TRA105 - Data-driven Product Realization
Spare Parts, Forecasting, Clustering, Machine Learning & advanced analytics

K. Dorner, R. Maruthappan, A. Piastowski, J. Sintes
Chalmers University of Technology
Gothenburg, Sweden

ABSTRACT

Volvo Cars has previously improved on its new spare parts forecasting through implementation of artificial intelligence (AI) based classifier models. This paper builds on prior efforts by evaluating these, and suggesting improvements to the data preparation and performance measures. A simplified approach towards building the dataset is introduced, while expanding the datasets' features among others with indicators of the part's function and physical location. The F1 score was introduced as a more correct measure of prediction accuracy and precision. Based on the improvements made, several classifier models were benchmarked on monthly and semestral period datasets. Improvements in F1 scores for these periods of 73% and 76% respectively have been achieved when compared to prior approaches.

Keywords: spare parts, demand forecasting, installed base data, classifier, random forest, binary classification, clustering

1. INTRODUCTION

Spare parts planning is highly challenging through infrequent and intermittent demand, that is characterized by high variance in demand magnitude and intervals, often with large periods without demand [1]. Volvo Cars distributes spare parts from three central warehouses in Sweden, United States, and China, to 51 regional warehouses. Therefore, it is of high interest for the company to have an accurate and dependable forecast for efficient parts distribution to its regional warehouses. This is important not only due to distribution lead-times, which can vary highly between regions, but also for efficient inventory management that avoids high cost of dead stock as well as inefficient transportation planning. With 90.000 part numbers and a stock value of 3 billion SEK, improvements to the current forecasting approach of Volvo Cars can create significant impact for the company. Therefore, the company was interested in seeing whether there are other ways to group parts and forecast demand, which would improve their forecasting accuracy.

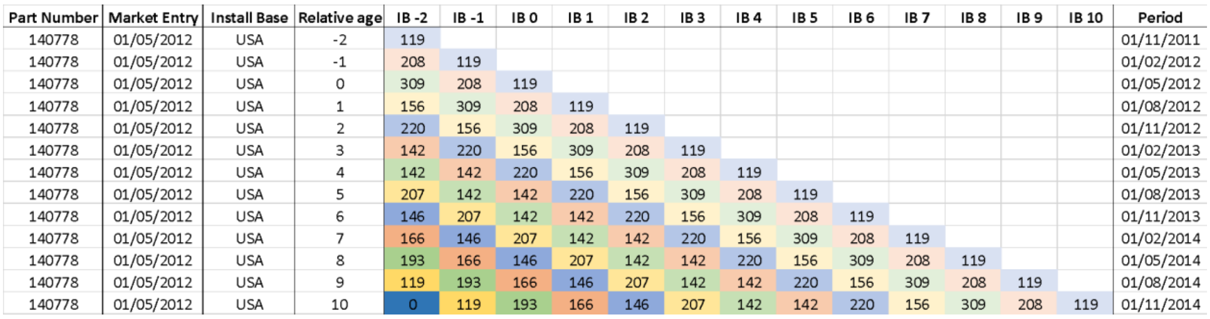
Volvo Cars' journey to improve their traditional time-series forecasting approach began with the master's thesis project [2], which explored machine learning (ML) based forecasting of new spare parts in the US market, using a classifier model. This effort used installed base data as a predictor towards demand, since this was described as an effective approach [2, 3]. Furthermore, in later stages of the spare parts' sales lifecycle, a hybrid approach was implemented which added the time-series forecast as a feature to the classifier model. These approaches have proved to be highly successful in improving prediction accuracy, therefore Volvo Cars have implemented them while also seeking further improvements. As parts have a parameter indicating grouping respective to the parts' function, called function group, the company chose to train a separate machine learning model for each function group. This has proved problematic due to the lack of data for certain groups.

This paper presents the improvements made towards new spare parts forecasting at Volvo Cars, which have been conducted during the second learning period of academic year 2021-2022. Hereby, the process and results are presented, which were obtained through the course project in TRA105: Data-driven Product Realization at Chalmers University of Technology. These results contain the evaluation of past efforts, introduction of F1 as a new performance measure, implementation of a new approach towards preparing data for the model, as well as benchmarking of several classifiers with the new approaches implemented. In the following section, the past efforts are described in more detail, followed by the approaches used in this project. Thereafter, the results and comparisons of new approaches are presented, followed by a discussion regarding these. Lastly, the paper is summarized under Conclusions.

2. FRAME OF REFERENCE

Initially, Volvo Cars was forecasting demand for spare parts using time-series, namely an exponentially smoothed moving average (EMA) [2]. This approach was of a reactive nature, hence sudden peaks in demand introduced assumption of future sales, without insight regarding the part properties. [3] have likewise described how such traditional approaches are not appropriate for intermittent demand. Therefore, [2] has set out to examine forecasting approaches using installed base data, whereby the number of parts of a certain age was obtained from the Volvo Cars' parts data generated during assembly of new cars. Herewith, the initial launch phase of new spare parts was targeted, which constituted 0-24 months from the installation time of a new part. The authors have further split this period in three stages, where Stage 1 represents the time before any sales occur for a certain part. At this stage it is of interest whether the part will sell in the following 6 months, in which case, Volvo Cars would like to stock the part at the respective regional warehouse in anticipation of sales. Stage 2 represents the period of low sales, which is the time between the first sale and the point where there have been a total of 3 sales periods for a part, which are not required to be consecutive periods. Here it is of interest to predict the quantity of sales for the following 3 months. Lastly, Stage 3 represents parts with relatively established sales, since this period consists of the time after Stage 2 until the part has reached 24 months since the product launch. In this period, it is of interest to predict the sales quantity for the following month.

When preparing the dataset, [2] used the installed base data of spare parts in the North American market, composed of the United States and Canada. In this dataset, for each part number and each period, the quantity of parts with a certain relative age was included in the respective bin, in a structure similar to Table 1. Hereby, as the current period progresses, each relative age bin "IB x" gets populated in a cascading style, where x represents the age in periods relative to the current period of a row. So, as seen in Table 1, in the last row that represents the most recent period, each bin of relative age is populated with the number of parts in each respective age group.



Part Number	Market Entry	Install Base	Relative age	IB -2	IB -1	IB 0	IB 1	IB 2	IB 3	IB 4	IB 5	IB 6	IB 7	IB 8	IB 9	IB 10	Period
140778	01/05/2012	USA	-2	119													01/11/2011
140778	01/05/2012	USA	-1	208	119												01/02/2012
140778	01/05/2012	USA	0	309	208	119											01/05/2012
140778	01/05/2012	USA	1	156	309	208	119										01/08/2012
140778	01/05/2012	USA	2	220	156	309	208	119									01/11/2012
140778	01/05/2012	USA	3	142	220	156	309	208	119								01/02/2013
140778	01/05/2012	USA	4	142	142	220	156	309	208	119							01/05/2013
140778	01/05/2012	USA	5	207	142	142	220	156	309	208	119						01/08/2013
140778	01/05/2012	USA	6	146	207	142	142	220	156	309	208	119					01/11/2013
140778	01/05/2012	USA	7	166	146	207	142	142	220	156	309	208	119				01/02/2014
140778	01/05/2012	USA	8	193	166	146	207	142	142	220	156	309	208	119			01/05/2014
140778	01/05/2012	USA	9	119	193	166	146	207	142	142	220	156	309	208	119		01/08/2014
140778	01/05/2012	USA	10	0	119	193	166	146	207	142	142	220	156	309	208	119	01/11/2014

Table 1. Dataset structure of initial clustering approach

In Stage 1, [2] have trained a CatBoost classifier model with the respective dataset. For Stage 2 and Stage 3, a combined approach has proved most successful. In this approach, [2] have included the EMA forecast of each part number to each respective period of the dataset, whereby the EMA forecast became an additional predictor feature for the classifier model. While these approaches included little information regarding the spare parts, they still managed to reach significant improvements when compared to the original forecasting approach used by Volvo Cars, as seen in [2].

Through these results, the company had a proof of concept regarding installed base forecasting, which was thereafter improved with additional data regarding the parts. Having access to a parameter that shows what functional group each part belongs to, Volvo Cars have decided to employ the general approach of [2]. However, the spare parts data was split for each regional market and each function group, which aided in more specific forecasting through additional data. Nonetheless, due to some function groups having insufficient data, this approach proved problematic. Through these efforts, a solid foundation was created for the improvements presented in this paper, which would have been difficult to achieve without this prior knowledge.

3. METHODS

3.1. Framework based on Crisp-DM

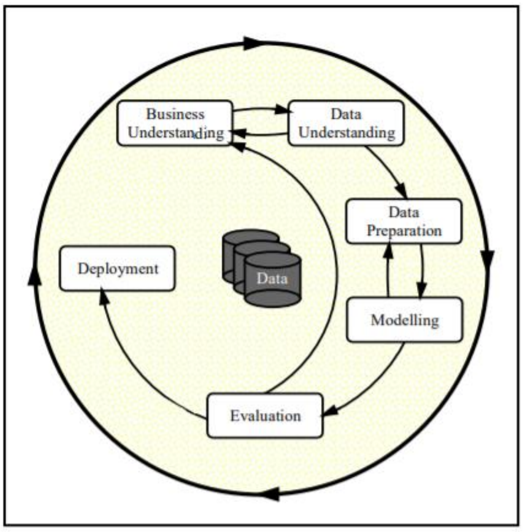


Figure 1. Phases of the Crisp-DM process for data mining [4, pg. 5]

The cross-industry process for data mining (CRISP-DM) model provides a complete overview of a data mining project's life cycle. It contains the phases of a project, as well as their associated tasks and outcomes [4]. The life cycle of the project can be divided into 6 phases as shown in Figure 1. The arrows which connect the phases show the importance and dependency between the phases. The outer circle in Figure 1, shows that data mining is a cyclical process, as the lessons learned from each phase results in more or new business questions [4].

The initial phase of the Crisp-DM focuses on the importance of project objectives and the requirements of the business perspective. This is then converted into the data mining problem and also the initial project plan is developed to achieve the required objective [4]. During our initial meeting with the stakeholders from Volvo Cars, the background of the project, business objective and the goal of the project was well defined. There was also a quick description of the many resources that were accessible and might be utilized as part of our project. Finally, during the business understanding process, there were discussions regarding potential project constraints, one of which being the limited time due to the short duration of approximately 8 weeks to complete the project.

The second phase of Crisp-DM, Data Understanding, is where the required data is collected and the activities for familiarizing with the data are conducted. To find some correlation with the data, or to form hypotheses for hidden information [4]. The clustering of spare parts project included three different master datasets: one that contained part properties for all of the parts, one that contained sales data for the parts since 2007, and one that contained all of the necessary information for the parts, as well as the information for each spare part that was transported to respective country warehouses. The total quantity delivered, total sales per region, and the pareto principle, among others, were analysed by analysing the various datasets. Some insights emerged from these many actions, which aided in the formulation of a new dataset for model preparation. The process of understanding the data led to deeper formulation of the data mining problem and the project plan [4].

The third phase included the data preparation, where a final dataset was prepared which runs on the modelling tool. The phase consists of data cleaning, construction of new attributes and transformation of data for the modelling tool [4]. As previously stated, the data interpretation process aided in the creation of a new dataset that comprised sales at various places (country, district), as well as the geographical location of spare parts. In order to reduce the model's error, data cleaning was performed to remove NaN values from the raw data.

Various modelling techniques are chosen and applied throughout the modelling phase, and the parameters are calibrated to produce the best result [4]. Several comparisons were made between the various architectures. The two datasets (semestral dataset and monthly dataset) were evaluated in several models, and the findings were compared to those obtained by Volvo Cars. Due to lack of time for the remaining elements of the model, we decided to keep the detailed evaluation and development aspects of the model for future research.

To conclude, Crisp-DM has proven to be a highly beneficial model for guiding us through the entirety of our project work. The Crisp-DM various phases helped in prioritizing the tasks that needed to be completed to achieve the desired objective.

3.2. Data Analysis

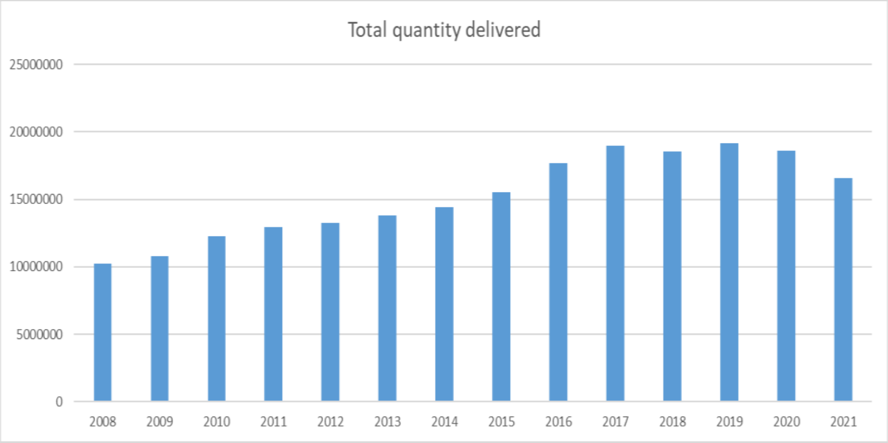


Figure 2. Total spare part quantities delivered

During the data analysis approach, there were certain trends in the dataset. As seen in Figure 2, from 2008 through 2019, the overall quantity delivered has increased linearly. The number of deliveries has decreased since 2019. One such factor could be the impact of covid-19; since vehicles were likely less used, there may have been reduced wear and tear on the consumers' automobiles.

Furthermore, the Pareto principle was examined for the initial set of data to understand the behaviour of how parts are purchased as shown in Figure 3. According to the pareto principle, 20% of the parts are responsible for 80% of the spare parts. In the case of the Volvo Cars sales, it’s apparent that a few part numbers alone already constitute 80% of the total sales. In fact, 3.6% of parts account for 80% of sales. Thus, it's important for often selling parts to be accurately forecasted.

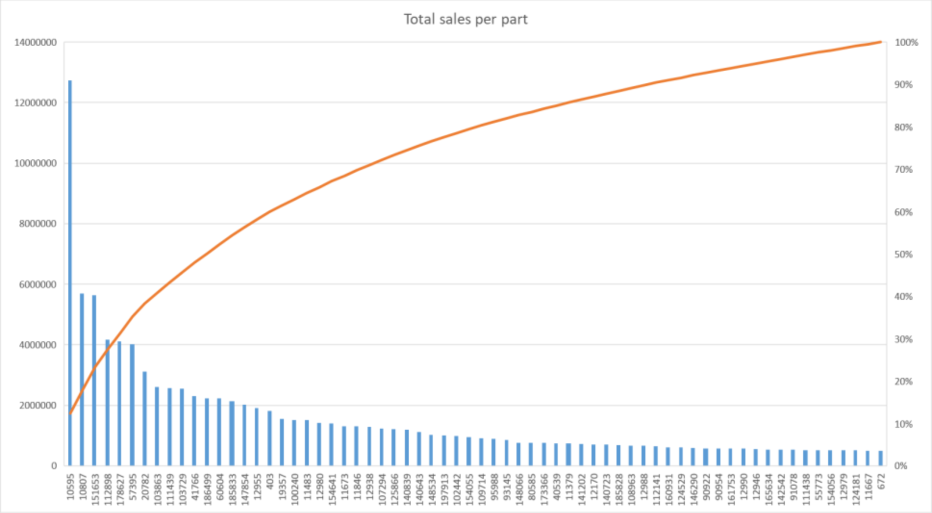
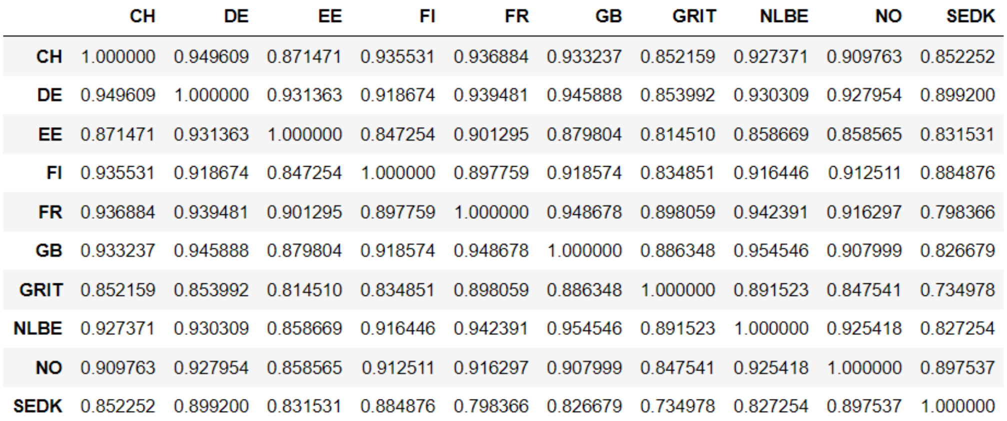


Figure 3. Pareto diagram

Finally, to see the correlation of demands between countries, a correlation matrix was calculated based on the initial data, as shown in Figure 4. This helps to see if there are any relations between parts sales behaviours between different countries. The correlation coefficients had a rather high value, as almost all the countries had correlation between 0.73 - 1.00. This shows that sales behaviours between countries are highly similar.



	CH	DE	EE	FI	FR	GB	GRIT	NLBE	NO	SEDK
CH	1.000000	0.949609	0.871471	0.935531	0.936884	0.933237	0.852159	0.927371	0.909763	0.852252
DE	0.949609	1.000000	0.931363	0.918674	0.939481	0.945888	0.853992	0.930309	0.927954	0.899200
EE	0.871471	0.931363	1.000000	0.847254	0.901295	0.879804	0.814510	0.858669	0.858565	0.831531
FI	0.935531	0.918674	0.847254	1.000000	0.897759	0.918574	0.834851	0.916446	0.912511	0.884876
FR	0.936884	0.939481	0.901295	0.897759	1.000000	0.948678	0.898059	0.942391	0.916297	0.798366
GB	0.933237	0.945888	0.879804	0.918574	0.948678	1.000000	0.886348	0.954546	0.907999	0.826679
GRIT	0.852159	0.853992	0.814510	0.834851	0.898059	0.886348	1.000000	0.891523	0.847541	0.734978
NLBE	0.927371	0.930309	0.858669	0.916446	0.942391	0.954546	0.891523	1.000000	0.925418	0.827254
NO	0.909763	0.927954	0.858565	0.912511	0.916297	0.907999	0.847541	0.925418	1.000000	0.897537
SEDK	0.852252	0.899200	0.831531	0.884876	0.798366	0.826679	0.734978	0.827254	0.897537	1.000000

Figure 4. Correlation Matrix

3.3. Evaluation of approach used by the company

Having limited time, which might not be enough for the whole process of new model development, we've decided to evaluate the approach and the model used by Volvo. This could have been done in parallel with our model development and could benefit us with a better understanding of the problem and present issues which we might come across. First thing that caught our interest was the amount of trained models as for each function group a separate model was trained, which resulted in over 2000 models being prepared. At first sight it seems like a proper approach but after deeper investigation having such a large amount of models sometimes results in having really small datasets for training due to the fact that a small amount of selling records in specific function groups. Another issue that was not considered before is the falsity of the results. Volvo's model was claimed to achieve accuracy at around 80% but it was not checked how imbalanced the distribution of the predicted classes is. Therefore, we introduced a new measure of performance called F1 score, which can be calculated using Equation 3.

$$Precision \ = \frac{True\ Positives}{True\ Positives + False\ Positives} \ (1); \qquad Recall \ = \frac{True\ Positives}{True\ Positives + False\ Negatives} \ (2)$$

$$F1 \ = \frac{2 * Precision * Recall}{Precision + Recall} \ (3)$$

Precision as presented in Equation 1, describes the percentage of correctly classified positives from all classifications of positives, whereas recall, obtained according to Equation 2, answers the questions of what part of all possible positives were classified as positive. The F1 score, obtained from Equation 3, is a harmonic mean of Precision and Recall. We tried out Volvo's model and checked its average performance over a few runs. To benefit the metrics, we've excluded function groups which did not have enough parts for testing so that the results were more reliable. During the experiment, we've achieved the claimed 80% of accuracy, to be exact 79,86% but the F1 score was equal to 0,479 which is a rather poor result. These results showed that this model was biased towards predicting that a part will not be sold, as the majority of the classes in the dataset were negative. Such an approach led to achieving 80% of accuracy but was not really trying to answer the asked question. What is more, we found out that the way the data was prepared resulted in having a lot of empty columns just to convey the information about exactly how many parts were installed in a particular quarter. Getting to know that, we were able to look for a different approach, since we believed this could be improved.

3.4. New approach and data-preparation

According to the evaluation of previous modelling, improvements on the method are largely possible at every step of the process. The modelling can be improved, but most important is the data-preparation and how the dataset is built. The previous approach was answering the question "Will these spare parts sell in the next 6 months?". Thus, it does not include the location of the part. Since this work is for logistic purposes, the location at different possible scales such as market (group of countries), country, and district is highly important and needs to be considered in the prediction of sales. Hence, our formulation of the problem is slightly different, and the question we are trying to answer with our model is: "Will a specific part sell in the following period in a given country or market?*"*

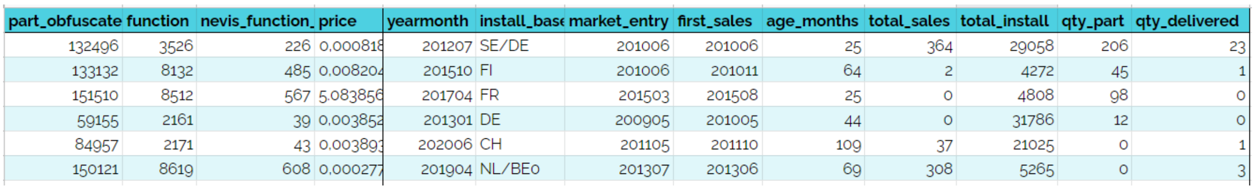
Since the previous dataset does not include any geographical information about the spare parts, we had to build a new dataset from the raw data (part general data, install base and sales). Moreover, the previous dataset includes for each part the whole past install base data, which makes it redundant and difficult to process with many NaN (undefined) values. This has led us to build a new dataset upon different key ideas:

- Simpler and less redundant architecture
- Add descriptive features of parts
- Add location data at the market scale (country / group of countries)
- Have both global historical and current data as features to keep track of sales behaviour

Having these key ideas, we were able to build a new dataset from the raw data. Since the data comes from 3 different databases, we had to make different assumptions. The most important is the consideration that the 3 datasets (part master, install base, sales) are coherent, which means that dates and parts information correspond with each other. We also decided to use the different markets as our location information which means that each part is located by the group of warehouses where it was sold. As in the previous approach, the market entry date is defined as the first install base data available for each part. Finally, we considered that no data on *install base* or *sales* for a given month or semester means no installed parts or sales.

Following the key ideas and using these assumptions, we've built a dataset that needed to be cleaned up in different ways. Firstly, one of the main problems was about market entry date. In fact, install base data and sales start in different years, namely in 2007 and 2010, and the collection of data could have been unreliable in the early years. Thus, often the first sales of a given spare part occur several years before the market entry date (computed according to our assumption), therefore we needed to remove all the parts where sales and market entry date mismatched. Fortunately, this cleaning had no major impact on our dataset since it represented a few parts that are largely sold. For these parts, the sales do not need to be predicted, since they are likely not new spare parts, but ones that have been sold for several years. Secondly, we use the nevis function group as a descriptive feature of parts. Nevis function group can be described as the function on the car in which the piece is involved, which is related to the physical location of the part on the car. It often happens that parts are involved in several nevis function groups, even up to 30, and thus the feature in our dataset is a list of these functional groups. So in order to process the data, we needed to encode the combinations of the nevis function group for each part, such that each combination became a distinct category. This particular encoding might be improved to keep track of the most important function group or close combinations of function groups.

Finally, the final dataset can be built with different time granularity. We considered in our work a monthly based dataset, which can be seen in Table 2, and a semestral one. The dataset includes 12 features and aims at predicting the sales in the next period. It includes more than 17 000 000 entries with monthly granularity.



part_obfuscate	function	nevis_function_	price	yearmonth	install_base	market_entry	first_sales	age_months	total_sales	total_install	qty_part	qty_delivered
132496	3526	226	0,000818	201207	SE/DE	201006	201006	25	364	29058	206	23
133132	8132	485	0,008204	201510	FI	201006	201011	64	2	4272	45	1
151510	8512	567	5,083856	201704	FR	201503	201508	25	0	4808	98	0
59155	2161	39	0,003852	201301	DE	200905	201005	44	0	31786	12	0
84957	2171	43	0,003893	202006	CH	201105	201110	109	37	21025	0	1
150121	8619	608	0,000277	201904	NL/BE0	201307	201306	69	308	5265	0	3

Table 2. Sample of the monthly dataset

4. RESULTS

4.1. Classifier comparison

Since the dataset prepared by the company consisted only of numerical values describing the amounts of installed parts, there were no real limitations towards selecting model architecture. To realize how complex the problem was, we ran different model architectures on the dataset so that we had an idea which has a tendency of finding good correlations for this particular problem and on which architectures we should focus on. In the comparison we also included the CatBoost architecture used by the company.

Classifier \ Metric	**CatBoost**	**SVM**	**Decision Tree**	**Logistic regression**	**Naive Bayes**	**SGD**	**KNN**	**Random Forest**	**XGBoost**
Accuracy	0,799	0,781	0,760	0,768	0,716	0,754	0,774	0,790	0,785
F1 score	0,480	0,370	0,496	0,373	0,377	0,385	0,461	0,504	0,506

Table 3. Comparison of different classifiers on the dataset provided by Volvo Cars

In the comparison of 9 different architectures, which can be seen in Table 3, the best results were achieved for Random Forest and XGBoost. What is worth mentioning is that CatBoost and Decision Tree also presented quite similar results, and this could have been expected as Random Forest is based on Decision Tree and CatBoost along with XGBoost is based on Random Forest.

The process of building a decision tree is based on a selected impurity function like ‘gini’, which is often chosen. Firstly, it is checked how well each feature alone can predict a decision and appropriate coefficient is being calculated. The feature with the lowest coefficient (if using the gini function) is chosen as a root node. Each node can give two answers so in case of categorical data like “Gender is male” it can divide the dataset for True or False and in case of numerical data it sets some threshold for differentiation. Subsequently, impurity coefficients are calculated repeatedly for each new division made by the previous node which then becomes a factor for selecting the next node.

Random Forest is a classifier made of many decision trees, where each one is trained on a randomly selected subset of the dataset, with the possibility of getting duplicate rows. Another thing is that each decision tree selects the best feature out of randomly selected two different features (it is controlled by setting hyperparameters). Thanks to that, every tree can choose different features as the most relevant, leading to different decisions. Depending on the size of the dataset, the appropriate amount of trees is being made. Each prediction is run though all the decision trees and the class with the most votes is chosen. This approach gives a much better generalization of solving the problems as each trained tree had a little different dataset and had to find the best way of classification with bootstrapped dataset.

There is a substantial difference between the dataset prepared by Volvo Cars and ours, which is due to the fact that we’ve decided to have categorical features and not all of the classifiers accept these. Which is also why we decided to choose the classifiers built on decision trees as they not only accept those but can efficiently operate with them.

4.2. Performance analysis

Selected classifiers were trained on both types of prepared datasets. The anticipated findings on the monthly dataset, as shown in Table 4, were good, since XGBoost and CatBoost both and Random Forest had around 92.8 percent accuracy. Random forest and CatBoost both scored 0.8305 and 0.8292 on the F1 scale, respectively. XGBoost's F1 score was 0.8281, which is also really close to other results.

Classifier \ Metric	**Random Forest**	**CatBoost**	**XGBoost**	**Simple DNN**
Accuracy	0.9289	0.9287	0.9288	0.8649
F1 score	0.8305	0.8292	0.8281	0.5527

Table 4. Results on monthly dataset

There was a minor gain in the F1 score when the model was applied to the semestral dataset, as seen in Table 5 but a bit more significant decrease in accuracy. Semi-annual datasets provide more general information with each row. In this case the F1 score is really close to accuracy, which assures us that the model is making a so called conscious decisions. Comparing our results with Volvo Cars' approach, there is a significant improvement. When compared to Table 3, the CatBoost model which was analysed on the dataset of Volvo Cars had an accuracy of 79.86 percent and F1 score of 0.4794. There is a substantial difference between the accuracy and F1 score obtained from the semestral dataset model and the monthly dataset & Volvo model, with a 73% and 76% improvement respectively.

Classifier \ Metric	XGBoost	Random Forest	CatBoost	Simple DNN
Accuracy	0.8866	0.8863	0.8842	0.7557
F1 score	0.8421	0.846	0.8386	0.6616

Table 5. Results on semestral dataset

All the trained models were run with many parameter configurations so that they could achieve the best possible scores but more often than not the improvement was not significant. This showed that classifiers tested by us are well optimized by default. Nonetheless, it is hard to say which configuration will perform the best on the future forecasts, however what is worth noting, is that over all runs the difference in accuracy and F1 score was not more than 0.01 which proves the consistent performance of the models.

5. DISCUSSION

5.1. *Time limitation*

Even though we would love to do each experiment with a proper preparation and focus it was simply not possible because of the time limitation. Developing a whole new approach with an AI model sometimes can take over a year, and in this case we had only a few weeks. This made us make some shortcuts in the whole process of model development and we are aware of some lacking standard steps. For example, the entire process of literature review was omitted, which is a typical step before the start of developing a new approach. We felt that with such a time limit it would cost us a lot of time and might not be that beneficial especially as the problem of the project was rather specific. Therefore, we've decided to use the resources given by Volvo Cars and the support we've received from the course. This as well as other decisions had to be made based on our intuition and experience as this is what was needed to achieve any satisfying results.

5.2. *Model evaluation*

It can be easily seen that the performance of different architectures is really similar, which is expected due to the fact that all of them base their logic on Random Forest. The difference is that XGBoost and CatBoost are expected to have a shorter time of training and are said to cope better with unbalanced datasets. Another aspect worth mentioning is that these architectures are well optimized by default and our tweaking of parameters improved the results only slightly. Nonetheless, we still believe that each fraction of improved performance is worth implementing. What has to be borne in mind is that the accuracy metric is not always a great indicator, as the main focus should be on generalization of the model so that it can benefit the company in all circumstances especially when dealing with a strongly unbalanced dataset. Having experience of training all of these architectures with different parameters on two types of datasets we believe that the most promising model is XGBoost. It constantly delivered good results and most of the time gives a slight edge over the other models. Moreover, this classifier is GPU compatible, which reduces computing times by an order of magnitudes, allowing for more runs.

5.3. Expanding and validating the method with North American market

The current approach is only focused on the European market. However, to better evaluate the model, it should be tested also on the North American data used in [2]. Moreover, our model should be trained and tested on the data available around the world to check the accuracy of the model in all markets.

Furthermore, it could be beneficial to verify the data with distinct warehouses instead of countries, in order to improve the prediction precision. This can assist in the movement of parts to the specific warehouses, thus the lead time could be further reduced. Finally, a comparison test for the intended model is recommended, because it can aid in understanding the model developed as well as having distinct models for different countries, which can increase the accuracy of each country's individual model.

5.4. Further evolution of the data preparation and modelling

The new approach developed with this project seems to be quite efficient and there are wide possibilities to improve it and actually develop a data-driven product. Firstly, as mentioned previously, modelling might be still slightly improved by parameter tweaking. Secondly, new algorithms could be tested to improve the quality of our model. However, it seems that larger improvement can be achieved by improving the data preparation as well as data cleaning.

During the project we worked with two data-sets: one was on a monthly basis and one on semestral basis. So the time granularity was different and we obtained different results. The quantity prediction by semester was pretty accurate but the granularity might be too large for logistic purposes. On the other hand, the modelling based on the monthly dataset has a little more accurate predictions but at a shorter time range. Thus, it could be interesting to build a dataset with intermediate time granularity, for example on a 3-months basis. The time step would be more optimal for logistic purposes without losing too much quality on prediction.

The sales of spare parts are directly linked with the type of part itself. Obviously, some parts have to be changed more often than others (breaks, air filter, etc…). Thus having more descriptive features for the parts could lead to great improvements in the prediction. Currently, the model is mainly based on sales and install base historical data and there are only a few descriptive features for the part, but coupling descriptive features to historical data in the data preparation could be very powerful for predicting sales.

Moreover, the main goal of the spare part sales prediction is for logistic purposes. Therefore, knowing precisely where a specific part will be needed is crucial. During this project, we've decided to work with the different markets in Europe (country/group of countries) and this has had great results. However, there could be a lack of precision, and maybe, more precise prediction would be needed. Thus, the dataset could be rebuilt using more precise location features such as the warehouses or the district.

6. CONCLUSIONS

In this paper, several improvements were presented to the new spare parts forecasting of Volvo Cars. First, the past efforts by the company and a master project have been presented, which formed the starting point of this project. Thereafter, the framework used for conducting the project and the project's approach were presented. An exploratory data analysis was conducted on the received datasets, after which the prior approaches used by Volvo Cars have been evaluated. One main improvement of the project to previous efforts was the introduction of the F1 score as a performance measure, which is a more correct approach towards measure of accuracy and precision.

Furthermore, a new approach towards the process of data preparation has been introduced, which not only simplifies the prior approach but also achieves significant improvements in performance. These improvements were also facilitated by inclusion of additional features to the dataset, including a part's function group, nevis function group besides other changes to the approach regarding data preparation. Volvo Cars' prior approach was evaluated with multiple classifiers, against which the new approach has been benchmarked on both monthly and semestral periods in the dataset. Improvements in F1 scores for these periods of 73% and 76% respectively have been achieved.

REFERENCES

[1] Andersson, J. (2019). Enhancing aftermarket demand planning with product-in-use data. *Licentiate thesis*, Department of Technology of Management and Economics, Chalmers University of Technology

[2] Lundh, A. and M. Marklund (2020). Forecasting Initial Phase Spare Part Demand Using Installed Base Data. *Master's thesis in Supply Chain Management.* Department of Technology of Management and Economics, Chalmers University of Technology

[3] Andersson, J. and P. Jonsson (2018). Big data in spare parts supply chains: The potential of using product-in-use data in aftermarket demand planning, *International Journal of Physical Distribution & Logistics Management*, Vol. **48,** Issue: 5, pp.524-544

[4] Wirth and Hipp, (2000). A process model for data mining—CRISP-DM. *Customer and Business Analytics*, 43–56. https://doi.org/10.1201/b12040-8