11             S = add(S; Find); //add Find to S
12             F = del(F; Find); //remove Find from F
13         else
14             break; //stop feature selection
15     return S;

```

Figure 21: The SFS algorithm.

```

1.1 Tune/train the model on the training set using all predictors
1.2 Calculate model performance
1.3 Calculate variable importance or rankings
1.4 for Each subset size Si, i = 1 . . . S do
1.5     Keep the Si most important variables
1.6     [Optional] Pre-process the data
1.7     Tune/train the model on the training set using Si predictors
1.8     Calculate model performance
1.9 end
1.10 Calculate the performance profile over the Si
1.11 Determine the appropriate number of predictors
1.12 Use the model corresponding to the optimal Si

```

Figure 22: The RFE algorithm.

5.1.5 K-Fold cross validation and parameter optimization

When constructing models, simply splitting the data into a training-set and test-set, building a model with the training-set and evaluating its performance with the test-set once (holdout method) would be rather pessimistic. Although being an easy to understand method, it can result in a biased performance estimate and over-fitting due to the fact that the training-set might be unrepresentative of the test-set or new instances in general (Chandrashekar and Sahin, 2014). To mitigate these effects and provide a more robust model validation technique, cross-validation is widely used (Efron and Fron, 2012).

Instead of a single split (holdout), the training-set is split into *k* mutually exclusive subsets of (roughly) the same size. Essentially this is a repeated holdout method that is repeated for *k* times where in each iteration, 1 of the *k* subsets is used as the validation set and the remaining *k*-1 sets are used for training. This mutually exclusive splitting of the dataset ensures that each data point is only used for validation once (and *k*-1 times for training). This technique allows for efficient use of all the available data and is therefore especially valuable when working with small datasets. Although *k* can be chosen arbitrarily, extensive research has shown that 10 splits are recommended as this minimizes variance and bias (Han et al., 2011)(Witten, I. H. , Frank, E., 2016). Now, using 10-fold cross validation, the created model is trained on 10 different subsets and scored on the remaining validation set, where the performance of the model is simply calculated by:

$$CVP = \frac{1}{k} * \sum_{i=1}^k P_i \quad (16)$$

Where k is the number of folds, CVP is the cross validated performance metric score and P is the performance metric score of the model. Figure 23 provides a visualization of the cross-validation process.

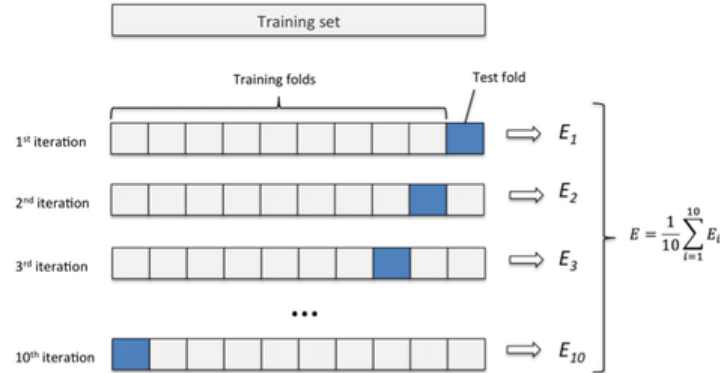


Figure 23: 10-fold cross validation where E is the performance metric score.

Besides evaluating different machine learning techniques for the problem at hand, the k -fold cross validation technique is also used to determine the optimal set of hyperparameters of models (when applicable). Using the same logic as described above, cross validation can be used to evaluate the performance of models with different hyperparameter settings over the same training data. Subsequently, the model with the best average performance is selected as the final model. More details on hyperparameters and their tuning are given per model in Chapter 6.

Note that the final selected model, because it has already been validated, can be retrained over all folds of the training data (thus including the validation fold). The final (trained) model can then be used on the test set which has been held out during the entire modeling process to evaluate the performance on new, unseen data.

5.1.6 Experimental setup

As a summary of the chapter, the experimental setup is given. This experimental setup is used for two different model setups which are given below:

- Base model - Model without fuzzy histograms or feature selection applied.
- Base model with SFS - Model without fuzzy histograms but with feature selection applied.
- Extended model - Model with fuzzy histograms but no feature selection applied.
- Extended model with SFS - Model with fuzzy histograms and feature selection applied.

As explained, for both the datasets (8 and 11 months) the dataset is divided into three subsets, being the training, validation and test set respectively. The holdout method is used to split the data into a training set and test set where 70% is used for training and 30% is used for testing. The training set is subsequently split into validation and training sets using 10-fold-cross-validation. The number of instances (rows) per resulting subset is given for both datasets in Table 17 (note that, as mentioned, each row contains 324 features). Furthermore, the categorical features in each of the datasets are one-hot encoded before being used for training and analysis as this provides the models with a numerical representation of the categorical data that can be used as input directly.

Subsequently the models are built using the training sets. The model hyper-parameters are tuned using the (cross validated) accuracy on the validation sets. Lastly, the best scoring model is subsequently retrained using both the training and validation data before being used to perform predictions on the

Table 17: The number of rows (instances) per subset for both datasets

Dataset 8 months subset	Nr. of rows	Dataset 11 months subset	Nr. of rows
Training set(s)	499	Training set(s)	216
Validation set(s)	56	Validation set(s)	24
Test set	238	Test set	102

test sets, whose results are used for evaluation in the next chapter. As mentioned, the specific parameters on which the models are tuned are explained for each model in Chapter 6. The general test design setup is given in Figure 24.

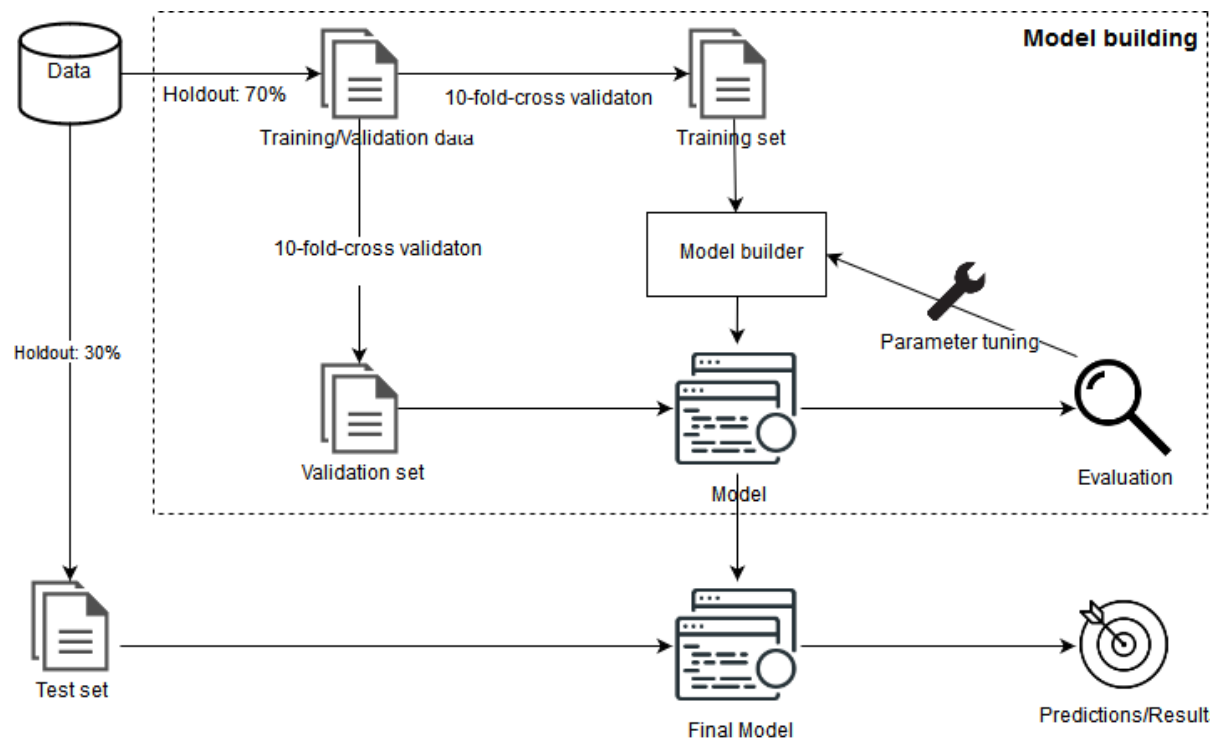


Figure 24: The general experimental setup for the different prediction models.

6 Modeling results

In this chapter the used modeling techniques are evaluated. Their performance is evaluated using the test set that has been untouched during training and tuning of the models. The results are listed and relevant variables are given and discussed. In this chapter, the holdout between the test-set and training/validation set is select a-priory in order to compare the results between the different models on the same dataset. The best scoring models using 10-fold-cross validation have been reported in this section. However, one has to take into account that the models are not stable per definition. Different runs and data splits resulted in slightly different performances. The average results using different splits are given in Appendix L

6.1 Feature selection

As explained in the previous chapter, three different feature selection methods have been proposed and tested on the final datasets in an effort to increase the modeling performance. The methods were tested using both decision trees and random forests. All 324 features including the features that have been derived using Fuzzy histograms (Appendix O) were used as input for the feature selection methods. The methods were specified to optimize the performance metric of *accuracy* as this has was the most relevant scoring metric for DAF (see Section 5.1.1). During testing it became apparent that the Sequential Forward Selection had superior results compared to RFE and MI. Where SFS showed a clear optimal number of features for the problem at hand, RFE and MI had poor results as the optimal number of features selected was unstable and no clear optimum could be identified when visualizing the methods. Furthermore, their performance (based on the validation accuracy scores) was significantly lower than the scores reached with SFS. A possible explanation for this phenomenon could be that MI and RFE base their feature selection on a feature importance metric while SFS bases its selection purely on the increase in performance of the model. Now, as the prediction power of the models is limited, the relations within the data are not as strong either. It could be possible that the metric as used by MI and RFE feature elimination do not adequately capture these relations within the data. To provide a better understanding of the above described limitations , an example for the three methods is visualized in Figure 25, where their performances for one of the used random forest models are compared. It shows the 10-fold-cross-validated validation scores against the number of features selected.

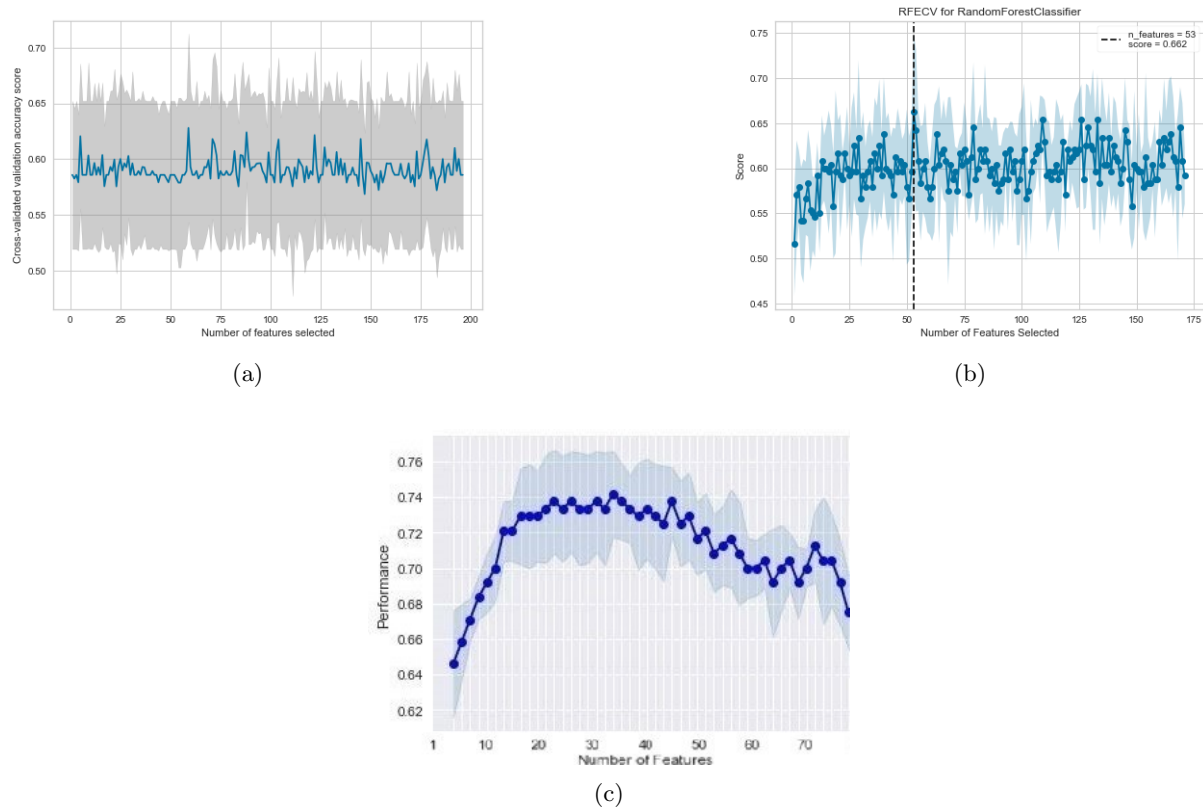


Figure 25: The performance of MI (a), RFE(b) and SFS (c) on one of the used random forest models.

As shown, The SFS technique outperforms the other methods on its performance score (8% higher accuracy score) and shows a much clearer optimal number of features due to the obvious trend that is visible in the figure. Based on the lack of performance of the other two methods, there has been decided to focus on SFS as the feature selection method during the final modeling and evaluation. Now, as each of the evaluated machine learning modeling techniques operates differently, applying SFS results in a different subset of features for each individual modeling technique. Thus, a single definitive subset of features cannot be given here. For readability purposes, the relevant, selected features are given for each modeling technique individually in the subsections of this chapter.

In the remainder of this chapter, the results for each of the modeling techniques are elaborated.

6.2 Logistic Regression

First, the results for Logistic Regression are presented. For the logistic regression model there are two parameters that can be chosen, where the first parameter is the regularization method. There is Lasso Regression (L1) and Ridge Regression(L2), where the difference between the two is found in the penalty term. L1 minimizes the sum of absolute differences between the targets and prediction, while L2 minimizes the sum of the square of differences between the two. Thus, one can image that the L2 method is much more sensitive to outliers. Furthermore, L1 is better capable in handling many (irrelevant) features (Andrew, 2004). A downside of the L1 method is that results can be unstable (i.e. feature correlation scores are not always consistent), while this is not the case for L2 regression. The second parameter to be chosen is the regularization strength C, where a larger C corresponds to a larger penalization of 'large' weight coefficients. The goal is to prevent the model from over-fitting by picking up 'noise' and 'specific cases'.

Furthermore, before being used as input for the logistic regression models, the feature values of the input are normalized. Thus, the values for each feature are scaled between the range 0 and 1 (Han et al., 2011). The reason to do this is twofold. On the one hand, this is done because the regularization methods are

'not equivariant under scaling of the inputs, and so one normally standardizes the input before solving' Hastie et al. (2001). I.e. they are sensitive to the scale of the features, as they penalize large weights of coefficients. On the other hand it allows for direct comparison of the beta coefficients to derive feature importance as they are formed over the same range (scale) of data for each feature.

The parameter optimization as described above has been done by a grid-search over the range of available parameters. This method searches for the optimal set of parameters for the model under consideration by an exhaustive search of the given parameter space (or grid). A (ten fold cross validated) model is constructed for each possible set of hyper-parameters after which the best performing set is given as output and used for final modeling. The reader is referred to Appendix Q for the range of parameters that has been searched. For the 8 months ahead models, a regularization strength C of 100 and Regularization method L1 was found to be the optimal combinations of features, while for the 11 month ahead datasets a regularization strength C of 1 and regularization method L2 was found to be optimal.

6.2.1 Base models

The results for the base model are given in Table 18.

Table 18: Logistic Regression base model results

Setup	Accuracy	Precision	Recall	F1	Kappa
Logistic Regression 8 months ahead	0.55	0.61	0.50	0.55	0.14
Logistic Regression 11 months ahead	0.57	0.45	0.46	0.46	0.12

As one can see, the base model provides only limited results. Besides only making a correct prediction for 55% (8 months ahead) and 57% (11 months ahead) of the samples, only a limited number of trucks having many repairs are correctly identified (recall) and of each truck that has been classified as having many repairs, only 50% or less actually were of that class (precision). Note that we are mostly interested in the accuracy as we simply need the model to be make as much correct classifications as possible. For contracting purposes both the trucks with many repairs as the trucks with few repairs are equally interesting. From a business side of view, classifying a truck with many repairs as a truck with few repairs has equal consequences as classifying trucks with few repairs as trucks with many repairs.

6.2.2 Extended models

Next, SFS has been applied in an effort to remove redundant variables and improve model accuracy. The SFS method was run using 10 fold cross validation and the results for the 8 months ahead model are given in Figure 26. The cross-validated accuracy scores on the validation sets are plotted against the number of features used in the model. The solid line shows the averaged validation accuracy score while the standard deviation is given by the shaded area around the average accuracy scores. The performance of the model increases up until 12 features, after which the performance of the model decreases again (due to noisy and irrelevant features). The 12 features are subsequently used to train the final model after which its performance is evaluated over the test set, which is given in Table 19. The same procedure has been applied for the other Logistic Regression models.

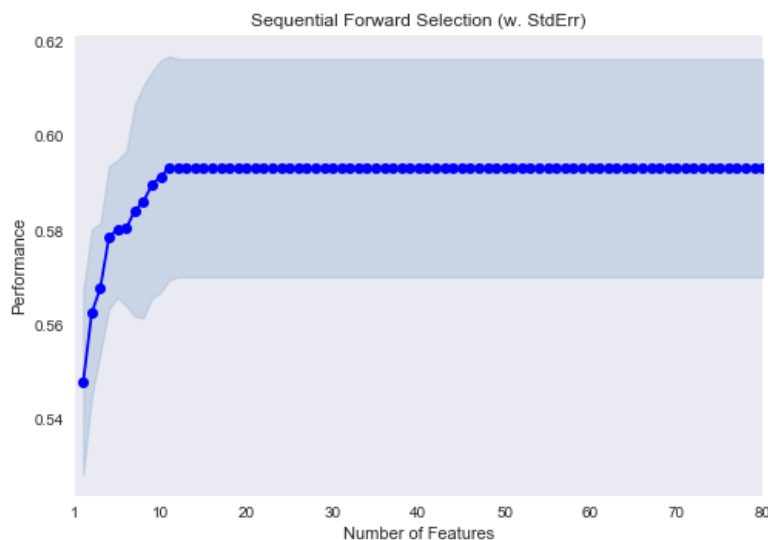


Figure 26: Cross-validated 8 months ahead Logistic Regression accuracy scores plotted against the number of features selected by the SFS.

Table 19: Modeling results for the base Logistic Regression model with SFS.

Setup	Accuracy	Precision	Recall	F1	Kappa
Logistic Regression SFS 8 months ahead	0.56	0.57	0.54	0.56	0.10
Logistic Regression SFS 11 months ahead	0.66	0.66	0.45	0.51	0.30

The best results are obtained using the 11 months ahead model and the features as found by SFS. Therefore, the feature importance in terms of beta scores are given in Figure 27. They provide insights in the features that are relevant for the prediction task. The features can then be compared to the relevant features in the other models to validate their importance in general. Now, the performance of the 11 months ahead model is significantly better than that of the 8 months ahead model. This is expected as the repairs become more predictable over time (i.e. the repairs that will occur on a truck during its lifetime are more predictable than the repairs during its early life only). The average number of repairs per truck jumped from 1 to 2 in the three months of time difference between the two datasets. Thus, trucks from the 8 months ahead dataset only need a single incidental repair to be classified as having 'many' repairs, while trucks from the 11 months ahead dataset need at least two. This likely allows for the model to find more robust relations between the number of repairs and truck features as a single incidental repair no longer classifies a truck as having many repairs. This holds for the results of the remaining other models as well.

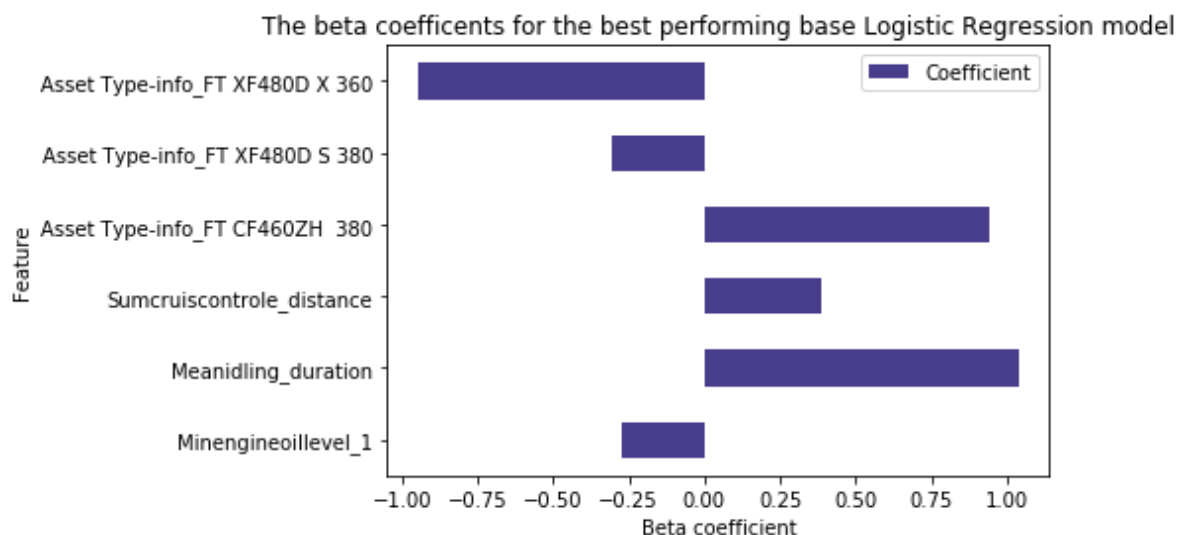


Figure 27: The beta coefficients for the best performing base Logistic Regression model.

Looking at the coefficients from the model, there can be seen that the model uses a combination of Connect data and product specification information for the classification task. It shows that a CF truck with a specific setup is expected to have more repairs than an XF truck with specific setups. Simultaneously, the total cruise control distance and mean idling duration have a positive effect on the number of repairs. The positive effect of the total cruise control is logically explained by the fact that a higher value for this feature indicates a more intensive and/or frequent use of the truck. The cause of the positive effect of the mean idling duration is a bit more speculative. It is most likely that the trucks with a high relative idling duration make frequent stops or are used for specific heavy tasks. Lastly, the negative effect of the minimum oil level in the engine is a feature that is directly related to the truck's condition. Thus, a lower oil level results in a higher chance of a truck having many repairs (as a higher minimum oil level leads to a negative value for the repair predictions).

Next, the extended model is evaluated, which follows the exact same procedure as the base model with the difference that fuzzy histograms have been used to derive additional variables from the Connect data (as explained in section 4.4.2). The models are run including the new variables and their results are given in Table 20 and Table 21.

Table 20: Modeling results for the extended Logistic Regression model without SFS.

Setup	Accuracy	Precision	Recall	F1	Kappa
Extended Logistic Regression 8 months ahead	0.56	0.58	0.53	0.55	0.11
Extended Logistic Regression 11 months ahead	0.58	0.45	0.49	0.47	0.13

Table 21: Modeling results for the extended Logistic Regression model with SFS.

Setup	Accuracy	Precision	Recall	F1	Kappa
Extended Logistic Regression SFS 8 months ahead	0.57	0.57	0.52	0.54	0.10
Extended Logistic Regression SFS 11 months ahead	0.69	0.61	0.48	0.54	0.31

The Extended models consistently outperform the base models. Especially the extended 11 months ahead model with SFS proves to outperform the model without SFS by 3 percent-point (see Appendix M for validation). To gain insights in the working of this model, the SFS results and the feature importances of the new best performing model are given in Figure 28.

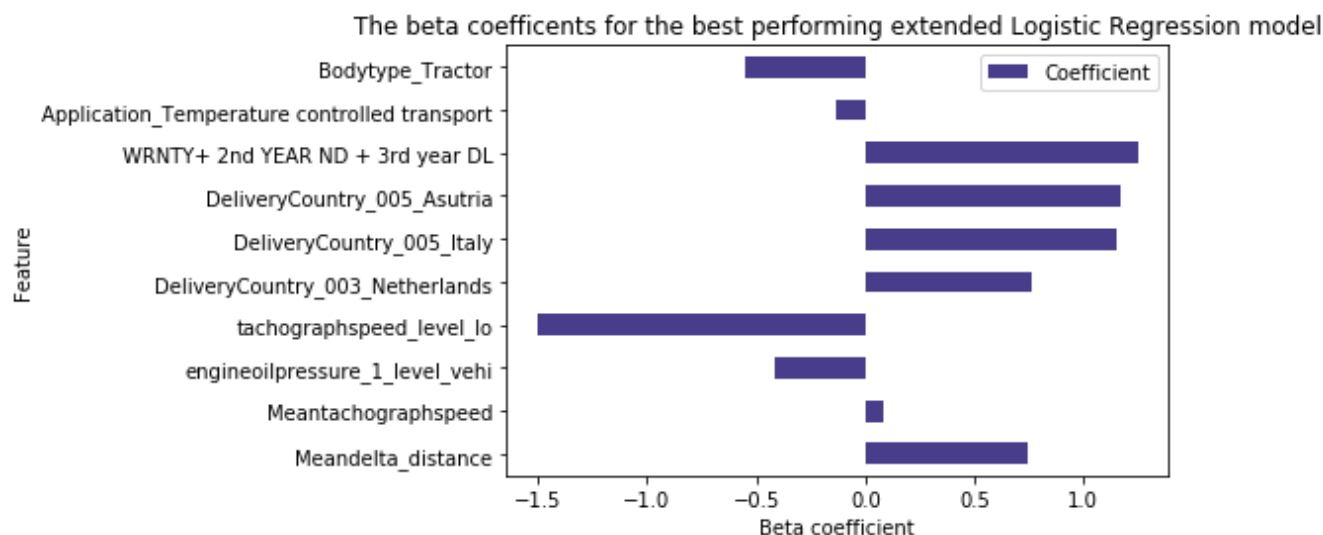


Figure 28: The beta coefficients for the best performing extended Logistic Regression model.

As can be seen, the extended model chooses quite some different variables for its predictions compared to the base model. Thus, the fuzzy histograms added some complexity to the model but improved it nonetheless. Most of the variables in Figure 28 are straightforward, with the exception of *WRNTY + 2nd YEAR ND + 3rd year DL*. One could derive that trucks that have this type of warranty contract can be expected to have a higher risk of many repairs during their first year of operation (as seen by the positive beta coefficient for that feature). However, trucks with this type of warranty package do not enjoy further benefits compared to the trucks with standard warranty packages. Therefore, the direct cause of the increased risk for these trucks cannot be directly derived. The *meandelta_distance* feature indicates the distance driven between the measured snapshot intervals and thus provides information about the driving behavior of trucks. A higher average distance between the measured snapshots relates to high driving speeds (and probably long distance trips).

6.3 Decision Tree

The decision tree is the next model under consideration. The maximum depth of the trees, minimum number of samples per split and minimum samples per leaf are tuned using cross validation in order to avoid over-fitting (and underfitting) of the models. A higher number of samples per split and per leaf results in a more pruned tree. Thus, sections in the trees that provide little classification powers are removed (sections that are only capable of classifying very specific instances). In general pruned trees have better generalization capabilities than large trees (Kantardzic, 2011). For the analysis, the same procedures as for the Logistic Regression models have been followed. For the searched parameter space during parameter optimization, the reader is referred to Appendix Q. Also, the optimal parameters that have been found for each of the models are given in Appendix R.

6.3.1 Base models

First the base model is evaluated, for which the results are given in Table 22

Table 22: The results for the base models without SFS.

Setup	Accuracy	Precision	Recall	F1	Kappa
Decision Tree 8 months ahead	0.54	0.59	0.41	0.48	0.08
Decision Tree 11 months ahead	0.64	0.50	0.56	0.53	0.21

Next, SFS is used as the feature selection method to see if results improved. It's cross-validated validation scores are plotted against the number of features selected and shown in Figure 29. It shows that the performance increases fast when a few features are added, after which it stagnates. This shows that just a few features are needed to reach peak performance. In fact, after only 14 features the model showed no further (significant) performance improvements.

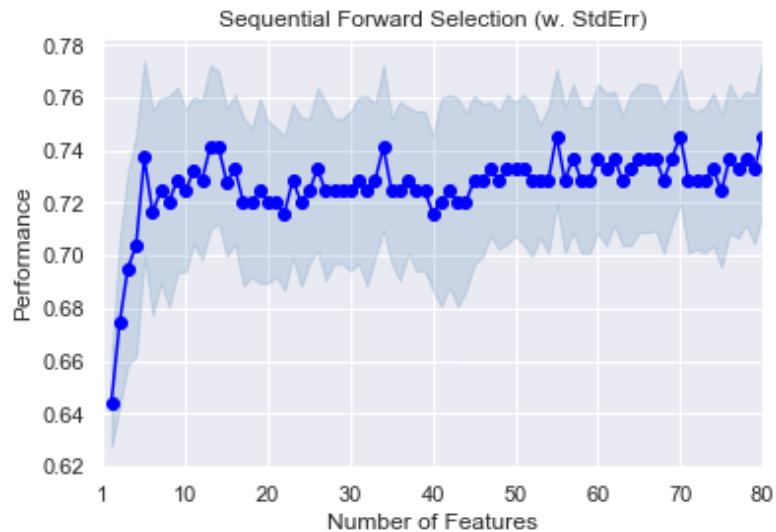


Figure 29: Results of the Decision Tree model performance using SFS and 11 months of data.

The results for the base model using SFS are given in Table 23.

Table 23: The results for the base models with SFS.

Setup	Accuracy	Precision	Recall	F1	Kappa
Decision Tree 8 months ahead	0.55	0.56	0.56	0.57	0.11
Decision Tree 11 months ahead	0.64	0.54	0.44	0.47	0.19

As one can see, this did not significantly improve the performance of the models. The highest accuracy remains to be 64%. Using the feature importance from the corresponding decision tree (11 months ahead with SFS), insights about relevant features can be gained. They are given in Figure 30.

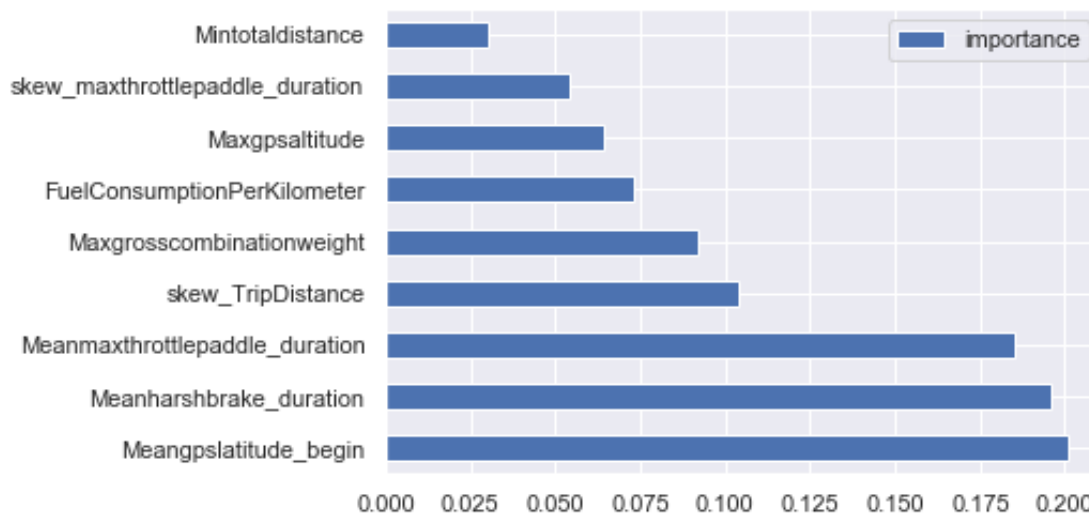


Figure 30: Feature importance for the best 11 months ahead decision tree with SFS.

From these features it can be derived that the Connect database is important when predicting the repair category of trucks. In fact, not a single feature from the CCM database is used in the decision tree. This is striking as the logistic regression model did include CCM database features. However, this can be explained by the different method with which the decision tree decides on the importance of its features compared to the Logistic Regression model. As elaborated in Section 2.3.2, the feature importance is derived based on the decrease in node impurity, weighted by the probability that a sample reaches the node. Now, analyzing the important CCM features from the logistic regression models revealed that chosen features often only had a limited number of appearances in the dataset. For example, the feature *delivery country Italy* was only present in 9 instances of the dataset, for which 8 trucks were labeled as having many repairs. Thus its relevant beta coefficient is explained for the logistic regression model. Its discriminating effect in the overall decision tree model, however, is limited (due to its few number of occurrences in the dataset) and thus the decision tree doesn't classify this feature as important. The same logic applies for the other CCM database features that do not have a presence in the decision tree models. Furthermore, the decision tree also considers non-linear relations between (combinations of) features, which the logistic regression model does not. For example, a high acceleration and braking duration per trip combined with a high weight of the truck (combination weight) likely influences the number of repairs more significantly than a high acceleration duration by itself.

6.3.2 Extended models

Using the same procedure, the extended models are analyzed next. Again, for the best performing model, the most important features are given below. The results for the extended models without SFS are given in Table 24

Table 24: The results for the extended models without SFS.

Setup	Accuracy	Precision	Recall	F1	Kappa
Extended Decision Tree 8 months ahead	0.58	0.58	0.72	0.64	0.15
Extended Decision Tree 11 months ahead	0.66	0.54	0.44	0.47	0.19

Subsequently the results for the extended models with SFS are given in Table 25.

It is interesting to see that the extended models (including the features derived with fuzzy histograms) perform better in terms of accuracy. For the 11 months ahead predictions with and without SFS, the

Table 25: The results for the extended models with SFS.

Setup	Accuracy	Precision	Recall	F1	Kappa
Extended Decision Tree SFS 8 months ahead	0.60	0.59	0.68	0.63	0.19
Extended Decision Tree SFS 11 months ahead	0.66	0.55	0.61	0.58	0.19

extended models show an consistent increase of about 2 percentage-points compared to the base models (see Appendix M for validation). To provide insights about the added features, the feature importance for the best performing extended decision tree is given in Figure 31.

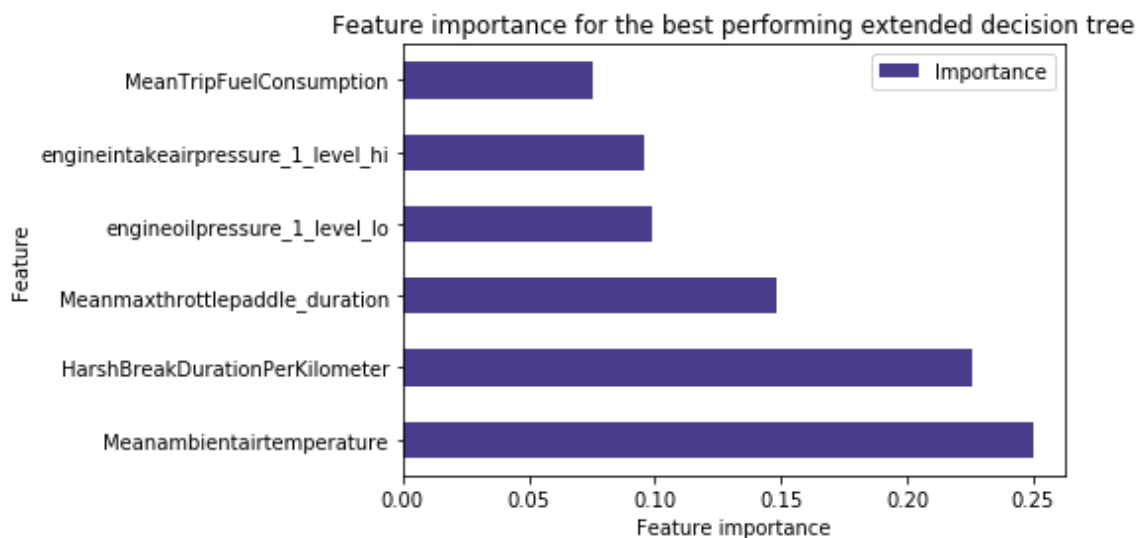


Figure 31: Feature importance for the best performing extended decision tree.

As one can see, two features that have been derived with the fuzzy histograms have been incorporated in the best performing model. These are *engineintakeairpressure_1_level_hi* and *engineoilpressure_1_level_lo* respectively. Again, the features that have been used for the predictions do not contain any truck specification features. Only Connect variables have been selected, where especially specific pressure levels inside the trucks appear to contain additional information for the classification task. The *meanambientairtemperature* could be of influence for two reasons. For one, a higher temperature could increase or decrease the strain on the engine. However, it is more likely that it is an indication for the country/area in which the truck operates. Realizing that the datasets contain trucks delivered in Europe only, the average measured temperature is region/country dependent. The throttle and breaking durations are direct indicators of a truck's usage, which confirms that at least some potential can be found in the Connect data at DAF regarding repair predictions. Especially as they are chosen in favor of the truck specification features.

6.4 Random Forests

As explained, a random forest is basically an ensemble of trees. As such, the same parameters as for decision trees can be tuned with the addition of the *number of estimators*. This parameter represents the number of decision trees that is used in the forest. In principle, adding more trees allows the model to obtain a better generalization performance, at the cost of increased computational complexity. Furthermore, as random forests provide generalization by majority voting, pruning of the trees is less important compared to single decision trees. Again, for the searched parameter space during parameter optimization, the reader is referred to Appendix Q. Also, the optimal parameters that have been found for each of the models are given in Appendix R.

6.4.1 Base models

First, the results for the base models are given in Table 26.

Table 26: The results for the base models without SFS.

Setup	Accuracy	Precision	Recall	F1	Kappa
Random Forest 8 months ahead	0.59	0.63	0.54	0.58	0.19
Random Forest 11 months ahead	0.64	0.67	0.64	0.66	0.30

Next, SFS is applied. The cross validated validation scores are plotted against the number of features selected in Figure 32. The performance (accuracy) follows a nice trend where the model performance increases as more features are added until it reaches a peak at 22 features. Adding more features both increases complexity and reduces the modeling performance, which is shown by the (mostly) decreasing performance when more than 22 features are added.

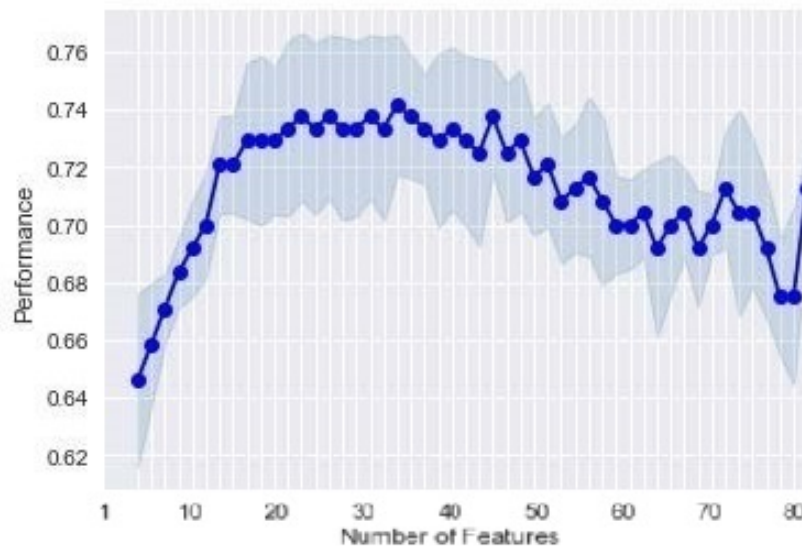


Figure 32: Results of the Random Forest validation performance using SFS and 11 months of data.

The results for the base model using SFS are given in Table 27.

Table 27: The results for the base models with SFS.

Setup	Accuracy	Precision	Recall	F1	Kappa
Random Forest SFS 8 months ahead	0.63	0.64	0.69	0.66	0.26
Random Forest SFS 11 months ahead	0.64	0.52	0.41	0.46	0.18

The random forest that predicts 8 months ahead benefits from SFS while the method does not increase the accuracy for the 11 months ahead predictions. The best performing model is the Random Forest 11 months ahead without SFS. Although it does have the same accuracy (64%) as the model with SFS, it has a higher Kappa score (0.30, which corresponds to a fair agreement), recall and precision. To provide insights in the decision making process, the top 10 most important features as identified by the model are given in Figure 33:

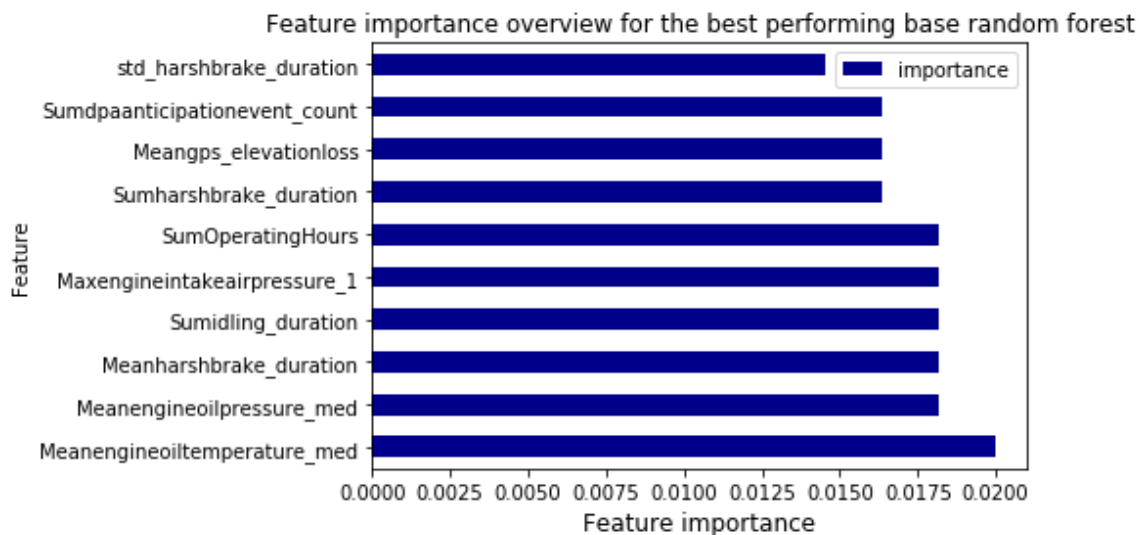


Figure 33: Feature importance for the best performing base random forest.

As can be seen, individual features do not provide much information about the classification of the trucks in this model. Even the top 10 of features only have an importance of 0.02 (normalized) or less. Nonetheless, they are the most important features for this model and can be used to verify the selected features against the other modeling methods used.

6.4.2 Extended models

Next, the extended models are analyzed to check if they are able to increase the performance of the random forests. Again, both including and excluding SFS. First, the results without SFS are given in Table 28.

Table 28: The results for the extended models without SFS.

Setup	Accuracy	Precision	Recall	F1	Kappa
Extended Random Forest 8 months ahead	0.61	0.63	0.63	0.63	0.23
Extended Random Forest 11 months ahead	0.64	0.66	0.64	0.65	0.29

Lastly, the performance for the extended models with SFS are reported in Table 29.

Table 29: The results for the extended models with SFS.

Setup	Accuracy	Precision	Recall	F1	Kappa
Extended Random Forest SFS 8 months ahead	0.62	0.67	0.52	0.59	0.24
Extended Random Forest SFS 11 months ahead	0.66	0.58	0.36	0.43	0.23

The best performing model when looking solely at accuracy is the Extended Random Forest SFS 11 months ahead. However it's recall is exceptionally poor. Thus, although providing a high accuracy, it is only capable of identifying a limited number of trucks that have many repairs. In contrast, based on the relatively high accuracy, it is therefore exceptionally good in recognizing trucks that will not have many repairs. The corresponding feature importances are given in Figure 34.

The model uses a combination of truck specifications and usage characteristics to derive the class of the truck. However, the importance of the usage characteristics and thus there can be stated that

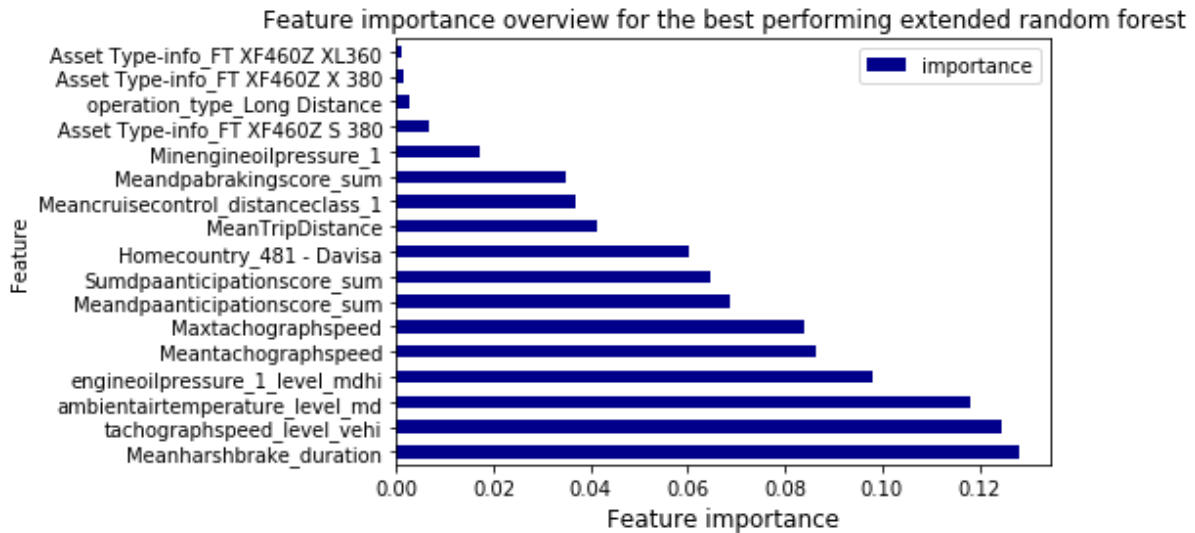


Figure 34: Feature importance for the best performing extended random forest.

they provide superior information compared to the truck specifications and contract information (CCM variables).

6.5 MLP-NN results

The last model under consideration is the MLP-NN. Although it will be less informative for DAF compared to the other methods, it has been implemented nonetheless, in order to see if it has significantly better performance. For this model, the number of hidden layers, the activation function, learning rate and maximum number of iterations have been tuned. The number of layers determines, as the name suggests, the number of hidden layers in the model. Too few layers could result in a high training and generalization error (under-fitting), while too many hidden layers result in a slow learner with poor generalization. The activation function for the hidden layers is tuned by simply trying different types of functions (logistic sigmoid, and hyperbolic tangens). Furthermore, the learning rate can be adapted in order to change the step-size in weight updates. Too high of a learning rate can cause the model to 'skip' optimal settings, while too slow of a learning rate could result in the model taking too long to reach the optimum. Lastly, the number of iterations is tuned as too many iterations could result in an over-fitted model while too few iterations cause the opposite (LeCun et al., 2012). As for the other models, the reader is referred to Appendix Q for an overview of the searched parameter space during parameter optimization. Also, the optimal parameters that have been found for each of the models are given in Appendix R.

6.5.1 Base models

Conform the other models, the performance of the base models is reported first. They are given in Table 30

Table 30: The results for the base models without SFS.

Setup	Accuracy	Precision	Recall	F1	Kappa
NN 8 months ahead	0.62	0.58	0.70	0.64	0.15
NN 11 months ahead	0.61	0.49	0.49	0.49	0.22

The accuracy and recall on the *8 months ahead base* dataset are slightly higher compared to other models.

However, as a trade-off, information about the feature importances cannot be derived in a straightforward manner due to the black box working of the NN. Furthermore, its performance on the other datasets is not significantly higher (or even worse) than that of the random forest models. They have a low precision and recall score while not achieving a significant increase in accuracy. Thus, feature importance derivations (e.g. sensitivity analysis) are not further explored for the NN's. It can be concluded that the results of the NN's are not useful for DAF with the current amount of available data.

Next, SFS has been applied for the MLP-Neural network. However, the method seemed less suitable for the MLP-NN than for the other methods. In fact, the accuracy of the models was hardly improved while the precision and recall decreased. An example of the SFS results for the 11 months ahead model is given in Figure 35.

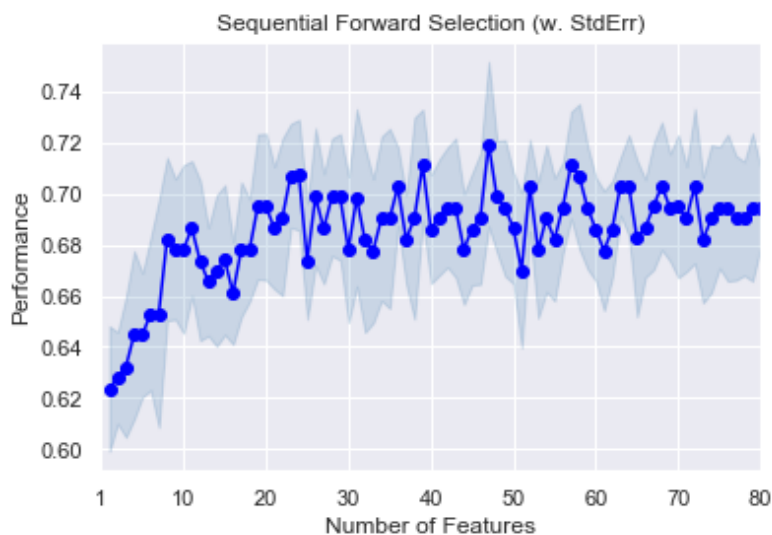


Figure 35: SFS applied to one of the MLP-NN's.

As one can see, there is no clear optimum for the number of features to select, which explains the poor results when applying it to the MLP-NN's. Nevertheless, the results with SFS are given in Table 31

Table 31: The results for the base models with SFS.

Setup	Accuracy	Precision	Recall	F1	Kappa
Base NN SFS 8 months ahead	0.61	0.61	0.73	0.66	0.17
Base NN SFS 11 months ahead	0.63	0.51	0.41	0.46	0.18

6.5.2 Extended models

Next, the results when including the additional features from the fuzzy histograms are analyzed. The results for these extended models are given in Table 32.

Table 32: The results for the extended models without SFS.

Setup	Accuracy	Precision	Recall	F1	Kappa
Extended MLP-NN 8 months ahead	0.61	0.60	0.73	0.66	0.20
Extended MLP-NN 11 months ahead	0.65	0.55	0.44	0.49	0.22

Subsequently, the results using SFS on the extended models are given in Table 33.

Table 33: The results for the extended models with SFS.

Setup	Accuracy	Precision	Recall	F1	Kappa
Extended MLP-NN SFS 8 months ahead	0.61	0.45	0.46	0.46	0.11
Extended MLP-NN SFS 11 months ahead	0.66	0.56	0.49	0.52	0.26

Again, SFS did not provide any significant improvements for the NN-models. The extended models without SFS however did have some improved results with respect to the base models. Although accuracy did not always increase, the recall, precision and Kappa score improved slightly.

In general, the MLP-NN did not show an improved accuracy or kappa score compared to the other models. In addition, some relevant information is lost in the models because the feature importance cannot be analyzed. As DAF wanted to receive insights about the factors that influence the numbers of repairs (such that they could act on it in the future) and the performance of the models is not significantly better, there can be concluded that the MLP-NN is not suitable for the problem at hand.

6.6 Summary of results and relevant features

6.6.1 The most suitable models to predict repairs

In this section, the best modeling results and most suitable prediction methods are stated. Furthermore, the relevant features for the prediction of repairs are given.

Four different types of machine learning techniques have been tested on their ability to predict if trucks will have few or many repairs during their early life. They are Logistic Regression, Decision Trees, Random Forests and MLP-Neural Networks respectively. In general, there is concluded that there is some predictive power present in the Connect data of DAF, but the few numbers of repairs in the early life of the trucks limited the prediction performances. The highest achieved accuracy is 63% for the 8 months ahead prediction, which has been reached using random forests in combination with SFS. In addition, the random forest had a decent precision and recall score of 64% and 69% respectively.

Furthermore, an accuracy score of 69% has been reached for the 11 months ahead prediction, which is achieved by the Logistic Regression model. However, the recall score (ability to recognize trucks having many repairs) of this model was limited, being 48%. This indicates that some bias toward the 'few repairs' class exists in the model. Therefore, the random forest technique is preferred. Although having a lower accuracy (64%), it has a significantly higher recall score 64% (and a precision score of 66%, which is 5% higher than that of the Logistic Regression model). Lastly, the models had comparable kappa scores as the Logistic Regression model scored 0.31 and the Random Forest model scored 0.30, showing a fair agreement level.

6.6.2 Relevant features for repair predictions

To assess the relevant features in the prediction models, the most important features have systematically been visualized in this chapter, where the feature importance based on gini impurity has been used for the random forests and decision trees, and beta coefficients have been used for the Logistic Regression models. Analysis of the features revealed that each modeling technique selects a different subset of most important features. Therefore, an unambiguous set of relevant features cannot be given. Instead, the most recurring features among each of the models is given as their presence across multiple models provides a nice validation of their importance. The top 10 of most used features across the best performing models is given in Table 34.

The asset type is found most often across the models. Depending on the specific type, this can either cause the predictor to expect less or more repairs. For example, the asset type *FT_XF480D X 360* has

Table 34: The 10 most used features across the best performing models

Feature	Presence in nr. of models	Total types present across models
Asset Type	2	6
Harsh brake duration	4	6
Tachograph speed	2	5
Engine oil pressure	4	5
Trip distance	3	4
DPA Anticipation scores	2	3
Max throttle duration	2	3
Ambient air temperature	2	2
Engine intake air-pressure	2	2

a negative beta coefficient in the Logistic Regression model while the asset type *FT_Cf460ZH 380* is associated with a positive beta coefficient. Implying that certain truck setups are expected to have more repairs than others, which could be explained by the fact that certain setups are more often used for special purposes than others. Harsh brake duration metrics such as *harsh brake duration per kilometer* and mean harsh brake duration per trip are used by the decision trees and random forest, indicating that they have a significant effect on the expected number of repairs. The tachograph speed is another feature that is often used by the models, where, according to the Logistic Regression model, metrics such as *low average tachograph speed* contribute to a lower expected number of repairs while *high average tachograph speed* contributes to a higher number of expected repairs. Next, the engine oil pressure is often used to determine the repair class of the trucks. In the Logistic Regression models a high pressure is related to less repairs and a low oil level is related to more repairs. Thus, a low oil level has a negative impact on the truck (engine). This is likely to be true as a low oil level causes the engine to run suboptimal or even fail when the oil level becomes too low. Next, the trip distance is another important metric. Although the direction of the relations cannot be directly derived from the models (large decision trees and random forests), it is highly likely that short trip distances contribute to more repairs as this means that the trucks have to make many stops and startups, while long trip distances relate to trucks for long distance transport, which are known at DAF to have less repairs. The DPA anticipation score is a metric that shows how well the driver is able to anticipate on traffic, based on the trucker's behaviour on the road. Thus, it is interesting to see that this score has an actual effect on the predicted number of repairs. Although it cannot be derived directly from the models, a higher DPA anticipation score likely contributes to a fewer number of repairs as it relates to a less aggressive driving style (i.e. better and more active anticipation). Furthermore, the ambient air temperature is incorporated in the decision tree, random forest and in the additional logistic regression model as given in Appendix K. From this model, there could be derived that a lower ambient air temperature is related to more repairs. Now, although it is possible that a low temperature results in strain on the engine and related parts (as it needs to get up to running temperature), it is also possible that specific regions and countries simply have a higher claim rate than others. Because the trucks in the dataset are from Europe only, the average ambient air temperature could be related to the region of operation. To check this, the claim rate has been derived for each country in the dataset, which is given in Table 35 (the claim rate is the average number of claims per truck).

Table 35: The claim rate per country, where the claim rate is the average number of claims per truck.

Country	ClaimCount	Nr. of trucks	Claim rate	Average temperature
Austria	12	9	1.33	6.35
UK	72	56	1.29	8.45
Netherlands	14	14	1.00	9.25
Italy	10	14	0.71	13.45
Germany	208	376	0.55	8.5
Switzerland	2	4	0.50	5.5
Spain/Portugal	153	326	0.47	13.3
Poland	13	29	0.45	7.85
France	35	87	0.40	10.7
Bulgaria	6	15	0.40	10.55
Romania	2	4	0.5	8.8
Belgium	5	13	0.38	9.55
Slovakia	1	3	0.33	6.8
Hungary	20	77	0.26	9.75
Czech Republic	0	5	0.00	7.55

Indeed, with the exception of Italy, the countries with a claim rate higher or equal to 0.50 generally have a relatively low average temperature (although not exclusively lower). This somewhat supports our suggestion that the ambient air temperature is related to the area of operation and thus the expected number of claims. However, as the number of trucks in each country (with the exception of Germany and Spain/Portugal) is rather limited in the available datasets, this conclusion cannot be properly validated and thus should be investigated further by DAF when more data becomes available.

Lastly, the intake air pressure from the engine has some influence on the expected number of repairs. According to the Logistic Regression model in Appendix K, a higher air-pressure is slightly related to less repairs. However, this cannot be further validated by the other models (due to the size of the decision trees and the limited comprehensibility of the random forests).

In conclusion, there can be said that the Connect variables have a relatively high prediction power compared to the truck specifications and contract data that is currently used by DAF. They are consistently favored by the prediction models when predicting the repair class of the trucks. However, due to the limited prediction power within the data, unambiguous relations between them cannot be given. Instead, based on the beta coefficients of the logistic regression models and the number of appearances as important features within the tested models, the top 10 of most informative repairs has been presented, which could be further explored upon usefulness in the future. This could be done by thoroughly consulting knowledge experts and by derivation of additional properties from the available Connect data regarding these features.

6.7 Comparison to the work of Goudsmits (2018)

As elaborated in Section 2.4, work on the prediction of the number of truck repairs has been executed at DAF before, with the difference that DAF Connect could not be used at that time. Instead, only truck specification information and contract details were used. To verify the added value of DAF connect, the performance of his models (without Connect) is compared to that of the models in this research. Goudsmits (2018) used a similar approach as this research by classifying trucks as having few of many

repairs based on the average number of repairs over time. The accuracy of his models and the models in this research are given in Table 36

Table 36: Comparison of the model performance (accuracy) of the work of Goudsmits (2018) and this research.

Model	Accuracy Goudsmits (2018)	Accuracy current work
Logistic Regression	0.68	0.69
Decision Tree	0.66	0.66
Random Forest	0.69	0.64
Neural network	Not used	0.66

As one can see, the modeling results are rather similar. At a first glance this seems illogical, as the addition of DAF Connect introduced much more detailed information about the trucks and their usage. However, the amount of available data has to be taken into account. DAF Connect is a new feature for DAF's trucks and thus, only a limited number of trucks has been fitted with the system at the time of this research.

Where Goudsmits (2018) used information from 2433 trucks to achieve the above described prediction performance, only 342 (11 months ahead dataset) trucks with DAF Connect have been used to reach a similar performance in this research (as more truck with DAF Connect simply were not available). Furthermore, he was able to do these predictions over a time horizon of two years, while with DAF connect, only 11 months of repair data was available. Thus, it can be concluded that with approximately 7 times less data instances (2433/342 trucks) and a roughly 2 times smaller time horizon (24/11 months), a similar prediction performance could already be reached by the addition of DAF Connect features for modeling. This indicates the potential of DAF Connect in the future. When more trucks and a larger time horizon become available it is expected that the models as used in this research will outperform the models from previous work.

7 Conclusions and recommendations

The last chapter of this document concludes the thesis. First, the individual sub-questions and finally the main research question are addressed. Secondly, recommendations, academic relevance, limitations and future research are discussed.

7.1 Research conclusions

In this subsection, the research subquestions and main research question are addressed and answered, based on the results of the research.

1. What cost/repair data is available?

Detailed information about the repairs on trucks is available in DAF's claim database. Among obvious information such as costs and the corresponding trucks, more detailed information about the repairs such as the type of repair, the associated components, labour hours involved and truck specifications can be found. This data has been used to define the number of repairs and associated costs per truck.

2. Which factors are currently used to predict repairs?

In Chapter 2 the current factors are explained in detail. Basically, DAF uses static truck specification data in combination with predefined usage profiles that a customer has to specify before buying a repair and maintenance contract at DAF. Based on this information, DAF derives expected costs and a maintenance category for the trucks. A downside of this method is that there is no method available to check if the users actually use their trucks as stated in the contract, which limits DAF's capability to derive specific costs per individual truck.

3. Which variables can be extracted from the data?

With the arrival of DAF Connect (the telemetry systems inside DAF's trucks) DAF has been able to collect detailed, real time, data about their trucks' operating state and health status. The factors that can be extracted are threefold:

- Truck specification data can be extracted from the repair and maintenance contract information of the trucks.
- Truck usage data can be derived using the data from DAF Connect, with which actual truck use and operating conditions could be derived.
- Details about the repairs of the trucks can be derived from the claim database at DAF.

The collected Connect data from DAF's trucks are basically sets of time series that are unique for each truck. In an effort to capture this, usage profiles have been derived for each truck by the extraction of variables from their time series. Minimums, maximums, averages, standard deviations and skewness of measurements have been derived for each truck and in addition, fuzzy bins have been applied to some of the relevant variables in order to try and capture detailed information about the probability distribution of the truck measurements (e.g. the distribution of truck speed and engine oil pressure). These features have subsequently been combined with the available truck specifications and contract data per truck such that their relevance compared to the currently used information for cost predictions could be verified.

4. Which prediction method is most suitable for the problem at hand?

Literature study showed that there is a range of methods that is commonly used in the field of predictive maintenance. Of these methods, MLP-Neural Networks and Random Forest were most often used and thus also incorporated in this research. In addition, Decision Trees and Logistic Regression has been included. As they are rather interpretable models, they can be used to provide DAF with insights about the specific features that had an influence on the expected number of repairs. Feature importance and beta coefficients of these models can be analyzed in order to provide DAF with insights on the effects of specific variables on the expected number of truck repairs. Based upon the results of the different models, the most suitable method has been identified, which is explained in the next section.

5. What is the performance of the most suitable models?

For the 8 months ahead prediction, the random forest proved to be the most suitable model. It achieved an accuracy of 63% in combination with a precision score of 64% and recall of 69%. Furthermore, relative feature importance could be derived from the model that provided DAF useful insights about the features used for these predictions.

Furthermore, for the 11 months ahead prediction, based solely on accuracy, the Logistic Regression model obtained the best result with an accuracy score of 69%. However, this came at a cost of a very limited recall score of 48%, which was most likely caused by the slight imbalance (40%-60% ratio) in the 11 months ahead dataset. This caused the model to bias itself towards the majority class. A more robust result was obtained by the random forest which had a lower accuracy of 64%, but a decent precision score of 67% and recall score of 64%.

In conclusion, the random forest shows the most potential for repair predictions on DAF's trucks, it has the most consistent results over both of the datasets (8 months- and 11 months ahead) compared to the other models. Furthermore, it had the best precision and recall scores and lastly, it provided some insights into the features that it selected based on the derived feature importances from the model.

6. What variables provide information about the expected number of truck repairs?

As each of the models used their own set of features for the prediction of expected repairs, an unambiguous answer to this question cannot be given. However, some features were used by multiple models and therefore, some form of feature importance validation has been done by counting the occurrence frequency of the variables that have been used by the best performing models. This resulted in a top 10 of most important features which is presented in Table 34 of the previous chapter. With the exception of one truck configuration feature (*Asset type*), the top 10 of most important features is comprised of features that have been derived from DAF Connect, indicating the potential of DAF Connect for the prediction of expected repairs on DAF's trucks. In short, these features can be divided into three categories, being driving related features, truck status features and indirect features. The features *Harsh braking duration*, *Tachograph speed*, *DPA anticipation scores* and *Max throttle duration* are directly influenced by the driver of the truck and thus categorized as driving related features. The features *Engine oil pressure* and *Engine intake air-pressure* are information about the truck's status and are categorized as truck status features. Lastly, the features *Asset type*, *Ambient air temperature* and *Trip distance* are features that are not directly influenced by the truck or the driver (assuming that a driver doesn't choose the trip distance, but receives this information from an external entity/company) and labeled as passive features. In conclusion these features provide the best information about the expected number of truck repairs.

With the sub-questions answered, the main question can be answered, which serves as a nice summary of the research.

Main research question: How can the number of truck repairs be predicted based on telemetry truck data and truck usage information?

To be able to use the telemetric truck data for repair predictions, useful features had to be extracted from the multivariate time-series data (telemetric truck data), as it couldn't be used as input for the prediction of the number of truck repairs directly. Instead, usage profiles per truck were derived by the extraction of global features from the telemetry data and using fuzzy histograms to approximate the probability distribution of the measurements as a feature vector.

As a next step, the time horizon over which to predict the number of repairs was established. Based on the available data at the time of research, a prediction horizon of 11 months was determined, simply because DAF Connect was still new and any data from before that time was deemed to be unreliable and inaccurate. Furthermore, an additional prediction horizon of 8 months ahead had been established to compare the effects of different time horizons on the prediction performance of the models.

Subsequently, four different modeling techniques were constructed and tested according to the CRISP-DM framework. These techniques were selected based on relevance in the literature and their ability to provide DAF with insights about the features that were used to make the predictions. Although the constructed Logistic Regression models, Decision Trees, Random Forests and MLP-Neural networks showed rather similar performances, the Random forests provided the most robust results. Simultaneously, it

allowed for (limited) insights into the relevant features that had been used to get those results based on the relative importance of the features in the model.

Lastly, The feature importance from the constructed decision trees and random forests, together with the beta coefficients derived from the Logistic Regression models were used to determine the most relevant features for the prediction of the number of truck repairs, which provided DAF with interesting insights about their data and the potential for repair predictions.

In conclusion, there has to be stated that the prediction performance of the models was limited. This was mainly caused by the fact that the available data only covered the early life of DAF's trucks, in which not many repairs actually occurred. Furthermore, the repairs that did occur were of many different origins, preventing the models from deriving robust patterns for specific or common repairs. Nonetheless, there is definitely potential in the available data from DAF Connect. Compared to the previous work of Goudsmits (2018), who predicted truck repairs for DAF without the available data from DAF Connect, the developed models in this research proved to be able to reach a similar prediction performance using approximately 7 times fewer data instances (trucks) and a roughly 2 times smaller time horizon for the number of repair predictions. Thus it is expected that in the future, when more trucks and data are available for analysis, the models in this research will outperform the methods from previous research.

7.2 Recommendations

Based on the findings in the research, several recommendations can be made for DAF.

At first, there was shown that the features that have been derived from DAF Connect show a better potential in the prediction of the numbers of repairs than the currently used truck specifications and contract details. Analysis of the constructed models showed that the Connect features were consistently favored above the currently used features. However, the current amount of available Connect data is too limited to make useful predictions for individual trucks. Therefore it is recommended for DAF to wait until more data is available before revisiting the problem of predicting the number of repairs. Furthermore, due to the many different origins and costs of truck repairs it is recommended to simplify the problem. Instead of predicting the exact number of repairs, dividing trucks in categories such as requiring *few repairs*, *average repairs* and *many repairs* respectively showed more potential. Ratings can then be applied to trucks (e.g. such as applied in credit ratings) to identify the risky and less risky trucks regarding the expected number of repairs. To provide DAF with an indication of the growth potential regarding the prediction of the number of repairs, learning curves for the Decision Tree model and the Random Forest (which proved to be the most suitable method for DAF) have been constructed. The results are shown in Figure 36

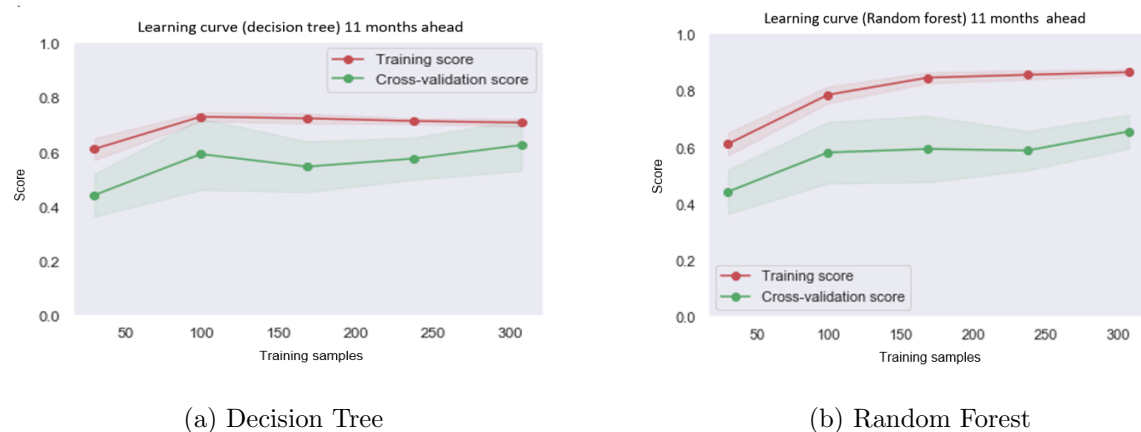


Figure 36: The learning curves for the 11 months ahead predictions

The learning curves show the 10-fold-cross-validated accuracy scores on the validation set over the number

of samples used for the predictions. As can be seen, a slight positive slope can be detected and thus, for more data, there can be expected to be better results. However, the slope only increases slightly, further indicating that DAF should only revisit the problem when significantly more training samples (trucks) have been collected.

Based on the results, it is recommended to DAF to focus on more narrowed down problems than the prediction of the total number of repairs over a given time horizon (at least for individual trucks). Instead, it is recommended to DAF to focus on the prediction of specific, expensive repairs before they actually happen. Extensive literature is available on the prediction of Remaining Useful Life, Future state and Health status of specific components or machines. For DAF, a prime example are the turbo and battery. They tend to fail rather regularly and are costly (at least the turbo) to replace. With the right sensors in place, these failures could be detected before they happen, allowing for the appropriate preventive actions to take place. It is of vital importance that DAF thinks about the implementation of the right sensors to do so now, at this point in time because information that can be derived with these sensors only becomes useful after months or even years of data collection. This is because currently available prediction methods need to learn from the past (or at least have reference values) in order to come up with meaningful predictions.

Furthermore, it is recommended to DAF to investigate the possibility to incorporate usage information as derived in this research into their current cost calculation methods when more data has become available. Based on the top 10 of most important features as described in the previous chapter, there can be concluded that some aspects of the Connect data are more informative than the currently used contract truck specifications and contract information. However, in this research, the usage information is derived over the first month of operation and therefore, their characteristics are known only after the sale of a truck and R&M contract. This information could be used to correct cost expectations 'on the go'. I.e. the driving behaviour of the connected trucks can be monitored while its operating and based on this, the expected costs and/or number of repairs for specific trucks can be adjusted in the forecasts. Subsequently, specific discounts could be offered to customers that prove to have a beneficial usage profile, either when they buy additional trucks or during the agreed contract period itself. As the harsh braking duration, tachographspeed, max throttle duration and DPA Anticipation score are directly influenced by the drivers and have been found to have an influence on the expected number of repairs, they could be used to derive such usage profiles.

Lastly, some general recommendations regarding the data can be given. Much information that was needed for this research was hidden in different databases and different sources (e.g. Mi database, CCM database and the HUE Hive for Connect data). In order to obtain it, many different and specific queries/scripts had to be written. Afterwards, they had to merged requiring another set of manual operations. This made the collection of data time consuming and prevents possibilities for automated analyses and data collection. It is therefore recommended for DAF to standardize their data and documents in a single (cloud) location such as provided by Microsoft Azure or Amazon web services (which are currently explored by DAF), allowing for much faster analyses and automation of data collection and processing.

7.3 Limitations

The available data at DAF was limited. At the time of research, the Connect data (telemetric data) had been recently introduced and only 11 months of reliable data was present. This time is considered to be the early life of a truck in which only a limited number of repairs occur. As trucks have an expected lifetime of at least 10 years, ideally we would want to have the same amount of information. Of course 4 years of data could also show relevant relations but at least there can be stated that less than a year of data is far from ideal. Additionally, the repairs that do happen are of a wide range of origins, preventing the models to learn robust patterns based on common or specific repairs.

As a result, the prediction performance of the initially constructed models was limited. Therefore, there was decided to make the problem a binary one, which was used to predict if a truck would have few or many repairs. Thus, no insights have been gained on the modeling performance regarding the prediction of the actual number of repairs.

Another limitation is that the number of trucks that was available for data analysis was limited. For the

8 months ahead predictions, 793 trucks were available, while for the 11 months ahead predictions only 342 trucks remained. As a result, the predictions might contain bias (e.g. if 10 trucks are operating in the Czech Republic and they all have a repair, the model will bias future trucks from the Czech Republic based on these 10 instances). Therefore, DAF has to be careful with the implementation of specific decision rules, which should be verified by experts and additional data analysis first.

Lastly, there are other machine learning techniques available that have not been used in this research. Deep learning for example, is gaining popularity fast and can be used to solve complex problems. However, they were not further investigated due to a few reasons. For one, the techniques used in this research were more prominently indicated to be useful for the problem at hand (predictive maintenance). Furthermore, compared to e.g. deep learning, the currently used models were able to provide more insights into their working and the features used. Lastly, time constraints (e.g. long computation times for deep neural networks) did not allow for further exploration of such methods.

7.4 Contribution to literature

To the best of our knowledge, this is the first research in which telemetric, real time, truck usage data on customer trucks has been used to predict the number of repairs over an extended period of time. Although much research about predictive maintenance is available, it focuses on RUL, future state and health status over short periods (hours, days, weeks) in time. Therefore, this research provides a unique approach on predictive maintenance predictions for machines by combining both telemetric data and truck (machine) specification information to derive the number of repairs over a given time period.

Secondly, fuzzy histograms have been applied to derive detailed information about the probability distribution of the real time truck measurements by approximating them with fuzzy bins. It showed that the prediction performance increased for both Logistic Regression and Decision Tree models. As fuzzy histograms are not widely used in literature, this research provides a unique use case and shows that fuzzy histograms can be used for predictive maintenance in order to improve the prediction accuracy.

7.5 Future research

First of all, there has been determined that the number of repairs compared to the range of repair types during the early life (first year) of truck operations is too low to make accurate predictions on exact number of repairs or exact costs per truck. It is recommended that DAF does not invest in these type of predictions for now, as they are too broad. In the future, (i.e. a few years) when more relevant data has been collected, the problem could be revisited. The findings in this research can then be used as a starting point and methodical explanation for the method to predict the number of repairs on trucks.

Furthermore, the research was performed by extraction of global features from the telemetric time series data, while more detailed local features might provide additional insights. In an effort to improve the modeling results and analyze the effects of more detailed features on the prediction performance, fuzzy bins were constructed. They improved the prediction performance of the models, which indicated that the models might further benefit from additionally derived features. Abnormal local patterns in the data could be derived or features could be combined in an effort to improve the model input. However, deriving useful local features and combinations is a complex task on which a lot of expert knowledge might be needed (about the truck's operating conditions and abnormal behavior patterns) and was not further explored due to time constraints. In the future this could be investigated further.

Lastly, this research was about the prediction of the number of repairs in general. However, as the results have proved to be of limited accuracy there could be decided to narrow the scope of future research projects regarding repair predictions. For example, when more data has become available, there could be focused on the prediction of specific expensive or commonly encountered drive-line related repairs over a given time horizon as these might be more directly related to the usage profiles of the trucks (e.g. the prediction of turbo failures or battery breakdowns). However, having the right sensors in place is vital for the successful execution of such predictions.

8 Appendices

A Overview of the available CCM truck data at DAF.

The variables that are present in the CCM truck dataset are given in table 37.

Table 37: Overview and explanations of the available CCM truck data.

Variable	Explanation
Forecasting Run Date	Date at which the contract cost forecast is run.
Forecasting Report Date	Date at which the contract cost forecast is reported (mostly same date).
Subsidiary	Subsidiary with whom the contract is closed.
Country	Country in which the subsidiary is located.
Contract Number	Unique contract identification number.
Contract Version	Version of the current R&M Contract
Contract Group	Group based on different contract settings.
Contract Name	Contract name, as a combination of the contract number, contract group and chassis-number.
current Contract (version) Status	Indicator of the contract status (active, canceled, expired or on hold).
snapshot Contract (version) Status	Indicator of the DAF Connect contract status (active, canceled, expired or on hold).
Contract Birthdate	Date at which the contract is drafted.
Contract (1st) activation date	Date at which the contract becomes active.
Contract (version) Start date	Date at which the contract is drafted.
Contract (version) activation date	Date at which the contract should become active.
Contract (version) End date (original)	Date at which the contract should terminate.
Contract end-date (actual)	Date at which the contract has been terminated (if so).
Contract end-year (actual)	Year at which the contract has been terminated (if so).
Contract closing date	Date at which the contract has been closed with the customer.
Contract closing year	Year at which the contract has been closed with the customer.
Contract (version) duration in months (original)	Contract duration according to terms.
Contract (version) duration in months (actual)	Actual duration of the current contract.
Contract (overall) duration in months (actual)	Duration of all contract versions combined.
FinVehAge	Vehicle age at the time of contract information inquiry.
Contract (overall) Age in months (actual)	Vehicle age at the time of contract information inquiry.
Contract (version) start kms	Kilometers driven by the truck at the date of contract drafting.

Contract contracted yearly mileage	Estimate of the truck's yearly mileage during the contract time.
Contract Origin	Boolean to tell if the contract concerns a new or used truck.
Contract package	Chosen R&M contract type.
Selling dealer	Dealer that sold the truck.
Selling dealer Location code	DAF internal location code for the selling dealer.
Default Service Dealer	Service dealer that is responsible for the truck's repair and maintenance.
Default Service dealer Location code	DAF internal location code for the service dealer.
Currency	Currency used in the truck and contract sale.
Claim delay Date	The date after which claims could potentially be filed but not yet processed by the forecast.
Last date invoiced	The last date at which the customer has received a claim reimbursement.
Model	Truck model specification (Axle and truck type).
Series	Model series (truck type and model year)
Sub series	Model year
Chassis number	Unique truck chassis number.
Brand	Truck brand.
Engine power	Fitted engine's power in kilowatt.
Axle configuration	Type of axle fitted.
Emission	European emission class of the truck.
Asset Description	Truck type, engine power in horsepower and axle type fitted.
Asset Type-info	Axle type, truck type, engine power in horsepower, engine type, cabin type and wheelbase size respectively.
Delivery date	Date of delivery from the factory to the selling dealer.
First Registration Date	The data upon which the license plate is registered.
Vehicle Park Number	Reference number for the truck as used by the customer.
Delivery Country	Country in which the truck is delivered.
S&M Inspection Interval ('O' licence)	Interval duration for service and maintenance inspections in weeks.
Vehicle Safety Features	Type of vehicle safety features fitted.
Soot Filter	Soot filter fitted (True/False).
Service Interval Engine	Type of engine service interval chosen (standard/extended).
Retarder System	Type of retarder system (braking aid) fitted.
Fuel Specification	Type of fuel for the truck.
Factory External Camera System	Camera system fitted (True/False).
Body Specification	Body type of the truck.
Taillift Fitted	Tail lift fitted on the truck (True/False).
ADR Specification	Allowed to carry hazardous material (True/False)
(Semi-) Trailer Coupling	Type of trailer coupling fitted.
Rear Axle Oil	Type of rear axle oil in the truck.

Gearbox Oil	Type of gearbox oil in the truck.
Engine Oil	Type of engine oil in the truck.
Driven Axle Suspension	Type of driven axle (shaft) suspension (if any).
Body Type	Type of truck body fitted.
Axle configuration	Axle type fitted.
Engine Type	European emission class of the truck.
Engine	Engine power in horsepower.
Gearbox	Type of gearbox fitted.
Number of Drops per day	Number of expected drops per day of the truck during contract time.
Area of Operation	Geographical area that the truck will operate in (Western europe, North Africa etc.).
Type of Operation	Operation class of the truck (long distance, regional or local).
Power Take Off (PTO)	Power take off installed (True/False).
Static PTO Hours per day	Expected number of hours that the truck will use the PTO per day.
Rear Axle Type	Type of rear axle fitted.
Road Type	Expected percentage of time that the truck will drive offroad.
Application	Expected cargo type that the truck will transport (e.g. pallets, containers, livestock).

B Overview of the available Mi claim data at DAF.

The variables that are present in the Mi claim dataset are given in table 38.

Table 38: Overview and explanations of the available Mi claim data.

Variable	Explanation
ChassisNr	The chassisnumber of the truck.
TypeName	Body type, model and series of the truck.
ProductRange	Model and eninge type of the truck.
productionsite	City where the truck is manufactured.
prodDate	Date of production.
prodMonth	Month of production.
DeliveryDate	Date that the truck is delivered to the customer.
deliveryCountry	Country of delivery.
deliveryDealer	Dealer to which the truck is delivered.
ClaimCountry	Country from which the claim is made.
ClaimDealer	Dealer that has claimed the repair.
Claimnr	Claim identification number.
ClaimSort	Indicates on which type of contract the claim is made.
FieldReportYN	Boolean that indicates if the claim is a field action or not.
warrantycategory	Indicates on which type of contract the claim is made.
LastClaimStatus	Indicates whether (if present) a previous claim has been handled or not.
Laststatusstartdate	Indicates when the last claim status update has been altered (for previous claims).
DefectCode	Code of the defect that is claimed
defectcodedescription	Description of the code.
defectcause	Reason for claim.
CasualPart	Part that has caused the claim/repair.
DefectDate	Date on which the defect occurred.
defectmonth	Mont in which the defect occurred.
ClaimReceiveDate	Date on which the claim was received.
Claimfinalised	Date on which the claim was handled.
KmChassis	Km count on the chassis.
KmPart	Km count on the repaired part.
KindofPart	Type of part that has been repaired
MATT_CLAIMED	Total costs of the material that has been claimed.
LABOUR_CLAIMED	Total costs of the labour that has been claimed.
MISC_CLAIMED	Total costs of miscellaneous that has been claimed.
TOTAL_CLAIMED	The sum of the costs of the claim.

MATT_PAID_DTNV	The amount as reimbursed by DAF for the material claim.
LABOUR_PAID_DTNV	The amount as reimbursed by DAF for the labour claim.
MISC_PAID_DTNV	The amount as reimbursed by DAF for the miscellaneous claim.
PAID_MIN_LANDED	The total amount paid by DAF deducted by the landed costs.
LANDED_COSTS_DTNV	A percentage of materail costs paid to cover miscellaneous (tiny) part costs.
TOTAL_PAID_DTNV	The total amount as reimbursed by DAF.
ArticlePaidDN_LocalPolicy	The amount as reimbursed by the dealer itself for the material claim.
LabourPaidDN_LocalPolicy	The amount as reimbursed by the dealer itself for the labour claim.
MiscPaidDN_LocalPolicy	The amount as reimbursed by the dealer itself for the miscellaneous claim.
HandlingPaidDN_LocalPolicy	The amount as reimbursed by the dealer itself for the handling claim
TotalPaidLocalPolicy	The total amount as reimbursed by the dealer itself.
HoursClaimed	The total number of labour hours claimed.
Artnr1	Number of the article that has been replaced.
Artnr2	Number of the article that has been replaced.
Artnr3	Number of the article that has been replaced.
Artnr4	Number of the article that has been replaced.
Artnr5	Number of the article that has been replaced.
labourcode1	Labour code associated with the repair action.
labourcode2	Labour code associated with the repair action.
labourcode3	Labour code associated with the repair action.
labourcode4	Labour code associated with the repair action.
labourcode5	Labour code associated with the repair action.
Misc1	Free text field
Misc2	Free text field
Misc3	Free text field
Misc4	Free text field
Misc5	Free text field
ITSCode	Code for the International Truck Service action when applicable.
MonthInService	Months that the trucks has been in service.
DriveLineYN	Boolean that indicates if the repair is on the driveline or not.
rejectioncode	Reason of claim rejection if applicable.
acceptedclaimwarranty	Boolean that indicates if the claim is fully accepted by DAF or not.
AcceptedClaimLocalPolicy	Boolean that indicates if (parts of) the claim is paid by the dealer itself or not.
serviceproduct	Contract on which the repair has been reimbursed.

C Overview of the available connect interval data at DAF.

The measurements that are sent from each individual truck as snapshots (5-minute intervals) are listed in Table 39, together with their variable explanation.

Table 39: Overview and explanation of the snapshot data.

Snapshot variable	Description
snapshotkey	Unique hash for each snapshot message (primary key).
datetime	Date/time of snapshot event in UTC.
unixtimestamp	Unixepoch of snapshot event in seconds since 1-1-1970.
gpsdatetime	Date/time when gps position was recorded in UTC.
gpslatitude	Gps latitude in degrees.
gpslongitude	Gps longitude in degrees.
totaldistance	All time vehicle distance since the start of operation (odometer) in meters.
eventid	Integer value identifying the type of snapshot.
distanceuntilservice	Remaining distance until service is required in meters.
totalfuelconsumption	All time fuel consumption since start of operation in milliliters.
idle_duration	Duration of vehicle speed <0.1 km/h since the beginning of a trip in seconds.
dcmserialno	Unique serial number of the DCM (Daf Connect Module).
dcmswversion	Software version of the embedded DCM software.
gpsaltitude	Altitude according to GPS in meters above sea level.
gpsheading	Heading according to GPS in degrees (0 is north, 180 is south).
fuellevel	Truck fuel tank level in percentage.
aftertreatmentlevel	Truck adblue tank level in percentage.
grosscombinationweight	Vehicle gross combination weight in kg.
wheelbasedspeed	Vehicle speed based on wheel sensor.
tachographspeed	Vehicle speed based on tachograph sensor.
gps_distance	Distance traveled based on high frequency GPS location algorithm since the start of the trip in meters.
enginecoolantlevel_1	Engine coolant level in percentage in one minute intervals (minute 1).
enginecoolantlevel_2	Engine coolant level in percentage in one minute intervals (minute 2).
enginecoolantlevel_3	Engine coolant level in percentage in one minute intervals (minute 3).
enginecoolantlevel_4	Engine coolant level in percentage in one minute intervals (minute 4).
enginecoolantlevel_5	Engine coolant level in percentage in one minute intervals (minute 5).
engineoillevel_1	Engine oil level in percentage in one minute intervals (minute 1).
engineoillevel_2	Engine oil level in percentage in one minute intervals (minute 2).
engineoillevel_3	Engine oil level in percentage in one minute intervals (minute 3).
engineoillevel_4	Engine oil level in percentage in one minute intervals (minute 4).

engineoillevel_5	Engine oil level in percentage in one minute intervals (minute 5).
aftertreatmentlevel_1	Adblue (diesel exhaust gas cleaner) tank level in percentage in one minute intervals (minute 1).
aftertreatmentlevel_2	Adblue (diesel exhaust gas cleaner) tank level in percentage in one minute intervals (minute 2).
aftertreatmentlevel_3	Adblue (diesel exhaust gas cleaner) tank level in percentage in one minute intervals (minute 3).
aftertreatmentlevel_4	Adblue (diesel exhaust gas cleaner) tank level in percentage in one minute intervals (minute 4).
aftertreatmentlevel_5	Adblue (diesel exhaust gas cleaner) tank level in percentage in one minute intervals (minute 5).
barometricpressure_1	Ambient air pressure in kilo-pascal in one minute intervals (minute 1).
barometricpressure_2	Ambient air pressure in kilo-pascal in one minute intervals (minute 2).
barometricpressure_3	Ambient air pressure in kilo-pascal in one minute intervals (minute 3).
barometricpressure_4	Ambient air pressure in kilo-pascal in one minute intervals (minute 4).
barometricpressure_5	Ambient air pressure in kilo-pascal in one minute intervals (minute 5).
fuellevel_1	Fuel tank level in percentage in one minute intervals (minute 1).
fuellevel_2	Fuel tank level in percentage in one minute intervals (minute 2).
fuellevel_3	Fuel tank level in percentage in one minute intervals (minute 3).
fuellevel_4	Fuel tank level in percentage in one minute intervals (minute 4).
fuellevel_5	Fuel tank level in percentage in one minute intervals (minute 5).
fueltemperature_1	Fuel temperature in degrees celcius in one minute intervals (minute 1).
fueltemperature_2	Fuel temperature in degrees celcius in one minute intervals (minute 2).
fueltemperature_3	Fuel temperature in degrees celcius in one minute intervals (minute 3).
fueltemperature_4	Fuel temperature in degrees celcius in one minute intervals (minute 4).
fueltemperature_5	Fuel temperature in degrees celcius in one minute intervals (minute 5).
engineoiltemperature_1	Engine oil temperature in one minute intervals (minute 1).
engineoiltemperature_2	Engine oil temperature in one minute intervals (minute 2).
engineoiltemperature_3	Engine oil temperature in one minute intervals (minute 3).
engineoiltemperature_4	Engine oil temperature in one minute intervals (minute 4).
engineoiltemperature_5	Engine oil temperature in one minute intervals (minute 5).
engineoilpressure_1	Engine oil pressure in kilo-pascal in one minute intervals (minute 1).
engineoilpressure_2	Engine oil pressure in kilo-pascal in one minute intervals (minute 2).
engineoilpressure_3	Engine oil pressure in kilo-pascal in one minute intervals (minute 3).
engineoilpressure_4	Engine oil pressure in kilo-pascal in one minute intervals (minute 4).
engineoilpressure_5	Engine oil pressure in kilo-pascal in one minute intervals (minute 5).
enginecoolanttemperature_1	Engine coolant temperature in degrees celcius in one minute intervals (minute 1).
enginecoolanttemperature_2	engine coolant temperature in degrees celcius in one minute intervals (minute 2).

enginecoolanttemperature_3	engine coolant temperature in degrees celcius in one minute intervals (minute 3).
enginecoolanttemperature_4	engine coolant temperature in degrees celcius in one minute intervals (minute 4).
enginecoolanttemperature_5	engine coolant temperature in degrees celcius in one minute intervals (minute 5).
servicebrakeairpressure_1	Service-brake air pressure in kilo-pascal in one minute intervals (minute 1).
servicebrakeairpressure_2	Service-brake air pressure in kilo-pascal in one minute intervals (minute 2).
servicebrakeairpressure_3	Service-brake air pressure in kilo-pascal in one minute intervals (minute 3).
servicebrakeairpressure_4	Service-brake air pressure in kilo-pascal in one minute intervals (minute 4).
servicebrakeairpressure_5	Service-brake air pressure in kilo-pascal in one minute intervals (minute 5).
engineload_1	Engine load in percentage in one minute intervals (minute 1).
engineload_2	Engine load in percentage in one minute intervals (minute 2).
engineload_3	Engine load in percentage in one minute intervals (minute 3).
engineload_4	Engine load in percentage in one minute intervals (minute 4).
engineload_5	Engine load in percentage in one minute intervals (minute 5).
enginespeed_1	Engine speed in rpm in one minute intervals (minute 1).
enginespeed_2	engine speed in rpm in one minute intervals (minute 2).
enginespeed_3	engine speed in rpm in one minute intervals (minute 3).
enginespeed_4	engine speed in rpm in one minute intervals (minute 4).
enginespeed_5	engine speed in rpm in one minute intervals (minute 5).
engineintakeairpressure_1	Air pressure in the intake manifold of the engine in kilo-pascal in one minute intervals (minute 1).
engineintakeairpressure_2	Air pressure in the intake manifold of the engine in kilo-pascal in one minute intervals (minute 2).
engineintakeairpressure_3	Air pressure in the intake manifold of the engine in kilo-pascal in one minute intervals (minute 3).
engineintakeairpressure_4	Air pressure in the intake manifold of the engine in kilo-pascal in one minute intervals (minute 4).
engineintakeairpressure_5	Air pressure in the intake manifold of the engine in kilo-pascal in one minute intervals (minute 5).
tachographspeed_1	Vehicle speed based on tachograph sensor in one minute intervals (minute 1).
tachographspeed_2	vehicle speed based on tachograph sensor in one minute intervals (minute 2).
tachographspeed_3	vehicle speed based on tachograph sensor in one minute intervals (minute 3).
tachographspeed_4	vehicle speed based on tachograph sensor in one minute intervals (minute 4).
tachographspeed_5	vehicle speed based on tachograph sensor in one minute intervals (minute 5).
totaldistance_1	All time vehicle distance since the start of operation (odometer) in meters in one minute intervals (minute 1).
totaldistance_2	All time vehicle distance since the start of operation (odometer) in meters in one minute intervals (minute 2).
totaldistance_3	All time vehicle distance since the start of operation (odometer) in meters

	in one minute intervals (minute 3).
totaldistance_4	All time vehicle distance since the start of operation (odometer) in meters in one minute intervals (minute 4).
totaldistance_5	All time vehicle distance since the start of operation (odometer) in meters in one minute intervals (minute 5).
gpspeed	Speed of the truck in km/h measured by the GPS unit.
ambientairtemperature	Air temperature outside of the truck in degrees celsius.
enginecoolanttemperature	Temperature of the engine coolant fluid in degrees celsius.
gpshdop	Accuracy indication for the gps position.

The set of measurements that is sent to DAF after each completed trip is given in Table 40. It contains aggregated data about the trips of individual trucks.

Table 40: Variable explanations of the trip data.

Trip data variable	Explanation
datetime_begin	Date/time at the start of the trip in UTC.
unixtimestamp_begin	Unixepoch timestamp at the start of the trip in seconds since 1-1-1970.
datetime_end	Date/time of end of trip in UTC.
unixtimestamp_end	Unixepoch of end of trip in seconds since 1-1-1970.
gpsdatetime_end	Date/time when gps position was recorded at end of the trip in UTC.
gpslatitude_end	Gps latitude at end of the trip in degrees.
gpslongitude_end	Gps longitude at end of the trip in degrees.
gpsdatetime_begin	Date/time when the gps position was recorded at the start of the trip in UTC.
gpslatitude_begin	Gps latitude at the start of the trip in degrees.
gpslongitude_begin	Gps longitude at the start of the trip in degrees.
totaldistance_begin	All time vehicle distance since the start of operation (odometer) at the start of the trip in meters.
totaldistance_end	All time vehicle distance since the start of operation (odometer) at the end of the trip in meters.
tripkey	Unique hash for each unique trip message (trip identifier).
dcmserialno	Unique serial number of the DCM (Daf Connect Module).
dcmswversion	Software version of the embedded DCM software.
brake_duration	Duration of active braking during the trip in seconds.
cruisecontrol_distance	Distance with cruise control enabled during the trip in meters.
harshbrake_duration	Duration of active breaking and deceleration $>2.5\text{m/s}^2$ during trip in seconds.
idling_duration	Duration of the vehicle speed being $<0.1\text{ km/h}$ during the trip in seconds.
gps_elevationloss	Total elevation loss during the trip, according to the GPS, in meters.
gps_elevationgain	Total elevation gain during the trip, according to the GPS, in meters.
pto_count	Count of the number of PTO enabled / disabled cycles, where PTO stands for Power Take Off. The PTO is used to e.g. power waste crushers in garbage trucks which uses energy from the engine.
pto_distance	Distance traveled with the PTO enabled during the trip in meters.
pto_duration	Duration of the PTO enabled during the trip in seconds.
totalfuelconsumption_begin	All time fuel consumption since start the of operation
fuellevel_begin	Level of the fuel tank at the start of the trip in percentage.
totalfuelconsumption_end	All time fuel consumption since the start of operation at the end of the trip in milliliters.

fuellevel_end	Level of the fuel tank at end of the trip in percentage.
gps_distance	Total trip distance based on high frequency GPS location algorithm in meters.
idling_fuelconsumption	Total fuel consumption with vehicle speed <0.1 km/h during the trip in milliliters.
acceleration_duration	Total duration of vehicle acceleration >0.1 m/s ² during the trip in seconds.
maxthrottlepaddle_duration	Total duration of the throttle paddle activation >95% during the trip in seconds.
dpabrakingscore_sum	The sum of all DPA (Driving performance assistant) braking event scores during the trip.
dpaaanticipationscore_sum	The sum of all DPA anticipation event scores during the trip.
dpabrakingevent_count	The number of dpa braking events during the trip.
dpaaanticipationevent_count	The number of dpa anticipation events during the trip.
cruisecontrol_fuelconsumption	Total fuel consumption while cruise control is enabled during the trip in milliliters.
gpsspeed	Vehicle speed based on GPS.
cruisecontrol_distanceclass_1	Total distance traveled during the trip in meters, with cruise control enabled in km/h class 1: 0-25.
cruisecontrol_distanceclass_2	Total distance traveled during the trip in meters, with cruise control enabled in km/h class 2: 25-50.
cruisecontrol_distanceclass_3	Total distance traveled during the trip in meters, with cruise control enabled in km/h class 3: 50-75.
cruisecontrol_distanceclass_4	Total distance traveled during the trip in meters, with cruise control enabled in km/h class 4: 75-100.
cruisecontrol_distanceclass_5	Total distance traveled during the trip in meters, with cruise control enabled in km/h class 5: 100-125.

The variables that are present in the trigger dataset are given in Table 41.

Table 41: Variable explanations of the trigger data.

Trigger data variable	Explanation
triggerkey	Unique hash for each trigger message (primary key).
inputfiledate	Date of the trigger event.
datetime	Date/time of the trigger event in UTC.
unixtimestamp	Unixepoch timestamp of the trigger event in seconds since 1-1-1970.
gpsdatetime	Date/time according to the gps upon trigger activatin, recorded in UTC.
gpslatitude	Gps latitude in degrees
gpslongitude	Gps longitude in degrees
eventid	Integer value identifying the type of trigger (decoded in theeventname database).
dcmserialno	Unique serial number of the DCM.
dcmswversion	Software version of the embedded DCM software.
gpsaltitude	Altitude according to GPS in meters above sealevel.
gpsheading	Heading according to GPS in degrees (0 is north, 180 is south).
dm01spn	Suspect parameter number decoded from the J1939 DM01 CAN (Controller Area Network) message.
dm01fmi	Failure mode identifier decoded from J1939 DM01 CAN message.
dm01occ	Occurance count decoded from J1939 DM01 CAN message.
dm01red	Red stop lamp status (True / False) decoded from the J1939 DM01 CAN message.
dm01yellow	Amber warning lamp status (True / False) decoded from the J1939 DM01 CAN message.
dm01mil	Malfunction indicator lamp status (True/False) decoded from the J1939 DM01 CAN message.
totaldistance	All time vehicle distance since the start of operation (odometer) in meters.
dm01sa	Source address of J1939 DM01 CAN message.
dm01active	Indicator (True\False) of an active lamp (yellow/red/malfunction).
ttblockid	Telltale block id decoded from J1939 FMS1 CAN message.
ttstatusid	Telltale status id decoded from J1939 FMS1 CAN message.
ttvalue	Vehicle identification number.
gpsspeed	The truck's speed in km/h measured by the GPS.
ambientairtemperature	Air temperature outside of the truck in degrees celsius.
enginecoolanttemperature	Engine coolant temperature in degrees celcius.
servicebrakeairpressure1	Service-brake air pressure in kilo-pascal in one minute intervals.
servicebrakeairpressure2	Service-brake air pressure in kilo-pascal in one minute intervals.
fuellevel	Truck fuel tank level in percentage.
aftertreatmentlevel	Truck adblue tank level in percentage.

grosscombinationweight	Vehicle gross combination weight in kg.
wheelbasedspeed	Vehicle speed based on wheel sensor.
tachographspeed	Vehicle speed based on tachograph sensor in one minute intervals.
retardertorqueactual	Retarder rpm (the retarder is a torque converter that helps the truck to break (e.g. while going downhill)).
retardertorquemode	The current retarder mode.
pedalbreakposition1	Tells how far down the break pedal is pressed in percentage.
pedalacceleratorposition1	Tells how far down the acceleration pedal is pressed in percentage.
enginespeed	Engine speed in rpm.
engineload	The engine power used in percentage of total power.
gearcurrent	Gear that the truck is in.
gearselected	Gear that is actually selected.
ptoengaged	Indicator (True/False), telling if the PTO is activated or not.
cruisecontrol	Indicator, telling if cruise control is active (True/False).
cumulatedfuel	All time fuel consumption since start of operation in milliliters.
distanceuntilservice	Remaining distance until service is required in meters.
enginototalhours	All time number of hours that the truck engine has been running (since the start of operation).
powerbatteryvoltage	Current battery voltage.
warningclass	Class from 1-10, indicating which warning light is activated on the driver dashboard.
warningnumber	Unique hash, used as identifier for the specific warning message.
warningstate	Indicator (True/False) of an active warning light on the dashboard.
vevtcause	Description of the event observed that generated the message.
eventname	Indicator of lamp event type (lamp raised/lamp still active/lamp cleared)
tripkey	Unique hash for each unique trip message (trip identifier).
vin	Vehicle identification number.
month	Month in which the trigger event occurred.

D Numerical summary of the snapshot data

Table 42: Quantitative summary of the numerical snapshot data.

Variable	min	max	mean	median	std
aftertreatmentlevel	0	102	77.6	82	21.19
aftertreatmentlevel_1	0	102	76.69	81.2	20.88
aftertreatmentlevel_2	0	102	76.77	81.2	20.84
aftertreatmentlevel_3	0.4	102	76.79	81.6	20.83
aftertreatmentlevel_4	0	102	76.81	81.6	20.81
aftertreatmentlevel_5	0	102	76.84	81.6	20.78
ambientairtemperature	0	65,535.00	2,217.60	12	11,818.71
barometricpressure_1	81	104	98.92	100	2.43
barometricpressure_2	80	103	98.9	100	2.44
barometricpressure_3	80	104	98.89	100	2.45
barometricpressure_4	80	105	98.89	99	2.45
barometricpressure_5	81	105	98.88	99	2.46
distanceuntilservice	-32,767,000.00	32,765,000.00	2,192,938.76	-1,000.00	14,906,454.89
enginecoolantlevel_1	0	100	100	100	0.06
enginecoolantlevel_2	0	100	100	100	0.09
enginecoolantlevel_3	0	100	100	100	0.07
enginecoolantlevel_4	0	100	100	100	0.07
enginecoolantlevel_5	100	100	100	100	0
enginecoolanttemperature	-1	255	79.89	85	17.94
enginecoolanttemperature_1	0	255	81.84	86	13.88
enginecoolanttemperature_2	0	255	82.88	86	12.19
enginecoolanttemperature_3	1	104	83.29	86	11.55
enginecoolanttemperature_4	3	105	83.65	86	10.94
enginecoolanttemperature_5	4	103	84.26	86	9.76
engineintakeairpressure_1	80	510	155.14	122	77.19
engineintakeairpressure_2	82	510	158.72	126	77.01
engineintakeairpressure_3	82	510	159.94	128	76.98
engineintakeairpressure_4	84	510	160.84	130	76.86
engineintakeairpressure_5	82	510	162	132	76.7
engineload_1	0	100	29.19	22	28.31
engineload_2	0	100	28.05	18	28.83
engineload_3	0	100	28.57	20	29.05
engineload_4	0	100	28.95	21	29.18

engineload_5	0	100	29.43	22	29.32
engineoillevel_1	8	102	100.38	100.4	0.74
engineoillevel_2	100	102	100.39	100.4	0.21
engineoillevel_3	100	102	100.39	100.4	0.2
engineoillevel_4	100	102	100.39	100.4	0.2
engineoillevel_5	100	102	100.39	100.4	0.2
engineoilpressure_1	0	744	236.22	252	94.73
engineoilpressure_2	0	748	257.95	252	67.29
engineoilpressure_3	0	716	258.53	252	67.03
engineoilpressure_4	0	724	259.05	252	66.81
engineoilpressure_5	0	672	259.38	252	66.5
engineoiltemperature_1	-6	1,774.00	116.86	104	190.07
engineoiltemperature_2	-6	1,774.00	118.12	105	185.08
engineoiltemperature_3	-4	1,774.00	118.01	105	181.72
engineoiltemperature_4	-2	1,774.00	117.92	105	178.66
engineoiltemperature_5	3	1,774.00	118.29	106	175.65
enginespeed_1	0	2,792.00	934.86	1,045.00	320.05
enginespeed_2	0	2,800.00	989.76	1,059.00	262.06
enginespeed_3	0	2,660.00	997.5	1,064.00	258.36
enginespeed_4	0	2,554.00	1,004.05	1,068.00	255.03
enginespeed_5	0	2,616.00	1,012.85	1,073.00	250.48
enginotalhours	-1	1,358.00	140.14	100	179.1
fuellevel	0	100	69.13	70	25.66
fuellevel_1	0	100	69.21	70	25.55
fuellevel_2	0	100	68.87	69	25.53
fuellevel_3	0	100	68.83	69	25.55
fuellevel_4	0	100	68.83	69	25.55
fuellevel_5	0	100	68.81	69	25.55
fueltemperature_1	-4	215	35.39	32	23.94
fueltemperature_2	-3	215	33.19	31	22.19
fueltemperature_3	-3	215	32.76	31	21.76
fueltemperature_4	-3	215	32.57	31	21.43
fueltemperature_5	-3	215	32.54	31	21.09
gps_distance	0	418,631.00	4,054.18	5,213.00	3,126.70
gpsaltitude	0	32,500.00	204.41	131	205.17
gpslatitude	36.03	255	55.97	49.18	40.64
gpslongitude	-9.43	255	15.6	6.96	48.99
gpsspeed	0	118.74	15.94	18.91	17.12

grosscombinationweight	3,200.00	655,350.00	209,453.05	38,800.00	282,667.96
idle_duration	-5	282,441,616.00	303.17	0	241,632.35
servicebrakeairpressure_1	0	1,216.00	178.51	120	254.12
servicebrakeairpressure_2	0	1,216.00	177.47	120	258.96
servicebrakeairpressure_3	0	1,216.00	177.82	120	259.43
servicebrakeairpressure_4	0	1,216.00	178.06	120	259.63
servicebrakeairpressure_5	0	1,216.00	176.77	120	256.67
servicebrakeairpressure1	0	65,535.00	3,902.19	128	15,149.48
servicebrakeairpressure2	0	65,535.00	3,895.32	128	15,135.91
tachographspeed	0	123	49.11	67	39.36
tachographspeed_1	0	121	54.7	76	37.6
tachographspeed_2	0	123	58.57	79	35.62
tachographspeed_3	0	123	60.05	80	34.91
tachographspeed_4	0	125	61.3	81	34.29
tachographspeed_5	0	126	62.95	82	33.41
totaldistance	0	94,235,735.00	12,760,532.30	9,846,645.00	11,261,477.27
totaldistance_1	0	94,228,464.00	12,819,211.99	9,909,165.00	11,281,552.03
totaldistance_2	0	94,229,904.00	12,869,253.59	9,959,315.00	11,299,494.93
totaldistance_3	0	94,231,344.00	12,888,000.57	9,979,300.00	11,304,084.72
totaldistance_4	0	94,232,784.00	12,907,043.66	9,998,165.00	11,312,532.97
totaldistance_5	0	94,234,240.00	12,944,436.19	10,044,005.00	11,316,709.97
totalfuelconsumption	200,372.00	31,615,098.00	3,764,006.44	2,870,086.50	3,410,584.40
wheelbasedspeed	0	124	49.1	67	39.33

E Overview of the tasks that belong to each phase of the CRISP-DM model.

Each phase of the CRISP-DM model comes with its own set of tasks. An overview of these tasks is given in Figure 37.

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives Background Business Objectives Business Success Criteria Assess Situation Inventory of Resources Requirements, Assumptions, and Constraints Risks and Contingencies Terminology Costs and Benefits Determine Data Mining Goals Data Mining Goals Data Mining Success Criteria Produce Project Plan Project Plan Initial Assessment of Tools and Techniques	Collect Initial Data <i>Initial Data Collection Report</i> Describe Data <i>Data Description Report</i> Explore Data <i>Data Exploration Report</i> Verify Data Quality <i>Data Quality Report</i>	Select Data <i>Rationale for Inclusion/Exclusion</i> Clean Data <i>Data Cleaning Report</i> Construct Data <i>Derived Attributes Generated Records</i> Integrate Data <i>Merged Data</i> Format Data <i>Reformatted Data Dataset Dataset Description</i>	Select Modeling Techniques <i>Modeling Technique Modeling Assumptions</i> Generate Test Design <i>Test Design</i> Build Model <i>Parameter Settings Models Model Descriptions</i> Assess Model <i>Model Assessment Revised Parameter Settings</i>	Evaluate Results <i>Assessment of Data Mining Results w.r.t. Business Success Criteria Approved Models</i> Review Process <i>Review of Process</i> Determine Next Steps <i>List of Possible Actions Decision</i>	Plan Deployment <i>Deployment Plan</i> Plan Monitoring and Maintenance <i>Monitoring and Maintenance Plan</i> Produce Final Report <i>Final Report Final Presentation</i> Review Project <i>Experience Documentation</i>

Figure 37: Overview of the tasks that belong to each phase of the CRISP-DM model (Wirth, 2000).

F Distribution of contracts for the connected trucks

The figures below show the contracts that are closed with the trucks.

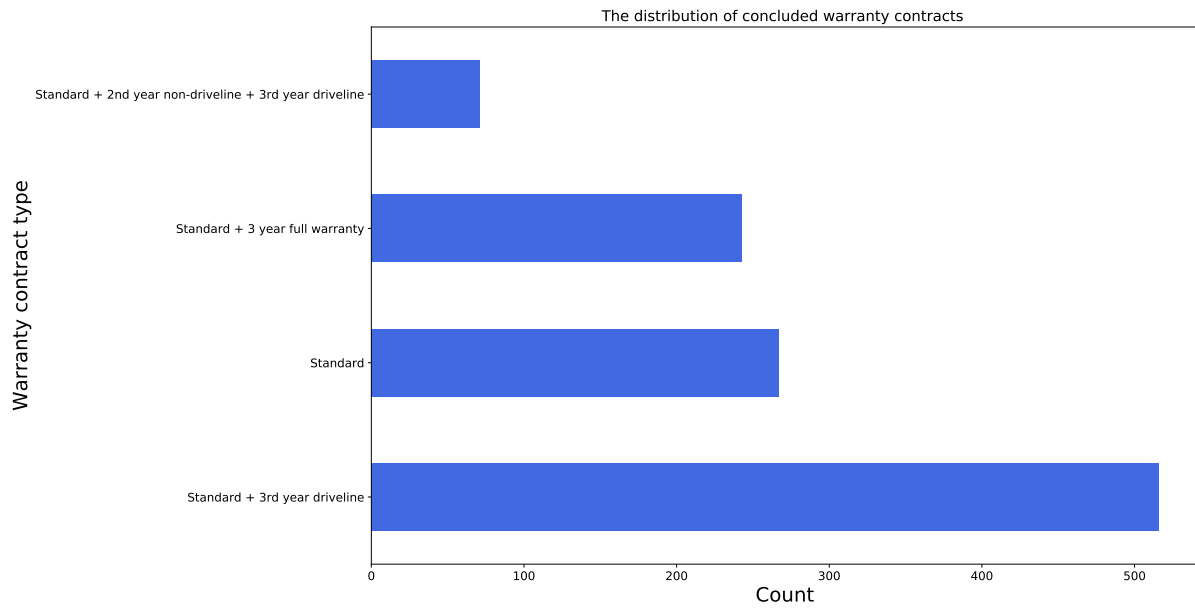


Figure 38: Distribution of the type of warranty contracts that are sold with the connect trucks.

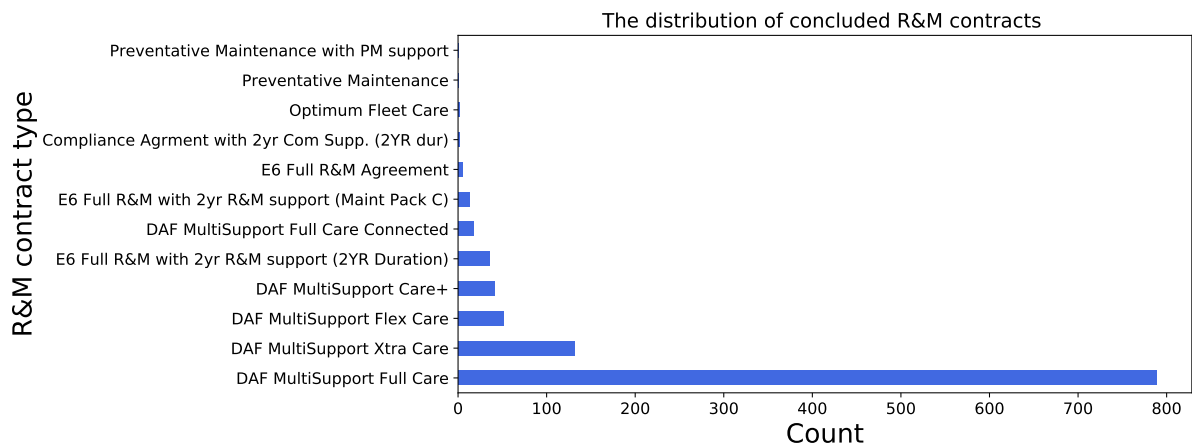


Figure 39: Distribution of the type of R&M contracts that are sold with the connect trucks.

G Fleet information

DAF allows its customers to choose from a range of different engines. For each model, a range of engine types is available. The MX-13 and MX-11 engines are by far the most common as they are the engines of choice for the DAF XF and CF. The remaining type of engines are installed into the smaller DAF LF, which is the least popular model of the three. An overview is given in Figure 40.

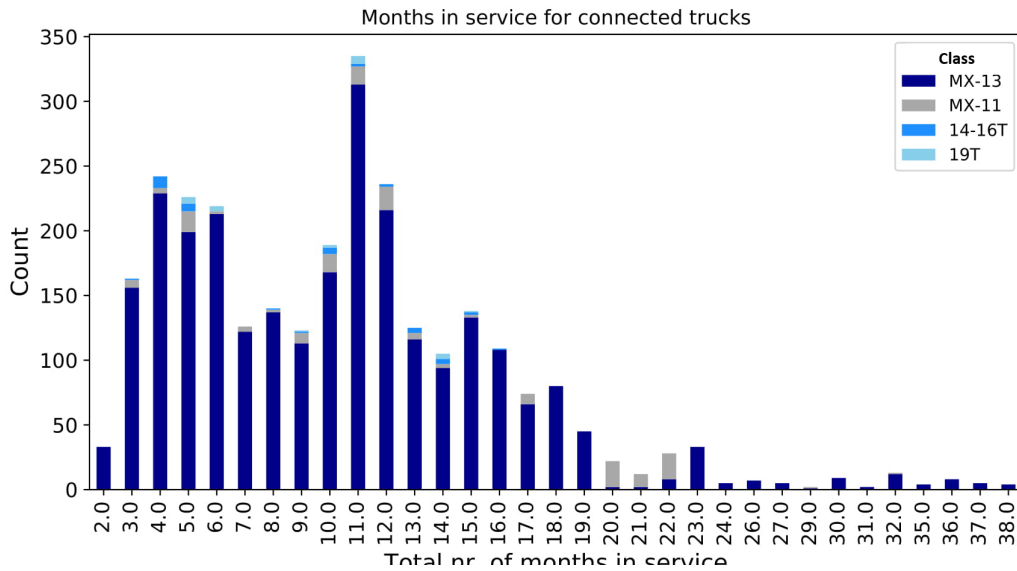


Figure 40: The number of months that connect trucks are in service divided per truck class.

The number of trucks that have DAF connect included is limited. This is due to the fact that the system is not included by default. Since 2017, of the roughly 7000 trucks that were produced per month (240 per day), 5.5% had DAF connected installed. This amounts to 385 trucks per month on average. An overview of the produced connected trucks per month is given in Figure 41.

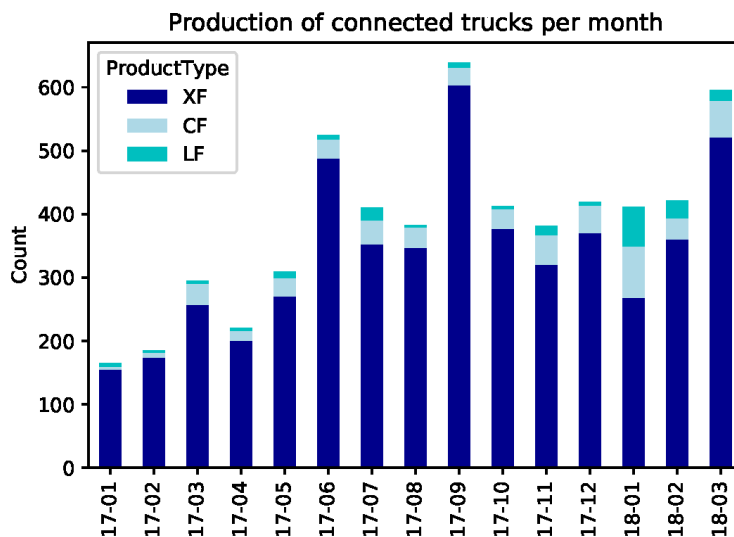


Figure 41: The production of connected trucks per month, divided over product type (model).

H Overview of box-plots for the Connect Data.

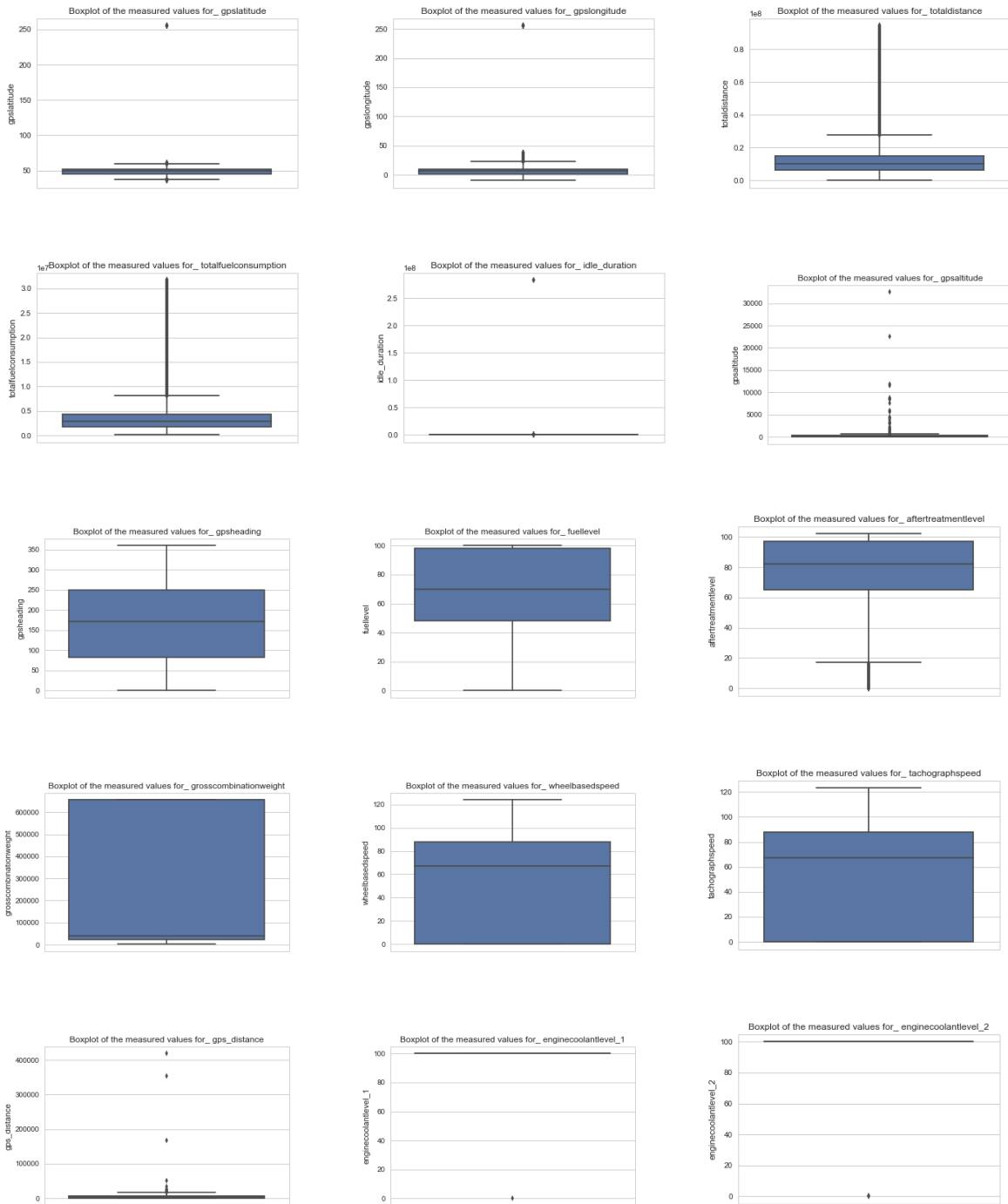


Figure 42: Snapshot data histograms (1)

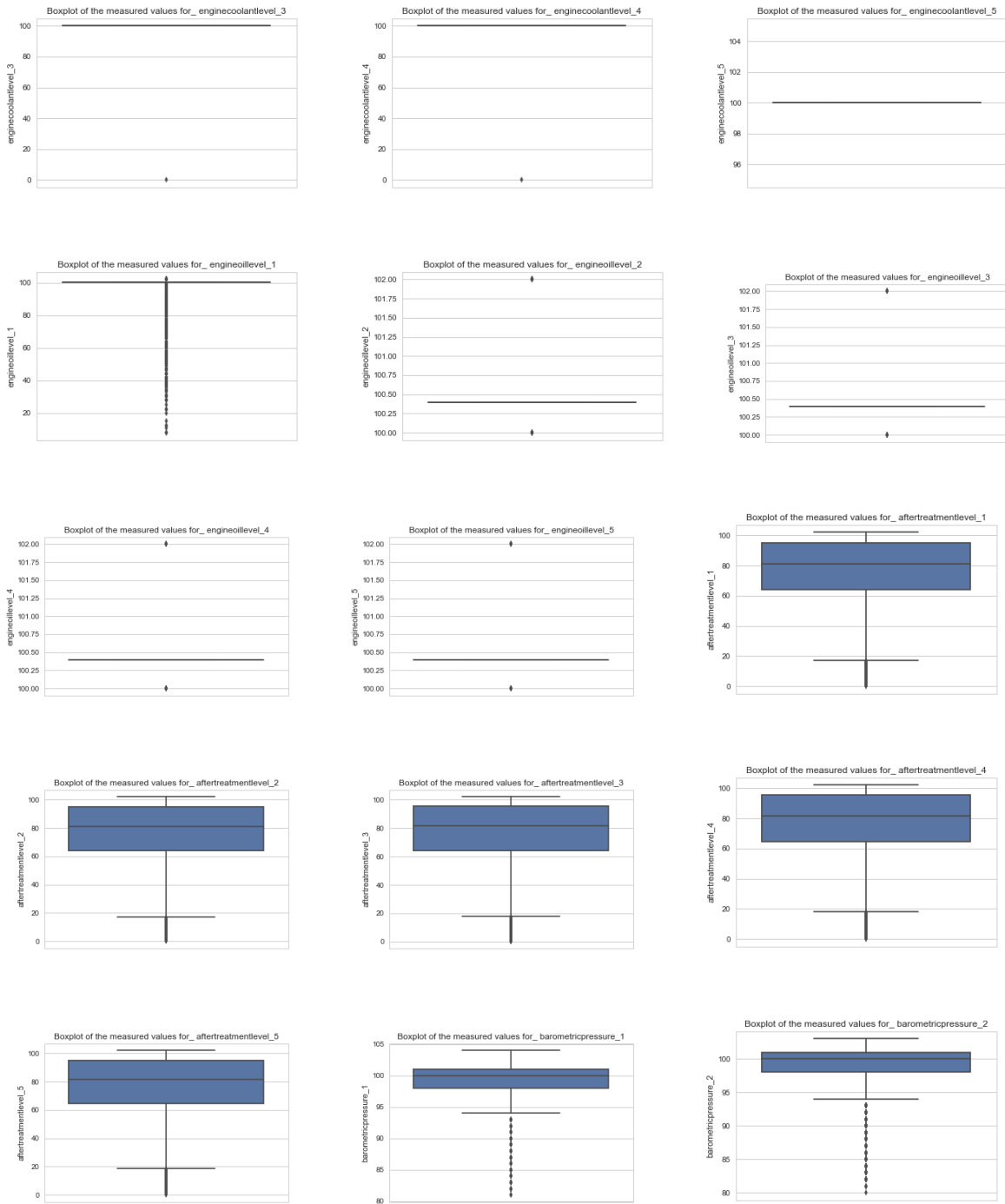


Figure 43: Snapshot data histograms (2)

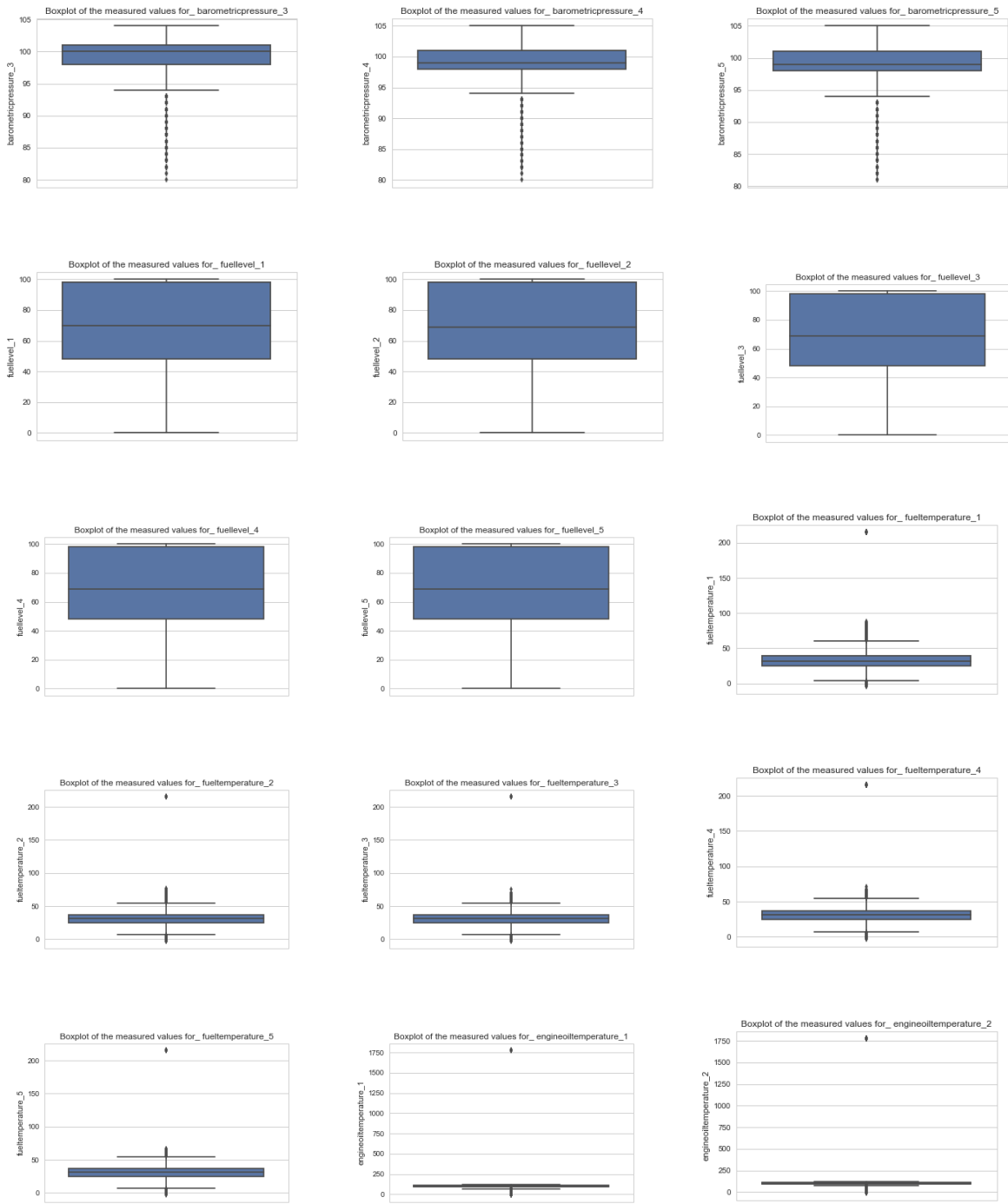


Figure 44: Snapshot data histograms (3)

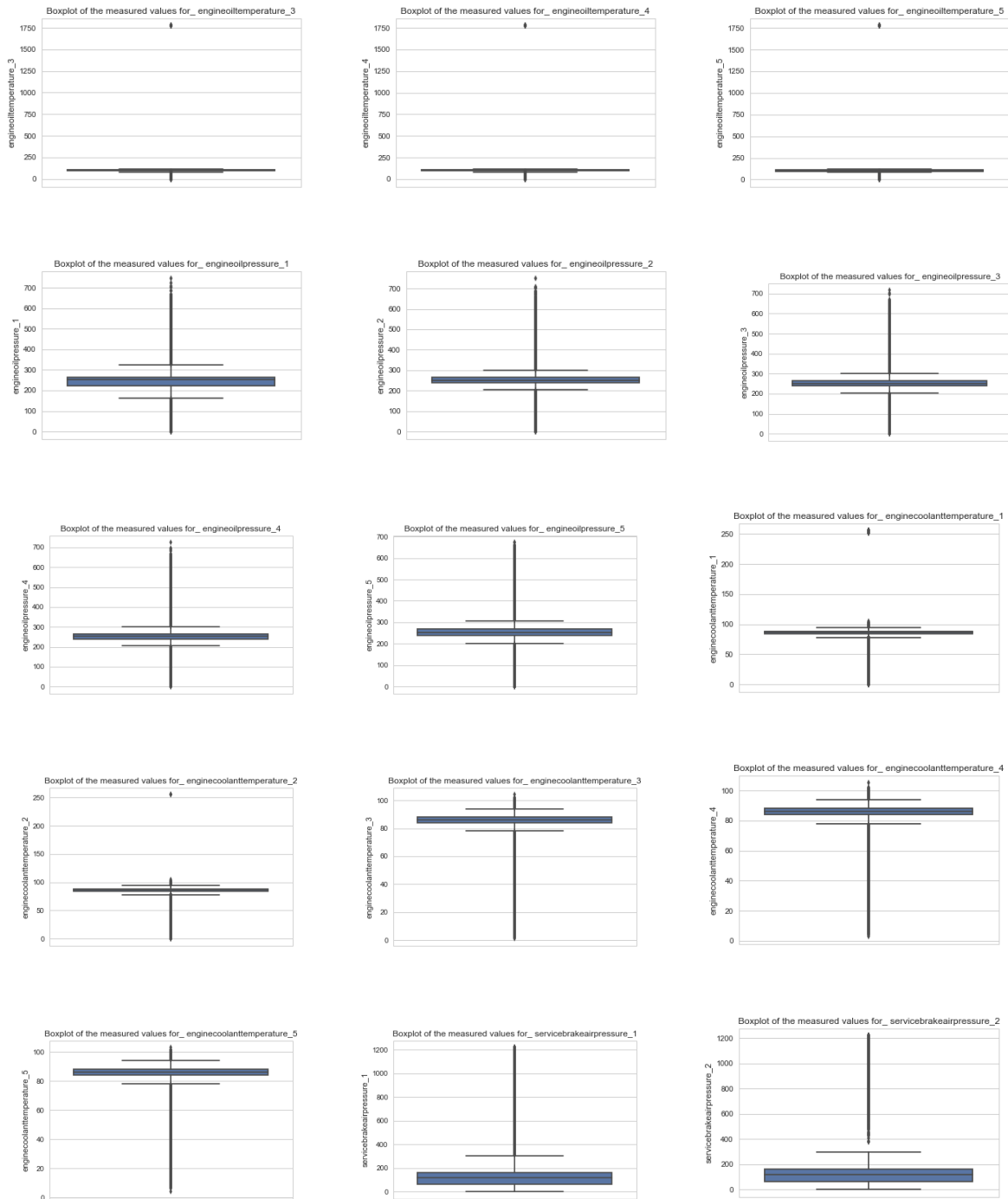


Figure 45: Snapshot data histograms (4)

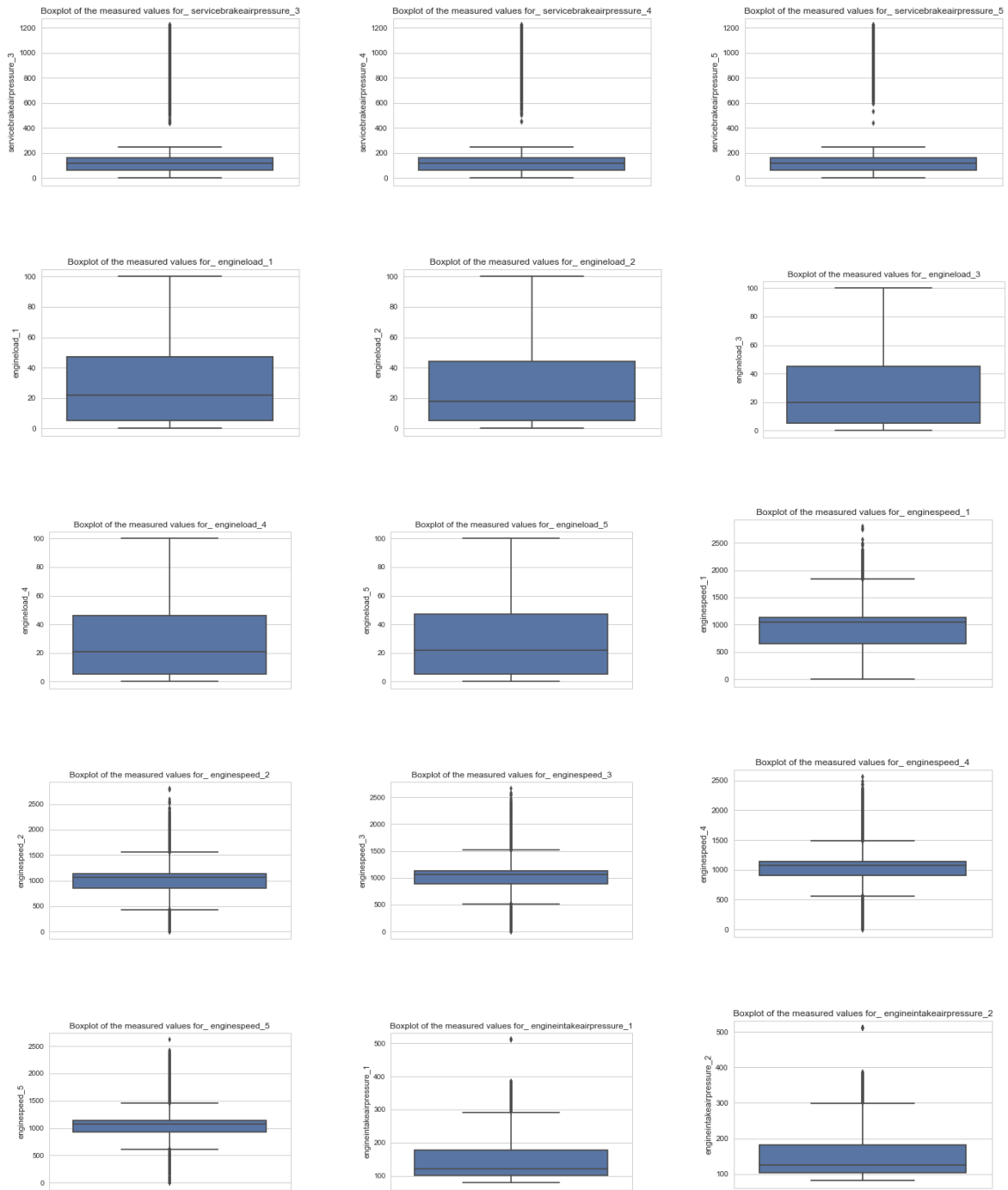


Figure 46: Snapshot data histograms (5)

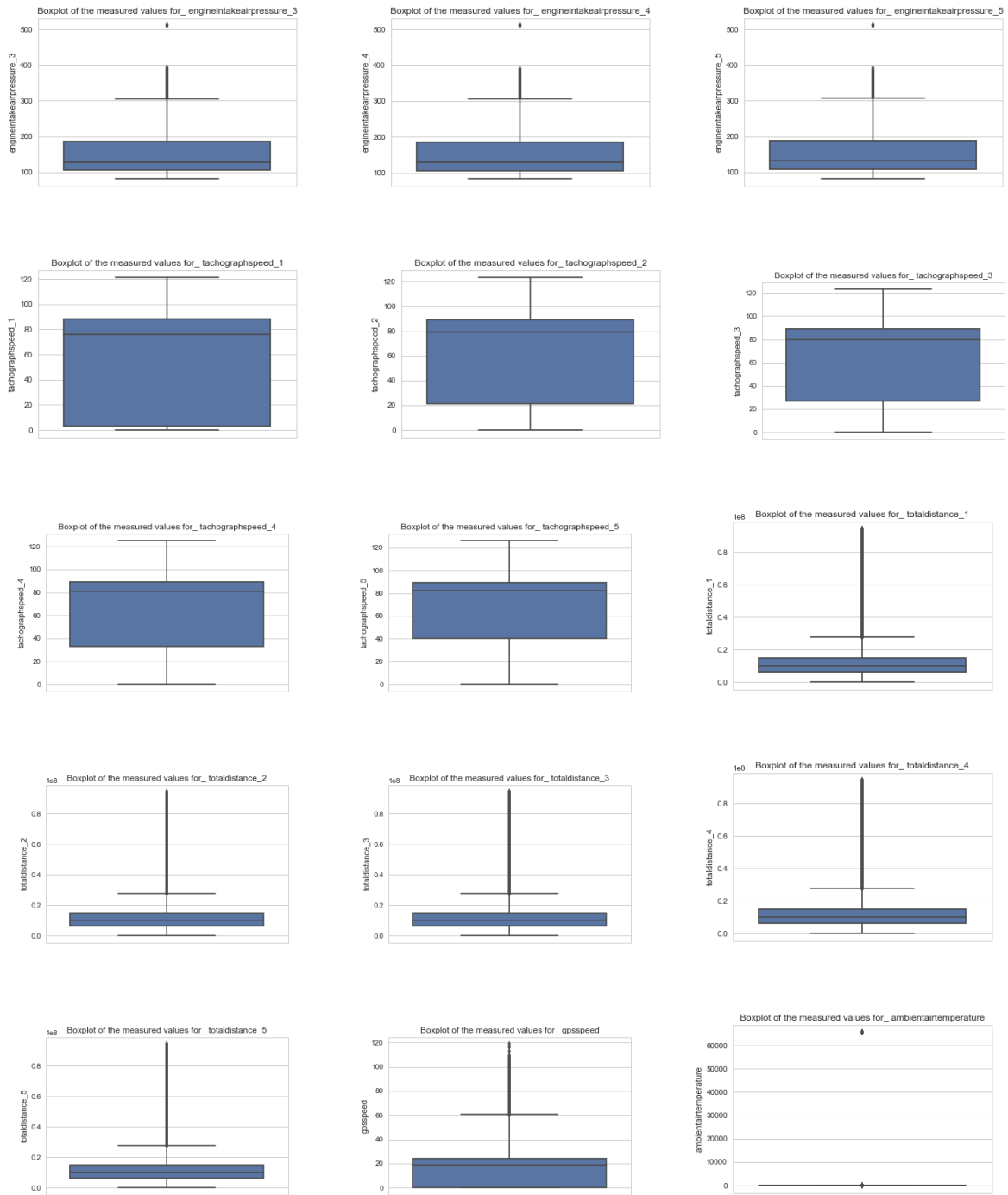


Figure 47: Snapshot data histograms (6)

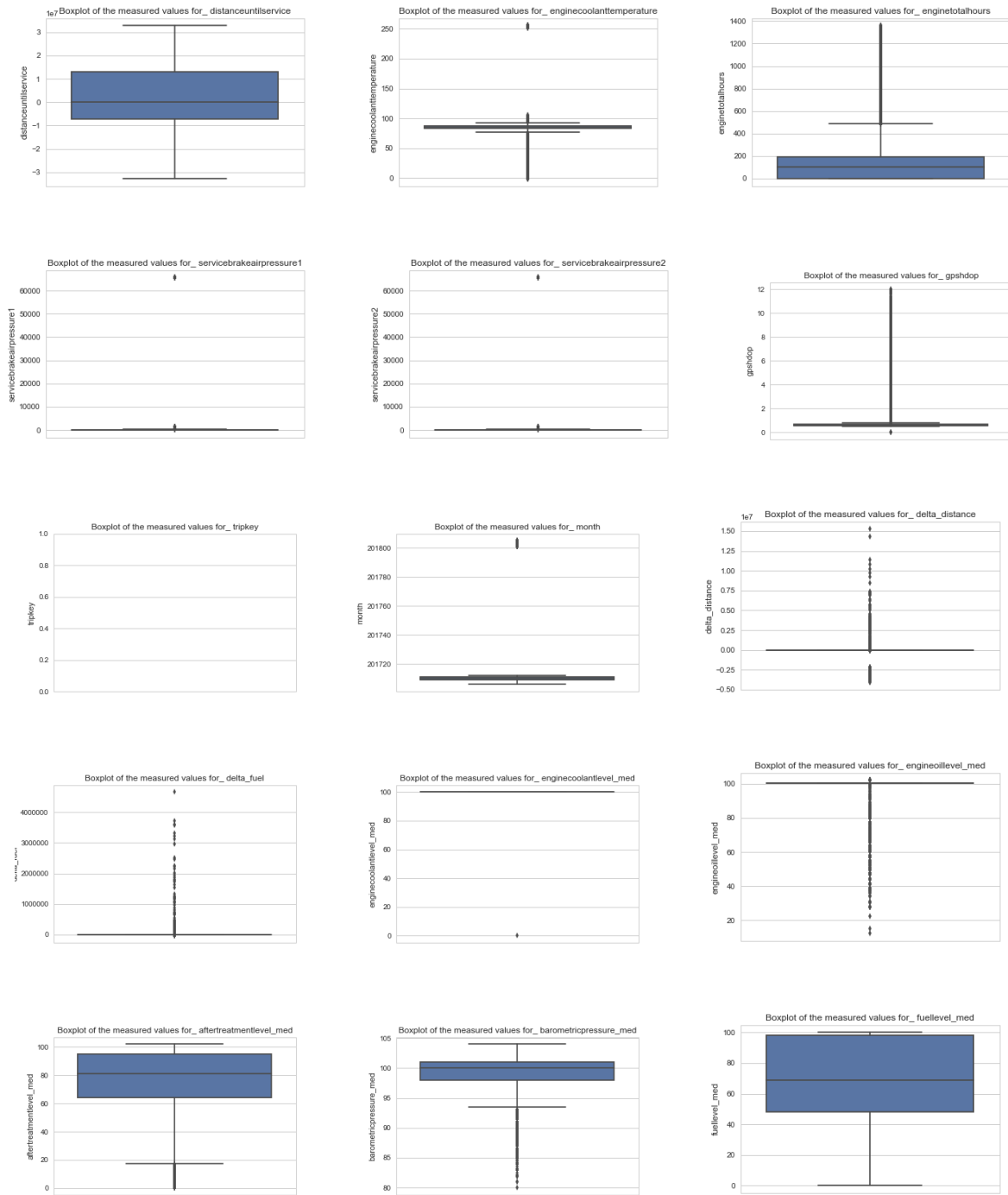


Figure 48: Snapshot data histograms (7)

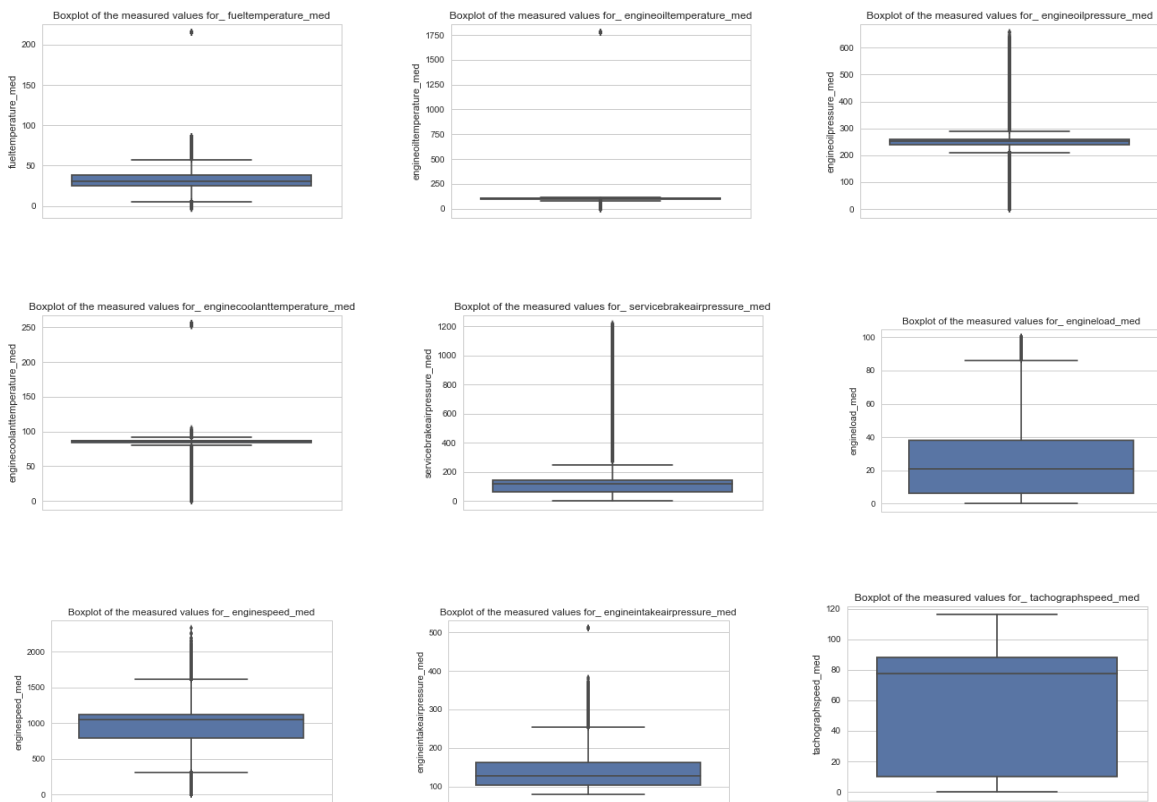


Figure 49: Snapshot data histograms (8)

H.1 Box-plots of the trip database features.

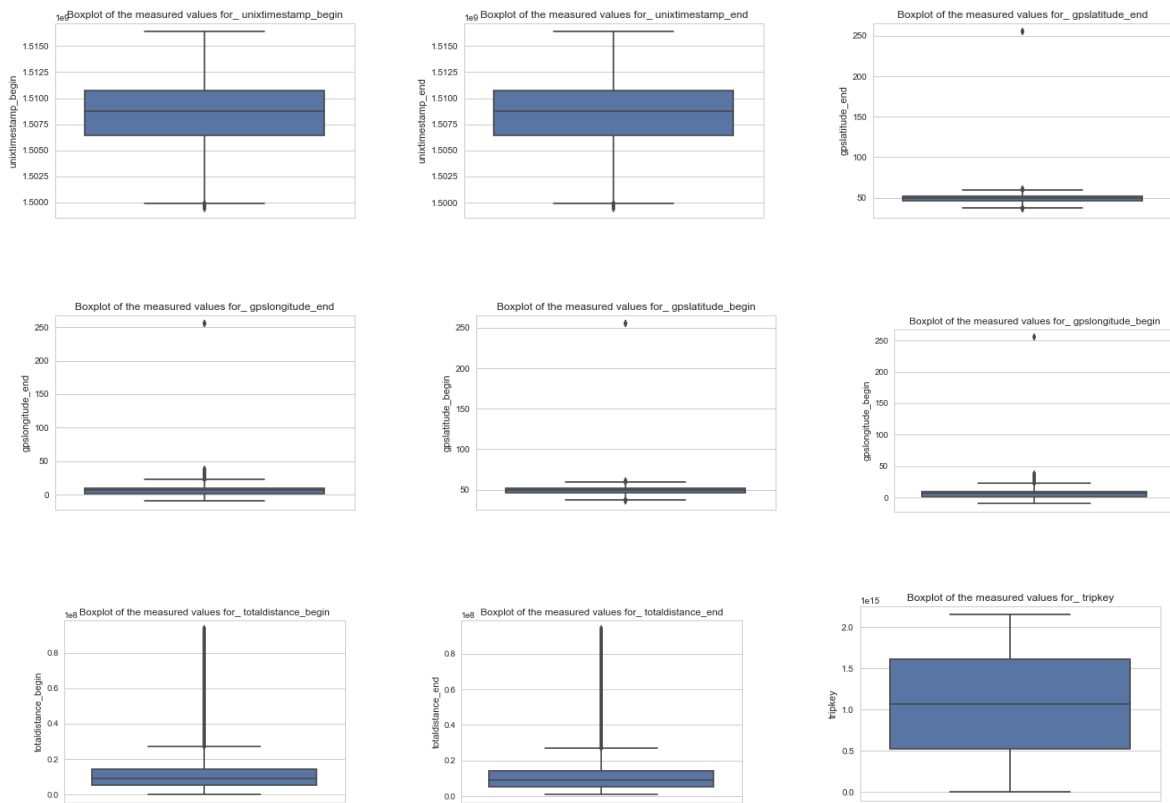


Figure 50: Trip data histograms (1)

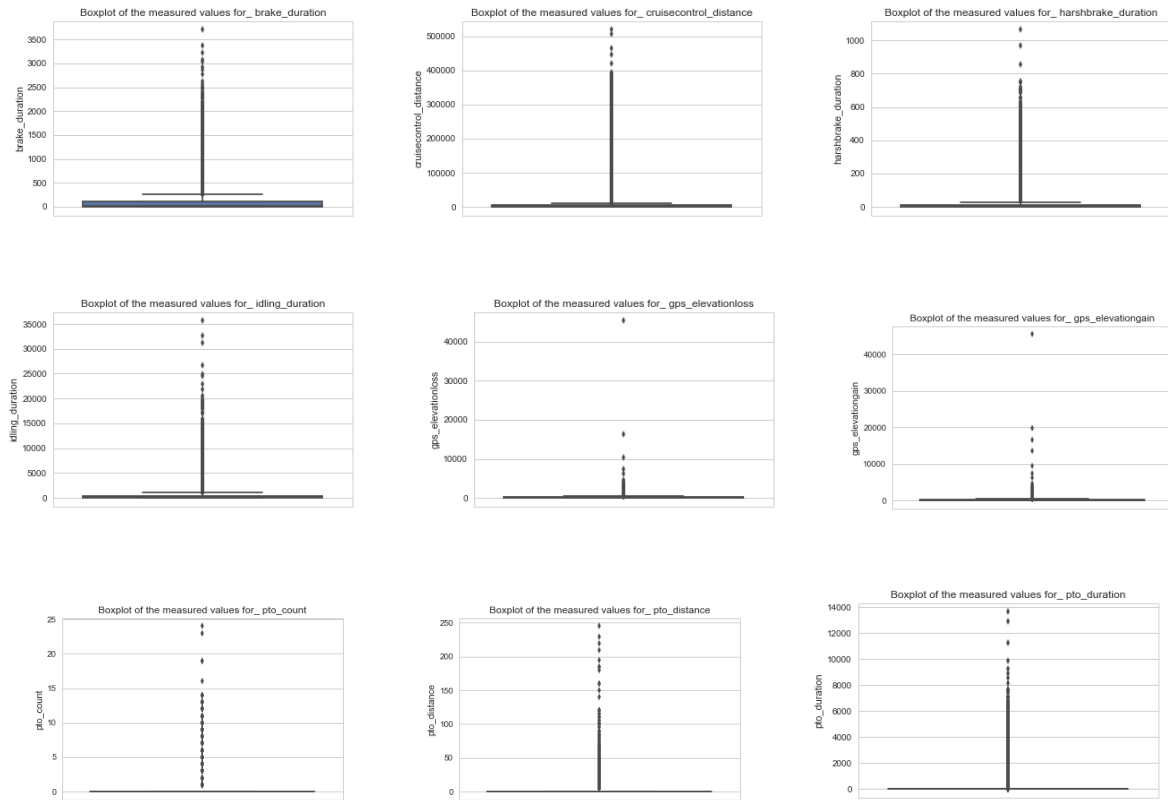


Figure 51: Trip data histograms (2)

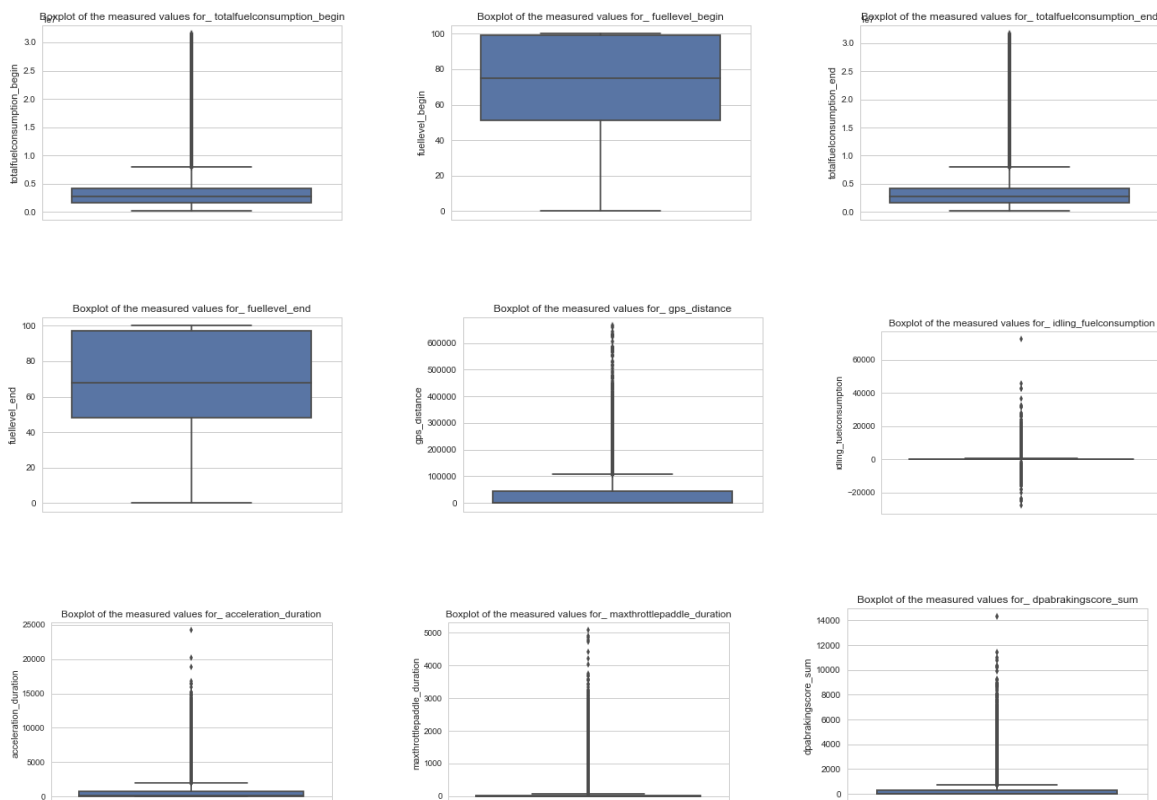


Figure 52: Trip data histograms (3)

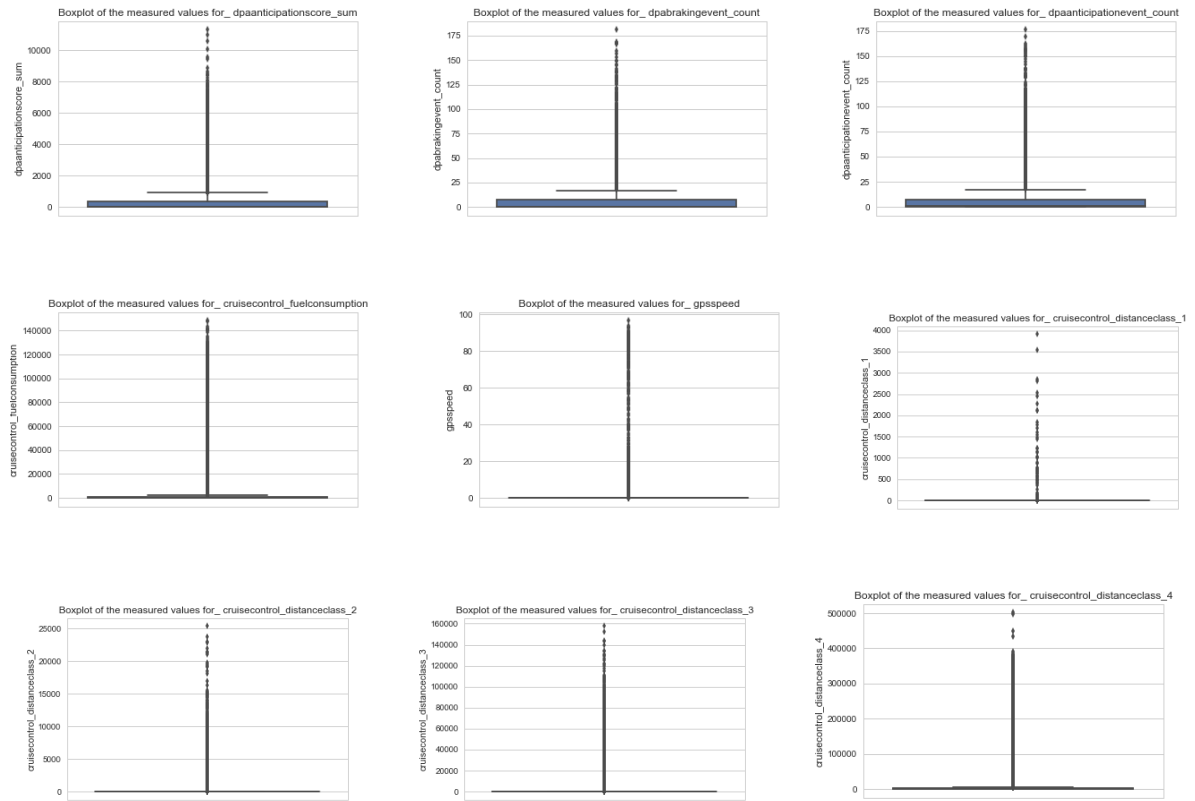


Figure 53: Trip data histograms (4)

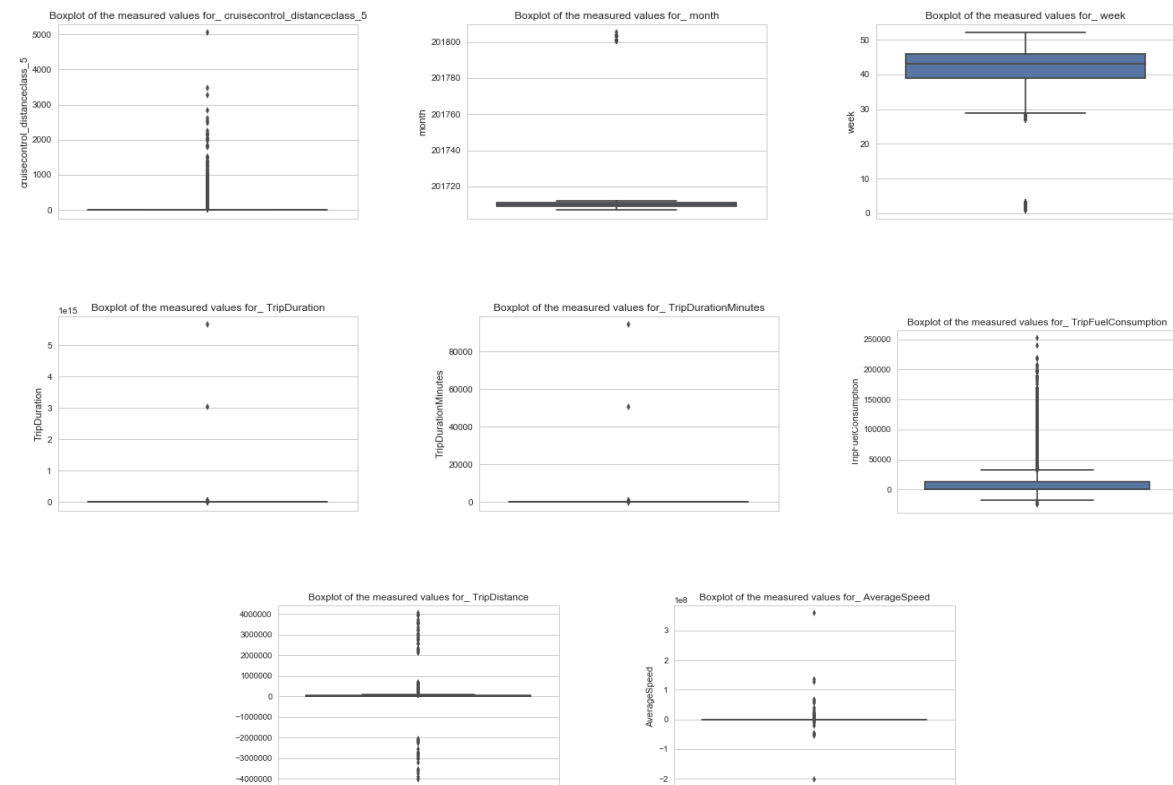


Figure 54: Trip data histograms (5)

I The 1.5 IQR for the Connect Data

The upper and lower boundaries as established using the 1.5 IQR rule are given below. When the lower bound of the IQR is lower than zero it is replaced by zero as none of the measurements can become negative by default. Note that these boundaries were only used when they could not be established by expert knowledge, analysis of the outlier visualization or known physical boundaries.

Table 43: The 1.5 IQR boundaries for the snapshot data.

Feature	Lower_bound	Upper_bound
gpslongitude	0.0	23.05
totaldistance	0.0	27720146.25
totalfuelconsumption	0.0	8207048.75
idle_duration	0.0	40.0
gpsaltitude	0.0	646.5
gpsheading	0.0	502.31
fuellevel	0.0	173.0
grosscombinationweight	0.0	1602975.0
wheelbasedspeed	0.0	220.0
tachographspeed	0.0	220.0
gps_distance	0.0	17526.5
fuellevel_1	0.0	173.0
fuellevel_2	0.0	173.0
fuellevel_3	0.0	173.0
fuellevel_4	0.0	173.0
fuellevel_5	0.0	173.0
servicebrakeairpressure_1	0.0	304.0
servicebrakeairpressure_2	0.0	304.0
servicebrakeairpressure_3	0.0	304.0
servicebrakeairpressure_4	0.0	304.0
servicebrakeairpressure_5	0.0	304.0
engineload_1	0.0	110.0
engineload_2	0.0	102.5
engineload_3	0.0	105.0
engineload_4	0.0	107.5
engineload_5	0.0	110.0
enginespeed_1	0.0	1839.0
engineintakeairpressure_1	0.0	292.0
engineintakeairpressure_2	0.0	299.0
engineintakeairpressure_3	0.0	306.0

engineintakeairpressure_4	0.0	306.0
engineintakeairpressure_5	0.0	308.0
tachographspeed_1	0.0	215.5
tachographspeed_2	0.0	191.0
tachographspeed_3	0.0	182.0
tachographspeed_4	0.0	173.0
tachographspeed_5	0.0	162.5
totaldistance_1	0.0	27773182.5
totaldistance_2	0.0	27820067.5
totaldistance_3	0.0	27839368.75
totaldistance_4	0.0	27860200.0
totaldistance_5	0.0	27894052.5
gpspeed	0.0	60.64
ambientairtemperature	0.0	29.5
distanceuntilservice	0.0	43815000.0
enginetotalhours	0.0	486.5
servicebrakeairpressure1	0.0	364.0
servicebrakeairpressure2	0.0	364.0
delta_distance	0.0	17740.0
delta_fuel	0.0	4334.0
fuellevel_med	0.0	173.0
servicebrakeairpressure_med	0.0	264.0
engineload_med	0.0	86.0
tachographspeed_med	0.0	205.0
gpshdop	0.4352	0.80
eventid	1.0	1.0
fueltemperature_1	4.0	60.0
fueltemperature_med	5.5	57.5
fueltemperature_2	7.0	55.0
fueltemperature_3	7.0	55.0
fueltemperature_4	7.0	55.0
fueltemperature_5	7.0	55.0
engineintakeairpressure_med	14.0	254.0
aftertreatmentlevel	17.0	145.0
aftertreatmentlevel_1	17.19	142.0
aftertreatmentlevel_2	17.19	142.0
aftertreatmentlevel_med	17.19	142.0
aftertreatmentlevel_3	17.60	142.39

aftertreatmentlevel_4	17.60	142.39
aftertreatmentlevel_5	18.20	141.4
gpslatitude	36.96	59.75
engineoiltemperature_1	70.5	130.5
engineoiltemperature_med	75.5	127.5
enginecoolanttemperature	77.0	93.0
engineoiltemperature_2	78.0	126.0
enginecoolanttemperature_1	78.0	94.0
enginecoolanttemperature_2	78.0	94.0
enginecoolanttemperature_3	78.0	94.0
enginecoolanttemperature_4	78.0	94.0
enginecoolanttemperature_5	78.0	94.0
enginecoolanttemperature_med	79.5	91.5
engineoiltemperature_3	80.5	124.5
engineoiltemperature_5	81.5	125.5
engineoiltemperature_4	83.0	123.0
barometricpressure_1	93.5	105.5
barometricpressure_2	93.5	105.5
barometricpressure_3	93.5	105.5
barometricpressure_4	93.5	105.5
barometricpressure_5	93.5	105.5
barometricpressure_med	93.5	105.5
enginecoolantlevel_1	100.0	100.0
enginecoolantlevel_2	100.0	100.0
enginecoolantlevel_3	100.0	100.0
enginecoolantlevel_4	100.0	100.0
enginecoolantlevel_5	100.0	100.0
enginecoolantlevel_med	100.0	100.0
engineoillevel_1	100.4	100.4
engineoillevel_2	100.4	100.4
engineoillevel_3	100.4	100.4
engineoillevel_4	100.4	100.4
engineoillevel_5	100.4	100.4
engineoillevel_med	100.4	100.4
engineoilpressure_1	164.0	324.0
engineoilpressure_5	198.0	310.0
engineoilpressure_2	204.0	300.0
engineoilpressure_3	204.0	300.0

engineoilpressure_4	204.0	300.0
engineoilpressure_med	210.0	290.0
enginespeed_med	303.75	1609.75
enginespeed_2	425.0	1561.0
enginespeed_3	508.0	1516.0
enginespeed_4	562.5	1486.5
enginespeed_5	615.5	1459.5
month	201706.0	201714.0
gpslatitude_end	37.77	59.45
gpslongitude_end	0.0	22.42
gpslatitude_begin	37.75	59.47
gpslongitude_begin	0.0	22.44
totaldistance_begin	0.0	26978247.5
totaldistance_end	0.0	27027932.5
brake_duration	0.0	269.5
cruisecontrol_distance	0.0	10787.5
harshbrake_duration	0.0	30.0
idling_duration	0.0	1056.0
gps_elevationloss	0.0	376.0
gps_elevationgain	0.0	453.5
pto_count	0.0	0.0
pto_distance	0.0	0.0
pto_duration	0.0	0.0
totalfuelconsumption_begin	0.0	7973692.25
fuellevel_begin	0.0	171.0
totalfuelconsumption_end	0.0	7996560.0
fuellevel_end	0.0	170.5
gps_distance	0.0	108274.5
idling_fuelconsumption	0.0	667.0
acceleration_duration	0.0	1966.0
maxthrottlepaddle_duration	0.0	72.5
dpabrakingscore_sum	0.0	735.0
dpaanticipationscore_sum	0.0	932.5
dpabrakingevent_count	0.0	17.5
dpaanticipationevent_count	0.0	17.5
cruisecontrol_fuelconsumption	0.0	2487.5
gpsspeed	0.0	0.07
cruisecontrol_distanceclass_1	0.0	0.0

cruisecontrol_distanceclass_2	0.0	0.0
cruisecontrol_distanceclass_3	0.0	0.0
cruisecontrol_distanceclass_4	0.0	5200.0
cruisecontrol_distanceclass_5	0.0	0.0
month	201706.0	201714.0
week	28.5	56.5
TripDurationMinutes	0.0	131.53
TripFuelConsumption	0.0	32904.0
TripDistance	0.0	111722.5
AverageSpeed	0.0	130147.13

J Fuzzy bins experiment results

To test the effect of different sizes of bins on the modeling performance, a small experiment using fixed model parameters and data splits has been run for the different modeling methods that have been used. The cross-validated validation scores are reported in Table 44. There can be seen that the modeling performance in general does not increase (or decrease) with an increased number of bins. Therefore, there has been chosen to work with fuzzy histograms with 5 bins during the modeling and evaluation phase of the project.

Table 44: The cross-validated experimental model performance using different sizes of bins

Nr. of bins	Decision Tree	Random Forest	MLP Neural Network	Logistic Regression
5	0.65 (+- 0.05)	0.65 (+- 0.07)	0.60 (+- 0.02)	0.55 (+- 0.07)
7	0.59 (+-0.07)	0.60 (+- 0.08)	0.60 (+- 0.01)	0.54 (+- 0.08)
9	0.59 (+- 0.12)	0.60 (+- 0.08)	0.60 (+- 0.01)	0.54 (+-0.13)

K Similarly performing Logistic Regression model beta coefficients

in this appendix, the beta coefficients for one of the found Logistic Regression trees during the modeling phase are given. It had a slightly worse accuracy (68%) than the best performing model. But as it incorporated some additional features, it can be used to extract some useful information about the beta coefficients and their effect on the expected number of repairs.

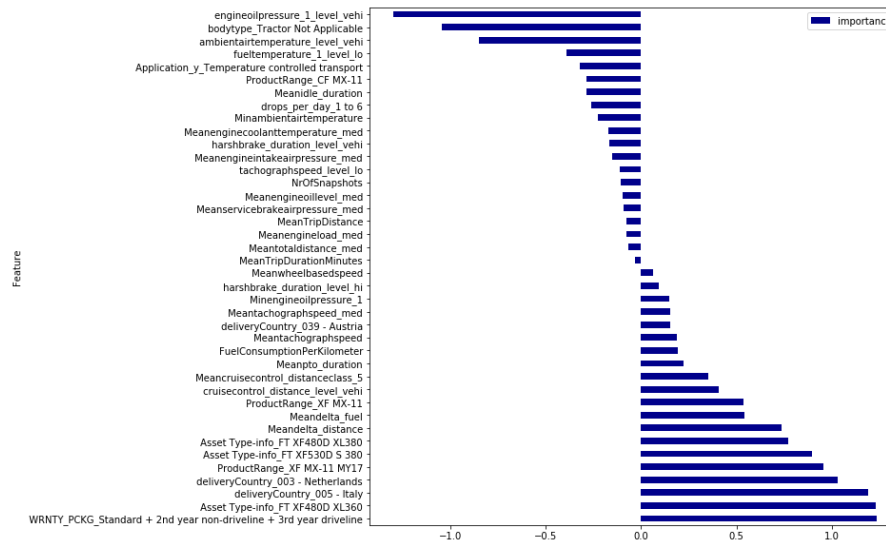


Figure 55: The beta coefficients for one of the found Logistic Regression models.

L Average accuracy over 10 different modeling runs

In this section, the average performance per model is given when run over 10 different splits of the initial data, in order to provide an insight in the modeling instability that is present.

Table 45: The average accuracy and it's standard deviation for 10 runs of the Logistic Regression modeling process.

Model	Mean	STD
Base LR 11 months	0.56	0.01
Extended LR 11 months	0.57	0.01
Base LR 8 months	0.55	0.00
Extended LR 8 months	0.56	0.00
Base LR 8 months SFS	0.55	0.01
Extended LR 8 months SFS	0.57	0.00
Base LR 11 months SFS	0.66	0.01
Extended LR 11 months SFS	0.69	0.01

Table 46: The average accuracy and it's standard deviation for 10 runs of the Decision Tree modeling process.

Model	Mean	STD
Base DT 11 months	0.57	0.05
Extended DT 11 months	0.61	0.03
Base DT 8 months	0.52	0.03
Extended DT 8 months	0.54	0.03
Base DT 8 months SFS	0.53	0.01
Extended DT 8 months SFS	0.54	0.03
Base DT 11 months SFS	0.58	0.04
Extended DT 11 months SFS	0.64	0.02

Table 47: The average accuracy and it's standard deviation for 10 runs of the Random Forest modeling process.

Model	Mean	STD
Base RF 11 months	0.63	0.02
Extended RF 11 months	0.62	0.02
Base RF 8 months	0.59	0.02
Extended RF 8 months	0.61	0.02
Base RF 8 months SFS	0.62	0.01
Extended RF 8 months SFS	0.61	0.01
Base RF 11 months SFS	0.62	0.02
Extended RF 11 months SFS	0.65	0.02

Table 48: The average accuracy and it's standard deviation for 10 runs of the MLP-NN modeling process.

Model	Mean	STD
Base NN 11 months	0.57	0.03
Extended NN 11 months	0.61	0.03
Base NN 8 months	0.55	0.06
Extended NN 8 months	0.58	0.02
Base NN 8 months SFS	0.55	0.04
Extended NN 8 months SFS	0.58	0.02
Base NN 11 months SFS	0.58	0.03
Extended NN 11 months SFS	0.61	0.02

M T-test on model improvements using Fuzzy Bins

To verify that the fuzzy bins indeed did increase the accuracy performance of the Logistic Regression and Decision Tree models, a one-sided t-test has been performed over 10 different runs and data-splits. The results are given below. First an F-test on equal variance has been performed after which subsequently the one sided T-test to test for a significant increase in accuracy has been performed. The Extended models are the models including features derived with fuzzy bins while the base models are the models without these features. In the tables, base models are indicated with a B while the Extended models are indicated with an E.

Table 49: Results on one sided t-test for increase of accuracy for the Logistic Regression models based on 10 different runs/datasplits

Model	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Mean	STD	F test	T-test	P-Value
B DT 11 months	0.55	0.64	0.62	0.58	0.52	0.56	0.64	0.54	0.5	0.57	0.572	0.05	Equal		
E DT 11 months	0.61	0.56	0.66	0.57	0.6	0.58	0.61	0.64	0.63	0.63	0.609	0.03		True	0.03
B DT 8 months	0.54	0.53	0.48	0.52	0.54	0.54	0.54	0.54	0.50	0.47	0.52	0.03	Equal		
E DT 8 months	0.58	0.58	0.51	0.54	0.51	0.51	0.52	0.58	0.52	0.57	0.54	0.03		False	0.08
B DT 8 months SFS	0.52	0.53	0.55	0.55	0.52	0.53	0.51	0.55	0.52	0.53	0.531	0.01	Unequal		
E DT 8 months SFS	0.55	0.54	0.56	0.53	0.52	0.56	0.48	0.6	0.49	0.54	0.537	0.03		False	0.41
B DT 11 months SFS	0.55	0.62	0.55	0.62	0.55	0.64	0.61	0.54	0.55	0.59	0.582	0.04	Equal		
E DT 11 months SFS	0.66	0.62	0.66	0.65	0.6	0.65	0.6	0.65	0.65	0.66	0.64	0.02		True	0

Table 50: Results on one sided t-test for increase of accuracy for the Decision Tree models based on 10 different runs/datasplits

Models:	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Mean	STD	F-Test	T-Test	P-Value
B LR 11 months	0.56	0.56	0.56	0.57	0.55	0.55	0.55	0.55	0.56	0.55	0.56	0.01	Equal		
E LR 11 months	0.57	0.57	0.57	0.57	0.58	0.58	0.57	0.57	0.58	0.58	0.57	0.01	Equal	True	0
B LR 8 months	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0	Equal		
E LR 8 months	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0	Equal	True	0
B LR 8 months SFS	0.55	0.56	0.55	0.56	0.55	0.56	0.56	0.55	0.55	0.55	0.55	0.01	Equal		
E LR 8 months SFS	0.57	0.57	0.56	0.57	0.57	0.56	0.57	0.57	0.57	0.57	0.57	0	Equal	True	0
B LR 11 months SFS	0.66	0.65	0.66	0.66	0.65	0.65	0.66	0.66	0.66	0.65	0.66	0.01	Equal		
E LR 11 months SFS	0.68	0.69	0.68	0.68	0.69	0.69	0.68	0.68	0.69	0.69	0.69	0.01	Equal	True	0

N The derived truck usage features over the first month of operation.

The relevant trip measurements of the trucks have both been averaged per kilometer and summed over the first month of operation, in order to derive general usage features.

Table 51: Overview of the derived truck usage features over the first month of operation.

Average value per driven kilometer	Aggregated (sum of all measurements)
TripFuelConsumption	TripFuelConsumption
dpaanticipationscore_sum	dpaanticipationscore_sum
brake_duration	brake_duration
harshbrake_duration	harshbrake_duration
cruisecontrol_distanceclass_1	cruisecontrol_distanceclass_1
cruisecontrol_distanceclass_2	cruisecontrol_distanceclass_2
cruisecontrol_distanceclass_3	cruisecontrol_distanceclass_3
cruisecontrol_distanceclass_4	cruisecontrol_distanceclass_4
cruisecontrol_distanceclass_5	cruisecontrol_distanceclass_5
cruisecontrol_fuelconsumption	cruisecontrol_fuelconsumption
dpaanticipationevent_count	dpaanticipationevent_count
dpabrakingevent_count	dpabrakingevent_count
dpabrakingscore_sum	dpabrakingscore_sum
maxthrottlepaddle_duration	maxthrottlepaddle_duration
acceleration_duration	acceleration_duration
idling_fuelconsumption	idling_fuelconsumption
idling_duration	idling_duration
TripDistance	TripDistance
TripDurationMinutes	TripDurationMinutes
TripFuelConsumption	TripFuelConsumption
pto_distance	pto_distance
pto_count	pto_count
pto_duration	pto_duration
gps_elevationloss	gps_elevationloss
gps_elevationgain	gps_elevationgain
cruisecontrol_distance	cruisecontrol_distance
gpslatitude_begin	gpslatitude_begin
gpslongitude_begin	gpslongitude_begin

O The features over which fuzzy bins have been used for feature extraction.

Table 52: The features over which fuzzy bins have been used for feature extraction.

Feature	Explanation
Fuel temperature	Temperature of the fuel inside the truck
Ambient air temperature	Air temperature outside of the truck
Engine intake air pressure	Air pressure at the engine intake
engine oil pressure	Oil pressure inside the engine
Tachographspeed	Speed of the truck in kmph
Harsh brake duration	Harsh braking duration per trip in seconds
Acceleration duration	Acceleration duration per trip in seconds
cruise control distance	Cruise control distance driven per trip
gps elevation gain	Elevation gain during a trip in meters
dpa-anticipation score	Calculated score for the anticipation skills of the driver during the trip (higher score equals better anticipation).

P An example of an instance in the final datasets.

To provide the reader with an overview of all the features that have been used for modeling, an example instance is given here. It contains all of the 324 derived features and their corresponding value for the truck under consideration.

Table 53: Example of an instance of the final datasets.

Feature	Value
NrOfSnapshots	1960
Meandelta_distance	4981.607143
Meandelta_fuel	1258.666327
Meanidle_duration	32.19897959
Meangrosscombinationweight	28566.73469
Meanambientairtemperature	15.36326531
Meangpsaltitude	30.78622449
Meangpslatitude	51.09156005
Meangpslongitude	4.506652781
Stddelta_distance	2102.417085
Stddelta_fuel	499.8493657
Stdidle_duration	117.5326319
Stdgrosscombinationweight	13599.80315
Stdambientairtemperature	3.420306526
Stdgpsaltitude	43.69880102
Stdgpslatitude	0.398866996
Stdgpslongitude	0.499523447
Skewdelta_distance	-0.64456125
Skewdelta_fuel	0.169818936
Skewidle_duration	7.306793124
Skewgrosscombinationweight	0.58745172
Skewambientairtemperature	0.151904669
Skewgpsaltitude	3.877692727
Skewgpslatitude	-3.318319163
Skewgpslongitude	-0.019503855
Meanenginecoolantlevel_med	100
Meanengineoillevel_med	100.4
Meanaftertreatmentlevel_med	78.75316327
Meanbarometricpressure_med	100.9043367
Meanfuellevel_med	70.92729592
Meanfueltemperature_med	31.21479592

Meanengineoiltemperature_med	98.26403061
Meanengineoilpressure_med	306.4377551
Meanenginecoolanttemperature_med	86.30280612
Meanservicebrakeairpressure_med	83.46530612
Meanengineload_med	25.59285714
Meanenginespeed_med	1063.029847
Meanengineintakeairpressure_med	142.9464286
Meantachographspeed_med	60.55127551
Stdenginecoolantlevel_med	0
Stdengineoillevel_med	0
Stdaftertreatmentlevel_med	18.58228037
Stdbarometricpressure_med	0.993165295
Stdfuellevel_med	26.82836748
Stdfueltemperature_med	5.607598529
Stdengineoiltemperature_med	10.31325622
Stdengineoilpressure_med	39.43783373
Stdenginecoolanttemperature_med	5.56623006
Stdservicebrakeairpressure_med	52.95606566
Stdengineload_med	15.62897076
Stdenginespeed_med	179.8935537
Stdengineintakeairpressure_med	31.11065872
Stdtachographspeed_med	27.36614923
Skewenginecoolantlevel_med	0
Skewengineoillevel_med	0
Skewaftertreatmentlevel_med	-0.658027484
Skewbarometricpressure_med	-0.581256587
Skewfuellevel_med	-0.452849222
Skewfueltemperature_med	0.153947941
Skewengineoiltemperature_med	-3.572221887
Skewengineoilpressure_med	-1.5873581
Skewenginecoolanttemperature_med	-7.233155903
Skewservicebrakeairpressure_med	1.365496146
Skewengineload_med	0.522099409
Skewenginespeed_med	-1.54314209
Skewengineintakeairpressure_med	1.755173623
Skewtachographspeed_med	-0.881706624
DistanceTravelled	10055155
Maxidle_duration	1861

Minidle.duration	0
Maxambientairtemperature	25
Minambientairtemperature	8
Maxtotaldistance	22688255
Mintotaldistance	12633100
Maxtotalfuelconsumption	6169847
Mintotalfuelconsumption	3504778
Maxwheelbasedspeed	94
Minwheelbasedspeed	0
Maxgrosscombinationweight	61600
Mingrosscombinationweight	6800
Maxgpslongitude	5.8441896
Mingpslongitude	3.0013466
Maxgpslatitude	51.702446
Mingpslatitude	48.7856
Maxenginecoolantlevel	100
Minenginecoolantlevel	100
Maxengineoillevel	100.4
Minengineoillevel	69.2
Maxaftertreatmentlevel	100
Minaftertreatmentlevel	36.4
Maxbarometricpressure	102
Minbarometricpressure	97
Maxfuellevel	100
Minfuellevel	9
Maxfueltemperature	76
Minfueltemperature	14
Maxengineoiltemperature	108
Minengineoiltemperature	15
Maxengineoilpressure	544
Minengineoilpressure	0
Maxenginecoolanttemperature	97
Minenginecoolanttemperature	16
Maxservicebrakeairpressure	248
Minservicebrakeairpressure	0
Maxengineload	100
Minengineload	0
Maxenginespeed	1957

Minenginespeed	147
Maxengineintakeairpressure	334
Minengineintakeairpressure	98
Maxtachographspeed	94
Mintachographspeed	0
MeanTripDurationMinutes	37.96256983
MeanTripDistance	28090.83799
std_TripDistance	48941.88263
skew_TripDistance	2.034432475
MeanTripFuelConsumption	7448.636872
std_TripFuelConsumption	12539.7079
skew_TripFuelConsumption	2.172959231
Meandpaanticipationscore_sum	366.5921788
std_dpaanticipationscore_sum	634.1488803
skew_dpaanticipationscore_sum	2.733583701
Meanbrake_duration	120.4888268
std_brake_duration	201.9308161
skew_brake_duration	3.13727088
Meanharshbrake_duration	15.30726257
std_harshbrake_duration	36.8586596
skew_harshbrake_duration	4.896012667
Meancruisecontrol_distanceclass_1	0.782122905
std_cruisecontrol_distanceclass_1	4.075961362
skew_cruisecontrol_distanceclass_1	5.133734591
Meancruisecontrol_distanceclass_2	58.35195531
std_cruisecontrol_distanceclass_2	324.0205135
skew_cruisecontrol_distanceclass_2	10.32158778
Meancruisecontrol_distanceclass_3	1360.670391
std_cruisecontrol_distanceclass_3	3465.15853
skew_cruisecontrol_distanceclass_3	4.651801326
Meancruisecontrol_distanceclass_4	14933.05866
std_cruisecontrol_distanceclass_4	31059.51845
skew_cruisecontrol_distanceclass_4	2.177355712
Meancruisecontrol_distanceclass_5	0
std_cruisecontrol_distanceclass_5	0
skew_cruisecontrol_distanceclass_5	0
Meancruisecontrol_fuelconsumption	3654.391061
std_cruisecontrol_fuelconsumption	7511.960345

skew_cruisecontrol_fuelconsumption	2.274831871
Meandpaanticipationevent_count	5.983240223
std_dpaanticipationevent_count	10.45302945
skew_dpaanticipationevent_count	2.849309093
Meandpabrakingevent_count	4.824022346
std_dpabrakingevent_count	8.4760188
skew_dpabrakingevent_count	2.783364283
Meandpabrakingscore_sum	319.4581006
std_dpabrakingscore_sum	569.9482921
skew_dpabrakingscore_sum	2.792204072
Meanmaxthrottlepaddle_duration	27.07821229
std_maxthrottlepaddle_duration	93.36284763
skew_maxthrottlepaddle_duration	8.976156083
Meanacceleration_duration	540.2011173
std_acceleration_duration	910.197503
skew_acceleration_duration	3.472129571
Meanidling_fuelconsumption	335.8268156
std_idling_fuelconsumption	476.7854936
skew_idling_fuelconsumption	3.514352277
Meanidling_duration	586.9106145
std_idling_duration	740.679333
skew_idling_duration	2.845569394
std_TripDurationMinutes	51.62018679
skew_TripDurationMinutes	2.037718417
Meanpto_distance	0.377094972
std_pto_distance	2.13290361
skew_pto_distance	5.883998601
Meanpto_count	0.055865922
std_pto_count	0.294119212
skew_pto_count	6.284201674
Meanpto_duration	22.66480447
std_pto_duration	181.6412706
skew_pto_duration	14.43365291
Meangps_elevationloss	53.01117318
std_gps_elevationloss	108.8023155
skew_gps_elevationloss	5.135947049
Meangps_elevationgain	53.5
std_gps_elevationgain	110.2918398

skew_gps_elevationgain	5.419655752
Meancruisecontrol_distance	16352.86313
std_cruisecontrol_distance	32778.98891
skew_cruisecontrol_distance	2.090047768
Meangpslatitude_begin	51.18865017
std_gpslatitude_begin	0.33472457
skew_gpslatitude_begin	-3.603809265
Meangpslongitude_begin	4.543775949
std_gpslongitude_begin	0.605398171
skew_gpslongitude_begin	-0.071885818
SumTripDistance	10056520
SumTripFuelConsumption	2666612
SumTripFuelConsumptionPerKilomter	0.265162502
Sumdpaanticipationscore_sum	131240
Sumdpaanticipationscore_sumPerKilometer	0.01305024
Sumbrake_duration	43135
Sumbrake_durationPerKilometer	0.004289257
Sumharshbrake_duration	5480
Sumharshbrake_durationPerKilometer	0.00054492
Sumcruisecontrol_distanceclass_1	280
Sumcruisecontrol_distanceclass_1PerKilometer	2.78E-05
Sumcruisecontrol_distanceclass_2	20890
Sumcruisecontrol_distanceclass_2PerKilometer	0.002077259
Sumcruisecontrol_distanceclass_3	487120
Sumcruisecontrol_distanceclass_3PerKilometer	0.048438227
Sumcruisecontrol_distanceclass_4	5346035
Sumcruisecontrol_distanceclass_4PerKilometer	0.531598903
Sumcruisecontrol_distanceclass_5	0
Sumcruisecontrol_distanceclass_5PerKilometer	0
Sumcruisecontrol_fuelconsumption	1308272
Sumcruisecontrol_fuelconsumptionPerKilometer	0.13009192
Sumdpaanticipationevent_count	2142
Sumdpaanticipationevent_countPerKilometer	0.000212996
Sumdpabrakingevent_count	1727
Sumdpabrakingevent_countPerKilomter	0.000171729
Sumdpabrakingscore_sum	114366
Sumdpabrakingscore_sumPerKilomter	0.011372324
Summaxthrottlepaddle_duration	9694

Summaxthrottlepaddle_durationPerKilometer	0.000963952
Sumacceleration_duration	193392
Sumacceleration_durationPerKilomter	0.019230509
Sumidling_fuelconsumption	120226
Sumidling_fuelconsumptionPerKilometer	0.01195503
Sumidling_duration	210114
Sumidling_durationPerKilometer	0.020893311
SumTripDistancePerKilometer	1
SumTripDurationMinutes	13590.6
SumTripDurationMinutesPerKilometer	0.001351422
Sumpto_distance	135
Sumpto_distancePerKilometer	1.34E-05
Sumpto_count	20
Sumpto_countPerKilometer	1.99E-06
Sumpto_duration	8114
Sumpto_durationPerKilometer	0.00080684
Sumgps_elevationloss	18978
Sumgps_elevationlossPerKilometer	0.001887134
Sumgps_elevationgain	19153
Sumgps_elevationgainPerKilometer	0.001904536
Sumcruisecontrol_distance	5854325
Sumcruisecontrol_distancePerKilometer	0.582142232
Sumgpslatitude_begin	18325.53676
Sumgpslatitude_beginPerKilometer	0.001822254
Sumgpslongitude_begin	1626.67179
Sumgpslongitude_beginPerKilometer	0.000161753
AverageSpeed	44397.68664
BreakDurationPerKilometer	0.004289257
HarshBreakDurationPerKilometer	0.00054492
FuelConsumptionPerKilometer	0.265162502
SumOperatingHours	226.51
AverageSpeedinKmH	44.39768664
DrivenKilometers	10056.52
NrOfTrips	358
ProductRange	XF MX-11
WARRANTY_PACKAGE	Standard + 3rd year driveline
Month_in_service	15
engineoil	Synthetic (ext)

Engine (DMSC ->CCM)	440
Homecountry	002 - Belgium
Application_y	General (dry freight, pallet loads)
bodytype	Tractor Not Applicable
drops_per_day	1 to 6
operation_type	Long Distance
Vehicle Safety Features	Not fitted
Asset Type-info	FT XF440Y X 380
Gearbox	AS Tronic 12 speeds
Contract_status	Active - Active
Retarder System	MX engine brake
Contract contracted yearly mileage	110000
Contract (version) start kms	0
Area of Operation	W.-Europe (excl. Scandinavia)
ChassisType	FT
fueltemperature_1_level_lo	0.004218122
fueltemperature_1_level_md	0.537751946
fueltemperature_1_level_mdhi	0.449920788
fueltemperature_1_level_hi	0.007699049
fueltemperature_1_level_vehi	0.000410095
ambientairtemperature_level_lo	0.003520408
ambientairtemperature_level_md	0.469846939
ambientairtemperature_level_mdhi	0.513418367
ambientairtemperature_level_hi	0.013214286
ambientairtemperature_level_vehi	0
engineintakeairpressure_1_level_lo	5.33E-06
engineintakeairpressure_1_level_md	0.549338002
engineintakeairpressure_1_level_mdhi	0.378398769
engineintakeairpressure_1_level_hi	0.066263183
engineintakeairpressure_1_level_vehi	0.005994718
engineoilpressure_1_level_lo	0.015085361
engineoilpressure_1_level_md	0.316457964
engineoilpressure_1_level_mdhi	0.658157943
engineoilpressure_1_level_hi	0.010298152
engineoilpressure_1_level_vehi	5.80E-07
tachographspeed_level_lo	0.102363946
tachographspeed_level_md	0.166666667
tachographspeed_level_mdhi	0.330238095

tachographspeed_level_hi	0.400459184
tachographspeed_level_vehi	0.000272109
acceleration_duration_level_lo	0.911695846
acceleration_duration_level_md	0.087355749
acceleration_duration_level_mdhi	0.000948405
acceleration_duration_level_hi	0
acceleration_duration_level_vehi	0
harshbrake_duration_level_lo	0.943695244
harshbrake_duration_level_md	0.055009276
harshbrake_duration_level_mdhi	0.00129548
harshbrake_duration_level_hi	0
harshbrake_duration_level_vehi	0
cruisecontrol_distance_level_lo	0.841935893
cruisecontrol_distance_level_md	0.149606353
cruisecontrol_distance_level_mdhi	0.008457754
cruisecontrol_distance_level_hi	0
cruisecontrol_distance_level_vehi	0
gps_elevationgain_level_lo	0.995292565
gps_elevationgain_level_md	0.004707435
gps_elevationgain_level_mdhi	0
gps_elevationgain_level_hi	0
gps_elevationgain_level_vehi	0
dpaanticipationscore_sum_level_lo	0.870778432
dpaanticipationscore_sum_level_md	0.124833229
dpaanticipationscore_sum_level_mdhi	0.004388338
dpaanticipationscore_sum_level_hi	0
dpaanticipationscore_sum_level_vehi	0

Q The searched hyperparameter space for each of the models

For the Logistic regression the hyper-parameter search space consisted out of the following:

- $C = [0.001, 0.01, 0.1, 1, 10, 100, 1000]$
- Regularization penalty method: [L1,L2]

For the Decision tree models, the hyper-parameter search space consisted out of the following:

- Maximum depth: [1,2,3...16]
- Minimum number of samples per split: [10, 15, 20...50]
- Minimum number of samples per leaf: [10, 12, 14.....30]

For the Random forest models, the hyper-parameter search space consisted out of the following:

- Maximum depth: [1,2,3...30]
- Minimum number of samples per split: [2, 4, 6...50]
- Minimum number of samples per leaf: [1, 3, 5.....31]
- Number of estimators: [50, 100, 150...1000]

For the MLP-Neural network models, the hyper-parameter search space consisted out of the following:

- Maximum number of iterations: [100, 200, 500, 1000, 1500, 2000]
- Number of hidden layers: [1,2,3...20]
- Learning rate: [constant, inverse scaling, adaptive]
- Activation function: [Logistic sigmoid, hyperbolic tangent, rectified linear]

R The optimal hyperparameters for each of the models.

Table 54: The optimal parameters for the decision tree models.

Model	Min. samples per leaf	Min. samples per split	Max. tree depth
Decision tree base 8 months	22	45	6
Decision tree base 11 months	10	25	5
Decision Tree extended 8 months	18	35	7
Decision tree extended 11 months	16	35	6

Table 55: The optimal parameters for the random forest models.

Model	Min. samples per leaf	Min. samples per split	Max. tree depth	Nr. of estimators
Random Forest base 8 months	7	14	15	500
Random Forest base 11 months	5	14	18	500
Random Forest extended 8 months	5	8	18	700
Random Forest extended 11 months	6	12	17	600

Table 56: The optimal parameters for the MLP-Neural Network models.

Model	Nr. of hidden layers	Activation function	Learning rate	max. nr. of iterations
MLP-NN base 8 months	4	rectified linear ($f(x) = \max(x,0)$)	inverse scaling (decreasing over time)	1000
MLP-NN base 11 months	5	rectified linear ($f(x) = \max(x,0)$)	inverse scaling (decreasing over time)	1000
MLP-NN extended 8 months	5	hyperbolic tangent	inverse scaling (decreasing over time)	1000
MLP-NN extended 11 months	7	rectified linear unit ($f(x) = \max(0,x)$)	inverse scaling (decreasing over time)	1000

S List of abbreviations

Table 57: The list of abbreviations.

DAF	DAF Trucks N.V.
R&M	Repair and maintenance
CBM	Condition based maintenance
PLM	Product life-cycle management
RUL	Remaining useful life
CCM	Customer contract management
CRISP-DM	Cross-industry standard process for data mining
SEMMA	Sample, Expolore, Modify, Model and Assess
KDD	Knowledge discovery in databases
FTE	fulltime-equivalent
AUC	Area under the curve
ROC-curve	receiver operating characteristic curve
TLFN	Time-lagged feedforward network
FIS	Fuzzy inference system
FFBPNN	Feedfword backpropagation neural network
ML	machine learning
RMSE	root mean square error
WILD	wheel impact load detectors
M-IFN	multi-target information fuzzy network
PCA	Principal component analysis
WGS	Wrapper with genetic search
WBFS	Wrapper with best-first search
BTA	Boosting tree algorithm
MLP	Multilayer perceptrons neural network
AE	Absolute error
MAE	Mean absolute error
RE	Relative error
MRE	Mean relative error
RNN	Recurrent neural network
ANFIS	Adaptive neuro fuzzy inference system
ARNFIS	Adaptive recurrent based neuro fuzzy inference system
MLFN	Multilayer feedforward neural network
NRMSE	normalized root mean squared error
DTW	Dynamic time warping

SVM	support vector machine
MTS	multivariate time-series
DMLP	Deep multilayer perceptron
FCN	Fully convolutional network
DCE	Dealer claim entry system
VIN	Vehicle identification number
CAN	Controller Area Network
DCM	Daf Connect Module
DPA	Driving Performance Assistant
PDF	Probability Distribution Function
TP	True positive
TN	True negative
FP	False positive
FN	False negative
ROC-curve	receiver operating characteristic curve
MI	Mutual Information
IQR	Inter Quartile Range
SFS	Sequential Feature Selection
RFE	Recursive Feature Elimination

References

- Andrew, Y. (2004). Feature selection, L 1 vs. L 2 regularization, and rotational invariance BT. In *Proceedings of the twenty-first international conference on Machine learning*, pages 379–387.
- Anthony J Viera and Joanne M. Garrett (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine*, 37(5):360–363.
- Arts, J. (2017). Maintenance, Modeling and Optimization. *Beta Working Paper*, 526:1–130.
- Azevedo, A. and Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADIS European Conference Data Mining*, (January):182–185.
- Barua, A. and Kosheleva, O. (2014). Why Trapezoidal and Triangular Membership Functions Work So Well : Towards a Theoretical Explanation. *Journal of uncertain systems*, 8(2013).
- Bastos, P., Lopes, I., and Pires, L. (2014). Application of data mining in a maintenance system for failure prediction. *Safety, Reliability and Risk Analysis: Beyond the Horizon-Steenbergen et al.(Eds)*.
- Baydogan, M. G., Runger, G., and Tuv, E. (2013). A Bag-of-Features Framework to Classify Time Series. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):245–252.
- Bolón-Canedo, V., Sánchez-Marño, N., and Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and Information Systems*, 34(3):483–519.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45:5–32.
- Brettel, M., Friederichsen, N., Keller, M., and Rosenberg, M. (2014). How-Virtualization-Decentralization-and-Network-Building-Change-the-Manufacturing-Landscape–An-Industry-40-Perspective. 8(1):37–44.
- Byington, C. S., Watson, M., and Edwards, D. (2004). Data-driven neural network methodology to remaining life predictions for aircraft actuator components. *IEEE Aerospace Conference Proceedings*, 6:3581–3589.
- Canizo, M., Onieva, E., Conde, A., Charramendieta, S., and Trujillo, S. (2017). Real-time Predictive Maintenance for Wind Turbines Using Big Data Frameworks. pages 1–8.
- Chandrashekar, G. and Sahin, F. (2014). A survey on feature selection methods. In *Computers and Electrical Engineering*, volume 40, pages 16–28. Elsevier Ltd.
- Chaochao Chen, Bin Zhang, Vachtsevanos, G., and Orchard, M. (2011). Machine Condition Prediction Based on Adaptive NeuroFuzzy and High-Order Particle Filtering. *IEEE Transactions on Industrial Electronics*, 58(9):4353–4364.
- Chen, P., Wilbik, A., Loon, S. V., Boer, A.-k., and Kaymak, U. (2018). Finding the optimal number of features based on mutual information. In *Advances in Intelligent Systems and Computing*, volume 641. Springer.
- Chinnam, R. B. and Baruah, P. (2004). A neuro-fuzzy approach for estimating mean residual life in condition-based maintenance systems. *International Journal of Materials and Product Technology*, 20(1/2/3):166.
- DAF (2018a). Facts and Figures. <https://www.daf.com/en/about-daf/facts-and-figures>.
- DAF (2018b). Products. <http://www.daf.com/en/products>.
- DAF (2018c). Services. <http://www.daf.com/en/services>.
- Efron, B. and Fron, B. E. (2012). The Estimation of Prediction Error The Estimation of Prediction Error : Covariance Penalties and Cross-Validation. 1459(2004).
- Estabrooks, A., Jo, T., and Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, 20(1):18–36.
- Fallis, A. (2013). *Applied Missing Data Analysis*, volume 53.

- Frisk, E., Krysander, M., and Larsson, E. (2014). Data-Driven Lead-Acid Battery Prognostics Using Random Survival Forests. pages 1–10.
- Fulcher, B. D. and Jones, N. S. (2014). Highly comparative feature-based time-series classification. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):3026–3037.
- Goudsmits, E. (2018). Predicting the number of repairs of a truck : a first step towards predictive maintenance - Thesis.
- Guyon, I. and Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research (JMLR)*, 3(3):1157–1182.
- Guyon, I., Weston, J., and Barnhill, S. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(3):389–422.
- Han, J., Kamber, M., and Pei, J. (2011). *Data mining concepts and techniques*. Morgan Kaufmann, Boston, 3 edition.
- Hartzell, A. L., Mark, G., Herbert, S., Silva, M. G., and Shea, H. R. (2011). *MEMS Reliability*. Springer, Boston, 2 edition.
- Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer, New York, 2 edition.
- Hong, Y. and Meeker, W. Q. (2010). Field-failure and warranty prediction based on auxiliary use-rate information. *Technometrics*, 52(2):148–159.
- Jackson, C. and Pascual, R. (2008). Optimal maintenance service contract negotiation with aging equipment. *European Journal of Operational Research*, 189(2):387–398.
- Kantardzic, M. (2011). *Data Mining: concepts, models, methods, and algorithms*. Wiley, Hoboken, 2 edition.
- Kaymak, U., den Bergh, W.-M. V., and Berg, J. V. D. (2003). A Fuzzy Additive Reasoning Scheme for Probabilistic Mamdani Fuzzy Systems. *The IEEE International Conference on Fuzzy System*, 12(I):331–336.
- Khaleghi, A., Ryabko, D., Mary, J., and Preux, P. (2016). Consistent Algorithms for Clustering Time Series. *Journal of Machine Learning Research*, 17(3):1–32.
- Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 69(6):16.
- Kurt, I., Ture, M., and Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications*, 34(1):366–374.
- Kusiak, A. and Verma, A. (2012). Analyzing bearing faults in wind turbines: A data-mining approach. *Renewable Energy*, 48:110–116.
- Last, M., Sinaiski, A., and Subramania, H. S. (2011). Condition-based Maintenance with Multi-Target Classification Models. 29(New Generation Computing, Ohmsha, Ltd. and Springer):245–260.
- Lawless, J. F. (1998). Statistical analysis of product warranty data. *International Statistical Review*, 66(1):41–60.
- LeCun, Y., Bottou, L., and Muller, K. (2012). Efficient BackProp. In Montavon, G. and Orr, G., editors, *Neural Networks: Tricks of the trade.*, pages 9–48. Springer, Berlin, 2 edition.
- Lee, J., Kao, H. A., and Yang, S. (2014). Service innovation and smart analytics for Industry 4.0 and big data environment. *Procedia CIRP*, 16:3–8.
- Lee, J., Lapira, E., Bagheri, B., and an Kao, H. (2013). Recent advances and trends in predictive manufacturing systems in big data environment. *Manufacturing Letters*, 1(1):38–41.
- Lee, S. G., Ma, Y. S., Thimm, G. L., and Verstraeten, J. (2008). Product lifecycle management in aviation maintenance, repair and overhaul. *Computers in Industry*, 59(2-3):296–303.

- Li, J., Tao, F., Cheng, Y., and Zhao, L. (2015). Big Data in product lifecycle management. *International Journal of Advanced Manufacturing Technology*, 81(1-4):667–684.
- Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R news*, 2(December):18–22.
- Marinelli, M., Lambropoulos, S., and Petroutsatou, K. (2014). Earthmoving trucks condition level prediction using neural networks. *Journal of Quality in Maintenance Engineering*, 20(2):182–192.
- Mathew, V., Toby, T., Singh, V., Rao, B. M., and Kumar, M. G. (2017). Prediction of Remaining Useful Lifetime (RUL) of Turbofan Engine using Machine Learning. pages 306–311.
- Meeker, W. Q. and Hong, Y. (2014). Reliability meets big data: Opportunities and challenges. *Quality Engineering*, 26(1):102–116.
- Murthy, D. N. P. and Djameludin, I. (2002). New product warranty:A literature review. *International Journal of Production Economics*, 79(3):231–260.
- Nanopoulos, A., Alcock, R., Manolopoulos, Y., Mastorakis, N. E., Nikolopoulos, S. D., and Manolopoulos, Y. (2001). Feature-based Classification of Time-series Data. *Information processing and technology*, 0056:49–61.
- Negnevitsky, M. (2005). *Artificial Intelligence - A guide to intelligent systems*. Pearson, Essex, 2 edition.
- Nelson, P. R., Coffin, M., and Copeland, K. A. F. (2003). *Introductory statistics for engineering experimentation*. Academic Press.
- Peng, Y., Dong, M., and Zuo, M. J. (2010). Current status of machine prognostics in condition-based maintenance: A review. *International Journal of Advanced Manufacturing Technology*, 50(1-4):297–313.
- Prytz, R., Nowaczyk, S., Rögnvaldsson, T., and Byttner, S. (2015). Predicting the need for vehicle compressor repairs using maintenance records and logged vehicle data. *Engineering Applications of Artificial Intelligence*, 41:139–150.
- Prytz, R., Nowaczyk, S., Thorsteinn, R., and Byttner, S. (2013). Analysis of Truck Compressor Failures Based on Logged Vehicle Data. In *9th International Conference on Data Mining*, number Lvd, Las Vegas.
- Rodríguez, J. J. and Alonso, C. J. (2004). Interval and Dynamic Time Warping-based Decision Trees . pages 1–5.
- Rohani, A., Abbaspour-Fard, M. H., and Abdollahpour, S. (2011). Prediction of tractor repair and maintenance costs using Artificial Neural Network. *Expert Systems with Applications*, 38(7):8999–9007.
- Ross, B. C. (2014). Mutual information between discrete and continuous data sets. *PLoS ONE*, 9(2).
- Rutkowski, L., Jaworski, M., Pietruczuk, L., and Duda, P. (2014). The CART decision tree for mining data streams. *Information Sciences*, 266:1–15.
- Singhal, A. and Seborg, D. E. (2006). Clustering multivariate time-series data. (January):427–438.
- Tchakoua, P., Wamkeue, R., Ouhrouche, M., Slaoui-Hasnaoui, F., Tameghe, T. A., and Ekemb, G. (2014). Wind turbine condition monitoring: State-of-the-art review, new trends, and future challenges. *Energies*, 7(4):2595–2630.
- Triebel, R., Mozos, O., Burgard, W., Preisach, C., Burkhardt, H., Schmidt-Thieme, L., and Decker, R. (2008). *Studies in Classification, Data Analysis, and Knowledge Organisation*.
- van den Berg, J., Kaymak, U., and van den Bergh, W. M. (2004). Financial markets analysis by using a probabilistic fuzzy modelling approach. *International Journal of Approximate Reasoning*, 35(3):291–305.
- Vieira, S. M., Sousa, J. M., and Kaymak, U. (2012). Fuzzy criteria for feature selection. *Fuzzy Sets and Systems*, 189(1):1–18.
- Walfish, S. (2006). A Review of Statistical Outlier Methods. *Pharmaceutical Technology*, (4):1–5.

- Waltman, L., Kaymak, U., and Berg, J. V. D. (2005). Fuzzy Histograms : A Statistical Analysis. In *Proceedings of the Joint 4th Conference of the European Society for Fuzzy Logic and Technology and the 11th Rencontres Francophones sur la Logique Floue et ses Applications*, volume 2, pages 605–610, Barcelona.
- Wang, L., Wang, Z., and Liu, S. (2016). An effective multivariate time series classification approach using echo state network and adaptive differential evolution algorithm. *Expert Systems with Applications*, 43:237–249.
- Wirth, R. (2000). CRISP-DM : Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, (24959):29–39.
- Witten, I. H. , Frank, E., & H. (2016). *Data mining - Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, Burlington, 3 edition.
- Wu, S. (2012). Warranty data analysis: A review. *Quality and Reliability Engineering International*, 28(8):795–805.
- Wu, S. and Akbarov, A. (2011). Support vector regression for warranty claim forecasting. *European Journal of Operational Research*, 213(1):196–204.
- Wu, S. and Akbarov, A. (2012). Forecasting warranty claims for recently launched products. *Reliability Engineering and System Safety*, 106:160–164.
- Wu, S.-j., Gebrael, N., Lawley, M. A., and Yih, Y. (2007). System for Condition-Based Optimal Predictive Maintenance Policy. 37(2):226–236.
- Yang, C. and Létourneau, S. (2005). Learning to predict train wheel failures. *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining - KDD '05*, page 516.
- Ye, Z. S. and Murthy, D. N. P. (2016). Warranty menu design for a two-dimensional warranty. *Reliability Engineering and System Safety*, 155:21–29.
- Zhang, S., Zhang, C., and Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, 17(5-6):375–381.
- Zheng, Y., Liu, Q., Chen, E., Ge, Y., and Zhao, J. L. (2014). Time Series Classification Using Multi-Channels Deep Convolutional Neural Networks. pages 298–310.