TRUSTAI

TRANSPARENT, RELIABLE
& UNBIASED SMART TOOL

D2.1: User studies on the realisation of explanations

Lead participant: University of Tartu

December 31, 2021

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant agreement No 952060.

DOCUMENT CONTROL PAGE

DOCUMENT	**D2.1 – User studies on the realisation of explanations**
TYPE	**Report**
DISTRIBUTION LEVEL	Public
DUE DELIVERY DATE	31/12/2021
DATE OF DELIVERY	08/01/2022
VERSION	0.6
DELIVERABLE RESPONSIBLE	UT
AUTHOR (S)	Marharyta Domnich (UT), Raul Vicente Zafra (UT), Eduard Barbu (UT)
OFFICIAL REVIEWER/s	Gonçalo Reis Figueira (INESC TEC), Fábio Neves-Moreira (INESC TEC), Zehra Cataltepe (TAZI), Ozden Gur-Ali (TAZI)

DOCUMENT HISTORY

VERSION	AUTHORS	DATE	CONTENT AND CHANGES
0.1	University of Tartu	24/12/2021	First draft
0.2	INESC TEC	07/01/2022	Final revision
0.3	University of Tartu	23/03/2022	Added Annex 3 containing the interview with Sónia Germano. Updated the section 4.3 and 4.4.
0.4	University of Tartu	14/04/2022	Added one response to decision-maker questionnaire in Annex 2; added subsection 4.2.4 and 4.2.5 with description and discussion of decision-maker questionnaire. Integrated some suggestions by LTP.
0.5	University of Tartu	25/11/2023	Operated the modifications asked by the reviewers in the project review meeting on 21.11.2021 We have added a new section called "**Design guidelines**

			emerging from the user studies” and slightly altered the text in the introduction to accommodate the new section.
0.6	INESC TEC	29/11/2022	Final revision (mainly adjustments in the introduction section).

ACKNOWLEDGEMENTS

NAME	PARTNER
Nikos Sakkas	AIT
Francisco Amorim	LTP
Fábio Neves Moreira	INESC TEC
Peter Bosman	CWI
Evi Sijben	CWI

DISCLAIMER:

The sole responsibility for the content lies with the authors. It does not necessarily reflect the opinion of the CNECT or the European Commission (EC). CNECT or the EC are not responsible for any use that may be made of the information contained therein.

Executive Summary

The present deliverable *D2.1, "User studies on the realisation of explanations,"* has the purpose of determining the best way to present explanation content for each of the three use cases (healthcare, retail, and energy). In collaboration with each use case partner, we developed specific questionnaires to test the preferred modality (graphical, textual, table, charts) and type of explanation (causal, kind of contrast, feature importance). Interviews with experts from each user case were also conducted to complement the information obtained from the questionnaires.

The questionnaires have been distributed through Google Forms. The interviews have been conducted over Zoom.

The received answers to the questionnaires and interviews have been instrumental in determining the specific end users, their desired accuracy-explainability trade-off, their familiarity with different types of explanation, and their preferred complexity and format of explanation. Together with deliverables D5.1, D6.1, and D7.1, this study helps to constrain and provide specific recommendations to the user interfaces (Task 2.2) and continue guiding WP3 and WP4 (on cognitive models symbolic learning systems).

This deliverable has the following structure. First, the introduction section presents the nature of the explanations we are looking to explore. Then each use case has a dedicated chapter. In each chapter, we:

1. Describe the types of questions asked
2. Introduce the key persons that took part in the interviews
3. Discuss the results of the questionnaires and interviews
4. Summarize the key points taken from questionnaires and the interviews

The questionnaires and the transcribed interviews are in the Annexes of this deliverable. They are linked from each chapter for completion and intelligibility.

The deliverable ends with the conclusions abstracting the insights applicable to all use cases.

Table of Contents

Abbreviations and Acronyms

AI	Artificial Intelligence
EC	European Commission
EU	European Union
HCXAI	Human-centered Explainable AI
KPI	Key Performance Indicators
MS	Milestones
PM	Person Month
PR	Press Release
SMEs	Small and Medium-sized Enterprises
WP	Work Package
XAI	Explainable Artificial Intelligence

1. Execution

The delivery requires close cooperation with all partners and their end users. Due to communication constraints and availability of customers, the delivery was delayed until January 2022.

Use-Case 1: Healthcare:

- A questionnaire for physicians was prepared and filled by 3 medical doctors.
- One interview was conducted with a medical doctor.

Use-Case 2: Online Retail:

- A questionnaire for customers was prepared and filled by 7 respondents.
- A questionnaire for decision-makers was prepared, but left unanswered due to unavailability of respondents.
- No interviews were conducted due to unavailability of decision-makers.

Use-Case 3: Energy:

- A questionnaire for operational managers was prepared and filled by 5 experts in the energy field.
- Four interviews were conducted with energy experts.

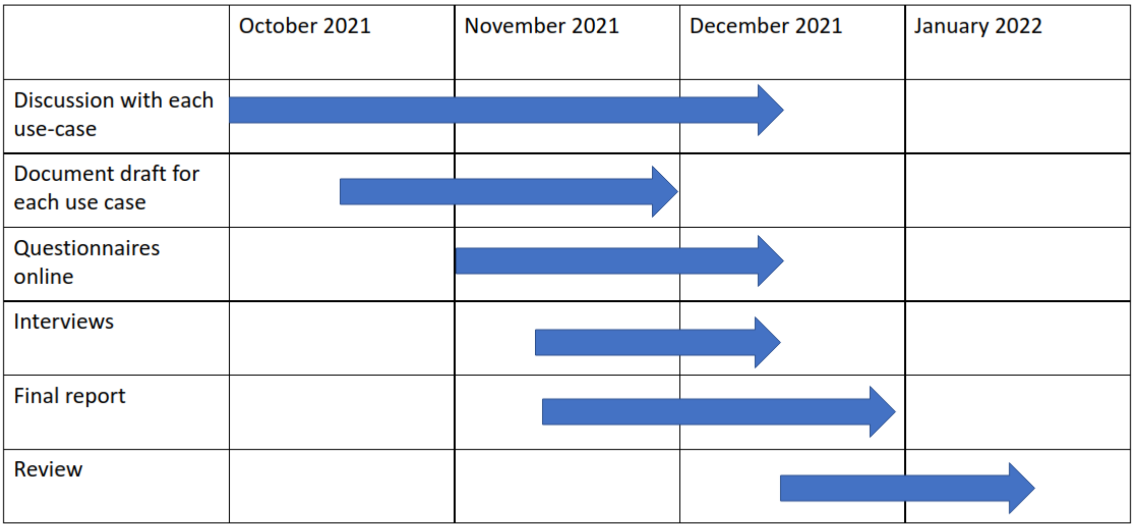


Figure 1. Chronogram for D2.1 "User studies on the realization of explanations".

2. Introduction

Explanations are needed for informing and supporting human decision making. Creating a shared understanding between an algorithm and a human by increasing transparency. Moderating trust and verifying generalization ability. Behavioral studies have extensively investigated how humans produce and perceive explanations for diverse mechanistic and social phenomena [1]. Explanatory understanding includes not only the processes of creating and discovering explanations but also the processes of providing and receiving them [2].

Explainability can be achieved on different levels (local, global). Explanations may answer different questions (What, Why, Why not, How to, What if) and be presented in different forms (textual, visual, symbolic expression, feature relevance etc.). Forms might have further division, like textual explanations can be divided into causal, contrastive, counterfactual, transactional, and visual explanations can be a graph, dashboard, pie chart and many more [3]. Altogether there are many combinations for building an explanation and this delivery aims to find the most efficient explanation approach for each use-case individually by performing questionnaires and interviews.

At this stage in the project the consortium has identified three categories of users:

1. **Domain Experts.** The domain experts have a thorough knowledge of the domain to be modeled. Examples include medical doctors, operations managers, and policy makers.
2. **AI Experts.** The AI experts have a profound knowledge of inner workings of the machine learning algorithms. In particular, they are able to knowledgably adjust the parameters of the algorithms or even extend their logic.
3. **Model Developers.** The model developers bridge the gap between the domain experts and the AI experts. They have an instrumental understanding of AI, while also understanding the problem at a reasonable level. The are responsible for translating the needs of the problem (raised by the domain experts) into the inputs of the AI system, as well as communicating the output (the model) of the latter to the domain experts. Adjustments to the model can also be done, since in the case of symbolic learning, the models are more easily manipulated.

The TRUST-AI project aims to bridge the gap between humans and machines, and promote a collaborative learning process, where humans are involved in the loop. The target users are therefore the model developers, who bridge the gap and bring human knowledge to the learning loop. Since they bridge this gap, understanding each of those users is important. However, as we know quite well the AI experts and the model developers (the consortium is composed of these users), the user studies described in this deliverable focus on the domain experts. The studies aim to inform the design of the TRUST-AI tool by surveying some domain experts through questionnaires and interviews.

The questionnaire was produced collectively together with each partner to narrow down the huge number of combinations for selecting the preferred format of explanations (tables, charts, interactive graphics, text) and type of explanations (causal, contrastive, counterfactual, prototype) depending on the aim of explanation (transparency, trust, accuracy). The questionnaire was built to assess the need of explanations and evaluate combinations of different forms of explanations in each use case. The important evaluations of explanations are circularity, relevance and coherence [2]. Recent works show that people prefer explanations that are coherent (they are consistent with prior knowledge) and simpler [4]. However, alongside simplicity, explanation should appeal to multiple causal mechanisms in order to be convincing, whereas the complexity of desired explanation may differ depending on the personal background. Additionally, the effectiveness can be estimated (whether explanation helps to make a better decision) together with explanation trust and bias. The choices made in questionnaires are investigated further with the help of interviews which allows us to talk with end users and assess the overall need for explanations.

3. Trust instantiation in Healthcare

This chapter of the deliverable is concerned with surveying questionnaires and interviewing healthcare specialists to understand the explanation and user interface requirements better. The key persons and their role is stressed in the following list and Table 1 below:

1. **Dr. Jeroen Jansen** is a professor of ENT and Head and Neck surgery, particularly Head and Neck Oncology and Skull base Surgery. Jeroen Jansen is a consultant Head and Neck surgeon at Leiden University Medical Center. He is vice-chairman of the department of ENT and chairman of the multidisciplinary head and neck cancer working group of the University Cancer Center Leiden- the Hague.
2. **Dr. Mischa de Ridder**, M.D., Radiation Oncologist at the University Medical Center Utrecht (UMC Utrecht), the Netherlands. Mischa obtained his PhD in 2017 on "Quality indicators in head and neck oncology" and finished his radiation oncologist training in 2019. After working as a radiation oncologist at the Instituut Verbeeten in 2019 and at the Leiden University Medical Center between 2019 and 2021, he recently started working at the UMC Utrecht. His main skills and expertise are in the treatment of head and neck oncology, including paraganglioma.
3. **Mr. Jelmen Roorda**, M.D., Resident Radiation Oncologist at the Amsterdam University Medical Centers (Amsterdam UMC), the Netherlands. After completing his master's degree program in general medicine in 2019, Jelmen gained 2 years of experience working as a resident not in training in obstetrics and gynecology. He recently started his radiation oncologist training in combination with a PhD student position on a project that is a collaboration between the Leiden University Medical Center, the Amsterdam UMC, and the national research institute for mathematics and computer science in the Netherlands (CWI).

Key Person	Interview	Questionnaire
Dr. Jeroen Jansen	yes	yes
Dr. Mischa de Ridder	-	yes
Mr. Jelmen Roorda	-	yes

Table 1. The key persons and their role in the Healthcare surveys

3.1. Problems to solve

A detailed presentation of the healthcare case and the first user requirements elaborated can be consulted in the deliverable D5.1 ("Use Case 1 - Health care requirements"). The following discussion of the etiology and evolution of the paraganglioma case is based on the ZOOM presentation **Dr. Jeroen Jansen** gave the consortium before the interview.

It is known that paraganglioma has a varying natural course. In Figure 2 below, one can see the evolution of the tumor from 2011 to 2019. One can notice that the tumor has hardly grown eight years after the tumor was identified.

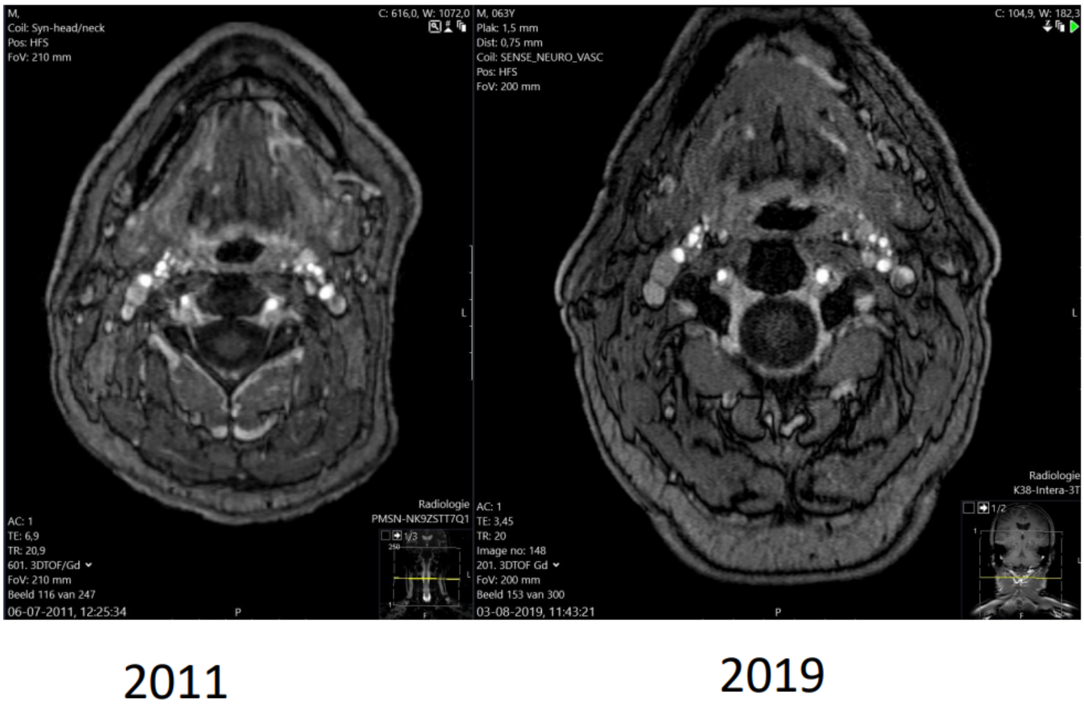


Figure 2. The evolution of the paraganglioma from 2011 to 2019 for a patient (the tumor has hardly grown)

It can be argued that the tumor can be removed in all cases. However, because the tumor has not grown, a doctor thinks we should not perform any surgery. Unlike other malignant tumors that spread in the body, paraganglioma does not spread; its aggressively manifests at a local site.

Moreover, not all paraganglioma cases are like this; the case in Figure 2 was a fortunate occurrence. There are cases when a tumor grows slowly over time, as shown in Figure 3.

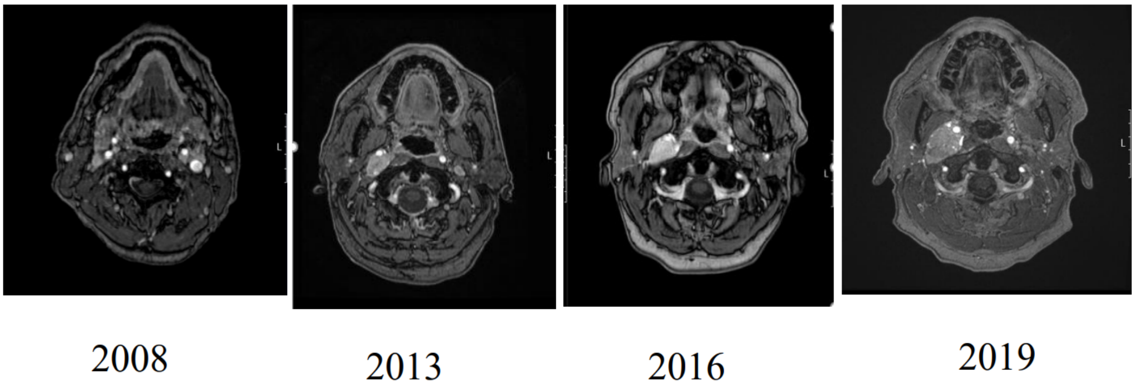


Figure 3. The evolution of the paraganglioma from 2008 to 2019 for a patient (the tumor growth induced vocal cord paralysis in the patient)

In 2019, the patient had vocal cord paralysis due to tumor growth. The paraganglioma has destroyed the nerve leading to the vocal cord, and the patient has problems speaking and swallowing. Therefore, one wished the patient had been irradiated before the problem arose. Unfortunately, a doctor cannot tell if or when the paralysis of the vocal cords occurs. At the time of diagnosis, a choice between no intervention and treatment needs to be made. Treatment of paraganglioma can be done either by surgery or radiotherapy. However, it is not always necessary to use these treatments on the (most often benign) tumors. This is the case if the tumors stop growing and do not cause symptoms. Whether treatment is needed somewhere in the future and when that will be, is most often uncertain up front. Therefore, treatment is postponed until at least persistent growth of the tumor is demonstrated or until the tumor starts to cause (irreversible) symptoms. Forecasting the future development of paraganglioma allows us to make better decisions about the moment of treatment and follow up [D5.1].

Based on the experience with paraganglioma cases, an AI tool has to predict the following things. The first two things are paramount, and the last two things would be nice to have but are not critical.

1. Tumor growth. In the graph below (Figure 4), we see a tumor that proliferates in an interval of time and then stops growing. We want to forecast the eventual size of the tumor.
2. The symptoms. This second problem is related to the first. If the tumor grows, we would like to know the consequences of the growth (for example, the paralysis of the vocal cords). From the MRI scans based on our best knowledge, we cannot tell when this will happen.

3. Surgery result. We want to know the effects of the surgical intervention on the patient: will the surgery work for a patient, or will there be complications?
4. Influencing factors. It is perhaps possible to understand from the model which factors affect tumor growth and the symptoms. It could be the size of the tumor, the blood velocity, the patient's age, or factors related to genetics. Maybe the model can give some clinical clues to help the doctors treat the patients.

The benefits for the patient and the society will be the personalization of the follow-up. If a doctor knows the tumor will not grow, the yearly scans are unnecessary. Increasing the intervals between scans is good for economics (reduces the costs) and reduces the anxiety of the patients who are not so eager to go into the scans. The model could guide the treatment decisions: the model per se will not decide but can give the doctor some clues.

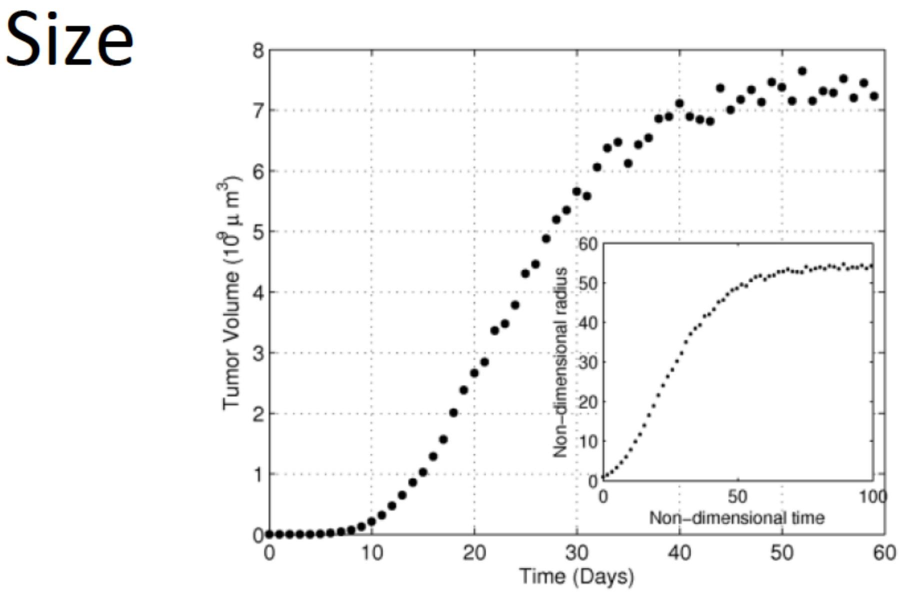


Figure 4. A graph depicting the evolution of a size of a tumor during the time

3.2. Questionnaire

3.2.1. Assumptions

The creation of the medical questionnaire was based on the following assumptions.

- The Recipients of explanations are the doctors.
- Doctors would like to have a broader overview of how (combinations of) different features contribute to the model prediction. They have an expert background and

receive a more technical and complex explanation. The aim is to have doctors validate, accept, and use the model (if it shows enough accuracy) by providing explanations to enhance their trust in the model prediction. Simplified and non-technical informative explanations are also included.

At questionnaire completion, no trained model questions for the paraganglioma case were available. Therefore, we have built a questionnaire for a familiar scenario of diabetes risk. The diabetes data set contains patients from a cohort likely to have diabetes. The AI systems are trained to predict whether these patients have diabetes or not. By using diabetes risk estimation, we avoid biasing doctors' answers about the preferred explanation type because of wrongly formulated explanations.

The resulting questionnaire can be accessible by the following link: https://docs.google.com/forms/d/e/1FAIpQLSelSljsNYbTPq_qlDgCANmPOFCNm7XFtNiBZ01PIDvYx1P3WQ/viewform?usp=sf_link

The medical questionnaire and the corresponding answers are also available as an appendix of this deliverable: Appendix 1. Healthcare questionnaire questions and answers.

The questionnaire was filled by 3 medical doctors that were introduced in Table 1.

3.2.2. Types of questions

The following types of questions were asked in this questionnaire:

1. **Accuracy-Explainability trade-off** asks how much would the doctors inclined to trade model performance over explainability
2. **Formats of presenting model logic**: Graphs based on symbolic regression, Protocols, SHAP feature importance graph, Table with coefficients, Textual explanations and Counterfactuals. These forms were ranked by understandability and effectiveness.
   a. **Graphs based on symbolic regression:** Two graphs are presented that describe the model behavior: Graph A and Graph B. Graph B has two times more operators and used terms than Graph B.
   b. **Protocols** based on genetic programming models were generated. Similarly to the graphs question, two protocols were described: one with a set of 3 rules having one cause in each rule and other with 4 rules and having several causes in some of the rules.
   c. **SHAP feature importance graph** is based on the input features (patient's characteristics like gender, age, obesity, etc.). It is important to know which

features contribute the most to the final model decision and what will be the outcome of changing the value of any feature. Feature importance is a common way to express how single features influence the model prediction. SHAP feature importance graph shows not only features sorted by importance, but data points distribution of each feature.

d. **Table with coefficients** gives a different overview of model feature importance, providing insight with the numerical coefficients of each characteristic in a certain regression model predicting the risk for diabetes. A positive coefficient value indicates an increase in the probability of diabetes when the associated characteristic takes a larger value and vice versa for a negative coefficient. A larger magnitude of the coefficient indicates a larger influence on the prediction due to a change in the associated characteristic. Hence, these coefficients also reflect the influence of each single characteristic on the model's prediction. Another important column is the p-value. Attributes that have significant importance for the model's prediction of diabetes have less than 0.05 in P>|z| column.

e. **Textual explanations** are an effective form for providing short explanations of specific situations. The explanation types that we want to assess are: causal, counterfactual (which alternative characteristic values lead to different predictions by the model) and contrastive (how this patient is different from other, what is now different from the past).

All explanation forms were assessed by interpretability and effectiveness on a scale from 1 to 5. For interpretability 1 means not interpretable, while 5 means very intuitive. And for effectiveness 1 means the explanation is not helping to make a better decision and 5 means very effective.

3.2.3. Discussion of answers

This subsection discusses the answers given by the doctors who completed the Questionnaire.

Accuracy-Explainability trade-off

All doctors agree that some part of model performance can be traded in favour of explainability. However, the amount of accuracy that can be donated to improve explainability and transparency of model decisions differs from 5% to 20%. (link to full version of question and answers: **Accuracy-Explainability trade-off**).

Graphs based on symbolic regression

Two graphs were presented: Graph A and Graph B. They differ by the level of complexity (Graph B has twice as many terms and operators as Graph A).

According to the scores, both graphs were hard to interpret by doctors. It wasn't clear for two respondents what "NOT" meant in the graph. As other comments "conditions that are linked with AND may be combined to a separate condition" and "They are clear, but limited. I am also missing the clear endpoint ". However, the effectiveness answers vary from doctor to doctor from 1 to 5 for both Graph A and B. (link to individual answers: Graphs based on symbolic regression).

Protocols

Two protocols were presented that differ in complexity. Two out of three doctors preferred more complex protocol B giving reasons that "it is more balanced" and "includes more exclusion criteria which might help the predictive value of this graph", however the third respondent preferred protocol A, because it has less variables. The assessment of usefulness of each Protocol corresponds to these preferences. Two respondents agreed that protocols are more intuitive to interpret than visual representations of the respective graphs.

SHAP feature importance graph

We asked respondents to evaluate the understandability and usefulness of the following graph. Surprisingly, the graph got a higher score for understandability and effectiveness than symbolic expression graphs. In order to verify that graph was correctly interpreted, we asked follow-up questions about how some concrete features affect model decision. In general, two questions were answered correctly, however, there was one question that was answered wrongly by 2 out of 3 respondents. Therefore, we can conclude that graph presentation should be simplified. An additional suggestion from one doctor was to explain the meaning of SHAP values.

Since we were afraid that SHAP feature importance graph might be overly complicated, the **simplified version of the feature importance graph** was presented to doctors that don't have feature points distributions. We asked to compare these two graphs to see which version is more preferable. After asking which graph is better we receive very different responses. One doctor strongly prefers graph with feature point distributions. Another doctor prefers simplified graph, it takes less time to read and the relative impact differences between values are more clear. The last participant said that a simplified graph is better for demonstration purposes, however a graph with points distribution is better for interpretation, because you can appreciate the number of events (even though the meaning of shap values are not fully clear).

Table with coefficients

The understandability and effectiveness of the table were assessed low (3 and 2.67 on average). One person pointed out that coefficients do not tell which value means what, i.e. gender has negative coefficient but it is not clear whether woman or man push diabetes risk higher. However, all follow-up interpretability questions were answered correctly by all participants. The suggestion was to include only the coefficient and p-value columns.

Textual explanations

Desired complexity:

One of the questions focused on the complexity of explanations, where answers varied on the number of provided causes. Two out of three doctors choose the most complex explanations with the longest number of causes and the third participant commented that “it depends on how big the difference in risk is between the lady without other conditions and the others. If it is 60% percent risk vs 65% risk it doesn't make sense to add the conditions”.

Preferred type of textual explanations:

Another question provided explanations for the same situation, but in different textual form (contrastive over patient, contrastive over time, counterfactual and causal). Interestingly, that 2 out of three doctors preferred the usual causal explanations. While the third doctor commented that comparing patients is not often useful. It very much depends on the intended use and “it would be used to select patients for a screening test and for that you would want to know how high the risk is. Another use could be to inform patient (if you develop polydipsia you have to return to my office)”.

Counterfactuals

Doctors were asked to evaluate the usefulness of counterfactuals. All doctors agree that counterfactuals are very useful. Additional comments were: “This could be used to increase compliance to therapy. The size of the effect in graphic display, could be used to convince people to comply” and other doctor would want to see in addition values that influence the most and the most clinically relevant.

Final comparison of all forms

The question asks to order previously presented forms in order of personal preference, which resulted in two out of three doctors selecting feature importance as their first choice putting **feature importance view** the highest in ranking. Second place took **textual explanations** that include causal, contrastive and counterfactual explanations being the first choice of one physician and second and fifth of other participants. **Rule-based protocol** took the third place with 2, 4 and 4 choices. And the least preferred form is divided between **genetic programming graph** and **table with coefficients**. We believe that reasoning behind selecting the last two forms as least appealing correlates with understandability of these forms. There were follow-up comments to the genetic programming graph, such as “what does the operator NOT mean in the graph”.

3.3. Interviews

Doctor Jeroen Jansen has been interviewed by the consortium members. The interview has been recorded as a video file and transcribed in Annex 1 (Interviews for Use Case 1- Healthcare).

The following problems have been discussed with doctor Jansen:

1. Clues in the image that can be indicative of a tumor
2. Statistical models to predict the tumor growth
3. Genetic factors that can influence tumor evolution
4. The protocol used by the specialist to train a new doctor in the paraganglioma case
5. The expectations the doctors have from an AI tool for this case
6. The necessity and the nature of explanation for the paraganglioma case

Link to appendix: Appendix 4. Interviews for Use Case 1 -Healthcare

3.4. Conclusion

The aim of the questionnaire was to assess the readiness to trade accuracy over explainability, to explore 5 different explanation forms and gather insights about what healthcare people care the most.

Key insights from the questionnaire:

- Physicians are ready to trade up to 10% accuracy on average to understand the model decision.
- Doctors will trust complex explanations and choose more complex graphs or rule-based protocols over simpler ones.
- Doctors prefer regular causal explanations over contrastive explanations.
- All participants agree that counterfactual explanations are a powerful tool.
- The feature importance graph with instance distributions was the most preferred among all forms.

Key insights from the interview:

- There are no machine learning models to benchmark the accuracy of our developed models against. The ML work for the paraganglioma case is pioneering.
- The genetic information relevant to the patient might play a role in paraganglioma evolution.
- A doctor has no tools to know if a tumor grows or not. The models we trained can be regarded as pioneering work.
- The doctors will assign a weight to the model prediction and decide how to present it to the patients.
- It is crucial for a doctor that the automatic models explain the decisions and not only give a number-based prediction.

4. Trust instantiation in Online Retail

4.1. Problem to solve

Although in-store pickups are available in grocery retail, companies often also perform home deliveries to attend customers in previously agreed delivery timeslots. Managing home deliveries entails a trade-off between operational efficiency, as the company will want to address its customer base with the least cost possible, and customer satisfaction, as the customer would like to choose a convenient timeslot at a fair price.

On the one hand, if the company cannot manage the demand and purely attend to customer preferences, there will be an unbalanced logistic load through time as customers will tend to choose similar timeslots, e.g., late afternoon timeslots after work hours. Moreover, geographically dispersed customers may place orders over the same timeslot [D6.1].

Solving the dynamic time slot pricing problem requires balancing customer preferences and operation parameters. The balance would be maintained by two models: Willingness-to-pay and Cost-to-serve (Figure 5).

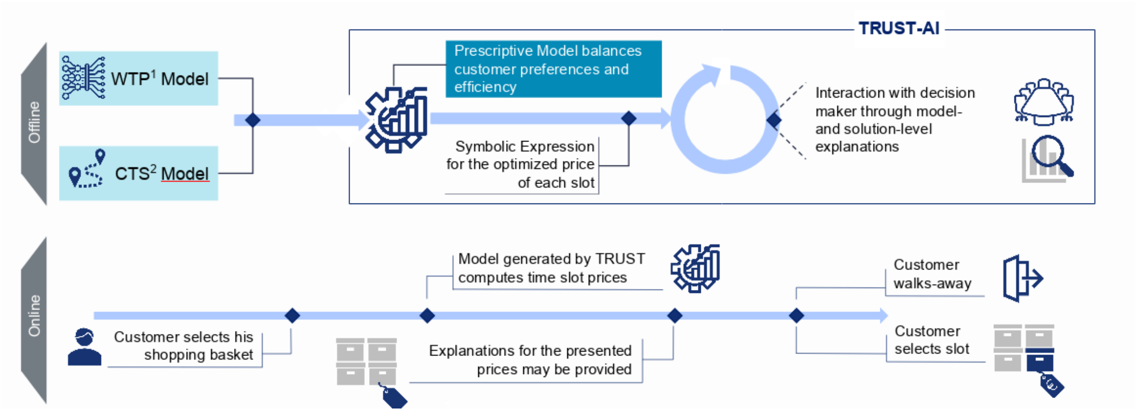

TRUST-AI
Offline
WTP¹ Model
CTS² Model
Prescriptive Model balances customer preferences and efficiency
Symbolic Expression for the optimized price of each slot
Interaction with decision maker through model- and solution-level explanations
Online
Customer selects his shopping basket
Model generated by TRUST computes time slot prices
Explanations for the presented prices may be provided
Customer walks-away
Customer selects slot

Figure 5. Future approach for dynamic time slot pricing

4.2. Questionnaire

4.2.1. Assumptions

The creation of the retail questionnaire was based on the following assumptions.

- **Receivers** of explanations are two distinct groups: **decision-makers** and **customers**.

- **Decision-makers** would like to have a **broader overview** of how different features contribute to the model prediction and optimize the system. They have an expert background and receive a more technical and complex explanation. The aim is to have decision-makers accept, validate, and use the model (if it shows enough accuracy) by providing explanations to enhance their trust in the model prediction.
- **Customers** would like to have a **relatively simplified** and non-technical yet informative **explanation**. The aim is to help them understand/accept the main cause behind the offer they are receiving.

Two questionnaires have been created, one for the decision-maker (Decision-maker questionnaire) and one for the customer (Customer questionnaire).

The customer questionnaire was filled by 7 respondents and the operational manager questionnaire was filled by 1 respondent (Sónia Germano).

4.2.2. Types of questions for Customers questionnaire

The following types of questions were asked in this questionnaire:

1. Price breakthrough aim to determine if atoms (location, demand, your basket, compatibility) of explanations are meaningful for the customer.
2. Types of explanation assess the preferred type of explanation and the customer's ability to understand the explanation provided
3. Summarizing expressions assess the need for a summarizing explanation on top of per item price breakthrough and assess the interpretability and effectiveness of the final form for the customer.

All explanation forms were assessed by interpretability and effectiveness on a scale from 1 to 5. For interpretability 1 means not interpretable, while 5 means very intuitive. And for effectiveness 1 means the explanation is not helping to make a better decision and 5 means very effective.

4.2.3. Discussion of answers for Customers questionnaire

This subsection discusses the answers given by the doctors who completed the Questionnaire.

The first question showed that more than 85% of respondents would be keen to have explanations for the time slot price.

Price breakthrough

Customers understood that **location** and **demand** significantly impact slot pricing, with the same score of 4.6 out of 5. **Compatibility** average score is 3.7 out of 5, and **Your Basket** received 2.4 out of 5. It indicates that customers do not see why larger orders should have an increased price.

Types of explanation

Two situations were provided with a textual explanation, feature importance, population-based and mathematical expression. In a situation with a higher than expected price explanation, 57% of respondents preferred the **feature importance** view, 29% selected **textual explanation**, and 14% selected **population-based** view and no one chose **mathematical expression**. Another situation explains the low price of the time slot, and 71% selected the textual explanation “Someone near your location made an order in this time slot.” The rest, 29%, chose the feature importance view, and no one went for population-based and mathematical expression. An additional comment was that mathematical expressions are very complex and hard to understand.

More situations were provided with textual explanations for advising another slot. The clearest advice was when the actual price difference was provided: “This time slot usually has a price of 4.99, but due to high demand, it’s at 6.99”. This explanation received an effectiveness score of 5 out 5 on average. Similar explanation without showing the actual price: “You are paying a premium for this delivery as demand for the time slot is very high.” get 3.8 out of 5.

Summarizing expressions

The questionnaire showed that 86% found a summarizing expression compelling and that it helped to understand price formation better. Based on the provided visual in the question, the customer is very confident that his basket's content contributes to making the time slot price higher. At the same time, the customer is not confident that demand pushes the time slot high in this situation. Therefore, the customer showed a good understanding of price breakthrough items.

4.2.4. Types of questions for decision-maker questionnaire

The following types of questions were asked in this questionnaire:

1. Accuracy-Explainability trade-off asks how much would the doctors inclined to trade model performance over explainability
2. Model division overview seeks the format of presenting global model behaviour between feature contributions to outcome graph, feature importance and symbolic expression.
3. Comparing different time slots in the same region questions present different formats of comparing customer situations including local visual feature importance, symbolic expression, natural language contrastive and counterfactual explanations.
4. Taking action based on explainable symbolic expressions evaluates the understanding of symbolic expression formulas for decision makers.
5. Evaluate the type of operators that the decision maker is more comfortable with section assess the readability of mathematical operators by decision-makers.
6. Counterfactuals vs Symbolic expressions block assess the need of counterfactual explanations and check if this type of explanations are preferred over symbolic expressions.
7. Critical task interactions where explanations have potential to improve performance check the areas that require insights from an operational managers point of view.
8. Customer behaviour modeling evaluates the interest for explanations of customer behaviour.

4.2.5. Discussion of answers for decision-maker questionnaire

This subsection discusses the response given by the decision-maker who completed the Questionnaire.

The respondent is interested in receiving explanations for all levels: customer behaviour, transportation cost and profit maximisation.

Accuracy-Explainability trade-off

The respondent is ready to trade 10-15% of model performance in favour of explainability. (link to full version of question and answers: Decision-maker questionnaire).

Model division overview

For global model explanation the decision maker finds it more intuitive to understand *feature contributions to the outcome graph* and *feature importance graph* rather than symbolic expression.

Comparing different time slots in the same region

Among the presented types of explanation the situation where custom of a certain region is very different in two adjusted time slots the decision-maker prefers to see feature importance comparison and contrastive natural language explanation.

In the situation where customer drop-outs are high in a certain region, respondents want to explore the situation with interactive graph where it is possible to change prices and see how drop-out rate can change.

In the situation where operational efficiency is privileged completely over customer satisfaction the best supporting visualisation would be a dashboard where system parameters could be altered to expect the impact on the profit.

Taking action based on explainable symbolic expressions

The situation to assess the understanding of a symbolic expression with multiplication and minus sign. Decision maker was presented with action that can be taken to improve the on-time deliveries (OTD). The respondent selected the correct term that will lead to increasing the number of OTD and showed the understanding of such expression.

Evaluate the type of operators that the decision maker is more comfortable with

The decision-maker claims to be comfortable with all presented operators (sum, subtraction, multiplication, division, minimum, maximum, exponential, if then else clauses) except logarithms.

Counterfactuals vs Symbolic expressions

The respondent says that it would be ideal to have a combination of a system with counterfactual explanations that allows to try different scenarios and to see the symbolic expression that helps to understand the direction of the impacts.

Critical task interactions where explanations have potential to improve performance

Decision-maker is willing to have insights on all proposed scenarios: how many customers will be attracted if I reduce the delivery price on a slot, the impact on profit if I nudge a customer from one slot to another, if my fleet should be relocated during the week to adapt to demand patterns, what can be done in terms of slotting and pricing to increase online retail penetration.

Customer behaviour modeling

On a scale from 1 (meaning not interested at all) to 5 (very interested) 4 is the operational manager's interest in having a tool for explaining why some slots are more preferred in certain regions. Such a tool should provide two things: first - a score and the expression used to score each time slot / region combination and second - the classes of customers that are more interested in each time slot / region combination.

4.3. Interviews

Sónia Germano, a Team Lead for E-Commerce Transportation at Sonae MC, the largest e-commerce retailer in Portugal, has been interviewed. The interview has been recorded as a video file and transcribed in Annex 3 (Appendix 6. Interviews for Use Case 2 -Retail).
The following problems have been discussed with Sónia Germano:

1. The acceptable trade-off between accuracy and explainability
2. The use of graphs and mathematical formulas in the explanation
3. Customer behavior based on the dynamic interplay between the slot availability and price
4. The design of a dynamic dashboard for controlling the interplay between operational efficiency and customer behavior.

4.4.Conclusion

The aim of the questionnaires was to assess the need for explanations, the understanding of items that contribute to the price formation and explore the most preferred form.

The key insights from customer's questionnaire:

- Customer questionnaires showed that customers are eager to receive explanations for their time slot pricing formation.
- Textual explanation and feature importance price breakthrough are the most preferred forms of explanations. In contrast, mathematical expressions are hard to understand for the end-user.
- Customers understand that "location" and "demand" contribute to the price; however, "compatibility" and "Your Basket" impact is not clear.

The decision-maker's questionnaire was prepared and explored the need for explanations for experts. The interview with Sónia Germano has addressed the issues in the decision-maker's questionnaire. The following key insights emerged from the interview :

- There is a willingness to trade some accuracy for explainability. However, given that the respondent has no expertise in machine learning, the precise amount of F1 score that can be traded should be taken cum grano salis.
- The explanation should focus on the visual elements, but could also include simple mathematical formulas (the logarithm function could be hard to understand)
- The operational efficiency and customer behaviour should be modelled in a graphical way using a dashboard
- A counterfactual explanation is the best way to model the impact of some decisions (e.g. opening a new time window)
- Modelling the characteristics of the customers based on the region and contrasting the customer behaviour across regions is an important business tool.

5. Trust instantiation in Energy

This chapter of the deliverable is concerned with surveying questionnaires and interviewing energy specialists to understand the explanation and user interface requirements better. The key persons and their role is stressed in the following list and Table 2 below:

1. Mr. **Christopher Moutoulas**, Head of Potomac Trading and Engineering (GR, USA), Industrial Consultant. Relation to APINTech: provider of real time WT technology for building and irrigation applications
2. Mr. **Alfio Galata**, building energy expert, manager of managerial positions in the area (Airport of Milano) with 30 year experience in energy related apps all over the world. CEO of AG Savings (IT) will 2020. Relation to APINTech: Provider as AG Savings of real time WT technology for building applications
3. Dr. **Stavros Chatzigiannis**, Manager at Cyric SA (CY), head of new product development in the area of building utilities (water, hot water, energy). Relation to APINTech: Co-development of assistive technology for a Swiss contractor (2015-1017).
4. Prof. **Nikos Zarkadis**, professor at the university of Geneva (HESGE), expert in the area of building energy. Relation to APINTech: Collaborator with APINTech for the co-development of behavioral technology. Provider of WT technology at the HESGE campus in Geneva.

Key Person	Interview	Questionnaire
Mr. Christopher Moutoulas	yes	yes
Mr. Alfio Galata	yes	yes
Dr. Stavros Chatzigiannis	yes	yes
Prof. Nikos Zarkadis	yes	yes

Table 2. The key persons and their role in the Energy surveys

In addition to the key persons in Table 2, other experts have answered the energy questionnaire, but they were not interviewed.

5.1. Problem to solve

This section is based on Nikos Sakkas's presentation before the interviews to give the necessary background to the interview participants. For more details about the energy case, one should consult the deliverable D2.1- Distributed, multiple data source and trust securing API protocol.

The idea of this project is that the environment presented in Figure 6 will reach the market three years from now.

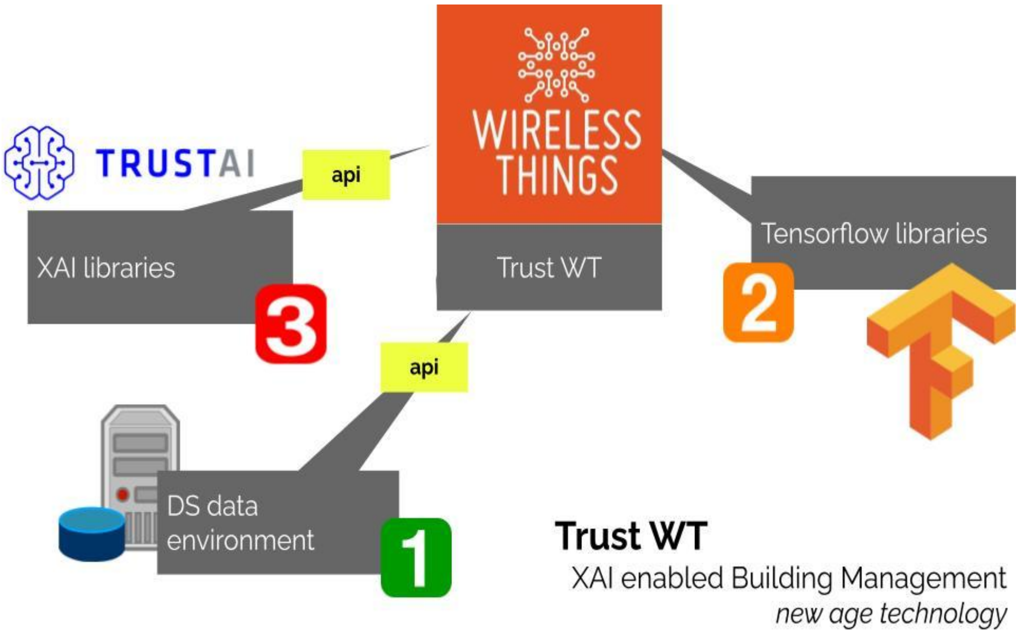


Figure 6. The solution that will reach the market 3 years from now

The real-time data related to the energy consumption of hotels, houses, and offices will be integrated into the environment. Data collected from the private buildings is generally accurate, while data for offices and hotels is, for the time being, distorted because of the COVID-19 pandemics. The data will be processed by the well-known machine learning environment TensorFlow that will generate predictions related to energy consumption. We will also test a genetic programming-based framework that we develop in the TrustAI framework.

But the most critical part of our project is the component related to the explanation. Basically, machine learning forecasting should be explainable to the user. We will test two explanation solutions based on the genetic programming framework and those developed by Google based on TensorFlow.

The milestones for the completion of the project are the following ones:

1. Setting up of the WT environment.
2. The WT environment will send the data to the Trust AI framework
3. The Machine Learning framework TensorFlow will be integrated into TRUST WT
4. The explanations generated by the Trust-AI will also be integrated into the TRUST WT

Regarding the interpretation, we are mainly interested in how the price management of different time intervals affects user behavior. Of course, the total energy consumption will remain the same, but we hope that different price schemas will entice the users to shift their energy consumption.

5.2. Questionnaire

5.2.1. Assumptions

The idea of the questionnaire is to select the preferred format of explanations (tables, charts, interactive graphics, text) and the type of explanations (causal, contrastive, counterfactual, prototype) depending on the aim of the explanation (transparency, trust, accuracy). To narrow down the huge number of combinations to study and keep the questionnaire relatively short, we have made the following assumptions.

Assumptions:

- In the questionnaire we are focusing on building sub-case
- Receivers of explanations that are possible to question at the moment are operational managers
- Operational managers would like to have a broader overview of how different features contribute to the model prediction and how to optimize the system. They have an expert background and can receive a more technical and complex explanation. The aim is to have decision-makers to accept, validate, and use the model (if it shows enough accuracy) by providing explanations that can enhance their trust in the model prediction.
- Customers would like to have a relatively simplified non-technical yet informative explanation. The aim is to help them to understand/accept the main cause behind the offer they are receiving.

5.2.2. Types of Questions

The following types of questions were asked in this questionnaire:

1. **Accuracy-Explainability trade-off** asks how much would the operational managers be inclined to trade model performance over explainability.
2. **Importance of user explanations** questions seek to register your appreciation of the importance of some possible explanations the system can provide to its users as to why the forecasting works the way it does.
    a. Assessment of possible user explanations measure to what extent explanations are needed in different situations for customers.
    b. Assessment of timing and type of explanations identify the desired frequency of providing explanations.
3. **What-if explanations** (counterfactuals): In addition to explaining the algorithm's workings it is equally important to explain to the users how they can affect the forecasting. Several questions were asked to evaluate the importance of counterfactuals.
    a. Assessing possible user action seeks to register the importance of making changes in various scenarios.
    b. Assessment of the type of what-if explanations identifies the most suitable form of providing what-if explanations.
    c. Frequency of what-if explanations identify the desired frequency of providing counterfactual explanations.
4. **Facility managers' explanations** should have a higher overview of data and models from all distinct building spaces. Questions in this category seek the importance of explanations for different situations.
    a. What-if explanations for Facility managers assess how valid counterfactual explanations are for facility managers.
    b. Feature importance graph is based on the input features (patient's characteristics like gender, age, obesity, etc.). It is important to know which features contribute the most to the final model decision and what will be the outcome of changing the value of any feature. Feature importance is a common way to express how single features influence the model prediction. SHAP feature importance graph shows not only features sorted by importance, but data points distribution of each feature.

5.2.3. Discussion of answers

Questionnaire for operational manager

We received six responses: four of the answers are from the people interviewed and who are listed in Table 1.

Accuracy-Explainability trade-off

Operational managers are ready to trade from 2% to 15% of accuracy to receive a more transparent model solution. However, one person commented that “It depends on who we're talking about: I'd say as high as 20% if it's about explainability to the end user / prosumer. If we're talking engineers/domain specialists, 2-5% maybe.”

Importance of user explanations

- Assessment of possible user explanations

The importance of providing explanations for two different situations was measured. **The first situation** asked what caused the consumption increase and following slope reduction. The importance was estimated with the scale from 1 to 5, where 1 means not important and 5 means very important. The resulting average importance score for such a situation is 3.3.

Another situation asked about the difference between two consumption days. The importance score for this problem was graded higher (3.5 on average).

There was a *comment* related to graphs in general “In my experience, non-tech background people (the majority of end users) have trouble reading through graphs (Sometimes graphs can literally scare and put off non tech people!). I'd reckon images with icons and meaningful and carefully chosen key-numbers would be more efficient to pass across messages to most” and proposing giving simpler textual explanations, such as “your predicted energy cons. is 20% higher than yesterday/last week or that 20% corresponds to leaving the TV on all the time for a year or something”.

- Assessment of timing and type of explanations

Four out of 6 participants selected the ***once per week*** option. However, once per day, once per month and once per hour were chosen by some of the respondents. The suggestion was to allow the user to set their own preferences for the information they receive. And additionally, alert when something unusual happens. All participants agree that providing information in a graphical way and showing statistics is important. The comment was that it might be important “to propose the user a bunch of potential changes (quantified and compared to something they can relate to) and let them choose the one(s) that would be the most convenient or suitable for them. Then, display them the impact of their chosen, personal scenario."

What-if explanations

- Assessing possible user action

The first situation evaluates the importance of possibility to see which changes in behavioural pattern are needed to reduce forecast consumption and achieve energy cost/consumption reduction. The situation received a high importance score with an average 4 out of 5.

The second situation evaluates the possibility to see which changes are needed in demand response patterns to reduce forecast consumption and achieve energy cost/consumption reduction. The demand counterfactuals were graded very high (4.5/5).

- Assessment of the type of what-if explanations

All participants agreed that providing what-if explanations in a graphical way by showing two curves (forecasted and the one that follow user action) alongside with statistics is the most effective way.

- Frequency of what-if explanations

The selected frequencies were quite different from once per day ro once per month. Participants suggest allowing managers to choose their own schedule and to alert when something unusual happens.

Facility managers' explanations

- What-if explanations for Facility managers

All respondents agree that presented above user counterfactuals will be also valid to facility managers. They highlight the importance of offering such service to decision-makers. One of the possible blockers that see one of the respondents is having wrong records from the monitoring devices. It will not be possible to provide counterfactuals if the model can not predict visitors behaviour (however it should be possible) or if there is no environmental approach in hotels, offices. Another comment was that different stakeholders (occupants, facility manager, realty agents, owners, energy providers....) need different kinds of information as regards: details, granularity, frequency, presentation etc.

- Feature importance graph

Feature importance graph provides a broad overview of model decision with feature distribution. The understandability of such a graph was graded as 4 out 5 on average. And effectiveness as 4.1 out of 5. Additional comments were that this kind of information would be useful in the hands of experts. And that all relevant information is there.

5.3. Interview

The interviews have focused on different aspects of the energy solution depending on the expertise of the interviewee **:**

1. The Interview with Christopher Moutoulas, focused on some market issues related to the solution and the need of explanation.
2. The Interview with Alfio Galata touched on issues related to disaggregation of the energy consumption and the appropriateness of using modeling tools like genetic programming.
3. The Interview with Stavros Chatzigianni focused on using some insights from a machine learning solution for monitoring water consumption.
4. The Interview with Nikos Zarkadis discussed some aspects of the user interface for users and facility managers.

Appendix 5. Interviews for Use Case 3 -Energy

5.4. Conclusion

The aim of the questionnaire was to assess the readiness to trade accuracy over explainability, to explore the importance of user and facility manager explanations, investigate the desired frequency and forms. Four interviews were conducted and a questionnaire was filled by 6 respondents.

Key insights from the questionnaire:

- Facility managers are interested in receiving a more transparent model with the possibility to trade from 2% to 15%.
- The timing of provided explanations should be up to the user/operational manager's choice, i.e., once per week and alerting when something unusual happens
- What-if explanations are a powerful tool to navigate user behavior and correct demand response for users. Also, such service is important to offer to decision-makers.
- Feature importance graph proves to be effective and useful in the hands of experts.
- Visual graphs alongside statistics are the preferred way of showing explanations, however, there is a concern that graphs might drag off non-tech end customers.

Key insights from interviews:

- A forecasting solution that contains an explanation is better than a simple forecasting solution.
- Users change their behavior if they perceive material gains and the solution is communicated understandably.
- Parts of the forecasting solution and explanation solution could be successfully exported to other domains like gas and energy consumption.
- The graphical solution should contain simple elements for the end-users (an interactive knob could be the way to go)
- The TrustAI solution, at least in the energy case, should contain a smartphone component that conveniently notifies end-users.

6. Design guidelines emerging from the user studies

This section presents the guidelines informed by the user studies relevant to the TRUST-AI tool's architecture.

Even though the use cases cover different domains like medicine, energy, and retail, they have commonalities. The following insights should be taken into consideration when designing the Trust-AI tool.

- All parties were ready to trade some accuracy for better explainability. This insight informs the model developers that there is a trade-off between explainability and accuracy and that they have to consider and analyze the models considering the explainability and accuracy dimensions.
- A feature importance graph is the most effective way to present overall model behavior to the domain experts and the decision-makers.
- Counterfactuals are seen as the most potent cognitive explanation tool. The ability of counterfactuals to model the change of features to achieve the desired result is appealing. Besides counterfactuals, other types of explanations should be implemented for each case.

Based on these insights, the TRUST-AI tool should implement the following modules: a counterfactual module, and a global importance module. These modules are briefly described below and presented in Figure 7. The proposed modules should be integrated into the tool's design by defining the input and outputs and the connection to other modules.

1. The **Counterfactual Module** computes the minimal modifications to the model to change the model decision to the value desired by the domain experts. In the case of classification problems, the value is a class different from the current one. In the case of regression problems, the value could be any number in the admitted range. The counterfactual module can allow “What-if” user queries that answer questions about what will happen if the input is different by customer settings.
2. The **Global Importance** module has to show the most important feature contribution to the result. The user studies suggest a graphical presentation showing the feature contributions to the outcome.

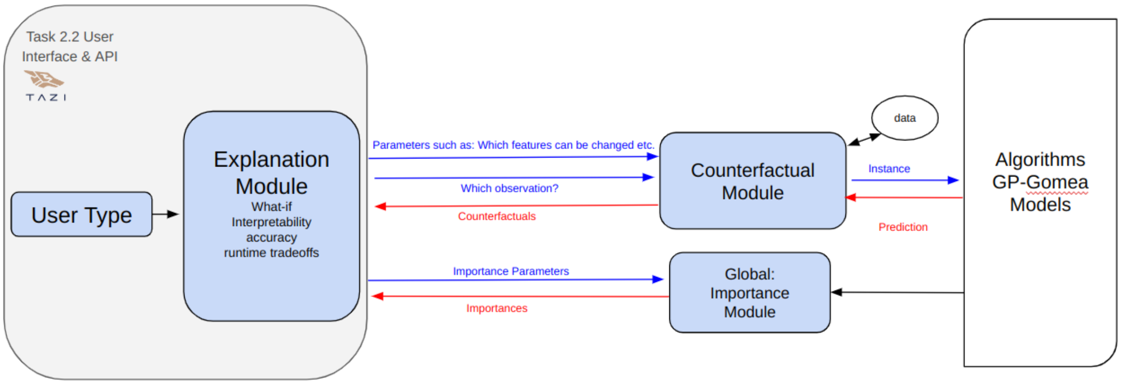


Figure 7. Visualization of TRUST-AI development modules.

7. Conclusion

In this deliverable, we have described the initial user studies performed with domain experts and other relevant users to inform the design of the TRUST-AI tool. Based on the typology of the users in the TRUST-AI project (see the Introduction for details), three modules have been identified and described: the Counterfactual Module, the Global Importance module, and the Scoring Module.

The user studies have been performed by questionnaire and interviews. The received answers to the questionnaires and interviews have been instrumental in determining the specific end users, their desired accuracy-explainability trade-off, their familiarity with different types of explanation, and their preferred complexity and format of explanation. Together with deliverables D5.1, D6.1, and D7.1, the results of this study provide specific recommendations for the TRUST-AI tool design.

8. References

[1] Google, “AI Explainability Whitepaper,” Jul. 2020. Accessed: Oct. 12, 2021. [Online]. Available: https://cerre.eu/wp-content/uploads/2020/07/ai_explainability_whitepaper_google.pdf

[2] F. C. Keil, “Explanation and Understanding,” *Annu. Rev. Psychol.*, vol. 57, no. 1, pp. 227–254, Jan. 2006, doi: 10.1146/annurev.psych.57.102904.190100.

[3] C. Tantitha Thavorn and J. Jiarpakdee, *Explainable AI for Software Engineering*. Monash University, 2021. doi: 10.5281/zenodo.4769127.

[4] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” *Artif. Intell.*, vol. 267, pp. 1–38, Feb. 2019, doi: 10.1016/j.artint.2018.07.007.

9. Appendix 1. Healthcare questionnaire questions and answers

Accuracy-Explainability trade-off

The goal of the question is to identify how the physician values the accuracy-explainability trade-off. Even though symbolic expressions could offer a desirable increase in explainability, its application could be undesired in case the decision maker is not willing to sacrifice some performance.

Question: How much would you be inclined to trade model performance (ability to maximise prediction accuracy) over explainability (increased comprehension of the model's prediction)?

Imagine that a black-box model (e.g., neural network) is our best, although not perfect model to predict the tumour growth over the next few years. This model offers the most accurate prediction among all models available but we cannot say how it does it, which characteristics of the patients are responsible for the model prediction. How much increase in relative error percentage: 100x(true growth - predicted growth)/true growth would you be willing to accept for a model (e.g., a relatively simple analytical expression) in which you could clearly understand how the patient's characteristics influence the model prediction?

I would be willing to work with a clearly understandable model that would increase the relative error (compared to the most accurate model available) up to:

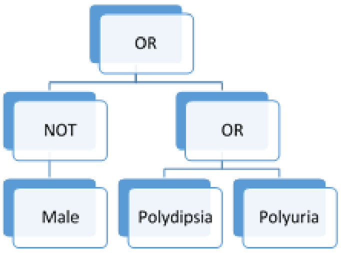
0%   2%   5%   10%   15%   20%   More than 20 %

The answers to this question were quite different **(20%, 10% and 5%)**, but all of them confirm that the party is ready to trade model performance over explainability and transparency of model decision.

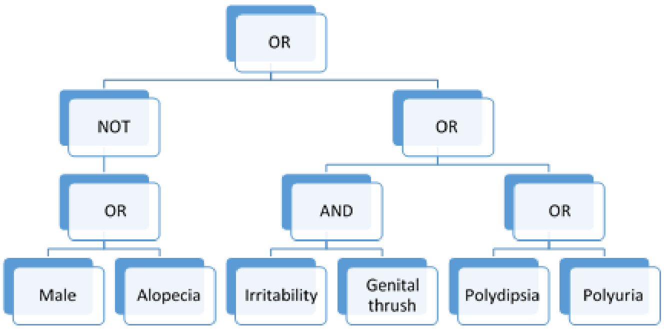
Graphs based on symbolic regression

Consider having the possibility of looking at an ***entire*** prediction model, which estimates the likelihood that a patient is at high risk of diabetes (recall that this cohort contains patients who already show signs of diabetes). Here, the entire model is a graph of logic operations, and should be read from the bottom to the top. If a condition is true (e.g., having Alopecia) then that information is passed to the operation above it (e.g., OR returns true if at least one of the conditions below it are true, while AND returns true only if both of the conditions below it are true). The prediction of the model (high risk=True/False) is given by the last operation, at the top of the graph.

We show two of such graph models below.



Graph A



Graph B

Question: Estimate the effectiveness and interpretability of each graph.

The effectiveness of Graph A was assessed as 1, 2 and 5. While Graph B has marks 1, 4 and 5. At the same time all respondents agreed that these graphs are hard to interpret (1, 3, 4 and 1, 2, 4 for Graph A and B respectively). It wasn't clear for two respondents what "NOT" meant in the graph. As other comments "conditions that are linked with AND may be combined to a separate condition" and "They are clear, but limited. I am also missing the clear endpoint ".

Protocols

Protocols based on genetic programming models. We can present the models (i.e., graphs) shown above in a different format, that is, as a set of rules called a *protocol* rather than graphically. Polydipsia

Protocol A:

We predict that the patient is at high risk of diabetes when one or more of the following applies:

- The patient is Female
- The patient has Polydipsia
- The patient has Polyuria

Protocol B:

We predict that the patient is at high risk of diabetes when one or more of the following applies:

- The patient is Female and does not have Alopecia
- The patient suffers from Irritability and Genital thrush

- The patient has Polydipsia
- The patient has Polyuria

Question: Estimate the effectiveness and interpretability of each protocol.

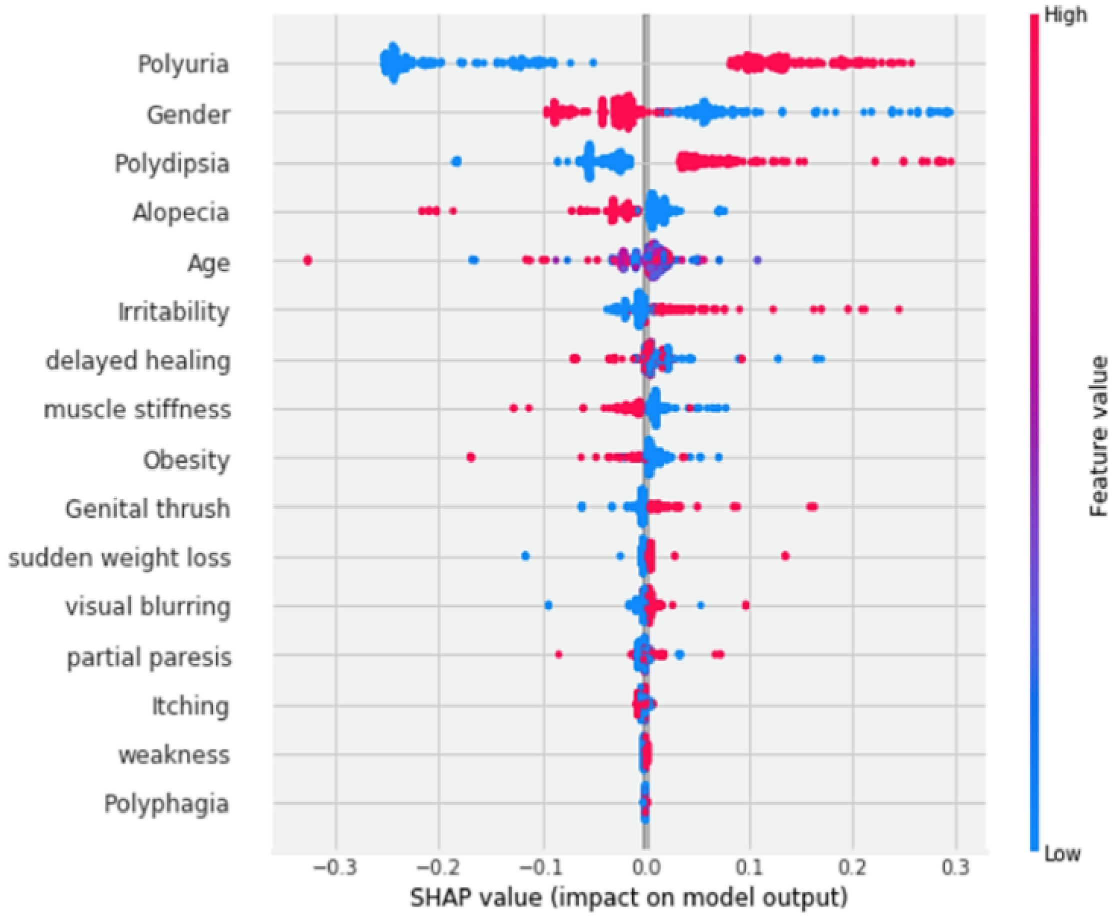
Two respondents prefer protocol B, because it is more balanced and includes more exclusion criteria which might help the predictive value of this graph and one person prefers protocol A because it has less variables. The assessment of usefulness of each Protocol corresponds to these preferences. Protocol A has grades 2, 2 and 5, while protocol B 3, 4, 5. Both Protocols interpretability is high with 4, 4, 5 marks. Two respondents agreed that protocols are more intuitive to interpret than visual representations of the respective graphs.

SHAP feature importance graph

The model prediction is based on the input features (patient's characteristics like gender, age, obesity, etc.). It is important to know which features contribute the most to the final model decision and what will be the outcome of changing the value of any feature. Feature importance is a common way to express how single features influence the model prediction. Below we present different forms of showing feature importance. Here we would like to test how a physician values this type of "explanation" and which format is easier to understand.

How well SHAP is understood? There exist AI models predicting the risk of diabetes based on the patients characteristics. Imagine a patient for which we would like to understand why our AI model is predicting a certain risk of diabetes. The model takes into account many characteristics (age, polyuria, etc...) of the patient but not all of these characteristics are equally important for the model to make its prediction.

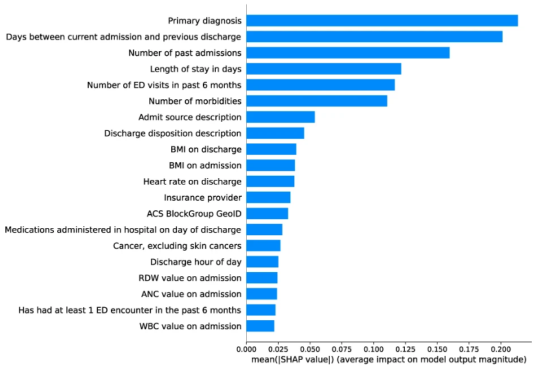
The following graph is a common way to show the most influential characteristics. It shows one characteristic per row (polyuria, gender, etc...), in descending order of their influence to change the model's prediction (diabetes risk). For each type of characteristic the corresponding row shows a distribution of points (one per patient) in which the colour indicates the value of such characteristic for the patients (**red** for a high value of the characteristic or "yes", and **blue** for a low value or "no"), and their position in the x-axis shows how much that characteristic pushes the risk of diabetes higher (right) or lower (left). For instance, for Polyuria the “Yes” value (red) is the most significant contributor to the risk of diabetes (red points are on the right side) while the “No” value (blue) strongly reduces the probability of diabetes (blue points are on the left side).

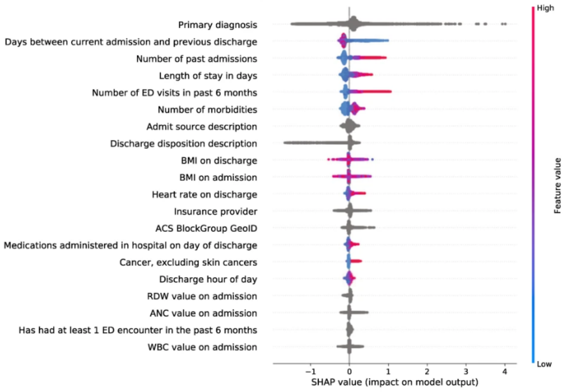


Question: Estimate the effectiveness and interpretability of feature importance graph. Further questions to evaluate the understanding of the graph meaning.

We asked respondents to evaluate the understandability and usefulness of the following graph. Surprisingly, the graph got a higher score for understandability than symbolic expression graphs with 2, 4 and 5 marks and helped to make a better decision with estimation 3, 3 and 5. However, the question “Based on the graph do you agree that Women (coded in red color at the gender row) are at higher risk of diabetes?” was answered wrongly by 2 out of 3 respondents. While the other question “Gender is a more significant risk estimator than age” was answered correctly by all participants. What makes us think that either the description or the presented figure should be simplified? An additional suggestion was to explain the meaning of SHAP values.

Next, the simplified version of the feature importance graph was presented to doctors.





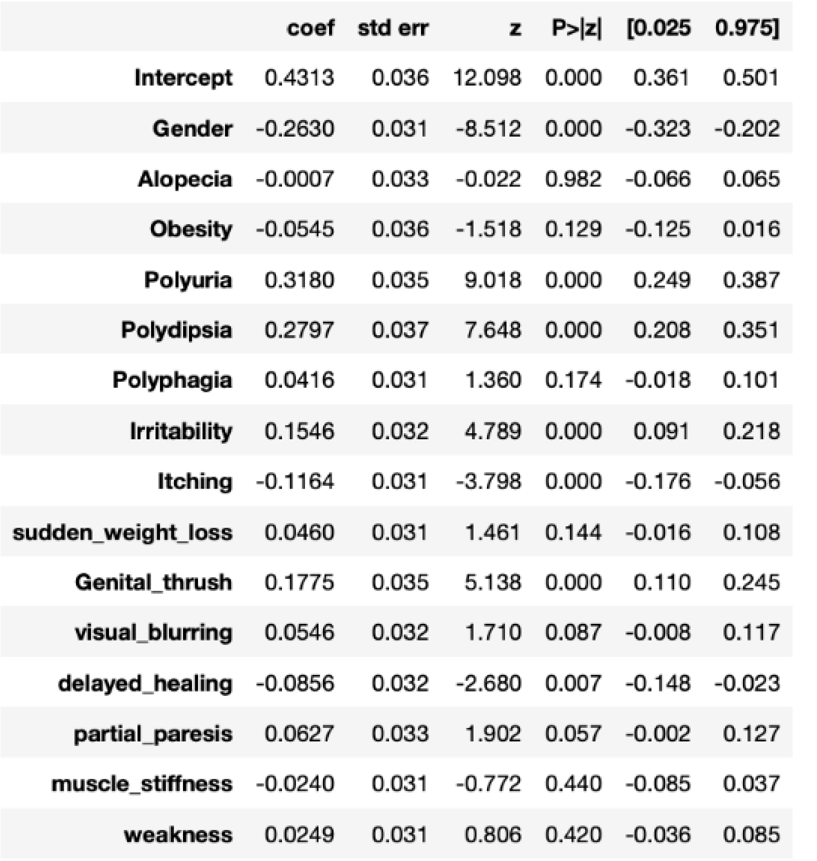
a) Feature importance simplified **b)** Feature importance with
value distribution

Question: Which of these two graphs is more preferable and why?

After asking which graph is better we receive very different responses. Two participants strongly prefer Graph B, another Graph A, because it takes less time to read and the relative impact differences between values are more clear. And another response pointed out that graph A is better for demonstration purposes, but B is better for interpretation, because you can appreciate the number of events (even though the meaning of shap values are not fully clear).

Table with coefficients

Would a table with coefficients be a better form? Now we propose another format. Below you will find a table with the numerical coefficients of each characteristic in a certain regression model predicting the risk for diabetes. A positive coefficient value indicates an increase in the probability of diabetes when the associated characteristic takes a larger value and vice versa for a negative coefficient. A larger magnitude of the coefficient indicates a larger influence on the prediction due to a change in the associated characteristic. Hence, these coefficients also reflect the influence of each single characteristic on the model's prediction. Another important column is the p-value. Attributes that have significant importance for the model's prediction of diabetes have less than 0.05 in P>|z| column.



	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.4313	0.036	12.098	0.000	0.361	0.501
Gender	-0.2630	0.031	-8.512	0.000	-0.323	-0.202
Alopecia	-0.0007	0.033	-0.022	0.982	-0.066	0.065
Obesity	-0.0545	0.036	-1.518	0.129	-0.125	0.016
Polyuria	0.3180	0.035	9.018	0.000	0.249	0.387
Polydipsia	0.2797	0.037	7.648	0.000	0.208	0.351
Polyphagia	0.0416	0.031	1.360	0.174	-0.018	0.101
Irritability	0.1546	0.032	4.789	0.000	0.091	0.218
Itching	-0.1164	0.031	-3.798	0.000	-0.176	-0.056
sudden_weight_loss	0.0460	0.031	1.461	0.144	-0.016	0.108
Genital_thrush	0.1775	0.035	5.138	0.000	0.110	0.245
visual_blurring	0.0546	0.032	1.710	0.087	-0.008	0.117
delayed_healing	-0.0856	0.032	-2.680	0.007	-0.148	-0.023
partial_paresis	0.0627	0.033	1.902	0.057	-0.002	0.127
muscle_stiffness	-0.0240	0.031	-0.772	0.440	-0.085	0.037
weakness	0.0249	0.031	0.806	0.420	-0.036	0.085

Question: Evaluate interpretability and effectiveness of table with coefficients. Further questions on understanding table meaning.

The understandability of the table was assessed between from 2 to 4. And effectiveness as 2, 3 and 3. One person pointed out that coefficients do not tell which value means what, i.e. gender has negative coefficient but it is not clear whether woman or man push diabetes risk higher. The questions “Based on the graph do you agree that Gender is a more significant risk estimator than Obesity for diabetes?” and “Based on the graph do you agree that Polyuria is a more significant risk estimator than Polyphagia?” were answered correctly by all participants. The suggestion was to include only the coefficient and p-value.

Textual explanations

The goal is to assess if a textual explanation is desired and which types of textual explanations are preferred. The explanation types that we want to assess are: causal, counterfactual (which alternative characteristic values lead to different predictions by the model) and contrastive (how this patient is different from other, what is now different from the past).

Question: How complex should a textual explanation be?

Option 1: Patient X has a high risk of diabetes, because she is a 55 year old overweight woman.

Option 2: Patient X has a high risk of diabetes, because she is a 55 year old woman.

Option 3: Patient X has a high risk of diabetes, because she is a 55 year old overweight woman with Polyuria, delayed healing and muscle stiffness.

Answer 1: it depends on how big the difference in risk is between the lady without other conditions and the others. If it is 60% percent risk vs 65% risk it doesn't make sense to add the conditions.

Answer 2: choose Option 3 with the biggest amount of causes.

Answer 3: Patient X has a higher risk of diabetes, because she is a 55 year old woman with polyuria and is overweight.

Question: Which type of textual explanation works better? (Contrastive over patients/Counterfactual/Causal/Contrastive over time)

Expl1. Patient X has a higher risk of diabetes than Patient Y, because Patient X has symptoms of polydipsia.

Expl2. If patient X did not have the symptom of polydipsia, his risk of diabetes would be lower.

Expl3. Patient X has a high risk of diabetes because he has polydipsia symptoms.

Expl4. Now the diabetes risk is higher for patient X, because he has developed polydipsia.

Among all types of explanations 2 medical doctors selected a simple causal explanation and one commented that the explanation will depend on the intended use: “it very much depends on the intended use. comparing patients is not often useful. it would be used to select patients for a screening test and for that you would want to know how high the risk is. Another use could be to inform patient (if you develop polydipsia you have to return to my office)”.

Counterfactuals

The situation was presented where counterfactuals were extracted.

A neural network model is trained to predict the risk for diabetes depending on age, BMI, number of pregnancies, and so on for women of Pima heritage. The counterfactuals below answer the question: Which characteristics of the patient should change to increase or decrease the risk score of diabetes to 0.5?

Person 1: If your 2-hour serum insulin level was 169.5, you would have a risk score of 0.51

Person 2: If your Plasma glucose concentration was 158.3 and your 2-hour serum insulin level was 160.5, you would have a risk score of 0.51

How useful is such formulation of counterfactuals to explain the model's prediction? Why? How would you change it to be more understandable and useful?

Answer 1: this could be used to increase compliance to therapy. the size of the effect in graphic display, could be used to convince people to comply

Answer 2: I would change it to show values that influence the most and are the most clinically relevant.

Final comparison of all forms

Finally, the summarization question was to rank different forms that were presented above (textual explanations, feature importance graph, rule-based protocols, genetic programming graphs and table with coefficients).

Question: Please order forms that you saw in this questionnaire in order of your personal preference?

Two respondents selected a feature importance graph as a first choice and one textual explanation.

Average ranking looks the following:

Feature importance 1+1+3 / 3 = 1.67

Textual explanation 1+2+5 / 3 = 2.67

Rule-based protocol 2+4+4 / 3 = 3.33

Generic programming graph 2+4+5 / 3 = 3.67

Table with coefficients 3+3+5 / 3 = 3.67

10. Appendix 2. Retail case questionnaire questions and answers

Decision-maker questionnaire

The questionnaire was prepared and one response was collected.

Time-slot pricing decisions concern the maximization of profit, while trading-off customer satisfaction and operational efficiency. On the one hand, price adjustments need to take into account customer choice behavior, as it is undesirable to set prices that would lead to customer walkaways. On the other hand, knowing that pricing decisions influence customer decisions, the price offerings will attempt to drive the customer to select a slot that contributes to a lower transportation cost.

Q1. At which level do you wish to receive explanations?

Customer Behavior: Determine what affects the slot selection probability

Transportation Cost: Decipher which elements are crucial in determining the price of serving a customer

Profit Maximization: Receive explanation on the problem as a whole

Answer: The respondent wishes to receive explanations at all three levels.

Accuracy-Explainability trade-off

Imagine that a **black-box** model is our best estimator for time slot price. This model offers the most accurate prediction among all models available but we cannot say how it does it, which characteristics of the patients are responsible for the model prediction. On the other hand, we can use **transparent** model which can explain the contributors to time slot prices, but the performance of such model can be lower.

Q2. How much would you be inclined to trade model performance (ability to maximize prediction accuracy) over explainability (increased comprehension of the model's prediction)? I would be willing to work with a clearly understandable model that would increase the relative error (compared to the most accurate model available) up to:

0% 2% 5% 10% 15% 20% More than 20%

Answer: 10-15%

Model division overview

The model prediction is based on the input features. It is important to know which features contribute the most to the final model decision and what will be the outcome of changing the value of any feature. Feature importance is a common way to express how single features influence the model prediction.

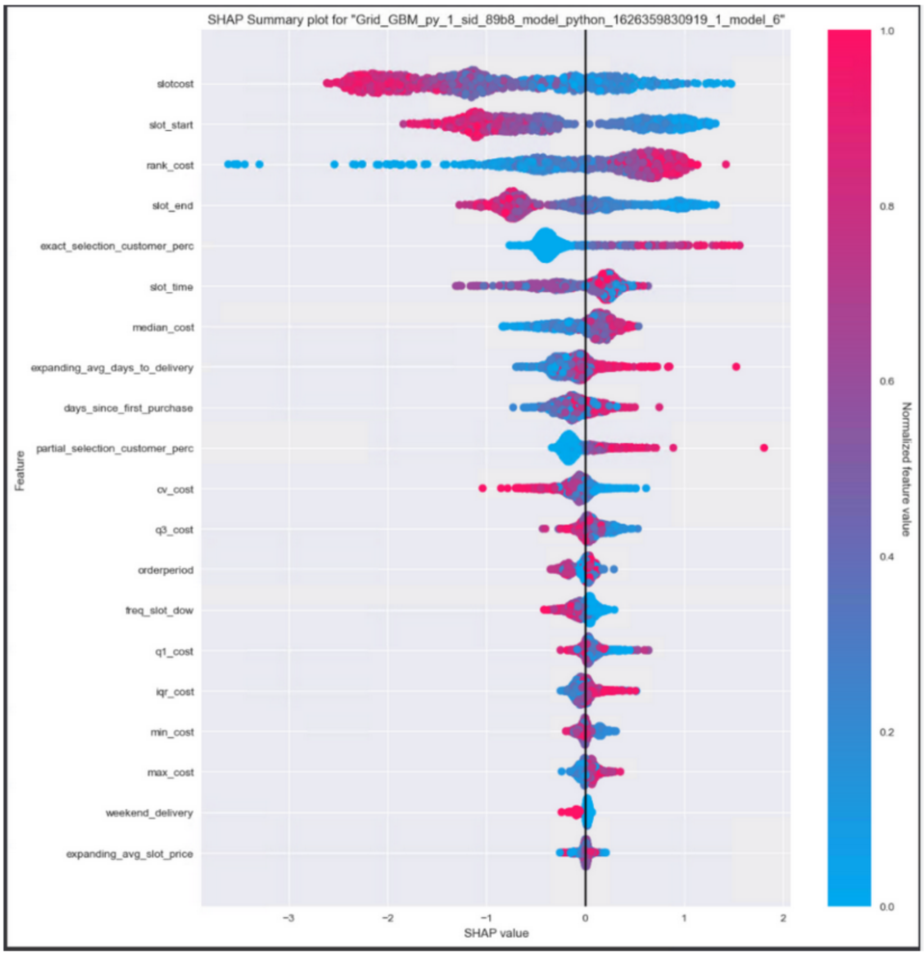
SHAP feature importance graph

The following graph is a common way to show the most influential characteristics. It shows one characteristic per row (slottime, days since first purchase, etc...), in descending order of their influence to estimate price of the slot.

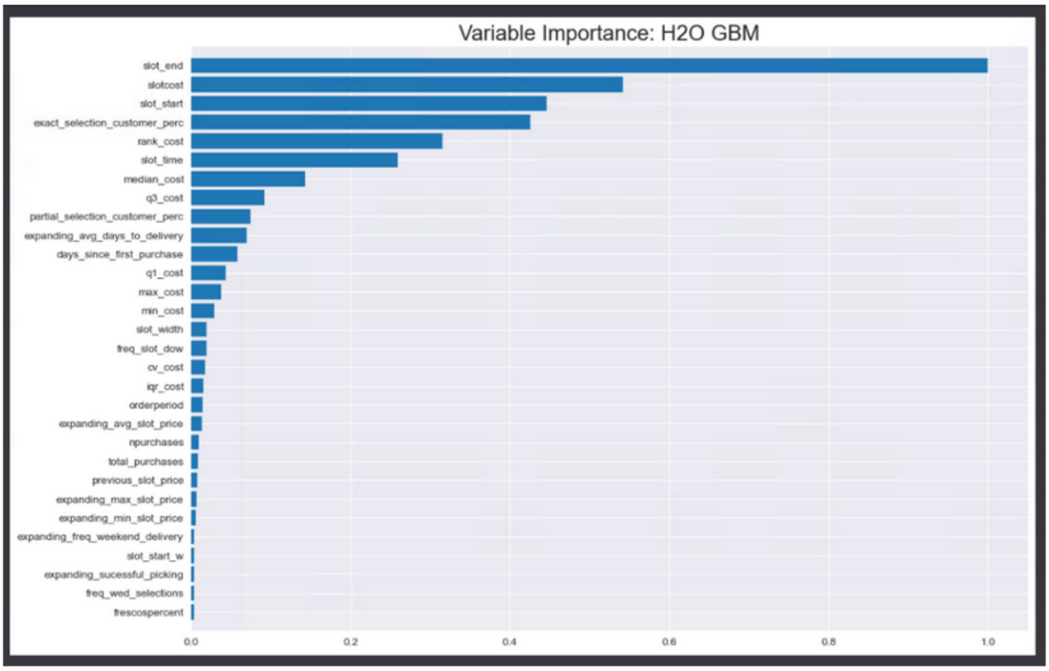
For each type of characteristic the corresponding row shows a distribution of points in which the color indicates the value of such characteristic for the patients (**red** for a high value of the characteristic or "yes", and **blue** for a low value or "no"), and their position in the x-axis shows how much that characteristic pushes the time slot higher (right) or lower (left).

Q3. Focusing on consumer behavior, which format of explanations do you consider more intuitive to understand the main factors that affect the probability of selecting a specific slot?

Feature Contributions to Outcome



Feature importance graph



Symbolic Expression

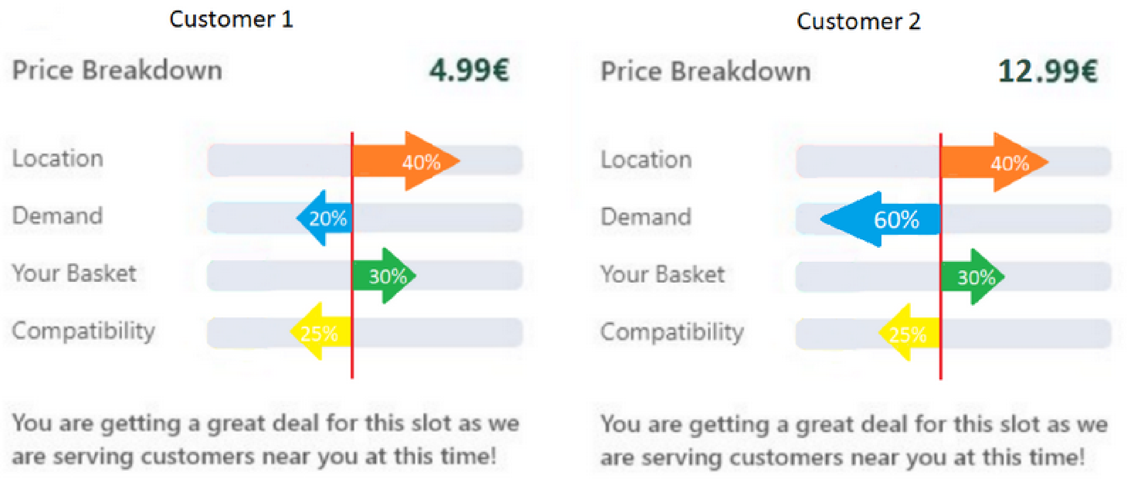
$$Prob_{customer,slot} \propto PastSelection_{customer} \times \frac{\widetilde{Cost}}{if(slot\ is\ 10h30 - 18h00, PastSelection_{customer}, Cost_{slot} - 1)}$$

Anwer: *Feature Contributions to Outcome* and *Feature importance graph*.

Comparing different time slots in the same region

Q4. The support system suggests that the average price charged to customers of a certain region is very different in two adjacent slots. Since a large gap is not to be expected, how would you prefer to check the behavior of the model?

Feature importance:



You are getting a great deal for this slot as we are serving customers near you at this time!

You are getting a great deal for this slot as we are serving customers near you at this time!

Symbolic expression

Natural language: constative

Natural language: counterfactual

Answer:Feature importance and contrastive natural language.

Q5. You notice that customer drop-outs are unexpectedly high in a certain region. Would it be enough to get the description of potential causes or would you like to explore with interactive "what-if" solution?

Options:

- Natural language listing of reasons
- Visualization of two regions with different drop-out rates and explanation of characteristics
- Interactive graph where it is possible to change prices and see how drop-out rate can change

Answer: Interactive graph where it is possible to change prices and see how drop-out rate can change

Q6. In a scenario where operational efficiency is privileged completely and customer satisfaction is neglected, the profit corresponds to only 60% of the optimal profit. For this type of explanation on global solution results, the best means of supporting it would be:

Options:

- A graph performing on sensitivity analysis of profit on different positions in the satisfaction vs. efficiency trade-off
- A dashboard where system parameters could be altered to observe the impact on profit
- A causal graph to understand the underlying relationships between all factors that ultimately affect the global profit

Answer: A dashboard where system parameters could be altered to observe the impact on profit

Taking action based on explainable symbolic expressions

Q7. A certain region has a low number of on-time deliveries (OTDs). You need to take action based on an explanation provided by your explainable decision support system (shown in the picture). What should you do to improve the OTD for this region?

$$OTD = \alpha NumberOfVehicles - \beta AverageDeliveryPrice$$

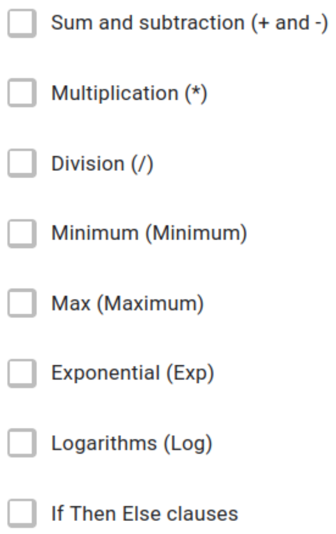
Options:

- Increase the number of vehicles serving the region
- Increase the average delivery price of delivering an order to that region
- The expression does not make any sense

Answer: Increase the number of vehicles serving the region

Evaluate the type of operators that the decision maker is more comfortable with

Q8. Which operators are you more comfortable with?



□ Sum and subtraction (+ and -)

□ Multiplication (*)

□ Division (/)

□ Minimum (Minimum)

□ Max (Maximum)

□ Exponential (Exp)

□ Logarithms (Log)

□ If Then Else clauses

Answer: Everything except logarithms.

Counterfactuals vs Symbolic expressions

Q9. Counterfactual explanations (CFEs) provide "what if" feedback of the form "if an input datapoint were x' instead of x, then the model output would be y' instead of y. For instance, an example of a data point can be tested to understand the impact on a certain measure: "If I open a new time slot in the morning and add one more vehicle to my fleed, what will be the impact on my delays?" Which of the following options applies in your case?

Options:

- I prefer a system with counterfactual explanations that allows me to try different scenarios.
- I prefer a symbolic expression that allows me to understand the direction of the impacts.
- I do not need explanations, I can always understand what is not right by myself.

Answer: A combination of 1 and 2 would be ideal.

Critical task interactions where explanations have potential to improve performance

Q10. Online retail comprises several competitive objectives that need to be faced by operations managers. The interactions between Transportation planning, Demand management, and Pricing are very complex. Which of the following questions would you like to have insights on?

Options:

- I would like to know how many customers will be attracted if I reduce the delivery price on a slot.
- I would like to know what is the impact on profit if I nudge a customer from one slot to another.
- I would like to know if my fleet should be relocated during the week to adapt to demand patterns.
- I would like to know what can be done in terms of slotting and pricing to increase online retail penetration.

Answer: all options were selected

Customer behaviour modeling

Q11. Customer behaviour modeling is a very important issue in online retail. It can provide better understanding on how customers behave and improve demand forecast for each region and time slot offered (crucial for planning logistics resources).

At the moment, there are slots that are more desirable among customers. Would you be interested in a tool for explaining why some slots are more prefered in certain regions?

	1	2	3	4	5	
Not interested.	○	○	○	○	○	Very interested.

Answer: 4

Q12. How would you like this tool to work?

Options:

- Provide a list of the most important features (average age, average income, etc.)
- Provide a score and the expression used to score each time slot / region combination.
- Provide the classes of customers that are more interested in each time slot / region combination.

Answer: both 2 and 3.

Customer questionnaire

Context:

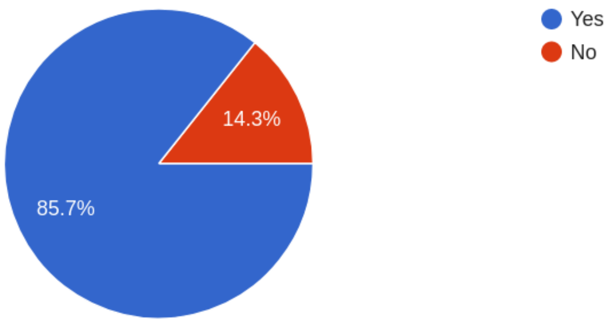
The very last step when making an order on an e-grocery retailer's website is typically choosing a time slot for the delivery. Time slots can vary in price and length. An example of this decision screen is shown below. From the set of available options, the customer chooses a single one.

The following questionnaire serves to select the preferred type and forms of explanations for providing insights into the time slot prices.

The explanations are aimed at increasing transparency regarding the prices shown and to assist the customer in making selections that benefit all parties.

Any feedback in the open comments is treasured.

Q1: Would you be keen on having an explanation for the asked time slot prices?



Yes - 6 answers; No - one answer

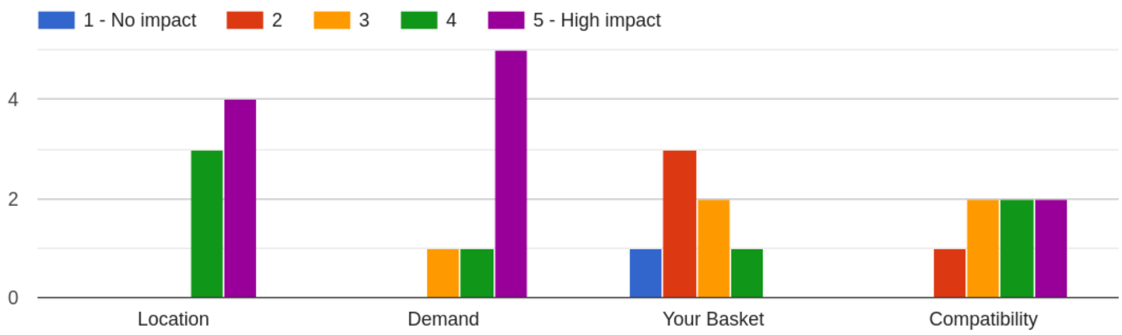
Price breakthrough

For the following questions, assume that the delivery slot price is the sum of several distinct contributions that include the delivery location, the number of items in your shopping basket, the overall demand for each time slot, among others.

Q2: Assume that the following factors influence the time slot price in the following manner: **Location**: "Customers placed in remote locations might see higher delivery prices" **Demand**:

"Time slots with high demand might have higher prices" **Your Basket**: “Larger orders might have a higher delivery cost" **Compatibility**: “Lower slot prices, due to people nearby placing orders for the same time slot". Describe the contribution you expect from each factor on the time slot price:

Answers (impact score from 1 to 5):



Location impact was graded as 4, 4, 4, 5, 5, 5, 5

Demand: 3, 4, 5, 5, 5, 5, 5

Your Basket: 1, 2, 2, 2, 3, 3, 4

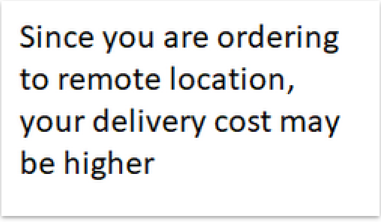
Compatibility: 2, 3, 3, 4, 4, 5, 5

Q3: Any comments regarding the question? - No comments

Types of explanation

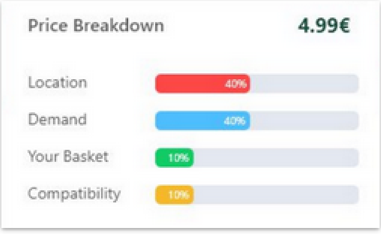
Q4: Assume that you live in a remote place and, thus, the retailer is likely to charge a higher price for delivering an order to your door. Which of the following explanations is clearer to you?

○ Textual explanation

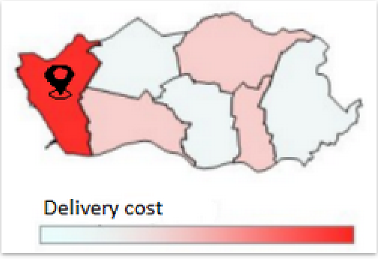


Since you are ordering to remote location, your delivery cost may be higher

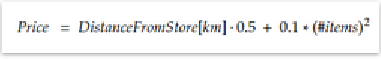
○ Feature importance



○ Population-based

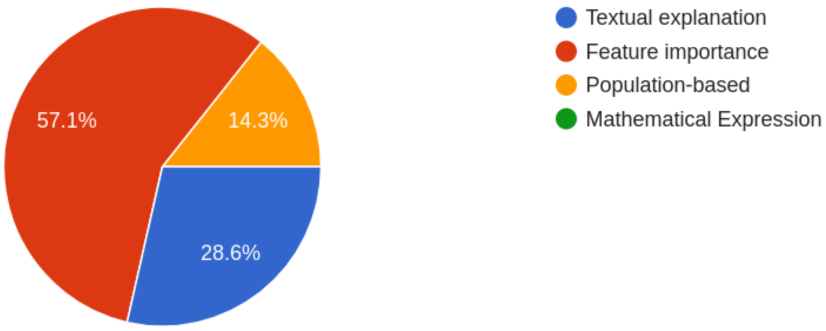


○ Mathematical Expression



$$Price = DistanceFromStore[km] \cdot 0.5 + 0.1 * (\#items)^2$$

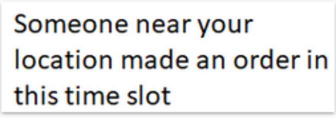
Answer:



Feature importance: 4 people, Textual explanation: 2 people, Population-based: 1 person.

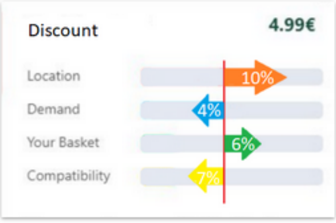
Q5. You notice that a time slot that is not convenient to you has a very low price. Which of the following explanations is clearer to you?

○ Textual explanation

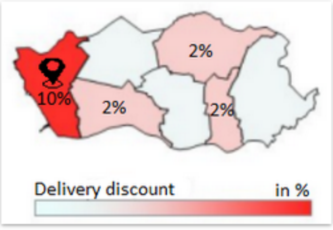


Someone near your location made an order in this time slot

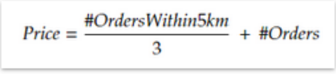
○ Feature importance



○ Population-based

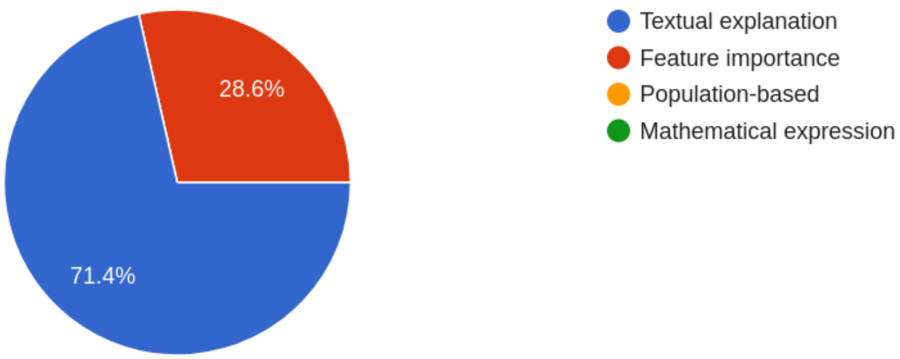


○ Mathematical expression



$$Price = \frac{\#OrdersWithin5km}{3} + \#Orders$$

Answer:



Textual explanation: 5 answers, Feature importance: 2 answers

Q6. Assume that the online retailer wants to advise you to pick another time slot, as the one that you selected is in high demand and, thus, has a higher price. Rank how clear or confusing each suggestion is to you:

A – You are paying a premium for this delivery as demand for the time slot is very high.

B – Other customers selected the Thursday at 10a.m slot and saved 2 euros.

C – This time slot usually has a price of 4.99, but due to high demand it's at 6.99

Answers (confidence score from 1 to 5):

A: 1, 3, 4, 4, 5, 5, 5

B: 2, 2, 3, 3, 5, 5, 5

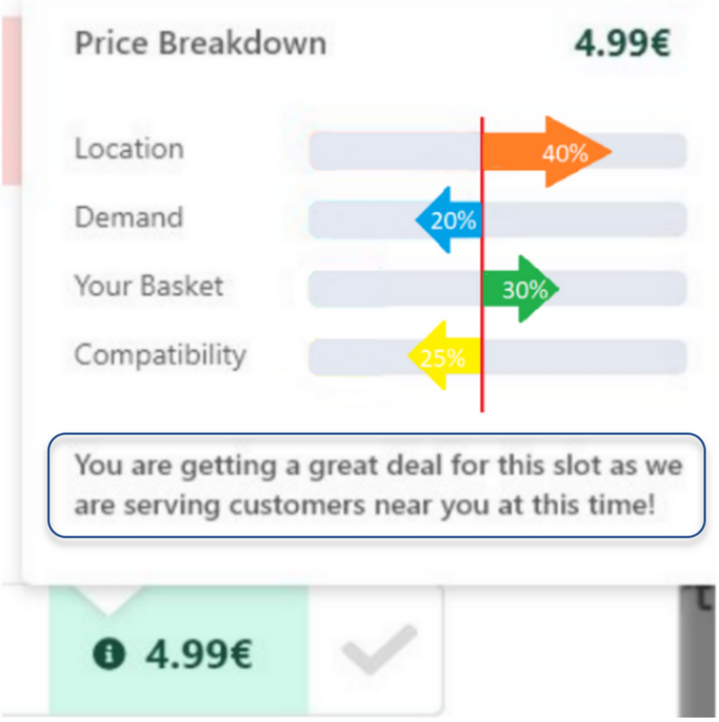
C: 5, 5, 5, 5, 5, 5, 5

Q7. Any comments regarding these questions:1 response

The mathematical expressions are very complex and hard to understand; On Q6, option C is not making a suggestion

Summarizing expressions

Q8. In the illustration a graphical explanation is combined with a summarizing expression. Would a summarizing expression help you understand the explanation better?



Price Breakdown 4.99€

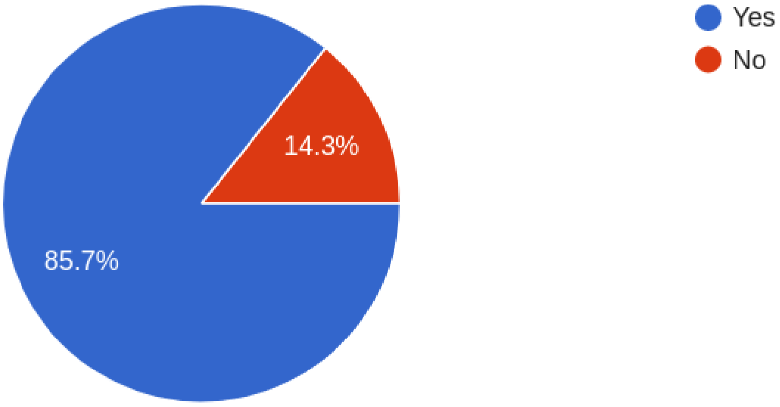
You are getting a great deal for this slot as we are serving customers near you at this time!

○ Yes

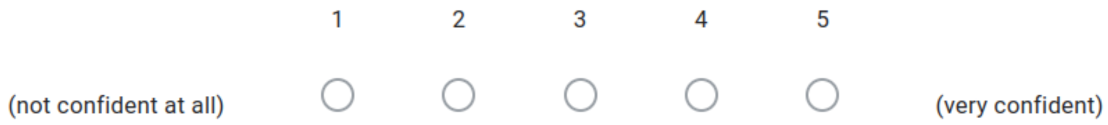
○ No

○ Other...

Answer:



6 answers “Yes” and 1 “No”



	1	2	3	4	5	
(not confident at all)	○	○	○	○	○	(very confident)

Q9. **a**) Your current shopping basket is contributing to a higher slot price

4, 4, 5, 5, 5, 5, 5

Q9. **b**) Demand for the time slot is high

1, 1, 2, 2, 3, 3

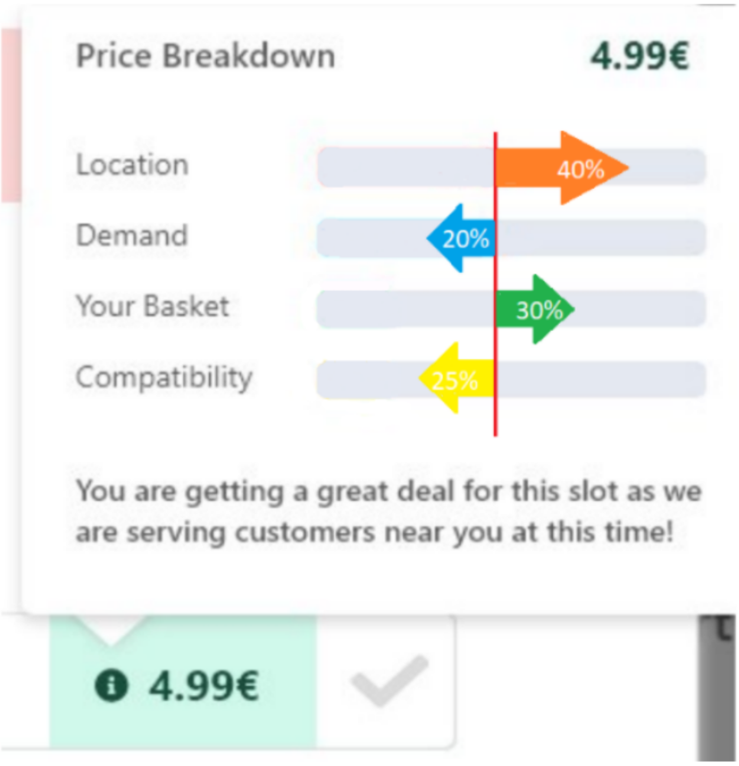
Q9. **c**) Your location isn't favorable for a delivery during this time slot

1, 2, 4, 5, 5, 5, 5

Q9. **d**) Your location and a high demand are the driving the asked slot price up

1, 1, 1, 2, 2, 2, 2

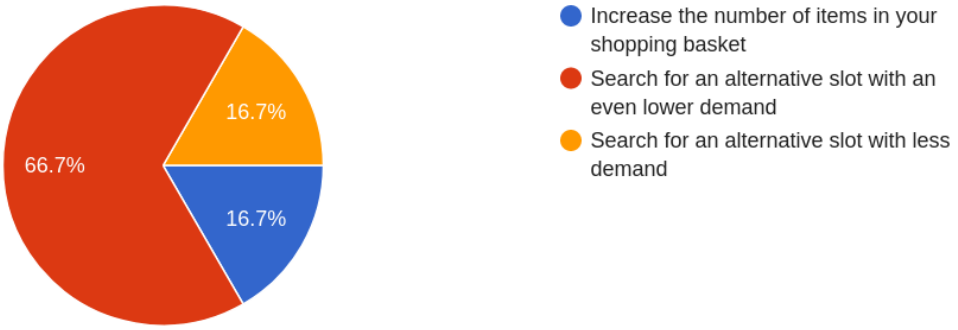
Q10. Given the illustration, what actions could you take to decrease the delivery cost?



You are getting a great deal for this slot as we are serving customers near you at this time!

○ Increase the number of items in your shopping basket

○ Search for an alternative slot with an even lower demand



5 - Search for an alternative slot with an even lower demand, 1 - increase the number of items in your shopping basket

Q11. Any comments regarding these questions:1 response

The Price breakdown graph is confusing. It mixes importance with delta from normal pattern

11. Appendix 3. Energy case questionnaire questions and answers

We are designing a next generation smart home controller that will be able:

- to provide insights on building user behavior especially on observed malpractices (e.g high set-points) . Users can then respond by reconsidering such aspects of their behavior. This will result in lower user energy consumption
- to provide for demand response, i.e., support users in making use of electricity at moments of low price (e.g. by shifting loads to night). Users can then respond by making use of this advice. This will result in lower user energy costs.

The outcomes of such observations might be lower user energy cost by making smarter decisions and lower user energy consumption with better planning.

Accuracy-Explainability trade-off

The goal of the question is to identify how the operational manager values the accuracy-explainability trade-off. Even though symbolic expressions could offer a desirable increase in explainability, its application could be undesired in case the decision maker is not willing to sacrifice some performance.Imagine that a black-box model (e.g., neural network) is our best although not perfect model to predict the energy consumption. This model offers the most accurate prediction among all models available but we cannot say how it does it, which characteristics are responsible for the model prediction.

Question: How much would you be inclined to trade model performance (ability to maximise prediction accuracy) over explainability (increased comprehension of the model's prediction)?

I would be willing to work with a clearly understandable model that would increase the relative error (compared to the most accurate model available) up to:

0% 2% 5% 10% 15% 20% More than 20 %

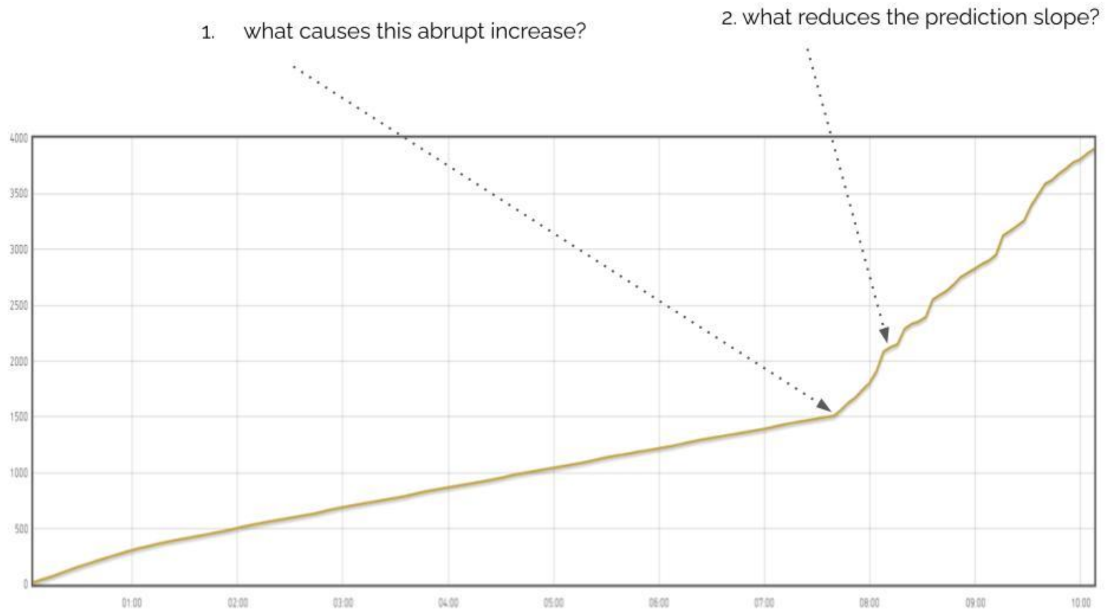
The answers to this question were quite different: **15%, 10%, 10%, 2%, 2%** and it depends "It depends on who we're talking about: I'd say as high as 20% if it's about explainability to the end user / prosumer. If we're talking engineers/domain specialists, 2-5% maybe.”.

Importance of user explanations

The goal of these questions is to register your appreciation of the importance of possible explanation the system can provide to its users as to why the forecasting works the way it does.

Assessment of possible user explanations

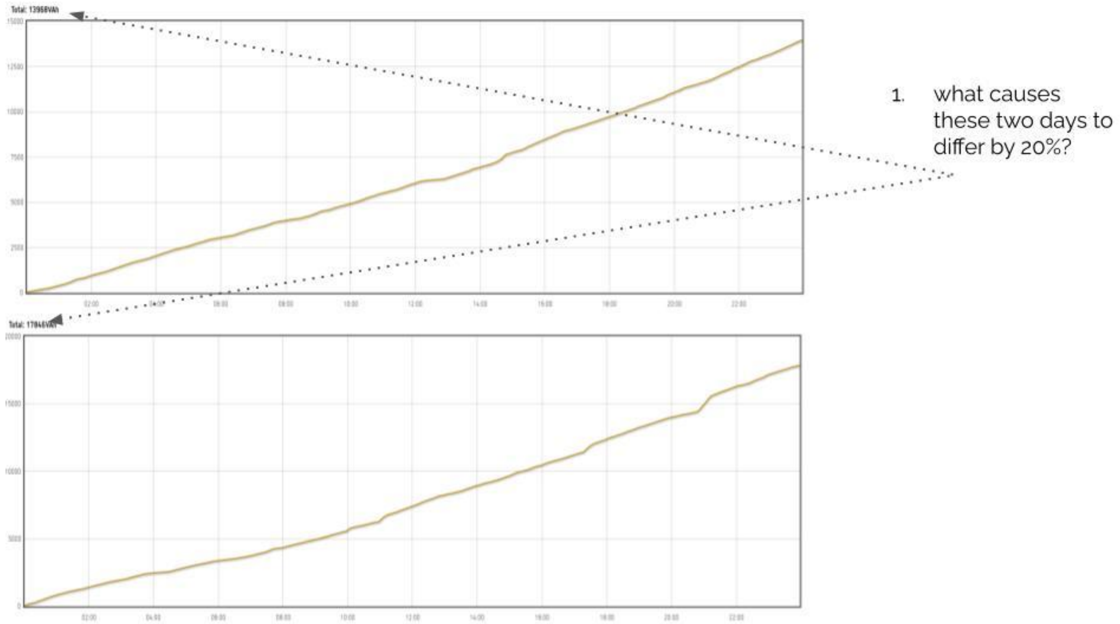
1. What causes this abrupt increase or what reduces the prediction slope?



(not important) 1 2 3 4 5 (very important)

Answers: 2, 2, 3, 4, 4, 5

2. Why was the 1-day ahead prediction less than 20% of the 2 -day ahead prediction?



1. what causes these two days to differ by 20%?

(not important) 1 2 3 4 5 (very important)

Answers: 2, 2, 4, 4, 4, 5

3. What other explanation crosses your mind, if any?

Answer 1: In my experience, non-tech background people (the majority of end users) have trouble reading through graphs (Sometimes graphs can literally scare and put off non tech people!). I'd reckon images with icons and meaningful and carefully chosen key-numbers would be more efficient to pass across messages to most (i.e. think of simple and tangible comparisons which "talk" to people, like your predicted energy cons. is 20% higher than yesterday/last week or that 20% corresponds to leaving the TV on all the time for a year or something).

Answer 2: Significant deviations of predictions could be generated by a low accuracy of the dataset used to train the model or by a real change of the boundary conditions of the system to which the individual measurement is related.

Answer 3: The user pattern was not well understood

Assessment of timing and type of explanations

1. What is the timeframe such explanations should be provided?

Answer 1: Once per day, Once per month, A good idea might be to have the user set their own preferences for the info they receive (=less frustration by untimely notifications)

Answer 2: Once per hour, It depends from the type of measurement

Answer 3: Once per week, alert when something unusual happens

Answer 4: Once per week

Answer 5: Once per week

2. The plan is to provide these explanations in a graphical way, showing the curve (s) and some statistics. Do you agree or do you think that some other way (text, table) could offer any advantage?

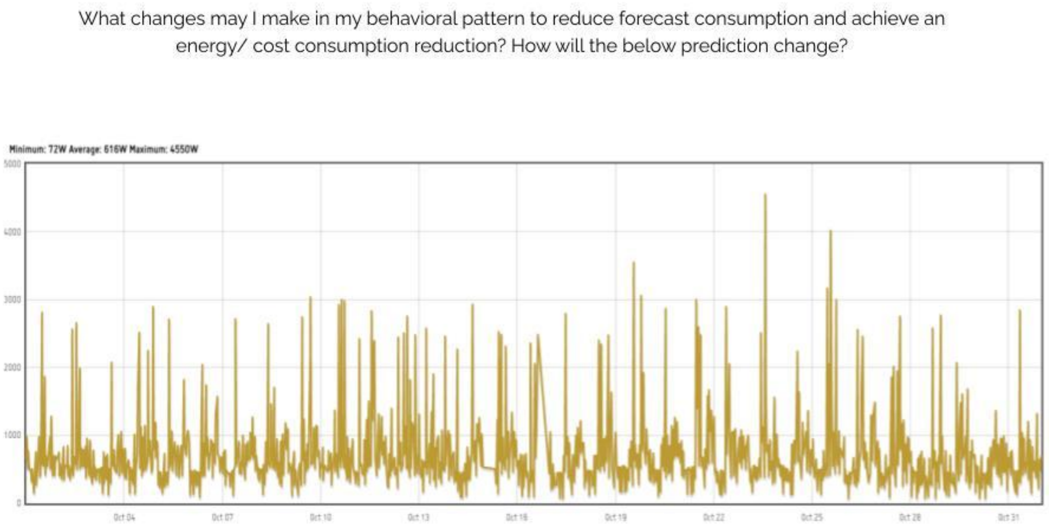
5 respondents agree with this and other participant commented “Here too, it might be a good idea to propose to the user a bunch of potential changes (quantified and compared to

something they can relate to) and let them choose the one(s) that would be the more convenient or suitable for them. Then, display them the impact of their chosen, personal scenario."

What-if explanations

Assessing possible user action

1. What changes may I make in my behavioral pattern to reduce forecast consumption and achieve an energy/ cost consumption reduction? *(example: reduce the setpoint of space 'x' between hour$_1$ and hour$_2$ / benefit estimated at 'y' (kWh. Euros)*



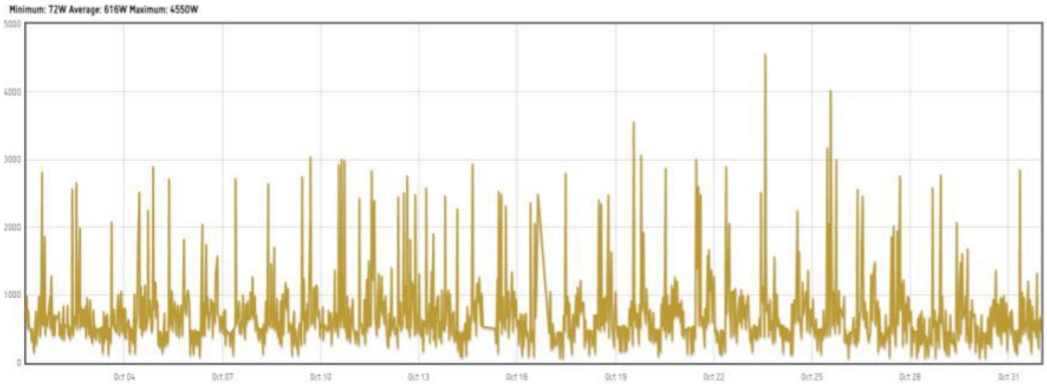
What changes may I make in my behavioral pattern to reduce forecast consumption and achieve an energy/ cost consumption reduction? How will the below prediction change?

(not important) 1 2 3 4 5 (very important)

Answers: 3, 3, 4, 4, 5, 5

2. What changes should I make in my demand response pattern to reduce forecast consumption and achieve a specific overall cost consumption reduction? *(example: shift device 'x' out of peak tariff / benefit estimated at 'y' (kWh. Euros)*

What demand response changes can I introduce to reduce forecasted energy cost? How will the below prediction change?



(not important) 1 2 3 4 5 (very important)

Answers: 4, 4, 4, 5, 5, 5

Assessment of the type of what-if explanations

The plan is to provide these what-if explanations in a graphical way, showing the two curves (forecasted and tweaked- following user action) and some statistics. Do you agree or do you think that some other way (text, table) could offer any advantage?

All respondents agree with this.

Facility managers explanations

What-if explanations for Facility managers

Question: Can you think of any reason that the above user explanations/ counterfactual explanations are not valid any more for the facility manager?

Answer 1: NO, it's always important to offer services to decision-makers.

Answer 2: No

Answer 3: Wrong records from the monitoring services

Answer 4: 1)can not predict visitors behavior (i disagree), 2)no environmental approach in hotels,offices .

Answer 5: I think different stakeholders (occupants, facility manager, realty agents, owners, energy providers....) need different kind of information as regards to: details, granularity, frequence, presentation etc.

Frequency of what-if explanations

1. What is the most suitable time-frame for delivering the explanations in the case of the facility manager?

Answer 1: Once per day, Once per week, Once per month, Yearly, As before managers could also choose their own schedule

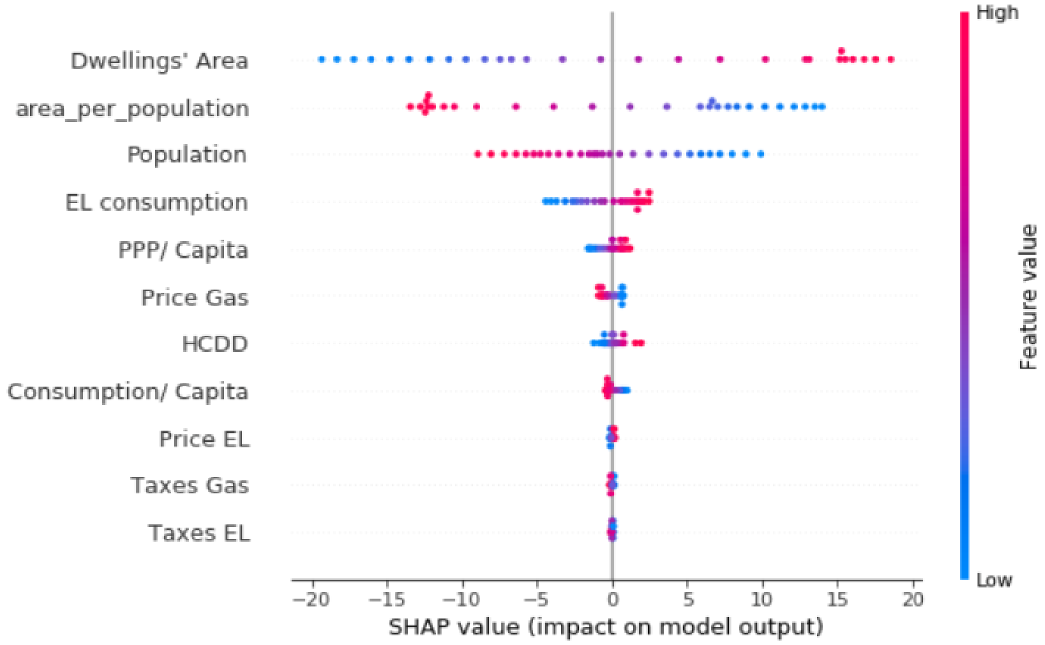
Answer 2: Once per month + alert for unusual energy demand

Answer 3: Once per week

Answer 4-6: Once per day

Feature importance graph

The following graph is illustrative example for energy consumption, but on a country level. Such graph shows the most influential characteristics. It shows one characteristic per row (Dwellings' Area, Population, etc...), in descending order of their influence to change the models prediction (high risk). For each type of characteristic the associated row shows a distribution of points in which the color indicated the value of such characteristic for the customer (red for a high value of the characteristic or "yes" and blue for a low value or "no") and their position in the x-axis shows how much that characteristic pushes the energy consumption higher (right) or lower (left).



Comments: It's ok. All the relevant information are there.

Comments: Yeap, this kind of info may be useful in the hands of experts.
How understandable is the feature importance graph?
(How understandable is graph?) 1 2 3 4 5 (very intuitive)
Answer: 3, 3, 4, 4, 5
How effective is the feature importance graph?
(I don't see any value) 1 2 3 4 5 (very effective)
Answer: 3, 4, 4, 5, 5

12. Appendix 4. Interviews for Use Case 1 - Healthcare

Interview with doctor

Interview with doctor **Jeroen Jansen**

Interview Metadata

The interview took place over Zoom on October 21, 2021, and have been recorded by Peter Bosman.

The **interviewee** is medical doctor Jeroen Jansen.

The **interviewers** are members of the TRUST-AI consortium who have participated in the Zoom meeting with doctor Jensen. The members who addressed questions have been: Eduard Barbu, Raul Vicente, Peter Bosman, Zehra Cataltepe. In the interview transcript, they are collectively named "interviewer."

The interview has been automatically transcribed with IBM speech to text API and has been edited for brevity, clarity, and intelligibility by Eduard Barbu.

The **interviewee**

Jeroen Jansen is a professor of ENT and Head and Neck surgery, particularly Head and Neck Oncology and Skull base Surgery. Jeroen Jansen is a consultant Head and Neck surgeon at Leiden University Medical Center. He is vice-chairman of the department of ENT and chairman of the multidisciplinary head and neck cancer working group of the University Cancer Center Leiden- the Hague.

Before the interview, doctor Jeroen Jensen has given a presentation of the paraganglioma case. The presentation and the interview have been recorded and stored on the INESC TEC drive.

Specialized vocabulary

Before reading the interview, it is helpful to have the following specialized knowledge:

1. "Paraganglioma is a type of neuroendocrine tumor that forms near certain blood vessels and nerves outside of the adrenal glands. The adrenal glands are important

for making hormones that control many functions in the body and are located on top of the kidneys. The nerve cells involved in paraganglioma are part of the peripheral nervous system, meaning the part of the nervous system outside of the brain and spinal cord. These tumors can also be called extra-adrenal pheochromocytomas. " (https://www.cancer.gov/pediatric-adult-rare-tumor/rare-tumors/rare-endocrine-tumor/paraganglioma)

2. Genes are transformed into proteins through transcription and translation. The journey from gene to protein is complex and tightly controlled within each cell. It consists of two major steps: transcription and translation. Together, transcription and translation are known as gene expression.
3. Indolent tumors either stop growing or grow very slowly. Such tumors are often reported in high numbers in autopsy studies because people usually die with them, not of them.
4. The Gompertz (https://en.wikipedia.org/wiki/Gompertz_function) and Von Bertalanffy (https://en.wikipedia.org/wiki/Von_Bertalanffy_function) curves are mathematical models for time series.

The Interview

Interviewer: I assume you are looking at a specific part of an image representing the tumor. I suppose there are particular features of the image that help you diagnose the paraganglioma and predict its evolution. Is there any way to formalize your expertise?

Jeroen Jensen: If I have a patient with tumor scans (at various times) and if the tumor does not vary and there are no complaints from the patient, I expect that the paraganglioma does not grow in the following scan. I think that this simple assumption is a good prediction. Usually, I'm looking at a history of four, five scans, but I assume I could apply the same procedure with fewer scans.

Often there is not so much indication for the treatment. For example, there is not much you can do in the tumor case that I have presented, which leads to the paralyzing of the vocal cords. In this case, radiotherapy is not an appropriate instrument to deal with this tumor. Regarding the decision to perform or not a scan, this depends on how anxious the patient is. The patient's opinion or attitude in front of the tumor is an essential factor when I decide.

But if you ask if something in the picture helps me predict the tumor growth, I can tell you there is nothing. There are no holes in the picture, nor a particular shape, blurring, or any other clues that would help me predict the evolution of the paraganglioma.

Interviewer: Your last observation is crucial as I assumed that by looking at the image as it is the case with some types of cancer, you could say if the tumor grows.

Jeroen Jensen: In general, if there is an enhancement of the contrast or more outgrow, you could expect that the tumor is more aggressive. But in the case under study, the tumors are looking the same. The same observation holds for histology. If you look at the microscope, the aggressive tumors look the same as the indolent ones. The only things that we have found to be predictive are the size of the tumor and the patient's age. The older the patient is, the slower the tumor grows, and the bigger the tumor is, the slower it grows. That is what we have found in a regression study.

Interviewer: The fact that you have tested a linear model is helpful for us.

Jeroen Jensen: We have fit a set of measurements with two curves famous in statistics. The first one is the Von Bertalanffy curve and the second one is the Gompertz curve. They are sigmoid shape curves and they best fit the data, but we do not make any predictions with these statistical models. Probably the S-shaped curves I've shown in the presentations fit this type of tumor as well.

Interviewer: Do the genetic make-up of the patient (for example, the fact that a patient has a mutation in a particular gene) play a role in paraganglioma growth? Is there any effort to acquire data about the specific mutations in genes?

Jeroen Jensen: The study about tumor growth was performed with a homogenous subject group. We did not notice any effect of the different genes on tumor growth in our practice, but we did not study this aspect in depth. Recently, we have got some clues that the phenotype might play a role but also the mutations of particular genes. Different gene mutations give barely functioning proteins, and maybe, in this case, the phenotype is less aggressive than if the protein is not produced at all. Maybe there is something to be found in there. We have the data, and we can use it if we want to answer this question.

Interviewer: How would you explain to a new doctor that you train a patient case?

Jeroen Jensen: I would advise the new doctors to start with the protocol. The protocol is rigorous (e.g., perform scans every year, talk to the patients, etc.). But once the protocol is internalized, you should move on. For example, when the patient does not care and in other particular cases, you might give reasons for not following the protocol to the letter. If the patient is too old or has other diseases that are far more threatening (e.g., the patient has a carcinoma), why would you bother making the scans for paraganglioma? In sum, there are psychological and physical factors of the patient (comorbidity and age) that would influence the actuation of the protocol. Another aspect is related to hospital logistics. For example, if a patient needs to be scanned by an endocrinologist and also needs the scan for paraganglioma, I would advise that the scans be performed together.

But take into account that we are treating tumors, and "tumor" is a scary word for the patient and the doctor. Therefore it is beneficial to assure the patient that the tumor is not dangerous.

Interviewer: It is essential to follow the protocol and to know how to talk to the patient. What other knowledge can you pass to a novice doctor?

Jeroen Jensen: We can pass other scientific data based on our experience: for example, we know in some instances what time to expect for the tumor to double its size. But if I tell you that the average time for a tumor to double its size is 4.5 years, this might not be valuable when the patient in front of you is not the average patient. In any case, you do not use this kind of figure when talking to the patient. I would tell a patient s/he has a small or big chance to develop this or that, but I would be afraid to talk about percentages. The figures might be exact, but we know that human beings (including myself) do not quantify the risk well when we see percentages. If your question is: is th*ere* exact knowledge about this tumor? My answer is there is. But if you ask: do you use that knowledge in the clinic? My answer is: not so much.

Interviewer: What would you then expect from an AI prediction tool, given that this kind of tool will most likely give you numbers, percentages? In light of what you have said before, I think that the doctor (but not the patient) will see the tool prediction. The doctor will then produce an explanation to the patient.

Jeroen Jensen: I think that you are hitting a crucial spot. It would be nice to have independent confirmation of your intuition that this tumor will be indolent. If the model says there is an 84 percent chance that the tumor will not grow, it will help me. But still, there is a 16 percent chance that the tumor is going to grow. Therefore, you will use the prediction in a guiding way: it will help you but definitely but will not decide for you. If a patient will ask me: will I develop vocal cord paralysis? And the model will give me a 95 percent chance that this will not happen; this fact might be communicated to the patient. In general, though, the patient should not see its model prediction.

The doctors will assign a weight to the model prediction and choose how to present it to the patients. It might sound paternalistic, but we know the patients better.

Interviewer: In addition to numbers, the model will also tell you the causes for the prediction. It will say to you, for example, that a tumor of a particular patient has an 80 percent chance to grow because this and that happened. What will you do with this explanation of the prediction?

Jeroen Jensen: Indeed, the difference between this project and other projects is that you have an explanation for the prediction. Knowing the basis for the prediction is valuable because it allows me to elaborate a more informed explanation for the patient. For example, suppose the model will tell me that the tumor color is white and has a specific size. That information corroborated with a patient's age gives an 84 percent chance that the tumor will not grow. In that case, it will be easier for me to elaborate advice for the patient than if I only knew that the model says that there is an 84 percent chance that the tumor will not grow.

But I'm anxious to see this model working and making supported predictions. I can only hope that the explanation is in the data we are collecting. However, I don't know the explanation, so finding it will be a significant achievement evolving from this project.

Interviewer: Evolve is the right word. We hope too that our evolutionary algorithms will find the explanation.

Thank you, doctor Jensen. Your presentation and the following QA session were precious.

Appendix 5. Interviews for Use Case 3 -Energy

Interview with **Christopher Moutoulas**

Interview Metadata

The interview took place over Zoom on November 17, 2021, and have been recorded by Nikos Sakkas.

The **interviewee** is the Head of Potomac Trading and Engineering, Christopher Moutoulas.

The **interviewer** is Nikos Sakkas

The interview has been extracted from a larger discussion that have taken place between Mr. Sakkas and Mr. Moutoulas and ha been edited for brevity, clarity, and intelligibility by Eduard Barbu.

The **interviewee**

Mr. Christopher Moutoulas is the Head of Potomac Trading and Engineering (GR, USA), Industrial Consultant. His relation to APINTech: provider of real- time WT technology for building and irrigation applications.

The **interview**

Nikos Sakkas

Now that you have seen and completed the questionnaire, can you tell us, please, what is your opinion about the framework and the explainability idea?

Christopher Moutoulas

I first want to know if I've understood correctly. You are telling me that the users will be informed randomly about the price intervals of the electricity (e.g., they will not know what the electricity price offer will be in 6 months from now). Therefore, there will be much price variation.

Nikos Sakkas

Yes, you are right but keep in mind that the electricity contract also has a fixed price component.

Christopher Moutoulas

If you ask me about building a system that will help the retailer and the users, I think this is a brilliant idea if your assumptions work. If you can get the information and the information is valid, and you have a vast number of users, and other few conditions. Regarding your observations that the users will be the primary beneficiaries, I agree, but I also think the energy retailers will benefit from the flexible price schemes. If the suppliers can offer better prices, but there is no way to communicate this to the users, then the users cannot take advantage of the price scheme.

Nikos Sakkas

All flexible price schemes seem not to have traction in Europe and USA because of the perceived risk. Adding an explanation to the user can help us to mitigate the risk. Our target users are homeowners and offices. The characteristic of our users is that they engage in repetitive consumption patterns. Do you think that this way of explaining things can be generalized to water and gas consumption?

Christopher Moutoulas

Simply put, I think that what you have is an excellent idea to pursue, and I also believe that it can be extended in principle to water and gas consumption. I would certainly enter a scheme where the explanation is part of the contract.

Nikos Sakkas

Do you think we should sell the technology to the homeowner or the electricity provider?

Christopher Moutoulas

My first reaction is always: stay away from the retail. I would try to sell the technology to the energy provider (I distinguish between the retailer and the energy provider, but if for you, they are equivalent, you should conclude). The reason is this: I would want to deal with a big guy than with thousands of small ones. All these little energy providers that are in Greece are tiny. You should ask yourself: are they going to be here tomorrow? We learned our lesson in the telecommunication area, where there were a bunch of providers that eventually did not get anything from the market share. Even if it is faster to deal with small providers, you should ask yourself if they will be growing fast enough for your technology to become profitable.

Nikos Sakkas

What do you think about the explanation component? Is it crucial or not?

Christopher Moutoulas

As a user, if I had to choose between two approaches, one that is a black box and the other would make even an attempt to explain how the prediction works, I would go for the explanation approach. When your competition arrives, you will have an advantage if they provide only a black box system. The user will perceive that your product is better than what the competitor offers if your product has an explanation. An explanatory component makes even more sense in the medical domain because people, including medical professionals, tend to make judgments based on feelings rather than logic.

Interview with **Alfio Galata**

Interview Metadata

The interview took place over Zoom on November 19, 2021, and have been recorded by Nikos Sakkas.

The **interviewee** Mr. Alfio Galata is a building energy expert, manager of managerial positions in the area (Airport of Milano) with 30 year experience in energy related apps all over the world. CEO of AG Savings (IT) will 2020. His relation to APINTech: Provider as AG Savings of real time WT technology for building applications

The **interviewer** is Nikos Sakkas. A question have been addressed by Eduard Barbu. The interview has been extracted from a larger discussion that have taken place between Mr. Sakkas and Mr. Galata and has been edited for brevity, clarity, and intelligibility by Eduard Barbu.

The Interview

Nikos Sakkas

Alfio, what do you think about our idea of building an explanation framework on top of a forecasting tool?

Alfio Galata

First, I would like to understand if you do not want to predict variables other than energy consumption. For example, I think of the variables like the comfort of the house or issues affecting the maintenance of the devices.

Nikos Sakkas

Maybe later, we will integrate other variables, but the project's target is energy consumption.

Alfio Galata

You will predict the energy consumption considering that you have a closed system; we fix boundary conditions. If, for example, we are talking about a house, you are not considering the addition of new devices that consume electricity. When relating energy consumption to cost, you have to consider the tariffs (e.g., the fact that you have a tariff for the day and another one for the night). It is crucial to notice that the tariffs can differ depending on the type of building, the country's economic policy in the electricity domain, etc. The global energy prediction is relatively easy, but let me ask you something: are you using neural networks for prediction or other algorithms?

Nikos Sakkas

We will test different models: neural networks will be one of them, but the project focus is genetic programming.

Alfio Galata

That's good. It is a fascinating area. In a European project called Edificio, we have used a mix of neural networks, genetic algorithms, and fuzzy logic to beat the state-of-the-art algorithms by a 10 percent accuracy. Will you perform source disaggregation? It is essential to know how much energy is consumed and what devices are consuming energy.

Nikos Sakkas

We will not go too much into source disaggregation. I have talked to an Irish expert you also know, who claims that source disaggregation works only in very simple environments.

Alfio Galata

But if you want to perform forecasting, this will not be of much value to the user.

Nikos Sakkas

You are right! But we would like to go beyond forecasting, anticipating the users' behavioral patterns, and issuing advice for demand response schemes.

Alfio Galata

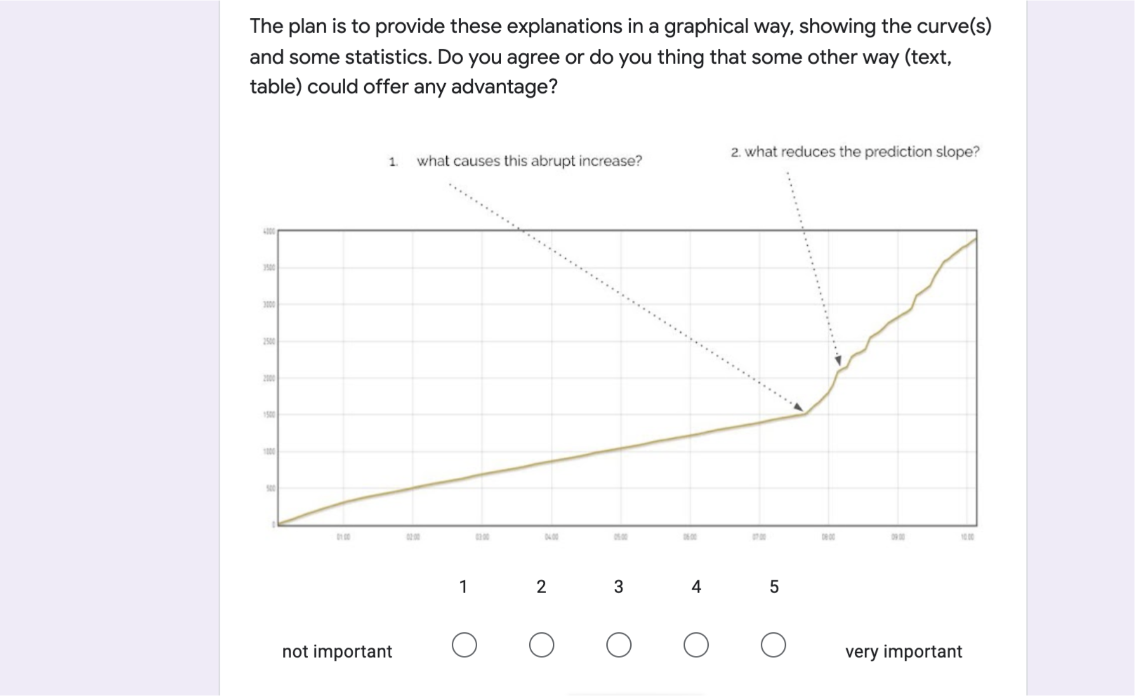
But I still see an essential aspect in forecasting the user's energy consumption based on the electricity devices he uses. For example, you might predict that 10% of a user's energy consumption is due to the washing machine, 15 % of the user's energy consumption is due to using other appliances, etc.

Nikos Sakkas

I do not think that knowing how the energy is distributed leads to savings.

Alfio Galata

But in this case, you cannot answer the question of why is an increase in energy consumption. And I would say you do not know what is responsible for the rise in energy consumption. If we look, for example, at the first graph in the questionnaire, I wonder how you can explain it.



The plan is to provide these explanations in a graphical way, showing the curve(s) and some statistics. Do you agree or do you thing that some other way (text, table) could offer any advantage?

	1	2	3	4	5	
not important	○	○	○	○	○	very important

First graph in the questionnaire (reproduced for intelligibility)

Indeed, I cannot understand why the spike in the graph happened. It might be the case that the user has been using the coffee machine in the morning.

Nikos Sakkas

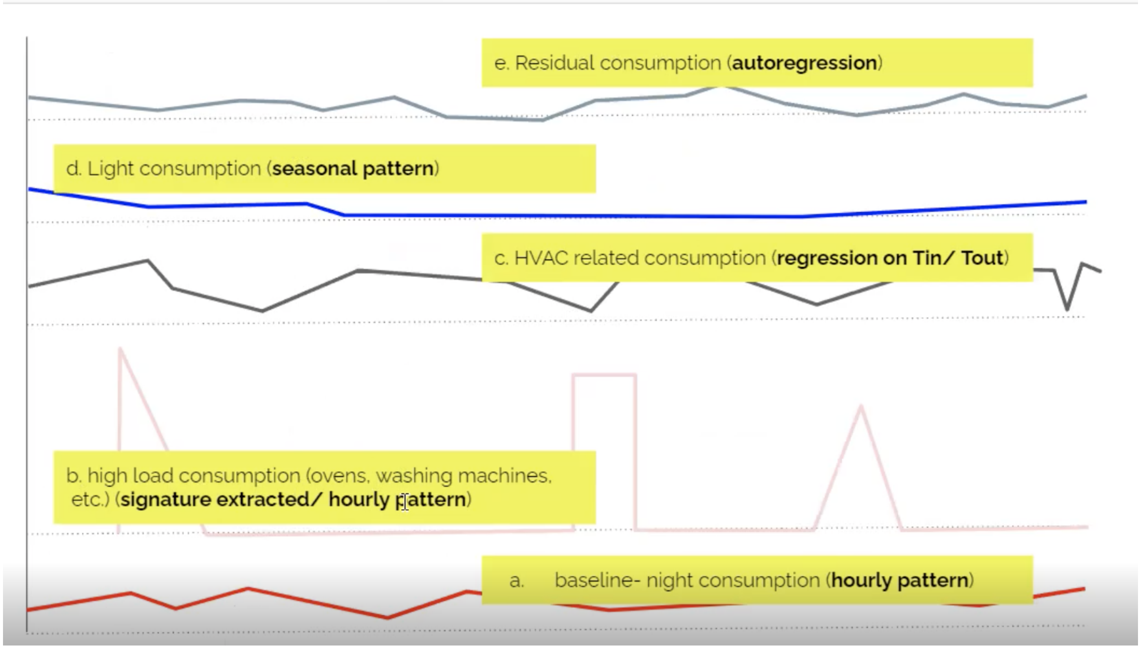
I think this is the direction: behavioral advice coupled with the explanation. In some cases, I do not know why a spike has happened, but there are many cases where I can give a bit of advice.

Alfio Galata

But if you want to provide advice, you also want to know the reason for a surge in electricity consumption. I think a deterministic model that you know and has solved the problem of source disaggregation is more appropriate than a neural network or a genetic programming model where such information is not considered.

Nikos Sakkas

Again I do not see where the problem is, and I do not think we need source disaggregation. Let’s look together at this graph (the graph below). If you look at the spikes, there is a surge of electricity related to extra-energy consumption. And our purpose is to learn these spikes from the data. But I think you are right: it was not evident in the question where the energy has been measured(at the flat level or the level of a block of flats). I want to mention that the sensors for recording the electricity at the apartment level are very cheap. They cost less than 100 Euros.



The second graph referred by the question above (reproduced for intelligibility)

Eduard Barbu
From your experience, if we tell a user to shift their energy consumption from high-cost intervals to low-cost intervals, would they follow the advice?

Alfio Galata
My answer is a strong yes. I can also give you an example of what we do in our family. My wife will turn on the washing machine and dish machine and iron whenever possible after 19:00 or during the weekend where the tariffs are lower. We consume 70% of electricity when the tariffs are lower and 30 percent during peak hours with this family policy.

Nikos Sakkas
Who do you think should be the priority of the explanation: the energy retailer or the house owner?

Alfio Galata
With the advent of the free energy market, many hundreds of energy retailers have popped up in Italy. Unfortunately, these hundreds of energy retailers who call you ten times per day to sign a contract with them will propose a unique tariff. They will try to convince you that their tariff is cheaper on average than their competitors. But according to my simulation, you will not have cost savings unless you have two tariffs. Therefore, I think the focus should be the house owner but not the condominium if this one is part of a condominium.

Nikos Sakkas
The one tariff is a consequence of risk, but retailers worldwide are offering more flexible pricing schemes. The explanations could play a role in convincing retailers to change the pricing scheme.

Alfio Galata
I would have a piece of advice for data collection. Please, train the model on data collected on at least one year and do not take a month there (winter month) or here (summer month) where the energy consumption is very different and influenced by the weather.

Interview with Dr. Stavros Chatzigianni

Interview Metadata

The interview took place over Zoom on November 26, 2021, and have been recorded by Nikos Sakkas.

The **interviewee Dr. Stavros Chatzigiannis**, Manager at Cyric SA (CY), head of new product development in area of building utilities (water, hot water, energy). Relation to APINTech: Co-development of assistive technology for a Swiss contractor (2015-1017).

The **interviewer** is Nikos Sakkas. Eduard Barbu has participated in the interview, clarifying some background information

The interview has been extracted from a larger discussion that have taken place between Mr. Sakkas and Dr. Chatzigiannis and has been edited for brevity, clarity, and intelligibility by Eduard Barbu.

The Interview

Nikos Sakkas

I will start directly: I know you have developed a water management system. We can borrow some ideas for the electricity management system we would like to build. In our system case, we will advise shifting loads on a daily or a three-monthly basis. I remember that in your case, you studied water volume reduction (e.g., when a person was bathing). Did you develop your system based on an ML forecasting model?

Stavros Chatzigianni

We did not implement forecasting, but we estimated algorithmically the volume of water available and how many persons can be served by hot water depending on the boiler's temperature. Our price schema depends on electricity, gas, central heating. I've tested five competitor solutions, but none of them consider the price of electricity or outdoor conditions. Although it does not contain a forecasting system, our system is the only one that lets the user know the volume of water remaining (e.g., a warning can be issued to the effect that there is water left for four persons).

Also, there are technical problems to implement a forecasting model: the system should not be in use. Our system considers the user's flexibility and gives them a piece of advice: at this moment, you have 80 liters of water available. Considering what I have said, what do you think about asking for feedback from the user in the case of electricity consumption?

Nikos Sakkas

I think it is a very good idea, Stavros. I added it to my list.

Stavros Chatzigianni

Working in this sector in the last years, I noticed two types of users. An average user does not need any detailed reading. For a high-level user, you need to install more controllers and supply them with various readings. In any case, if you want to have a reliable forecast, you need a lot of data. Can you make a discretization of energy consumption at the level of appliances?

Nikos Sakkas

In certain cases, we can.

Stavros Chatzigianni
We have worked with AI experts to develop discretization algorithms. If you want, we can arrange a meeting with them. I think you should make the user part of the product development. Of course, to engage the user, you need to give them incentives. The communication with the user is done today through mobile applications. Forget about email or SMS; they are not trendy anymore.

Nikos Sakkas
Do you have any time horizon for providing the disaggregation algorithms?

Stavros Chatzigianni
We plan to launch a product incorporating the disaggregation algorithms in the next six months. Please consider that we have managed to measure a volume of water as low as 0.05l/minute.

Nikos Sakkas
We look forward to your market application.

Interview with Professor Nikos Zarkadis

Interview Metadata

The interview took place over Zoom on November 29, 2021, and have been recorded by Nikos Sakkas.
The **interviewee** is **Nikos Zarkadis**, professor at the university of Geneva (HESGE), expert in the area of building energy. Relation to APINTech: Collaborator for the co-development of behavioral technology. Provider of WT technology at the HESGE campus in Geneva.

The **interviewer** is Nikos Sakkas. The interview has been extracted from a larger discussion that have taken place between Prof. Zarkadis and Nikos Sakkas, and has been edited for brevity, clarity, and intelligibility by Eduard Barbu.

The Interview

Nikos Sakkas

As you have seen from the questionnaires, the main innovation of this project is the forecasting component coupled with an explanatory framework. In this context, the concept of demand-response is crucial for the whole idea of the smart grid and energy production. You know certainly that some users do not understand the concept of flexible price schemes. Therefore, we think that the explanation will help us give good feedback to the user. If there is any comment about the questionnaire or the whole idea you want to make, we are happy to hear it.

Nikos Zarkadis

I think that the concept of explainability and transparency is crucial today when there is much concern about privacy issues. Of course, the explanation should be targeted to the user background, as most users do not understand overly complicated formulas and graphs.

Also, the users will want to collaborate if they know how the data you have collected about them is used. See the recent scandals with Google and Facebook and how it is not clear how these big companies are using the user data. But given all of these, I think it is crucial to communicate the information or explanation to users, given that they do not have technical degrees. The graphs were easy to understand, but I doubt that ordinary people can make sense of them. The user interfaces must be more straightforward.

Nikos Sakkas

You are right. We should think about this.

Nikos Zarkadis

Regarding simplicity, if you want to implement this in a smartphone, a circle with an interactive knob should be the way to go. In general, we need simple things: etiquettes, a number, a knob that turns, and something happens.

Nikos Sakkas

Please, consider that the graphs you have seen were not intended for the consumption of end-users, but were taken from the work of the facility managers.

Nikos Zarkadis

The energy providers need more of these data, of course. They cannot operate without it.

Nikos Sakkas

I will then come back to you when we have a more elaborate design and look for your feedback.

Nikos Zarkadis

I have a comment about the notifications. Because we are flooded with notifications from our preferred applications, I think that the GUI interface should allow the user to set the time when s/he receives the notification. In general, any stakeholders in the project will need their type of information: a more straightforward interface for the end-user, richer controllers for the facility managers. These categories of people have different needs, and the program should cater to their needs. Also, the program should let each class of users choose the granularity and the frequency of the details they need.

Appendix 6. Interviews for Use Case 2 -Retail

Interview Metadata

The interview took place over Microsoft Teams on January 27, 2022 and was recorded by Francisco Amorim. The interviewee is Sónia Germano, Team Leader of E-Commerce Transportation at Sonae MC.

The **interviewers** are members of the TRUST-AI consortium who have participated in the meeting with Sónia. The members who addressed questions have been Francisco Amorim and André Morim. In the interview transcript, they are collectively named "interviewer." The interview was conducted in Portuguese and subsequently transcribed using an automatic speech recognition tool obtain a first draft of the text. Then, the obtained text was translated and edited for brevity, clarity, and intelligibility by Daniela Fernandes.

The **interviewee** Sónia Germano (SG in the transcript) is a Team Lead for E-Commerce Transportation at Sonae MC, arguably the largest e-commerce retailer in Portugal. Sónia has 18 years of experience in Operations and Supply Chain Management functions. Currently, Sónia Germano is responsible for coordinating tactical and operational planning of e-commerce deliveries. The presentation and the interview have been recorded and stored on the INESC TEC drive.

The Interview

Entrevistador – Olhando para este problema do *pricing* das *sots*, existem três grandes níveis de análise. Em primeiro lugar, queremos perceber o comportamento do cliente para lhe oferecer *slots* que sejam apelativas. De seguida, temos a dimensão do custo de transporte e da eficiência operacional, que não pode ser descartada. Por fim, há uma terceira dimensão que combina estas duas e que visa a maximização do lucro ou a minimização das perdas. Estarias mais interessada a olhar para questões sobre o comportamento do cliente, sobre a eficiência operacional ou ambas as coisas? Nota que podes escolher mais do que uma dimensão.

Interviewer - Looking at this slot pricing problem, there are three major levels of analysis. Firstly, we want to understand customer behaviour in order to offer slots that are appealing to clients. Next, we have the dimension of transport cost and operational efficiency, which cannot be ruled out. Finally, there is a third dimension that combines these two and aims to maximize profit or minimize losses. Would you be more interested in looking at questions about customer behaviour, operational efficiency, or both? Note that you can choose more than one dimension.

SG – Do meu ponto de vista, obviamente, tudo que envolva os transportes é mais relevante, pois é o foco da nossa atividade. Contudo, entender o cliente também pode ajudar a perceber o que é que pode levar o cliente optar por uma determinada alternativa. Assim sendo, um bocadinho o conjunto das duas coisas.

SG - From my point of view, obviously, everything that involves transport is more relevant, as it is the focus of our activity. However, understanding the customer can also help to understand what might lead the customer to choose a particular alternative. So, a bit of both combined.

FA – Vou então assumir que gostavas de ter uma explicação nas três vertentes, pois pareceu que a tua resposta foi nesse sentido.

Agora, eu quero colocar uma situação em que nós temos um modelo que é capaz de modular a resposta de um cliente a um conjunto de *slots* com diversos preços. Esse modelo tem uma determinada percentagem de acerto, por exemplo, de 60%. O desafio aqui é perceber quanto dessa percentagem de acerto é que estarias disposta a perder se tivéssemos um modelo alternativo que fosse mais explicativo. Portanto, nós temos um modelo pouco explicável, mas com uma boa performance e um outro modelo que é mais interpretável, mas que tem pior performance. Nesta perspetiva, estarias disposta a abdicar de 5, 10, 15, 20%?

FA – So, I'm going to assume that you would like to have an explanation framed in all three levels, because it seemed to me that your answer suggested so.

Now, I want you to imagine a situation where we have a model that is able to modulate a customer's response to a set of slots with different prices. This model has a certain percentage of success, for example, 60%. The challenge here is figuring out how much of that correct prediction percentage would you be willing to lose if we had an alternative model that was more explanatory. Therefore, we have an unexplainable model, but with a good performance, and another model that is more interpretable, but with a worse performance. In this perspective, would you be willing to give up 5, 10, 15, 20%?

SG - Eu diria que, tendo em conta aquilo que é a nossa forma de trabalhar, que 10 a 15 porcento é razoável.
SG - I would say that, taking into account our work method, 10 to 15 % is reasonable.

FA –Vamos então passar a um conjunto de explicações relativo ao comportamento do cliente. Portanto, imagina que nós estamos a analisar a resposta às variações de preço. Temos um conjunto de variáveis que podem explicar essa resposta, sendo que uma delas é o custo que estamos a impor numa *slot* em específico e nós estamos a analisar porque é que o cliente selecionou ou não selecionou essa mesma *slot*. Agora, temos um conjunto de visualizações ou elementos textuais que, no fundo, mostram-nos o comportamento do modelo. O primeiro elemento visual é um gráfico deste género [3:00]. Aqui tens as variáveis mais importantes, neste caso é o tal custo da *slot* e à medida que a cor das bolinhas vai ficando mais azul nós estamos a baixar o preço. Isto significa que, ao baixar o preço estou a aumentar a probabilidade de seleção do cliente. Nesta outra variável acontece o contrário. Este é o primeiro nível de explicações. Podemos ter um segundo [4:12] relativo à importância das variáveis, ou seja, neste caso atribuo 40% da probabilidade de seleção a esta variável. Temos uma terceira forma [4:25] de apresentar explicações que é através de uma expressão matemática, portanto expressar matematicamente que a probabilidade de seleção é proporcional, neste caso, à seleção que ele teve no passado mais o custo, etc. No fundo ter uma expressão mais ou menos deste género em termos de complexidade. Dentro destas três formas de explicação qual é que te pareceu mais intuitiva de explorar?
FA - Let's then move on to a set of explanations regarding customer behaviour. So, imagine that we are analysing the response to price changes. We have a set of variables that can explain this answer, one of which is the cost that we are imposing on a specific slot and we are analysing why the customer selected or did not select that same slot. Now, we have a set of visualizations or textual elements that, in essence, show us the behaviour of the model. The first visual element is a graphic like this [3:00]. Here you have the most important variables, in this case, it is the cost of the slot and as the colour of the dots gets bluer we are lowering the price. This means that by lowering the price I am increasing the probability of customer selection. In this other variable, the opposite happens. This is the first level of explanations. We can have a second level

[4:12] relative to the importance of the variables, that is, in this case I assign 40% of the selection probability to this variable. We have a third way [4:25] of presenting explanations which is through a mathematical expression, that is, to express mathematically that the probability of selection is proportional, in this case, to the selection the customer has made in the past plus the cost, etc. Basically, having an expression more or less of this kind in terms of complexity. Among these three forms of explanation, which one did you find the most intuitive to explore?

SG – Eu efetivamente sou uma pessoa muito mais visual do que matemática e, portanto, a fórmula matemática era aquela que eu excluiria em primeiro lugar. Eu diria qualquer um dos dois gráficos… até acho que eles são complementares.
SG – I have a much more of a visual perception than a mathematicial one, and therefore the mathematical formula is the one I would rule out first. I would say either of the two graphs… I actually think they are complementary.

FA – Pronto, então vou selecionar estes dois (os dois gráficos). Avançando para uma questão em que estamos a comparar comportamentos de clientes que estão em duas zonas geográficas adjacentes. Vamos tentar perceber porque é que o preço de um cliente é superior ao do outro. Na primeira visualização, temos um conjunto de variáveis que estão a influenciar o preço. Depois, podemos ter uma explicação, através de uma equação, no fundo que nos está a dizer que o preço é proporcional à distância. Temos também duas explicações, que são textuais; a primeira “Demand for one customer is three times higher than for customer with lower price” e a segunda, “ if demand for customer 1 is three times lower, the price would be the same as for customer 2”. No fundo a primeira é de carácter descritivo e a segunda é o que chamamos de counterfactual.
FA – Okay, so I'm going to select these two (the two graphs). Moving on to an issue where we are comparing behaviour from customers who are in two adjacent geographic zones. Let's try to understand why the price presented to one customer is higher than the other. In the first picture, we have a set of variables that are influencing the price. Then, we can have an explanation, through an equation, which tells us that the price is proportional to the distance, etc. We also have two explanations, which are textual; the first reads “Demand for one customer is three times higher than for customer with lower price” and the second, “if demand for customer 1 is three times lower, the price would be the same as for customer 2”. Basically, the first is descriptive and the second is what we call counterfactual.

SG – Como digo, para mim a nível de interpretação de dados, gosto sempre das questões mais visuais acho que são sempre mais percetíveis e fáceis de enquadrar. Obviamente, podem ser sempre complementadas porque o que o gráfico representa é também aquilo que as frases vão transcrever. (Não há…)

SG – As I said, for me in terms of data interpretation, I always prefer visual elements. I think they are always more understandable and easier to frame. Obviously, they can always be complemented because what the picture represents is also what the sentences will transcribe.

FA - Então o que é que sugeres?

FA – So, what do you suggest?

SG – O gráfico em baixo já tem uma frase, não é?

SG – Below the chart there is already a sentence, right?

FA - Não queria para já focar nesta frase, só na parte do gráfico.

FA - I do not want you to focus on this sentence right now, just on the graphic part.

SG – Só queria perceber se o gráfico estava a complementar por esta expressão simbólica que estava a ser referenciada, era só essa a dúvida(???). Porque se fosse, era o complemento perfeito, não é?

SG – I just wanted to understand if the graph was complementing this symbolic expression, that was my only doubt. Because if it were, it would be the perfect complement.

FA – Então vou colocar a feature importance (1ª visualização). Relativamente às frases, colocarias no mesmo patamar de facilidade de interpretação?

FA – So, I am going to select the feature importance option (first option). Regarding the sentences, would you place them on the same level of ease of interpretation?

SG – Prefiro a 1ª opção, mas atenção isto é mesmo algo muito pessoal ao nível de interpretação. Eu diria é que o e-commerce é que deve ter mais interesse em preencher este questionário.

SG - I prefer the first option, but this is really something very personal in terms of interpretation. I would say that e-commerce would have been more interested in filling out this questionnaire.

FA - Até era algo que te ia perguntar no final se podemos ter mais pontos de contacto…
FA – That is something I intended to ask you at the end of the interview, if we can have more points of contact.

SG- Acho que era giro
SG – That would be great

FA – Aqui, nós temos um conjunto de clientes que chegam ao painel de seleção e desistem da compra. E nós vamos tentar perceber o que se está a passar. Que tipo de explicações seriam melhores? Uma lista de frases com as razões que o modelo interpreta com sendo as principais por detrás desse dropout. O segundo, uma visualização do conjunto de regiões e conseguimos comparar um par de regiões em termos desta probabilidade dos clientes que desistem, sendo possível comparar as características que estão a variar entre essas mesmas duas regiões. Ou uma terceira opção, em que podíamos fazer uma simulação e, no fundo, variar os preços e ver no mesmo gráfico como variam as probabilidades de o cliente desistir?
FA – Here, we have a set of customers who arrive at the selection panel and dropout of their purchase. And we're going to try to understand what's going on. What kind of explanations would be better? A list of sentences with the reasons that the model interprets as being the main ones behind this dropout. The second is a visualization of the set of regions which enables us to compare a pair of regions in terms of this probability of customers' dropouts, and where it is also possible to compare the characteristics that differ between these same two regions. Or a third option, where we could run a simulation and, basically, vary the prices and see, on the same graph, how the probabilities of the customers' dropouts change?

SG – Esta última opção parece-me mais importante, só acho que é importante perceber também se o cliente desiste porque já não há disponibilidade para as slots que o cliente pretende. Isto porque, nós estamo-nos a basear muito no facto de o cliente desistir pelo preço, mas ele pode também desistir, e acredito que seja provável de acontecer, porque não tem o horário que pretende independentemente do preço.

Como também já referi, temos uma percentagem relevante de clientes 0, que não pagam independentemente da escolha do slot, e aqui eu diria que tem de haver muita comparação entre a disponibilidade da slot e o preço. Se houver alguma coisa mais dinâmica que possa simular estes cenários é algo que eu diria que faria sentido apresentar.

SG – This last option seems more important to me, I just think it is also important to understand if the customer drops out because there is no longer availability for the slots that he wants. This is because, we are relying a lot on the fact that the reason behind the customer's dropout is the price, but it can also be because, and I believe it is often the case, the customer is not presented with the slots he wants, regardless of the price. As I have already mentioned, we have a significant percentage of "0" customers, who do not pay regardless of their choice of slot, and here I would say that we have to consider both slot availability and price. If it is possible to simulate these scenarios incorporating this dynamic element, then I would say it makes sense to introduce this solution.

FA – (Vou assinalar essa opção e ter esse comentário em consideração) Depois, temos uma questão mais relacionada com análise de sensibilidade de parâmetros. Portanto (imagina) tens uma alavanca que diz se queres premiar mais o cliente ou ter mais eficiência operacional. Para fazer uma análise dos impactos disto, preferias um gráfico que te mostrasse a fronteira, ou seja, um gráfico com 2 eixos, um de satisfação de cliente o outro de eficiência operacional, e um conjunto de pontos, e se variares um determinado parâmetro sabes em que ponto da curva é que estás. Segundo, um dashboard com a visualização do impacto, ou seja, se afetares este parâmetro desta forma o resultado esperado é este. Um gráfico causal, isto é, um conjunto de caixa textuais ligadas e que dizem: este parâmetro tem um impacto nesta variável do comportamento do cliente e por isso causa um impacto x no resultado final. Destas três soluções, existe alguma que …

FA – I will have that into consideration. The following question is related to the sensitivity analysis of parameters. So, consider a lever between rewarding the customer more or having more operational efficiency. To analyse the impacts of this balance, would you prefer a frontier graph, that is, a graph with 2 axes, one for customer satisfaction and the other for operational efficiency, and a set of frontier points, and by varying a certain parameter, you would know in what point of the curve you are at. Second, a dashboard with visualization of the impacts, namely, if the parameters are adjusted in this certain way, the expected result is this. Or also, a causal graph, specifically, a set of connected boxes that say: this parameter has an impact on this variable of customer behavior and therefore, it causes an impact "X" on the final result. Of these solutions, is there any that you prefer?

SG – A dashboard é mais intuitiva até pelo que fazemos à data de hoje. Existem momentos em que temos de tomar decisões neste sentido, pode não ser da forma que aqui foi identificada. Como já partilhei convosco, em alguns momentos e operações em específico temos de ver que a própria baixa de preço acaba por aumentar a eficiência operacional, não é só o efeito contrário. Se eu tiver uma baixa taxa de ocupação, a transportar em vazio, o custo daquelas entregas é também mais elevado.
SG – The dashboard is the most intuitive option, especially considering what we currently do. There are times when we have to make this sort of decisions, although it may not be in the exact way you described. As I have previously shared with you, at certain moments and in specific operations we have to consider that decreasing the price leads to an increased operational efficiency, it is not just the opposite effect. If I have a low capacity utilization, and deliver orders with almost empty vehicles, the cost of the deliveries is also higher.
FA – Ou seja ao influenciar uma coisa não quer dizer que necessariamente estejamos a influenciar a outra no sentido contrário?
FA – You are saying that by influencing one aspect (either customer satisfaction or operational efficiency), the other does not necessarily change in the opposite direction?

SG – Sim. Por isso eu diria que, conforme alterando os parâmetros, eu punha um dashboard com as várias ações que podiam ser…
SG – Yes. For that reason, I suggest a dashboard that, as you change the parameters, shows the various actions that could be taken.

FA – Ok. Estas expressões fazem sentido para ti? Não estas em específico, mas se umas expressões do género fariam sentido para ti. Se gostarias de ter algumas métricas deste género, algo que te direcionasse a tomada de decisão. Aqui, as nossas entregas atempadas (on time deliveries), é um kpi, que é proporcional ao número de veículos e é inversamente proporcional ao preço médio que é cobrado. A pergunta é a seguinte: se quiseres melhorar este indicador o que farias? 1. Aumentar o número de veículos; 2. aumentar o preço médio cobrado para essa região; ou 3. esta expressão não faz sentido e não a quero usar.

FA – All right. Do these expressions make sense to you? Not these specifically in practice, but if an expression like this would make sense to you. Do you want some measurements of this kind, something that drives decision making? Here, On-Time-Deliveries is a KPI, which is proportional to the number of vehicles and inversely proportional to the average price that is charged. My question is: if you want to improve this indicator, what would you do? 1. Increase the number of vehicles; 2. increase the average price charged for that region; or 3. this expression doesn't make sense and I don't want to use it.

SG – Esta expressão é curiosa porque, consoante a área do negócio, defende perspectivas diferentes. Para mim, que estou associada aos transportes e não ao pricing, dir-te-ia que quero aumentar a média do custo de entrega. Eu não quero aumentar o número de veículos porque isso só vai aumentar o meu custo.
SG – This expression is curious because, depending on the business area, it defends different perspectives. For me, because I am in the transport area and not in the pricing, I would say that I want to increase the average cost of delivery. I don't want to increase the number of vehicles because that will only increase "my" cost.

FA – Sim, mas do ponto de vista de aumentar o on time deliveries…
FA – Yes, but if you want to increase On-Time-Deliveries…

SG – Se só nos preocupássemos com o serviço ao cliente, aumentar o número de veículos era a solução.
SG – If our only concerned were customer service, increasing the number of vehicles would be the solution.

FA – Aqui o ponto era perceber se conseguias interpretar esta expressão. (Continuando). Temos um conjunto de operadores, e queria saber, se te dessemos uma expressão, com que operadores te sentirias mais confortável. Temos soma e subtração, multiplicação, divisão, o mínimo e o máximo, exponencial, logarítmico e if then else. Podemos criar expressões matemáticas com todos estes operadores.
FA – The purpose of this question was to understand if you could interpret this expression. We now have a set of operators, and I was wondering, if we have an expression, which operators would you feel most comfortable with? We have addition and subtraction, multiplication, division, the minimum and maximum, exponential,

logarithmic and if-then-else. We can create mathematical expressions with all these operators

SG – Eu diria que o único com o qual não estou tão tao confortável é o logaritmo.
SG – I would say the only thing I'm not so comfortable with is the logarithm.

FA – Temos aqui um exemplo de explicações que são “what if”,ou seja, se eu fizer isto qual será o impacto. Estamos a testar, por exemplo, se abrirmos uma nova janela temporal de manhã e adicionarmos um veiculo à nossa frota, qual vai ser o impacto esperado nos atrasos. Qual das seguintes opções se aplicam neste caso: preferes uma explicação deste tipo e que permita testar diferentes cenários; preferes uma explicação simbólica, uma equação, que permite medir o impacto direto que uma variável tem; ou não precisas de explicações de todo porque consegues perceber autonomamente qual a melhor decisão a tomar?
FA – Here we have an example of explanations that are “what if”, that is, if I do this what will the impact be. We are testing, for example, if we open a new time window in the morning and add a vehicle to our fleet, what is the expected impact on delays. Which of the following applies in this case: do you prefer an explanation of this type and which allows you to test different scenarios; you prefer a symbolic explanation, an equation, which allows you to measure the direct impact that a variable has; or, do you not need explanations at all because you can autonomously understand the best decision to make?

SG – Deve haver sempre sustentação, uma justificação, por mais percetível que seja a decisão. Eu diria que um bocado das duas primeiras. É importante vermos os cenários na perspetiva de explorar. (Por exemplo) abro mais uma slot, e para não atrasar tenho de conjugar nos restantes turnos desta determinada forma para ter atratividade, e também tração na procura e o meu custo do veículo ser eficiente. Por isso, acho sempre importante estas duas conjugações. Hipoteticamente, se puseres toda a capacidade numa slot eu consigo dizer-te que não é viável, mas acho importante conseguirmos sustentar as decisões em factos e dados.
SG – There must always be support, a justification, no matter how obvious the decision may be. I would say a bit of the first two option. It is important to look at scenarios from the perspective of exploring. [For example,] I open a new slot, and in order not to delay deliveries, I have to combine the remaining shifts in this particular way for it to be attractive, to manage demand and to keep my vehicle cost efficient. That is why I think

these two [options] conjugated are important. Hypothetically, if you put all the capacity in a slot I can immediately tell you that it is not viable, but I think it is important that we can support decisions on facts and data.

FA – Temos aqui um trade-off entre satisfação de cliente e custos de transporte e temos também a parte do comportamento do cliente. Vou-te ler um conjunto de frases e vou-te perguntar se gostavas de ter mais insights sobre este tópico ou não. Estás interessada em saber quantos clientes terias a mais se baixasses o preço de uma determinada slot? (I would like to know how many customers will be attracted if I reduce the delivery price on a slot)
FA – Here we have a trade-off between customer satisfaction and transport costs and we also have this topic regarding customer behaviour. I am going to read you a set of sentences and I am going to ask you if you would like to have more insights into this topic or not. Are you interested in knowing how many more customers you would have attracted if you lowered the price of a particular slot?

SG - Eu não mas o e-commerce ia adorar.
SG – I wouldn't, but e-commerce would love it.

FA –Gostarias de perceber qual é o impacto no lucro se conseguisses desviar a seleção do cliente de uma slot para a outra. (I would like to know what is the impact on profit if I nudge a customer from one slot to another)
FA – I would like to know what is the impact on profit if I nudge a customer from one slot to another.

SG – Sim se conseguir tracioná-los para slots mais caras porque assim vou conseguir diluir o meu custo de entrega (ela refere mais caras no sentido de menos eficientes, com menos gente, penso)
SG – Yes, if I can nudge them to more expensive slots, because then I will be able to attenuate my delivery costs.

FA – Gostarias de perceber se poderia ser necessário realocar parte da frota durante a semana para adaptares ao padrão da procura? (I would like to know if my fleet should be relocated during the week to adapt to demand patterns)
FA –I would like to know if my fleet should be reallocated during the week to adapt to demand patterns.

SG – Isso é fundamental na ótica dos transportes

SG – This is fundamental from the point of view of transport.

FA – I would like to know what can be done in terms of slotting and pricing to increase online retail penetration?

FA – I would like to know what can be done in terms of slotting and pricing to increase online retail penetration?

SG – Isto também é muito relevante, porque estamos a trabalhar a nossa taxa de ocupação, que é o nosso primeiro driver de eficiência. Com uma boa taxa de ocupação, conseguimos trabalhar o €/entrega mais competitivo.

SG – This is also very relevant, because we are working on our capacity utilization rate, which is our first efficiency driver. With a good capacity utilization rate, we are able to work on the most competitive €/delivery.

FA- Vou saltar para o comportamento do cliente. Vamos imaginar que temos slots que são mais atrativas para os clientes do que outras e, sobretudo existem slots que estão a ser mais selecionados por clientes de uma região do que por clientes de outra. Portanto são mais atrativas numa região do que noutra. Numa escala de 1 a 5, sendo 1 não interessada e 5 muito interessada, quão interessada estarias em receber insights sobre as razões que estão a levar clientes de uma determinada região a preferir uma determinada slot em detrimento de outra?

FA- I will move on to customer behavior. Let's imagine that we have slots that are more attractive to customers than others and, above all, there are slots that are being selected more by customers in one region than by customers in another region. Therefore, they are more attractive in one region than in another. On a scale of 1 to 5, with 1 being not interested and 5 being very interested, how interested would you be in receiving insights into the reasons that are driving customers in a particular region to prefer a particular slot over another?

SG – Eu reitero que esta questão da slot está muito associada à disponibilidade. Existem zonas geográficas que têm uma disponibilidade de slots curtas mais reduzida, e por isso, essas slots vão ser sempre preenchidas com mais rapidez. E nós temos essa consciência no modelo que temos atualmente. O interesse é elevado, mas vem constatar aquilo que nós já fazemos no planeamento. Nós já sabemos que, em muitas

das vezes, prejudicamos de certa forma os clientes de determinadas áreas geográficas a favor da eficiência na ótica de transportes. Por isso, acho sempre interessante perceber se aquilo que estamos a tomar como decisão é comprovado e se de facto é traduzido em eficiência ou não. Se calhar até estamos a tomar uma decisão em certo planeamento no sentido de reduzir a disponibilidade de slots de curta duração, que o cliente prefere, e isso o leve a desistir da encomenda. Como um trade-off com aquilo que vimos anteriormente... Por isso, acho que é uma informação bastante interessante até para perceber se a estratégia que seguimos ao fazer o planeamento, se comprova.
SG – I reiterate that this slot issue is very much associated with availability. There are geographic areas that have a lower availability of narrow slots, and therefore, these slots will always fill up faster. And we are aware of that in the model we currently follow. The interest is elevated, but these insights will only verify what we are already doing in the planning phase. We already know that, in many cases, we impair customers in certain geographic areas in favour of efficiency in terms of transport. Therefore, I always find it interesting to understand whether the decisions we make are proven [to be good decisions] and whether it actually translates into efficiency or not. Maybe, in some planning moment, we have decided to reduce the availability of narrow slots, which the customer prefers, and this has lead him to drop out. As in the trade-off that we saw earlier... Therefore, I think it is a quite interesting piece of information, even to verify the strategy we have been following.

FA- Dirias então que interessante é um 4 ou um 5?
FA- So would you say that “interesting” is a 4 or a 5?

SG – É um 4. É só para comprovar.
SG – It is a 4. It is just to verify.

FA – Nós tínhamos esta ferramenta que nos permitia analisar o comportamento do cliente e as suas preferências por determinadas slots, e agora nos temos várias formas de montar essa ferramenta. Uma delas seria conhecer quais as variáveis que estariam a impactar mais a escolha do cliente, por exemplo a idade média, o rendimento médio... Depois outra forma de o fazer, seria através de uma expressão matemática que nos daria um score atribuído à combinação da região-slot de forma a percebermos a sua atratividade. E um terceiro ponto, seria fazermos um drill down para compreender que características é que têm os clientes de uma determinada região. Relativamente aos clientes que habitualmente selecionam uma slot e que estão

inseridos numa determinada região, que características têm eles, e permitir a comparação dessas mesmas características com as de outra combinação slot região.
FA – Hypothetically, we have this tool that allowed us to analyse customer behaviour and preferences for certain slots, and now we have several ways to build this tool. One of them would be to know which variables would have a greater impact on the customer's choice, for example average age, average income... Then another way of doing it, would be through a mathematical expression that would give us a score assigned to the combination of the slot-region in order to understand its attractiveness. And a third option, would be to drill down in order to understand what characteristics customers in a given region have and allow the comparison of these same characteristics with those of another slot-region combination.

SG – Tentando pôr-me no papel negócio e transportes, eu diria que, para o transporte a segunda opção é bastante interessante porque ajuda-nos a trabalhar no contexto de região e a trabalhar na ótica dos transportes. A terceira, diria muito na ótica de negócio.
SG – I will try to speak from the perspective of both business and transport. I would say that, for the transport section, the second option is quite interesting because it allows us to work in the context of a region. The third, I would say is very interesting from a business perspective.

FA – Relativamente ao questionário estamos terminados. Podemos só ter mais uns minutos para vermos umas questões mais abertas?
FA – Regarding the questionnaire, we are finished. Can we just have a few more minutes to approach some open questions?

SG – Sim, sem dúvida.
SG – Yes, without a doubt.

FA - Queria começar por saber, apesar de não estares inserida na área de negócio, qual é a tua perspetiva sobre sermos mais transparentes na comunicação do preço de entrega ao cliente. Ou seja, ao cliente estão a mostrar um preço bastante alto em relação à média que ele tem pago nas últimas semanas e, teríamos uma indicação da razão pela qual nós estamos a cobrar esse valor adicional; pode prender-se, por exemplo, com uma procura excessiva pela slot, ou porque a procura dessa slot está concentrada numa região geográfica distante do cliente.

FA - I would like to start by asking you, despite not being in the business area, what is your perspective on being more transparent in communicating the delivery fee to the customer. In other words, the customer is being presented with a very high price compared to the average he has paid in recent weeks and we would have an indication of why we are charging this additional amount; it may be related, for example, to excessive demand for the intended slot, or because the demand for that slot is concentrated in a geographic region far from the customer.

SG – Não sendo a minha área core mas pensando muito numa ótica de cliente, nós em Portugal vivemos num contexto no qual o cliente acredita que a entrega deve ser sempre o mais baixa possível, ou seja, o cliente não tem a visibilidade efetiva de aquilo que é o custo de entrega. Ele já hoje não tem a visão de que aquilo que está a pagar é manifestamente um valor mais baixo do que o custo real da entrega. Portanto, eu não sei se para o cliente essa perceção é evidente ou seria benéfica. Tenho a expectativa de que o modelo nos ajude a perceber ou a direcionar o cliente para a escolha de uma outra slot, porque a procura para a slot preferencial já tende para a capacidade. Aí, acho que é importante o cliente perceber o porquê de eu lhe estar a oferecer aquele preço mais competitivo - porque tenho mais oferta para essa slot e mais procura para outras, nesse momento.

SG – It is not my core area but thinking from a customer perspective, in Portugal, we live in a context in which the customer believes that delivery should always be as low as possible. That is to say, the customer does not really have a notion of what the cost of delivery is, and he does not know that what he is paying for is a significantly lower value than the actual cost of delivery. Therefore, I don't know if this perception is evident to the client or if it would be beneficial. What I expect from the model is that it will help us understand or nudge the customer to choose another slot, because the demand for the preferred slot already tends towards [maximum] capacity. So, I think it's important for the customer to understand why I am offering that slot at a more competitive price - because I have more supply for that slot and more demand for others at that moment.

FA – Então no fundo ser um encorajamento mais positivo, mais do que um negativo. Os preços mais atrativos serem assim por este determinado conjunto de razões?

FA – So basically, you intended to have more of a positive encouragement than a negative one. And you would like to justify the most attractive prices with a particular set of reasons?

SG – Sim

SG - Yes

FA – Da tua experiência, qual é a alavanca que permite direcionar mais o comportamento dos clientes. É a oferta de slots que são disponibilizados, a largura, o preço, existe algum outro fator?

FA – From your experience, what is the lever that allows you to better redirect customer behavior? Is it the availability of slots, the time length, the price, is there any other factor?

SG – Eu diria que é um misto. Temos alguma predominância de entregas 0, em que o valor da slot não é tido em consideração na escolha. Eu diria que é a disponibilidade dos slots. Isto é muito notório em momentos que estamos com taxas de procura mais elevadas. As slots mais atrativas, que são aquelas com 2 horas de duração, são as que mais rapidamente são preenchidas. Adicionalmente, as slots da noite enchem primeiro, pois é um horário que o cliente valoriza porque por defeito está em casa. Portanto, diria que quando olhamos para momentos de procura elevada, em que as pessoas querem é encomenda para esse dia, o fator preço é pouco valorizado. O que querem é fazer a encomenda. Quando estamos num contexto em que há disponibilidade, aí sim, querem é conseguirem colocar a encomenda na slot pretendida.

SG – I would say it is a mix. We have some predominance of “0” deliveries, in which the value of the slot is not considered when choosing. I would say it is the availability of slots. This is very noticeable at times when we have higher demand rates. The most attractive slots, which are 2 hours long, are the ones that fill up the fastest. Additionally, the night slots fill up first, as it is at a time of the day that the customer values because by default he is at home. Therefore, I would say that when we look at days of high demand, when people want to order for that day, the price factor is not significant. What they want is to place their order. When we are in a context where there is more availability, then yes, they want to be able to place the order in the desired slot.

FA – Voltando então para a questão mais operacional da gestão das entregas. Vou te mostrar um mock-up de um dashboard e a ideia é, definirmos 2 cenários alternativos sobre um conjunto de parâmetros. Estes parâmetros poderiam ser relativos ao serviço ao cliente - estabelecer limites mínimos e máximos do preço que quero mostrar ao

cliente - como mais operacionais - definir o número de veículos disponíveis que tenho para fazer as entregas. Parametrizamos 2 cenários alternativos e depois vamos perceber que impactos resultam desses 2 cenários ao nível dos KPIs que estamos a analisar. Podemos estar a olhar para o lucro, o número de horas de carga no entreposto, o tempo médio entre encomendas de clientes, o número de clientes que estou a servir, quantos clientes estão a desistir da compra... E consigo ver isto por geografia; consigo ver onde tenho mais clientes a desistir. Tens algum comentário relativamente a uma dashboard deste género? Se seria algo útil para gerir a informação?
FA – Let us return to the more operational view of transport management. I will show you a mock-up of a dashboard and the idea is the following, we define two alternative scenarios on a set of parameters. These parameters could be related to customer service - establish minimum and maximum price limits that I want to show the customer – or more operational - define the number of available vehicles that I have to make deliveries. We have parameterized two alternative scenarios and then we will understand what impacts result from these two scenarios in terms of the KPIs we are analysing. We could be looking at profit, the number of loading hours in the warehouse, the average time between customer orders, the number of customers I'm serving, how many customers are dropping out... And I can see this by geography; I can see where I have more customers dropping out. Do you have any comments regarding a dashboard of this kind? Would it be something useful to manage the information?

SG – Provavelmente adicionaria aqui outro tipo de métricas, mas sim esta visão onde podemos mexer em parâmetros e ver o impacto em KPIs acho muito funcional.
SG – I would probably add another type of KPIs here, but this way of visualizing information where we can change parameters and see the impact on KPIs, I think it is very functional.

FA – Que tipo de parâmetros achas que têm mais influência e que deverias ter mais controlo?
FA – What kind of parameters do you think have more influence and that you should have more control over?

SG – É muito importante, para nós, perceber de que forma a baixa dos preços e a amplitude das slots se traduz em veículos. Depois, tenho sempre de fazer esse movimento no meu dia como um todo; trabalho por turnos, tenho de garantir que

mexer nas slots é algo transversal, que mexo em todos os turnos. Não me posso concentrar só nas slots que têm até mais atratividade, porque isso pode desequilibrar a minha viatura, ou melhor, a minha taxa de ocupação da viatura. Por isso, eu entraria em detalhe ao nível da disponibilidade pois quando falamos no preço máximo e mínimo da slot, temos de casar isto com o número real de encomendas associado. E se esse número real estiver devidamente nivelado, quantos veículos preciso para fazer realizar esse número real de encomendas? Cada encomenda é um cliente, podemos sempre ver dessa forma. E depois, existe obviamente essa segunda derivada que observamos quando olhamos para o impacto no benefício, ou seja, o que ganhamos com o custo das slots versus o investimento que estamos a fazer. Acho que faz todo o sentido. O tempo de carga não sei se seria relevante, talvez onde é que estamos a centralizar a nossa …a nossa…ou seja, quais as slots que….
SG – It is very important for us to understand how decreasing prices and the length of slots translates into vehicles. Then, I always have to have this in mind during my whole day; I work in shifts, I have to make sure that tinkering with slots is something transversal, that I do on every shift. I cannot just focus on the slots that are more attractive, because that can unbalance my vehicle, or rather, my vehicle utilization rate. Therefore, I would go into detail about availability because when we talk about the maximum and minimum price of the slot, we have to match this with the actual resulting number of orders. And if that actual number is properly levelled, how many vehicles do I need to carry out that actual number of orders? Every order is a customer, we can always see it that way. And then, there is obviously this second derivative, when we look at the impact on the benefit, namely, what we gain from the cost of the slots versus the investment that we are making. I think it makes perfect sense. I don't know if the loading time would be relevant, maybe where we are centralizing ours…that is, which slots….

FA – Se existe muita disparidade entre as slots?
FA – If there is a lot of disparity between the slots?

SG – É por aí. Relativamente à zona, é interessante se tivermos de mudar a tal disponibilidade. Por vezes, temos estado a apostar num determinado código postal ou a apostar numa determinada zona geográfica e a oferecer-lhes mais serviço e não temos benefício nenhum com isso e aqui, podemos fazer essa aprendizagem.
SG - That's it. Regarding the zone, it is interesting if we have to change the availability. Sometimes, we have been investing on a certain postcode or investing on a certain

geographic area by offering them better service and we have not collected benefit from that and here, we can understand that.

FA – À data não têm uma ferramenta que vos auxilie na gestão?
FA – Don't you currently have a tool to help you manage the operation?

SG – O que nós temos hoje é…Atenção que eu falo na ótica operacional e de transportes. Eu sugiro ativar alguma medida, seja da oferta da taxa de entrega, seja de redução de taxa de entrega. Uso isso como argumento quando sinto que a procura que temos não está alinhada com o nosso forecast e que estamos a perder eficiência. Peço ajuda ao negócio para responder com esse tipo de ações, mas não são ações que estejam sobre a minha alçada. Como tal, aquilo que acompanho é a taxa de ocupação de cada slot por zona geográfica. Isso é que é a base da minha eficiência operacional. Como vos dizia, se tenho um determinada slot, se tenho um dia em que a taxa de ocupação ou se tendencialmente já vamos em alguns dias com uma taxa de ocupação abaixo daquela para a qual nos dimensionamos, e tendo em conta que trabalhamos com a nossa frota que é fixa, lançamos o desafio ao negócio de mexer no pricing. Outra coisa, e que de certa forma esta associada ao pricing, mas não estamos com essa preocupação… nós sabemos que temos slots mais caras e mais baratas, e as mais caras são as de 2 horas. E o cliente valoriza essas slots de 2 h. Se eu tiver mais tração nas slots de 2 h e tiver ainda muita disponibilidade em slots de 4 h, ajustamos consecutivamente de forma equilibrada e transitamos capacidade entre slots para conseguirmos melhorar a nossa taxa de ocupação. Acompanhamos muito mais a taxa de ocupação da slot e do dia, não estamos tao ligados a esta parte do pricing. O pricing, pedimos muito, quando vemos que em traços gerais a ocupação quebra. Aí pedimos para ajustarem os preços.
SG – What we currently have is… Please note that I speak from the operational and transport point of view. I suggest activating some action, either offering free delivery fee or a delivery fee reduction. I use this as an argument when I feel that the demand we have is not in line with our forecast and that we are losing efficiency. I ask the business sector for help, to respond with these types of actions, but these are not actions that are under my responsibility. As such, what I keep track is the capacity utilization rate of each slot by geographic zone. This is the basis of my operational efficiency. As I told you, if I have a certain slot, a certain day or even some days with a capacity utilization rate tendentiously below the value for which we are dimensioned, and considering that we work with our fleet which is fixed, we launch the challenge to the business to

change the pricing. Another thing, which is somehow associated with pricing, but we are not concerned about that… we know that we have more expensive and cheaper slots, and the most expensive are the 2-hour slots. And the customer values these 2-hour slots. If I there is more demand for the 2-hour slots and I still have a lot of availability in the 4-hour slots, we consecutively adjust and transfer capacity between slots in order to improve our capacity utilization rate. We monitor the capacity utilization rate at a slot and day level, we are not so connected to this part of pricing. We normally ask for pricing decisions, when we see that, in general terms, capacity utilization is very low. Tin those situations, we asked them to adjust the prices.

FA – Esta decisão conjunta de alterar o pricing para ajustar a taxa de ocupação é algo que só acontece em casos extremos?
FA – Is this joint decision to change the pricing to adjust the capacity utilization rate something that only happens in extreme cases?

SG – Sim. Diria que é uma decisão que, tendencialmente, é a oferta da taxa de entrega no seu todo. É a ação que tomamos recorrentemente em situações em que estamos com pouca atratividade na procura. Oferecemos ao cliente a taxa de entrega ou restituímos o valor em cartão, é algo que usamos muito. Em situações muito específicas, como é o caso da operação Algarve que está muito associada ao comportamento do cliente, o cliente desloca-se muito ao sábado. E, ao sábado de manhã normalmente temos uma taxa de ocupação mais baixa e o cliente só procura a tarde e a noite. Mas nós não podemos, de forma a ter uma frota eficiente, meter toda a capacidade neste turno e por isso mexemos no pricing das slots da manhã. Mas são coisas muito especificas. Atenção, que isto é aquilo que é a minha perceção na ótica operacional. De certeza que os colegas de negócio podem fazer uma leitura e interpretação, e mesmo algumas ações, das quais eu não tenho visibilidade e que possam ter impacto. E como cliente com entrega 0, também não consigo perceber esta dinâmica em termos de pricing no website.
SG – Yes. I would say that it is a decision that tends to be the offering of the delivery fee as a whole. It is the action that we frequently recur to in situations where we are running low on demand. We offer the customer the delivery fee or refund the amount in card, it is also something we do a lot. In very specific situations, such as the Algarve operation, which is closely associated with customer behaviour, the customer travels a lot on Saturdays. And on Saturday morning we usually have a lower capacity utilization rate and the customer only wants the afternoon and evening slots. But we cannot, in

order to have an efficient fleet, put all the capacity in this shift and that is why we change the pricing of the morning slots. But these are very specific situations.

Attention, this is my perception from an operational point of view. Of course, business colleagues can read and interpret, and even take some actions, of which I have no knowledge and that could have an impact. And as a customer with “0” delivery, I cannot understand this dynamic in terms of pricing on the website either.

FA – Penso que podemos terminar por aqui a entrevista. André, não sei se tens alguma questão a acrescentar…

FA – I think we can end the interview here. André, I don't know if you have any questions to add...

AM – Sim, eu gostei da parte em que estávamos a falar do dashboard, em que a Sónia tinha dito que no fundo queria perceber consoante os preços e o serviço que queríamos apresentar ao cliente, em que necessidade de veículos é que isso se traduzia. Ou seja, tentando resolver um problema destes, gostaria que ele fosse resolvido com uma frota fixa de x veículos e iriamos tentar otimizar a partir daí ou gostaria que fosse um grau de liberdade, e calcularmos também qual o número de veículos que seria efetivamente usado.

AM – Yes, I would like to refer to the moment where we were talking about the dashboard, when Sónia said that she would like to understand, depending on the prices and the service we want to present to the customer, in what need for vehicles this translates to. That is, when trying to solve this problem, would you like it to be solved with a fixed fleet of X vehicles and we would try to optimize from there or would you like it to be a degree of freedom, and we would also calculate the number of vehicles that would be effectively used?

SG – Exato, esta segunda hipótese é muito interessante. Até para percebermos se temos elasticidade para isso ou não.

Agora o que eu acho é que, este tema que vocês estão a trabalhar, é muito relevante para o e-commerce. Acho mesmo, e seria muito interessante entrarem em contacto. Se quiserem posso falar com eles e ver quem é que do lado deles vos podia ajudar aqui. Eles têm agora uma pessoa que tem estado a trabalhar muito neste conceito de expansão, do pricing das slots e da disponibilidade. Acho que essa pessoa poderia ajudar muito naquilo que vocês estão a estudar. E se quiserem ou acharem relevante,

posso falar com essa pessoa, e tenho a certeza de que ele teria todo o gosto em ajudar-vos. Até talvez com opiniões que são mais válidas do que as minhas.
SG – Exactly, that second option is very interesting. Even to see if we have elasticity for that or not.
Now what I think is that this topic that you are working on is very relevant to e-commerce. I really think so, and it would be very interesting for you to get in touch. If you want I can talk to them and see who on their side could help you here. They now have a person who has been working hard on this expansion concept, slot pricing and availability. I think that person could help a lot in what you are studying. And if you want or find it relevant, I can talk to that person, and I am sure he would be happy to help you. Maybe even with opinions that are more valid than mine.

FA – Sim, e temos de olhar um bocadinho mais para a perspetiva da satisfação do cliente. Para já estamos só a olhar para eficiência operacional. Acho que era muito relevante, se nos conseguisses pôr em contacto.
FA – Yes, and we have to look a little more from the perspective of customer satisfaction. For now, we are just looking at operational efficiency. I think it would be very relevant, if you could get us in touch.

SG – Claro que sim. Vou falar com ele. Depois partilho aqui convosco os contactos. Acho que seria a conjugação perfeita. Olhando até para este questionário que realizaram, acho que eles vão ter muito interesse naquilo que vocês estão a estudar. Mesmo muito. Acho muito relevante para eles. Portanto se concordarem com o que estou a dizer, vou falar com eles.
SG – Of course. I will talk to him. Then, I will share the contacts with you. I think it would be the perfect combination. Even looking at this questionnaire that you did, I think they will be very interested in what you are studying. I think it is very relevant to them. So, if you agree with what I am saying, I will talk to them.

FA- Sim, por favor. Do nosso lado temos todo o interesse.
FA- Yes, please. We are very interested.