Surfacing AI Explainability in Enterprise Product Visual Design to Address User Tech Proficiency Differences

Sara Tandon
King's College London, Microsoft
London, United Kingdom

Jennifer Wang
Microsoft
Redmond, WA, United States

ABSTRACT

This case study presents an investigation on explainable artificial intelligence (AI) visualization in business applications. Design guidelines for human-AI interaction are broad and touch on a range of user experiences with AI. Oftentimes, guidelines are not specific to enterprise scenarios with late-stage end users with limited AI knowledge and experience. We present a three-phase study on a visual design of a machine learning (ML) algorithm output. We conducted a user study on an existing design with limited visual AI explanation cues, ran a redesign workshop with various design and data experts, and conducted a reassessment with systematically applied AI explanation guidelines in place. We surface how users with various tech proficiency and AI/ML backgrounds interact with designs and how visual explanation cues increase understanding and effective decision making of users with low AI/ML familiarity. This design process corroborated the application and impact of existing guidelines and surfaced specific design implications for AI explainability within enterprise design.

CCS CONCEPTS

• **Human-centered computing** → **User interface design**; **Visualization design and evaluation methods**; *Empirical studies in interaction design.*

KEYWORDS

Artificial intelligence, user experience, user perceptions, practitioners, enterprise, explainable AI

ACM Reference Format:
Sara Tandon and Jennifer Wang. 2023. Surfacing AI Explainability in Enterprise Product Visual Design to Address User Tech Proficiency Differences. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3544549.3573867

1 INTRODUCTION

In fast-paced business scenarios, it is fundamental to increase efficiency for enterprise users and provide time to value insights – artificial intelligence (AI) and machine learning (ML) algorithms are often sought out to help reach these goals. Businesses are increasingly focused on leveraging AI solutions to optimize spending, inventory cycles, customer experience, and more. Thus, a high value is placed on design methods around human-AI/ML experiences so that organizations can make efficient and effective decisions [4]. The growth in complexity and performance of AI/ML systems results in challenges for visual analytic designs to be clear and understandable to humans [5]. This is especially true for enterprise users who may be experts in their industry but vary in their AI/ML familiarity and data/statistics knowledge – imperative to engaging with AI experiences in products. Research and design guidelines in explainable AI seek to empower designers to create transparency and ease of use for business users of all levels of AI/ML familiarity.

1.1 AI design guidelines and real world application in product

Human-AI interaction guidelines offered by Microsoft [1], Google [8], and business management organizations [9] relate mostly to the building, creation, and use of an AI. However, for human-AI interaction design guidelines to be successful, they must be understood and applied by enterprise end users. Additionally, research demonstrates individual backgrounds and experiences result in performance differences when interacting with data that can affect visual design choices [11]; this implies enterprise users who approach AI/ML systems with various levels of proficiency and familiarity could be influenced by design of visual AI systems. Thus, product designers face specific challenges around human-AI/ML interactions including collaboration with stakeholders, explainability of capabilities for differing levels of AI familiarity, and trust of data and visual systems [6, 13]. Visual analytic systems of AI/ML model output are a critical piece of the human-AI interaction pipeline – these systems can aid in trust building by making computations transparent and providing explanations for results [3]. Visual systems are often where end users encounter AI/ML systems for the first time and what they base decisions and recommendations on; thus, these systems need to be robust to all levels of technological familiarity of end users. Research demonstrates that users look for quick heuristic routes to confirm or discredit their working theories around AI/ML systems and users look for information when they have limited knowledge and/or when a system goes against expectations [5, 10]. Thus, design recommendations include principles like progressive disclosure [10]: the process of providing advanced information on an as needed basis, only when the user requests it. Progressive disclosure can help calibrate trust, increase user control of AI/ML features, improve acceptance of algorithmic systems, and promote learning and insight from complex data [10, 15]. The aim is to design visuals that support enterprise users in building appropriate trust, and making effective and time-efficient decisions utilizing AI/ML systems, regardless of individual familiarity and proficiency with AI/ML [2].

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3573867

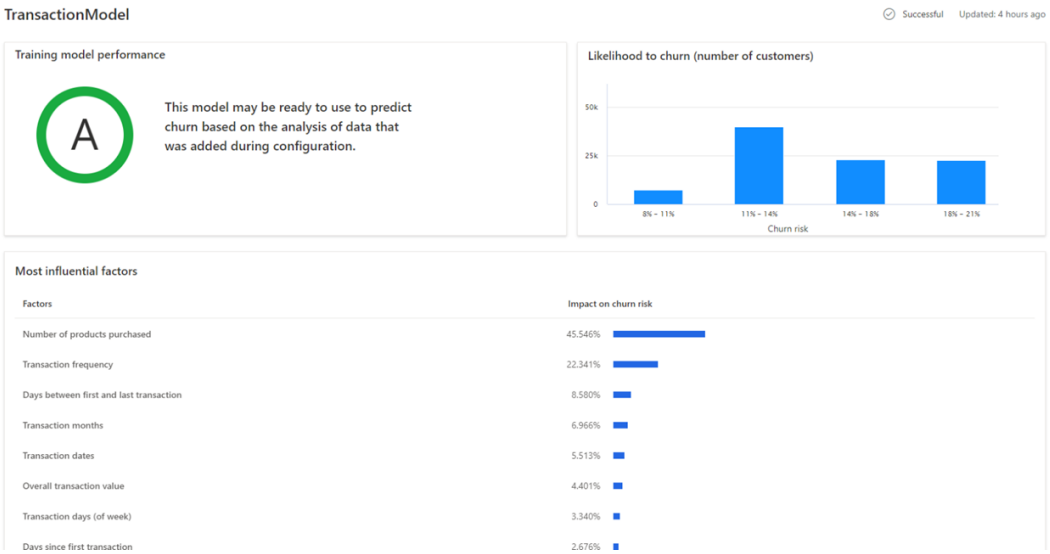


Figure 1: The initial dashboard contained three charts. Top left displays a training model performance metric with a letter grade, often used to simplify interpretation for enterprise product users. Top right displays model output of 4 bins of customers by estimated churn risk. The bottom chart displays customer behavior factors and their individual impact on the model's prediction of churn risk. This is a static dashboard with no interaction design. Full-scale images in supplementary material.

1.2 Present Study

With these principles in mind, we set out to test and design visual explanation cues for an AI/ML output system in an enterprise application. In this case study, we adapted an experience from real-world enterprise product scenarios with a range of end users who may engage with the product. We present a visual system created for end users without access to a data/applied scientist. The product takes customer behavior and product information and generates a predictive model segmenting customers by their predicted likelihood to churn. Additionally, we targeted participants with diverse AI/ML familiarity and expertise to gain an understanding of how explanations and design needs might differ between users. We present a three-phase study in which we assess a current visual dashboard with an AI/ML visual system (Phase 1), run redesign workshops with varied stakeholders in the product and user journey (Phase 2), and finally assess the redesigns for differences in user response and experience (Phase 3), surfacing design implications for AI/ML explainability in visual designs for enterprise.

2 PHASE 1: VISUAL DASHBOARD EVALUATION

To gauge baseline reactions and needs of enterprise users, we began by evaluating interaction with a visual design common across many enterprise products offering limited AI/ML explanation (e.g., top factors influencing model prediction) before design feedback and systematically applying explainable design principles. The dashboard we assessed displays the results of a predictive ML model of likelihood to churn (see Fig. 1). These visuals sit across many products in our group displaying similar model outputs. We aim to understand customer interaction and understanding of the data and underlying AI/ML methods through these charts.

2.1 Methods

Interviews were conducted using a third-party user testing service where we screened for representative users identified to potentially use the adapted product. We conducted semi-structured un-moderated interviews to gain user insight and feedback on the initial visual dashboard designs.

Participants. We interviewed 15 participants (5 from of each persona) screened to suit our primary enterprise personas (below) in June 2022. Participants were selected to represent target business personas across a range of functional roles within enterprise; recruitment screeners consist of questions including role, goals, and typical work experiences – they have been well established to target participants with expertise and needs around our product.

Marketer (MT): Semi-technical workers who are responsible for customer campaigns, content editing, and strategy implementation for customer engagement.

Business Analyst (BA): Experts involved in analyzing and interpreting relevant data quickly and effectively to identify key insights to share with stakeholders.

Data Wrangler (DW): Data and statistics experts responsible for data manipulation, unifying, and configuring into a digestible format. Comfortable running statistical models and troubleshooting for downstream users.

2.1.2 Interview Protocol. Interviews lasted 21 minutes on average, following the persona screener. The interview consisted of a scenario description, user understanding/feedback and a recommendation. Finally, Likert scale questions were asked on trust, understanding, and technological/AI/ML experience and interaction.

Scenario & Recommendation: Participants were asked to imagine they worked in the data department for a subscription

Table 1: Tech Proficiency dimensions and items. See supplementary material for all scores and interview guides.

Dimension	Tech Proficiency Items
AI/ML use and comfort for job role (2 items)	I use AI-powered insights or information to help make business decisions for my job. I am comfortable using AI features or capabilities for my job.
Coding / SQL (2 items)	I am experienced in coding and scripting languages. I am experienced in querying data using SQL or other querying languages.
Data & statistical interpretation for job role (2 items)	I am experienced in synthesizing and communicating findings from complex data types to make business decisions. I can easily interpret data from various outputs, including charts and graphs.
Resource to explain complex concepts and help (2 items)	I can explain complex technology topics to someone who is not familiar with the technology. People ask me for help with understanding complex topics related to technology in my job role.

foods business. Their objective was to utilize our customer data platform to send coupons to customers that were likely to churn. Participants were told their historical subscription data (e.g., customer transactions, subscription dates, etc.) had been uploaded to the product and it had created a transactional churn prediction, which helps predict if a customer will no longer purchase their products or services in a given time window. Participants were then shown the dashboard in Fig. 1 (see caption for chart descriptions). Participants were first asked to describe each of the three charts and what they understood the charts to represent. Subsequently, they were asked if they had enough information to make a recommendation on which/how many customers to give coupons to and/or what information they needed to make such a recommendation.

Trust & Tech Proficiency: Following the recommendation, participants were asked 7-point Likert scale questions to rate confidence in their recommendation, trust, and understanding of the dashboard and its AI/ML features. Additionally, participants were asked questions surrounding demographics and Tech Proficiency – an internally constructed multi-dimensional model of attributes to identify categories of technology proficiency across enterprise products. Tech Proficiency allows for a more holistic understanding of enterprise users and their backgrounds to tailor solutions for more meaningful and trustworthy experiences and enables consistency in cross-product research insights. While we expected differences in performance between MTs, BAs, and DWs, we aimed to examine if performance varied with concrete dimensions regarding tech proficiency and AI/ML experience.

2.2 Findings

Interview transcripts and recordings were analyzed using an inductive approach. This process produced both opportunities for redesign, and reinforcement of current design principles for data visualization and explainable AI/ML over the three displayed charts. Additionally, we identified disparate needs and understandings of ML data modeling between personas that affected insights and recommendations. To protect anonymity, participants are referred to by using the abbreviation for their persona (i.e., BA, MT, DW), followed by a participant number.

2.2.1 Participant Profile. Evaluating Tech Proficiency responses, we found that MTs had the lowest self-reported agreement with the dimensions in Table 1 followed by BAs and DWs having more experience. Responses between individuals with high and low Tech Proficiency are described in the sections below.

2.2.2 Understanding of Charts. The **Training Model Performance** chart surfaced redesign opportunities. Five of all 15 respondents (1 MT, 2 BAs, 2 DWs) understood that "A" referenced a grade but noted a lack of scale or how low/high the grade system went. An additional six respondents (4 MTs, 2 BAs) did not interpret the chart as a grade at all, rather they were either confused by it or assumed it represented something else altogether (e.g., product profile, dashboard status, etc.); it may be interesting to note that most of these participants were not located in the United States where the highest grade-level is commonly "A". When prompted the chart displayed a grade, four participants (2 MTs, 2 BAs) over-relied on the metric interpreting the green circle surrounding the "A" meant the model had no error at all. However, four participants (1 BA, 3 DWs) with higher Tech Proficiency noted that a high grade does not necessarily indicate the model is completely without error; *"a lower grade just means to take it with more of a grain of salt...there's never 100% certainty in modeling, there's always susceptibility"* [DW2]. This chart was easily misread, contained a sociocultural specific design, and did not allow for effective understanding of uncertainty in model performance.

The **Likelihood to Churn** chart caused the most confusion and/or misinterpretation for respondents. This could be due to lack of labeling of the y-axis, ambiguous notation, or misunderstanding of model structuring. Five participants (3 MTs, 2 BAs) with lower Tech Proficiency completely misunderstood the chart – some thought it represented purchase data, product saturation, financial information, or even admitted to not understanding the chart at all. However, participants with higher Tech Proficiency (1 BA, 5 DWs) eventually grasped the chart.

It notably took respondents extra time to formulate an understanding of the Likelihood to Churn chart with eight respondents taking anywhere from 30 seconds to 1 minute to respond, either correctly or incorrectly. The initial Likelihood to Churn chart lacked basic labeling that affected interpretation. Additionally, there seemed to be a lack of understanding of which customers were most at risk of churn, leaving a clear need for redesign.

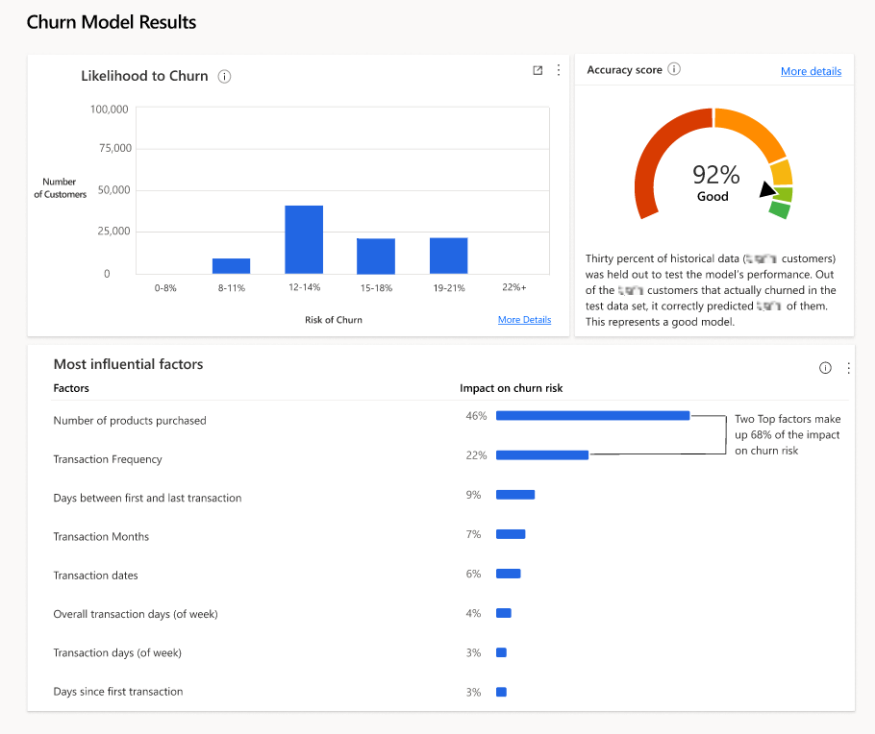


Figure 2: Redesign 1 designed in Phase 2 Workshop, tested in Phase 3. Blurred details were shown to participants.

The **Most Influential Factors** chart was the simplest for participants to understand, regardless of persona or Tech Proficiency. Many respondents (1 MT, 1 BA, 4 DWs) preferred to base their recommendation on (or included) the customer behaviors shown in this chart rather than the Likelihood to Churn chart (see below). We believe this chart was the easiest to interpret given the clear labeling and layout of highest influence to lowest – common principles found in data visualization and explainable AI/ML design recommendations [3, 7].

2.2.3 Recommendation & Trust. Recommendations varied between the 15 participants with some similarities depending on persona and Tech Proficiency. Participants were asked if they had enough information to make a recommendation on which customers to target for coupons or what they would need if they could not make a recommendation. Three MTs and two BAs did not give a recommendation due to confusion/misunderstanding of the data. One MT and two BAs recommended targeting the largest group of customers, which fell in the 11-14% churn risk group. This strategy implied they might not have understood the chart or did not think to target the two groups with higher churn risk. One BA and two DWs (with higher Tech Proficiency) wanted more data on customer behavior and demographics and/or more historical data on the model and churn likelihood to give a recommendation. The remaining three DWs (with higher Tech Proficiency) mentioned combining customers in groups with higher churn risk with their behaviors on "number of products purchased" and "transaction frequency." They all additionally mentioned diving into customer data themselves after choosing customers to target from the model output.

Notably, most participants readily trusted AI/ML generated data. Four MTs and two BAs (with lower Tech Proficiency) mentioned trusting AI generated data/predictions more than humans as *"data never lies"* [MT2], *"there is no error in AI data"* [MT1], and *"machines make less mistakes than humans do"* [MT4]. These statements imply an incomplete understanding of model prediction and reveal opportunities for visual design and communication to improve clarity on model building and uncertainty for less Tech Proficient users. All five DWs mentioned trusting the output since the performance metric was high and models are *"generally robust and trustworthy for predictive analysis"* [DW5], ultimately they interpreted model output with *"a grain of salt"* [DW2, DW4]. Higher Tech Proficient participants understood that *"none of this is an exact science"* [DW1], and that humans should be involved in decision making and data exploration in tandem with model predictions to make business recommendations.

3 PHASE 2: REDESIGN WORKSHOPS

Building on findings from Phase 1, we set out to explore how current designs could be improved for user understanding. To that end, we gathered various experts to collaborate on new designs grounded in human-AI/ML interaction design practices and principles. We ran workshops as research shows it is important for interdisciplinary teams to make design decisions together in AI/ML contexts as design methods are still developing around these constructs [13, 14].

3.1 Methods

Workshops lasted 1 hour and were conducted on July 22nd and July 29th, 2022. We utilized Miro, an online collaborative whiteboard

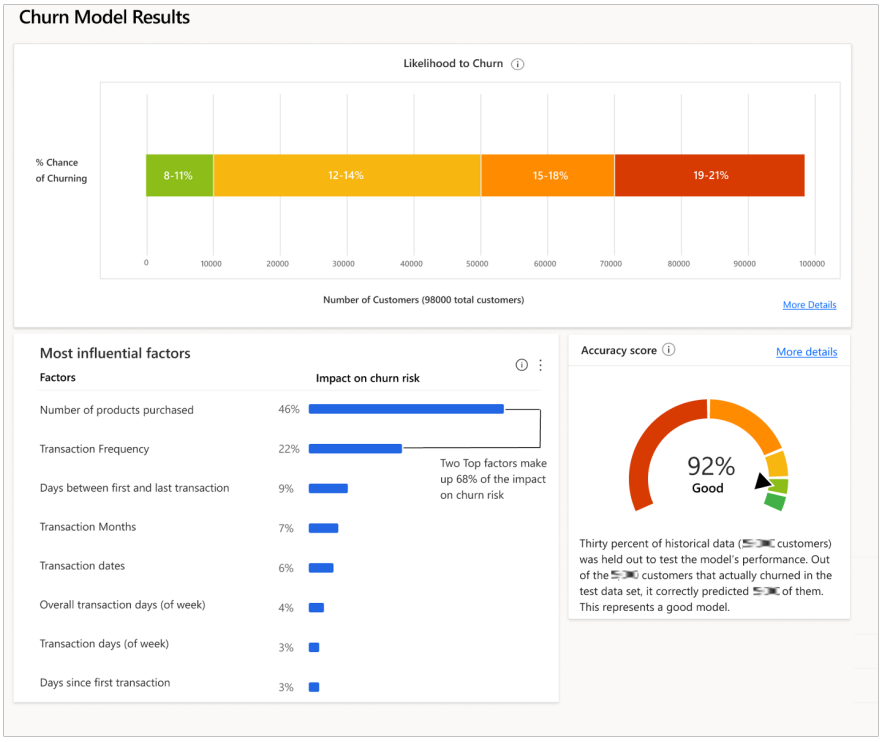


Figure 3: Redesign 2 designed in Phase 2 Workshop, tested in Phase 3. Blurred details were shown to participants.

Table 2: Chart descriptions available on information hover

Chart	In-line information description
Likelihood to Churn	The model grouped customers by their risk to churn percentage based on customer behaviors on the factors influencing churn
Model Performance	This score describes the accuracy of the AI model created using your historic customer data to predict the risk of your current customers churning
Most Influential Factors	The model identified these customer behaviors as the factors contributing to churn risk and assigned each an influence making up 100%

platform, to engage experts in the redesign process. Workshops consisted of experts in front-end (2 interaction designers, 1 user experience researcher, 1 content designer) and back-end practitioners (1 data scientist, 1 product manager): all held expertise crucial for product development and user experience. Participants were familiar with similar dashboards and had experience with AI/ML design in enterprise products.

3.1.1 Workshop Protocol. Workshops began with an overview of internal and external guidelines for human-AI/ML interaction design that referenced research on data visualization, progressive disclosure, and transparent representations of model performance. Labeling and annotation are commonly used tools for increased understanding and engagement with data visualization [7, 12]. Progressive disclosure design principles include on-demand information, hierarchically organized explanatory information (i.e., available information ranging from simple to complex), and context tracking (i.e., remaining in-context to relevant data with conversational and simple explanations) [10]. Progressive disclosure can support contextual understanding, build appropriate trust for lower Tech Proficiency, and offer more details for those with experience while not disturbing users' workflow (i.e., it is unique from documentation and does not require users to perform additional searches). Finally, we discussed internal research on model performance representations that included numerical representations with basic color-coded indicators of performance.

During design collaboration experts were asked to distinguish critical/needed changes from future/larger design changes. This distinction was encouraged so as to maintain a similar visual experience to original designs while including expanded visual explanation cues for assessment. The working definition for minimal changes included simple design implementation that would not disrupt users but enough to monitor impact – reported below.

3.2 Redesigns

Workshop collaboration and input were synthesized by the authors and two prototypes were created for testing (see Figs. 2 & 3). Redesign priorities included (1) rearranging the hierarchy of visuals to draw attention to model prediction output, (2) redesigning model

performance for transparency, (3) increasing labeling and clarity of the Likelihood to Churn chart to aid understanding, and (4) implementing AI explainability design principles. These priorities fall under Microsoft guidelines [1] around making clear what and how well the AI system can perform its task.

In both redesigns, across all charts, there is in-line information on each chart for users when hovering over the information *"i"* icon (see Table 2 for verbiage). Annotations follow AI interaction guidelines suggestions to show contextually relevant information as users interact with the model.

The **Model Performance** chart reflects a numerical score with an implied scale of 0-100 and displays a gauge ranging from red to green corresponding with the accuracy score, now clearly labeled. The visual includes a basic description of how the predictive ML model was created and what the score represents. Progressive disclosure is included wherein the "more details" hyperlink link cue opens a side-panel with further information on model history and model performance scores (see Fig 4(a)).

The **Most Influential Factors** chart is also the same across both redesigns. Users were able to understand and utilize this chart effectively in Phase 1, so required the least redesign. The chart now uses whole numbers to curb over-reliance and includes annotation of the two factors with the largest influence on churn to draw user attention and ease interpretation. Additionally, each bar displays information on-hover with a basic phrase stating the corresponding factor makes up x% of the factors influencing churn risk (e.g., *Transactions Months makes up 7% of the factors contributing to churn risk*).

In Redesign 1 (Fig. 2), the **Likelihood to Churn** chart is similar to the chart in Phase 1 with updated labeling on both axes. The x-axis now begins at 0 and ends at 22%+, although no data falls in those categories, the complete scale could aid comprehension. This chart also includes information on-hover over each bar with a phrase stating how many customers are predicted to churn at what risk (e.g., *It is predicted that 24,000 customers have a 15-18% risk of churning*). See Fig. 4(b) for the progressive disclosure side panel.

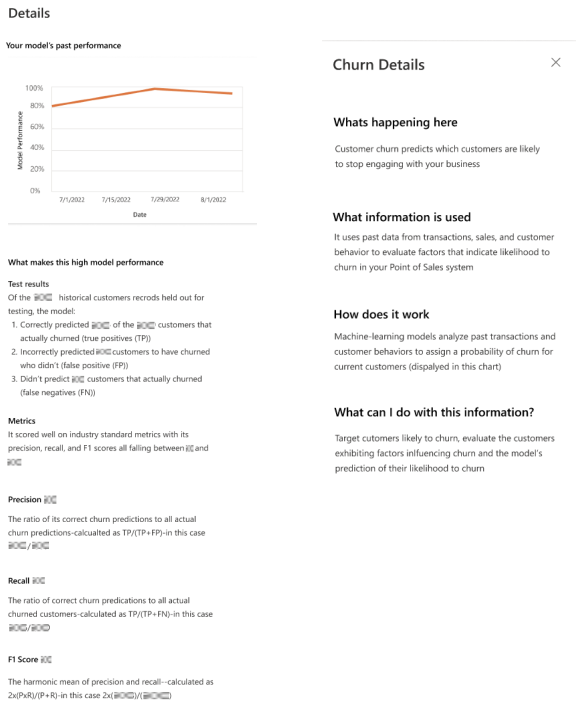
The only difference between redesigns is the Likelihood to Churn chart (the side panels, information on hover, and annotations remain the same). We chose to manipulate this chart alone in Redesign 2 (Fig. 3) as a complete redesign of the chart was suggested by experts but would require further testing and user interface overhaul. The design is a horizontal stacked bar with a green to red color scale to indicate increased churn risk. The y-axis shows the number of customers while each section is labeled with the churn risk for that bin of customers. This display facilitates clarity concerning breakdown of total customers by churn risk and which customers have the highest risk of churn.

4 PHASE 3: REDESIGN EVALUATION

We aimed to assess how added visual explanation and transparency cues affect user understanding, recommendations, and trust of the dashboard. For both redesigns we conducted the same assessment from Phase 1 on the same personas.

4.1 Methods

Interviews were conducted on the same user testing service as Phase 1. In this phase, we included un-moderated and moderated



(a) Side Panel for Accuracy Score chart.

(b) Side Panel for Likelihood to Churn

Figure 4: Progressive disclosure side panels available to users in both redesigns for surfacing explanations and model performance transparency. Panels appear overlapping the dashboard so users are not navigated away from the page.

interviews to collect more nuanced feedback on how the two designs compared to one another. We recruited 18 participants for un-moderated interviews – three respondents from each persona for each design (i.e., 9 participants for each design). We subsequently conducted three moderated interviews over the two designs, one from each persona.

4.1.1 Interview Protocol. Interviews lasted 17 minutes on average and followed the same scenario and flow of Phase 1 displaying one of the redesign prototypes. We encouraged participants to click around the prototype, but it was not explicitly required to do so – this allowed us to see if/when participants chose to interact with the prototype. For the three moderated interviews, we conducted the same interview using Redesign 1 (Fig. 2) but followed the interview by showing all three participants Redesign 2 (Fig 3). We asked participants to share initial thoughts on Redesign 2 and if they preferred for one, both, or neither design to complete the recommendation task.

4.2 Findings

Interview transcripts and recordings were analyzed analogous to Phase 1. This process produced insight into how the added explanation cues facilitated understanding of the ML model as well as opportunities for future work and research. We report findings across both designs for common changes: Model Performance chart

and Most Influential Factors chart. Findings for the Likelihood to Churn chart for Redesign 1 and Redesign 2 follow.

4.2.1 Participant Profile. Participants fell within similar ranges of Tech Proficiency and AI/ML familiarity as in Phase 1; MTs had the lowest followed by BAs with DWs having the highest levels of Tech Proficiency. We again identified some differences in response and interaction by persona and Tech Proficiency outlined below.

4.2.2 Understanding of Charts. The **Model Performance** color scale and numerical representation of model accuracy prompted confidence in all participants. Notably, participants with lower Tech Proficiency said the chart helped them feel *"the data is trustworthy"* [MT1, MT2], *"confident about where to go from here"* [MT3], and *"knowledgeable about the accuracy of the data"* [MT4]. Two MTs, three BAs, and two DWs clicked 'more information' on the chart. These participants noted their interest in, and usefulness of, the information: *"this is great – I know I can come here when I want to learn more about my model in the future"* [DW1]. MT2 mentioned feeling empowered by the information, *"now I know I can trust it and share with others who ask about the accuracy."* Some participants mentioned the color scale, stating it is *"easy to see the status of the data with the colors"* [DW2]. The redesign of the Model Performance chart increased clarity across the board and allowed participants to dive into specifics if needed.

The addition of on-hover annotation and indication of the top two factors in the **Most Influential Factors** chart aided participants through plain language descriptions. Participants equally relied on the factors as an important part of understanding customer behavior as in Phase 1 but were able to reinforce their initial impressions of the charts through annotation. The annotation drew attention to the influential factors such that two MTs, with lower Tech Proficiency, mentioned the top two factors as part of their recommendation.

Additional labeling in **Redesign 1 (Fig 2)** of the **Likelihood to Churn** chart led to increased understanding of this chart compared to Phase 1. Most participants read the on-hover bar information and gained an immediate understanding of what the data represented. Of all nine participants that saw this design, only one participant misunderstood it. We believe the reduced overall interview length reflects the effect of increased labeling on understanding of this chart as it was the most difficult chart for interpretation in Phase 1. Only two participants clicked 'more details' on this chart – this implies the chart may have been clear enough that participants did not feel the need for more information.

Redesign 2 (Fig 3) differed the most from Phase 1 design. The new color scale led to instantaneous comprehension of customers with the highest churn risk. However, this design initially confused two participants. On the other hand, all three participants in the moderated interviews preferred this design to Redesign 1. They all felt the color scale aided in quick and effective interpretation, making essential information instantly clear. The DW noted seeing how the churn risk is distributed amongst the total number of customers helped them contextualize the information. All participants, including the two initially confused participants, noted the color scale helped them target customers.

4.2.3 Recommendation & Trust. Between both redesigns, only one MT (who saw Redesign 1) and one BA (who saw Redesign 2) of 18 participants did not make a recommendation – both wanted more information and/or did not grasp the data enough to make a clear recommendation. The rest of the respondents targeted customers with the highest risk to churn in Redesign 1 or customers in the red (with 2 DWs including the orange) section of Redesign 2. Overall, we observed an increase in the number of participants that were able to give a recommendation on both designs regardless of Tech Proficiency. Additionally, more participants mentioned combining risk with customer behavior, and mentioned wanting to access customer demographics of those with high churn risk (4 BAs, 3 MTs, and 4 DWs). In moderated interviews, the BA and the MT did not feel confident making a recommendation based on Redesign 1, however when shown Redesign 2, they both stated it was easier to read and quickly understand which customers to target. All 21 participants mentioned trusting the data as the accuracy score was strong and they had resources to understand how the data was being generated. Implications of our findings are discussed below.

5 DISCUSSION & CONCLUSION

As AI/ML features become commonplace in enterprise products, design needs and challenges increase [6]. We recruited participants from three distinct enterprise personas that often make recommendations to increase business impact but may have limited access to data scientists – Business Analysts, Marketers, and Data Wranglers. While experts in their industry, these personas vary in their AI/ML familiarity and data knowledge, which are imperative in engaging with AI experiences in products. We ran a user study on visual output of a predictive ML model with limited transparency and model explanation (Phase 1, Fig. 1); this constituted a *Human + Machine* scenario wherein humans receive the baseline performance/output of a system without explanation [2]. We asked participants to share their understanding of the dashboard, make a recommendation, and rate their interactions and competency around relevant technological constructs. We examined Tech Proficiency (Table 1) as design should consider a range of user proficiencies and implement methods that reduce barriers to entry while supporting those with more experience. We found participants with higher Tech Proficiency were able to make recommendations, while those with limited to no experience were not able to make a strong recommendation and struggled to understand the ML output. We noted limited labeling and ambiguous notation and metrics led to incorrect interpretation, and over-reliance, when no indication of uncertainty was present. Additionally, we found participants reported high levels of trust in AI/ML generated data regardless of their understanding.

This prompted us to conduct redesign workshops with experts to increase impact of the visuals (Phase 2). Workshops followed human-AI/ML design guidelines and suggestions [1, 9] while adapting them to our specific design problems from Phase 1. We distilled workshops into two designs for testing that included increased interaction, annotation, progressive disclosure, and design overhaul of two charts (Fig. 2, Fig. 3, Fig. 4, Table 2). Redesigns were then assessed (Phase 3) with consonant methods to Phase 1. These designs reflect a *Human + Machine + Explanation* scenario where the human receives explanation along with ML output [2]. We compared

user understanding and recommendations to those without explanations to investigate the impact of the design changes. While trust remained high, recommendations and understanding of lower Tech Proficient participants was similar to those with more experience.

5.1 Reflection

The fact that level of trust in AI/ML output despite misunderstandings should be important to the academic community creating and implementing AI experiences. Approachable explanations, fairness, and transparency of methods and output should be of the highest importance. Our work also exposed the utility of expert collaboration – each focused on a different piece of the user experience and product with an eye for their domain: focusing on user understanding of data, model performance, chart engagement & aesthetics, verbiage, and follow-up action prompting. Bringing together user research and varying domain experts in iterative design can lead to avenues of opportunity and insight that can empower users beyond the immediate interface. Future changes were discussed as part of the workshops we have not yet implemented and tested, including offering data and model customization tips, linking customer data, next-action prompts, real-world application examples, and positive and negative impacts of trusting AI/ML output.

Design Implications for Explainable AI/ML in Enterprise. While human-AI interaction guidelines call for clarity in what and how systems do what they do, matching social norms, and offering users explanations [1], they often do not specify *how* to go about this in a visual design scenario. We found on-hover annotation utilizing plain language empowers users to make clear, concise, and simple conclusions. Additionally, on-hover annotation aids in clearing the visual field and offering information as users desire it. To avoid over-reliance and misinterpretation due to sociocultural differences, we recommend displaying model performance metrics as a numerical score with an explicit scale. We found this eases interpretation due to transparent communication of uncertainty and removes cultural ambiguity surrounding an alpha-numeric grade system. Finally, we recommend utilizing progressive disclosure that does not navigate users away from the visual context of the model output. Designs should include in-line entry points, connecting to panels with hierarchical information and jargon-free descriptions of features. We found these visual design changes together with progressive disclosure gave an opportunity for users with limited AI/ML experience to come away with stronger recommendations and increased understanding of AI/ML output while allowing high Tech Proficient users to gain knowledge efficiently without distraction.

5.2 Outlook

Our study complements human-AI/ML design guidelines but is limited to design for end users in specific enterprise environments. We also limited our study to qualitative feedback, and we would like to gather quantitative data from our personas toward a holistic understanding. Additionally, it is important to work toward pinpointing perceived versus tangible expertise and understanding around AI/ML outside of a qualitative interview space. In this case study, we found users with low AI/ML experience and data literacy perform similarly in recommendations and understanding to those with more expertise when visual explanations are used. Our work resulted in design implications that we hope can empower enterprise designers and academics to create effective AI/ML interactions for their users and inspire research in explainable AI design for a range of users with diverse needs, expertise, and resources.

REFERENCES

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13.

[2] Brittany Davis, Maria Glenski, William Sealy, and Dustin Arendt. 2020. Measure Utility, Gain Trust: Practical Advice for XAI Researchers. In *2020 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*. 1–8.

[3] Wenkai Han and Hans-Jörg Schulz. 2020. Beyond Trust Building — Calibrating Trust in Visual Analytics. In *2020 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*. 9–15.

[4] Sarah Hanses and Jennifer Wang. 2022. How Do Users Interact with AI Features in the Workplace? Understanding the AI Feature User Journey in Enterprise *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 36, 7 pages.

[5] Alexander John Karran, Theophile Demazure, Antoine Hudon, Sylvain Senecal, and Pierre-Majorique Leger. 2022. Designing for Confidence: The Impact of Visualizing Artificial Intelligence Decisions. *Frontiers in Neuroscience* 16 (2022).

[6] Clara Kliman-Silver, Oliver Siy, Kira Awadalla, Alison Lentz, Gregorio Convertino, and Elizabeth Churchill. 2020. Adapting User Experience Research Methods for AI-Driven Experiences. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8.

[7] Tamara Munzner. 2015. *Visualization analysis & design.* Taylor & Francis Group.

[8] Google Research. 2019. People and AI Guidebook. https://pair.withgoogle.com/

[9] SAP SE. 2022. SAP Fiori Design Guidelines. https://experience.sap.com/fiori-design-web/explainable-ai/

[10] Aaron Springer and Steve Whittaker. 2020. Progressive Disclosure: When, Why, and How Do Users Want Algorithmic Transparency Information? *ACM Trans. Interact. Intell. Syst.* 10, 4 (oct 2020).

[11] Sara Tandon, Alfie Abdul-Rahman, and Rita Borgo. 2022. Measuring Effects of Spatial Visualization and Domain On Visualization Task Performance: A Comparative Study. *IEEE Transactions on Visualization and Computer Graphics* (2022), 1–11.

[12] Colin Ware. 2004. *Information visualization: Perception for design.* Elsevier Science & Technology.

[13] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13.

[14] Sabah Zdanowska and Alex S Taylor. 2022. A Study of UX Practitioners Roles in Designing Real-World, Enterprise ML Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 531, 15 pages.

[15] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 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* (Barcelona, Spain). Association for Computing Machinery, New York, NY, USA, 295–305.