

Are you sure? Prediction revision in automated decision-making

Nadia Burkart¹  | Sebastian Robert¹ | Marco F. Huber^{2,3}

¹Interactive Analysis and Diagnosis, Fraunhofer IOSB, Karlsruhe, Germany

²Institute of Industrial Manufacturing and Management IFF, University of Stuttgart, Stuttgart, Germany

³Center for Cyber Cognitive Intelligence (CCI), Fraunhofer IPA, Stuttgart, Germany

Correspondence

Nadia Burkart, Interactive Analysis and Diagnosis, Fraunhofer IOSB, Karlsruhe, Germany.

Email: nadia.burkart@iosb.fraunhofer.de

Abstract

With the rapid improvements in machine learning and deep learning, decisions made by automated decision support systems (DSS) will increase. Besides the accuracy of predictions, their explainability becomes more important. The algorithms can construct complex mathematical prediction models. This causes insecurity to the predictions. The insecurity rises the need for equipping the algorithms with explanations. To examine how users trust automated DSS, an experiment was conducted. Our research aim is to examine how participants supported by an DSS revise their initial prediction by four varying approaches (treatments) in a between-subject design study. The four treatments differ in the degree of explainability to understand the predictions of the system. First we used an interpretable regression model, second a Random Forest (considered to be a black box [BB]), third the BB with a local explanation and last the BB with a global explanation. We noticed that all participants improved their predictions after receiving an advice whether it was a complete BB or an BB with an explanation. The major finding was that interpretable models were not incorporated more in the decision process than BB models or BB models with explanations.

KEYWORDS

experiment, explainable ML, interpretability, prediction revision

1 | INTRODUCTION

Many automated decision-making systems are used in practice with little human interference. With the advances in machine learning (ML), the proportion of automated decision-making will further increase. The recent progress made in ML allows machines to take over various tasks that were previously performed solely by human experts. Especially the domain of automated decision making benefits from these advances. New algorithms and approaches such as deep learning, for example, deep artificial neural networks were discovered that are characterized by increased accuracy. Furthermore, the number of domains in which the algorithms are applicable grows steadily. New use cases are constantly developed that change the requirements to be met by the algorithms. In addition to the accuracy of the predictions, their explainability becomes essential. The algorithms are based on mathematical models that can become highly complex. This causes insecurity to the predictions made by these algorithms. The insecurity gives rise to the need for equipping the algorithms with explanations.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2020 The Authors. *Expert Systems* published by John Wiley & Sons Ltd.

In contrast to machines making decisions, humans can be asked to explain the decisions they make. Humans can answer the question by arguing. They can give reasons for their decisions. In human decision-making, this is the major advantage over artificial systems. However, systems outperform humans when it comes to the amount of data they can process in a short time period.

For instance, FICO¹ developed the FICO Score to improve strategies and processes in the lending business of financial service providers in the United States. Sesame Credit2 is a private credit scoring system, an affiliate of the Chinese Alibaba Group. The underlying algorithms use thousands of user data points that include what the users' purchase, their social network affiliations, or their ability to honour an agreement to determine a particular credit score between 350 (bad) and 900 (excellent). While it might seem obvious to claim for greater transparency, with the use of ML and large data sets it is extremely difficult to locate a potential bias.

Systems, for example, based on ML algorithms can make more consistent decisions since their only source for making decisions is the given data (Davenport & Harris, 2005). Nevertheless, if the data contains opinions that include bias in form of discrimination, the machines learn this bias as well. Due to aforementioned advantages of ML over humans in automated decision making, approaching the challenge of explainability of ML algorithms is of great interest. Accordingly in the context of ML, interpretability or explainability is the ability to explain or provide meaning in understandable ways to a human for a certain prediction or a model (Doshi-Velez & Kim, 2017).

This work examines how prediction revision, for example, adjusting an initial prediction on the basis of new information, is affected by explainability. In our user experiment we chose a task where participants need to predict the students' grades (regression problem). This type of task has been chosen because it can be easily transferred to other domains like, for example, credit scoring. After the participants have submitted their first prediction of the grade, they get support to predict the students' grade by means of an automated decision support systems (DSS). Our research question is to examine how users adapt their initial predictions based on the type of support. The four treatments differ in the degree of explainability to understand the prediction of the system. Section 2 gives an overview of the related work in explainable ML and human subject studies in the field. Section 3 describes the methodology of the experiment and Section 4 illustrates the results. Section 5 discusses the results and concludes.

2 | THEORETICAL BACKGROUND

In this section, we first give a short overview of what explainable ML means and describe the main concepts that are approached by our experiment. Furthermore, we want to overview relevant human subject studies in the field of prediction revision. Our experiment is new, because it addresses the problem of prediction revision in the context of explainable ML. To our knowledge this is the first kind of work that assesses this topic in detail with different degrees of explainability.

2.1 | Explainable machine learning

Supervised ML approaches can produce interpretable models (white boxes) or complex opaque models (black boxes). For example, sparse linear models are considered more interpretable than dense ones Tibshirani (1996). Explainability offers many approaches that addresses the problem of explanation generation. The survey paper of Guidotti et al. (2018) gives a detailed overview. We further differentiate only between *Interpretable By Nature* and *Surrogate models* because those are the used concepts within the experiment.

2.1.1 | Interpretable by nature

Prediction models being interpretable by nature can be understood in themselves and how they work can be grasped by humans. In Henelius, Puolamäki, Boström, Asker, and Papapetrou (2014), small decision trees and decision lists are considered to belong this group of models. Singh and Guestrin (2016) consider decision trees, rules, additive models, attention based networks and sparse linear models as interpretable models. However, with increasing complexity, for example, measured in size or number of nodes of the tree, or number of rules or conditions in a rule, these models loose interpretability. For instance, we train a sparse linear regression model directly on the data set. The weights $a = (a_1, \dots, a_n) \in \mathbb{R}^n$ of a linear regression model

$$y = a^T x + b \quad (1)$$

can be interpreted as the importance of a feature, for a feature vector $x \in \mathbb{R}^n$. The bigger the absolute value of a component a_i , the more does a change in the input x_i affect the output. For example does the so-called partial dependency plot show the effect of a single (or two) features on the output, when marginalized over the other features. Through this plot, a user can get an idea of how this feature interacts with the output over

the whole range of possible values. Model internals can sometimes also be used to convey insights about the model. Other examples could be for instance the decision nodes in a small decision tree.

2.1.2 | Black boxes and surrogate models

A random forest is an ensemble technique that can perform regression and classification tasks using multiple decision trees and bootstrap aggregation (bagging). Bagging involves training each decision tree on a different data sample. The idea behind is not to rely on individual decision trees but to combine multiple decision trees to determine the final output with higher accuracy. The model and the decisions cannot be understood by themselves. Therefore, surrogate models can be applied. A surrogate model translates the underlying prediction model into an approximate model (Henelius et al., 2014). This technique is applied whenever the prediction model is not interpretable by itself. An interpretable surrogate model, for example, a linear regression model, is built to complement the black box. Separating the black-box prediction model from its explanation introduces flexibility, accuracy and usability (Singh & Guestrin, 2016).

A wide range of explanation techniques exists that extract rules from black boxes or build decision trees on top of the black boxes. Such techniques use an interpretable model and apply it to the predictions returned by the black box (Hall, Phan, & Ambati, 2017). Currently many different explanation approaches exist, for example, providing logical statements (Lakkaraju, Bach, & Leskovec, 2016; Su, Wei, Varshney, & Malioutov, 2015; Wang et al., 2015; Wang & Rudin, 2014), local models (Rüping, 2005; Turner, 2016) or feature importance (Adler et al., 2018; Datta, Sen, & Zick, 2016; Goldstein, Kapelner, Bleich, & Pitkin, 2015; Puolamäki & Ukkonen, 2017).

We are focusing on LIME that is a local explanation approach and sp-LIME, as a global explanation. Local explanations are concerned about an individual's decision (Phillips, Chang, & Friedler, 2017) and provide the reason behind a specific decision (Doshi-Velez & Kim, 2017). This is helpful for justifying a specific decision. According to (Robnik-Šikonja & Kononenko, 2008) the local scope of explanations can also be called instance explanations. Global explanations are concerned about the system's actions overall (Phillips et al., 2017) and provide a pattern that the prediction model discovered in general. The system can convey the behaviour of a classifier as a whole without regarding predictions of individual instances (Lakkaraju, Kamar, Caruana, & Leskovec, 2017).

Local model agnostic explanations (LIME): From the original instance $x \in \mathbb{R}^d$ we initially have to create an *interpretable* representation of a lower dimensional binary form, that is, $x' \in \{0, 1\}^d$ (x' needs to be specified by the user). This is an essential step since one has to decide which representation x' to use in order to make it interpretable or understandable to the targeted user and rich enough to still allow meaningful deduction. The goal is then to train a fairly simple model $g \in G$ of a potentially interpretable class of models G that acts on the interpretable representation x' and is trying to resemble the decision of the black box model $f: \mathbb{R}^d \rightarrow \mathbb{R}$ around the sample x . To optimize the interpretability and fidelity tradeoff the model g is acquired through:

$$\xi(x) = \arg \min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \quad (2)$$

where π_x is a proximity measure to define locality around x and Ω is a measure of complexity of a model g .

Submodular-pick LIME (sp-LIME): Out of a set of instances X with their explanations, choose a subset of reasonable size ($A \subseteq X$, $|A| < B$, where B is the Budget). Intuitively, we want to choose instances with features that are also globally important, as in they explain many different instances. If we achieve an optimal pick the user can trust the model to perform well and faithful in most cases. In short, we can summarize sp-LIME as follows: First, we select B instances for the user to be inspected. We then aim to obtain non-redundant explanations that represent how the model behaves globally. Given a so-called explanation matrix of n explanations using d features, the features are ranked such that the feature which explains more instances gets a higher score. When selecting instances, the algorithm avoids instances with similar explanation and tries to increase the coverage.

In order to evaluate the performance of a regression model either white box or black box the metrics mean absolute error and root mean squared error are important. The mean absolute error (MAE) measures the average scale of the errors in the test set of predictions. It is the average over the test set of the absolute differences between prediction and the true value. The lower the mean absolute error, the better the model accuracy. The root mean squared error is the square-root of the average of squared differences between prediction and actual observation.

2.2 | Human subject studies in the context of prediction revision and explainable machine learning

In many organizations, decision makers rely more and more heavily on information provided by DSSs. In these cases, the usefulness of the information a decision maker receives from the system will not only depend on the systems features to analyse and communicate the relevant information needed by the decision maker, but also on the decision maker's trust in the information acquired (Miller, 2017). From a rational perspective, processing information provided by DSSs can be characterized by a Bayesian belief revision process, that is, using the decision

maker's prior prediction about relevant determinants of the decision to be made and about the relevance and reliability of the information from the DSS (Fischhoff & Beyth-Marom, 1983). However, prior research has demonstrated that a decision maker's actual use of information may be biased in various ways (Ashton & Ashton, 1988). A many times proven to be robust bias is egocentric advice discounting (Yaniv & Kleinberger, 2000). Although decision makers benefit on average from the advice, they do not benefit fully due to the sub-optimally low level of adjustment to the information. Even the reference to an expected higher quality of decision making with greater consideration of the information leads to the fact that the decision maker does not approach the advice by more than 20–30% on average. (Goodwin & Fildes, 1999; Lim & O'Connor, 1995). Further empirical studies have shown that inadequate adjustment to the information goes hand in hand with reduced decision quality. For example, Diamantopoulos and Mathews show in (Diamantopoulos & Mathews, 1989) that the strength of the subjective adjustment of the sales forecasts they examined is positively correlated with the forecast quality. Fildes et al. also report in a field study that strong adjustments to demand forecasts significantly improve the quality of decisions more than weak adjustments (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009). Various reasons are discussed in the literature for the inadequate processing of the information. On the one hand, easier access to one's own arguments is discussed as a reason (Yaniv, 2004): decision makers can understand the basis on which they made their own decision. However, they lack this basis in the decision of a consultant. The over-weighting of their own arguments in comparison to those of others is referred to as egocentric bias (Kruger, 1999). On the other hand, the average insufficient consideration of advice can be explained on the basis of the anchor heuristic (Tversky & Kahneman, 1974). Following this explanation, decision makers anchor on their own estimates and thus take insufficient account of the advice given by the DSS in the decision-making process. The assumed anchor can result directly from the problem or be the result of an incomplete calculation. To get closer to the desired value, individuals check the respective values successively in an iterative process starting from the anchor until a plausible result is reached (Quattrone, 1982).

Poursabzi-Sangdeh, Goldstein, Hofman, Vaughan, and Wallach (2017) conducted a user study to measure trust. In their view, confidence can be measured by determining the difference between the prediction of the prediction model and the prediction of the participant. In their investigation they forecast real estate prices. The task of the prediction model is therefore a regression approach. Participants receive different information about the underlying forecast model and are asked to make their own forecast of the property price. This estimate is compared with the house price predicted by the forecast model. Their absolute deviation is a measure of confidence, while smaller values meaning higher confidence. Lage et al. (2019) conducted a user study to find out what makes explanations interpretable for humans by systematic variation of the properties of an explanation to measure their effect on the performance of several tasks. The tasks were to simulate the system's response, to verify a suggested response, and counterfactual reasoning. One of their findings included that counterfactual questions had significant lower accuracies across the experiments. Schmidt and Biessmann (2019) introduce a quantitative measure of confidence in ML decisions and conducted an experiment. In their experiment they examined two methods, COVAR, a glass-box method and LIME. They noticed that COVAR yielded more interpretable explanations. Thereby, they highlighted the usefulness of simple methods. Yin et al. (Yin, Wortman Vaughan, & Wallach, 2019) conducted a human-subject experiment to examine laypeople's trust in a model.

3 | METHODOLOGY

Our experiment was designed to measure how participants adapt their initial prediction if they will get support from an automated DSS. We designed four kind of treatments. The four treatments (T1, T2, T3, T4) offer participants different approaches of an explanation, for example, based on an interpretable "white box" model or "black box" models equipped with or without explanations (local or global). The goal is to gain insights about which type of explanation participants trust and favour the most.

3.1 | Experimental design

The advice was either based on a linear regression model (T1) (see Figure 1) as white box, a random forest model (T2) as entire black box (see Figure 2), a random forest model with local explanation (LIME) (T3) (see Figure 3) or a random forest model with global explanation (sp-LIME) (T4) (see Figure 4). Participants in each treatment were told basic facts about their corresponding model and how each is used for producing a response (e.g., predicting a student's grade) when presented with a set of predictor values (e.g., academic and personal characteristics). Please see Appendix A (Figure A1 - Figure A13) for a detailed overview of the experimental setup in the online experiment system.

Moreover, in treatment 2–4 (T2–T4) participants were explained that random forests are sometimes treated as a black box, meaning their prediction techniques are opaque and we cannot say with certainty how the prediction was derived from the model.

In the explanation treatments (T3, T4), participants were introduced to the basics of explainable models (surrogates) and how they could help in the decision making process. Together with an example, they were shown that an interpretable model is used to explain individual predictions of random forests (either globally or locally) and that they highlight those academic and personal characteristics that were most important for the specific prediction.

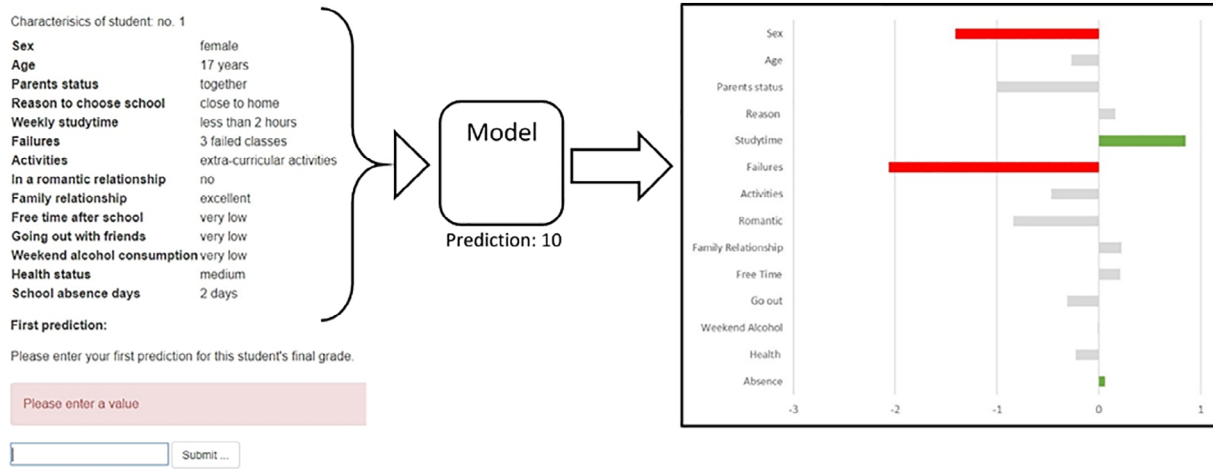
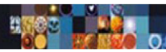
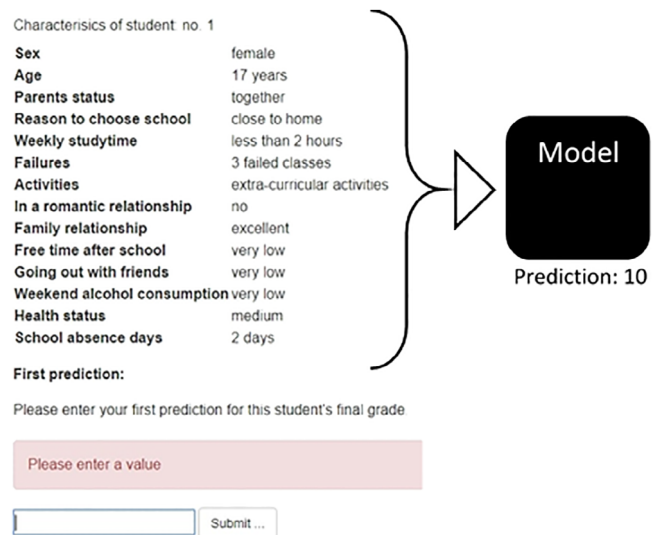


FIGURE 1 T1: Naturally interpretable model (global). Participants were explained that the linear regression found that the following characteristics have a significant (positive: green or negative: red) impact on the student's grades. The other characteristics have—according to the model—no significant impact (grey)

FIGURE 2 T2: Black box with prediction. Participant solely received the prediction (grade) for a given student



We ran the experiment using Amazon Mechanical Turk using Sophie Labs, as a platform for designing online experiments. We recruited 80 participants and randomly assigned them to one of our four treatments. Before starting the experiment, they were given a short introduction to the scenario and the task. They were told that their payment will be adapted according to how good their predictions on the student's grades are (i.e., average absolute deviation from the true value). Participants could only participate in one of our four treatments (between subjects design).

Participants received a show-up fee of \$0.50. In the experiment they collected points with their estimates with 1 point = \$ 0.01. The maximum number of points a participant could earn with an estimate was 200. It would be reduced by the absolute deviation of the predictions from the true value. The payment model was chosen to ensure that participants are encouraged to give their best estimation effort.

3.2 | Task

The participants' task in this study was to predict the final grade of different students, an integer between 0 (worst score) and 20 points (best score). We chose the well-known student performance data set from University of Minho,3 Portugal that gives insights into the student's achievements in secondary education of two Portuguese schools. It contains 649 instances (e.g., students) and includes student's grades as well as demographic, social and school related features. We only used a subset of 14 features. The attributes are easy to understand for the participants and

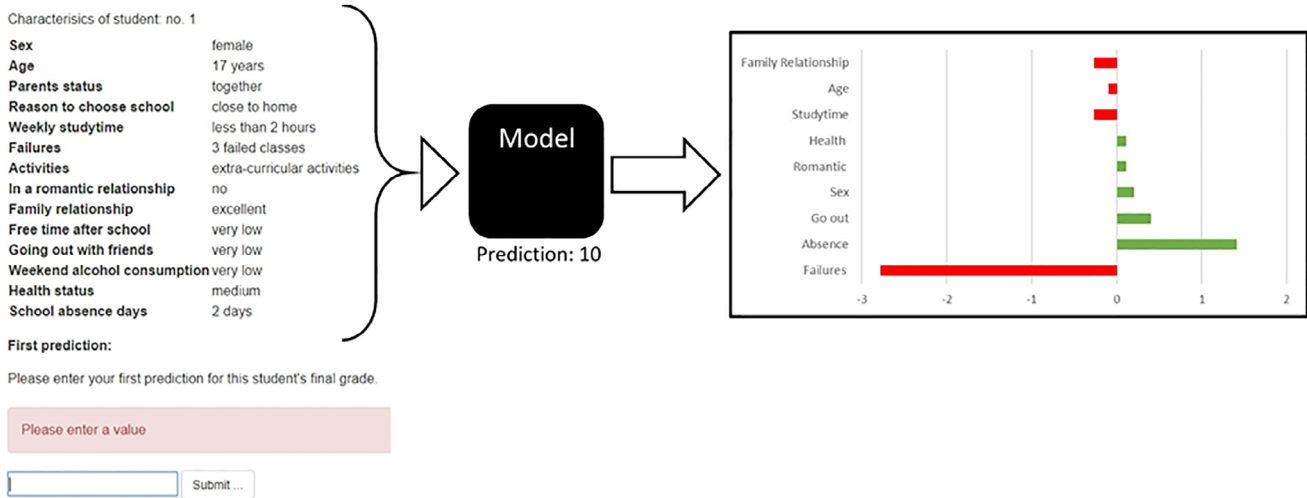


FIGURE 3 T3: Black box with local explanation. Participants were explained that in order to gain an insight into the decision-making process of the Random Forests, we have introduced an additional interpretable surrogate model which highlights (positive: green or negative: red) those academic and personal characteristics that were most important solely for the specific instance (student)

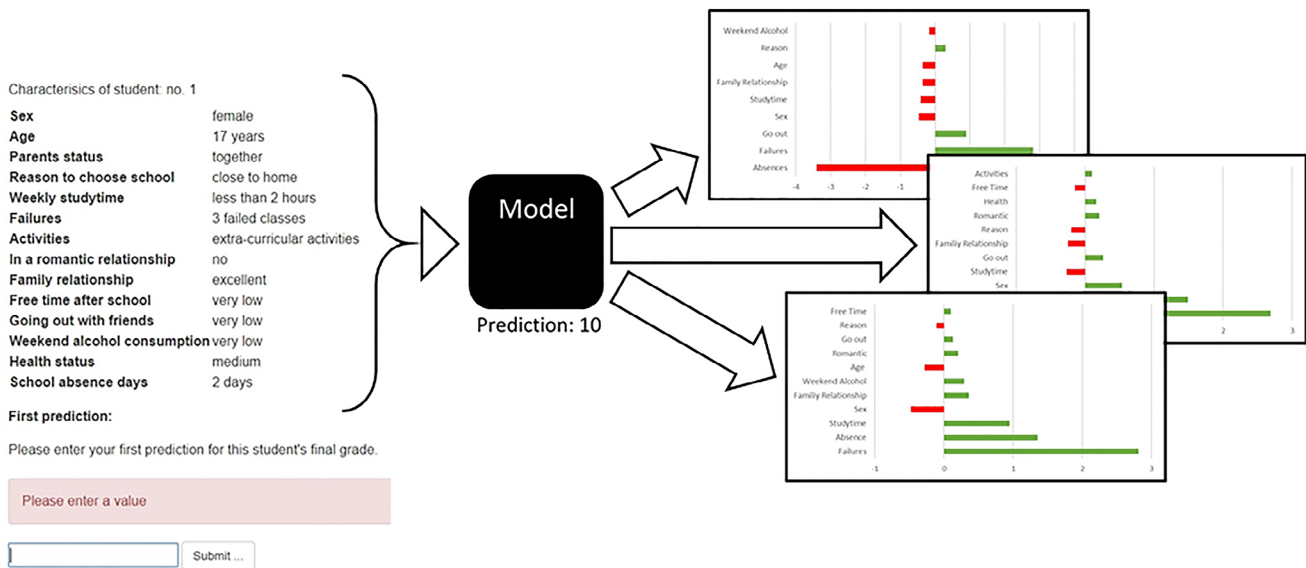


FIGURE 4 T4: Black box with global explanation. Participants were explained that we selected three representative students. If they grasp the concept on how the decision making is done for those three, this knowledge helps them to understand and probably give better predictions. Note that also the characteristics of the three students were shown to the participants

should not overwhelm them with a lot of information. This was ensured in a pre-study that was conducted at our institution. We chose the following subset: sex, age, parents status, reason to choose school, study time (weekly), failures, activities, romantic relationship, family relationship, free time, go out, weekend, health, absence.

To make the predictions, participants were given information such as academic and personal characteristics of the students. After the first prediction, participants received advice from a DSS. For the white box model (T1) we stated that it has a MAE of 4.64. Further we explained it by describing that the average advice from the DSS is 4.64 points away from the true grade of a given student. For the Random Forest (T2–T4) we had a MAE of 2.90. Therefore, the Random Forest had a better accuracy for predicting the student grades than the linear regression model. Then they had the chance to adjust their first prediction, for example, to give a second prediction that may or may not differ from their first prediction. After the second prediction, the prediction round was over and a new round began that was exactly the same as the previous one with a new target student to predict. Overall, there were five prediction rounds for every participant.

We designed the estimation task in this way, because our task is a true prediction task in the sense that participants are aware that, although there is a true answer to the question, no one except the experimenters could possibly know this answer. Moreover, the questions are related to a participant's personal world of experience, which we expected to ensure that the participant could easily understand the task, to enhance the participant's perception of self-efficacy and thus making cognitive efforts invested by the participants improve the prediction.

3.3 | Metrics

For the analysis, we have a total of 400 observations from 80 participants over 5 rounds (20 participants per treatment). Estimation accuracies are measured by the prediction errors, defined as the absolute deviation of the respective prediction from the actual answer (true value). Absolute prediction errors are denoted by $absErr(\mu)$ for the first prediction μ and $absErr(\hat{\mu})$ for the second prediction $\hat{\mu}$. Correspondingly, the relative prediction error $absErr(\hat{\mu})/absErr(\mu)$ is denoted by $relErr$.

In order to quantify the trust in the treatment, the difference between the initial estimation of the participants and the respective adjustment given by a corresponding advice was calculated. It was assumed that a higher adaptation rate for an advice signals more trust in the respective DSS. Based on a first estimate μ_{ti} , β_{ti} records the adjustment of the second estimate μ'_{ti} to the information $\hat{\mu}_{ti}$:

$$\beta_{ti} = \frac{\mu'_{ti} - \mu_{ti}}{\hat{\mu}_{ti} - \mu_{ti}}, \quad (3)$$

with i being the participant and t the prediction round. β_{ti} is the weight of advice, that is, the weight the participant assigns to the advice he receives, and $1 - \beta_{ti}$ is accordingly the weight he places on his own initial prediction. The more precise the participant expects the advice to be, the more weight will $\hat{\mu}_{ti}$ receive in his revised estimate μ'_{ti} .

4 | RESULTS

In total 80 participants attended the experiment. No specific background or requirements were demanded. The average age of the participants was 37 years, the youngest 21 and the oldest 64. In total more men than women—48 compared to 31—and 1 diverse were surveyed. Overall 46 white collar workers, 12 blue collar workers, 1 PhD, 13 self-employed and 8 unemployed participated in the experiment.

Figure 5 presents descriptive data on prediction errors and adjustments. It shows that absolute prediction errors $absErr(\mu)$ for the first prediction are on comparable levels for all four treatments, but slightly higher for the white-box treatment (see Table 1 with 6.48, 6.32, 6.01 and 6.12, respectively). All differences are statistically not significant (Mann–Whitney U and Bonferroni correction, all p -levels .33 or higher). Furthermore, relative adjustment indicate a lower extent of information usage in T2 compared to T1, T3 and T4 (with T2 significantly lower than T1, T3 and T4, all p -levels <.05).

Absolute errors in the second prediction ($absErr(\hat{\mu})$) are highest in the white-box treatment (T1), little lower in the black-box treatment (T2) and are on a comparable level for the two treatments with explanation treatments (T3, T4) (5.12, 4.51, 3.7, 3.93, respectively). The prediction error of the white-box model (T1) is significantly higher than the prediction error of all the black-box models (T2, T3 and T4; Mann–Whitney U test with p -values .041, .000 and .000 one-sided, respectively). Moreover, the prediction error of the black-box model (T2) is significantly higher than the prediction error of the black-box model with local explanation (T3 Mann–Whitney U test with p -values .038 one-sided) but not higher

TABLE 1 Absolute and relative estimations errors

	T1	T2	T3	T4
$absErr(\mu)$ est 1	6.48	6.32	6.01	6.12
$absErr(\hat{\mu})$ est 2	5.12	4.51	3.70	3.93
Difference scores	1.36	1.81	2.31	2.19
$relErr(\mu/\hat{\mu})$	0.79	0.71	0.61	0.64
Rel adjustment	0.59	0.42	0.56	0.55

Note: Absolute errors indicate the average absolute deviation from the true value in the first (est 1) and second prediction (est 2). Difference scores describe mean absolute change between first and second error, for example, a positive value meaning a better second prediction. The relative adjustment indicates the extent to which the information was used for the second estimate (i.e., 0 meaning not at all and 1 meaning 100%).

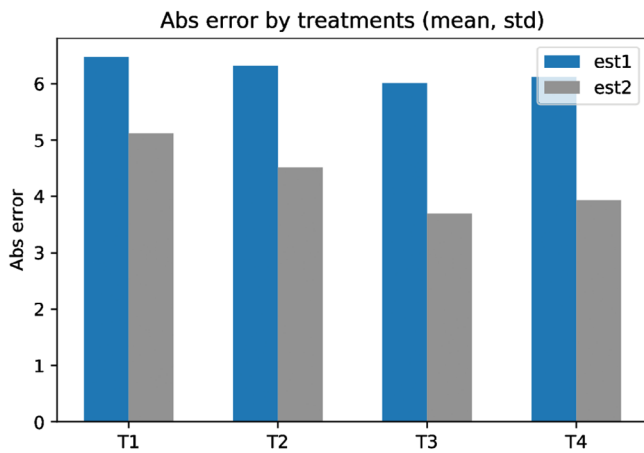


FIGURE 5 Absolute errors per treatments

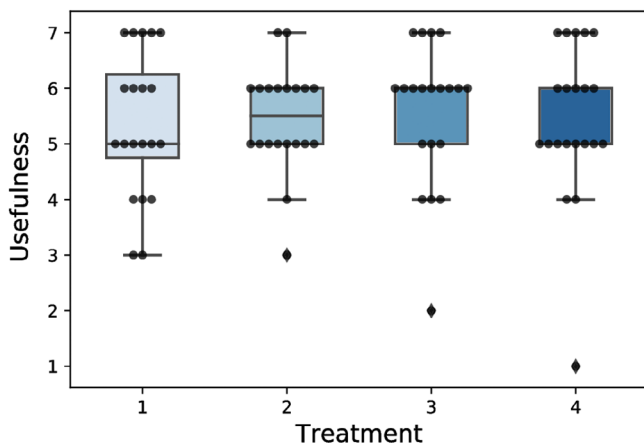


FIGURE 6 Box-plots for subjective usefulness per treatment T1(5.35), T2(5.45), T3(5.55), T4(5.43)

than the prediction error of the black-box model with global explanation (T4; Mann-Whitney U tests with p -values .091 one-sided). All other differences are statistically not significant (Mann-Whitney U tests, all p -levels .36 or higher).

As expected, participants improve their predictions after receiving advice in all four treatment conditions. The differences between the absolute errors of the first and the second prediction are all highly significant (all p -levels below .001).

Additionally, we used Cohen's d as an appropriate effect size for the comparison between two means of the relative adjustments $relErr(\mu/\hat{\mu})$. We found medium effects for the comparison of T1 versus T2 ($d = 0.406$) and small effects for T2 versus T3 (-0.365) and T2 versus T4 (-0.281), whereas only marginal effects for T1 versus T3 (0.074), T1 versus T4 (0.083) and T3 versus T4 (0.022). Therefore, Cohen's d suggests a treatment effect of all treatments compared with T2.

The participants were also asked to directly provide a ranking on how useful the given explanations were via a Likert scale. The scale was from 1 (strongly disagree) to 7 (strongly agree). The box-plots for T1, T2, T3, T4 are shown in Figure 6 per treatment with the means of 5.35, 5.45, 5.55, 5.43 respectively, showing no significant difference. Although there are only slight differences between the treatments, T3 was rated as the most useful.

Wang, Yang, Abdul, and Lim (2019) have described that different users have different needs for explanation. In this case lay users preferred local explanations to understand each specific model prediction over the global explanations. For example, developers may find global explanations more useful to get a high-level inspection of the model in order to debug it. Berry and Broadbent (1987) explored the effect of explanations on the user's performance at a simulated decision task where participants received a global explanation and in one condition an additional local explanation. Participants who received the additional local explanation performed better. Hohman, Head, Caruana, DeLine, and Drucker, (2019) also mentioned in their work that global and local explanation paradigms are complementary.

Thereby, it might be an optimal solution to combine both types of explanations. First of all, a global explanation could be shown to each participant and additionally a local explanation if required (or vice versa). This could be a solution for DSS designed for domain experts.

5 | CONCLUSION

In this paper, we investigated whether the advice of an automated decision-making system affects participants' willingness to adapt their initial prediction revision via a randomized, human between-subject design experiment. We noticed that as expected all participants improved their predictions after receiving an advice whether it was a complete black box or an black box with an explanation. The major findings were that naturally interpretable models were not incorporated more heavily in the decision process than black box models and explanations from them. The black box treatment (T2) had a little lower absolute error, which could be related to a smaller MAE. The black box model without an explanation was used significantly less compared to a black box with local explanation. Participants used local explanations more than global explanations. Probably participants preferred this type of explanation more than the global one because they are tailored to the specific instance. Thereby, the participants did not have to figure out the entire reasoning process and transfer the learnings to the specific instance. For global explanations like SP-LIME it probably was hard for the participants to grasp the general decision-making process by just showing them three example explanations with a high coverage. For interpretable models or explanations we want the possibility to reproduce the decision-making process. This gives us more trust in the model, if we can predict how the model will behave beforehand. This is basically referred to global interpretability as in naturally interpretable models or global surrogates (mimic learning).

Our future work will be to assess more explanation approaches with domain experts. Especially, we want to examine local against global explanations in detail. Furthermore, the focus will be on procedures to generate a more human-like explanation (entire sentences) with decision rules.

ORCID

Nadia Burkart  <https://orcid.org/0000-0003-4700-057X>

ENDNOTES

¹For more information on FICO see <https://www.fico.com/en/products/fico-score>

²For more information on Sesame Credit see <https://www.creditsesame.com/>

³<http://archive.ics.uci.edu/ml/datasets/Student2BPerformance>

REFERENCES

- Adler, P., Falk, C., Friedler, S., Rybeck, G., Scheidegger, C., Smith, B., & Venkatasubramanian, S. (2018). Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54, 95–122.
- Ashton, A. H., & Ashton, R. H. (1988). Sequential belief revision in auditing. *Accounting Review*, 63, 623–641.
- Berry, D. C., & Broadbent, D. E. (1987). Explanation and verbalization in a computer-assisted search task. *The Quarterly Journal of Experimental Psychology*, 39, 585–609.
- Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *IEEE symposium on security and privacy (SP)*, 2016 (pp. 598–617). San Jose, CA: IEEE.
- Davenport, T. H., & Harris, J. G. (2005). Automated decision making comes of age. *MIT Sloan Management Review*, 46, 83.
- Diamantopoulos, A., & Mathews, B. (1989). Factors affecting the nature and effectiveness of subjective revision in sales forecasting: An empirical study. *Managerial and Decision Economics*, 10, 51–59.
- Doshi-Velez, F. and Kim, B. (2017) Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, 25, 3–23.
- Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a bayesian perspective. *Psychological Review*, 90, 239–260.
- Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24, 44–65.
- Goodwin, P., & Fildes, R. (1999). Judgmental forecasts of time series affected by special events: Does providing a statistical forecast improve accuracy? *Journal of Behavioral Decision Making*, 12, 37–53.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(93), 42. <https://doi.org/10.1145/3236009>
- Hall, P., Phan, W. and Ambati, S. (2017) Ideas on interpreting machine learning. Retrieved from <https://www.oreilly.com/ideas/ideas-on-interpreting-machine-learning>.
- Henelius, A., Puolamäki, K., Boström, H., Asker, L., & Papapetrou, P. (2014). A peek into the black box: Exploring classifiers by randomization. *Data Mining and Knowledge Discovery*, 28, 1503–1529.
- Hohman, F., Head, A., Caruana, R., DeLine, R., & Drucker, S. M. (2019). Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–13). New York, NY: Association for Computing Machinery.
- Kruger, J. (1999). Lake wobegon be gone! the "below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology*, 77, 221–232.
- Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S. and Doshi-Velez, F. (2019) An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006*.

- Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International conference on knowledge discovery and data mining* (pp. 1675–1684). New York, NY: Association for Computing Machinery.
- Lakkaraju, H., Kamar, E., Caruana, R. and Leskovec, J. (2017) Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*.
- Lim, J. S., & O'Connor, M. (1995). Judgemental adjustment of initial forecasts: Its effectiveness and biases. *Journal of Behavioral Decision Making*, 8, 149–168.
- Miller, T. (2017) Explanation in artificial intelligence: Insights from the social sciences. *arXiv preprint arXiv:1706.07269*.
- Phillips, R. L., Chang, K. H. and Friedler, S. A. (2017) Interpretable active learning. *arXiv preprint arXiv:1708.00049*.
- Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. and Wallach, H. (2017) Manipulating and measuring model interpretability.
- Puolamäki, A.H.K. and Ukkonen, A. (2017) Interpreting classifiers through attribute interactions in datasets.
- Quattrone, G. A. (1982). Overattribution and unit formation: When behavior engulfs the person. *Journal of Personality and Social Psychology*, 42, 593–607.
- Robnik-Šikonja, M., & Kononenko, I. (2008). Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, 20, 589–600.
- Rüping, S. (2005) Learning with local models.
- Schmidt, P. and Biessmann, F. (2019) Quantifying interpretability and trust in machine learning systems. *arXiv preprint arXiv:1901.08558*.
- Singh, M. T. R. S. and Guestrin, C. (2016) Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.
- Su, G., Wei, D., Varshney, K. R. and Malioutov, D. M. (2015) Interpretable two-level boolean rule learning for classification. *arXiv preprint arXiv:1511.07361*.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.
- Turner, R. (2016) A model explanation system: Latest updates and extensions. *arXiv preprint arXiv:1606.09517*.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131.
- Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–15). New York, NY: Association for Computing Machinery.
- Wang, F. and Rudin, C. (2014) Falling rule lists. *arXiv preprint arXiv:1411.5899*.
- Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E. and MacNeille, P. (2015) Or's of and's for interpretable classification, with application to context-aware recommender systems. *arXiv preprint arXiv:1504.07614*.
- Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, 93, 1–13.
- Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, 83, 260–281.
- Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1–12). New York, NY: Association for Computing Machinery.

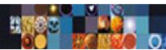
AUTHOR BIOGRAPHIES

Nadia Burkart is a research scientist at the Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB in Karlsruhe. She received her master degree in information systems from the University of Applied Science in Karlsruhe. Currently she is working on her PhD thesis in the field of Explainable Machine Learning.

Sebastian Robert is group leader at Fraunhofer Institute of Optronics, System Technology and Image Exploitation IOSB in Karlsruhe, Germany. He holds a master degree in Information Systems and a Ph.D. from the University of Osnabrück. His current research interests include clinical decision making and behavioral sciences.

Marco Huber received his diploma, Ph.D., and habilitation degrees in computer science from the Karlsruhe Institute of Technology (KIT), Germany, in 2006, 2009, and 2015, respectively. From June 2009 to May 2011, he was leading the research group “Variable Image Acquisition and Processing” of the Fraunhofer IOSB, Karlsruhe, Germany. Subsequently, he was Senior Researcher with AGT International, Darmstadt, Germany, until March 2015. From April 2015 to September 2018, he was responsible for product development and data science services of the Katana division at USU Software AG, Karlsruhe, Germany. At the same time he was adjunct professor of computer science with the KIT. Since October 2018 he is full professor with the University of Stuttgart. At the same time, he is director of the Center for Cyber Cognitive Intelligence (CCI) and of the Department for Image and Signal Processing with Fraunhofer IPA in Stuttgart, Germany. His research interests include machine learning, planning and decision making, image processing, data analytics, and robotics. He has authored or co-authored more than 80 publications in various high-ranking journals, books, and conferences, and holds two U.S. patents and one EU patent.

How to cite this article: Burkart N, Robert S, Huber MF. Are you sure? Prediction revision in automated decision-making. *Expert Systems*. 2020;1–19. <https://doi.org/10.1111/exsy.12577>



APPENDIX A.

SoPHIE

Procedure

In total, you will predict the performance of five students. Each round of prediction will target one student and will be made according to this scheme:

1. You will predict the final grade of a student based on information such as academic and personal characteristics. The prediction must be a whole number between 0 and 20.
2. You will get advice from an assistance system. The advice is a whole number between 0 and 20.
3. After having received the advice, you may or may not adjust your first prediction. In any case, you have to predict a whole number between 0 and 20.

After your second prediction, the round is over and a new round begins that is exactly the same as the previous one.

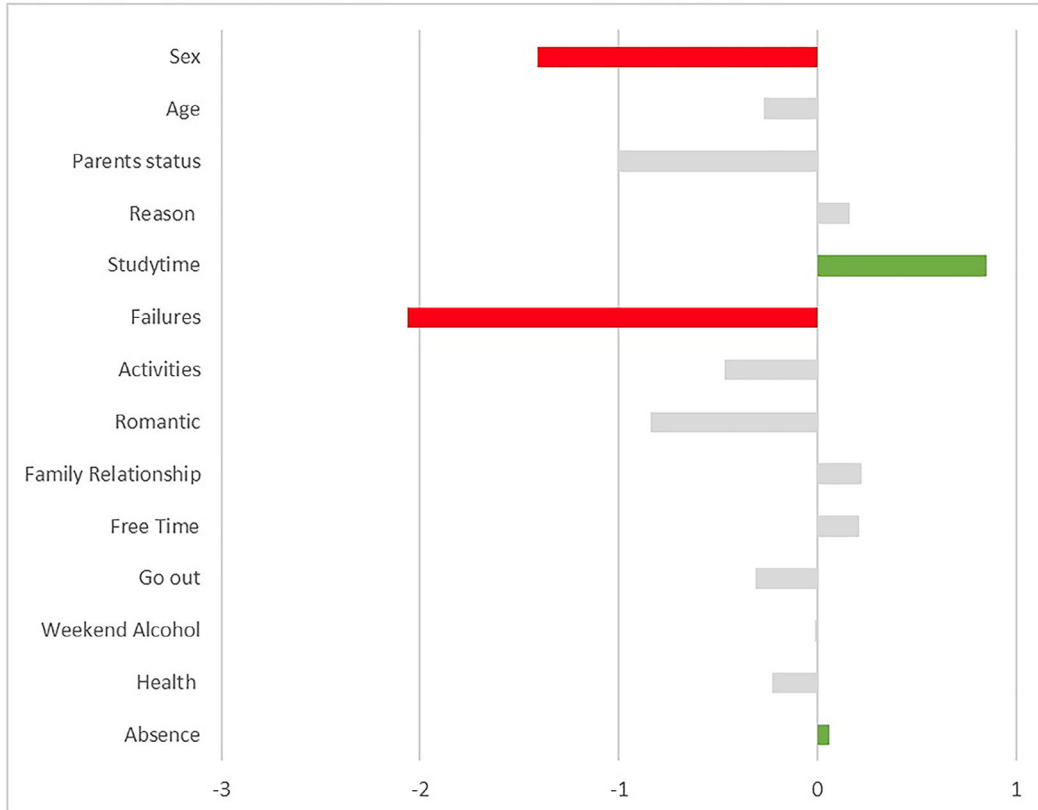
[Continue ...](#)

FIGURE A1 Description of the procedure



Advice explanation

The advice from the assistance system is based on a linear regression. After evaluating all samples, the linear regression found that the following characteristics to have a significant (positive: green or negative: red) impact on the student's grades. The other characteristics have - according to the model - no significant impact. These characteristics are shown in grey.



This means, that e.g., female students have better grades than male students (*sex*); more studytime leads to better grades (*studytime*); failing a class means worse grades (*failures*) and absence school days have a weak negative impact in grades (*absence*).


The analysis is based on 649 samples, i.e., students. It has a mean absolute error (MAE) of 4.64. This means, that the average advice from the assistance system is 4.64 points away from the true grade of a given student.

Note: This is the only time we will show you this explanation of the assistance system.

Continue ...

FIGURE A2 Advice explanation for Treatment 1

FIGURE A3 First prediction for Treatment 1

 **SoPHIE**

First prediction in round 1


Characteristics of student: no. 1

Sex	female
Age	17 years
Parents status	together
Reason to choose school	close to home
Weekly studytime	less than 2 hours
Failures	3 failed classes
Activities	extra-curricular activities
In a romantic relationship	no
Family relationship	excellent
Free time after school	very low
Going out with friends	very low
Weekend alcohol consumption	very low
Health status	medium
School absence days	2 days

First prediction:

Please enter your first prediction for this student's final grade.

FIGURE A4 Second prediction (revision) for Treatment 1

 **SoPHIE**

Second prediction in round 1

Characteristics of student: no. 1

Sex	female
Age	17 years
Parents status	together
Reason to choose school	close to home
Weekly studytime	less than 2 hours
Failures	3 failed classes
Activities	extra-curricular activities
In a romantic relationship	no
Family relationship	excellent
Free time after school	very low
Going out with friends	very low
Weekend alcohol consumption	very low
Health status	medium
School absence days	2 days

For this student, the assistance system has calculated an advice of: **5**

Your first prediction was: **10**

Please enter your second prediction:

SOPHIE

Instructions (3/3)

After your first prediction you will receive advice from an assistance system. The advice is based on a predictive analysis called Random Forests. This is a machine learning approach based on so called decision trees - each capable of producing a response (e.g., predict a student's grade) when presented with a set of predictor values (e.g., academic and personal characteristics). The different decision trees are combined by "voting" for the most popular prediction. Random Forests have proven to be one of the most accurate learning algorithms available.

Typically, Random Forests consist of a large number of deep trees, where each tree is trained on a lot of data. Therefore, Random Forests are treated as a black box, meaning their prediction techniques are hidden from us and we cannot say with certainty how a prediction was derived.

The advice given by the assistance system will always be a number between 0 and 20.

Continue ...

FIGURE A5 Advice explanation for Treatment 2

SOPHIE

Instructions (3/3)

After your first prediction you will receive advice from an assistance system. The advice is based on a predictive analysis called Random Forests. This is a machine learning approach based on so called decision trees - each capable of producing a response (e.g., predict a student's grade) when presented with a set of predictor values (e.g., academic and personal characteristics). The different decision trees are combined by "voting" for the most popular prediction. Random Forests have proven to be one of the most accurate machine learning algorithms available.

Typically, Random Forests consist of a large number of deep trees, where each tree is trained on a lot of data. Therefore, Random Forests are sometimes treated as a black box, meaning their prediction techniques are hidden from us and we cannot say with certainty how a prediction was derived.

In order to gain an insight into the decision-making process of the Random Forests, we have introduced an additional interpretable model. This interpretable model is used to explain individual predictions of Random Forest. In general, the interpretable model will highlight those academic and personal characteristics that were most important to the Random Forest for the specific prediction. Therefore, this model can help you to understand why the Random Forest made a certain prediction.

Here is an example of an interpretable model, where the characteristics A and B have a positive and the characteristic C a negative influence on the prediction.

Characteristic	Influence
Characteristic A	3
Characteristic B	1.8
Characteristic C	-2.2

The advice given by the assistance system will always be a number between 0 and 20.

Continue ...

FIGURE A6 Additional instructions for Treatment 3

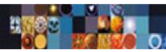


FIGURE A7 Second prediction for Treatment 3

Second prediction in round 1

Characteristics of student: no. 1

Sex	female
Age	17 years
Parents status	together
Reason to choose school	close to home
Weekly studytime	less than 2 hours
Failures	3 failed classes
Activities	extra-curricular activities
In a romantic relationship	no
Family relationship	excellent
Free time after school	very low
Going out with friends	very low
Weekend alcohol consumption	very low
Health status	medium
School absence days	2 days

Interpretable model of your assistance system:

Characteristic	Value (approx.)
Family Relationship	-0.2
Age	-0.1
Studytime	-0.2
Health	0.1
Romantic	0.1
Sex	0.2
Go out	0.4
Absence	1.4
Failures	-2.5

For this student, the assistance system has calculated an advice of: **10**

Your first prediction was: **12**

Please enter your second prediction:

SoPHIE

FIGURE A8 Additional instruction for Treatment 4

Advice explanation

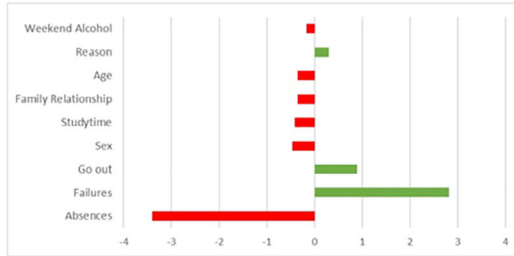
The advice from the assistance system is based on a Random Forest. The machine learning approach is based on 649 samples, i.e., students. It has a mean absolute error (MAE) of 2.50. This means, that the average advice from the assistance system is 2.50 points away from the true grade of a given student.

After evaluating all samples, the interpretable model selected the following 3 students with their grades as being representative for the Random Forest:

Representative student 1 with a grade of 3

Sex	male
Age	1 years
Parents status	together
Reason to choose school	school reputation
Weekly studytime	2 to 3 hours
Failures	0 failed classes
Activities	no extra-curricular activities
In a romantic relationship	yes
Family relationship	good
Free time after school	medium
Going out with friends	very high
Weekend alcohol consumption	low
Health status	medium
School absence days	0 days

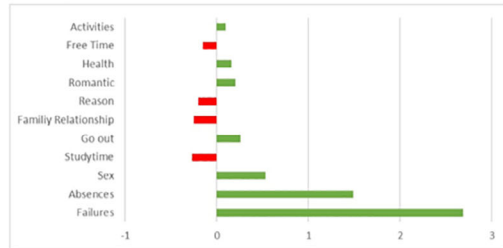
Explanation of interpretable model for student 1



Representative student 2 with a grade of 11

Sex	female
Age	17 years
Parents status	together
Reason to choose school	close to home
Weekly studytime	less than 2 hours
Failures	0 failed classes
Activities	no extra-curricular activities
In a romantic relationship	no
Family relationship	good
Free time after school	high
Going out with friends	medium
Weekend alcohol consumption	medium
Health status	very good
School absence days	4 days

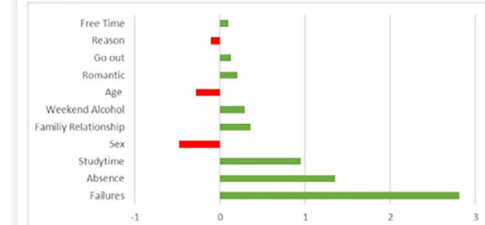
Explanation of interpretable model for student 2



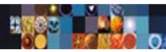
Representative student 3 with a grade of 16

Sex	male
Age	18 years
Parents status	together
Reason to choose school	close to home
Weekly studytime	more than 10 hours
Failures	0 failed classes
Activities	no extra-curricular activities
In a romantic relationship	no
Family relationship	excellent
Free time after school	medium
Going out with friends	medium
Weekend alcohol consumption	very low
Health status	very bad
School absence days	7 days

Explanation of interpretable model for student 3



The interpretable model only lists the most important characteristics for the representative students in each case. Note: This is the only time we will show you this explanation of the assistance system.



SOPHIE

Instructions (3/3)

After your first prediction you will receive advice from an assistance system. The advice is based on a predictive analysis called Random Forests. This is a machine learning approach based on so called decision trees - each capable of producing a response (e.g., predict a student's grade) when presented with a set of predictor values (e.g., academic and personal characteristics). The different decision trees are combined by "voting" for the most popular prediction. Random Forests have proven to be one of the most accurate machine learning algorithms available.

Typically, Random Forests consist of a large number of deep trees, where each tree is trained on a lot of data. Therefore, Random Forests are sometimes treated as a black box, meaning their prediction techniques are hidden from us and we cannot say with certainty how a prediction was derived.

In order to gain an insight into the decision-making process of the Random Forests, we have introduced an additional interpretable model. This interpretable model selects a set of 3 representative students with explanations that describe the Random Forest decision process behind. The procedure picks explanations of students that cover as many characteristics as possible. You can analyse these 3 representative students to generalize how the Random Forest makes predictions. The 3 representative students will be shown to you later.

Here is an example of an interpretable model, where the characteristics A and B have a positive and the characteristic C a negative influence on the prediction.



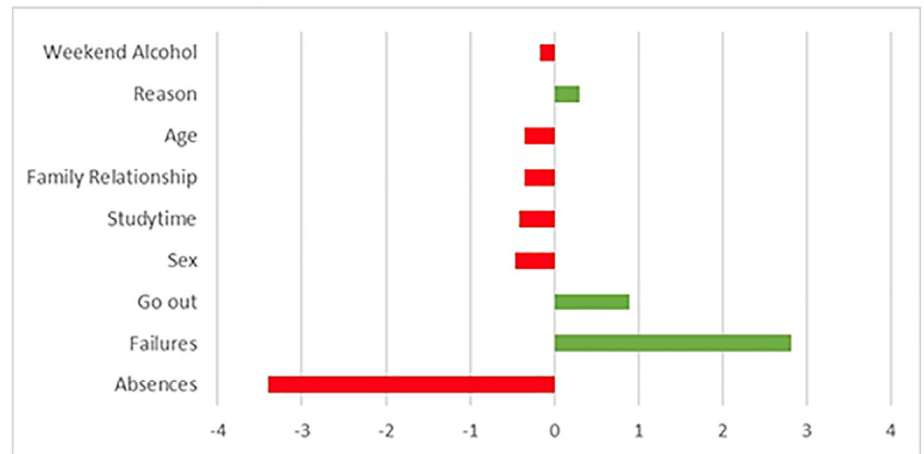
The advice given by the assistance system will always be a number between 0 and 20.

Continue ...

FIGURE A9 Advice explanation for Treatment 4 (Part 1)

FIGURE A10 Advice explanation for Treatment 4 (Part 2)

Explanation of interpretable model for student 1

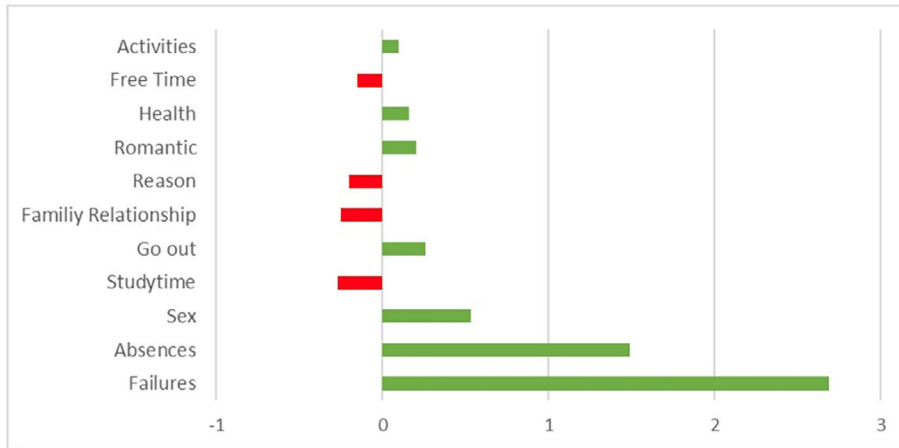


Representative student 2 with a grade of 11

Sex	female
Age	17 years
Parents status	together
Reason to choose school	close to home
Weekly studytime	less than 2 hours
Failures	0 failed classes
Activities	no extra-curricular activities
In a romantic relationship	no
Family relationship	good
Free time after school	high
Going out with friends	medium
Weekend alcohol consumption	medium
Health status	very good
School absence days	4 days

FIGURE A11 Advice explanation for Treatment 4 (Part 3)

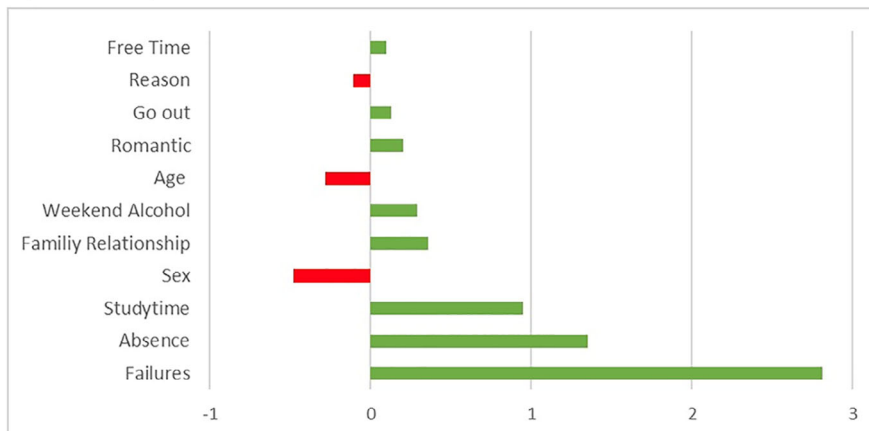
Explanation of interpretable model for student 2



Representative student 3 with a grade of 16

Sex	male
Age	18 years
Parents status	together
Reason to choose school	close to home
Weekly studytime	more than 10 hours
Failures	0 failed classes
Activities	no extra-curricular activities
In a romantic relationship	no
Family relationship	excellent
Free time after school	medium
Going out with friends	medium
Weekend alcohol consumption	very low
Health status	very bad
School absence days	7 days

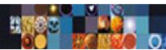
Explanation of interpretable model for student 3



The interpretable model only lists the most important characteristics for the representative students in each case.

Note: This is the only time we will show you this explanation of the assistance system.

FIGURE A12 Advice explanation for Treatment 4 (Part 4)

**FIGURE A13** Second prediction for Treatment 4 (Part 4)

Second prediction in round 1

Characteristics of student: no. 1

Sex	female
Age	17 years
Parents status	together
Reason to choose school	close to home
Weekly studytime	less than 2 hours
Failures	3 failed classes
Activities	extra-curricular activities
In a romantic relationship	no
Family relationship	excellent
Free time after school	very low
Going out with friends	very low
Weekend alcohol consumption	very low
Health status	medium
School absence days	2 days

For this student, the assistance system has calculated an advice of: **10**

Your first prediction was: **12**

Please enter your second prediction: