Association for Information Systems

AIS Electronic Library (AISeL)

ICIS 2022 Proceedings

Human Computer /Robot Interaction

Dec 12th, 12:00 AM

Effectiveness of Example-Based Explanations to Improve Human Decision Quality in Machine Learning Forecasting Systems

Tobias Benjamin Fahse
University of St.Gallen, tobias.fahse@unisg.ch

Ivo Blohm
University of St. Gallen, ivo.blohm@unisg.ch

Benjamin van Giffen
University of St. Gallen, benjamin.vangiffen@unisg.ch

Follow this and additional works at: https://aisel.aisnet.org/icis2022

Recommended Citation

Fahse, Tobias Benjamin; Blohm, Ivo; and van Giffen, Benjamin, "Effectiveness of Example-Based Explanations to Improve Human Decision Quality in Machine Learning Forecasting Systems" (2022). *ICIS 2022 Proceedings*. 5.
https://aisel.aisnet.org/icis2022/hci_robot/hci_robot/5

This material is brought to you by the International Conference on Information Systems (ICIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in ICIS 2022 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Effectiveness of Example-Based Explanations to Improve Human Decision Quality in Machine Learning Forecasting Systems

Short Paper

Tobias Fahse
University of St. Gallen
St. Gallen, Switzerland
tobias.fahse@unisg.ch

Ivo Blohm
University of St. Gallen
St. Gallen, Switzerland
ivo.blohm@unisg.ch

Benjamin van Giffen
University of St. Gallen
St. Gallen, Switzerland
benjamin.vangiffen@unisg.ch

Abstract

Algorithmic forecasts outperform human forecasts by 10% on average. State-of-the-art machine learning (ML) algorithms have further expanded this discrepancy. Because a variety of other activities rely on them, sales forecasting is critical to a company's profitability. However, individuals are hesitant to use ML forecasts. To overcome this algorithm aversion, explainable artificial intelligence (XAI) can be a solution by making ML systems more comprehensible by providing explanations. However, current XAI techniques are incomprehensible for laymen, as they impose too much cognitive load. We contribute to this research gap by investigating the effectiveness in terms of forecast accuracy of two example-based explanation approaches. We conduct an online experiment based on a two-by-two between-subjects design with factual and counterfactual examples as experimental factors. A control group has access to ML predictions, but not to explanations. We report results of this study: While factual explanations significantly improved participants' decision quality, counterfactual explanations did not.

Keywords: Explainable AI, Explanation Interface, Sales Forecasting, Machine Learning

Introduction

Algorithmic forecasts outperform human forecasts by 10% on average (Grove et al., 2000). This disparity has expanded thanks to cutting-edge machine learning (ML) algorithms (Blohm et al., 2020). Because business decisions based on sales forecasting are regarded as particularly important and a variety of other activities rely on them, accurate sales forecasting is critical to a company's profitability (Sun et al., 2008; Mentzer & Bienstock, 1998). As a result, sales forecasting is one of the primary value propositions of Artificial Intelligence (AI) (Cam et al., 2019). Throughout this paper, we generally refer to ML, the most common technical implementation of AI (Engel et al., 2022), although the generic term AI is still used in some contexts. Despite the numerous benefits, people are hesitant to use ML predictions. There are various reasons for this so-called algorithm aversion: (1) Decision makers do not understand algorithmic forecasts (Herm et al., 2021; Poursabzi-Sangdeh et al., 2021), (2) Greater intolerance for algorithmic than for human errors (Dietvorst et al., 2014), (3) Lower willingness to use ML forecasts (Castelo et al., 2019; Prahl & Van Swol, 2017), (4) Low algorithmic literacy and experience of individuals (Burton et al., 2020).

Explainable AI (XAI) could be a solution to overcome algorithm aversion. Users interact with an explanation interface (XI) in a XAI system to request model predictions and explanations. This way, the behavior of ML systems becomes more understandable to humans (Gunning et al., 2019). However, for users with little statistical education, XAI approaches for time series forecasting are scarce (Ates et al., 2020) and underexplored (Liao et al., 2020). The majority of research focuses on algorithm-centric explainability for ML experts. As many decision makers in the sales forecasting domain are no experts in ML, XAI approaches for lay users are especially important. We focus on example-based explanations in this research, which are very promising for time series forecasts but lack evaluation (Rojat et al., 2021). The overarching goal of this research project is to aid decision makers who use ML forecasts in improving their decision quality, i.e. the accuracy of sales forecasts (Geva & Saar-Tsechansky, 2021). Following the concept of hybrid intelligence (Dellermann et al., 2019), we aim to improve the collaboration between humans and AI.

To address the outlined research problem, we investigate the effectiveness of two selected example-based explanation approaches, namely factual and counterfactual examples, which are embedded into an XI for sales forecasting. Against this backdrop, we conduct a lab experiment and report preliminary results which validate the effectiveness of the two selected example-based explanation approaches to increase the decision quality of lay users by making ML systems more understandable to humans. We use a two-by-two between-subjects design with factual and counterfactual examples as experimental factors. The control group used an XI that consists solely of the control panel, the ML forecast, and historical sales data. The remaining groups have access to only factual explanations, only counterfactual explanations, or both. Participants were recruited from an online platform (total of 151 participants). Using the XI, the task was to forecast the next day's sales of 14 products. A medium-sized Swiss bakery historical sales data is used as a real-world evaluation context.

We seek to contribute to theory by investigating the effectiveness of two selected example-based explanation approaches (factual and counterfactual examples) and answering the research question:

RQ: *Are factual and counterfactual explanations in combination with ML forecasts useful in helping people make better forecasts compared to when they receive only ML forecasts?*

Conceptual and Theoretical Background

This research project is grounded on two main streams of literature: XAI for sales forecasting and cognitive load theory.

Explainable Artificial Intelligence for Sales Forecasting

Forecasting is important for business strategy because forecasts are used to make many organizational decisions (Mentzer & Bienstock, 1998). ML allows for more accurate forecasting of future events than traditional approaches (Kelleher et al., 2015) and are often able to outperform humans (Grove et al., 2000), leading to the basic hypothesis of this study:

H1. *Users make better predictions when provided with the ML forecast compared to users who are not provided with the ML forecast.*

However, forecasting accuracy comes at the expense of predictability. By providing explanations, XAI makes the behavior of ML systems more understandable to humans (Gunning et al., 2019). The XIs are user interfaces that allow interaction with the model and asking for predictions and explanations (Mohseni et al., 2020). Effective explanations aid in the mitigation of human biases and the avoidance of erroneous decisions (Wang et al., 2019). Unmet explanation needs are inversely proportional to decision quality (Liao et al., 2020). Against this backdrop, we formulate hypotheses 1 and 2:

H2: *Providing users with factual examples increases the decision quality of users.*

H3: *Providing users with counterfactual examples increases the decision quality of users.*

However, research has yet to answer how XIs can effectively meet the needs of lay users in terms of explanations. Explainability can be divided into two types: global (explaining the inner logic of a ML model, i.e. the general mapping of inputs to outputs) and local (explaining specific model outputs) (Adadi &

Berradda, 2018). Local approaches can be based on examples, features, or rules. Local approaches can be based on examples, features, or rules. Among example-based explanations one can distinguish between factual and counterfactual explanations. Factual explanations are examples of similar situations in which similar outcomes occurred. For instance, a factual example could be: “The prediction was X because in the past, on days with similar characteristics, similar volumes were sold”. Counterfactual explanations provide examples of situations with highly diverging outcomes. An example is: “If the temperature on the day to be forecasted would have been X degrees, the forecast would have been Y instead of Z” (Kenny & Keane, 2021; Dandl et al., 2020)

Feature-based approaches (Apley & Zhu, 2020) specify how much each input feature contributes to a model's prediction, giving experts insight into the algorithm's workings. There are three main drawbacks of existing XAI approaches. First, most approaches focus on feature-based explanations and aim to explain time series classification (rather than forecasting) problems. In contrast, example-based explanations for time series forecasts are very promising, but lack evaluation (Rojat et al., 2021). The effectiveness of example-based explanations is due to their alignment with human reasoning processes and the minimal cognitive load they pose on the user. Second, the methods are often unable to help users with low technical literacy (Cheng et al., 2019), particularly when the data is highly dimensional, as in ML forecasting where many abstract, inherently incomprehensible features such as lag variables are used. Even explainable white-box models (e.g. regression models) produce outputs that are difficult to decipher by laypeople (Hagras, 2018). Third, existing research is algorithm-centric, despite the fact that these mostly feature-based approaches have limited value for improving the quality of lay users' decisions (Alufaisan et al., 2021) and are thus ineffective.

In this light, Abdul et al. (2018) identified trends in XAI and noted the need for improved usability, practical interpretability, and effectiveness for real-world users. In XAI research, most explanation approaches are static, assuming that there is only one message to be delivered. "Users can freely explore the system's behavior" with interactive approaches (Abdul et al., 2018). Cheng et al. (2019) investigate a non-expert stakeholder XI that provides counterfactual explanations but focuses on ML classification tasks. In this study, we address the disadvantages mentioned above and add to existing XAI research by validating the effectiveness of two interactive example-based explanations approaches for time series.

Cognitive Load Theory

Cognitive load is an additive construct that assumes humans process information in a short-term working memory and store it in a long-term memory. Working memory processes all conscious cognitive tasks, but the number of processable information items is limited (Sweller, 1988). There is a difference between intrinsic (task complexity) and extraneous (task presentation complexity) cognitive load (Blohm et al., 2016). The users’ intrinsic cognitive load is increased by the number of forecasts to be made, the number of explaining examples displayed, as well as the complexity of the explanations. In this study, we increase intrinsic cognitive load by incrementally adding factual and counterfactual examples to the base version of the XI without any explanations. However, too much information on an XI may have a negative impact. As a result, because cognitive load is diametral to task performance, XIs should keep it to a minimum. Users who actively contribute cognitive effort to process ML forecasts face significant cognitive load. Explanations add to that potentially leading to cognitive overload and poor decision quality. As a result, the goals of the explanations may not be met. Consequently, we formulate our third hypothesis as follows:

H4a: *Providing users with both factual and counterfactual examples decreases the decision quality of users compared to users who receive only counterfactual examples.*

H4b*: Providing users with both factual and counterfactual examples decreases the decision quality of users compared to users who receive only factual examples.*

Existing research shows that interfaces that place an excessive amount of cognitive load on users result in poor forecasting decision quality (Blohm et al., 2016). This situation is even worse for non-expert users who are unfamiliar with ML forecasting systems. Because lay users have less domain knowledge and experience, they have fewer cognitive structures that allow them to process multiple elements that cause cognitive load at the same time. As a result, according to cognitive load theory, specific XAI approaches are required for lay users. The threshold that delineates between the right amount of cognitive load and too much cognitive load depends on the individual cognitive structures of the users. The goal of this study is to find the threshold for the case of one-day-ahead sales forecasting of baked goods.

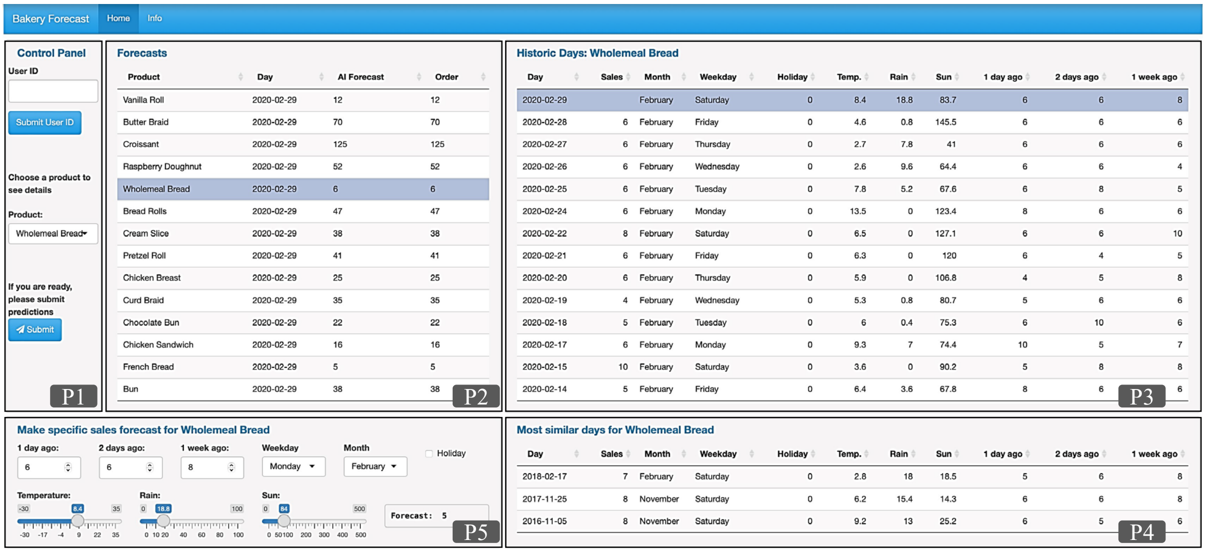
Machine Learning Model, Dataset and Artifact Description

A real-world context is used for this project: As evaluation context, historic transactional point-of-sale bakery data is provided by a medium-sized Swiss bakery chain and used by the ML forecasting model that is implemented in the XI. Out of all available products, we selected a stratified sample of 14 products to make sure that both high and low volume, both easy and hard to predict and both products for which ML outperforms/ does not outperform human predictions are included. The time series spans five years of sales which add up to over 400’000 transactions for all 14 products. It includes a timestamp as well as information on the product type, quantity, and price. Weather indicators (radiation, rainfall, temperature), national holidays, and school vacations were added using external data. Lag variables (lag 1, 2, and 7) were included by performing extensive autocorrelation analysis.

An XGBoost model (Chen & Guestrin, 2016) provides one-step ahead predictions of the next day's sales. On the basis of a performance comparison (root mean squared error, RMSE) of various ML methods, we chose the XGBoost model. XGBoost outperformed the other five models (linear regression, decision tree, random forest, ARIMA, and Neural Network). Our generic prescriptions, on the other hand, are not constrained by the limitations of a single algorithm. To establish a baseline for commercial value, we first assess the correctness of our model. The ML model outperforms the branch managers' predictions by 1.7 percent (RMSE). Based on a 70% margin (SBC, 2019), this increase in forecast accuracy equates to about CHF 13000 in annual savings due to reduced waste in the branch with the greatest unit sales. This amounts to 3.1 percent of the branch's annual revenue. As a result, simply following the ML predictions without making any changes would result in considerable savings. This confirmation is necessary in order to justify any effort to develop an XI for ML forecasting.

We conducted five expert interviews with decision makers of five different Swiss bakeries (two production managers and three CEOs) to validate the general effectiveness of example-based explanations. The interviews confirmed that factual and counterfactual explanations can help decision makers make better decisions and called for displaying predictions for all products at once to be most efficient when making decisions. When the user’s mental model is challenged, the interviewees appreciated the possibility to drill down to the details of the prediction’s explanation. Building upon these findings, we developed the XI to display all ML predictions for all products at once. The factual and counterfactual explanations adapt to the chosen product in the control panel to enable the user to access details for each product.

In total, the XI consists of five panels (Figure 1). Through the control panel (P1), users enter their user ID, choose the product for which they need details and submit their final predictions. The second Panel (P2) displays the products and respective days to-be-predicted and the ML forecast. The users enter their predictions into the “Order” column. Historic sales data is displayed in the third panel (P3). Depending on the chosen product in the control panel, P3 adapts and shows historic sales data for the selected product. The day for which predictions need to be made by the user is highlighted and in the first row. Factual explanations are provided in the fourth panel (P4). By this, the user is enabled to identify most similar historic days supporting the forecast using a KNN algorithm (MacQueen, 1967) that calculates the k nearest data points to the selected, to-be-predicted day. In this implementation, KNN uses the squared euclidean distance to identify the three most similar days. Users can thus construct a mental model of the forecast by creating analogies from representative cases for which the sales quantity is known. Counterfactual explanations are provided in the fifth panel (P5). Starting from the real values of the day to be predicted, the user can manipulate all input features to explore the behavior of the underlying ML model. This way, counterfactual explanations can be created by examining how different feature values affect the forecast.



Control Panel

User ID

Submit User ID

Choose a product to see details

Product: Wholemeal Bread

If you are ready, please submit predictions

Submit

Forecasts

Product	Day	AI Forecast	Order
Vanilla Roll	2020-02-29	12	12
Butter Braid	2020-02-29	70	70
Croissant	2020-02-29	125	125
Raspberry Doughnut	2020-02-29	52	52
Wholemeal Bread	2020-02-29	6	6
Bread Rolls	2020-02-29	47	47
Cream Slice	2020-02-29	38	38
Pretzel Roll	2020-02-29	41	41
Chicken Breast	2020-02-29	25	25
Curd Braid	2020-02-29	35	35
Chocolate Bun	2020-02-29	22	22
Chicken Sandwich	2020-02-29	16	16
French Bread	2020-02-29	5	5
Bun	2020-02-29	38	38

Historic Days: Wholemeal Bread

Day	Sales	Month	Weekday	Holiday	Temp.	Rain	Sun	1 day ago	2 days ago	1 week ago
2020-02-29		February	Saturday	0	8.4	18.8	83.7	6	6	8
2020-02-28	6	February	Friday	0	4.6	0.8	145.5	6	6	6
2020-02-27	6	February	Thursday	0	2.7	7.8	41	6	6	6
2020-02-26	6	February	Wednesday	0	2.6	9.6	64.4	6	6	4
2020-02-25	6	February	Tuesday	0	7.8	5.2	67.6	6	8	5
2020-02-24	6	February	Monday	0	13.5	0	123.4	8	6	6
2020-02-22	8	February	Saturday	0	6.5	0	127.1	6	6	10
2020-02-21	6	February	Friday	0	6.3	0	120	6	4	5
2020-02-20	6	February	Thursday	0	5.9	0	106.8	4	5	8
2020-02-19	4	February	Wednesday	0	5.3	0.8	80.7	5	6	6
2020-02-18	5	February	Tuesday	0	6	0.4	75.3	6	10	6
2020-02-17	6	February	Monday	0	9.3	7	74.4	10	5	7
2020-02-15	10	February	Saturday	0	3.6	0	90.2	5	8	8
2020-02-14	5	February	Friday	0	6.4	3.6	67.8	8	6	6

Make specific sales forecast for Wholemeal Bread

1 day ago: 6

2 days ago: 6

1 week ago: 8

Weekday: Monday

Month: February

□ Holiday

Temperature: -30 8.4 35

Rain: 0 18.8 100

Sun: 0 84 500

```
Forecast:  5
```

Most similar days for Wholemeal Bread

Day	Sales	Month	Weekday	Holiday	Temp.	Rain	Sun	1 day ago	2 days ago	1 week ago
2018-02-17	7	February	Saturday	0	2.8	18	18.5	5	6	8
2017-11-25	8	November	Saturday	0	6.2	15.4	14.3	6	6	8
2016-11-05	8	November	Saturday	0	9.2	13	25.2	6	5	6

Figure 1. The Explanation Interface with its five panels.

Methodology

This observational online study uses a controlled environment to observe the behavior of 151 online study participants while making one-day-ahead sales forecasts for 14 products using an XI that provides ML forecasts and, depending on the group, different types of explanations approaches. The goal is to validate the effectiveness of two example-based explanation techniques, namely factual and counterfactual explanations, by emulating how actual employees within a bakery interact with the XI. To this end, we conducted a lab experiment based on a 2x2 between-subjects design with factual and counterfactual examples being the experimental factors (Figure 2). We focused on the effect of example-based explanations on the participants' decision quality which is measured as RMSE as in the situation at hand, the goal of measuring accuracy is to compare the accuracy across all products jointly at a given point in time for different configurations (Koutsandreas et al., 2022). The control group (G1) used an XI that only consists of the control panel (P1), the second panel (P2) with the ML forecast, and historical sales data (P3). The other groups additionally had access to counterfactual explanations only (G2), factual explanations only (G3), or both (G4). To validate the general meaningfulness of ML forecasts, an external validation group (GO) only had access to P1, P3, and a modified second panel. GO was used as external validation to ensure that participants perform better when given the AI forecasts to justify any effort in validating explanation techniques. For GO, P2 was missing the second column that displays the AI forecast. In the third column ("Order"), the default values are set to the real sales of one week ago for each product. This is in accordance with the functionality of the production planning software that was used by the bakeries of our interview study.

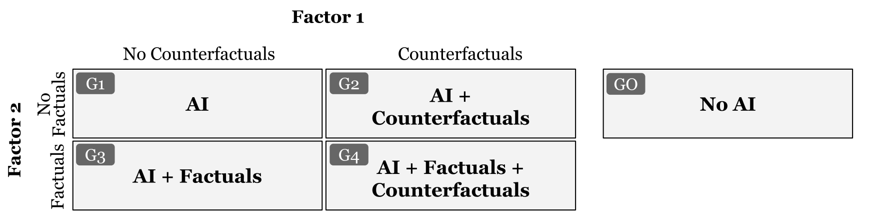


Figure 2. 2x2 between-subject experimental design

In line with existing research, participants were recruited from the online platform prolific (Palan & Schitter, 2018). Between 26 and 38 participants were recruited per condition (see Table 1), adding up to a total of 151 participants in all groups. Participants were randomly assigned to different groups. The participants' task for this experiment was to forecast the next business day's sales of 14 products using the XI. For that, the participants first got information on how to use the XI with its different panels. These instructions consisted of an explanatory video and multiple screenshots. To avoid confusion and leak of information, these instructions were tailored to the respective group the participants will be in. To test whether participants understood the task correctly, they were asked to describe the task in their own words. All participants understood the task sufficiently and were subsequently forwarded to the XI. They submitted their predictions through the XI, which saves the predictions to an online database for analysis. The quality of the participants' decisions is operationalized as RMSE and compared across groups. After using the XI and making the predictions, the participants were asked additional demographic questions as well as questions to measure the complexity of the task as an operationalization of the cognitive load posed on the user. For that, participants were asked to indicate the degree to which they agree with the statement "To make predictions, using the app was too complex." on a likert scale of 1 ("strongly disagree") to 7 ("strongly agree"). The perceived complexity of the task is an item of the proposed operationalization of cognitive decision effort by Pereira (2000) and positively related to the psychological costs of processing information. Using the outcome as a measure for the cognitive load posed upon the user (Hong et al., 2004), we examine the influence of cognitive load on the users' decision quality.

Results

We inspected the responses for outliers and particular response patterns to ensure that only valid responses were considered in the analysis (Rouse, 2015). More specifically, we checked that participants spent a reasonable amount of time (around 15 minutes) to make their predictions. This check led to the exclusion of 3 participants. 6 participants were excluded as their mean RMSE over all products was more than 3 standard deviations higher than the mean RMSE for the particular product and group. In total, we dropped 9 participants from our original set of 160 participants, leading to 151 participants for which results are reported. Based on 115 participants of the study, 50% were male, 47% female and 3% preferred not to say. 2% of the participants did not finish high school, 49% held a high school diploma, 44% a college or professional degree, and 4% a doctoral degree. 24% worked more than 40 hours, and 28% less than 40 hours per week. 47% were not employed out of which 49% were looking for work. 1% was retired. On average, the participants were 25.4 years old.

Effects of Example-Based Explanations on Forecast Accuracy

First, the external validation was successful: The group which did not receive any ML forecast (GO) performed significantly worse than the other groups. In general, all groups who received some type of explanation (G2, G3, G4) made better decisions in terms of forecast accuracy (RMSE) than the group who received ML forecasts but no explanations (G1). Table 3 shows the forecast accuracy for each group in descending order and the respective standard deviations (SD).

Group	**RMSE**	**SD**	**Complexity**	**SD**	**# Participants**
AI + Factuals (G3)	11.39	5.39	3.41	1.82	27
AI + Factuals + Counterfactuals (G4)	14.32	5.05	4.19	1.99	28
AI + Counterfactuals (G2)	14.46	2.15	3.96	1.90	29
AI (G1)	15.60	5.44	3.00	1.82	26
No AI (GO)	19.69	6.07	3.00	1.46	31

Table 1. Forecast Accuracy and perceived complexity per group

To check the significance of the differences in RMSE between groups, we performed an analysis of variance (ANOVA) between all pairs of groups. The results are depicted in Table 1, indicating that the prediction quality is significantly improved by including factual explanations in the XI (p-value < 0.01).

	AI + Factuals + Counterfactuals	AI	AI + Counterfactuals	AI + Factuals
No AI	.0002***	.011*	.0001***	.000***
AI + Factuals + Counterfactuals		.343	.889	.03*
AI			.32	.008**
AI + Counterfactuals				.008**

Table 2. p-values differences in decision quality (*:p<.05; **:p<.01; ***: p<0.001)

In summary, H1 is met: All groups where participants received ML forecasts performed significantly better than the group without help by the ML model. H2 is also met: G3 performed significantly better than the control group G1. However, H3 is not supported as we could not find a significant performance increase of participants when provided with counterfactual examples compared to the control group without explanations. Hypothesis H4b is accepted on a 5% significance level, as G3 performed significantly better than the control group. However, H4a is rejected. Counterfactual examples have not proven to be an effective approach to increase decision quality of lay users in our study.

Table 1 shows the complexity per group and the respective standard deviations. The full version of the XI (G4) is the most complex and significantly more complex than G1 (p-value: .02*) and GO (p-value: .01*). The difference between G2 and G3 is not significant as both were perceived similarly complex. G2 is perceived as marginally more complex than the control group G1 (p-value: .08). However, the perceived complexity of the XI with factual examples is not significantly higher than that of the control group (G1). The results of the perceived complexity are intuitive: The presence of ML advice does not lead to increased perceived complexity. When explanation approaches are implemented individually, the participants perceived the XI to be more complex.

Discussion, Contribution and Next Steps

In this study, we evaluated the effectiveness of factual and counterfactual examples as explanation approaches. To this end, we conducted an observational online study, in which we observed 151 online study participants while making one-day-ahead predictions for 14 products using an XI that provides ML forecasts. We focused on the effect of example-based explanations on the participant's decision quality measured. Depending on the group, participants received factual examples, counterfactual examples, or both. A control group only received ML forecasts and historical sales, but no explanations.

Our results indicate that factual examples are effective to increase decision quality in ML-aided sales forecasting. The forecast accuracy of participants was significantly higher when given factual examples compared to the control group. Counterfactual examples did not significantly increase accuracy study. Based on our results, it is yet unclear whether the counterfactual examples are generally ineffective or just poorly executed. Consequently, we aim to implement more sophisticated counterfactual explanations. Dandl et al. (2020) propose multi-objective counterfactual explanations, which we will test in the next study. We will also implement more sophisticated factual explanations, specifically the most influential instances in the past as proposed by Koh and Liang (2017) and influential time series characteristics (shapelets) as proposed by Ye and Keogh (2009). To explore the cognitive mechanism through which example-based explanations foster decision quality, we measured the cognitive load that was imposed on the users while using the XI and making predictions. The group that received both factual and counterfactual examples experienced the highest cognitive load. This might be a reason for the lower forecast accuracy of participants from this group compared to the group that only received factual examples. To further examine the boundary conditions in which XI may improve decision quality, we aim to focus on the cognitive processes influencing the decision quality of users in a third study. To validate the results of the online experiment, a fourth study will cover a field test with practitioners from the domain.

We contribute to theory by evaluating the effectiveness of two selected example-based explanation approaches (factual and counterfactual examples) in sales forecasting (Dandl et al., 2020; Ates, 2021) and investigating the cognitive load that is imposed on the users by the different explanation approaches (Miller, 2019). For practice, we aim to provide guidance for building XIs for lay users in sales forecasting by providing insights into which explanation approaches are most effective to enable decision makers to make better sales forecast decisions.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y. and Kankanhalli, M. 2018. "Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda.," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, New York, NY, pp. 1–18.

Adadi, A. and Berrada, M. 2018. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)", *IEEE Access*, 6, pp. 52138–52160.

Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y. and Kantarcioglu M. 2021. "Does Explainable Artificial Intelligence Improve Human Decision-Making?," in *Proceedings of the AAAI Conference on Artificial Intelligence* 35 (8), pp. 6618–6626.

Apley, D. W. and Zhu, J. 2020. "Visualizing the effects of predictor variables in black box supervised learning models." *Journal of the Royal Statistical Society. Series B: Statistical methodology* (82:4), pp. 1059–1086.

Ates, E., Aksar, B., Leung, V. J. and Coskun, A. K. 2021. "Counterfactual explanations for multivariate time series". in *2021 International Conference on Applied Artificial Intelligence (ICAPAI)*. IEEE.

Blohm, I., Antretter, T., Sirén, C., Grichnik, D. and Wincent, J. 2020. "It's a Peoples Game, Isn't It?! A Comparison Between the Investment Returns of Business Angels and Machine Learning Algorithms," *Entrepreneurship Theory and Practice*.

Blohm I., Riedl, C., Füller, J. and Leimeister, J. M. 2016. "Rate or Trade? Identifying Winning Ideas in Open Idea Sourcing," *Information Systems Research* (27:1), pp. 27–48.

Burton, J. W., Stein, M. K. and Jensen, T. B. 2020. "A systematic review of algorithm aversion in augmented decision making," *Journal of behavioral decision making* (33:2), pp. 220–239.

Cam, A., Chui, M. and Hall, B. 2019. "Global AI Survey: AI proves its worth, but few scale impact." URL: http://dln.jaipuria.ac.in:8080/jspui/bitstream/123456789/1323/1/Global-AI-Survey-AI-proves-its-worth-but-few-scale-impact.pdf (visited on 16. April 2022).

Castelo, N., Bos, M. W. and Lehmann, D. R. 2019. "Task-Dependent Algorithm Aversion," *Journal of marketing research* (56:5), pp. 809–825.

Chen, T. and Guestrin, C. 2016. "XGBoost: A Scalable Tree Boosting System.," un: *arXiv [cs.LG]*. Available at: http://arxiv.org/abs/1603.02754.

Cheng, H. F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M. and Zhu, H. 2019. "Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders.," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Glasgow, Scotland, UK.

Dandl, S., Molnar, C., Binder, M., Bischl, B. 2020. "Multi-Objective Counterfactual Explanations." *Parallel Problem Solving from Nature – PPSN XVI*. Springer International Publishing, pp. 448–469.

Dellermann, D., Ebel, P., Söllner, M., and Leimeister, J. M. 2019. "Hybrid intelligence.," *Business & Information Systems Engineering* (61:5), pp. 637-643.

Dietvorst, B. J., Simmons, J. P., Massey, C. 2015. "Algorithm aversion: people erroneously avoid algorithms after seeing them err.," *Journal of Experimental Psychology: General* (144:1), pp. 114–126.

Engel, C., Ebel, P., and Leimeister, J. M. 2022. "Cognitive Automation," *Electronic Markets*.

Geva, T. and Saar-Tsechansky, M. 2021 "Who is a better decision maker? Data-driven expert ranking under unobserved quality.," *Production and operations management* (30:1), 127–144.

Grove, W. M. et al. 2000. "Clinical versus mechanical prediction: a meta-analysis.," *Psychological assessment* (12:1), pp. 19–30.

Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., and Turini, F. 2019. "Factual and Counterfactual Explanations for Black Box Decision Making," *IEEE Intelligent Systems* (34:6), pp. 14–23.

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., Giannotti, F. 2018. "Local Rule-Based Explanations of Black Box Decision Systems.," in *arXiv [cs.AI]*. Available at: http://arxiv.org/abs/1805.10820.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.-Z. 2019. "XAI-Explainable artificial intelligence.," *Science robotics* (4:37).

Hagras, H. 2018. "Toward Human-Understandable, Explainable AI." *Computer* (51:9), pp. 28–36.

Herm, L. V., Wanner, J., Seubert, F. and Janiesch, C. 2021. "I Don't Get It, but It Seems Valid! The Connection Between Explainability and Comprehensibility in (X)AI Research.," in *Proceedings of the 2021 European Conference on Information Systems*.

Hong, W., Thong, J. Y. L., and Tam, K. Y. 2004. "The Effects of Information Format and Shopping Task on Consumers' Online Shopping Behavior: A Cognitive Fit Perspective," *Journal of Management Information Systems* (21:3), pp. 149–184.

Kenny, E. M., and Keane, M. T. 2021. "On Generating Plausible Counterfactual and Semi-Factual Explanations for Deep Learning," *Proceedings of the AAAI Conference on Artificial Intelligence* (35:13), pp. 11575–11585.

Kelleher, J. D., Namee, B. M. and D'Arcy, A. 2015. "Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies.," The MIT Press.

Koh, P. W., and Liang, P. 2017. "Understanding black-box predictions via influence functions.," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, Sydney, Australia.

Koutsandreas, D., Spiliotis, E., Petropoulos, F., and Assimakopoulos, V. 2022. "On the Selection of Forecasting Accuracy Measures," *The Journal of the Operational Research Society* (73:5), pp. 937–954.

Liao, Q. V., Gruen, D. and Miller, S. 2020. "Questioning the AI: Informing Design Practices for Explainable AI User Experiences.," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

MacQueen, J. 1967. "Some methods for classification and analysis of multivariate observations.," *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* (1:14), pp. 281–297.

Mentzer, J. T. and Bienstock, C. C. 1998. *Sales Forecasting Management: Understanding the Techniques, Systems and Management of the Sales Forecasting Process*. SAGE Publications.

Miller, T. 2019. "Explanation in Artificial Intelligence: Insights from the Social Sciences," *Artificial Intelligence* (267), pp. 1–38.

Mohseni, S., Zarei, N. and Ragan, E. D. 2020. "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems.," in *arXiv*. Available at: http://arxiv.org/abs/1811.11839.

Palan, S. and Schitter, C. 2018. "Prolific.ac—A subject pool for online experiments.," *Journal of Behavioral and Experimental Finance* (17), pp. 22–27.

Pereira, R. E. 2000. "Optimizing Human-Computer Interaction for the Electronic Commerce Environment," *Journal of Electronic Commerce Research* (1:1), pp. 23–44.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. 2021. "Manipulating and measuring model interpretability," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA.

Prahl, A. and Van Swol, L. 2017. "Understanding Algorithm Aversion: When Is Advice from Automation Discounted?," *Journal of Forecasting* (36:6), pp. 691–702.

Rojat, T., Puget, R., Filliat, D., Del Ser, J., Gelin, R. and Díaz-Rodríguez, N. 2021. "Explainable Artificial Intelligence (XAI) on Time Series Data: A Survey." in: *arXiv [cs.LG]*. Available at: http://arxiv.org/abs/2104.00950.

Rouse, S. V. 2015. "A Reliability Analysis of Mechanical Turk Data," *Computers in Human Behavior* (43), pp. 304-307.

SBC 2019. "Branchenspiegel für das Bäckerei-, Konditorei- und Confiserie-Gewerbe. Schweizerischer Bäcker-Confiseurmeister-Verband (SBC)." Available at: https://www.swissbaker.ch/inhalte/03_dokumente/verband/de/sbc_branchenspiegel_2019_d.pdf.

Sun, ZL., Choi, T. M., Au, K. F. and Yu, Y. 2008. "Sales forecasting using extreme learning machine with applications in fashion retailing.," *Decision support systems* (46:1), pp. 411–419.

Sweller, J. 1988. "Cognitive load during problem solving: Effects on learning.," *Cognitive science* (12:2), pp. 257–285.

Wang, D., Yang, Q., Abdul, A. and Lim, B. Y. 2019. "Designing theory-driven user-centric explainable AI," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

Ye, L., and Keogh, E. 2009. "Time series shapelets: a new primitive for data mining.," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Zhang, Y., Liao, Q. V. and Bellamy, R. K. E. 2020. "Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making.," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.