Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 162 (2019) 865–872

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Conference on Information Technology and Quantitative Management (ITQM 2019)

A fuzzy linguistic supported framework to increase Artificial Intelligence intelligibility for subject matter experts

Juan Bernabé-Moreno[a,b,1], Karsten Wildberger[b]

[a]*Department of Computer Science and A.I., University of Granada, Granada E-18071, Spain*
[b]*E.ON SE, Essen DE-45131, Germany*

Abstract

The application of artificial intelligence (AI) techniques in the decision making processes is more widespread in the industry than ever before. Yet, one of the most critical show-stoppers is the communication gap between the machine learning (ML) models and the experts community. On one hand, the output of ML is often not intelligible for experts, in spite of the latest advances in explainable AI. On the other hand, the expert knowledge, rarely completely present in the available data, but rather in the heads of the experts, needs to be connected to the data-driven insights created by the ML model. In this paper we first identify the most critical situations with a manifest intelligibility gap and then propose a framework supported by fuzzy linguistic modelling techniques to close this gap. In addition, we present its integration into the end-to-end decision making flow, from data gathering to the execution and evaluation and we show the output of our approach with practical examples.

Keywords: fuzzy linguistic modeling, expert knowledge modelling, decision making, intelligible AI

1. Introduction

In the recent years, AI has experienced an unprecedented adoption in the industry. After many years heavily investing in big data and cloudification initiatives, companies are implementing intelligent systems to harness the value of corporate data. Algorithms can support business decisions in a supposedly most reliable and most efficient manner than traditional business analysts or subject matter experts. ML techniques enable the creation of models to infer business knowledge from the historical data gathered for a particular process. These models typically provide a solid performance in isolated theoretical environments, but harnessing the real value requires embedding them as integral part of the business operations, which requires the interaction with subject matter experts.

In fact, the incorporation of ML output into decision making processes, especially when humans are intended to make sense of its output, has been identified as one of the critical points preventing a more widespread AI adoption. According to Weld et al [1] the key challenge for designing intelligible AI is communicating a complex computational process to a human, but the problem has different facets. First, not all outputs of machine learning

*Corresponding author: juan.bernabe@webentity.de

1877-0509
10.1016/j.procs.2019.12.061

models are equally useful and can be operated in the same way (e.g. propensity models providing high scores for an almost impossible marketing addressable audience, etc.). Second, there are always certain preconceived hypothesis or believes in experts' minds that, if compliant with the output of the machine learning model, can be substantially more valuable to the business than any other "new" insights, just because the mechanisms to act on them are clearer or even already existing (e.g. the ability to execute a particular campaign). Likewise, some expert knowledge items are often contradicted by the model and experts need to be made aware of them. Third, in spite of the advanced approaches to explain machine learning models [2, 3], the format of the explanations is usually too specific and too concrete for experts to thoroughly exploit them in decision making processes. In addition, experts usually can't express exact quantifications of their believes or knowledge, neither make sense of too exact quantifications of the importance of a particular combination of attributes. Rather, they typically use broader, less precise linguistic quantifiers, such as "more", "very likely", "less likely", "substantially more", etc. (which introduces the challenge of lack of standardization in the usage of these quantifiers [4]).

For example, if we consider a ML model to determine the propensity of a customer to churn, an expert can probably tell you that according to her/his experience, the customers in a particular contract, in a particular age range and living in a particular area are more likely to churn than others. A machine learning model, for example [5], can exactly assign a churn propensity score to any customer and an interpretability model, such as LIME [2], can provide us with the "rationale" of this particular score based on to which extend values for different attributes support or contradict the evidence found by the model. But even with the LIME explanations, a customer retention expert might struggle defining the proper audiences to target in a campaign, as the output lacks intelligibility.

In this paper, we propose a novel *fuzzy linguistic modeling based method* to increase the intelligibility of the ML model output to the community of experts and therefore speed up the adoption of AI in corporate environments. Fuzzy linguistic modeling works as a mechanism to enable the interaction between subject matter experts and the ML model, typically by making experts knowledge items intelligible to the model and model findings intelligible to the experts, always trying to minimize the information loss. Fuzzy linguistic modeling has proven its performance implementing algebraic operations on natural language experts opinions [6] and that's why we propose this technology to bridge the ML model-2-human communication gap.

The main contributions of this paper are listed below:

- We have identified 4 situations taking place in the process of operationalizing the output of the ML model, that present a substantial intelligibility gap between model/model explainers and the community of experts.
- We have proposed a mechanism to validate expert knowledge against a ML model using *linguistic fuzzification and defuzzification* as well as a fuzzy linguistic modeling based method to consolidate the expert knowledge related to the ML model of two or more experts.
- We have enhanced the standard ML explainers (LIME, SHAP and features importance) with a mechanism to extract experts-intelligible knowledge items.

This paper is organized as follows: after introducing the problem and explaining the novelty of our approach, we review the supporting research background. Then, we introduce the 4 identified scenarios we aim at increasing the intelligibility of the ML model, define the fuzzy linguistic components required to implement our system and describe how our method tackles the intelligibility issues with real examples. After discussing the results, we then finalize the paper providing the concluding remarks and pointing to further research lines.

2. Background

2.1. On Machine Learning Intelligibility

Given the huge performance improvement of the Deep Neural Networks and other black-box ML models and the need for understanding the rationale of the ML predictions, the field of explainable AI is experiencing an unprecedented research activity. Molnar in [7] provides a complete overview of the state of the art in this field, including implementation details of the most popular explainers (LIME [2], SHAP [3], etc). While the first explainers were model specific, Ribeiro et al. [2], explained the advantage of embracing model agnostic approaches to give ML developers the choice of the modeling method but also to enable the comparability of models without having to change the model explainer.

According to Dietterich et al. [8], in order to trust deployed AI systems, we must not only improve their robustness, but also develop ways to make their reasoning intelligible. Intelligibility will reportedly help us spot AI that makes mistakes due to distributional drift or incomplete representations of goals and features. Intelligibility will also facilitate control by humans in increasingly common collaborative human/AI teams, and Intelligibility will help humans learn from AI. In addition, these authors pointed out legal reasons implement intelligible AI, including the European GDPR and a growing need to assign liability when AI errors. Lundberg and his co-authors [3] establish the need to 1) ensure that the underlying reasoning or learned models are inherently interpretable and 2) if it is necessary to use an inscrutable model to prevent decreasing the predictive quality, such as complex neural networks or deep-lookahead search, then mapping this complex system to a simpler, explanatory model for understanding and control. Gilpin [9] on the other hand defends the need to design models that are inherently interpretable, outlining several key reasons why explainable black boxes should be avoided in high-stakes decisions and identifying challenges to interpretable ML.

The use of fuzzy logic in machine learning systems is experimenting a great adoption and being intensively reseached[10]. Couse et al. in [11] discuss the development the internal shift from largely knowledge-based to strongly data-driven fuzzy modeling and systems design due to the increasing integration of fuzzy logic and machine learning. Bonanno et al. in [12] demonstrated the use of fuzzy inference to turn deep neural networks into rule-based explained systems.

While the advances in ML interpretability are paving the way towards increased intelligibility, there is still a gap not being addressed as explained in the previous section, which we are covering in the present research paper.

2.2. On Fuzzy linguistic modelling

The fuzzy linguistic approach is a tool based on the concept of linguistic variable proposed by Zadeh [13]. This theory has given very good results to model qualitative information and it has been proven to be useful in many problems.

The 2-Tuple Fuzzy Linguistic Approach

The 2-Tuple Fuzzy Linguistic Approach [6] is a continuous model of information representation that allows reduction in the loss of information that typically arises when using other fuzzy linguistic approaches, both classical and ordinal [14]. To define it both the 2-tuple representation model and the 2-tuple computational model to represent and aggregate the linguistic information have to be established.

Let $\mathscr{S} = \{s_0, ..., s_g\}$ be a linguistic term set with odd cardinality. We assume that the semantics of labels is given by means of triangular membership functions and consider all terms distributed on a scale on which a total order is defined. In this fuzzy linguistic context, if a symbolic method aggregating linguistic information obtains a value $\beta \in [0, g]$, and $\beta \notin \{0, ..., g\}$, we can represent β as a 2-tuple (s_i, α_i), where s_i represents the linguistic label, and α_i is a numerical value expressing the value of the translation between numerical values and 2-tuple: $\Delta(\beta) = (s_i, \alpha)$ y $\Delta^{-1}(s_i, \alpha) = \beta \in [0, g]$ [6].

In order to establish the computational model negation, comparison and aggregation operators are defined. Using functions Δ and Δ^{-1}, any of the existing aggregation operators can be easily extended for dealing with linguistic 2-tuples without loss of information [6].

Multi-Granular Linguistic Information Approach

To accommodate the interaction with different experts, it's important to support different "granularity levels". For instance, an expert might use a term set of 3 categories only ("low", "mid", "high") to provide domain related expertise statements: (e.g. "customers who used the hotline more than 3 times have *higher* likelihood to churn"), while a different one might be used to more levels (e.g. "customers interacting more frequently with the call center after receiving a bill have a *substantially higher* likelihood to churn). Thus, supporting different granularities and providing tools to manage the multi-granular linguistic information is key to our approach. Formally, when different experts have different uncertainty degrees on the phenomenon or when a single expert has to evaluate different concepts, then several linguistic term sets with a different granularity of uncertainty are necessary [15]. In such situations we need tools to manage the multi-granular linguistic information. In [16] a multi-granular 2-tuple fuzzy linguistic modelling based on the concept of linguistic hierarchy is proposed.

A *Linguistic Hierarchy*, *LH*, is a set of levels $l(t, n(t))$, where each level t is a linguistic term set with different granularity $n(t)$. The levels are ordered according to their granularity, so that we can distinguish a level from the

previous one, i.e., a level $t+1$ provides a linguistic refinement of the previous level t. We can define a level from its predecessor level as: $l(t,n(t)) \rightarrow l(t+1, 2 \cdot n(t)-1)$. In [16] a family of transformation functions between labels from different levels was introduced. To establish the computational model we select a level that we use to make the information uniform and thereby we can use the defined operator in the 2-tuple model. This result guarantees that the transformations between levels of a linguistic hierarchy are carried out without loss of information. Using this LH, the linguistic terms in each level are the following (see also Fig. 4 and Fig. 3):

- $S^3 = \{b_0 = None = N,\ b_1 = Medium = M,\ b_2 = Total = T\}$
- $S^5 = \{c_0 = None = N,\ c_1 = Low = L,\ c_2 = Normal = N,\ c_3 = High = H,\ c_4 = Total = T\}$
- $S^9 = \{d_0 = None = N,\ d_1 = Very\_Low = VL,\ d_2 = Low = L,\ d_3 = More\_Less\_Low = MLL,\ d_4 = Medium = M,\ d_5 = More\_Less\_High = MLH,\ d_6 = High = H,\ d_7 = Very\_High = VH,\ d_8 = Total = T\}$

3. Fuzzy linguistic modelling system to increase machine learning intelligibility

Before we present our approach, we'd like to introduce some definitions that we will be using all along to explain our system:

- **Definition 1** An *Expert knowledge item* is a believe related to a knowledge domain, usually expressed as a set of quantified predictors and their impact on the class variable. Sometimes, a constraint statement might be added to limit the scope of the data. This constraint is usually built as a composition of quantified predictors.
- **Definition 2** A *Class variable* represents the problem we are trying to tackle or we are trying to provide expertise about in the context of an Expert Knowledge Item.
- **Definition 3** A *Predictor* is an attribute or feature in the data whose value impacts the likelihood of the class variable to take a value or another one.

We focus on the intelligibility problem once both ML Model and Explainers have been created. In Fig. 1, we can see the general ML process, from the *Problem* definition to the *Take-to-Action*. As we can appreciate, a new module between the *Model*, the *Prediction* and the *Explainers* has been added and connects with the *Action* part itself. This new module consists of 4 different situations where the lack of intelligibility manifests: a) existing expert knowledge compatibility with the machine learning model (*Expert-2-Model*), b) consolidation of knowledge from multiple experts in accordance with the model (*Expert-2-Expert*), c) output of model explainers to humans (*Model-2-Expert*), and d) feature importance to humans (*Feature-2-Expert*).

For each situation, we have specified a pipeline to transform the input into an intelligible output to support the experts-model interaction therefore the operationalization of artificially created intelligence. Fig. 2 depicts the different processes and components, with the proper Input and Output definition, required to address these 4 situations. The different components required to implement the information pipelines to address the identified

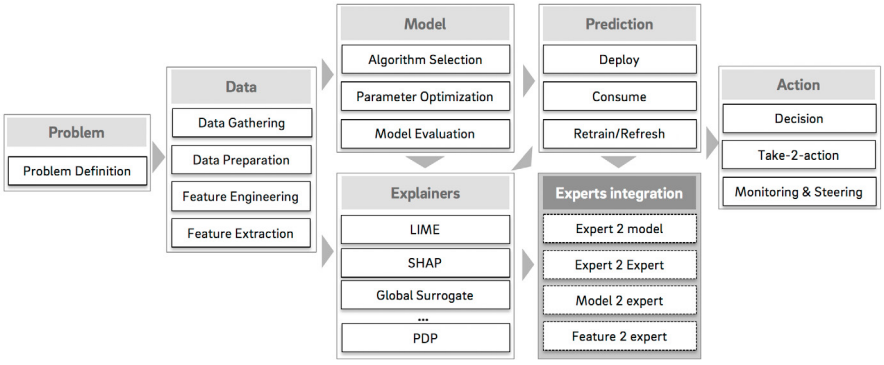

Fig. 1. From problem to actions - Integrated machine learning approach

situations are described below:

- *Quantifiers extractor*: identifies the linguistic quantifiers at attribute level. It can be applied to both predictors and predicted class. It is implemented as explained in 2.2 as a linguistic variable *LH* per qualifier supported by a linguistic hierarchy *LH*
- *Linguistic Defuzzifier* specifies a point in the attribute space $z*$ that best represents the linguistic term c_i. [17].There are a number of choices in determining the crisp output $z*$, all of them trying to fulfil following criteria: Plausibility ($z*$ should represent c_i from an intuitive point of view (e.g.: it may lie approximately in the middle of the support of c_i or has a high degree of membership in c_i.), computational simplicity and continuity (small changes in c_i should not result in a large change in $z*$ and dis-ambiguity (defuzzification should always produce an unique value for $z*$).
- *Model-based evidence checker*: applies explainability algorithms to quantify to which extend a particular statement is supported by the machine learning model. Returns a supporting value between -1 and 1 (-1 meaning a strong support for the opposite statement and 1 fully support in all cases).
- *Linguistic Fuzzifier* converts a crisp input to a term defined for a linguistic variable, as explained in the subsection 2.2. Therefore, the fuzzifier can be defined as a mapping from an observed input space to fuzzy set labels in a universe of specified input universe of discourse.
- *Knowledge-items matcher*: provides a score from 0 to 1 determining whether 2 knowledge items refer to the same entities, for both predictors and class variable (e.g. having *"Customers with a higher online activity are more likely to churn "* from Expert 1 and *"Customers with an increased online activity churn at least twice as much as the ones with not coming to the website"* from Expert 2, our module identifies the class variable *"churn / not churn"* and the predictor *"online activity"*).
- *Top features selector*: computes the most relevant features to explain a particular prediction, vs. the entire feature set provided by the Model Agnostic Explainers (LIME, SHAP, etc).
- *Feature Importance space practitioner* computes quantiles (as many as the cardinality of the output linguistic variable) based on the weight of each feature in the feature importance model and assigns each feature to a quantile.
- *Attribute space partitioner*: establish partitions of the attribute space to assign the values to a partition based on the quantile logic (as many quantiles as terms in the target linguistic variable)
- *Knowledge Statement Creator*: elaborates human intelligible statements compounding assertions on one or many linguistic variables referred to predictors and class variables.

Let's have a closer look at the different **Intelligibility increasing scenarios**:

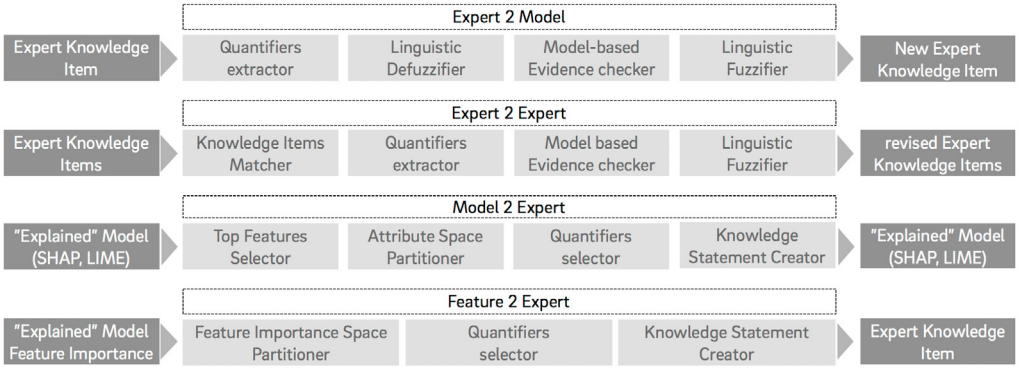

Fig. 2. Modular view of the Intelligibility increasing scenarios identified in this paper

3.1.1. Expert-2-Model

A ML model can prove experts' knowledge items wrong, can give them more or less weight or can reinforce existing believes to a point where take-2-action activities are triggered. On the other hand, experts' knowledge items can pinpoint some limitations in the ML model, can add knowledge not present in the underlying data and can challenge the correctness of the modelling choices. Taking a Knowledge Item KI_i as input, all quantifiers are

extracted (step 1) and deffuzified (step 2) to be then validated against the model, retrieving -1 if not supported at all to 1 if 100% supported (step 3). The *Linguistic Fuzzifier* quantifies to which extend the Knowledge Item is supported by the ML model selecting the proper term in a linguistic variable (step 4). In Fig. 3 we can see the KI_i *'Older customers with less income are less likely to churn'* and steps 1 and 2 represented. The evidence checker will retrieve for example 0.87 and the *Linguistic Fuzzifier* will map it to *highly supported*.

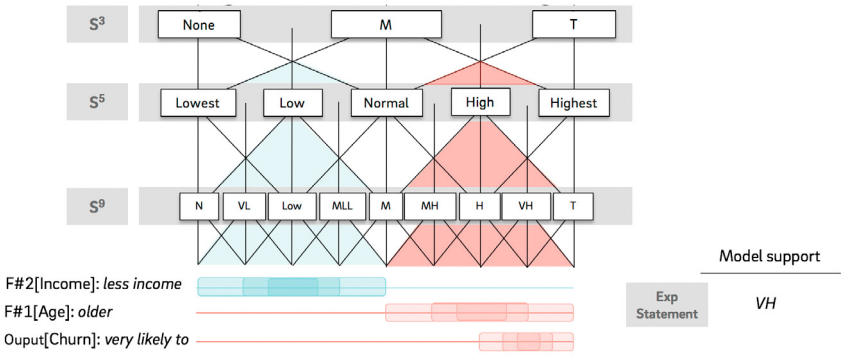

Fig. 3. Example of expert knowledge item representation in a multi-granular linguistic hierarchy

3.1.2. Expert-2-Experts

Having more than one expert available might be very useful to remove bias or generate more knowledge, but also introduces challenges, such as contradicting knowledge items, lack of comparability in supposedly overlapping statements, etc. Our approach checks the support for both statements as explained in the previous section, helping in precision gain but also support the construction of consolidated knowledge items by amending the quantifiers, as shown in Fig. 4, where we can also see how the model provides a higher support to the knowledge item from Expert 1.

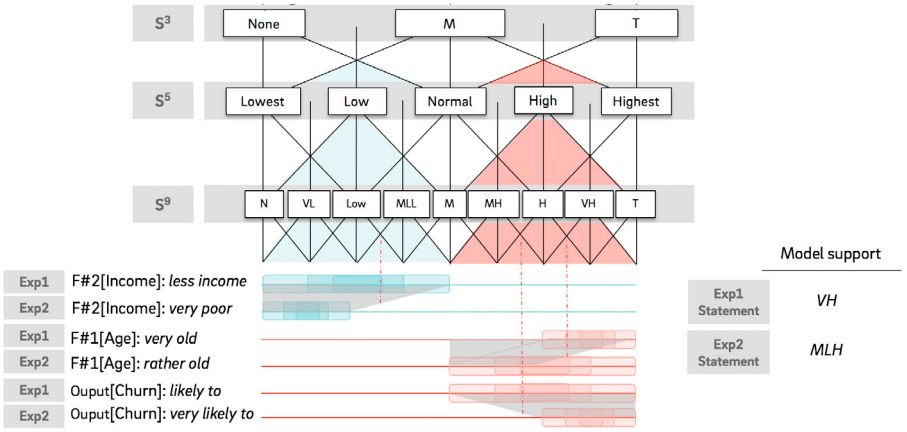

Fig. 4. Fuzzy linguistic matching of overlapping statements from 2 different experts

3.1.3. Model-2-Expert

Our approach here consists of taking the output of explainers (LIME, SHAP) and making it intelligible for our experts. The *Top Features Selector* computes the most relevant features to explain a particular prediction (step 1). The *Attribute Space Partitioner* extracts the quantiles for the attribute values space for each attribute (n quantiles, n is the cardinality of the linguistic variable) and assigns the values used in the explained model rules to one quantile to create intelligible rules (e.g. ąřFeature 1 $\leqslant$ 200ąř $\longrightarrow$ ąřFeature 1 is *high* to *very high*ąř) (step

3). The *Quantifiers selector* maps a particular term c_i in the linguistic variable to each quantile (step 3) and last but not list, the *Knowledge Statement Creator* formulates intelligible statement experts can understand (step 4). Alternatively, the class variable can also be fuzzified to increase the understanding.

In Fig. 5 we can see how 10 features have been selected out of 30 and how Features 1, 2, 8 and 10 have been turned into intelligible rules using for all of them following linguistic variable $S^5 = \{c_0 = VeryLow = VL,\ c_1 = Low = L,\ c_2 = Normal = N,\ c_3 = High = H,\ c_4 = VeryHigh = VH\}$. While *Feature 1* identifies "*very high*" already with 200, *Feature 10* would assign just "*high*" to 200 while giving "*very high*" to over 520. Predicted Risk of Churn can be also expressed in liguistic terms (e.g. "*High risk of churn*").

3.1.4. Feature-2-Expert

Feature importance methods compute to which extend a particular feature contributes to increase the quality of the prediction, typically a value from 0 to 1. The importance of a feature is the increase in the prediction error of the model after we permuted the featureąŕs values, which breaks the relationship between the feature and the true outcome [7]. Fisher et al. [18] developed a model agnostic version of the initial proposal by Leo Breiman in his seminal paper introducing the Random Forest[19].

The *Feature Importance Space Partitioner* computes n quantiles (being n the cardinality of the linguistic variable) based on the weight of each feature in the feature importance model and assigns each feature to a quantile (step 1). The *Quantifiers selector* maps a particular term c_i in the linguistic variable to each quantile (step 2). Finally, the *Knowledge Statement Creator* generates the statements in an intelligible way for expert users (step 3).

Fig. 6 shows the result of applying steps 1 and 2. Step 3 would create statements such as: *"Feature 1, Feature 2 and 3 have a very high impact predicting churn"* or *"Feature 16, 17, 18, etc. have a very low impact on predicting churn"*.

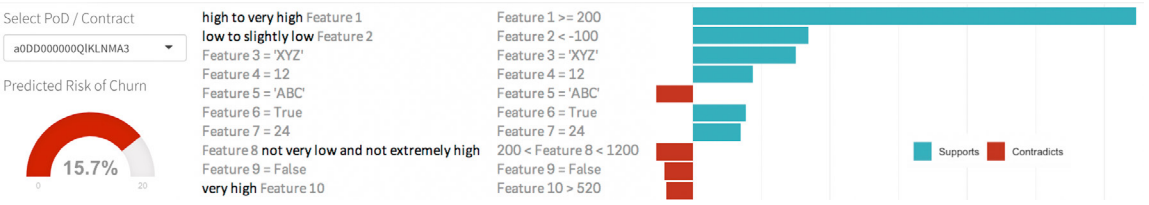

Fig. 5. LIME output modelled using a linguistic variable at attribute level

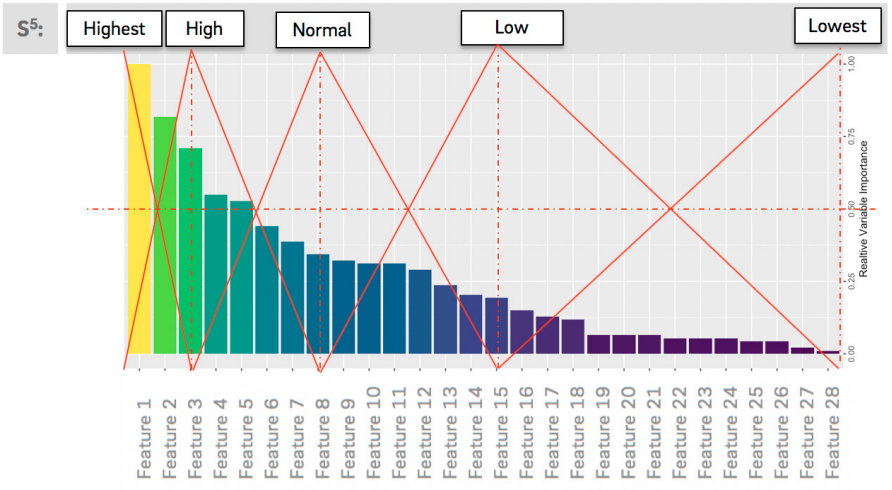

Fig. 6. Linguistic variable mapped to the LIME feature importance output for a particular model

4. Concluding remarks

In this paper we have presented a novel *fuzzy linguistic modelling based* approach to address the intelligibility problem of modern machine learning models. Explainers of all kinds (model specific, model agnostic such as LIME or SHAP) make the predictions of machine learning models transparent for humans (black-2-white box), yet there is still a gap to make these models intelligible enough. The lack of sufficient intelligibility compromises not only the adoption of AI techniques, but also prevents any potential knowledge exchange with existing subject matter experts communities.

In the definition of our method, we first identified 4 situations around machine learning output where the lack of intelligibility manifests: a) existing expert knowledge compatibility with the machine learning model (*Expert-2-Model*), b) consolidation of knowledge from many experts in accordance with the model (*Expert-2-Expert*), c) output of model explainers to humans (*Model-2-Expert*), and d) feature importance to humans (*Feature-2-Expert*).

For each situation, we have specified a pipeline to transform the input into an intelligible output to support the experts-model interaction therefore the operationalization of artificially created intelligence. The pipeline consists of different fuzzy linguistic components (such as quantifiers extractors, linguistic fuzzifier and defuzzifier, etc.) orchestrated to bridge the communication gap between machine and human. To illustrate the output of our method, we have provided concrete examples of all four situations using a real churn prediction model (yet obfuscating the name of the features to prevent revealing competitive information).

The field of AI intelligibility is increasing the research activity and will continue to play a crucial role to enable the AI breakthrough. We'd like to continue this paper focusing on embedding knowledge expertise at modelling time, not just to interpret the model: (*intelligibility by design*). In addition, we'd like to integrate our approach within the model agnostic explainers (e.g. creating a *fuzzy LIME* or a *fuzzySHAP*, etc).

References

[1] D. S. Weld, G. Bansal, The challenge of crafting intelligible intelligence, arXiv preprint arXiv:1803.04263.
[2] M. T. Ribeiro, S. Singh, C. Guestrin, Why should i trust you?: Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, ACM, 2016, pp. 1135–1144.
[3] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, 2017, pp. 4765–4774.
[4] I. Glockner, Quantifier selection for linguistic data summarization, in: 2006 IEEE International Conference on Fuzzy Systems, IEEE, 2006, pp. 720–727.
[5] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, ACM, 2016, pp. 785–794.
[6] F. Herrera, L. Martínez, A 2-tuple fuzzy linguistic representation model for computing with words, IEEE Transactions on Fuzzy Systems 8(6) (2000) 746–752.
[7] C. Molnar, A guide for making black box models explainable, URL: https://christophm.github.io/interpretable-ml-book/).
[8] T. G. Dietterich, Steps toward robust artificial intelligence, AI Magazine 38 (3) (2017) 3–24.
[9] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An overview of interpretability of machine learning, in: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), IEEE, 2018, pp. 80–89.
[10] E. Hüllermeier, Does machine learning need fuzzy logic?, Fuzzy Sets and Systems 281 (2015) 292–299.
[11] I. Couso, C. Borgelt, E. Hullermeier, R. Kruse, Fuzzy sets in data analysis: From statistical foundations to machine learning, IEEE Computational Intelligence Magazine 14 (1) (2019) 31–44.
[12] D. Bonanno, K. Nock, L. Smith, P. Elmore, F. Petry, An approach to explainable deep learning using fuzzy inference, in: Next-Generation Analyst V, Vol. 10207, International Society for Optics and Photonics, 2017, p. 102070D.
[13] L. Zadeh, The concept of a linguistic variable and its applications to approximate reasoning. Part I, Information Sciences 8 (1975) 199-249, Part II, Information Sciences 8 (1975) 301-357, Part III, Information Sciences 9 (1975) 43-80 (1975).
[14] F. Herrera, E. Herrera-Viedma, Aggregation operators for linguistic weighted information, IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems 27 (1997) 646–656.
[15] F. Herrera, E. Herrera-Viedma, L. Martínez, A fusion approach for managing multi-granularity linguistic term sets in decision making, Fuzzy Sets and Systems 114 (2000) 43–58.
[16] F. Herrera, L. Martínez, A model based on linguistic 2-tuples for dealing with multigranularity hierarchical linguistic contexts in multi-expert decision-making, IEEE Transactions on Systems, Man and Cybernetics. Part B: Cybernetics 31(2) (2001) 227–234.
[17] A. A. Márquez, F. A. Márquez, A. Peregrín, A multi-objective evolutionary algorithm with an interpretability improvement mechanism for linguistic fuzzy systems with adaptive defuzzification, in: International Conference on Fuzzy Systems, IEEE, 2010, pp. 1–7.
[18] A. Fisher, C. Rudin, F. Dominici, Model class reliance: Variable importance measures for any machine learning model class, from the âĂIJrashomonâĂİ perspective, arXiv preprint arXiv:1801.01489.
[19] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.