Seeing is Believing: The Role of Scatterplots in Recommender System Trust and Decision-Making

Bhavana Doppalapudi[1], Md Dilshadur Rahman[2], and Paul Rosen[2]

[1] University of South Florida, Tampa, FL 33620, USA
bhavanadoppalapudi@gmail.com
[2] University of Utah, Salt Lake City, UT 84112, USA
{dilshadur,prosen}@sci.utah.edu

Abstract. The accuracy of recommender systems influences their trust and decision-making when using them. Providing additional information, such as visualizations, offers context that would otherwise be lacking. However, the role of visualizations in influencing trust and decisions with recommender systems is under-explored. To bridge this gap, we conducted a two-part human-subject experiment to investigate the impact of scatterplots on recommender system decisions. Our first study focuses on high-level decisions, such as selecting which recommender system to use. The second study focuses on low-level decisions, such as agreeing or disagreeing with a specific recommendation. Our results show scatterplots accompanied by higher levels of accuracy influence decisions and that participants tended to trust the recommendations more when scatterplots were accompanied by descriptive accuracy (e.g., *high*, *medium*, or *low*) instead of numeric accuracy (e.g., *90%*). Furthermore, we observed scatterplots often assisted participants in validating their decisions. Based on the results, we believe that scatterplots and visualizations, in general, can aid in making informed decisions, validating decisions, and building trust in recommendation systems.

Keywords: Machine Learning · Visualizations · Trust · Decision-Making.

1 Introduction

Recommender systems are ubiquitous in diverse contexts, from low-risk activities like traffic prediction [20] and movie suggestions [13] to high-risk scenarios like delivering child care [7] and homeless services [22]. The widespread use of recommender systems highlights the need to understand the factors influencing users' trust in these systems. Significant research has been dedicated to studying and improving trust by enhancing transparency and providing information such as accuracy and explanations for recommendations [21, 23, 24, 39].

While providing accuracy offers insights into recommender system performance, research often overlooks the nuanced meaning of accuracy, which is context-specific (e.g., due to class imbalance, classifier type, i.e., binary vs. multiclass, etc.), and expecting the general population to possess this knowledge is

unreasonable. A simple scenario is to consider a condition that is observed only 10% of the time. A recommender system that always reports false would still exhibit a 90% accuracy! While experts possess the knowledge to employ additional metrics to illustrate this imbalance, the same cannot be expected from the general population, highlighting the need to investigate the effects of providing additional context, such as visualizations, to assist people in decision-making.

The use of visualizations in understanding people's trust in recommender systems has been relatively underexplored. Most of the current research focuses on interactive visualizations and tools [1, 8, 28] to explore, analyze, and validate the recommendations and performance of recommender systems. While offering tools for analysis and exploration proves advantageous for professionals and researchers, translating these to the general population seems unreasonable. Furthermore, there exists a lack of focus on how including simple visualizations that depict data and offer extra contextual information regarding the features utilized by the recommender system supports trust and decision-making.

We describe the findings of a crowdsourced study on how the inclusion of visualization, scatterplots in particular, with accuracy, shapes people's decisions and trust in recommender systems. Our first experimental task requires participants to consider the presented scatterplots along with accuracy and make a decision on selecting 1 of the 2 provided recommender systems, with the selection of 1 over the other serving as a proxy for trust. The second experimental task requires participants to consider the presented scatterplot and decide whether to agree or disagree with a recommendation made by the recommender system. Through these tasks, we focus on (1) how scatterplots presented with different accuracy influence people's decisions and (2) how scatterplots accompanied by accuracy presented in different formats impact trust.

Our findings show that scatterplots paired with accuracy do influence people's decisions and trust in recommendations. In addition, we found scatterplots accompanying descriptive accuracy (e.g., 'high,' 'low,' etc.) induced more trust. Finally, we observed that scatterplots not only assisted people in making decisions but also invalidating them. Taken together, the results provide new insights into how scatterplots assist in decision-making and trust in recommendations.

2 Related Works

Trust and Transparency in Recommendations. It has been shown that people prefer human over algorithmic advice even if the algorithm performs better [9]. Additionally, mistakes made by automated systems result in a quicker loss of confidence than those made by humans [25]. However, transparent and open explanations positively influence decision-making processes with machine learning [24], and precisely calibrating to transparency can significantly enhance user trust and the perceived quality of recommender systems [21, 23]. Various studies have investigated the role of transparency in decisions with machine learning models [29,32,38], but, to the best of our knowledge, the role of visualization, particularly scatterplots, in providing transparency has not been thoroughly studied.

Visualizations in Machine Learning. Chatzimparmpas et al. present a state-of-the-art report on enhancing trust in machine learning models through interactive visualizations and outline research opportunities in their surveys [5, 6]. Visualizations are crucial for deciphering black-box models by identifying key features and decoding complex behaviors. Tools such as the What If Tool [35] and SliceTeller [40] provide insights into model performance under various scenarios and identify data segments causing inaccuracies. Advances in interactive visualization techniques, such as RuleMatrix [26], EnsembleMatrix [33], and deep learning interpretation systems [19,34,36], have improved the comprehension, exploration, and validation of predictive models. Tools such as ModelTracker [2] and Squares [28] refine error analysis and debugging processes through interactive visualizations that display summary statistics and instance-level performance metrics. Additionally, ExplainExplore [8] and Melody [4] analyze and summarize explanations through interactive systems for better understanding. Tools such as Fairkit-learn [18], FAIRVIS [3], and FairSight [1] help data scientists assess fairness, offering methods for subgroup analysis and fairness evaluation in algorithmic decisions. DiscriLens [34], D-BIAS [15], and Visual Auditor [27] focus on uncovering bias, incorporating human insights, and providing analytical views on fairness issues. These efforts highlight a trend toward leveraging visual analytics for improved fairness and bias mitigation in machine learning, with less emphasis on improving general user trust in recommendations.

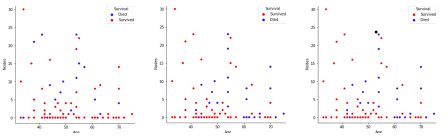
Visualization and Decision-Making. Visualizations guide users in making informed decisions [30, 41], communicate critical information [10, 30], aid in predictions with uncertainty [17], and help identify complex patterns [31]. Dimara et al. proposed an agenda to enhance decision-making through visualization research [11]. Limited work employs visualizations to investigate trust in recommendation system decision-making. Xiong et al. used map-based visualizations to study transparency and trust through factors like accuracy, clarity, disclosure, and thoroughness [37]. A study that closely relates to ours investigates how visual design impacts the perception of model bias, trust, and adoption [14]. Findings reveal that visualization design choices, such as explaining fairness and mentioning bias, significantly affect trust levels. While both studies use visualizations and examine model performance metrics like accuracy, our study investigates the influence of scatterplots and accuracy on trust and decision-making, whereas their study focuses on identifying fair models using various signals and visualization methods, making these studies complementary.

3 Study

We now describe our study of how scatterplots depicting recommender system predictions influence general users' trust and decision-making.

Procedure. The IRB-approved study was run on Amazon Mechanical Turk (AMT) using custom web pages and server built with Python. First, participants received an overview of the study's goals and provided consent. After demographic questions, they completed tutorial questions with a sample dataset

to avoid learning effects. Participants then engaged in 2 types of task: (1) high-level decision-making tasks (see sec. 4) and (2) low-level decision-making tasks (see sec. 5). An example experiment is at `https://shorturl.at/xEQR7`.

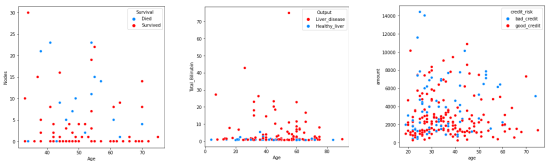


(a) (b) (c)

Fig. 1: Experimental plots: (a) reference plot of training data, (b) recommender plot of testing data, and (c) testing data with single recommendation in black.

Visualizations. For all datasets, the 2 continuous attributes with the strongest correlation to the predicted value were selected to visualize. To reduce color bias, we generated the visualizations with 4 distinct color set orientations: Red/Blue, Blue/Red (see fig. 2), Purple/Dogerblue, and Dogerblue/Purple (see fig. 4). Experiments used 3 types of scatterplots: (1) reference scatterplots (i.e., ground truth from training data), where color represents the class data points actually belong to (see fig. 1a); (2) recommender scatterplots, where the color of the data points represents the class the system predicted they belong to (see fig. 1b); and (3) single recommendation scatterplots, where recommender scatterplots had an extra data point, depicted in black, to represent the specific data point for which the recommendation was being made (see fig. 1c).

Recommender Models. We did not want to compare multiple models, but we did want to provide diversity in the recommendations. Therefore, we used 4 popular binary classification models: Decision Tree, Random Forest, K-Nearest Neighbors, and Support Vector Machine. As is the case with most real recommender systems, participants were not told which model they were being presented with during the experimental tasks. All models were built by randomly partitioning each dataset into training and test sets at a 75% to 25% ratio, respectively. Each resulted in similar levels of accuracy that are reported at `https://shorturl.at/xEQR7`.

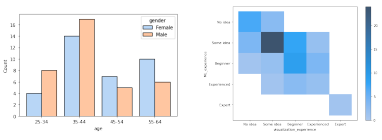


(a) Haberman (b) Liver Disease (c) South German

Fig. 2: Datasets

Data Sets. For our study, we used 3 datasets acquired from the UCI Machine Learning Repository [12,16]. The datasets were chosen to showcase some variety in data densities while avoiding complications of overdraw. *Haberman's Survival* (see fig. 2a) contains 306 instances of 3 attributes and deals with the survival of patients 5 years after undergoing surgery for cancer. *Indian Liver Patient* (see fig. 2b) comprises 583 instances of 10 attributes about the health of individuals

with and without liver disease. *South German Credit* (see fig. 2c) consists of 1000 instances of 20 attributes with the credit quality of customers for a requested loan. Details of feature selection and feature visualizations are reported in the supplement (see `https://shorturl.at/xEQR7`).



(a) Demographics (b) Background

Fig. 3: (a) Age and gender (female/blue and male/orange) and (b) experience in machine learning (vert.) and visualization (horiz.).

Participants. 120 subjects were paid $2 to participate in the study, which lasted about 15 minutes. We screened the participants for their location and only those in the US and Canada were considered for the study. We analyzed the data of 71 participants (35 female and 36 male) who passed the following inclusion criteria: (1) answered all the experimental tasks without any omissions; and (2) provided a correct response to at least 1 of 2 attention-checks. Demographics and self-reported experience with visualization and machine learning can be found in fig. 3.

4 Task 1: Trust in Recommender Systems

The primary aim of these tasks was to investigate the impact of scatterplots presented with the accuracy information on individuals' preferences for selecting a recommender system.

4.1 Procedure

Each participant was presented with 3 visualizations: a reference scatterplot and 2 recommender scatterplots (see sec. 3). Their objective was to select a recommender system from the 2 available options. We designed 3 versions of the task. Each participant was presented with all 3 versions, making it a within-subject stimuli. The variations were:

- The *Visualization-Only* version presented a reference scatterplot and 2 recommender scatterplot visualizations (example in supplement).
- The *Visualization + Numeric Accuracy* version presented a reference scatterplot with 2 recommender scatterplots, each accompanied by numeric accuracy (e.g., *91%*) for each of the recommender scatterplots (see fig. 4).
- The *Visualization + Descriptive Accuracy* version followed a similar setup, with a reference scatterplot and 2 recommender scatterplots, each accompanied by descriptive accuracy labels (e.g., *high*, *medium*, or *low*), instead of numeric accuracy values (example in supplement).

Recommender Systems Used. We employed all 4 recommender system models (see sec. 3). For every question, 2 models were chosen at random.

Accuracy. In the real-world, accuracy is often provided without significant context. We intentionally avoided providing participants with an explanation of

what the accuracy actually meant to determine better how participants naturally perceived it. Further, we opted to ignore the model output accuracy and instead systematically generated an accuracy for each stimulus. For binary classifiers like those we used, 50% accuracy indicates that the algorithm is essentially guessing. Using this value as a lower bound, we categorized accuracy into 3 tiers, aligning descriptive accuracy with numeric accuracy: *high* $\rightarrow [90\%, 95\%]$; *medium* $\rightarrow [75\%, 80\%]$; and *low* $\rightarrow [65\%, 70\%]$. The precise numeric accuracy displayed for each question was chosen at random from within the designated range.

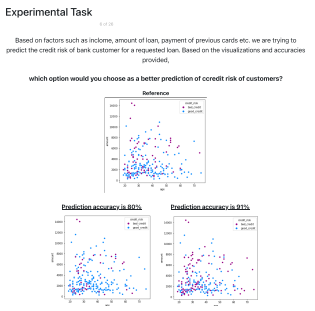


Fig. 4: *Visualization + Numeric Accuracy* version of the experiment 1 tasks.

Color. Participants were randomly assigned to 1 of the 4 distinct color treatments (see sec. 3), and all of their stimuli used that color combination.

Tasks. Each participant was assigned 9 tasks in random order, encompassing the 3 experimental conditions (i.e., Visualization-Only, Visualization + Numeric Accuracy, and Visualization + Descriptive Accuracy) across the 3 datasets (i.e., Haberman's Survival, Indian Liver Patient, and South German Credit).

4.2 Results

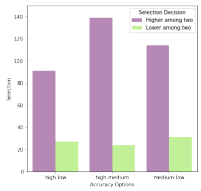
We first evaluated whether there was any potential association (i.e., bias) for recommender system selection and task versions. We conducted a 2-way ANOVA that showed no statistically significant interaction ($F(6,24) = 2.43$, $p=.057$), suggesting that participants did not have a preference for any specific recommender model but rather made decisions based on the information provided.

Hypothesis 1 [H1]: *Scatterplots accompanied by higher accuracy correspond to a greater chance of selection compared to those accompanied by lower accuracy.*

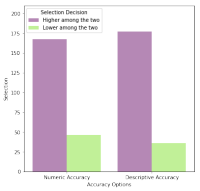
We investigated the influence of scatterplots accompanied by higher accuracy (both numeric and descriptive) on the selection of recommender systems, as compared to scatterplots accompanied by lower accuracy. We used Shapiro-Wilk test to check for normality, and the results indicated the data deviated from a normal distribution ($W=0.813$, $p<.001$). Thus, we conducted a Kruskal-Wallis non-parametric test, and the results of the test indicate that *scatterplot visualizations presented with higher accuracy significantly impact the selection of recommender systems* ($H(1, n=142)=161.19$, $p<.001$). For the post hoc analysis, we performed a Dunn's test and compared the mean ranks. The results indicated a statistical significance for the selection of scatterplots that are accompanied by higher accuracy ($Z(1, n=142)=12.69$, $p<.001$). The data was visualized to show the effects (see fig. 5a). These findings are consistent with our hypothesis, leading us to reject the null hypothesis and **accept H1**.

Hypothesis 2 [H2]: *Scatterplots accompanied by numeric accuracy will have less influence on selection than those accompanied by descriptive accuracy.*

For the subsequent analysis, we examined whether the method of presenting accuracy, specifically Numeric Accuracy (e.g., *90%*) versus Descriptive Accuracy (e.g., *high*), alongside scatterplot visualizations influenced the selection of recommender systems. We used the Shapiro-Wilk test to check for normality, and the results indicated the data deviated from a normal distribution (W=0.698, p<.001). Due to the non-normal distribution of data, we conducted a Wilcoxon test. The results revealed that the *method of presenting accuracy metrics alongside scatterplot visualizations did not exhibit a statistically significant association with the selection of recommender systems* (W=276, p=.189) (see fig. 5b). Therefore, we are unable to reject the null hypothesis and, therefore, **reject H2**.



(a) High, Medium, and Low Accuracy



(b) Numeric and Descriptive Accuracy

Fig. 5: Selection of recommender systems: (a) presents the proportion of time the recommender with higher accuracy was selected, while (b) presents the selection of recommender systems for numeric and descriptive accuracy conditions.

Discussion. **H1** validated that accuracy influences trust and decisions regarding the recommender system. On the other hand, we expected that the ambiguity associated with numeric accuracy would cause participants to "trust the visualization" less, as opposed to the less ambiguous descriptive accuracy. However, as **H2** was rejected, we cannot support that claim in the situation.

5 Task 2: Trust in Individual Recommendations

Building on the previous task, this task focuses on examining the way scatterplots, paired with accuracy, impact participants' trust in individual recommendations generated by the recommender systems.

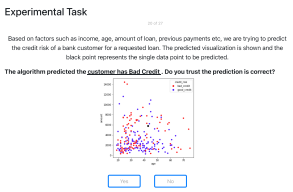
5.1 Procedure

Participants were presented with information about the dataset, including what the recommender system is recommending in a single recommendation scatterplot (see fig. 1c) coupled with the recommender systems' recommendation for the data point. They were then required to indicate if they would agree with the recommendation. The decision to agree with the recommendation was considered a proxy for trust in the recommendation. Additionally, in order to investigate the behavior of the participant's decision when additional information was given, we altered the sequence in which the information was provided.

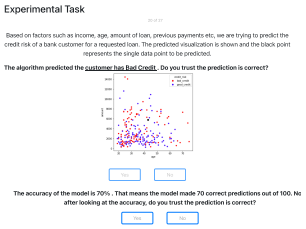
– The *Visualization + Accuracy* condition provided all information at once and asked for agreement with the recommendation (example in supplement).

– The *Visualization* → *Accuracy* condition first showed the recommendation and the single recommendation scatterplot and asked for agreement with the recommendation (see fig. 6a). Subsequently, accuracy information was provided, and participants were asked if they still agreed with the recommendation (see fig. 6b).
– The *Accuracy* → *Visualization* condition initially presented the recommendation and the accuracy and asked for agreement with it. Subsequently, the single recommendation scatterplot was shown, and participants were asked if they still agreed with the recommendation (example in supplement).

Participants were randomly assigned to 1 of 12 distinct treatments in a *4 * 3* design. As with Task 1, participants were randomly assigned to 1 of the 4 distinct color treatments (see sec. 3). All of their stimuli used that color combination, making it a between-subject stimuli. The task type (i.e., *Visualization + Accuracy*, *Visualization* → *Accuracy*, and *Accuracy* → *Visualization*) was the second between subject stimuli. On the other hand, participants were presented with both the numeric (e.g., *91%*) and descriptive (e.g., *high*) conditions, making it a within-subject stimuli. Other variables, such as recommender systems' recommendation and the accuracy shown, were randomly assigned to each participant for each task.



(a) First part of the task



(b) Second part of the task

Fig. 6: *Visualization* → *Numeric Accuracy* version of task 2. (a) illustrates the initial decision of the task, while (b) the second step provides an opportunity to reconsider their decision.

Task. Participants were given information about the dataset, including what the recommendation system is recommending, a single recommendation scatterplot, and the systems' recommendation of the data point and the accuracy. They were asked to record their agreement with the recommendation by selecting either *Yes* or *No* (see fig. 6a). For the *Visualization* → *Accuracy* and *Accuracy* → *Visualization* conditions, they were then shown additional information and asked the same question again (see fig. 6b). Each participant responded to a total of 6 tasks—3 datasets and 2 types of accuracy.

Recommender Systems and Recommendations. We employed all 4 recommenders (see sec. 3). 3 data points, excluded from both training and test sets, were chosen from each dataset for recommendation. All 4 recommender systems yielded identical recommendations for the data points. We presented participants with either the actual recommendations made by the system or the inverted recommendations to balance the behavior toward stimuli and avoid bias. The recommendations were randomly assigned to participants, meaning while

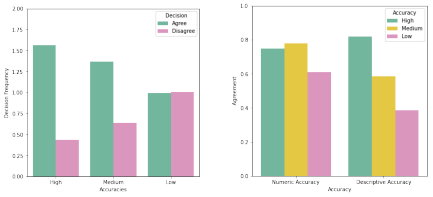
a participant was presented with the actual recommendation for 1 task, they might encounter the incorrect recommendation for the subsequent task.

Accuracy. To assess the impact of varying accuracy, we ignored the output accuracy and systematically generated accuracy similar to Task 1.

5.2 Results

Hypothesis 3 [H3]: *Scatterplots accompanied by higher accuracy correspond to a greater chance of trusting the recommendation compared to scatterplots accompanied by lower accuracy.*

We filtered the full data to test this hypothesis. Our initial analysis involved examining if the data satisfy the assumptions of a 1-way ANOVA. The results from the Shapiro-Wilk test (W=0.665, p<0.001) indicated that data deviated significantly from a normal distribution. This led us to choose non parametric Kruskal-Wallis test. The Kruskal-Wallis analysis (H(2, n=280)=15.352, p<.001) indicated the difference was statistically significant. This supported that *scatterplots accompanied by higher accuracy correspond to a greater chance of building trust in the recommendation compared to scatterplots accompanied by lower accuracy.* The visualization of the corresponding data also supports this finding (see fig. 7a). This led us to reject the null hypothesis and **accept H3**.



(a) Level of accuracy (b) Num. or Desc.

Fig. 7: Agreement with the recommendation: (a) across high, medium, and low accuracy and (b) when accuracy is presented as Numeric or Descriptive.

Hypothesis 4 [H4]: *Scatterplots accompanied by Descriptive Accuracy correspond to a greater chance of trusting the recommendation compared to scatterplots accompanied by Numeric Accuracy.*

We again filtered the data needed to test this hypothesis from the full data. As each participant was exposed to both scatterplots along with Numeric Accuracy and scatterplots along with Descriptive Accuracy an equal number of times, we decided to use a Paired Samples T-test. The Shapiro-Wilk test was used to test data normality, and the results (W=0.931, p<.001) indicated the distribution of data deviated from normality. We selected a non parametric alternative, Wilcoxon sign test on paired samples to test our hypothesis. The results indicated (V=419, p=.008) that the median *trust of scatterplots accompanied by Numeric Accuracy was less than that of scatterplots accompanied by Descriptive Accuracy.* The visualizations from the data also reflected a similar relation (see fig. 7b). This led us to reject null hypothesis and **accept H4**.

Hypothesis 5 [H5]: *The order of presentation of scatterplots and accuracy of information influence the participant's trust in the recommendation.*

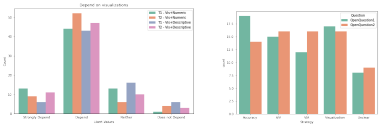
We utilized data for *Visualization + Accuracy*, *Visualization → Accuracy* and *Accuracy → Visualization* stimuli. We used Shapiro-Wilk to test for normality, and the results (W=0.852, p<.001) implied data deviated from a normal distribution. Therefore, we avoided 1-way ANOVA and instead used a non-parametric alternative, Kruskal-Wallis test. The Kruskal-Wallis test results (H (2, n=142)=1.005, p=.604) implied that *no strong evidence exists to suggest that the order of presentation of scatterplots and accuracy influenced decisions.* Based on the results, we could not reject the null hypothesis and therefore **reject H5**.

Discussion. In this task, **H3** showed that scatterplots accompanied by higher accuracy assist in making trustworthy decisions about a recommendation. Furthermore, unlike the previous experimental task, **H4** validated that scatterplots accompanied by subjective information like Descriptive Accuracy aid in instilling trust in a recommendation compared to Numeric Accuracy. Although we expected the order of presenting information to sway the trust in recommendations, we did not find evidence of it in our experiment.

6 Qualitative Tasks

We evaluated participants' self-reported dependence on scatterplots and examine the decision-making strategies they used.

Role of Scatterplots. At the end of each experimental task version, participants recorded their level of dependence on scatterplots using a 5-point Likert Scale ranging from 'Strongly Depend' to 'Strongly Does not Depend'. Our analysis revealed that over 75% of participants indicated a significant level of dependence on the visualization, with many reporting either 'Strongly Depend' or 'Depend' on scatterplots for decision-making (see fig. 8a).



(a) Dependency levels (b) Strategies

Fig. 8: (a) Self-reported dependence on scatterplots for task 1 (T1) and task 2 (T2). (b) The outcomes derived from the post-experiment questionnaire, where `OpenQuestion1` was for recommender systems, and `OpenQuestion2` was for individual recommendations (AV: Accuracy-Validation, VA: Visualization-Validation).

Post Experiment Questionnaire. After the experimental tasks, participants were asked 2 open-ended questions. The first prompted them to record the approach employed in deciding to pick a recommender system when a scatterplot was coupled with accuracy (`OpenQuestion1`), while the second asked about the approach utilized in determining whether or not to trust an individual recommendation (`OpenQuestion2`). The responses were reviewed and organized into 5 categories based on their decision strategies. (1) *Visualization-Based* included participants

who expressed that decisions were based primarily on scatterplots. (2) *Accuracy-Based* was assigned to participants who indicated accuracy as the primary determinant of their decisions. (3) *Visualization-Validation* was when the decisions were initially made using scatterplots and subsequently validated based on accuracy. (4) *Accuracy-Validation* first relied on accuracy to make decisions and then validated the decisions based on scatterplots. (5) Participants whose decision-making strategy lacked specificity, often involving factors such as gut feeling and instinct, were categorized as *Unclear*. The results suggest that when faced with the choice between 2 recommender systems, participants tended to give slightly higher priority to accuracy over scatterplots (see fig. 8b). However, our analysis revealed that when making a decision regarding their trust in individual recommendations, scatterplots played a slightly more significant role in influencing decisions.

7 Discussion and Conclusion

Discussion. We analyzed the way scatterplots presented, along with accuracy, influence trust in recommender systems. We observed that participants relied on the accuracy provided with the scatterplot to make decisions, and higher accuracy positively influenced both deciding on selecting a recommender system and trusting an individual recommendation. While the presentation of accuracy accompanying scatterplots as numeric and descriptive did not show a statistically significant difference in selecting a recommender system, Descriptive Accuracy exhibited a notable impact on trust in individual recommendations. We believe the lack of influence on the selection of a recommender system could be due to the multiway comparison of the scatterplot and accuracy of both recommender systems. Moreover, the findings also revealed that participants utilized scatterplots to validate their decisions when accuracy was higher, but they particularly relied on scatterplots to make decisions as accuracy values dropped.

Limitations and Future Work. While we primarily focused on scatterplot visualizations in our study, we plan to extend our research to examine the influence of additional visualization types on trust and decision-making. We visualized the relation between 2 attributes and aimed to explore how visualizing multiple attributes affects people's decisions on recommendations. We recognize that real-world datasets are often large, which can clutter visualizations and become difficult to interpret. Finally, we acknowledge the subjective nature of trust and the challenges in measuring it. In our work, we primarily consider relative trust by focusing on participants' selection of one option over another.

Conclusion. The inspiration behind this research was to explore and comprehend the role of visualizations, especially scatterplots presented alongside accuracy, in influencing trust in recommender systems and individual recommendations. Our observations show participants did depend on scatterplots for making decisions and validating them. In particular, participants tended to trust high accuracy alone, but visualizations became increasingly more important with lower accuracy. We also saw that the presentation of the accuracy in a descrip-

tive form, although vague, was easier to interpret than the numeric form, which also has contextualized meaning. Based on our findings, we advocate for incorporating visualizations to enhance trust and decision-making in recommender systems. This approach helps mitigate the issues that arise from relying solely on accuracy metrics, thereby increasing the ability of the general population to make more informed decisions.

References

1. Ahn, Y., Lin, Y.R.: Fairsight: Visual analytics for fairness in decision making. IEEE TVCG **26**(1), 1086–1095 (2019)
2. Amershi, S., Chickering, M., Drucker, S.M., Lee, B., Simard, P., Suh, J.: Modeltracker: Redesigning performance analysis tools for machine learning. In: ACM CHI. pp. 337–346 (2015)
3. Cabrera, Á., Epperson, W., et al.: Fairvis: Visual analytics for discovering intersectional bias in machine learning. In: IEEE VAST. pp. 46–56 (2019)
4. Chan, G.Y.Y., Bertini, E., Nonato, L.G., Barr, B., Silva, C.T.: Melody: Generating and visualizing machine learning model summary to understand data and classifiers together. arXiv (2020)
5. Chatzimparmpas, A., Martins, R.M., Jusufi, I., Kerren, A.: A survey of surveys on the use of visualization for interpreting machine learning models. Information Visualization **19**(3), 207–233 (2020)
6. Chatzimparmpas, A., Martins, R.M., Jusufi, I., Kucher, K., Rossi, F., Kerren, A.: The state of the art in enhancing trust in machine learning models with the use of visualizations. Computer Graphics Forum **39**(3), 713–756 (2020)
7. Chouldechova, A., Benavides-Prado, D., Fialko, O., Vaithianathan, R.: A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In: ACM FAccT. pp. 134–148 (2018)
8. Collaris, D., van Wijk, J.J.: Explainexplore: Visual exploration of machine learning explanations. In: IEEE PacificVis. pp. 26–35 (2020)
9. Dietvorst, B.J., Simmons, J.P., Massey, C.: Algorithm aversion: people erroneously avoid algorithms after seeing them err. J. Exp. Psych.: General **144**(1), 114 (2015)
10. Dimara, E., Bailly, G., Bezerianos, A., Franconeri, S.: Mitigating the attraction effect with visualizations. IEEE TVCG **25**(1), 850–860 (2018)
11. Dimara, E., Stasko, J.: A critical reflection on visualization research: Where do decision making tasks hide? IEEE TVCG **28**(1), 1128–1138 (2021)
12. Dua, D., Graff, C.: UCI machine learning repository (2017)
13. Furtado, F., Singh, A.: Movie recommendation system using machine learning algorithms. International Journal of Research in Engineering and Technology (2020)
14. Gaba, A., Kaufman, Z., Cheung, J., Shvakel, M., Hall, K.W., Brun, Y., Bearfield, C.X.: My model is unfair, do people even care? visual design affects trust and perceived bias in machine learning. IEEE TVCG **30**(1), 327–337 (2024)
15. Ghai, B., Mueller, K.: D-bias: a causality-based human-in-the-loop system for tackling algorithmic bias. IEEE TVCG **29**(1), 473–482 (2022)
16. Groemping, U.: South german credit data: Correcting a widely used data set. Rep. Math., Phys. Chem., Berlin, Germany, Tech. Rep **4**, 2019 (2019)
17. Guo, S., Du, F., Malik, S., Koh, E., Kim, S., Liu, Z., Kim, D., Zha, H., Cao, N.: Visualizing uncertainty and alternatives in event sequence predictions. In: ACM CHI. pp. 1–12 (2019)

18. Johnson, B., Bartola, J., Angell, R., Witty, S., Giguere, S., Brun, Y.: Fairkit, fairkit, on the wall, who's the fairest of them all? supporting fairness-related decision-making. EURO Journal on Decision Processes **11**, 100031 (2023)
19. Kahng, M., Andrews, P.Y., Kalro, A., Chau, D.H.: Activis: Visual exploration of industry-scale deep neural network models. IEEE TVCG **24**(1), 88–97 (2017)
20. Kay, M., Kola, T., Hullman, J.R., Munson, S.A.: When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In: ACM CHI. pp. 5092–5103 (2016)
21. Kizilcec, R.F.: How much information? effects of transparency on trust in an algorithmic interface. In: ACM CHI. pp. 2390–2395 (2016)
22. Kube, A., Das, S., Fowler, P.J.: Allocating interventions based on predicted outcomes: A case study on homelessness services. AAAI **33**(01), 622–629 (2019)
23. Kunkel, J., Donkers, T., et al.: Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In: ACM CHI. pp. 1–12 (2019)
24. Lai, V., Tan, C.: On human predictions with explanations and predictions of machine learning models. In: ACM FAccT. pp. 29–38 (2019)
25. Mahmud, H., Islam, A.N., Ahmed, S.I., Smolander, K.: What influences algorithmic decision-making? a systematic literature review on algorithm aversion. Technological Forecasting and Social Change **175**, 121390 (2022)
26. Ming, Y., Qu, H., Bertini, E.: Rulematrix: Visualizing and understanding classifiers with rules. IEEE TVCG **25**(1), 342–352 (2018)
27. Munechika, D., Wang, Z., Reidy, J., et al.: Visual auditor: Interactive visualization for detection and summarization of model biases. In: IEEE VIS. pp. 45–49 (2022)
28. Ren, D., Amershi, S., Lee, B., Suh, J., Williams, J.D.: Squares: Supporting interactive performance analysis for multiclass classifiers. IEEE TVCG (2016)
29. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: ACM KDD. pp. 1135–1144 (2016)
30. Savikhin, A., Maciejewski, R., Ebert, D.S.: Applied visual analytics for economic decision-making. In: IEEE VAST. pp. 107–114 (2008)
31. Scheepens, R., Michels, S., van de Wetering, H., van Wijk, J.J.: Rationale visualization for safety and security. In: Computer Graphics Forum. pp. 191–200 (2015)
32. Suresh, H., Lao, N., Liccardi, I.: Misplaced trust: Measuring the interference of machine learning in human decision-making. In: ACM Web Science (2020)
33. Talbot, J., Lee, B., et al.: Ensemblematrix: interactive visualization to support machine learning with multiple classifiers. In: ACM CHI. pp. 1283–1292 (2009)
34. Wang, Q., Xu, Z., Chen, Z., Wang, Y., Liu, S., Qu, H.: Visual analysis of discrimination in machine learning. IEEE TVCG **27**(2), 1470–1480 (2020)
35. Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., et al.: The what-if tool: Interactive probing of machine learning models. IEEE TVCG **26**(1), 56–65 (2019)
36. Wu, C., Qian, A., et al.: Feature-oriented design of visual analytics system for interpretable deep learning based intrusion detection. In: TASE. pp. 73–80 (2020)
37. Xiong, C., Padilla, L., Grayson, K., Franconeri, S.: Examining the components of trust in map-based visualizations. In: TrustVis Workshop. pp. 19–23 (2019)
38. Yang, F., Huang, Z., Scholtz, J., Arendt, D.L.: How do visual explanations foster end users' appropriate trust in machine learning? In: ACM IUI. pp. 189–201 (2020)
39. Yin, M., Wortman Vaughan, J., Wallach, H.: Understanding the effect of accuracy on trust in machine learning models. In: ACM CHI. pp. 1–12 (2019)
40. Zhang, X., Ono, J.P., Song, H., Gou, L., et al.: Sliceteller: A data slice-driven approach for machine learning model validation. IEEE TVCG **29**(1), 842–852 (2022)
41. Zhang, Y., Bellamy, R.K., Kellogg, W.A.: Designing information for remediating cognitive biases in decision-making. In: ACM CHI. pp. 2211–2220 (2015)