Contextualising local explanations for non-expert users: an XAI pricing interface for insurance

Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, Marcin Detyniecki

Contextualising local explanations for non-expert users: an XAI pricing interface for insurance

Clara Bove[a,b], Jonathan Aigrain[b], Marie-Jeanne Lesot[a], Charles Tijus[c] and Marcin Detyniecki[b,d]

[a]Sorbonne Université, CNRS, LIP6, Paris, France
[b]AXA, Paris France
[c]Laboratoire CHArt-Lutin, University Paris 08, France
[d]Polish Academy of Science, Warsaw, Poland

Abstract

Machine Learning has provided new business opportunities in the insurance industry, but its adoption is for now limited by the difficulty to explain the rationale behind the prediction provided. In this work, we explore how we can enhance local feature importance explanations for non-expert users. We propose design principles to contextualise these explanations with additional information about the Machine Learning system, the domain and external factors that may influence the prediction. These principles are applied to a car insurance smart pricing interface. We present preliminary observations collected during a pilot study using an online A/B test to measure objective understanding, perceived understanding and perceived usefulness of explanations. The preliminary results are encouraging as they hint that providing contextualisation elements can improve the understanding of ML predictions.

Keywords

Interpretability, Explainability, Interface Principle, Interaction Human-ML, Machine Learning, Contextualisation, Explanations

1. Introduction

The rise of Machine Learning (ML) has provided new business opportunities in the insurance industry. ML can for instance help improve pricing strategies, fraud detection, claim management or the overall customer experience. Yet, its adoption is for now limited by the difficulty for ML to explain the rationale behind predictions to end-users [1, 2]. This currently very active topic has led to the development of the so-called eXplainable Artificial Intelligence (XAI) domain. This need for explanation is indeed an important issue, as not being able to understand why a prediction is provided can decrease the trust of a potential customer for the offered product. From a legal point of view, Article 22 of the recent General Data Protection Regulation (GDPR) states that, for any decision made by an algorithm, customers have a "Right to Explanation" to allow them to make informed decisions [3]. For these reasons, explainability is a crucial issue for the insurance industry as well.

Joint Proceedings of the ACM IUI 2021 Workshops, April 13–17, 2021, College Station, USA

clara.bove@axa.com (C. Bove); clara.bove@lip6.fr (C. Bove); jonathan.aigrain@axa.com (J. Aigrain); marie-jeanne.lesot@lip6.fr (M. Lesot); tijus@lutin-userlab.fr (C. Tijus); marcin.detyniecki@axa.com (M. Detyniecki)
https://kmitd.github.io/ilaria/ (J. Aigrain)

The notion of explanation is a complex one that has led to many works and definitions, as discussed in Section 2.1. Among others, it can be defined as "an answer to a why-question" [4], which takes into account contextual and external information [5]. In the more specific context of interaction with a ML-based system, it has been shown that users mostly try to understand the predictions they receive, rather than the model or the training data [6]. In a nutshell, they usually consider the question "Why do I get this prediction?". Explaining such a specific prediction corresponds to the so-called *local* interpretability task in the XAI literature. It is the one addressed by most existing XAI interactive tools. In this paper, we address the challenge of building and presenting local explanations of ML predictions to non-expert users, i.e. users with expertise neither in the considered application nor in ML. Indeed, providing explanations to this target audience brings its own set of challenges, which we propose to address by contextualising the prediction, understood as providing explanations on elements that surround it.

The three contributions of this work are the following: (i) we propose guidelines to add contextual information in interfaces presenting local explanations of ML predictions for non-expert users, considering three levels: information about the Machine Learning system, about the application domain and about relevant external information. (ii) We apply these guidelines to create a smart insurance pricing interface, illustrating it in the case of car insurance. (iii) We conduct a pilot study with non-expert users to assess the effectiveness of our propositions regarding subjective understanding, objective understanding and perceived usefulness of the explanations.

2. State of the Art

This section briefly reviews the complex notion of explanation, first considering the cognitive perspective. It then provides a short overview of different local explanations that can be extracted from Machine Learning models. Finally, it presents how these explanations have been exploited for interfaces in the XAI literature.

2.1. What is an Explanation?

The notion of explanation has been widely studied, the objective of this section is not to provide a complete review of works on this topic, but only to point to some major elements, exploited later on in the paper.

First, offering an explanation requires to identify the underlying, most often implicit, question it should answer. It has been shown that an explanation can be defined as **an answer to a why-question** [4, 7] and that it should provide a reason that justifies what happens [8, 9]. Besides, it is dependent on the context, as it must be adapted to the specific user need [10]. The explanation is also the social process of someone explaining something to someone else [11]. This process is also shown to be multidimensional: in order to be well constructed and well perceived, an explanation answering a why-question needs to be completed with external required information regarding the context, as well as transparency over its pragmatic goals [5, 12].

The specific question of explanation in the context of interaction with a ML system has also been considered, in particular regarding the underlying question a user may be asking when requesting an explanation. Based on a rich user study, a question bank for explainability tools composed of 50 questions has been established [6], beyond the previously mentioned why-question defined from a cognitive point of view. It proposes to dis-

tinguish between questions related to the process understanding, the global understanding of the model, the local understanding of a prediction and the exploration for local contrasting explanations. Moreover, it has been observed in this user study that explanations are most frequently sought to gain further insights or evidence on why a prediction has been made for this instance and not another prediction, bridging the gap to the cognitive results about the notion of explanation.

2.2. What Local Explanations can be Extracted from Machine Learning Models?

As recent ML models have become increasingly accurate and complex, numerous interpretability methods have been developed to provide local explanations [13]. Based on the provided explanation type, one can for instance distinguish between counterfactual examples [14, 3] local rules [15] or local feature importance [16, 17]. The former highlight the minimum changes one needs to apply to a specific example to change its prediction. Local rules provide a combination of simple IF-THEN rules to approximate the decision boundary locally. Local feature importance approaches provide a weight for each feature describing its contribution to the final decision for a specific instance.

2.3. Which Explanations are Presented to Which Users?

In the XAI research community, several interfaces have recently been proposed for explaining to a user a specific prediction of a ML model [18, 19, 20, 21, 2, 22, 23, 24]. Many are aimed at users with an expertise in ML [21, 22, 23, 24] and propose visualization and interactive tools to help data scientists better understand ML models. Others are dedicated to users with advanced knowledge in the application domain, for instance in the medical domain [2].

On the other hand, several works tackle the challenge of presenting local explanations to non-expert users [18, 19, 20, 25] and provide information about the most appropriate type of explanations. In the case of image data, it has been shown that normative explanations lead to better ratings of the underlying ML models than comparative explanations [18] and that example-based explanations have a positive effect on users' trust in ML, regardless of their familiarity with it [25]. In the case of structured data, using counterfactual examples as explanations has been explored in the ViCE tool [20] but this proposition has not been evaluated experimentally. Local feature importance has also been considered [19] and it has been shown that combining these explanations with the possibility to interact with the ML model to explore its behaviour improves both subjective and objective understanding of non-expert users.

3. Interface Principles

This section describes the general principles of our propositions: after stating the purpose of the interface as compared to existing ones, it describes the guidelines we propose and gives an overview of the interface before presenting the three types of added contextual information that can be considered as missing in current systems: general information on ML, domain information and external information. It also describes the interface principles we propose to include each type of contextualised information.

3.1. Purpose of the Interface

Based on the state of the art study presented in Section 2, we consider the definition of an

explanation as **an answer to a why-question**. We focus on users having expertise neither about machine learning, nor about the application domain. In the context of car insurance pricing, our goal is to help a non-expert user answer the following question: "Why did I get this price?". We consider the user filled a form asking him/her for some personal information and is faced with a price proposition computed based on this information. We aim at providing the necessary contextual information to allow the user to make an informed decision about this price.

We believe that **local feature importance** is the most relevant type of available local explanations for such a why-question about a prediction made on tabular data. Indeed, we argue that counterfactual examples are more relevant to answer "Why-not" questions, i.e. to explain why another price has not been obtained and to provide indication about how to change the predicted price. As for local rules, they have mostly been applied to classification tasks, whereas the pricing scenario we consider constitutes a regression task. In addition, it has been shown that local feature importance is helpful to non-expert users [19], who are the target users we consider.

3.2. Interface Overview

A global view of the interface principles we propose is illustrated in Figure 1, it is commented in details in this section and the following ones. Figure 2 illustrates its implementation in the case of smart car insurance, as discussed in Section 4.

First, the proposed interface applies a card-based design: it contains an individual card for each of the fields the user is required to fill in when requesting a price prediction. Indeed, the rationale of this design choice is that it allows the user to get an overview of all information he/she entered. More importantly, the second motivation for this design choice is that it is straightforward when considering the feature importance approach, that considers features individually.

Each card is made of four parts containing different pieces of information related to the feature. The top part shows the name of the associated field, as present in the user-filled form, as a reminder of the latter. An icon in the middle of the card provides a more user-friendly visual representation of the feature. In the illustration of the general principles given in Fig. 1, these pictures are geometric shapes, see Fig. 2 for some examples for real features. Below the name, the card shows the effect of the feature, i.e. its individual contribution to the prediction, as derived from the feature importance method. Moreover, an intuitive color code helps the user get an immediate understanding of the feature effect, displaying in green the effects that help reduce the predicted price and in red the ones that increase it. Finally the bottom part of the card provides contextual information at the level of the application domain, as discussed in Section 3.4.

The order in which the cards are displayed follows a double principle: first they are grouped to define categories and are then ordered within these categories, as discussed in Section 3.3. This makes it possible to provide contextual information at the model level, while taking into account the user low expertise.

3.3. Contextualising with ML Information

Machine Learning tools are used at several levels of the user interaction with the system: to predict the proposed price and to explain the role of the attributes. Non-expert users do not know how the model has been trained and what basis it uses to make a prediction. They may also confuse the displayed local explanations with global ones, and thus erroneously think that the importance attached

to a field value is the same across the whole value domain. Consequently, it is important to give transparency on the model's purpose and basic operations. To that aim, the guidelines we propose for an XAI interface include the notion of **ML transparency** providing guidance regarding how to interpret the following explanations.

First, we propose to show, in the top part of the interface (region A in Fig. 1), contextual preliminary information that can help users build a mental model of how the ML system works and better interpret explanations. This additional explanation about the model should be visible at first, so as to act as an on-boarding guide to read adequately the local feature importance explanations provided below.

Second, as mentioned in Section 3.2, the card ordering follows a double principle. A natural choice would be to sort the cards in decreasing order of the absolute feature importance values, providing an obvious representation of the ML model behaviour. Note that the absolute value must be considered so that the attributes with major negative influence are not postponed to the end of the list, but shown at the top, together with the attributes with major positive influence. We propose to achieve a compromise between this approach and a categorical sorting more in concordance with a non-expert user. Indeed, in a classic user journey when interacting with the system, there is no transition between the input stage when he/she fills the application form and the output stage when the prediction is displayed. As a consequence, a non-expert user might have trouble finding a logical path between the information he/she gave and the provided explanation, if the order is completely different.

In order to facilitate this transition, we propose to contextualise the local feature importance sorting to match the input stage and exploit the field categories users encounter in the form. The local feature scores are aggregated at a category level, summing the values associated to all features in each category. Categories are then displayed showing the most influential ones first (in absolute values). Then, within each category, features are sorted by decreasing importance. This principle is illustrated in Figure 1 by the region denoted with letter B.

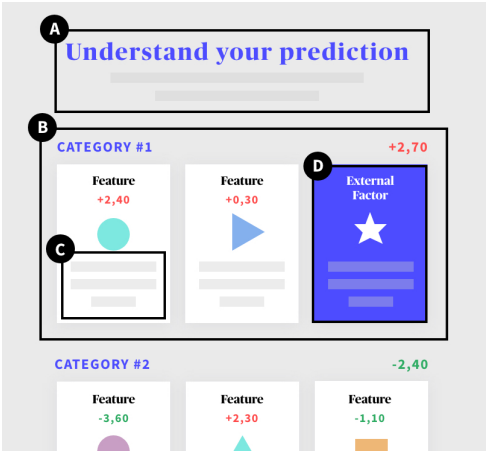

Figure 1: Contextualized Local Feature Importance explanations in XAI interface for non-expert users. (A) Contextual information on the ML system, (B) categorical sorting, (C) contextual information on the domain, (D) contextual external information.

3.4. Contextualising with Domain Information

As discussed in Section 2.1, from a cognitive point of view, an explanation should provide a rationale about why a prediction is made, which means it should be related with the notion of cause. Now the automatic extraction of causality relations by ML models is a very challenging task [26], it is not achieved by the local feature importance approach chosen for the proposed interface. Thus, a non-expert user might have difficulties understanding why

his/her specific input influences the output.

To compensate for this lack of rationale, we propose to associate local feature importance explanations with information provided by a domain expert, e.g. an actuary for an insurance pricing platform. This added information acts as a generic transparency over the domain, called **domain transparency**, providing some brief justification about how this feature might impact the outcome. In other words, instead of trying to extract automatically causality relations, which remains a very difficult task, we require an expert to provide this piece of information. This domain information is generic, i.e. applicable to all instances, it is displayed on each feature card (see region C in Fig. 1).

3.5. Contextualising with External Information

Whereas the ML interpretability approaches aim at providing explanations about a given prediction model, the conception of the model itself, prior to its training phase, also affects the outcome the user gets. For instance, some fields the user is requested to fill can be excluded from the ML model by design. The form may include a field about gender so that the system knows how it should address the user, although it is not taken into account by the prediction system so as to avoid any gender bias. We call external information this type of knowledge, which is not domain specific and differs from the information considered in the previous section. It is important to provide the users with this global information which may impact the outcome they get, even though this information is external to the model. We believe users can benefit from added transparency over the real-life context (e.g. external events such as the COVID crisis that indirectly influences the prediction through the dataset) and algorithmic processes (e.g. data that are collected and used or not). This **external transparency** can both improve understanding and the level of trust in the prediction and its explanations.

For the case of attributes requested in the form but excluded from the prediction model, we propose to highlight their specific role by a different type of feature-associated card, as illustrated by the card denoted D in Fig. 1.

4. Considered Application

This section describes the implementation of the principles described in the previous section to the case of a smart pricing insurance application.

4.1. Usage Scenario

We apply the proposed interface principles for contextualising local feature importance explanations for a fictive car insurance pricing platform. In this scenario, a user first provides several kinds of information, regarding his/her insurance background, the car to insure, its usage and parking as well as housing and personal information. The interface (see Fig. 2) then displays the price computed by the ML model on the left, and the proposed explanation interface on the right. This influence is computed using SHAP [16], a local feature importance method that provides the contribution of each feature value to the prediction as compared to the average prediction.

4.2. Implementing ML Transparency

As discussed in Section 3.3, we integrate the proposed model transparency principle an onboarding text. This text first states that the impact of each feature is expressed relatively to the average price predicted by the model and it makes explicit the difference between

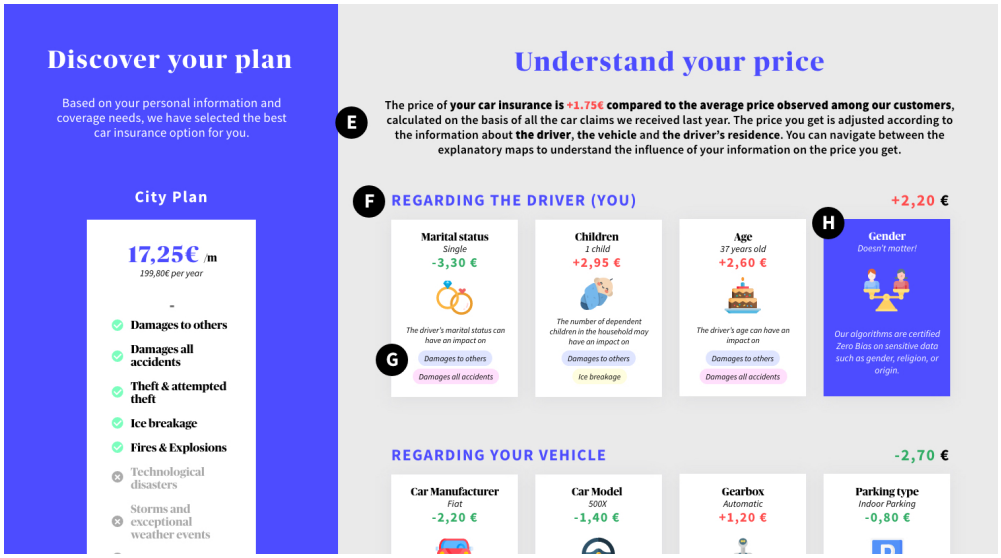

Figure 2: Implementation of the general principles on the left for a smart car insurance pricing application. (E) Onboarding on ML system for pricing and explanation calculation, (F) Driver's information categorical sorting, (G) Actuarial global information about risks associated with each feature, (H) External information about an algorithmic process excluding sensitive data from pricing strategy

the predicted and the average prices. Second, it explains that the price has been personalized based on the user's information. Finally, it introduces the feature-associated cards.

Regarding categorical sorting, we split features in three categories, distinguishing between features related to the driver, those related to the car and those related to residence.

4.3. Implementing Domain Transparency

As discussed in Section 3.4, the role of domain transparency is to provide users with a rationale about why the collected features are useful for the ML model. For the car insurance pricing platform, we implement this principle by complementing each feature associated card with the main kinds of risk that can be impacted by the feature (e.g. accident, theft or natural catastrophes to name a few). This pairs the local feature weight explanation with global domain information. In addition, each type of risk is highlighted with a different color to improve visualization (see region G in Fig. 2).

4.4. Implementing External Transparency

We implement external transparency by providing information on the gender feature that is not included in the ML model, but which is likely to be considered important by users. Indeed, it may be the case that users are suspicious about how their gender can be used to affect their prediction. Therefore, we display that the gender information is not used by the model in a feature-associated card. This card is presented in a different color from other

feature-associated cards to highlight the difference of purpose.

It is noteworthy that only removing the gender feature from the training data does not necessarily make a ML model fair [27]. Explaining to non-expert users more sophisticated fairness protocols is an important topic for the XAI community, but it is outside the scope of this paper.

5. Evaluation

This section presents the experimental evaluation framework we propose to assess the propositions described in the previous sections, describing in turn the considered evaluation metrics, the experimental design and the results of the conducted pilot study.

5.1. Evaluation Metrics

Evaluating the effectiveness of explanations is a challenging task [28], various methods and quality criteria to measure understandability and usefulness have been proposed. Two categories can be distinguished: objective understanding can for instance be evaluated through task completion [25] or quiz questions [19], measuring the answer correctness as well as answering duration. Subjective criteria, on the other hand, measure the perceived usefulness and understanding of explanations through self-reports [10]. We take into account these two types of criteria, as detailed below.

Regarding objective understanding, using a similar quiz approach as Cheng *et al.* [19], we first propose four types of questions to check the user objective understanding. The details of the questionnaire are provided in appendix A. (i) Feature Importance Questions measure the extent to which the user understands the relative influence of the attributes on the prediction, e.g. *"Does feature X impact more the prediction than feature Y ?"*. (ii) ML Information Questions measure the user's effective understanding of how the considered SHAP method generates the influence of each attribute over an average price, e.g. *"Are the explanations provided based on the average prediction?"*. (iii) Local Explanation Questions measure the user's understanding of the difference between the influence of his/her attributes and global explanations, e.g. *"Will the prediction remain for sure the same even if feature X is different?"*. (iv) Interpretation Questions measure the extent to which the user processes the explanations provided to understand the price rather than relying on potential cognitive biases, e.g. *"Does this information/event influence the prediction?"*.

We design two quiz questions for each of the four types. For statement questions, three answer options are provided: "true", "false" and "I don't know"; for one-choice questions, lists of possible answers are offered as well as an "I don't know" option. We measure the answer correctness and time to answer each question.

Regarding subjective understanding, we adapt two self-reporting questions from the Explanation Satisfaction Scale [10], to assess the perceived understanding and usefulness of explanations to make an informed decision. Participants are required to answer on a 6-point Likert scale, from "Strongly disagree" (0) to "Strongly agree" (5), as it has been shown that 6-point response scales are a reasonable format for psychological studies [29].

In addition, the questionnaire includes two questions regarding the participant literacy in artificial intelligence/machine learning and insurance, again using 6-point Likert scales, from "Not familiar at all" to "Strongly familiar". We also ask for basic demographic information such as age and education level. Finally, participants can share their insights and comments on the study in an open response question.

	Objective understanding	Feature Importance Questions	Interpretation Questions	Local Explanation Questions	ML Information Questions
Interface A	**0.73 (±0.20)**	0.71 (±0.24)	0.71 (±0.24)	0.71 (±0.39)	0.14 (±0.35)
Interface B	**0.88 (±0.16)**	0.63 (±0.26)	0.63 (±0.26)	1.00 (±0.00)	0.83 (±0.41)
	Self-reported understanding	**Self-reported usefulness**			
Interface A	**0.71 (±0.28)**	**0.63 (±0.37)**			
Interface B	**0.87 (±0.16)**	**0.91 (±0.10)**			

Table 1
Obtained results for the two interfaces, interface A without contextualisation and interface B with contextualisation: for objective questions, average and standard deviation of the percentage of correct answers (overall and for each question type), for self-reported questions, average and standard deviation of the scores on the Likert scale

5.2. Experimental Design

We conduct an A/B testing. Interface A, displayed in Figure 3, presents local feature importance explanations as extracted from SHAP, following the same card-based design described in Section 3.2 but it does not include the different elements of contextualisation. Interface B includes all of our propositions, it is displayed partly in Figure 2 and fully in Figure 4. Participants are randomly assigned to one version of the interface.

For the pilot experiment, we simulate a pricing ML model and the use of SHAP to extract local feature importance. Each participant acts as a female persona with a given set of 16 feature values related to the driver, the vehicle and the residence of the driver. Prior to the evaluation, participants are introduced to the persona and her need to understand the price she gets. We also explain the platform uses an algorithm to determine a personalized price based on her personal information. The evaluation starts with the objective understanding question quiz, which is displayed next to the interface to allow participants to look for the answers. Then, the subjective understanding questions and demographics information questions are asked.

Because of the COVID-19 situation, we were unable to conduct the pilot study in lab. Thus, we conducted the pilot on Useberry. 20 participants were recruited from university and professional social network.

5.3. Results

The obtained results are displayed in Table 1. For objective questions, the results are defined as the percentage of correct answers; for the subjective questions, the results are the average scores on the Likert scale, normalized to the [0,1] interval.

The data of 6 participants were not exploitable as they dropped off from the survey at the start. We also excluded the data of one more participant, who completed the test in an abnormally short time and who appeared not to scroll through the explanations to look for the answers. Out of the 13 remaining participants, 7 were assigned to interface A and 6 to interface B. Participants assigned to interface A (resp. interface B) are 29.6 years old on average (resp. 29.8) and reported an average artificial intelligence literacy score of 0.71 (resp. 0.37) and an average insurance literacy score of 0.47 (resp. 0.60).

Participants assigned to interface B obtain overall higher scores for the objective understanding questions (0.88) as compared to the

ones using interface A (0.73). When considering the different types of questions, it appears that interfaces A and B lead to comparable results for feature importance and local explanation questions. This could mean that providing local feature importance is enough for a non-expert user to correctly answer these questions, even without any contextualisation. The results hint that there is an improvement for participants assigned to interface B for interpretation and ML information questions, which hints that contextual information, especially external and ML transparency, may help non-expert users to answer this kind of questions. It is however noteworthy that the difference in ML Information question scores can be partly due to the fact that the answer was easier to retrieve in interface B thanks to the added ML transparency.

In this preliminary study, participants who used interface B report higher subjective understanding (0.87) compared to version A (0.71) and also rate higher the usefulness of the explanations (0.91 for interface B and 0.63 for interface A).

Overall, we observe that the contextualisation elements of interface B provide an improvement for all considered evaluation metrics: +0.14 for objective understanding, +0.15 for self-reported understanding and +0.29 for self-reported usefulness.

5.4. Discussion

These preliminary results indicate that interface B seems to improve the explanation understanding thanks to the three levels of added contextual information. More specifically, this improvement is especially important for the perceived understanding and usefulness of explanations. Unfortunately we cannot analyze whether these added information leads to participants spending more time on the interface, since the reported time data are too noisy, probably due to participants taking breaks during the test. We hope to mitigate this issue by performing this experiment in a lab setting.

Looking at the feedback collected through the open-question, it appears that participants using interface A, without contextual information, report more uncertainty regarding their answers and their understanding, as two of them explicitly state. On the other hand, 1 participant using interface B reports that the explanations are "pleasantly surprising and help choosing among different insurance plans", while another participant states that the explanations are clear.

To conclude, although the sample size is too small to provide strong and reliable insights backed up with statistical tests, the experimental results and the qualitative feedback lead us to believe that contextualisation can be an interesting solution to explore in order to enhance local explanations.

6. Conclusion

In this work, we study whether contextualisation can help non-expert users understand local explanations. We investigate three kinds of information for contextualisation, respectively regarding ML, the application domain and external factors. In the context of a smart pricing platform for car insurance, we conducted a pilot study using an A/B experiment online to measure objective understanding, perceived understanding and perceived usefulness of explanations. The preliminary results are encouraging as they hint that providing contextualisation elements can improve the understanding of ML predictions.

Future work will include a larger user study in a more controlled environment, to draw stronger conclusions regarding the effectiveness of our propositions. We also plan to run experiments using a ML model to extract actual local explanations instead of simulated ones, as it may influence our findings.

References

[1] C. Cai, M. Stumpe, M. Terry, E. Reif, N. Hegde, J. Hipp, B. Kim, D. Smilkov, M. Wattenberg, F. Viegas, G. Corrado, Human-centered tools for coping with imperfect algorithms during medical decision-making, in: Proc. of the Int. Conf. on Human Factors in Computing Systems, CHI'19, 2019.

[2] D. Wang, Q. Yang, A. Abdul, B. Y. Lim, Designing Theory-Driven User-Centric Explainable AI, in: Proc. of the Int. Conf. on Human Factors in Computing Systems, CHI'19, 2019.

[3] S. Wachter, B. Mittelstadt, L. Floridi, Why a right to explanation of automated decision-making does not exist in the general data protection regulation, International Data Privacy Law 7 (2017) 76–99.

[4] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38.

[5] B. F. Malle, How the mind explains behavior: Folk explanations, meaning, and social interaction, MIT Press, 2006.

[6] Q. V. Liao, D. Gruen, S. Miller, Questioning the ai: Informing design practices for explainable AI user experiences, in: Proc. of the Int. Conference on Human Factors in Computing Systems, CHI'20, 2020, pp. 1–15.

[7] J. Pearl, D. Mackenzie, The Book of Why: The New Science of Cause and Effect, Basic Books, Inc., 2018.

[8] D. C. Dennett, The Intentional Stance, MIT press, 1989.

[9] D. C. Dennett, From bacteria to Bach and back: The evolution of minds, WW Norton & Company, 2017.

[10] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for explainable AI: Challenges and prospects, arXiv preprint arXiv:1812.04608 (2018).

[11] D. J. Hilton, Conversational processes and causal explanation., Psychological Bulletin 107 (1990) 65.

[12] B. F. Malle, Attribution theories: How people make sense of behavior, Theories in social psychology 23 (2011) 72–95.

[13] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (xai), IEEE Access 6 (2018) 52138–52160.

[14] T. Laugel, M. J. Lesot, C. Marsala, X. Renard, M. Detyniecki, The dangers of post-hoc interpretability: Unjustified counterfactual explanations, in: Proc. of the Int. Joint Conf. on Artificial Intelligence, IJCAI'19, 2019, pp. 2801–2807.

[15] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: Proc. of the 32nd AAAI Conference on Artificial Intelligence, AAAI'18, 2018, pp. 1527–1535.

[16] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proc. of the Int. Conf of Advances in Neural Information Processing Systems, NeurIPS'17, 2017, pp. 4765–4774.

[17] M. T. Ribeiro, S. Singh, C. Guestrin, "why should I trust you?" explaining the predictions of any classifier, in: Proc. of the ACM Int. Conf. on Knowledge Discovery and Data Mining, SIGKDD'16, 2016, pp. 1135–1144.

[18] C. J. Cai, J. Jongejan, J. Holbrook, The effects of example-based explanations in a machine learning interface, in: Proc. of the 24th Int. Conf. on Intelligent User Interfaces, IUI'19, 2019, pp. 258–262.

[19] H. F. Cheng, R. Wang, Z. Zhang, F. O'Connell, T. Gray, F. M. Harper, H. Zhu, Explaining decision-making algorithms through UI: Strategies to help

non-expert stakeholders, in: Proc. of the Int. Conf. on Human Factors in Computing Systems, CHI'19, 2019.

[20] O. Gomez, S. Holter, J. Yuan, E. Bertini, Vice: visual counterfactual explanations for machine learning models, in: Proc. of the 25th Int. Conf. on Intelligent User Interfaces, IUI'20, 2020, pp. 531–535.

[21] Y. Ming, H. Qu, E. Bertini, Rulematrix: Visualizing and understanding classifiers with rules, IEEE Transactions on visualization and computer graphics 25 (2018) 342–352.

[22] D. K. I. Weidele, J. D. Weisz, E. Oduor, M. Muller, J. Andres, A. Gray, D. Wang, AutoAIviz: Opening the blackbox of automated artificial intelligence with conditional parallel coordinates, in: Proc. of the Int. Conf. on Intelligent User Interfaces, IUI'20, 2020, pp. 308–312.

[23] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viegas, J. Wilson, The what-if tool: Interactive probing of machine learning models, IEEE Transactions on Visualization and Computer Graphics 26 (2020) 56–65.

[24] X. Zhao, Y. Wu, D. L. Lee, W. Cui, iForest: Interpreting random forests via visual analytics, IEEE Transactions on visualization and computer graphics 25 (2018) 407–416.

[25] F. Yang, Z. Huang, J. Scholtz, D. L. Arendt, How do visual explanations foster end users' appropriate trust in machine learning?, in: Proc. of the Int. Conf. on Intelligent User Interfaces, IUI'20, 2020, pp. 189–201.

[26] R. Guo, L. Cheng, J. Li, P. R. Hahn, H. Liu, A survey of learning causality with data: Problems and methods, ACM Computing Surveys 53 (2020) 1–36.

[27] B. Ruf, C. Boutharouite, M. Detyniecki, Getting fairness right: Towards a toolbox for practitioners, in: Proc. of the Int. Conf. on Human Factors in Computing Systems, CHI'20 - Workshop on Fair & Responsible AI, 2020.

[28] Z. C. Lipton, The mythos of model interpretability, in: Proc. of the Int. Conf. on Machine Learning, ICML'16 - Workshop on Human Interpretability in Machine Learning, 2016.

[29] L. J. Simms, K. Zelazny, T. F. Williams, L. Bernstein, Does the number of response options matter? Psychometric perspectives using personality questionnaire data., Psychological assessment 31 (2019) 557.

A. Detailed Questionnaire for Objective Understanding Questions

This section gives the 8 questions asked to assess the participants' objective understanding.

Question 1: The model of your vehicle influences more your price than the number of children you have at charge.

True
False
I don't know

Question 2: What is the influence of the gearbox of your car on your price?

It increases my price
It doesn't change my price
It decreases my price
I don't know

Question 3: Even if you were older, you would get the same price for sure.

True
False
I don't know

Question 4: If you were living in another city, you would probably get a different price.

True
False
I don't know

Question 5: Which one of your information doesn't influence your price?

My age
My vehicle's power supply
My job occupation
My residence area
I don't know

Question 6: Again, which one of your information doesn't influence your price?

The model of my vehicle
The number of children at my charge
My gender
My job occupation
I don't know

Question 7: Your price is calculated based on an average price of 15.5€

True
False
I don't know

Question 8: Your information increases your price by 1.15€.

True
False
I don't know

B. A/B Testing: Interfaces A and B

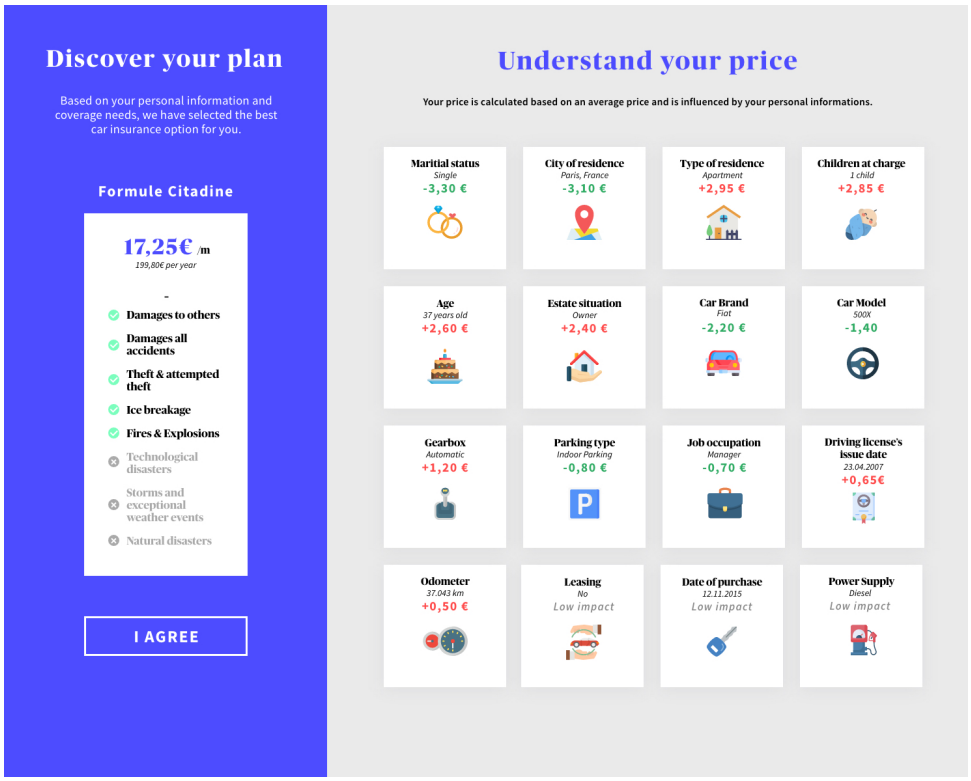

Figure 3: Interface A without contextualization principles

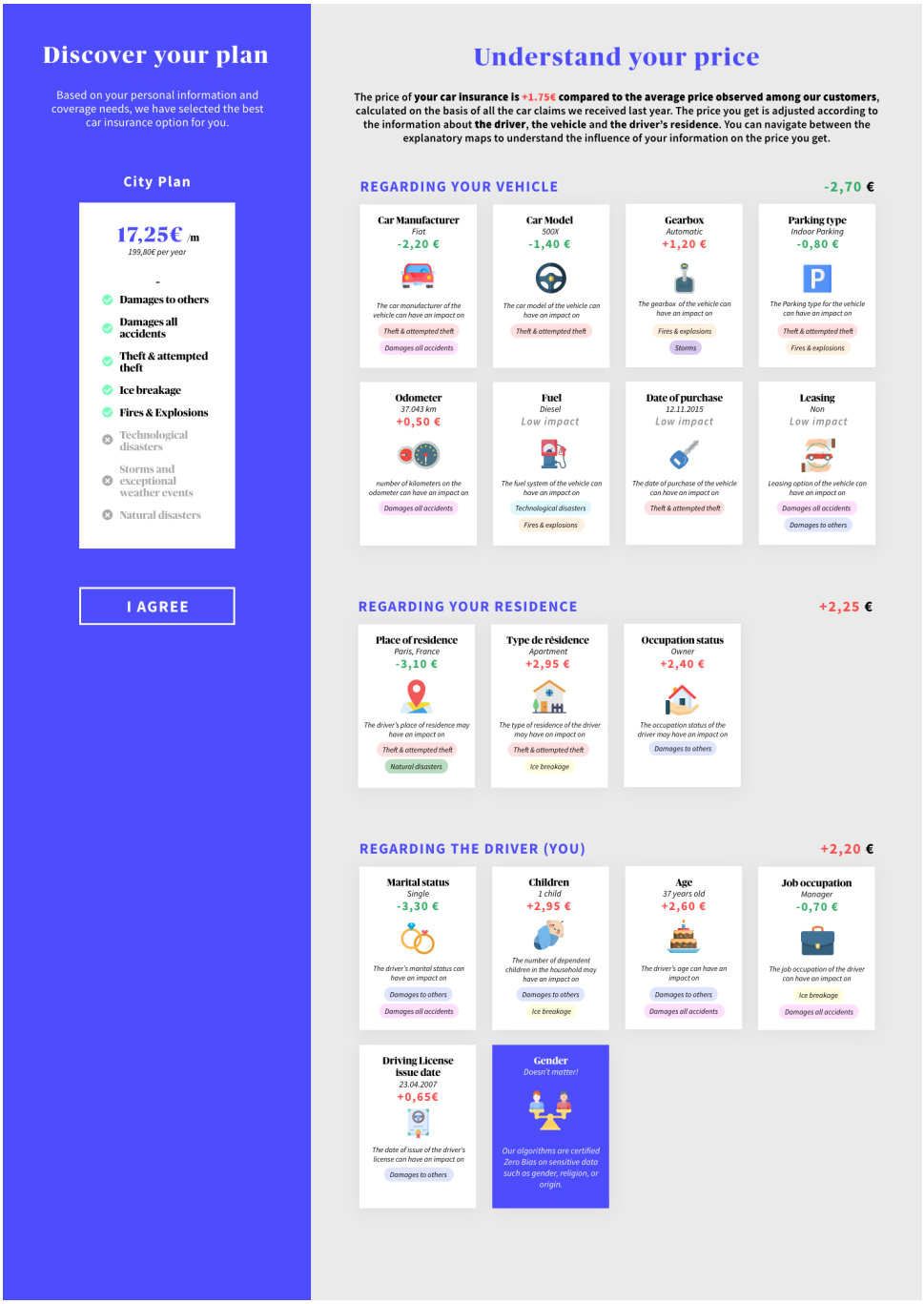

Figure 4: Interface B with contextualization principles