A Cautionary Tale for Machine Learning Design: why we Still Need Human-Assisted Big Data Analysis

A Cautionary Tale for Machine Learning Design: Why We Still Need Human-Assisted Big Data Analysis

Marco Roccetti, Giovanni Delnevo, Luca Casini, Paola Salomoni

Abstract Supervised Machine Learning (ML) requires that smart algorithms scrutinize a very large number of labeled samples before they can make right predictions. And this is not always true either. In our experience, in fact, a neural network trained with a huge database comprised of over fifteen million water meter readings had essentially failed to predict when a meter would malfunction/need disassembly based on a history of water consumption measurements. With a second step, we developed a methodology, based on the enforcement of a specialized data semantics, that allowed us to extract only those samples for training that were not noised by data impurities. With this methodology, we re-trained the neural network up to a prediction accuracy of over 80%. Yet, we simultaneously realized that the new training dataset was significantly different from the initial one in statistical terms, and much smaller, as well. We had reached a sort of paradox: We had alleviated the initial problem with a better interpretable model, but we had changed the replicated form of the initial data. To reconcile that paradox, we further enhanced our data semantics with the contribution of field experts. This has finally led to the extrapolation of a training dataset truly representative of regular/defective water meters and able to describe the underlying statistical phenomenon, while still providing an excellent prediction accuracy of the resulting classifier. At the end of this path, the lesson we have learnt is that a human-in-the-loop approach may significantly help to clean and re-organize noised datasets for an empowered ML design experience.

Keywords Smart data · Machine learning design · Human-Machine-Bigdata interaction loop · Human-in-the-loop methods · Water metering and consumption

Marco Roccetti
Department of Computer Science and Engineering, University of Bologna, Bologna, Italy
E-mail: marco.roccetti@unibo.it

Giovanni Delnevo
Department of Computer Science and Engineering, University of Bologna, Bologna, Italy
E-mail: giovanni.delnevo2@unibo.it

Luca Casini
Department of Computer Science and Engineering, University of Bologna, Bologna, Italy
E-mail: luca.casini7@unibo.it

Paola Salomoni
Department of Computer Science and Engineering, University of Bologna, Bologna, Italy
E-mail: paola.salomoni@unibo.it

1 Introduction

This time, we would like to begin our research paper with a clear question: What do we talk about when we talk about Machine Learning (or ML, for short)? An oversimplifying, yet precise, answer is that ML algorithms essentially scrutinize large piles of data to recognize hidden relationships buried deep within them, while using patterns, that are eventually found, to classify, interpret and make predictions on a variety of different real-world phenomena [1-3]. Nonetheless, it often goes disregarded that algorithms that learn are, at least for now, just another form of *machine instruction*, still driven and influenced by a continuous gathering of information, with which ML algorithms are fed and trained [4].

Yet, to come to our point: while we can recognize myriad uses for these smart algorithms, even more crucial is to harness the power of the data that can make them sensitive to past experiences. Needless to hide it any longer, one can use a clever ML algorithm, but it is not the elaborate design of that procedure that wins, at the end. Rather, the statistical validity, the sense, the references, the subtle implications, in one single word: the semantics of the data that are used to train the algorithm [5, 6].

Our experience in this field has gone far enough to justify the above statements of our thinking. What we have done, in fact, has been to work on a huge dataset comprised of over fifteen million water meter readings, provided by a company that distributes water in Northern Italy. The aim was to design a ML-based classifier, able to predict defective water meters, based on the history of the measurements of the water that was consumed over the years.

On a first phase of this research activity, we tried to train a neural network with (almost) all of the fifteen million water meter readings, without any specific attention to the quality of those data. What we have inevitably obtained has been a ML model essentially unable to predict, with a sufficient precision, when a given water meter would fail/need disassembly.

With a second step, we developed a methodology, based on the enforcement of a specialized data semantics, that allowed us to extract only those samples for training that were not noised by data impurities. Applying this data semantics, most of those impurities were filtered out, and our neural network was re-trained based on a more stable dataset up to a prediction accuracy of over 80%. Yet, we simultaneously realized that this new training dataset was significantly different from the initial one in statistical terms, and much smaller, as well. We had reached an apparent paradox: We had alleviated the initial classification problem with a better interpretable data model to be used by a ML algorithm, but we had changed the replicated form of the initial data [7-9].

In an effort to reconcile this paradox, with a third and final step, a further new training dataset was obtained, reaching a greater standard of data quality. In close collaboration with domain experts, in fact, a new data semantics was defined with the aim of: i) not only cleaning the initially noised dataset, but also ii) privileging for training just those data that human experts consider representative of what a normal history of water consumption can either be (regular meters) or not be (defective meters). This has finally led to the extrapolation of a training dataset able to represent the underlying statistical phenomenon, while still providing an excellent prediction accuracy of the classifier. Simply told, we have found out that, if we use a sufficient number of readings from this final dataset, there is no statistical paradox at all, and the accuracy of the correspondent ML classifier equals, or even outperforms, that of its predecessor. To summarize our results, we can maintain that a human-in-the-loop approach may significantly help to clean and re-organize noised datasets for an empowered ML design experience.

The remainder of this paper proceeds as follows. In Section 2, we discuss on the importance of the quality of data that are used for training ML algorithms, with some relevant examples, while in Section 3 and 4 we present the initial problem and the first naïve approaches we had adopted to design a ML model able to predict defective water meters, with its correspondent (and not fully convincing) results. Section 5 describes, instead, the final data semantics that was agreed upon with experts and has led both to an improved prediction accuracy of the classifier and to a statistical compliance with the initial dataset. Section 6, finally, concludes the paper.

2 Related Work

The importance of having data that provide an accurate and high-quality description of a given phenomenon is not that new for computer science, even before the advent of the current ML age. This motivates why data quality has been a subject of intense study for many years in the past [10]. Along this line of reasoning, it should be noticed that before the International Organization for Standardization (ISO) had standardized its own *definition* of what quality means for data (i.e., ISO 8000-8:2015, too long to be reported here), many researchers have contended for the most effective one [11].

And so, data quality has been intended in a variety of different ways over the years, ranging, for example, from: “(we have data of) high quality if it is fit for its intended uses in operations, decision making and planning” [12], to: “(quality as) the totality of features and characteristics of data that bears on its ability to satisfy a given purpose; the sum of the degrees of excellence for factors related to data” [13].

Further to a given general definition, more important is to note that, from a more practical standpoint, data quality can be returned in terms of the various dimensions that identify specific facets of the process that utilizes those data. Obviously, those dimensions can vary, depending on both the objectives and the context of use of the data. Nonetheless, to simplify this complex subject, we are often concerned with the quality of data expressed in terms of: accuracy, completeness and consistency [14-16].

While data quality, prior to the big data era, was essentially meant in terms of the aforementioned features (i.e., accuracy, completeness and consistency) yet with a specific attention to the databases that store them, the advent of what we have called *big data* has significantly changed the picture. The specific characteristics of this huge and uninterrupted flow of data (to be treated not only based on its volume, but also facing issues like velocity and value variety) have posed unique challenges to the interpretation of what quality has become. To convince skeptical readers to accept this new challenge, it should be enough to think about the diversity of application fields where the quality of data that are exploited can make a real difference, like for example: health and biology, social media and politics, logistics and transport, just to cite a few [18-22].

At this point, you will end up really disappointed to know that the most relevant ML initiatives disregard data quality as a crucial factor for success. Those who train learning machines, in fact, are still convinced that the focus has to be put more on algorithms and computational infrastructures, rather than on data and their quality [23]. In some sense, in the ML arena, many are those who think that: “Until we have all the data, any other issue comes after”. This is also the motivation behind the scarcity, in the specialized scientific literature, of research papers where issues of ML and data quality are treated with the same level of mindfulness. Fortunately, some happy few examples exist that contradict our previous narration.

For example, Sessions and Valtorta discuss on the influence that the quality of data may have if one wants to learn the structure of a Bayesian network, starting from data [24]. More interestingly, Foidl and Felderer try to assess the risk of using poor data for training a learning machine. Specifically, they discuss how much is to the harm of the final performance of a ML algorithm trying to learn a feature which is described only by low quality data [25].

Our present paper extends this last discussion to a more general ML setting, and also gives considerations to alternative solutions useful to overcome the problem of using low quality data for training a modern ML algorithm, in a realistic situation.

3 Learning from Big (Non-Quality) Data

At the beginning of our research adventure, we were provided with a huge dataset comprised of almost fifteen million readings (plus other contextual information) coming from almost one million water meters. This large dataset spanned a period in time, from the beginning of 2014 to the end of 2018. As already mentioned, all those data were supplied by a company that distributes water over a large area in Northern Italy. To guarantee its privacy, we keep it here anonymized.

The target of this operation was to train a neural network with those data, and then implement an intelligent classifier able to make a prediction on when a given meter would have failed/needed disassembly, based on a history of water consumption quantities (the readings, for short). An additional requirement was that of keeping the number of the consecutive readings, useful for a prediction for a given meter, as small as possible. All this with the intent to try to respond to the request, expressed by the company, to better organize the number and the type of interventions that *human operators* have to carry out *on site*, for water meter maintenance and repair.

Unfortunately, the problem with this initial dataset was that many of the provided readings came with numerous impurities, many of which traced down to a point where different business processes have conflicts that are too complex to be explained in this paper. As a simple example of this complexity, and of its reflection on data, take the fact that, at a certain point of one of these processes, a human operator has to validate a given water reading. In the absence of such positive validation, that reading is to be considered as non-valid and cannot be taken into consideration.

If you look at Table 1 (uppermost part), you can find reported the quantity of valid/non-valid readings, starting from the initial amount of circa fifteen million readings. Precisely: almost 2 million readings are to be considered as non-valid, based on the precise company terms we have explained above.

Not only, out of the 13 million readings *generically* considered as valid, the company further classifies those readings using several different categories (of validity), into which a given reading can fall, depending on the combination of the values that are assigned to three specific fields of the record implementing that reading. This brings to a total combination of some 45 different categories of validity for readings. Yet, just seven of them are the categories of validity into which almost the 99% of the readings, comprised in the dataset, do fall. These seven categories of readings are shown in the middle part of Table 1, defined through the combination of the different values that the three fields, termed X, Y and Z, can assume.

In essence, the higher the position of the category in (the middle section of) Table 1, the more reliable are to be considered the corresponding readings falling in that category. For the sake of secrecy, we have used codes in Table 1 (i.e., capital letters and numbers), that hide the meaning of both the record fields and their values.

Anyway, at this point, it should be clear that the company considers as highly reliable especially all those readings comprised in the category denoted by the following code combination: X=1, Y=Z=2 (see Table 1). For the sake of simplicity, from now on we will identify those readings as those enjoying the so-called *1-2-2 Factor*.

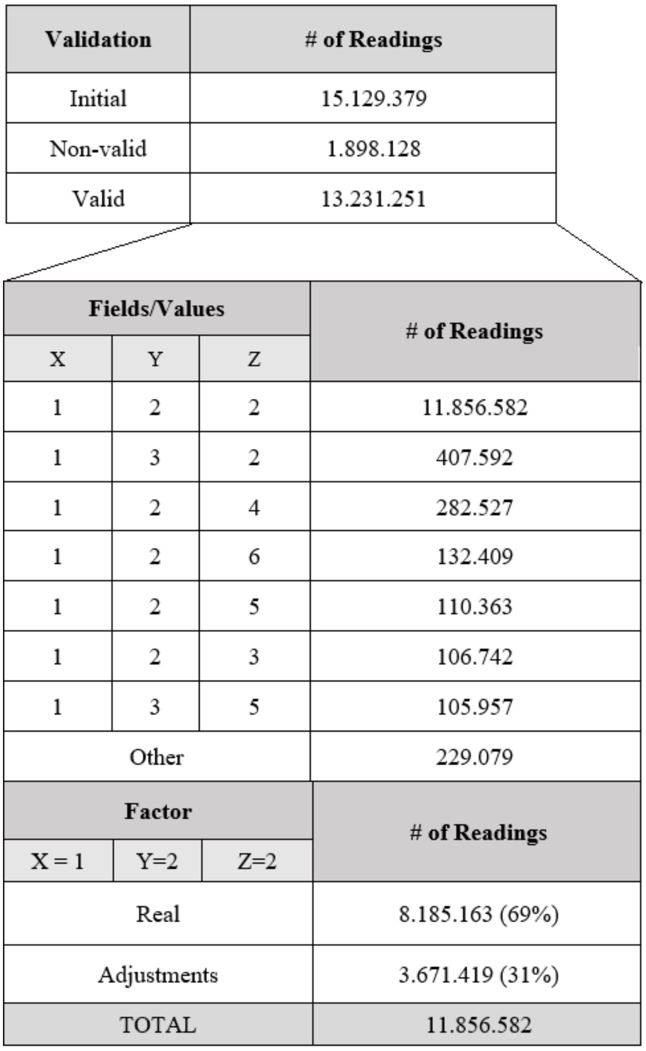
Needless to say, as the company considers all the readings in that category as reliable and safe, they were passed to us as the candidates to be used as positive/negative examples to train a supervised neural network.

Unfortunately, we are not yet at the end of this story. A further analysis of the total amount of readings considered valid by the company as enjoying the 1-2-2 Factor (that is, more than 11 million readings), has revealed that many of them were not genuine measurements of a real water consumption, rather just a kind of an *ex-post* correction of estimates of quantities of consumed water, computed on the basis of a certain mathematical model.

What we are saying is that not all the 11 million readings mentioned above correspond to real water consumption measurements taken on the field, rather they are complex mathematical re-adjustments of presumed values of water consumption; simply put, they are just *rough approximations* of real values. The balance between real measurements vs. re-adjustments is shown at the bottom of Table 1, yielding an amount of 8.186.163 *real* readings enjoying the 1-2-2 Factor.

Not only, prior to beginning our machine learning activity, we tried to define the sets of positive and negative examples to be shown to our algorithm. In essence, we used our 8.186.163 *real* readings, all enjoying the 1-2-2 Factor, to precisely identify the sets of both defective and non-defective meters to be used as positive and negative examples, respectively. The results of this latter activity are shown in the first two lines of Table 2, where counted is the total number of water meters with readings enjoying the 1-2-2 Factor (first line, Table 2), contrasted against the quantity of faulty meters with readings enjoying the 1-2-2 Factor (second line, Table 2).

Table 1. Water meter readings: their validity, as seen from a company perspective



Validation	# of Readings
Initial	15.129.379
Non-valid	1.898.128
Valid	13.231.251

Fields/Values			# of Readings
X	Y	Z	
1	2	2	11.856.582
1	3	2	407.592
1	2	4	282.527
1	2	6	132.409
1	2	5	110.363
1	2	3	106.742
1	3	5	105.957
Other			229.079

Factor			# of Readings
X = 1	Y=2	Z=2	
Real			8.185.163 (69%)
Adjustments			3.671.419 (31%)
TOTAL			11.856.582

Table 2. Water Meters (with readings enjoying the 1-2-2 Factor)

1-2-2 Factor	# of Meters
Total	1.177.806
Faulty	23.752
1-2-2 (at least 1 reading)	1.154.054
1-2-2 (at least 2 readings)	1.091.334
1-2-2 (at least 3 readings)	1.038.337
1-2-2 (at least 4 readings)	981.420
1-2-2 (at least 5 readings)	915.441

However, this kind of information is not yet complete, as we have to consider also *how many* valid readings we have for each meter; where valid, here, means that the 1-2-2 Factor is satisfied.

Within the five lines in the lowermost part of Table 2, we report the number of all the water meters comprised in our dataset, that have a history with, respectively, at least: 1, 2, 3, 4, and 5 readings, all enjoying the 1-2-2 Factor.

Finally, what is still missing in this narration is the role played by time.

In essence, of great importance is the time past between two consecutive and valid readings read on a given meter. In fact, to correctly train a neural network, crucial is the regularity of the frequency with which a reading is read over time. Unfortunately, we have no good news here.

Figure 1 provides insightful information with this regard. On the *x axis* of Figure 1, plotted are the differences of the two values (i.e., cubic meters of consumed water) recorded at two subsequent consecutive readings, while on the *y axis* we can see the time intervals (measured in days) between those readings.

This Figure 1 essentially summarizes some millions of recorded reading values taken over time, and shows very clearly the type of *dispersion* that is experienced both in terms of time and of measured values, even for readings enjoying the 1-2-2 Factor. (Note that negative values can be observed on the x axis, owing to the phenomenon of the ex-post re-adjustments of some readings explained before).

Summing up, looking at this problem from all the possible perspectives, the data that were made available to us, in their initial form, cannot be considered a good starting point to train a machine, as they can hardly provide unambiguous examples to be learnt by a learning algorithm.

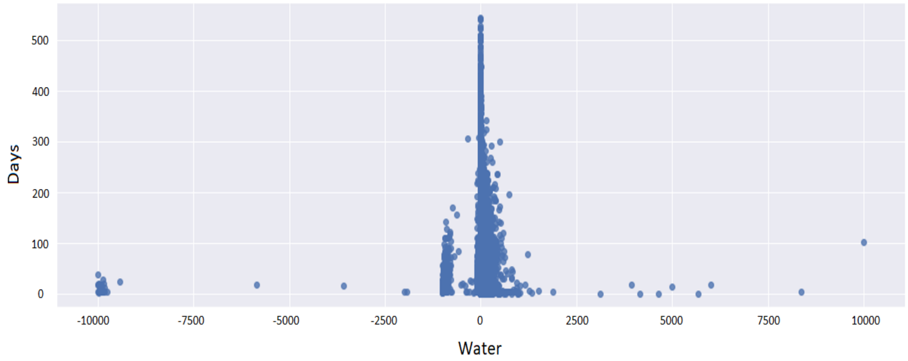


Figure 1. Time intervals vs differential water consumption (between two consecutive readings)

The confirmation of all our doubts concerning the appropriateness of these data for training a machine came definitely with our first attempt to train a recurrent neural network, using positive/negative examples assembled by sampling faulty and non-faulty meters, with their corresponding readings, from those mentioned in Table 2.

We employed a deep neural network with two parallel inputs: a feed forward one for the contextual attributes of a water meter and a recurrent one for its water reading series. We implemented it using the Keras library. We experimented with series of readings of different lengths, containing either two or three consecutive readings, using data taken from the period beginning 2014 – mid 2018.

Upon completion of the training activity, we moved to the testing phase that was carried out on that set of data that the network never looked at before (that is, readings taken in the period: mid 2018 - end 2018).

Unfortunate, yet well expected, was the final result, in terms of accuracy of the prediction. Using the standard AUC-ROC metrics, we never were able to surpass the AUC value of 61%. In simple words, our classifier, trained on the data at the center of our current discussion, was never able to predict a possible water failure/disassembly event with an accuracy larger than 0.61.

4 Enhancing Big Data with Semantics: Pros and Cons

To make all the complex piles of data described above genuinely valid for carrying out a learning activity, we moved towards an alternative approach, at the basis of which lies a procedure developed to clean our data. In essence a novel semantics of data validity was defined to be satisfied by our readings that can be summarized as follows.

A reading is valid only if each of the following requirements is satisfied:

- a human operator has read and confirmed that reading, on the reading site;
- that reading value has been correctly transmitted and recorded onto the company ERP;
- that reading value has been definitely certified by another human operator on the company ERP; and, finally,
- the instants in time when that reading was taken and then certified are coherent: i.e., certified as real and time-congruent, by a specific process.

Said simpler, this new semantics confirms the 1-2-2 Factor, with the addition that valid readings must possess a certification on the validity of the dates when a given reading has been read, and then processed by human operators. For the sake of simplicity, we will call this *enhanced data semantics*, from now on, as the *X-Factor*.

At this point, we enforced the aforementioned X-Factor semantics on our initial data, with the result of reducing the number of valid readings down to two million valid readings. After that, we re-trained our recurrent neural network on a sample of faulty and non-faulty meters, with their corresponding (circa 135.000) valid readings, enjoying that X-Factor.

After this new training activity, we subjected our neural network to a new testing phase, using again all the valid data of the period mid 2018 – end 2018.

Results from this test are plotted in Figure 2, where the AUC-ROC metrics is used to measure the accuracy of the obtained predictions. It is easy to see that now water meter failure events are predicted with an improved precision, in the range of [86 - 89] %, depending on the number of consecutive readings exploited to make the prediction (2 readings vs. 3 readings).

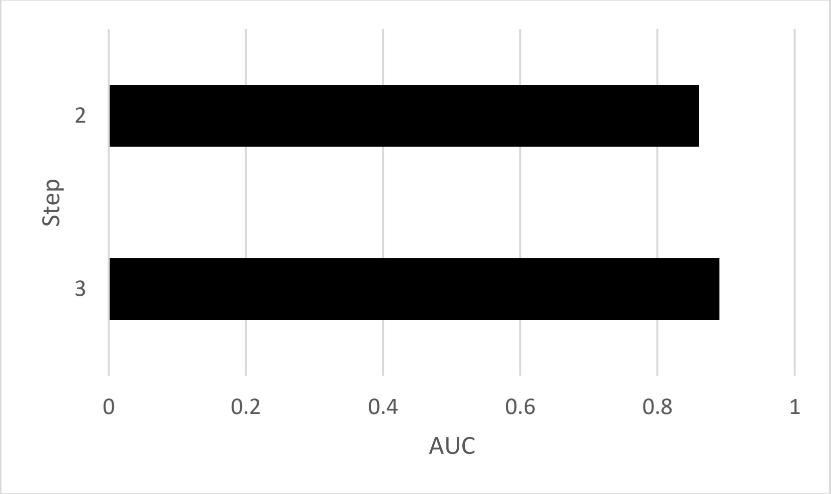


Figure 2. A dataset with an X-Factor: improved prediction accuracy

At this point, we were satisfied with the prediction accuracy reached after having enforced the so called X-Factor on our data, but we began to reflect on the statistical meaning and validity of the operations we had carried out to reach our result.

We simply asked ourselves: Did the X-Factor semantics just realign our data and clean them from the initial impurities, or has it actually transformed them from some (statistical) viewpoint? To answer to this question, we conducted various statistical tests.

To start the narration of this kind of analysis, we report in Table 3 the total number of readings, respectively: i) from the initial dataset, ii) from the dataset cleaned with our X-Factor semantics, and finally iii) belonging to the sample (with the X-Factor) that was used for training our neural network.

In Table 3, reported are also the average (μ) and standard deviation (σ) values for the correspondent set of water readings (in terms of cubic meters of consumed water).

Table 3. Water readings: statistics

Id	1	2	3
Dataset	Initial	Total with X-Factor	Sampled for training
# of Readings	13.231.251	1.973.493	135.018
μ	5.307	3.674	3.647
σ	86.450	17.796	11.852

The values of averages and standard deviations reported in Table 3, and also a simple observation of the shapes of the curves portrayed in Figure 3 below, have raised our doubts.

To understand, consider Figure 3. It shows two curves: they both aim to measure the number of water meter readings (on the y axis), whose average value equals a given value, say X (on the x axis). Gray is the color for the case of the initial (not filtered) dataset, while in Black we have the X-Factor case.

Comparing the two plots, we have a clear impression that the shapes of the two curves are somewhat different, because the value of each Gray reading (initial dataset) is always much larger than the value of the correspondent Black reading (X-Factor), for each given value X of consumed water.

To better understand this phenomenon, we conducted some statistical tests. In other words, we wanted to have a confirmation to the hypothesis that the average quantity of consumed water, as measured by the readings comprised in the initial dataset, was larger than that measured by those readings chosen by using the X-Factor semantics.

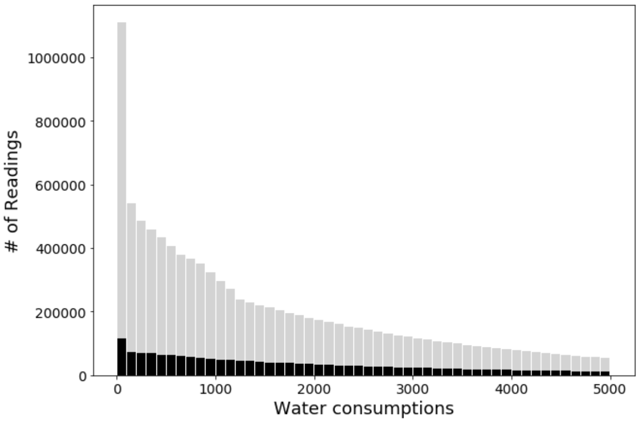


Figure 3. Water consumptions: initial dataset (gray); dataset with the X-Factor (black)

We assumed normal distributions (with known values for the average and standard deviation values) and proceeded with a Z Test, whose results are reported in the uppermost part of Table 4. We tested our null hypotheses (the two average values are equal) with two different significance α factors.

As seen from the results of Table 4, the null hypothesis that the average quantity of consumed water as per the readings of the initial dataset and the average quantity of consumed water as per the readings subjected to the X-Factor are equal is to be rejected. Instead, as expected, it cannot be rejected the null hypothesis that those specific readings with the X-Factor we used for training show an average quantity of consumed water equal to the average quantity of the consumed water as per all the readings with the X-Factor.

To have a further confirmation, we repeated the same kind of test, yet with a different statistic. Simply, we tried to use a Student's T test (with an unknown standard deviation). This has to be intended just as an additional attempt to confirm the previous results and, in fact, not surprisingly, we got very similar outcomes, as shown in the lowermost part of Table 4.

In conclusion, while it is true that training a recurrent neural network with just those data cleaned by our semantics has improved the prediction accuracy of the resulting classifier, on the other side we have reached a (statistical) paradox that can be expressed as follows:

As we aim to improve the ML performances in terms of accuracy of the predictions, enabling the transformation of the initial data through re-organization, we simultaneously change the replicated forms of those data.

Table 4. **Z** and **T** Tests: results

Test	μ1 = μ2			μ2 = μ3		
	p-value	α=0.05	α=0.01	p-value	α=0.05	α=0.01
Z Test	<10-5	Reject	Reject	<10-5	Fail to reject	Fail to reject
T Test	0,75	Reject	Reject	0,62	Fail to reject	Fail to reject

5 An Empowered ML Design with Humans in the Loop

Following the controversial phase described at the previous Section 4, we had a debate with the company experts and decided to agree upon a new data semantics (extending the one we had previously proposed); with the intent to reconcile the statistical paradox that emerged from the transformation of the dataset due to the use of the X-Factor.

To do that, we have to introduce first the concept of *plateau*. A plateau is to be intended as a series of two or more consecutive water readings, whose values do not change over time, since that meter is faulty.

This said, the set of validation criteria which were agreed upon with the company experts can be summarized as follows (note that those in italics were already comprised in the X-Factor):

1. *A human operator has read and confirmed a reading, on the reading site;*
2. *That reading value has been correctly transmitted and recorded onto the company ERP;*
3. *That reading value has been definitely certified by another human operator on the company ERP; and, finally,*
4. *The instants in time when that reading was taken and then certified are coherent: i.e., certified as real and time-congruent, by a specific process,*
5. The time separating two consecutive readings cannot exceed the value of seven months,
6. When a **plateau** is found in a defective meter, only the first reading of this plateau is taken into consideration for training, and then we go back in the series to choose other previous readings to be used for training our recurrent neural network.

To provide further motivations on the meaning of Rules 5 and 6 above, please consider what follows. Rule 5 has emerged from the consideration that water readings that are too far apart in time can lead our neural network to learn water consumption patterns whose discontinuity cannot be related to any reasonable human behavior.

Understanding the meaning of Rule 6 is, instead, somewhat trickier. The point is as follows. It goes undisputed that a long series of readings whose values do not change over time amount to a situation where either the client has quit to consume water or the meter has become defective. Nonetheless, there would be no need to instruct a complex learning algorithm, if we simply want to use this unique information to infer that a water meter has broken. Two lines of code would be enough: "if reading(n)=reading(n-1)=reading(n-2), then the meter is broken".

But we want much more than this. We want to anticipate as much as possible the instant in time when the meter is going to break (this is, actually, the meaning of "making a prediction"). Only an attempt to anticipate a meter failure event can motivate our idea to train a machine that learns by examining the readings that come before a *plateau*.

This said, we tried to train our recurrent neural network by exploiting two alternative usages of Rule 6. In one case, we used for training just the first reading of a given plateau (plus other readings that come before it). This situation was denoted with the abbreviation **1P**. In the second case, we used the first two (identical) values of a plateau, plus other readings that come before that value. This second situation was denoted with the abbreviation **2P**. The difference is clear. If we succeed with just the 1P approach, this means that our neural network has achieved a very good ability to *anticipate* a fault in the meter, almost without any need to know that a plateau has come.

At this point, we re-trained our network by applying this extension of our data semantics. In particular, 18.949 non-faulty meters were utilized with their corresponding readings, plus 2.279 defective meters with readings of type 1P and 2.260 meters with readings of type 2P. Then, we subjected our re-trained neural network to a subsequent testing phase. Experiments were conducted with series of consecutive readings that were respectively as long as: 2, 3, 4, 5.

The results are portrayed in Figure 4, where the black histograms represent the case where defective meters with readings of type 1P are employed, while gray histograms represent the case where defective meters with readings of type 2P are utilized. The adopted metrics for the prediction accuracy is the usual AUC-ROC.

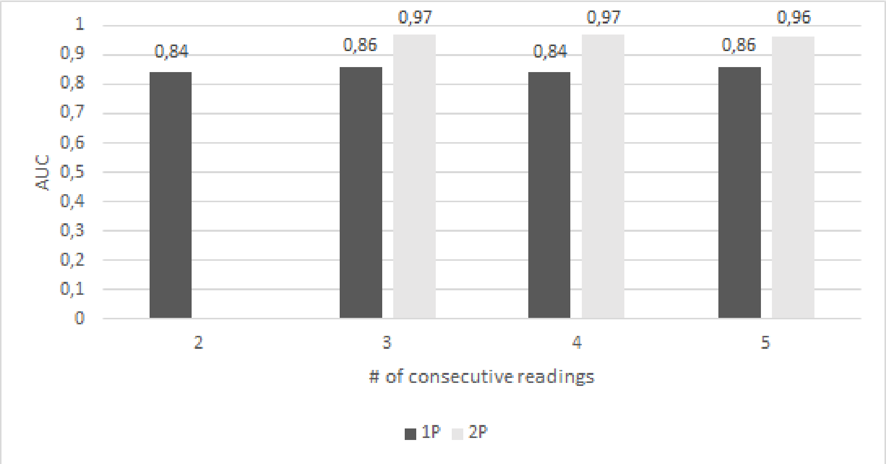


Figure 4. Accuracy of the classifier after the enforcement of the new data semantics

Two are the issues emerging in Figure 4 that deserve a comment.

The first one is that the classifier, obtained by training a recurrent neural network with the data subjected to our new extended semantics, achieves a very good performance, on average, in terms of prediction accuracy, with AUC values in the range [84 - 96] %.

The second one is that, while it is evident that the situation that makes use of the 2P case outperforms the one using the 1P type of information, nonetheless we can maintain that our ML model has an excellent prediction ability even in the case where no explicit information about the plateau is provided during the training phase. In fact, even with the 1P case we yield prediction accuracy values in the range [84 - 86] %.

To conclude, it is now the time to understand if the extension we have made to our data semantics has been also of some utility to overcome the statistical paradox we have illustrated in the previous Section 4.

To do that, we conducted a new Z test to verify if we can reject the null hypothesis that the readings of the initial dataset show an average quantity of consumed water that equals that exhibited by those readings that were used to train our neural network at this final phase of our experience.

The results are really interesting and portrayed in Table 5. They can be summarized as follows. With a number of water readings equal or less than three (P readings plus regular ones), we are still in presence of the statistical paradox discussed in Section 4, as the null hypothesis above comes rejected. Instead, if we use four readings or more (P readings plus regular ones), the null hypothesis cannot be rejected, and hence we cannot say that we have a clear discrepancy between the two datasets (namely, the initial one and that used for training our ML model) from a statistical viewpoint. What is convincing here is the fact that the more we add readings to our series selected for training the more we achieve a general statistical congruence.

Table 5: **Z Test**: initial vs sampled for training

Z Test	p-value	Result	**Z Test**	p-value	Result
1P + 1reading	0.007	Reject			
1P + 2readings	0.029	Reject	2P + 1reading	0.028	Reject
1P + 3readings	0.28	Fail to Reject	2P + 2 readings	0.29	Fail to Reject
1P + 4readings	0.65	Fail to Reject	2P + 3readings	0.70	Fail to Reject

Without any intention to further elaborate on the deep technical motivations behind this result, it goes unquestioned the fact that the intervention of field experts has given a fundamental contribution to the development of a data semantics on the basis of which water readings (and water meters) were finally chosen that have trained our neural network, adequately. We say *adequate*, in the sense that both the classifier makes, at the end, reliable predictions and no statistical discrepancy emerges between the huge dataset with all the readings we were provided initially with, and the subset of only those readings selected for the final training activity.

6 Conclusions

In this paper, we have presented our experience with a huge dataset comprised of over fifteen million water meter readings, with the intent to design a ML-based classifier, able to predict defective water meters based on the history of the measurements of the water that was consumed.

On a first phase, we trained a neural network with (almost) all of the fifteen million water meter readings, unfortunately resulting into a definitive failure of our attempt to predict with a sufficient precision when a given water meter will fail.

With a second step, we selected samples for training based on a naïve data semantics that has allowed us to achieve a good prediction accuracy (over 80%). Yet, simultaneously this new training dataset resulted significantly different from the initial one, in statistical terms.

With a third and final step, we elaborated a more complex data semantics, agreed upon with experts of this specific application field, that allowed us to use samples for training that still got excellent performances in terms of prediction accuracy, while guaranteeing a general statistical compliance of the underlying phenomenon. All this has confirmed our hypothesis that even complex form of machine intelligences can benefit by the role that human experts can play.

Acknowledgements

We are indebted towards the company that has provided us with the data of interest. Our thanks also go to Giuseppe Cappiello and Nicolò Zagni (University of Bologna)

for their participation to a previous phase of this research activity.

References

1. Pettersen, L. (2018) Why Artificial Intelligence will not outsmart complex knowledge work. Work, Employment and Society. Sage. To appear.
2. Jordan, M. I., Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.
3. Delnevo, G., Roccetti, M., Mirri, S. (2019). Intelligent and good machines? The role of domain and context codification, Mobile Networks and Applications, Elsevier. To appear.
4. Witten, I. H., Frank, E., Hall, M. A., Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.
5. Alkowaileet, W., Alsubaiee, S., Carey, M., Li, C., Ramampiaro, H., Sinthong, P., Wang, X. (2018). Enhancing Big Data with semantics: The AsterixDB approach. In Proc. of 12th IEEE International Conference on Semantic Computing, 314-315. IEEE.
6. Emani, C. K., Cullot, N., Nicolle, C. (2015). Understandable big data: a survey. Computer Science Review, 17, 70-81.
7. Casini, L., Delnevo, G., Roccetti, M., Zagni, N., & Cappiello, G. (2019, August). Deep Water: Predicting water meter failures through a human-machine intelligence collaboration. In International Conference on Human Interaction and Emerging Technologies (pp. 688-694). Springer, Cham.
8. Roccetti, M., Delnevo, G., Casini, L., Zagni, N., & Cappiello, G. (2019, September). A paradox in ML design: less data for a smarter water metering cognification experience. In Proceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good (pp. 201-206). ACM.
9. Roccetti, M., Delnevo, G., Casini, L., & Cappiello, G. (2019). Is bigger always better? A controversial journey to the center of machine learning design, with uses and misuses of big data for predicting water meter failures. Journal of Big Data, 6(1), 70.
10. Wang, R. Y., Storey, V. C., & Firth, C. P. (1995). A framework for analysis of data quality research. IEEE Transactions on Knowledge & Data Engineering, (4), 623-640.
11. ISO 8000-8:2015, https://www.iso.org/obp/ui/#iso:std:iso:8000:-8:ed-1:v1:en
12. Juran, J., & Godfrey, A. B. (1999). Quality handbook. Republished McGraw-Hill, 173-178.
13. Kodra, Y., De La Paz, M. P., Coi, A., Santoro, M., Bianchi, F., Ahmed, F., ... & Taruscio, D. (2017). Data quality in rare diseases registries. In Rare Diseases Epidemiology: Update and Overview (pp. 149-164). Springer, Cham.
14. Scannapieco, M., Missier, P., & Batini, C. (2005). Data quality at a glance. Datenbank-Spektrum, 14(January), 6-14.
15. Sidi, F., Panahy, P. H. S., Affendey, L. S., Jabar, M. A., Ibrahim, H., & Mustapha, A. (2012, March). Data quality: A survey of data quality dimensions. In 2012 International Conference on Information Retrieval & Knowledge Management (pp. 300-304). IEEE.
16. Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. Communications of the ACM, 45(4), 211-218.
17. Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. Journal of management information systems, 12(4), 5-33.
18. Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. Data science journal, 14.
19. Chen, H., Hailey, D., Wang, N., & Yu, P. (2014). A review of data quality assessment methods for public health information systems. International journal of environmental research and public health, 11(5), 5170-5207.
20. Chen, J. V., Su, B. C., & Widjaja, A. E. (2016). Facebook C2C social commerce: A study of online impulse buying. Decision Support Systems, 83, 57-69.
21. Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., & Bork, P. (2002). Comparative assessment of large-scale data sets of protein–protein interactions. Nature, 417(6887), 399.
22. Burggräf, P., Dannapfel, M., Förstmann, R., Adlon, T., & Fölling, C. (2018, January). Data quality-based process enabling: Application to logistics supply processes in low-volume ramp-up context. In 2018 International Conference on Information Management and Processing (ICIMP) (pp. 36-41). IEEE.
23. Breck, E., Polyzotis, N., Roy, S., Whang, S. E., & Zinkevich, M. (2018, January). Data Infrastructure for Machine Learning. In SysML Conference.
24. Sessions, V., & Valtorta, M. (2006). The Effects of Data Quality on Machine Learning Algorithms. ICIQ.
25. Foidl, H., & Felderer, M. (2019, August). Risk-based data validation in machine learning-based software systems. In Proceedings of the 3rd ACM SIGSOFT International Workshop on Machine Learning Techniques for Software Quality Evaluation (pp. 13-18). ACM.