IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Data-driven estimation of heavy-truck residual value at the buy-back

TANIA CERQUITELLI[1](Member, IEEE), ANDREA REGALIA[2], EMANUELE MANFREDI[2], FABRIZIO CONICELLA[3], PAOLO BETHAZ[1], ELENA LIORE[2],

[1]Department of Control and Computer engineering, Politecnico di Torino, Turin, Italy; (e-mail: name.surname@polito.it)
[2]Accenture S.p.A. (e-mail: name.surname@accenture.com)
[3]CNH Industrial (e-mail: name.surname@cnhind.com)

Corresponding author: Tania Cerquitelli (e-mail: tania.cerquitelli@polito.it).

ABSTRACT In a context of deep transformation of the entire automotive industry, starting from pervasive and native connectivity, commercial vehicles (heavy, light, and buses) are generating and transmitting much more data than passenger cars, with a much higher expected value, motivated by the higher costs of the vehicles and their added-value related businesses, such as logistics, freight, and transportation management. This paper presents a data-driven and unsupervised methodology to provide a descriptive model assessing the residual value estimates of heavy trucks subject to buy-back. A huge amount of telematics data characterizing the actual usage of commercial vehicles is jointly analyzed with different external conditions (e.g., altimetry), affecting the truck's performance to estimate the devaluation of the vehicle at the buy-back. The proposed approach has been evaluated on a large set of real-world heavy trucks to demonstrate its effectiveness in correctly assessing the real status of wear and residual value at the end of leasing contracts, to provide a few and quantitative insights through an informative, interactive and user-friendly dashboard to make a proper decision on the next business strategies to be adopted. The proposed solution has already been deployed by a private company within its data analytics services since (1) an interpretable descriptive model of the main factors/parameters and corresponding weights affecting the residual value is provided and (2) the experimental results confirmed the promising outcomes of the proposed data-driven methodology.

INDEX TERMS Business vs data-driven methodologies, Automotive industry, Commercial Vehicles, Residual value estimation.

I. INTRODUCTION

Many companies today are striving to become data-driven. This term refers to companies that are able to make strategic decisions based on data analysis and interpretation. A data-driven approach enables them to examine and organise their data with the goal of better serving their customers and consumers.

Data can offer a huge additional value, but must be appropriately analyzed with a robust business objective. Data science supports the business, and offers greater benefits when the data collected are properly filtered and aggregated. In fact sometimes data is collected without having a precise target: often in this case it turns out that the large amount of data collected is useless, and the investment in terms of hardware and software is not effective for the Company.

In the context of data science to support business, it is increasingly common in the automotive industry to install sensors on vehicles to constantly monitor mission profile and use, considering both internal and external parameters, the first are generated directly from the truck (e.g average speed, fuel consumption), the second are independent of the truck (e.g. altimetry). However, both sets of parameters affect the performance of the trucks and have to be jointly considered to estimate the truck's wear and corresponding residual value at the buy-back. Before the introduction of the Internet of Things paradigm, data in vehicles was collected and processed locally by the Engine Control Unit for routine operations managing and monitoring vehicle operativity. However, the limited ECU (Engine Control Unit) and storage capabilities of vehicles made the collection and the analysis of past data infeasible. The situation changes with the introduction of connected vehicles, like explained in [1] and [2], where it is presented a comprehensive framework of Internet of Vehicles (IoV) with emphasis on layered architecture, protocol stack and network model. In connected vehicles, data can be locally collected and sent to a remote storage

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2998940, IEEE Access

location (cloud) for later analyses with better performing tools. In this way, automotive manufacturers can use the data collected to offer additional value to their customers, both in terms of more transparency and additional services.

In this automotive context, there are three important objectives that can be supported by data-driven methodologies: *(1) predictive maintenance*, that aims at the identification of possible malfunctions ahead of time, allowing a prompt intervention before the actual failure; *(2) Remaining Useful Life (RUL)*, used more in the context of industry 4.0 rather than automotive, it estimates the remaining life time of a machine; *(3) Residual Value*, which is based on vehicle degradation as well as remaining life estimates, but the outcome is the estimate of the residual value. The latter is very important to make business decisions to target proper marketing strategies to improve the business revenue.

Focusing on this last point, estimating the residual value of a vehicle has been a challenge already addressed in some research activities. This is for example what happens in [3], where the authors propose an approach based on machine learning techniques in order to estimate the residual value of a car. And similar works are also done in [4] and [5], where although the used machine learning algorithms are different, the data that the authors use in order to estimate the residual value are very similar and traditional (such as model, mileage in km and year of manufacture). Whatever, all these works were always carried out considering cars, not heavier vehicles, such as heavy trucks, which are usually equipped with more sensors capable of monitoring additional parameters thus providing more challenging opportunities to design data-driven methodologies. Furthermore, the added-value business to heavy-trucks and their costs higher than cars increases the importance to address the research issue to estimate the residual value and wear of the heavy trucks. In [6] for example, authors present a new service paradigm for Vehicle Communication Platforms (VCPs) based on the Sensor Cloud concept, able to support real-time Intelligent Truck Monitoring (ITM) services on about one thousand tank trucks for fuel distribution in Europe. Moreover, due to the high cost of heavy trucks, one of the most common ways of contracting with this type of vehicle is leasing; and estimating the residual value can be particularly useful in this kind of contract, for both the company and the customer. In fact, when the contract expires, the residual value indicates the level of degradation of the vehicle and the consequent necessary maintenance actions that the company must do before signing another contract. This value could be also useful for third-part companies responsible of the perizia, in fact they can have a more detailed look at a truck that is in bad conditions, according to what has arisen from telematics data.

So, what we aim to do in this article is to offer an unsupervised and data-driven methodology capable of accurately evaluating residual value of heavy trucks, taking advantage of the benefits coming from a proper use of the vehicles monitored through telematics data, after their manipulation, statistical, and data mining analysis. The methodology proposed in this work offers two different formulas to estimate the residual value, a data-driven one and a business-driven one. Both are based on the same parameters collected by the same sensors, but the first gives greater importance to the punctual values of these parameters, while the second considers different business weights for each parameter depending on its importance. The results of these formulas were evaluated on real data collected by a multinational corporation, which has installed a black box on its trucks that is able to constantly monitor their use, considering both external (environmental characteristics and actual conditions of the travelled roads) and internal (directly related to the effective usage such as fuel consumption, loaded weights, hours of usage) parameters.

The paper is organised as follows. Section II introduces the use case we are going to analyse, showing how it is possible to constantly monitor heavy trucks and and how can it be useful to estimate the residual value after a period of truck's usage. Section III details the proposed methodology, introducing the two formulas and the other performed analyses, while section IV shows the results obtained by adopting the proposed methodology on a real dataset including a large set of heavy trucks. Finally, section V summarizes the findings of our study, discusses open issues, and presents possible future development and research directions of this work.

II. USE CASE DESCRIPTION

In the automotive industry, the term 'leasing' refers to a type of contract in which a company (known as 'lessee') grants the use of its own vehicle to a customer (known as 'lessor'), in exchange of a periodic monthly fee and an advance payment. The evaluation of the condition of the vehicle is very important in this case, because at the end of the contract (buy-back) the customer can decide whether to redeem it or not. Until now, this assessment is mainly based on common benchmarks (e.g. kilometers, years of purchase), but what we aimed to do in our study was to make the evaluation more accurate, taking advantage of the benefits coming from a proper use of telematics data, after their manipulation and statistical and data mining analysis. This is a key asset in the case of heavy trucks for their high costs and where a black box installed on-board constantly monitors the actual usage, considering both external and internal parameters.

To perform our study we analyzed real data collected by a multinational corporation that constantly monitors the use of their trucks through a telematics system. Since the kind of data monitored and the frequency of the transmission signals are different for each type of trucks and the application installed on-board, we focus our analysis on a single kind of heavy trucks. As these heavy trucks are very expensive, they are frequently sold through a leasing contract and for this reason it is very important to know the real conditions of the vehicle at the end of the contract. The residual value could be estimated by constantly monitoring many parameters, from the most classic ones such as the hours of usage and the fuel

consumption, to the less common ones such as the engine braking system.

We aimed at analyzing these parameters not to propose a new price for the truck (which is established by the leasing contract at the beginning and it is drastically influenced by the market dynamics), but to suggest an objective, interpretable, and data-driven evaluation/score of its status at buy-back, mainly based on its effective usage. This is particularly significant for the process of differentiation of trucks at the buy-back. Indeed, through their residual value, they can be allocated to three different categories (basic, comfort or premium) and be promoted for targeted marketing strategies. Trucks allocated to basic class are the most worn and it is not advisable to invest money on these, except for minimal interventions to ensure fundamental functionalities; those belonging to comfort class need also further technical and mechanical checks, but no aesthetic refinement. Finally, premium class, the class of the *elite* of trucks, is composed by those trucks which have few kilometers, four year of seniority at the most and for which investing money, not only for mechanical adjustment but also for improvement in aesthetic, makes sense.

Considered trucks got on-board a telematics system, that receives from sensors information relative to internal and external parameters, how the vehicle is driven and if some faults occur. The system collects with a high frequency a variety of telematics data and then sends them to a cloud platform, where data can be properly processed and analyzed in-depth.

Telematics system in this use case is able to constantly monitor the trucks, sending a lot of signals to the cloud. The application installed sends periodic messages, like GPS, temperature, odometer data and lamps signals and it is able to identify and transmit DTC (Diagnostic Trouble Code) at its occurrence. It also monitors classic parameters that can be useful in consumption optics to evaluate how the truck has been driven (like altimetry, idling time and weight), together with other more technical ones like the use of brakes, the gearbox creep time, the kick down switch time, the engine speed, the engine coolant temperature and the engine fuel temperature.

Real-time data collected through the telematics system are then integrated in the cloud with additional and relevant information available in other data sources (e.g., warranties, anagraphic data, features of the leasing contract). To carry out our study we analyzed a large dataset including more than 10,000 trucks monitored for three years. The history of each truck is characterized by 38 features, some of them are monitored over time (e.g., altimetry, fuel consumption), while others are static features (e.g., anagraphic data, origin, market and model year).

III. PROPOSED DATA-DRIVEN METHODOLOGY

Here we present the proposed data-driven methodology to fast classify heavy trucks into three main categories mainly based on the quantitative estimate of the residual value. The latter mainly relies on the analysis of telematics data monitoring effective truck usage. The proposed strategy is an unsupervised and descriptive method that assesses vehicle wear using fine-grained data collected in real time. To this aim, all supervised techniques cannot be used. The proposed methodology includes five main building blocks: (i) *Data exploration*, (ii) *Correlation analysis and feature selection*, (iii) *heavy truck residual value estimation*, (iv) *discovering worst trucks as outliers*, and (v) *informative dashboard*. A detailed description of each one is reported in the following sub-sections.

A. DATA EXPLORATION

Data exploration includes a set of preliminary analyses that allows exploring and characterizing the data. It is a very important step that aims to better design all the steps of the KDD (Knowledge Discovery from Data) pipeline. Understanding the input data allows the data scientist to make better decision on further and deeper analysis, saving time and future efforts.

The aim of this building block is to study the data distribution of numerical attributes separately. To do this, we exploited the boxplot, a method to graphically represent a numerical attribute and its variability through its quartiles. A boxplot is a graph that gives a good indication of how the data values are distributed, displaying the range in which the values are included and how they are distributed within this range. The graph obtained is based on 5 numbers, which are: *minimum*, the minimum observed value; *first quartile (Q1)*, the minimum observed value such that at least 25% of data is less than this or equal to this; *median*, the minimum observed value such that at least 50% of the data is less than this or equal to this; *third quartile (Q3)*, the minimum observed value such that at least 75% of data is less than this or equal to this; *maximum*, the maximum observed value. Boxplot is a method that allows to easily identify outliers, in fact after setting the Interquartile Range (IQR) as IQR = Q3-Q1, all the values less than Q1-1.5*IQR or greater than Q3+1.5*IQR are labelled as outliers. However, in our case, we used this method in line with their definition but with another purpose. In fact, we used it for finding the asymptotic value for each parameter useful to estimate the residual value. The asymptotic value is the limit value that each parameter can assume, due to mechanical constraints or intrinsic characteristics. The values found in this analysis will be subsequently verified with the engineering department's sensitivity and they will be used in the definition of the data-driven formula for the residual value.

B. CORRELATION ANALYSIS AND FEATURE SELECTION

After the first data exploration analyses, a second step of analysis allows us to study how the various numerical attributes are related to each other. Two strongly correlated attributes in fact do not bring any significant information to future analysis and for this reason it makes sense to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2998940, IEEE Access

consider only one of them. To improve the effectiveness of the next analytics steps, the proposed strategy leverages the correlation matrix to analyze the dependence between multiple parameters at the same time. For each couple of parameters (X, Y) monitored over time, the correlation coefficient is computed through the Pearson correlation, defined as $\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$, where $cov(X, Y)$ is the covariance between X and Y, σ_X is the standard deviation of X and analogously σ_Y for Y. $\rho_{X,Y}$ can assume values between -1 and 1, where +1 corresponds to perfect positive linear correlation, 0 corresponds to no linear correlation and -1 corresponds to perfect negative linear correlation.

Since the higher the absolute value of the coefficient, the stronger the correlation, for each pair of attributes with the Pearson correlation higher than 0.7 or less than -0.7, only one parameter is considered in the subsequent analytics steps. The choice of which parameter to consider and which not within a pair was made with the help of a domain expert, who has always chosen the variable that is easiest to interpret. For example, the length in km of a mission and its duration in hours are two parameters strongly correlated, thus, the domain expert selects the duration, that is the most consistent variable to measure the wear of an engine.

At the end of this step, the features of the dataset have been reduced, allowing us to work more easily with our data, which now contain only the most significant information.

C. HEAVY TRUCK RESIDUAL VALUE ESTIMATION

Once we have found the most significant variables to use in the analysis, the purpose of our work is to derive a mathematical approach to estimate the residual value of a truck. Here we discuss two formulas relying on the selected parameters: a *data-driven formula*, mainly based on theoretical life threshold for each parameter and a *business driven formula*, based on business expertise on parameter relevance.

We experimentally proved that the measured value for a given parameter, although it is a cumulative value (e.g., the number of km) over time in many cases, is a too fine-grained value to be able to estimate the contribution of the variable to the residual value. A coarse-grained estimation should be used. To this aim, for each of the selected variables, its distribution is discretized with a width-size strategy, to assign the same value for all trucks that assume values of the parameter in the corresponding bin. We proposed to divide the data distribution of each parameter into twenty ranges of equal size, in order to assign to each of them a specific score from 1 to 20. 20 represents the optimal condition (trucks with a parameter value in the range 20 represents a vehicle well-used and with a higher residual value) and 1 the worst.

1) Data-driven approach

The intuitive formula for calculating this value can be defined as:

$$score = 100 * \prod_{j=1}^{n} (1 - a_j)$$

where a_j is a value in the range between 0 and 1 representing the degradation of the truck due to the parameter j. The degradation effect of each parameter in the residual value is considered independently from the others. Furthermore, the product in the formula makes it possible to amplify the effect of each parameter in the overall estimation.

Based on this formula, the degradation relative to each parameter is calculated as the ratio between the measured value for that parameter and the asymptotic value that the parameter could assume (limit value). The formula obtained is as follows:

$$score = c * 100 * \prod_{j=1}^{n} (1 - \frac{y_j}{L_j} \frac{1}{x_j}) * (1 - \frac{obs}{12} * 0.09)$$

where c is a corrector factor that represents the min-max scaling, y_j is the measured value for the parameter j, L_j is the limit value for the parameter j and x_j is the score (from 1 to 20) for the parameter j. The factor $\frac{1}{x_j}$ has the scope of further enhance good/bad result for each single parameter, since x_j is the score obtained in it. It figures at the denominator because the best result ($x_j = 20$) is obtained when y_j is greatly far from L_j, so when their ratio is close to 0. In this case, x_j at the denominator further decreases the value of the ratio, obtaining an even better result. On the other hand, in the worst case the ratio between y_j and L_j it will be almost 1, and with $x_j = 1$ this ratio will keep the same.
The last term of the formula takes into account the fact that a truck may remain stationary in the yard for a long time. For this reason, the degradation of the truck must be taken into account even during its period of disuse. It is estimated a factor of depreciation of 9% (defined by the domain expert and it is a well-recognised estimation) each year and the *obs* parameter indicates for how long a truck is stopped in the yard in terms of months. If the value of this term is close to zero, the final result will remain almost the same; while a high number of this parameter will make the depreciation of the truck evident.

2) Business driven approach

The business driven formula takes into account a different weight for each parameter, in a range between 0 and 1. These business weights are defined thanks to the engineering department (domain experts), according to the influence of each parameter in the total degradation of the truck. For example, it has been estimated that the major impact on trucks usury and consequently on its value at buy-back is due to the fuel consumption and to the total hours of travelling, while the engine brake is not so relevant. For this reason, the number of hours and the fuel consumption are weighted α, while the engine brake is weighted β [1]. After having assigned the weights, each of them is divided by their sum, so that at

[1]The real values cannot be declared since they represent a key point in the corresponding business asset, but $\alpha > \beta$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2998940, IEEE Access

the end of this operation the total sum of the weights is 1. The formula obtained is as follows:

$$score = c * 100 * \sum_{j=1}^{n} \frac{w_j x_j}{20} * (1 - \frac{obs}{12} * 0.09)$$

where c is always a corrector factor and the last term always represents the degradation due to inactivity, as in the data-driven formula. The first term of the summation instead is a linear combination of the scores x_j relative to parameter j, weighted by the business weight w_j and divided by the maximum score, 20.

D. DISCOVERING WORST TRUCKS AS OUTLIERS

The main purpose of this analytics block is to quickly estimate the set of heavy trucks on which maintenance is required; and even if such a solution is only approximate, the results found are quite satisfactory. This can be very useful for companies because this strategy allows them to identify the worst set of trucks, that should be removed from the yard as soon as possible with ad-hoc marketing strategies.

To this aim, the building block relies on the DBSCAN algorithm [7], a density-based clustering algorithm that classifies points into core points and border points. Any point that is not recognized as either core point or border point is labelled as outlier and all these outliers correspond to the trucks with a lower residual value than the others. This algorithm is based on two parameters: a radius *eps* and a minimum number of points *minPoints*. To choose the most suitable value of minPoints, the sorted distances of every point with respect to their k-nearest neighbour are plotted starting from k=2 and increasing it of one each simulation. Comparing the various plots, we then look for the value of k after which the graphs do not change much (the situation is stable) and we choose that value as minPoints. With this value of k, we then choose the corresponding eps value by looking for an elbow point in the sorted k-dist graph, which is the graph representing the distances in descending order of each point from its neighbour k-th neighbour.

E. INFORMATIVE DASHBOARD

To make the results of this analysis more user-friendly, an interactive dashboard is delivered through the use of Power BI. The dashboard allows the selection of a specific truck (properly identified), year of production, Market and model year to instantly obtain its residual value and the values of the parameters collected by the sensors. Each of these parameters, according to its value, is assigned a score from 1 to 5 and all these parameters are then shown in a radar chart. Moreover, the residual value of a truck can be compared to its "similar", namely with the values obtained by the trucks with its same Market, year and model year. This is helpful for gathering information about the distribution of the score in a same category, so that the user can be aware if the truck under analysis belongs to the worst, the average or the best portion of trucks. These three classes are delimited using the first quartile Q1 and the third quartile Q3. The first class is marked by a red star and it includes all the trucks with a score inferior to Q1, the second is labelled by the yellow star and it characterizes trucks with value between Q1 and Q3 and, finally, the third, represented by the green star, is the elite of trucks, so those which have score's values above Q3. Figure 1 shows a graphical example of the dashboard.

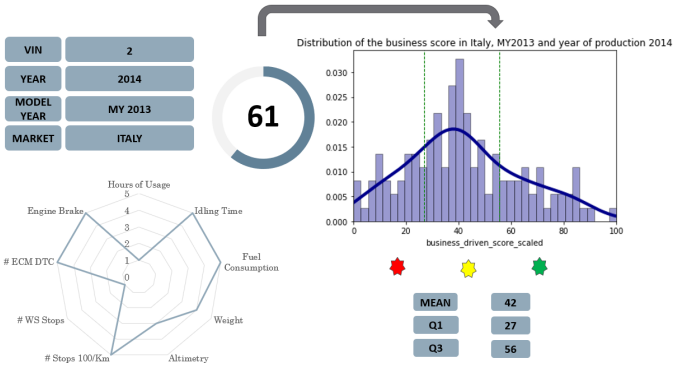

FIGURE 1. *Dashboard showing the residual value score of the selected truck (center), the radar chart containing the values of the parameters collected by the sensors (left) and a comparison with the values obtained by the other trucks with the same Market, year and model year (right)*

IV. EXPERIMENTAL EVALUATION

Here we discuss the results obtained by applied the proposed data-driven methodology presented in Section III on the data available for the use case described in Section II.

The correlation analysis and feature selection has allowed us to reduce the size of the original dataset, selecting only 9 numeric attributes from the original 38. This reduction has been made considering only the numerical attributes and eliminating those that are too correlated with each other; for example, the length in km of a mission and its duration in hours are two variables strongly correlated, so we consider only the duration in hours according to the suggestion of the domain expert. At the end of this step the nine variables selected are the *duration in hours* of a mission, the *average consumption of fuel each 100 km*, the *average load of a truck*, *Variable 1* and *Variable 2* that take into account the stops made during a journey; *Variable 3* and *Variable 4*, describing certain parameters relating to the vehicle's braking system; *Variable 5* and *Variable 6*, describing the maintenance work on the truck.

These nine variables are the ones we used to estimate the residual value of the trucks. Using the business-driven formula, each of these variables is associated with a business weight denoting their importance, while using the data-driven formula, we had to define for each variable its asymptotic value. Figure 2 shows the weights and limit values chosen for each parameter in order to use the two formulas. The threshold chosen for variables 1 to 6 is 1.5*min or 1.5*max depending on the nature of the variable concerned.
Using the value in this figure, we were able to calculate the residual values of the analysed trucks through the formulas

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2998940, IEEE Access

BUSINESS-DRIVEN WEIGHTS:		DATA-DRIVEN THRESHOLDS:
1 (normalized 0,19)	Fuel consumption	100l / 100Km
1 (normalized 0,19)	Hours of usage	20000
0,6 (normalized 0,11)	Loaded Weight	70000
0,6 (normalized 0,11)	Variable 1	1.5*min/max
0,5 (normalized 0,10)	Variable 2	1.5*min/max
0,4 (normalized 0,07)	Variable 3	1.5*min/max
0,5 (normalized 0,10)	Variable 4	1.5*min/max
0,3 (normalized 0,06)	Variable 5	1.5*min/max
0,4 (normalized 0,07)	Variable 6	1.5*min/max

FIGURE 2. *Weights (on the left) and thresholds (on the right) chosen for each of the 9 variables extracted after the feature selection*

we implemented. A first comparison between the results obtained, shows us that data-driven formula has a deeper level of granularity than the business-driven one, since it considers the values that the trucks assume in each variable in an exactly way, without grouping them into ranges (the score from 1 to 20). This aspect is evident in the two histograms and boxplots of Figure 3, that represent the distribution of the two scores: the business-driven one produces values that are distributed close to the mean, while the data-driven one ensures that values are more diversified. It means that the last ones have more variability and capture well the diversity among trucks.

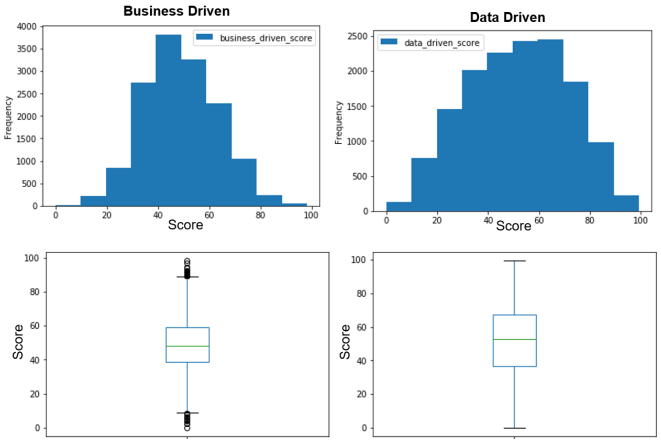

FIGURE 3. *Histogram and boxplot representing the distribution of the business-driven score (on the left) and the data-driven score (on the right)*

Comparison with respect to the traditional residual value estimation. Up to now, domain-experts rely on a two-dimensional matrix (named traditional estimation in the following) to assess the residual value of trucks. The residual value is determined through a grid-based approach, using the construction year and the number of travelled kilometers (properly discretized in bins). It is then slightly updated based on the visual inspection of the cabin trucks performed by domain experts. To assess the quality of the proposed methodology, we compared the obtained scores with the ones computed through the traditional estimation. We discovered the following interesting findings:

- The truck's usage is very varying in the considered large set of analyzed trucks, thus it is correct to rely on it to estimate the residual value.
- The proposed methodology provides more accurate estimates than traditional ones since a large number of parameters are analyzed.
- The new scores allow bettering differentiating among trucks with the same scores based on travelled kilometers and construction year, thus assessing the effective usage and the truck's wear.
- Based on the new scores, differentiating business strategies can be targeted to increase revenues coming from re-selling the trucks after buying back.
- The new scores are suited to promote differentiating leasing pricing.

A. COMPARISON THROUGH REGRESSION

In order to further compare the two formulas, considering the data driven score as the target variable and the parameters adopted in the formula as the predictors, we performed a multiple linear regression [8] to compare the coefficients obtained with the weights adopted in the business driven formula. To do this, the model used is as follows:

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n + \epsilon \quad (1)$$

where Y is the data driven score, x_i (i from 1 to 9) the scores from 1 to 20 obtained in the single variables, β_i their relative coefficients and ϵ the random error. The aim of this analysis is to estimate the values of the coefficients β_i, through the Least Squares approach that minimizes the RSS (Residual Sum of Squares) [9]. The result is that all the coefficients obtained are negative, except that for the use of the engine brake. It obviously means that there is an inverse relation between all of them (except the engine brake) and the final score, indeed the higher values they assume, the worse are the conditions for the truck.

According to the regression, the most significant variables are not only the fuel consumption, but also *Variable 2*; another discrepancy with the business-driven formula is that now the weight assumes the lowest importance. So, after this major insight thanks to the regression, the weight assigned to *Variable 2* can be increased, while the importance attributed to the *weight* should be decreased in the business driven formula.

B. IDENTIFICATION OF THE WORST TRUCKS AS OUTLIERS

Before carrying out this analysis to find the worst trucks among those we are considering, the trucks were divided by geographical market: Spain, Italy, Germany. Then, for each subgroup, DBSCAN was applied. What we can see is that in all the results found, the trucks labelled by the cluster algorithm as 'outliers' are actually those with a lower residual value. Figure 4 shows the results obtained: for each market there is a cluster set representation through SVD

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2998940, IEEE Access

[10] and boxplots of the data driven and business driven score. In each 3d scatter plot, the three dimensions reproduce the first three most representative attributes after the SVD decomposition and each point is typified by the colour of the cluster it belongs to. The first three values of the singular value decomposition explain more than 80% of the total variance, so they can be fairly used to reduce data dimension.

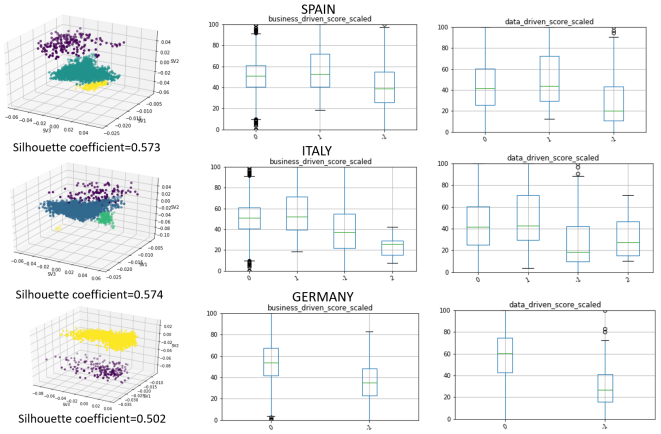

FIGURE 4. *From left to right: cluster set representation through SVD after the DBSCAN, boxplot of the business driven and data driven score scaled by Market in the clusters obtained; from top to bottom: Spain, Italy, Germany*

The representation of the data in a 3d way through SVD allows us to see how in all three cases the clusters are well separated, while from the boxplots we can easily see how in all cases the cluster with label -1 (the outliers) is the one with the lowest residual values. These results were obtained with minPoints=5 and eps=0.18 found by using the strategy proposed in Section III-D. In Figure 4, under each 3d representation, there is the silhouette coefficient that shows us how good the subdivision is. Silhouette is a technique used to evaluate how good a clustering solution is [11]. This technique measures how similar an object is to its own cluster and how far it is from other clusters. For each point, the Silhouette index can assume a value between -1 and 1, where 1 means that the point has been correctly assigned to the cluster, while -1 means that the point was not assigned to the right cluster. To evaluate the goodness of all our clustering algorithm, we use the Average Silhouette, which is the average of all the silhouettes of our single points.

V. RELATED WORK

In the past years different studies regarding the residual values of vehicles have been carried out, like in [3]. In this paper the authors propose an approach based on machine learning techniques in order to estimate the residual value of a car. There are two sets of parameters on which the learning algorithm has been trained on. The first set , which is referred to as standard domain variables (SDV), includes the brand, the model, the age and the mileage. The remaining set, which is called transaction specific variables (TSV), exploits more specific information, such as intended vehicle using and special equipment. Another study [4] concerning a similar topic has been carried out in 2014. This study is focused on determining the residual values of cars in Mauritius. The author tried four different techniques (k-nearest neighbour, decision trees, multiple regression analysis and naïve bayes) and compared the results. The data are collected from ads in daily papers and the parameters that the author used are similar to the ones that were exploited by [3], such as model, mileage in km and year of manufacture. In addition to the studies just mentioned there is [5], that takes on the residual value issue with a different approach. More specifically it applies an artificial neural network, a support vector machine and a random forest to work as an ensemble. Even if all of these studies use different techniques, the data that they use in order to estimate the residual value are very similar and traditional. In more recent works further analysis has been made: in [12] authors use an asymmetric cost functions in order to estimate the residual value, emphasizing the role of deliberate managerial decisions in cost behaviour, while in [13], authors examine the challenge of forecasting residual values for commercial vehicles, using a model based on an ANN approach and investigating in more detail the influence of the model age and a general time factor, where the model age measures the time between the market launch of a new model class and the resale date of a specific vehicle belonging to this class.

Our approach on the other hand, besides focusing on the residual value of heavy trucks instead of the residual value of cars, exploits the information retrieved from the electronic control module (ECM), instead of using traditional parameters, such as mileage and year of manufacture.

Other studies have been carried out on trucks, trying to exploit the telematics data collected by sensors installed on them, but the purpose of these studies is not to calculate their residual value. For example, in [14], the authors aim to predict the travel speed of trucks driving on urban express roads, basing the prediction on GPS data collected from vehicles. Another work illustrated in [15] explains the use of telemetrics and machine learning in order to model the fuel consumption by trucks. The data used to train the different machine learning models come from two different sources. The first source are the sensors installed on the trucks, which keep track of information such as fuel consumption, while the second source is the Highways Agency Pavement Management System (HAPMS) of Highways England, which provides information regarding the roads. Another study [16] focuses on how the fuel consumption of a truck is a crucial aspect to consider while purchasing a new truck. In order to determine the fuel consumption the authors consider different parameters, some of which are more traditional, such as total weight and average speed, while others are more peculiar, such as weather conditions. The authors, after comparing different models, proved that weather conditions are a relevant factor to take into account while dealing with fuel consumption. Most of the data used was obtained by telematics systems from more than 500 trucks over a two years period.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2998940, IEEE Access

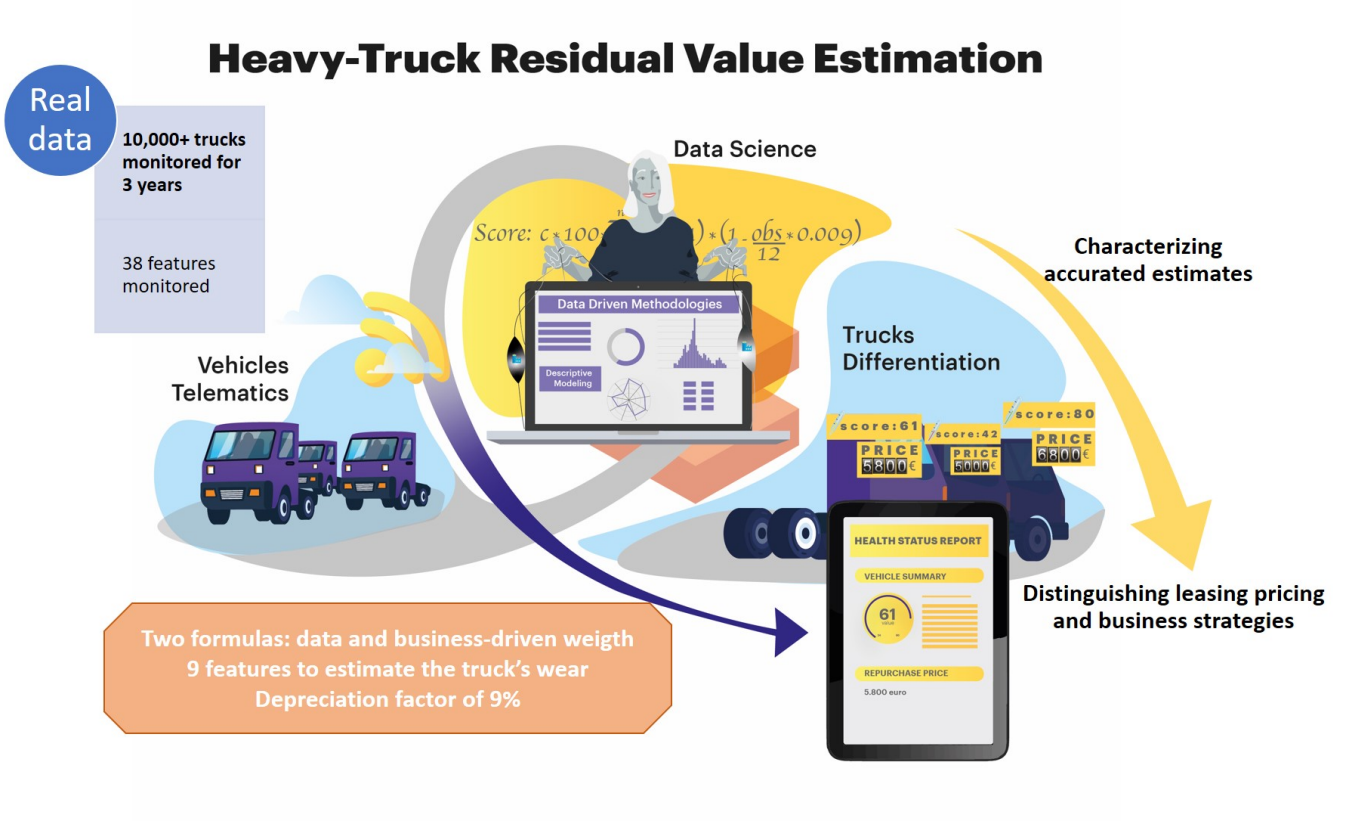

FIGURE 5. *Summary graphical abstract of the proposed methodology*

As the reader can see our work differs also from this second batch of studies. The residual value issue addressed with machine learning techniques focuses on cars and not trucks, like in [17]. On the other hand studies that focus on trucks are more concerned with the fuel consumption issue or with maintenance, such as [18], more than the residual value.

VI. DISCUSSION

The advent of Industry 4.0 promotes a constant evolution of companies through intelligent environments including an intensive and pervasive use of sensors of different nature located in different production and industrial contexts. This revolutionary industry model requires the integration of information and communication systems in the production process with the consequent generation of large amounts of data. In the automotive industry, this innovation has manifested itself with the introduction of connected vehicles, able to collect a large amount of data thanks to sensors and send it to a remote storage location for later analyses. Then, in order to extract additional value from the collected data, the recent Big-Data-enabled IT technologies allow efficient data management in terms of storage and processing and excellent data analysis libraries to support the knowledge extraction process.

To the best of the authors' knowledge, this work is the first attempt to study the estimation of the residual value of a large collection of heavy trucks by analyzing telematics data. Figure 5 provides a graphical summary of the work proposed in this paper enriched with (i) the most relevant characteristics of the real-data used to assess the effectiveness of the proposed data-driven methodology, (ii) the key innovative features of the proposed strategy, and (iii) the most relevant findings. Two formulas have been proposed with data-driven and business-driven weights to perform an accurate estimation of the trucks' wear and the corresponding values useful to support decision making-process.

The findings of this study can be exploited from both the managerial and the academic perspective.

From the academic perspective, the findings of this research activity demonstrate the ability of the proposed methodology to correctly analyze a huge volume of sensor data related to the effective usage of trucks through an original approach and for a new business objective (i.e., estimation of the residual value of heavy trucks at the buy-back).

From the managerial perspective, the findings of this study could be exploited to inform business users about the residual value of heavy trucks at the buy-back to promote targeted marketing strategies based on the identified truck categories. Furthermore, the algorithm can be run periodically to monitor the current estimates and made the proper decision to improve the effectiveness of the business strategies. Furthermore, we believe that the proposed data-driven approach can

be easily ported also to different contexts (e.g., agricultural machinery, industrial robotics, self-driving machines) where the original value of the vehicle/object is very high, there already exists ad-hoc data collection applications to effectively monitor the usage behaviour, and where the residual value assumes a key estimation to increase the next business activities alternative to the selling. The possible relation between this work and the examples of agricultural machinery, industrial robotics, self-driving machines is straightforward, while the overall benefit in terms of business opportunities could be very high.

Moreover, the formula proposed in this work can be extended to apply new business policies. Considering for example a case of car sharing, the formula can be useful to know not only the residual value of the vehicle, but the degradation caused by a user during its use. This information could give rise to new business policies, making consumers pay differently not only according to the time of use, but also considering other parameters collected telematically on the vehicle. This could encourage customers to maintain a safer and more balanced driving style, reducing vehicle degradation. This new managerial asset could also be applied to electrical machines, which still have excessive prices to be affordable. Leasing contracts could be offered to customers in this case, monitoring the use of the machine battery and basing the contract price on an analysis of the parameters collected, adapting the formula proposed in our work.

As future research direction there is still room for improvement the proposed methodology. In fact, one of its main drawback is that the data-driven formula requires the definition of the asymptotic value for each parameter under analysis to model the maximum value assumed (i.e., limit condition). A similar drawback also appears in the business-driven formula where the engineering department is responsible to define a different weight for each parameter according to the influence of each parameter in the total degradation of the truck. We are currently investigating novel statistics strategies to separately model the ratio between the measured value and the corresponding limit for each parameter and the business weight by analyzing a larger and heterogeneous set of heavy trucks used in Europe.

REFERENCES

[1] Omprakash Kaiwartya, Abdul Hanan Abdullah, Yue Cao, Ayman Altameem, Mukesh Prasad, Chin-Teng Lin, and Xiulei Liu. Internet of vehicles: Motivation, layered architecture, network model, challenges, and future aspects. IEEE Access, 4:5356–5373, 2016.

[2] Kazi Masudul Alam, Mukesh Saini, and Abdulmotaleb El Saddik. Towards social internet of vehicles: Concept, architecture, and applications. IEEE access, 3:343–357, 2015.

[3] Stefan Lessmann, Mariana Listiani, and Stefan Voss. Decision support in car leasing: a forecasting model for residual value estimation. ICIS 2010 Proceedings - Thirty First International Conference on Information Systems, page 17, 01 2010.

[4] Sameerchand Pudaruth. Predicting the price of used cars using machine learning techniques. International Journal of Information Computation Technology, 4:753–764, 01 2014.

[5] Enis Gegic, Becir Isakovic, Dino Kečo, Zerina Mašetić, and Jasmin Kevric. Car price prediction using machine learning techniques. TEM Journal, 8:113–118, 02 2019.

[6] Nicola Zingirian and Carlo Valenti. Sensor clouds for intelligent truck monitoring. In 2012 IEEE Intelligent Vehicles Symposium, pages 999–1004. IEEE, 2012.

[7] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, et al. Cluster analysis: basic concepts and algorithms. Introduction to data mining, 8:487–568, 2006.

[8] Per Kragh Andersen and Lene Theil Skovgaard. Regression with linear predictors. Springer, 2010.

[9] Thomas J Archdeacon. Correlation and regression analysis: a historian's guide. Univ of Wisconsin Press, 1994.

[10] Alan Kaylor Cline and Inderjit S Dhillon. Computation of the singular value decomposition. 2006.

[11] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20:53–65, 1987.

[12] Korbinian Dress, Stefan Lessmann, and Hans-Jörg von Mettenheim. Residual value forecasting using asymmetric cost functions. International Journal of Forecasting, 34(4):551–565, 2018.

[13] Christoph Gleue, Dennis Eilers, Hans-Jörg von Mettenheim, and Michael H Breitner. Decision support for the automotive industry: forecasting residual values using artificial neural networks. 2017.

[14] Jiandong Zhao, Yuan Gao, Zhenzhen Yang, Jiangtao Li, Yingzi Feng, Ziyan Qin, and Zhiming Bai. Truck traffic speed prediction under non-recurrent congestion: Based on optimized deep learning algorithms and gps data. IEEE Access, 7:9116–9127, 2019.

[15] Federico Perrotta, Tony Parry, and Luís C. Neves. Application of machine learning for fuel consumption modelling of trucks. In 2017 IEEE International Conference on Big Data, BigData 2017, Boston, MA, USA, December 11-14, 2017, pages 3810–3815, 2017.

[16] T. Bousonville, M. Dirichs, and T. Krüger. Estimating truck fuel consumption with machine learning using telematics, topology and weather data. In 2019 International Conference on Industrial Engineering and Systems Management (IESM), pages 1–6, 2019.

[17] Shen Gongqi, Wang Yansong, and Zhu Qiang. New model for residual value prediction of the used car based on bp neural network and nonlinear curve fit. Measuring Technology and Mechatronics Automation, International Conference on, 2:682–685, 01 2011.

[18] Slawomir Nowaczyk, Rune Prytz, Thorsteinn Rognvaldsson, and Stefan Byttner. Towards a machine learning algorithm for predicting truck compressor failures using logged vehicle data. Frontiers in Artificial Intelligence and Applications, 257:205–214, 01 2013.

TANIA CERQUITELLI has been an associate professor at the Department of Control and Computer Engineering of the Politecnico di Torino, Italy, since March 2018. Her research activities have been mainly devoted to fostering and sharing research and innovation on automated data science and machine learning in different real-life settings. Tania has been involved in many European and Italian research projects addressing different research issues related to machine learning and data analytics for energy-related data and Industry 4.0. She got the master degree with honours in Computer Engineering (in 2003) and the PhD degree (in 2007) from the Politecnico di Torino, Italy, and the master degree with honours in Computer Science (in 2003) from the Universidad de Las Américas Puebla, Mexico.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2998940, IEEE Access

ANDREA REGALIA is leading Smart Vehicle businesses at Accenture. With an innovator DNA and a technology life-passion, he's obsessed by applied research, bringing to the business world new Use Cases that helps companies to improve their services and differentiate from competitors. Early player in the Internet of Things, currently focuses on Artificial Intelligence for self driving vehicles, but doesn't stop searching the Holy Grail of Data Monetization

ELENA LIORE is currently working at Accenture as a data science analyst. Her main activities concern business intelligence for the Big Data, data analytics and data visualization in the field of Mobility X.0. She got the degree in Physical Engineering (in 2017) and the master degree in Mathematical Engineering with honors (in 2019); she took part in the Erasmus program at TU/e, the Netherlands, in the Data Science curricula.

EMANUELE MANFREDI is a senior data scientist working at Accenture. Always passionate by maths and numbers with a strong interest in technology, also very curious about business process, used to be a data scientist when the figure was not already born thanks to a mix of different experiences. With a strong background in Customer analytics and the opportunity to experience other topics is now focusing on automotive fields, always looking for the right balance between cutting edge technologies and traditional ones aimed to solve real life problems.

FABRIZIO CONICELLA is Head of Digital CV Segment within CNHi organization. He is driving Iveco digitalization and leading development and launches of new connected services for trucks and buses aiming lower TCO and higher uptime and productivity. Fabrizio got a Master Degree in Aeronautical Engineering and 24 years of experience into the automotive sector with roles of growing responsibility mainly focused on engine development, truck & powertrain integration, vehicle efficiency, emissions and digitalization.

PAOLO BETHAZ received the master's degree in computer engineering from the Politecnico di Torino, Italy, in 2019. His research interests are focused on the improvement of the automated data science in different contexts, including industry 4.0, the IoT, and NLP fields.