Analysis of Truck Compressor Failures Based on Logged Vehicle Data

Rune Prytz, Sławomir Nowaczyk, Thorsteinn Rögnvaldsson, *Member, IEEE*, and Stefan Byttner

***Abstract*— In multiple industries, including automotive one, predictive maintenance is becoming more and more important, especially since the focus shifts from product to service-based operation. It requires, among other, being able to provide customers with uptime guarantees. It is natural to investigate the use of data mining techniques, especially since the same shift of focus, as well as technological advancements in the telecommunication solutions, makes long-term data collection more widespread.**

In this paper we describe our experiences in predicting compressor faults using data that is logged on-board Volvo trucks. We discuss unique challenges that are posed by the specifics of the automotive domain. We show that predictive maintenance is possible and can result in significant cost savings, despite the relatively low amount of data available. We also discuss some of the problems we have encountered by employing out-of-the-box machine learning solutions, and identify areas where our task diverges from common assumptions underlying the majority of data mining research.

***Index Terms*— Data Mining, Machine Learning, Fault Prediction, Automotive Diagnostics, Logged Vehicle Data**

I. Introduction

With modern vehicles becoming more and more sophisticated cyber-physical systems, increased software and system complexity poses new development and maintenance challenges. For commercial ground fleet operators, including bus and truck companies, the maintenance strategy is typically reactive, meaning that a fault is fixed only after it has become an issue affecting vehicle's performance.

Currently, there is a desire for truck manufacturers to offer uptime guarantees to their customers, which obviously requires a shift in the paradigm. New ways of thinking about component maintenance, scheduling and replacement need to be introduced. Statistical lifetime predictions are no longer sufficient, and workshop operations need to be planned and their results analysed at the level of individual vehicles.

At the same time, it is slowly becoming feasible to analyse large amounts of data on-board trucks and buses in a timely manner. This enables approaches based on data mining and pattern recognition techniques to augment existing, hand crafted algorithms. Such technologies, however, are not yet in the product stage, and even once they are deployed, a significant time will be required to gather enough data to obtain consistently good results.

In the meantime, it is necessary to explore existing data sources. One example of that is Volvo's "Logged Vehicle Database" (LVD), that collects statistics about usage and internal workings of every vehicle. This data is stored on-board Electronic Control Units during regular operation, and uploaded to a central system during visits in authorised workshops.

The LVD is just one database among many that are of interest for predictive maintenance purposes. Others that are being currently used in related projects include "Vehicle Data Administration" (VDA) and "Vehicle Service Records" (VSR). These databases each contain different, but complementary information: usage statistics and ambient conditions, up-to-date information regarding vehicle equipment, design and configuration specifications, as well as history of all maintenance and repair actions conducted at Volvo Authorised Workshops.

In a typical data mining study, the underlying assumption is that a lot of information is available. For example, it is common in fault prediction research to be able to continuously monitor the device in question. In this regard, the automotive domain is much more restrictive. We are only able to observe any given truck a couple of times per year, at intervals that are unknown *a priori* and difficult to predict even during operation.

In this project we have decided to focus on analysing two components: compressor and turbocharger. Due to lack of space, in this work we only present results related to the compressor, but most of our discussions are valid for both subsystems. The main motivation of predictive maintenance is the possibility to reduce the unplanned stops at the road side. They can be very costly, both for the customer and for the OEM.

If the truck is under warranty or service contract the following expenses could typically be incurred: towing, disruption of garage workflow, actual repair, rent of replacement truck and loss of OEM reputation. During a service contract all maintenance and service costs are covered by a fixed monthly fee. A secondary motivation is to minimise the amount of maintenance that is done on trucks under service contract while still guaranteeing required level of uptime towards the customer.

Additionally, certain components, such as the turbocharger or timing belt, cause significant collateral damage to the vehicle when they fail. Such components are often already either designed to last the full lifetime of the vehicle or scheduled for planned maintenance. In practice, however, this is not enough to prevent all unexpected failures. In these cases predictive maintenance would also be very effective in reducing the excess cost, even though the number of

Rune Prytz is with the Volvo Group Trucks Technology, Advanced Technology & Research Göteborg, Sweden (email: rune.prytz@volvo.com).

Sławomir Nowaczyk, Thorsteinn Rögnvaldsson and Stefan Byttner are with the Center for Applied Intelligent Systems Research, Halmstad University, Sweden (emails follow firstname.lastname@hh.se pattern).

breakdowns is low.

Obviously, predictive maintenance not only saves money, it also introduces additional expenses in terms of unnecessary repairs for the wrongly diagnosed vehicles as well as wasted component life. The latter comes from the fact that the still working component gets exchanged.

The importance of this factor varies greatly depending on particular application. In this study we disregard it completely, since both turbocharger and compressor are exchanged at most once during a vehicles lifetime.

The other cost factor, incorrectly diagnosed failures, can never be completely avoided, but is expected to be surpassed by the savings obtained from finding vehicles before they have an unexpected breakdown. This expense will be the major focus of our discussions in this work.

From classification point view, this can be directly linked to the ratio between True Positive examples and False Positive ones. As mentioned previously, the cost of one on-the-road breakdown is far greater than the cost of one unnecessary component replacement. It is also important to notice that the number of False Negatives is almost irrelevant in this application. They represent "wasted opportunity," i.e. money that could potentially be saved but was not, however they do not incur any direct expenses.

The predictive maintenance solution we are proposing in this paper is designed to be used as an aid in the garage. Whenever a truck is in the workshop for whatever reason, logged data is collected and analysed. The classification algorithm then marks the vehicle as either normal or in need of compressor replacement (within a specified prediction horizon). The workshop will then either exchange the compressor right away, perform additional diagnostics, or schedule another visit in the near future.

This paper is organised as follows. In the next section we describe in more detail the type of data we are working with, as well as present the business constraints that dictate how we state the problem and how are we trying to solve it. We follow by a discussion of related research in Section III. We present our approach in Section IV and results of experiments we have conducted in Section V. We close with conclusions in Section VI.

II. Data and constraints

A typical quality measurement in the automotive industry is the fault frequency of a component. It's percentage of components that fail within a given time: most typically, either a warranty or service contract period. However, that is not a representative measure for our case. Our data consists of a number of data readouts from each truck, spread over long time, but compressor or turbocharger gets replaced at most once.

Most of the vehicles never have a failure of the components we are interested in. Even for those that do, many of the readouts come from the time when the compressor is in good condition, and only in some cases there is a readout from the workshop visit when it is exchanged.

In order to get representative data, we need to select our examples from three scenarios: some of the data should come from trucks on which compressor never failed, some should come from readouts shortly before compressor failure, and some should come from trucks on which the compressor failed far in the future. In order to ensure that, we also consider the number of readouts that is available from each vehicle. Trucks that have too few readouts or do not contain all the data parameters we are interested in are discarded at this stage.

One of the topics of our analysis is to investigate how does the relative ratio of positive and negative examples in train and test datasets influence machine learning results. It is obvious that component failures are an exception rather than a norm. However, there are different ways of measuring the precise ratio between "faulty" and "good" cases. Nevertheless, the fault frequency in the vehicle population does not necessarily translate directly into exactly the same level of imbalance between examples.

We are not able to disclose any real fault frequency data. However, as a guidance, high fault frequency is between 5-10% while a good components may have fault frequency in the range of 0 to 3%. In this paper we will construct the dataset in such way that the baseline fault frequency is 5%. It is important to be aware, however, that there are many factors affecting this and under different circumstances, the data can look very different. Examples include truck configuration and age, usage patterns, geographical location and many more

As a simple example, we can easily imagine a predictive maintenance system being deployed and not applied to all vehicles, but only to those that service technicians consider "high risk". Similarly, while compressor is an important component to monitor, the methodology itself is fully general, and there are other parts that could be targeted. Some of them are designed to be replaced regularly, and thus could have failures that occur on almost all trucks. Therefore, in several places in this paper, we will discuss how different fault frequencies affect classification results.

The vehicles in our dataset are all Volvo trucks, from the same year model, but equipped with three different compressor types. They also vary with respect to geographical location, owner, and type of operation, for instance long-haul, delivery or construction.

We have selected 80 trucks which had compressor failures and at least 10 LVD readouts, with the right number of parameters available. In addition we have chosen 1440 trucks on which, so far at least, no compressor had failed. They all fulfil the same requirements on LVD data. We could easily obtain more "non-faulty" vehicles, but it is the ones with compressor failures that are the limiting factor.

A. Logged Vehicle Data

Logged Vehicle Data is a Volvo internal database which gathers usage and ambient statistics collected from Volvo vehicles. The data is downloaded from the truck when it is serviced at an authorised Volvo workshop, or wirelessly through a telematics gateway. The database is used for

various tasks during product development, after market and even sales support.

A typical task for product development would be to support a simulation or validate an assumption with real usage statistics from the field. For instance, such questions could concern the relationship between average fuel economy and weight, altitude or engine type. During the sales process the database can provide usage statistics for already existing customers, which is helpful in configuring the right truck for a particular purpose.

This database contains data of varying types and has high number of dimensions. Typically a vehicle record contains hundreds of parameters and at most tens of readouts. The number of readouts directly depends on the availability of telematics equipment and on whether the vehicle has been regularly maintained at a Volvo workshop. For example, in our dataset the average number of readouts per vehicle is 4 per year. However, the variance is very high and many trucks have one or less readouts per.

There is also a problem with missing values, typically caused by connectivity issues or software updates. Modern on-board software versions log more parameters, which means that older readouts tend to include less data than newer ones.

Finally, the stored parameters are typically of cumulative nature. This means that the readouts are highly correlated and not *independently identically distributed*, as is usually assumed in machine learning. It could be interested to analyse, instead of the LVD data itself, the changes between subsequent readouts — but it can be complicated because there is a number of different aggregation schemes employed (for example, averages, accumulators and histograms).

B. VSR and VDA

The Volvo Service Records a database that keeps track of all maintenance and repair operations done on a particular vehicle. The database is mainly used by the workshop personnel for invoicing purposes, as well as for diagnostics, allowing to check previously carried out repairs.

A typical repair event contains date, current mileage, and a list of unique maintenance operation codes and exchanged part numbers. In addition to that there may be a text note added by the technician. For the purposes of this work, we are using VSR to find out whether and when a compressor was replaced on a given truck.

The VDA database contains vehicle specification for all vehicles produced by Volvo. It lists the included components such as gearbox model, wheel size, cab version, or engine and compressor type. All options have a unique label which makes it easy to use for classification.

III. Related Work

In a survey of Artificial Intelligence solutions being used within automotive industry, [1] discusses, among other things, both fault prognostics and after-sales service and warranty claims. An representative example of work being done in this area are [2] and [3], where authors present two data mining algorithms that extracts associative and sequential patterns from a large automotive warranty database, capturing relationships among occurrences of warranty claims over time. Employing a simple IF-THEN rules representation, the algorithm allows filtering out insignificant patterns using a number of rule strength parameters. In that work, however, no information about vehicle usage is available, and the discovered knowledge is of a statistical nature concerning relations between common faults, rather than describing concrete individual.

More recently [4] presented a survey of 150 papers related to the use of data mining in manufacturing. While their scope was broader than only diagnostics and fault prediction, including areas such as design, supply chain and customer relations, they have covered a large portion of literature related to the topic of this paper. The general conclusion is that the specifics of automotive domain make fault prediction a more challenging problem than in other domains: almost all research considers a case where continuous monitoring of devices is possible, e.g. [5] or [6].

It is more common to consider emergent solutions, where vehicles are able to communicate using telematic gateways. An early paper [7] shows a system architecture for distributed data-mining in vehicles, and discusses the challenges in automating vehicle data analysis. In [8] cross-fleet analysis, i.e. comparing properties of different vehicles, is shown to benefit root-cause analysis for pre-production diagnostics. In [9] and [10], a method called COSMO is proposed for distributed search of “interesting relations” among on-board signals in a fleet of vehicles, enabling deviation detection in specific components.

A method based on a similar concept of monitoring correlations, but for a single vehicle instead of a fleet, is shown in D’Silva [11]. In Vachkov [12], the neural gas algorithm is used to model interesting relations for diagnostic of hydraulic excavators. Contrary to our work, however, both the papers by D’Silva and Vachkov assume that the signals which contain the interesting relations are known *a priori*. In [13], a method for monitoring relations between signals in aircraft engines is presented. Relations are compared across a fleet of planes and flights. Unlike us, however, they focus on discovering relationships that are later evaluated by domain experts.

Even though not particularly recent, [14] and [15] are still excellent introductions to more general machine learning and artificial intelligence topics. In this paper we are also facing many challenges related to the imbalanced nature of diagnostics data. In order to make our initial investigations more widely accessible we have decided not to use any specialised solutions, but an overview of research on this area can be found, for example, in [16], [17] or [18].

IV. Approach

We have decided to base our initial analysis on using out-of-the-box supervised classification algorithms. From among the available attributes, 4 interesting VDA parameters and 8 LVD interesting parameters were chosen by experts within

Volvo. Those include, for example: compressor model, engine type, vehicle mileage, average compressed air usage per kilometre, etc.

At this stage of our research, we have decided to consider each data readout as a single learning example. Even though they definitely do not satisfy the basic *independent and identically distributed* assumption, this gives us flexibility in both the classifier choice and in deciding how to analyse actual faults.

When constructing the dataset we need to merge data from the three databases. First we find, in the VSR, all truck that had the compressor exchanged. To do that we use the unique maintenance code for compressor replacement. After that we find all the LVD and VDA data for the faulty vehicles, up to and until the aforementioned repair occurred. At this stage we discard some vehicles, either because they do not have sufficient number of readouts or because not all the interesting parameters selected by Volvo experts are available. After that we also select some number of "non-faulty" trucks.

For each LVD readout, we also create a new parameter denoting time to repair. It uses the timestamp of repair entry in VSR and this particular readout's date. In the case of non-faulty trucks we are assuming that they may break just after the latest readout available, so that the *time to repair* parameter can be calculated for all trucks. This parameter is later used for labelling examples as either positive or negative, based on the prediction horizon, but is of course not used for classification. This step is one of the areas where there is definitive room for improvement, since it is definitely not clear, however, when – if at all – the symptoms for the imminent failure become visible in the data.

When selecting examples for classification a prediction horizon and the desired fault rate must first be defined. The *time to repair* parameter is used to determine which readouts are considered as positive: those that fall within the prediction horizon. After that, at most two examples per vehicle are drawn to form the training and test datasets.

For the trucks marked as faulty, we select exactly one positive and one negative example, at random. Finally, we add one negative example from the remaining trucks until the desired fault frequency is archived. By selecting an equal (and small) number of positive and negative examples from each truck we avoid the problem of classifiers learning characteristics of individual vehicles rather than those of failing compressors.

The reason for choosing random readouts as examples is twofold. First of all, it is not entirely clear how to choose which data readout is the best one to use. It is important that there is sufficient distance between corresponding positive and negative example, in order for the data to be changed significantly. The further apart the two examples are, the larger the chance that symptoms of failing compressor are present in the positive example and are missing from the negative one. On the other hand, selecting dates close to the cutoff boundary would allow more precision in estimating *when* the components is likely to break.

The random approach avoids any systematic bias in either direction, but it means that actual training dataset only depends on the prediction horizon to a limited degree. It also means that we have no real control over how similar positive and negative examples actually are. It is an interesting question of how to find the appropriate cutoff point automatically, preferable on an individual basis.

In the final step, we remove 10% of the dataset, to be used as the test data, and use the rest as train data. Since we have few examples available, we use both out-of-bag evaluation on the training dataset, as well as the separate evaluation on the test data. In section V we sometimes present both evaluations, and sometimes only one of them, depending on which one is more appropriate for a particular purpose.

One of the issues with out-of-bag evaluations is that it is computationally intense. To speed up the processing, each classifier is only evaluated on a subset of the train data. The out-of-bag evaluation subset contains all the positive examples, but only a portion of negative examples. The resulting confusion matrix is then up-scaled for the *true negatives* and *false positives*.

As an evaluation of the business case for the predictive maintenance solution, we introduce measure of cost savings:

$$C_{save} = TP \cdot (C_u - C_p) - FP \cdot C_p$$

The method will be profitable if the correctly classified faulty trucks (i.e. *true positives TP*) save more money than the non-faulty trucks wrongly classified as faulty (i.e. *false positive FP* waste. Because an on-road *unplanned* breakdown costs (C_u) is much higher than the *planned* component replacement (C_p), every TP reduces costs.

A. Learning algorithms

In this work we have used the KNN, C5.0 and Random Forest learning algorithms. Each of them is evaluated in R using the Caret package as described in [19]. By default, the Caret package tunes the parameters of each classifier.

V. Experiments

In this section we present the results of early experiments we have performed. Throughout this presentation we have two main goals. First, we argue that those initial results are encouraging and promise a tangible business benefits, thus warranting further work, and hopefully inspiring others to investigate similar approaches in other applications. Second, we demonstrate difficulties we have encountered due to the type of data available and specifics of the domain.

As the first step towards familiarising the reader with our data, we present how the dataset size affects quality of classification. In Figure 1 we have plotted the classification accuracy, both using out-of-bag evaluation and a separate test set, for all three classifiers.

This figure is mainly useful to show the level of variance in classifier behaviour, since — even though it looks impressive — accuracy is not a particularly suitable measure for this

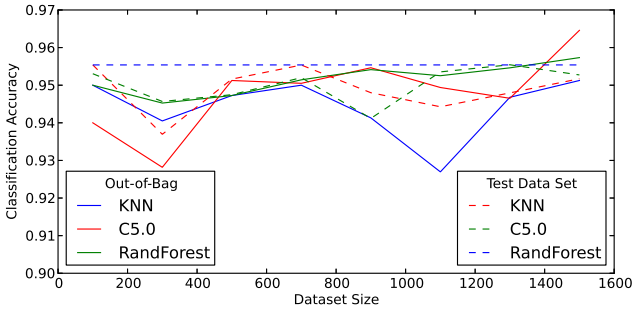

Fig. 1. Impact of dataset size on classification accuracy

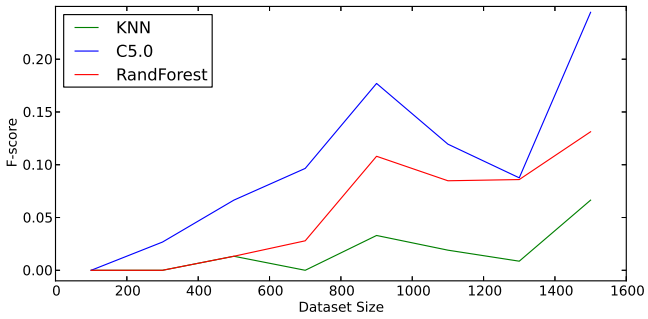

Fig. 2. Impact of dataset size on $F_{0.5}$-score

problem. As explained before, the baseline for our analysis is to assume 5% fault frequency, and this is the ratio between positive and negative examples in both training and test datasets.

Therefore, accuracy of 95% can be achieved in a very simple manner, by doing no generalisation whatsoever and simply answering "No" to every query. As can be seen from the plot, classification algorithms we are using are employing more complicated schemes, but only Random Forests consistently beats that simplest strategy, and only on the test data set — which in itself is not entirely conclusive, due to the limited size of the data we are working with.

Finally, this plot also shows that there is no significant difference in results between out-of-bag and test data evaluations. Therefore, in some of the subsequent plots we will limit ourselves to only presenting one of them, unless particular scenario makes both interesting.

In figure 2 we are presenting the F-score:

$$F = (1+\beta^2)\frac{\text{precision} * \text{recall}}{\beta^2 * \text{precision} + \text{recall}},$$

as this is one of the most popular measures that is actually suitable for highly imbalanced data sets. In our case we have decided to use parameter $\beta = 0.5$, because in this application, precision is significantly more important than recall: every compressor that we do not flag as needing replacement simply maintains *status quo*, while every unnecessary repair costs money.

By analysing this plot it is clearly visible that the dataset we have currently access to is very small, only barely sufficient for the analysis. Even when using all the data as the training set, the F-score of the best classifier barely exceeds 0.2. On the other hand, this plot clearly shows that we have not yet reached saturation levels, and it is reasonable to assume that as more data becomes available, the quality of classification will continue to increase. This also means that most of the results presented subsequently can be expected to improve in the future.

One of the most interesting questions with regard to predictive maintenance is how early in advance can faults be detected. In order to answer that, we have performed an experiment where we were interested in evaluating the influence of prediction horizon on the classification quality.

In this case we have decided to present the results in Figure 3 for three different values of fault frequency (colours correspond to different classifiers, while line styles denote 5%, 20% or 50% class distribution). The imbalanced nature of the data is obviously a problem, but as we have discussed in section II, there is significant flexibility in how the final product will be deployed, and that allows us some freedom. Therefore, it is interesting to see prediction quality in a number of settings. That said, the performance on highly skewed data sets is still the most important one, because other solutions typically involve various kinds of cost-incurring tradeoffs. In order to not clutter the figure, we only include F-score evaluated using out-of-bag method.

In most diagnostic applications the prediction horizon is a very, if not the most, important measure. In out case, however, it is both less critical and more difficult to define precisely. The former comes from the fact that one is only expected to exchange compressor once in a lifetime of a vehicle. Therefore, the precise time of when is it done, as long as it is reasonable, does not directly influence the costs. There are, of course, some benefits of minimising wasted remaining useful life, but they are difficult to measure since they mainly relate to customer satisfaction.

The difficulty in defining the prediction horizon, however, is definitely something we are interested in investigating further. One idea would be to take into account individual usage

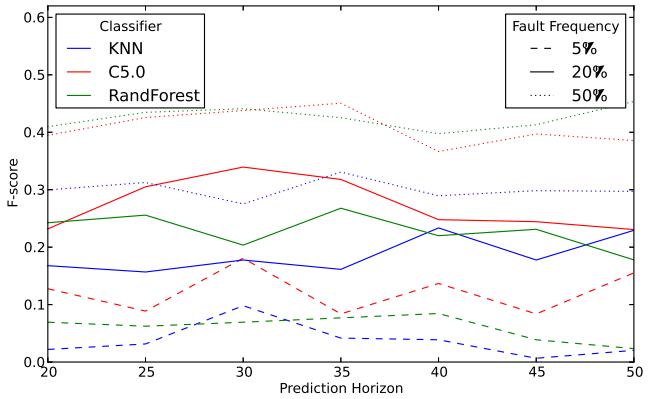

Fig. 3. $F_{0.5}$-score as a function of prediction horizon, for three different levels of fault frequency in vehicle population

patterns of trucks, for example by assuming that vehicles that are rarely in the workshop should have longer advance notice, while those that are maintained more regularly can wait until the failure is more imminent.

At the moment, however, we are treating all data readouts as individual and independent examples, and therefore each of them has to be marked as either positive or negative one. We use a very simple scheme of assuming that all examples closer to the failure than the prediction horizon are positive, and all examples further away are negative. This, however, makes analysing influence of prediction horizon on the classification quality more difficult, especially taking into account the irregular intervals at which we obtain vehicle data.

Moreover, during our first attempts of analysing the data (which we are not presenting here due to space constraints), we have encountered a situation that all machine learning algorithms learned to almost exclusively consider characteristics of particular trucks, instead of indicators of failing compressor. They would provide, for most of the vehicles, predictions that never changed over time. This resulted in classifiers that achieved good accuracy and F-score, but were completely useless from business point of view.

To this end we have decided to use exactly two data readouts from each vehicle on which we have observed compressor replacement: one positive and one negative example. This solves the aforementioned problem, since now there is no benefit to distinguishing individual, but it even further reduces the size of available data. In addition, it is not entirely clear how to choose which data readout to use, if we can only use one of them.

On the one hand, one would want to use readouts as close to the prediction horizon boundary as possible, to be highly precise in predicting wasted life of the components. On the other hand, it is not good to choose positive and negative examples that are too close in time, since it is very likely that the difference in logged data between those two points does not contain any new information about state of the compressor.

To this end, we have decided to choose one example from each side of the prediction horizon boundary at random. It means, however, that varying the prediction horizon only introduces small changes in the actual training and test datasets. It may even happen that for two significantly different values of the horizon, we end up with the same data. This explains the results that can be seen in Figure 3: prediction horizon has very little influence on the F-score.

Accuracy and F-score are important measures from research point of view. The inspiration for our work, however, arises from practical needs of automotive industry, and the major measure from the business perspective is clearly cost reduction. It is very expensive to have components fail during transport missions, because not only does it introduce disruptions in the workshop operations, it also incurs other costs, like towing, collateral damage, and customer dissatisfaction. Therefore, it is cheaper to replace components during

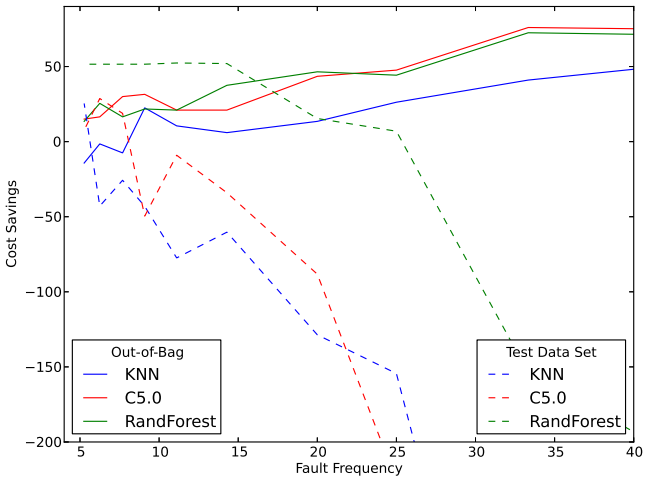

Fig. 4. Maintenance cost savings that can be achieved for varying fault frequency in training dataset (test set always has 5% of positive examples).

scheduled maintenance. The exact degree to which this is the case varies, of course, from component to component, and depends on which factors are taken into account: reputation, for example, is notoriously difficult to appraise.

Therefore, in order to be on the safe side, we have decided to use a factor of 2.5 to measure cost savings that can be provided by our solution. In other words, it costs on average two and a half as much to repair a truck in which compressor failed on the road, as it would cost to replace this component as a scheduled operation.

Figure 4 shows how the benefits of introducing our predictive maintenance solution depend on the fault rate in the vehicle population. The most interesting is, of course, the left side of the plot, because it shows that even the low quality classification results that we are able to obtain from our 1600 data samples are enough to offer tangible benefits. Both Random Forest and C5.0 classifiers are accurate enough to save expenses.

It is interesting to see how cost savings (at least looking at out-of-bag data) grow as the imbalance in the data decreases. This is consistent with results from Figure 2 and can be easily explained by the higher quality of classification.

On the other hand, the cost when measured on the test set drops very rapidly (except for the Random Forest classifier, the result which we are not able to explain just yet). The reason for this behaviour is that the test data always contains 95%–5% split of negative and positive examples. As the distribution of data in the training set become more and more different from the distribution in test set, the quality of classification drops.

Finally, in Figure 5 we present the relation between True Positives and False Positives, again as a function of fault frequency. We are only using out-of-bag evaluation here. This is the plot that actually contains the most information, since those are the two factors that directly affect the economical viability of our solution. As mentioned earlier, presence of False Negatives does not affect the cost in any direct way.

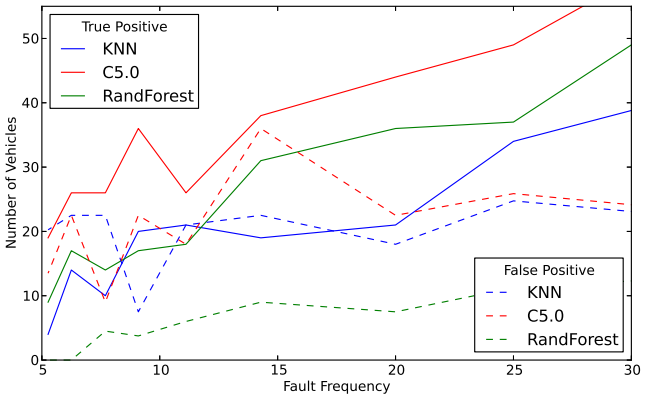

Fig. 5. True Positives and True Negatives

It is interesting to look at the differences between the three classifiers, and the potential tradeoffs that may be important from business perspective.

It is clear that KNN is not well-suited for this particular problem, although it can possibly be explained by the fact that we have not performed any data normalisation, and the large differences in absolute values of various parameters may be difficult for it to handle. Even for more balanced data sets, this classifier is struggling to obtain more True Positives than False Positives.

From the pure cost perspective, Random Forest seems to be better than C5.0, because the difference between True Positives and True Negatives is larger. On the other hand, C5.0 actually detects more faulty compressors, in simply makes more FP mistakes as well. In Figure 4 those two classifiers score very close, but if we would assume another relative costs for planned and unplanned component replacements, the difference between them could be significant. It would be interesting to investigate what is the reason for this difference, and possibly to identify parameters that would allow us to control this tradeoff.

VI. Conclusions and Future Work

The most important conclusion of this work is that using data mining based on Logged Vehicle Data as predictive maintenance solution in automotive industry is a viable approach. We will continue the work in this area, investigating more complex machine learning approaches. Current classification quality and cost avoidance is not great, but it is expected to increase as we get access to more data and as we replace generic algorithms with more specialised ones.

It is known that data availability will dramatically increase as the new Volvo truck reaches the customers. It is equipped with new and enhanced telematics platform, enabling larger and more frequent LVD readouts.

The second contribution of this paper is identifying a number of distinctive features of automotive industry, and discussion regarding to what degree do they fit typical machine learning and data mining research paradigms.

Ideas for future work include extending this analysis to other components, especially the ones where "exchange once in a lifetime" assumption does not hold, as well as evaluating known methods of dealing with imbalanced data sets.

It is also necessary to define the notion of prediction horizon in a better way, preferably allowing learning algorithm to choose the threshold in an individualised manner. Another approach to investigate is to use regression to predict *time to repair*. One possible solution would be to look at the differences between readouts, as this may decrease the correlation between examples and enhance classification performance.

Acknowledgement

Parts of this work have been supported by Halmstad University, Volvo GIB-T, VINNOVA (the Swedish Governmental Agency for Innovation Systems) and The Knowledge Foundation (KK-stiftelsen).

References

[1] O. Gusikhin, N. Rychtyckyj, and D. Filev, "Intelligent systems in the automotive industry: applications and trends," *Knowledge and Information Systems*, vol. 12, pp. 147–168, 2007.

[2] J. Buddhakulsomsiri, Y. Siradeghyan, A. Zakarian, and X. Li, "Association rule-generation algorithm for mining automotive warranty data," *International Journal of Production Research*, vol. 44, no. 14, pp. 2749–2770, 2006.

[3] J. Buddhakulsomsiri and A. Zakarian, "Sequential pattern mining algorithm for automotive warranty data," *Computers & Industrial Engineering*, vol. 57, no. 1, pp. 137 – 147, 2009.

[4] A. Choudhary, J. Harding, and M. Tiwari, "Data mining in manufacturing: a review based on the kind of knowledge," *Journal of Intelligent Manufacturing*, vol. 20, pp. 501–521, 2009.

[5] A. Kusiak and A. Verma, "Analyzing bearing faults in wind turbines: A data-mining approach," *Renewable Energy*, vol. 48, 2012.

[6] A. Alzghoul, M. Löfstrand, and B. Backe, "Data stream forecasting for system fault prediction," *Computers & Industrial Engineering*, vol. 62, no. 4, pp. 972–978, May 2012.

[7] H. Kargupta *et al.*, "VEDAS: A mobile and distributed data stream mining system for real-time vehicle monitoring," in *Int. SIAM Data Mining Conference*, 2003.

[8] Y. Zhang, G. Gantt *et al.*, "Connected vehicle diagnostics and prognostics, concept, and initial practice," *IEEE Transactions on Reliability*, vol. 58, no. 2, 2009.

[9] S. Byttner, T. Rögnvaldsson, and M. Svensson, "Consensus self-organized models for fault detection (COSMO)," *Engineering Applications of Artificial Intelligence*, vol. 24, no. 5, pp. 833–839, 2011.

[10] R. Prytz, S. Nowaczyk, and S. Byttner, "Towards relation discovery for diagnostics," in *Proceedings of the First International Workshop on Data Mining for Service and Maintenance*. ACM, 2011, pp. 23–27.

[11] S. D'Silva, "Diagnostics based on the statistical correlation of sensors," Society of Automotive Engineers (SAE), Tech. Rep., 2008.

[12] G. Vachkov, "Intelligent data analysis for performance evaluation and fault diagnosis in complex systems," in *IEEE International Conference on Fuzzy Systems*, July 2006, pp. 6322–6329.

[13] J. Lacaille and E. Come, "Visual mining and statistics for turbofan engine fleet," in *IEEE Aerospace Conf.*, 2011.

[14] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.

[15] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed. Prentice Hall Series in AI, 2003.

[16] G. M. Weiss, "Mining with rarity: a unifying framework," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 7–19, Jun. 2004. [Online]. Available: http://doi.acm.org/10.1145/1007730.1007734

[17] K. Napierala and J. Stefanowski, "Bracid: a comprehensive approach to learning rules from imbalanced data," *Journal of Intelligent Information Systems*, vol. 39, no. 2, pp. 335–373, 2012.

[18] J. Stefanowski, "Overlapping, rare examples and class decomposition in learning classifiers from imbalanced data," in *Emerging Paradigms in Machine Learning*, vol. 13. Springer, 2013, pp. 277–306.

[19] M. Kuhn, "Building predictive models in R using the `caret` package," *Journal of Statistical Software*, vol. 28, no. 5, pp. 1–26, 2008.