Attempt to better trust classification models: Application to the Ageing of Refrigerated Transport Vehicles

Marie Le Guilly, Claudia Capo, Jean-Marc Petit, Vasile-Marian Scuturici, Rémi Revellin, Jocelyn Bonjour, Gérald Cavalier

Attempt to better trust classification models: Application to the Ageing of Refrigerated Transport Vehicles

Marie Le Guilly[1], Claudia Capo[1,2,3], Jean-Marc Petit[1], Vasile-Marian Scuturici[1], Rémi Revellin[2], Jocelyn Bonjour[2], and Gérald Cavalier[3]

[1] Univ Lyon, INSA Lyon, LIRIS, UMR 5205 CNRS, Villeurbanne, France
[2] Univ Lyon, INSA Lyon, CETHIL, UMR 5008 CNRS, Villeurbanne, France
[3] Cemafroid, Fresnes, France

Abstract. CEMAFROID is a company with a french delegated public service, delivering conformity attestations of refrigerated transport vehicles. It studies the ageing of those vehicles, depending on several physiochemical and mechanical factors, for which physical models have been proposed by researchers in thermal engineering. The DATAFRIG® database records more than 300 000 attestations of vehicles over 80 attributes, opening the opportunity to predict the ageing by building a numerical model using machine learning methods. During the development of such a model, several classical questions appeared, regarding the data quality, the field reality and the mistrust of domain experts.
In this paper, we propose to use the notion of functional dependencies to address the aforementioned model's limitations. In particular, we investigate how FDs could help, especially using their counterexamples, that turn out to provide meaningful examples of such limitations, easily interpretable by domain experts. Interestingly, the existence of such counterexamples in the dataset is a way of demystifying the numerical model with the experts, by giving them back the control over their own data. This approach has been tested with domain experts from CEMAFROID, with many positive feedbacks. It is worth noting that this attempt to better trust classification models is not limited to a particular application, and could be generalized to others.

Keywords: Prediction model · Industrial machine learning · Applications · Functional Dependencies.

1 Introduction

Refrigerated transport vehicles' main objective is to supply consumers with good quality and safe perishable products. In this context, CEMAFROID offers testing and calibration services, and has been designated as an approved body for issuing conformity attestations for refrigerated transport vehicles. To this end, researchers in thermal engineering have access to a large database, DATATFRIG®, recording data about more than 300 000 attestations for refrigerated vehicles, with tables spanning over almost 80 attributes. Among all

these attributes, the thermal insulation of the body is a basic element of the refrigerated equipment, characterized by an "insulation coefficient" denoted K: it is very important as it may be controlled as required by the ATP (the international Agreement for the Transport of Perishable foodstuff). The lower the K coefficient the better, but due to the ageing of the vehicle, it increases over time. To control it and allow the vehicle to continue to transport perishable food, its thermal insulation is measured again after 12 years of service. The ratio between this value and the initial value allows too define the *ageing* of the vehicle.

To explain the variability in the ageing of refrigerated vehicles, several studies have looked at the problem to identify factors to explain it, and propose physical models for the ageing. [**?**] presented the factors playing a role in the ageing process. As stated by [**?**], the ageing of an insulated enclosure for a refrigerated vehicle is mostly due to the permeability of the foam to the gases, to the condensation of water into the foam cells and to the increase of the percentage of the broken cells. Ageing of refrigerated vehicles also has a mechanical component due to the movements on the road, the routes covered and the payload. A statistical analysis carried out in [**?**] highlighted the influence in the ageing rate due to the rails and to the refrigerating units.

But whereas all these studies rely on physical models to analyse the problem, the amount of data available in the DATATFRIG® makes it possible to build a predictive model, using machine learning techniques: such a new approach is a way to tackle the problem from a different angle, to possibly identify new causes for the ageing of vehicles, and to confront new results to the one obtained with physical models. Shifting from a physical to a numerical model is indeed a big change of paradigm, especially for the domain experts that mainly design and work with physical models. The ML techniques might even generate scepticism. In this study, we therefore paid strong attention to give guaranties about the model, and to explain its limitations. It was important to explain why the considered data could (or not) produce a satisfying model, so that the experts that would use it afterwards could trust it and understand why its decisions made sense. To this end, it appeared that functional dependencies could be a very powerful tool to model knowledge on the classifier [**?**].

Functional dependencies (FDs) are a well-known and widely studied notion over the years, at the foundation of the theory for relational database design (see [**?**]), expressing constraints and relationships between two sets of attributes. In the setting of this paper, we propose to use FDs to model the possible limitations of a predictive model, by considering that FDs can determine whether a model can actually exists over a relation, before actually trying to determine the model using a classification algorithm. Indeed, a classification model seeks an optimal function, with respect to a given error measure, that maps features to a class. This process relies on a strong hypothesis, which is the existence of a function from the features to the labels, and only seeks to determine such a function. However, in a complementary way, FDs can be seen as a way to determine if such a function even exists. A detailed study of the link between FDs and classification is available in [**?**]. We only underline here that when there exists

a function from a feature vector X to a label Y, it follows that the function dependency $X \to Y$ is satisfied.

Thus, for a classification problem, it seems logical to first verify the existence of a function using FD $features \to class$. Moreover, understanding what tuples prevent the FD from being satisfied can lead to improve the results, if the identified blockages can be explained and fixed by the domain experts. This is where the notion of counterexample is useful. Counterexamples identify pairs of tuples for which the classifier will never be able to perform correctly, as for the same input, it will always predict the same output. Understanding where these counterexamples come from, and what can be done to avoid having such tuples in the dataset, is then an important tool to improve the classification performances.

These notions were used in an application to CEMAFROID's data, for which the notion of counterexamples appeared to be central for the discussion with thermal engineers. In this paper, we expose how we presented such counterexamples to the experts in thermal engineering, in order to give them back the control over their own data, by explaining some of the limitations of the classifier. The discussion that followed allowed to identify several possible causes for the existence of counterexamples, that are exposed in this paper, along with possible solutions to remove them. We also measured the proportion of counterexamples in the dataset, by using metrics on the error of FDs in a relation. The feedback obtained during this discussion highlighted several interesting research problems related to classification, for which FDs could be useful, such as feature engineering and data cleaning.

Paper organisation Section **??** presents the available data and the first numerical model built on it. Section **??** presents the approach used, and the results of a discussion with thermal engineers, based on the analysis of counterexamples to a FD in their dataset, before concluding in section **??**.

2 Predicting the ageing of Refrigerated Transport Vehicles Data

2.1 Refrigerated transport vehicle's data

At 31 December 2017, based on the Datafrig® data, the French fleet counted 110 000 refrigerated transport equipment with a valid ATP certification. These equipments are divided into different categories, the main ones are: vans (vehicles with a total weight allowed in charge smaller than 3.5 tons), trucks (vehicles whose total weight allowed in charge varies from 3.5 to 29 tons) and semi-trailers (vehicles with total weight allowed in charge higher than 29 tons). The Datafrig® database is managed by Cemafroid and contains all the ATP in-service equipment in France.

From this database, sample of only 1158 highly curated data could be extracted from the 109 122 refrigerated equipment registered in the Datafrig® database. These 1158 data represent different in-service vehicles of different firms, tested after twelve years in the laboratory of Cemafroid. For each of these

vehicles more than 80 features are known. Ten of the most important features, from a domain expert point of view, were analyzed including the K_{12}, the corresponding K_p, the type of vehicle, the body manufacturer, the nature of the refrigerated enclosures and their isolation, the use of the vehicles and the different types of transport to which they may be subjected. Most of the available features are therefore categorical. Our selection criteria for the sample of 1158 vehicles were based on data quality requirements applied on the ten selected attributes (or variables): no null values, no outliers, and no duplicates. It is worth mentioning that we rely on knowledge domain provided by experts to identify our ten features. A statistical analysis of this dataset is available in [**?**].

2.2 Initial prediction results

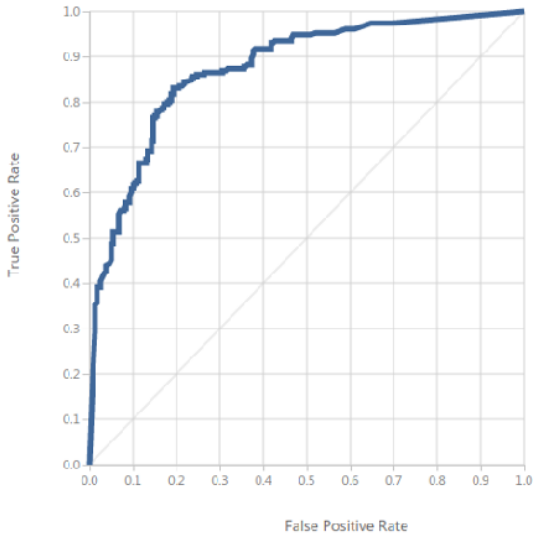

Fig. 1. ROC curve for the prediction of refrigerated vehicles ageing

Once the cleaned dataset obtained, a traditional classification workflow was applied: the data was split in a training (80%) and a testing set (20%). The objective was to sketch the construction of a first initial model, before any analysis of the dataset, using classical state-of-the-art methods. It was decided to build a first classification model using a decision tree [**?**], in order to obtain an interpretable model, that could easily be discussed with thermal engineers. To improve the results, a boosted version of the algorithm (see [**?**]) was used. The results were very encouraging, with a precision of 0.818 and a recall of 0.783. The F1 score was of 0.800. In addition, a ROC curve of this first classifier is presented on figure **??**, also showing good performances for the ageing prediction.

With this first model available, it was possible to engage in a discussion with the experts in thermal engineering. The purpose was to understand what these result meant, but also and especially what could be done to improve the results, and what were the blocking points in the dataset itself. For example, some tuples might require additional cleaning, or some records in the database might cause problem for the learning process. As checking the data manually is very tiresome, even on a small dataset, it appeared that the notion of functional dependency could be extremely useful to assist the data expert in this specific task.

3 Domain experts and trust in the classification model

3.1 Preliminaries

We first recall basic notations and definitions that will be used throughout the paper. It is assumed the reader is familiar with databases notations (see [**?**]).

Let U be a set of attributes. A relation schema R is a name associated with attributes of U, i.e. $R \subseteq U$. A database schema $\mathcal{R}$ is a set of relation schemas.

Let D be a set of constants, $A \in U$ and R a relation schema. The domain of A is denoted by $dom(A) \subseteq D$. A tuple t over R is a function from R to D. A relation r over R is a set of tuples over R. If $X \subseteq U$, and if t is a tuple over U, then we denote the restriction of t to X by $t[X]$. If r is a relation over U, then $r[X] = \{t[X], t \in R\}$. The active domain of A in r, denoted by $ADOM(A, r)$, is the set of values taken by A in r. The active domain of r, denoted by $ADOM(r)$, is the set of values in r.

We now define the syntax and the semantics of a FD.

Definition 1. *Let R be a relation schema, $X \subseteq R$ and $C \subseteq R \backslash X$. A FD on R is an expression of the form $R : X \rightarrow C$ (or simply $X \rightarrow C$ when R is clear from context)*

Definition 2. *Let r be a relation over R and $X \rightarrow C$ a FD on R. $X \rightarrow C$ is satisfied in r, denoted by $r \models X \rightarrow C$, if and only if for all $t_1, t_2 \in r$, if $t_1[X] = t_2[X]$ then $t_1[C] = t_2[C]$.*

Without loss of generality, the scope is limited to crisp FDs, i.e. using only the strict equality. However, many other extensions have been proposed (see for example [**?**]). In addition, this paper only uses canonical FDs, that only have a unique attribute in their right hand side.

Strictly speaking, it is only possible for a relation to satisfy entirely a FD: otherwise, the dependency is not satisfied, there is no middle ground. When dealing with real life data, it is very likely that the FD will not be satisfied, even though a classification model might then be built with satisfying performances. However, understanding what tuples prevent the FD from being satisfied can then lead to improve the results, if the identified blockages can be explained and fixed by the domain experts. Such tuples are called counterexamples, and are defined as follows:

Definition 3. *Let r be a relation over R and $X \to Y$ a FD f on R. The set of counterexamples of f over r is denoted by $CE(X \to Y)$ and defined as follows:*

$$CE(X \to Y, r) = \{(t_1, t_2) | t_1, t_2 \in r, t_1[X] = t_2[X] \text{ and } t_1[Y] \neq t_2[Y]\}$$

Counterexamples identify pairs of tuples for which the classifier will never be able to perform correctly, as for the same input, it will always predict the same output. These pairs are therefore very important, as their number directly impacts the quality of the classification. As a result, in order to evaluate the impact of counterexamples on classification, it is necessary to know their proportion in the dataset. Indeed, if a classifier only contains a few counterexamples, the impact on the classification will be marginal. On the opposite, a large counterexample set will significantly impact the accuracy results.

Evaluating the impact of counterexamples can be a little subtle. Indeed, a single tuple might cause many counterexamples, if it is in conflict with many other tuples that agree between them. On the opposite, on other relations, the counterexamples might be all due to many different tuples that each are in conflict with only a few other tuples. This problem is actually equivalent to estimating the error of the FD in a relation, a problem addressed in [**?**], in which three measures are presented, given a FD $X \to C$ and a relation r.

The first one, G_1, gives the proportion of counterexamples in the relation:

$$G_1(X \to C, r) = \frac{|\{(u, v) | u, v \in r, u[X] = v[X], u[C] \neq v[C]\}|}{|r|^2}$$

Following this first measure, it is also possible to determine the proportion of tuples involved in counterexamples. This measure G_2 is given as follows:

$$G_2(X \to C, r) = \frac{|\{u | u \in r, \exists v \in r : u[X] = v[X], u[C] \neq v[C]\}|}{|r|}$$

These two metrics are designed to evaluate the importance of counterexamples in the relation. Similarly, measure G_3 computes the size of the set of tuples in r to obtain a maximal new relation s satisfying $X \to C$. Contrary to [**?**] that present this measure as an error, we propose it as follows:

$$G_3(X \to C, r) = \frac{max(\{|s| | s \subseteq r, s \models X \to C\})}{|r|}$$

In [**?**], it is underlined that G_3 is a direct upper bound for the accuracy of a classifier trained on the considered data, when the FD between the features and the class is considered, exactly as for the application to CEMAFROID's data.

3.2 Presentation of counterexamples

Using the tools presented before, the classification dataset was analyzed to discuss with the experts. A toy sample of counterexamples is presented in table **??**. For sake of clarity, only five attributes are represented in this example. Each

id	Manufacturer	Cell Type	Insulation type	Vehicle type	Products	Ageing t_1	Ageing t_2
1	Firm 1	Integrated	Polyuréthane	Truck	Meat	High	Low
2	Firm 2	Integrated	Polyuréthane with cylopentane	Van	Fruits	High	Low
3	Firm 3	Rapportee	Polyuréthane with cylopentane	Panel truck	Frozen food	High	Low
4	Firm 4	Integrated	Polyuréthane with cylopentane	Trailer	Vegetables	High	Low
5	Firm 5	Rapportee	Polyuréthane with cylopentane without CFC	Truck	Cheese	High	Low
6	Firm 6	Rapportee	Polyuréthane with cylopentane	Remorque	Dairy products	High	Low

Table 1. Subset of counterexamples from the classificationd dataset, on the ageing of refrigerated transport vehicles.

line represents a counterexample, which is two tuples. As they both share the same values over the classification features, only the ageing column, on which they differ, is represented for both(Ageing t_1 and Ageing t_2). This is also how the counterexamples were presented to the thermal engineers, using an interface developped to present the counterexample. It is named LeaFF (Learning Feasibility with FDs), and can be used to query the data used for prediction, retrieve the counterexample, and obtain the metrics for FD satisfaction. A snapshot of this interface is presented on figure **??**, applied to CEMAFROID's data: it shows the part of the interface that can be used to enter the features and the class, so that the FD can be checked and counterexamples retrieved if necessary.

3.3 Measuring the counterexamples rate

The three aforementioned metrics were computed over the refrigerated vehicles dataset. The proportion of counterexamples, $G_1 = 9.02\%$, is low, showing that the pairs of tuples in the dataset are not a big proportion. However, as $G_2 = 100\%$, all tuples from the dataset are involved in at least one counterexamples: this is likely because a few tuples are in conflicts with almost all the other. This is confirmed by measuring $G_3 = 86.73\%$: this shows that to obtain a counterexamples-free dataset, the vast majority of the data can be saved.

When presented with the measures, the domain expert had many questions regarding what exactly each measure meant, and how to know whether the results were good or not: is it good if the score is low ? Or should it be higher to be better ? Indeed, these measures are not trivial, and the meaning they carry over the pairs of tuples is not intuitive right the first time. This interaction therefore highlighted that it will be interesting to find a convenient visualization for domain experts, to present this result in a more instinctive manner.

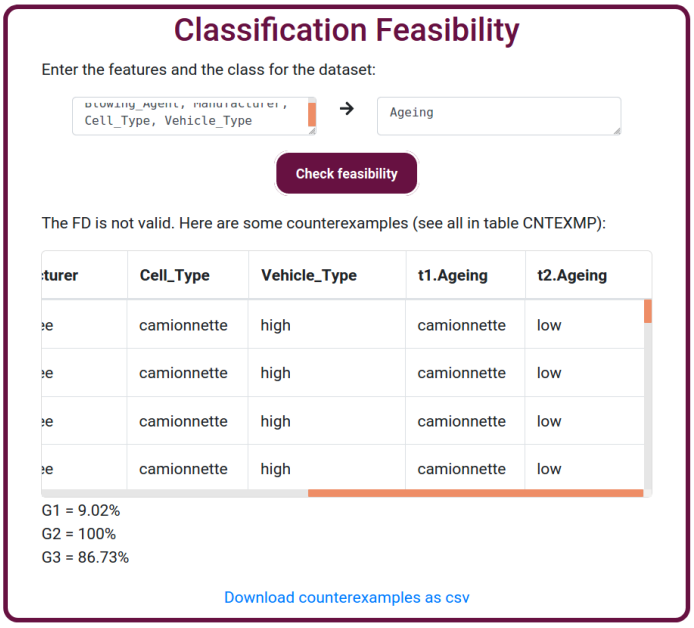

Fig. 2. Snapshot of the interface LeaFF used to explore FDs and counterexamples

3.4 Explaining counterexamples

It then appeared necessary to dive into counterexamples. After a thorough discussion, it appeared that several factors could explain them, each having a different solution. Ultimately, these factors were divided into three main categories.

Dirty Data A considerable amount of time had been spent on preparing and cleaning the dataset in order to use it for classification. However, when looking at counterexamples, it appeared that some tuples contained data that appeared to be incorrect. For counterexample 1 in table **??**, it appeared that the vehicle with a low ageing was actually a van instead of a truck. Similarly, for counterexample 2, the vehicle with a high ageing actually transported meat instead of vegetables. Such counterexamples are easy to fix by correcting the wrong values, once they have been detected using the FD-based approach. Alternatively, it is possible to take into account additional attributes that are more precise than the dirty ones: for example, the weight of the vehicle can better define the vehicle's type than the given label, so mistakes can be avoided if it is taken into account: with the weight, vans and trucks could be more easily differentiated.

Missing information For other counterexamples, it appeared that the attributes selected for classification were not enough to explain their existence. However, other attributes, that had not been kept for classification, allowed to discriminate between the two tuples involved in the counterexample. For instance, in table

??, counterexample 3 can be removed if the number of food cases in the vehicle is considered. For counterexample 4, a specific characteristic of the cooling unit differed between the two tuples. As a result, taking those additional attributes, that at first had not been considered relevant for the classification, allowed to remove counterexamples, that will then improve the classifier's performances. This is a way to do feature engineering in collaboration with the domain experts.

Human Factor Finally, for a last group of counterexamples such as number 5 and number 6 from table **??**, it appeared that the only explanation was a human factor, such as how the driver operates the vehicle. Indeed, this is susceptible of influencing the ageing of the vehicle, but it is hard to quantify, and may pose ethical issues. As a result, this last class of counterexamples is very difficult to fix, as data cannot be cleaned or completed. However, being aware that such counterexamples exist helps the data expert in understanding the limitations of the classification model. Finally, it also indicates other values that could be interesting to record: in this case for example the average speed of the vehicle. Similarly, during its life, the vehicle is subject to accidents of which nothing is known. Information about the nature and severity of such accidents could be useful for the study of ageing, and could also explain some counterexamples.

To assess the impact of these counterexamples on the predictive model, it was decided to remove the tuples that are the most involved in counterexamples from the dataset: each removed tuples was carefully checked with a thermic expert, and each removal allowed to gain accuracy for the classifier. It is necessary to be careful when removing such data to not create a biased model, but it can have a positive impact on the classifier's accuracy, directly linked to measure G_3.

3.5 Take away lessons

Using functional dependencies to explore the classification dataset was an enriching experience, that created fruitful discussions and results. First, the counterexamples allowed to identify limitations in the dataset, and to take concrete actions to get higher quality data, and therefore improve the results for a future new model. The counterexamples are a powerful notion to avoid the domain expert from being overwhelmed by the data, as she then only have a small but meaningful subset of tuples to study. The counterexamples are therefore a perfect starting point for a discussion between data scientists and domain experts: while the first gain knowledge on data they are not expert on, the others can point out important information more easily. The counterexamples are a way for domain experts to read a concrete information that have an impact on their day-to-day work. By helping to clean and improve the dataset, domain experts stay in the loop, and better understand the data used in the learning process. Last but not least, they gain some evidence regarding whether or not it makes sense to infer a numerical model from their data. This study also shows that in classification, it is not possible to ignore the field reality, and to only see the problem as a matrix of data: the physical model is as important as the data itself, and bridges have to be built between physical and numerical models.

4 Conclusion

This paper presents the results obtained to better trust classification models in the context of a study with CEMAFROID, a french company delivering conformity attestations for refrigerated transport vehicles. To this end, a classification dataset was built and a first classifier was elaborated using standard classification methods. It showed very promising results, but also raised interrogations from the thermal engineers, related to trust issues in machine learning. In order to understand the dataset and to demystify the classifier, we proposed to see the classification model from the features to the class to be predicted as a functional dependency. The counterexamples to this FD then give direct limitations of the classifier. The counterexamples were analyzed, and various causes for their existence were discovered, as well as potential solutions to remove them. More importantly, domain experts were able to better understand the pros and cons of the numerical model, influencing their trust in the whole machine learning process. This study was therefore really enriching, allowing to bridge a gap by using FDs, a well-known notion in databases, into a machine learning process. The methodology from this paper could be used in any classification problem, to first study the existence of a model before trying to learn from data.

References

1. Armstrong, W.: Dependency structures of database relationship. Information processing pp. 580–583 (1974)
2. Bosc, P., Dubois, D., Prade, H.: Fuzzy functional dependencies and redundancy elimination. Journal of the American Society for Information Science **49**(3), 217–235 (1998)
3. Breiman, L.: Classification and regression trees. Routledge (2017)
4. Capo, C., Petit, J.M., Revellin, R., Bonjour, G., Cavalier, G.: Ageing of in-service refrigerated transport vehicles: a statistical analysis. In: 25th IIR International Conference on Refrigeration (2019)
5. Kivinen, J., Mannila, H.: Approximate inference of functional dependencies from relations. Theor. Comput. Sci. **149**(1), 129–149 (Sep 1995). https://doi.org/10.1016/0304-3975(95)00028-U, http://dx.doi.org/10.1016/0304-3975(95)00028-U
6. Le Guilly, M., Petit, J.M., Scuturici, V.M.: Evaluating classification feasability over datasets using functional dependencies. In: BDA 2019 35ème conférence sur la Gestion de Données: Principes, Technologies et Applications. Lyon, France (2019)
7. Levene, M., Loizou, G.: A Guided Tour of Relational Databases and Beyond. Springer-Verlag, Berlin, Heidelberg (1999)
8. Panozzo, G., Alberti, O., Toniolo, B., Barizza, A., Boldrin, B.: Parameters affecting the ageing of insulated vehicles. Proc 20th International Congress of Refrigeration **IV** (1999)
9. Panozzo, G., B., B., Minotto, G., Toniolo, B., Biancardi, N.: Ageing of insulated vehicles: theoretical model and experimental analysis. Proc. 19th Int. Congr. Refrig., Den Hague, Netherlands, Vol. II **II**, 583–589 (1995)
10. Schapire, R.: The strength of weak learnability. Machine learning **5**(2), 197–227 (1990)