User Studies in Human-Feature-Integration

Yixin Li[1,†], Lucas Lefebvre[1,†], Sonali Parbhoo[2], Finale Doshi-Velez[3] and Isaac Lage[1,*]

[1]Colby College
[2]Imperial College London
[3]Harvard University

Abstract

Machine learning (ML) models make predictions using the same set of input features for all instances. Predictions made about people may be improved by considering individualized context, however this context must be supplied to the ML model. Lage et al. [1] describes the problem of human feature integration: querying each user of an ML system for values of a small, personalized set of additional features to improve the prediction made about the user. This work is based on assumptions, including that users have access to relevant information about themselves (i.e. human features), but not necessarily about the predicted outcome; and that selecting the set of features to query for each user is best approached algorithmically. In this work, we provide preliminary evidence from two user studies towards the validity of these assumptions, and lay out directions for future work exploring the human feature integration problem.

Keywords

Interactive machine learning, human-AI collaboration

1. Introduction

ML models are attractive partly because they promise personalized predictions and decisions at scale. But the outputs are typically produced based on the same set of pre-specified inputs for all users. The inputs, or features, that are informative for some instances or individuals may not be as informative for others and vice versa. Lage et al. [1] introduced the problem of human-feature-integration to address this issue by querying users of ML systems for the values of a small set of individualized features at test time. This work is based on three key assumptions: (1) users about whom predictions are made have access to relevant information about themselves, but not necessarily about the predicted outcome; (2) this relevant information can improve model predictions; and (3) selecting the set of features to query for each user is best approached algorithmically. In this work, we provide preliminary evidence from two user studies towards the validity of assumptions (1) and (3). Whether these features actually improve predictions in a realistic system (assumption 2) is an important open question that we plan to address in future work. We discuss this in Section 5.

The preliminary results from 2 user studies described in this paper address the following 4 research questions:

- RQ1 What does the human feature space look like?
- RQ2: Does asking users about human features impact trust or willingness to use the model?
- RQ3: How do the features chosen by users compare to the features chosen algorithmically?
- RQ4: Do people prefer to select their own features, or to have an algorithm select the features for them?

RQ1 and RQ2 provide evidence towards assumption (1) and are described in Section 3, and RQ3 and RQ4 provides evidence towards assumption (3) and are described in Section 4. Our key findings are that (1) a large space of task-relevant human features can exist; (2) asking users about human features does

Joint Proceedings of the ACM IUI Workshops 2025, March 24-27, 2025, Cagliari, Italy
*Corresponding author.
†These authors contributed equally.
yli25@colby.edu (Y. Li); lplefe26@colby.edu (L. Lefebvre); s.parbhoo@imperial.ac.uk (S. Parbhoo); finale@seas.harvard.edu (F. Doshi-Velez); ilage@colby.edu (I. Lage)

not appear to negatively impact trust in the model and may improve willingness to use it; (3) a simple operationalization of human feature integration described in [1] can provide qualitatively different feature queries than users; and (4) people report a subjective preference for a computer-generated set of queries instead of their own. While there remain important open questions, these findings provide a first set of human-factors evidence towards key aspects of the human feature integration problem outlined in [1].

2. Related Work

Human-AI Team Decision-Making Various methods have been proposed to facilitate joint human-ai predictions by combining predictions made independently by each party. Raghu et al. [2] learns a model of past human predictions to triage predictions between an ML model and a human. Wilder et al. [3] presents an end-to-end framework for learning a model that complements human decision makers in cases where they are less accurate. Madras et al. [4] presents a framework based on rejection learning that accounts for the knowledge of other decision makers–specifically humans, when the ML algorithm is deciding whether to make a prediction for an instance. In contrast, we study a method to incorporates *features* from users so that relevant knowledge from both decision makers can be leveraged in the *same* decision.

Algorithm Aversion Foundational work on algorithm aversion shows that allowing users to edit models can improve uptake, even when the edits are constrained [5]. Recently, Cheng and Chouldechova [6] found that feature edits were less effective at mitigating algorithm aversion than output edits. In their context, users are contributing features about someone else (a student who is having their grade predicted) while in our case, users are contributing features about themselves. This is a significantly different context and may affect how people perceive this intervention. Our goal is also not avoid algorithm aversion altogether, rather to avoid magnifying it.

Editing Machine Learning Models Whether and how users can edit machine learning models is a well studied question. Some approaches allow users to directly edit learned models, (e.g. Gillies et al. [7], Kulesza et al. [8]), affecting all decisions made by the model. Others edit the data going into the model, which can be done at the level of the raw inputs or concepts. Jacobs et al. [9] finds that clinicians are interested in being able to edit patient features in a clinical dashboard for antidepressant prescription. Koh et al. [10] finds that a set of intermediate concepts that they learn in order to base predictions on can be edited to correct any mistakes made while learning concept definitions. In contrast, the human feature integration problem assumes that the feature editing is guided by an algorithm rather than acquired from users in a fully open-ended way. In principle, explainable machine learning models allow arbitrary to the prediction process when it is transparent (e.g. Wang et al. [11], Lakkaraju et al. [12], Yang et al. [13]). However this requires significant mental effort and domain knowledge to understand how to appropriately change the prediction function to get the intended output. The human feature integration problem requires substantially less of both to be effective.

Instance-wise Feature selection Other machine-learning methods for choosing features to incorporate into predictions for specific instances exist beyond the approach described in Lage et al. [1]. E.g. Yoon et al. [14] and Jethani et al. [15] both solve a similar technical problem framed around a different goal (providing concise explanations of machine learning models). Future work comparing different technical solutions to the problem of querying human features would be interesting, however this work focuses on better understanding properties of the proposed problem rather than finding the most effective solution to it.

3. User Study: Exploring the Human Feature Space

In this section, we explore research questions 1 and 2 towards better understanding assumption (1) in the human feature integration problem formulation–users about whom predictions are made have access to relevant information about themselves.

- RQ1: What does the human feature space look like? How large is the human feature space, and are features suggested by people reasonable and plausibly predictive?
- RQ2: Does asking users about human features impact willingness to use the model? On the one hand, allowing people to input personal knowledge into an ML prediction made about them may improve willingness to use the model, or it may instead call attention to the model's limitations.

We demonstrate through a user study that the space of human features users suggest can be large and diverse, with different users suggesting different human features. We also provide evidence that people subjectively prefer to use a model that uses their suggested features, and that calling peoples' attention to human features does not appear to harm willingness to use the model.

Experiment setup We conducted a user study to determine what kinds of additional features users of machine learning models may wish to integrate into predictions made about them. In the study, we described a hypothetical task where an AI assists a hiring manager to predict the participant's salary as a way of making a well-calibrated job offer to the participant (i.e. with the salary higher than their current amount, but not too much higher). We told the user that the AI had access to a sparse set of features about them–age, highest level of education completed, number of years of experience, and job sector. We then asked the participant to provide additional features (phrased as "factors you believe the AI should consider to better predict your salary.") that could improve the AI's accuracy in their case. We provided 8 required and 2 optional free text boxes with a 3-word maximum.

We chose this prediction task because: 1) it is one many people are familiar with (from negotiating job offers and understanding their own salary), 2) there are many existing datasets for this prediction task that we can use to compare user-suggested features to features from existing ML datasets, and 3) it is a prediction task with a plausible amount of user-specific variance since different people's salary may depend on unique attributes of their job, their expertise, or their personal life, among other things. While this is a domain where many users arguably *are* domain experts, it allows us to explore the potential richness of the human feature space in a concrete setting. We addressed the issue of users being interested in receiving a higher salary by reminding them that they will not get a job offer if the manager does not think their expected salary is within the company's range.

In addition to this main task, we also asked users several questions to measure willingness to use the AI in order to observe whether this changes after providing human features. We report on the results of one variation of this question in the main text, and discuss 2 other variations on this question and their results in Appendix B.

Exclusion Criteria	Study 1 # Responses	Study 2 # Responses
Commitment question	3	N/A
Generative AI use	49	N/A
Attention question	10	4
Nonsense or repeated responses	20	16
Incorrect practice questions	N/A	25
Survey technical error	1	0

Table 1
Table of reasons for excluding responses. Responses excluded for multiple reasons are counted for each reason. We did not use identical exclusion criteria for both studies, so those do not apply are marked "N/A." For study 1, the majority of responses are excluded due to reported generative AI use, while for study 2, the majority were excluded for incorrectly answering our practice questions.

Demographics	Description	Study 1 Percentage	Study 2 Percentage
Age	under 25	5%	2%
	25 - 34	66%	65%
	35 - 44	22%	29%
	45 - 54	2%	2%
	55 - 64	5%	0%
	65 and above	0%	2%
Education Level	Bachelor's Degree	61%	57%
	Graduate Degree	29%	38%
	Some college or Associate Degree	5%	2%
	High School Diploma	5%	3%
Gender	Male	61%	69%
	Female	39%	31%

Table 2
Table of demographics for retained survey participants in both study 1 and study 2.

Participants We recruited 107 unique participants from Amazon Mechanical Turk to participate in this study in Summer of 2024. Of those, we retained data from 41 participants after controlling for quality. The demographics of the retained participants are shown in Table 2. All participants saw the same survey. The survey lasted for an average of around 12 minutes, and participants were paid $2.50 for participating in the study. This study was approved by our institution's IRB.

We filtered low quality responses through a commitment question, an instructed responses question [16] and a question where participants were asked if they used generative AI to complete the task (we emphasized that this would not affect their payment). We also excluded nonsense responses with non-words, or repeated identical responses. The exclusion criteria with corresponding exclusions (participants may be listed more than once if multiple criteria applied) are listed in Table 1. The high exclusion rate is mostly due to participants responding that they used generative AI at some point during the survey. While some of these responses might still accurately represent what the participant wished to convey, we excluded them all as it is not possible to tell which ones do, and which ones simply convey what generative AI responds to our survey questions.

Results–RQ1: The features suggested by users provide variations on features from existing salary datasets, as well as distinct concepts not covered by those datasets. Figure 1 shows the features suggested by users visualized in 2-dimensions using TSNE on the word embeddings from the Hugging Face all-MiniLM-L6-v2 [1]. The small purple points show where the feature names from 4 existing salary datasets (see AppendixA) appear in the space. Each cluster (using DBSCAN on the TSNE distance matrix, excluding the dataset features, and using 61 clusters and 22 noise points) has 1 random feature label shown on the graph. From this, we can see that many of the clusters of user suggestions appear near one or more purple points, suggesting that they are similar to features in existing datasets. They do not tend to appear directly on top of each other, perhaps because of variations in structure between feature names in datasets and free text responses from our survey. This suggests that many of the features proposed by users are similar to these existing features that we know to be reasonable, but not identical. They may provide variation that could be useful in different prediction contexts. It also confirms that many of the user suggestions are reasonable and task relevant. On the other hand, a significant number of the clusters of user suggested features do not appear near any features from the existing sets, like *benefits*, *communication skills*, and *industry certifications*. This suggests that users are not just providing variation on "obvious" features that might already appear in datasets, but are also providing unique but still intuitively task-relevant features. These results suggest that the human feature space can both more nuanced than one might expect from a set of features in a machine learning dataset, and more varied, including unique concepts that may not appear in existing datasets.

Results–RQ2: The majority of participants would prefer to use the model that uses their

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

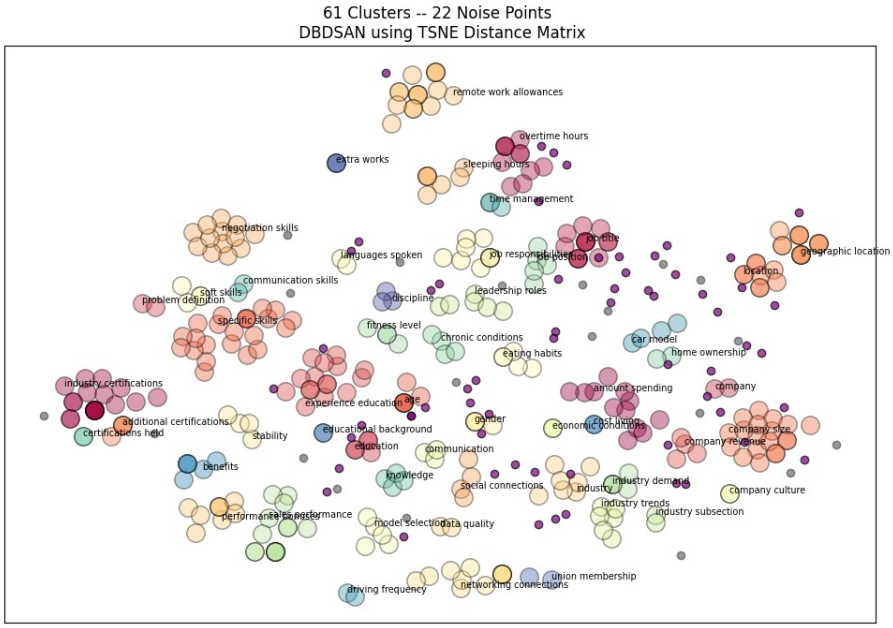


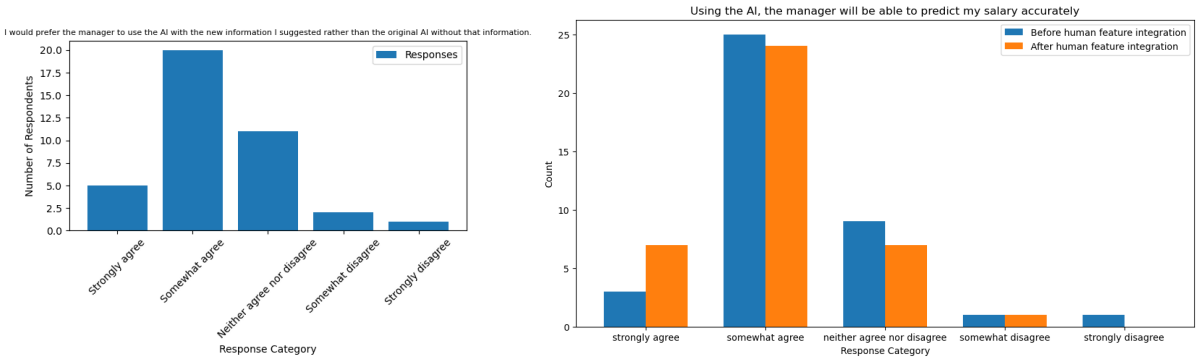
Figure 1: This figure shows user-suggested features visualized in 2-D using TSNE based on word embeddings. Colors corresponds to clusters of features derived from the TSNE distance used in the graph using DBSCAN, and are used to select one response from each cluster to show on the graph. Purple points correspond to features in the 4 salary prediction datasets we considered. These were used in TSNE but not in the clustering. While most of the purple features from existing datasets are near clusters of user-suggested features, several clusters like benefits, industry frequency, and networking connections do not appear similar to any features from these datasets.

suggested features, and willingness in the model is not harmed by asking people to suggest new features. Figure 2 shows the distribution of responses to the question "I would prefer the manager to use the AI with the new information I suggested rather than the original AI without that information.". Over half of participants (25/41) would either somewhat or strongly prefer to use the model with their suggested features, while the majority of those who don't prefer it feel neutrally (11/41). Only 3 participants actively disagree that they want to use the model with their suggested features. This suggests that acquiring human features is perceived positively for most people. Differences in willingness to use the model are also minimal and suggest that this is not negatively impacted by requesting additional human features. Figure 2 shows the distribution of responses before and after collecting human feature suggestions for the question "Using the AI the manager will be able to predict my salary accurately." The differences are not statistically significant, suggesting that belief in the model's accuracy is not harmed by allowing users to specify additional human features that a model should consider. We see similar trends across the 2 other ways we ask about willingness to use the model. See Appendix B for details on statistical tests and results on variations of this question.

4. User Chosen Features vs. Algorithmically Chosen Features

In this section, we explore research questions RQ3 and RQ4 towards validating assumption (3): selecting the set of features to query for each user is best approached algorithmically.

- RQ3: How do the features chosen by users compare to the features chosen by entropy selection? Are users and the proposed algorithm making similar or distinct selections?
- RQ4: Do people prefer to select their own features, or to have an algorithm select the features for them? Subjective preferences do not necessarily correspond to predictive performance, but can give indications of how usable an approach is.



(a) Bar chart of responses to the question: "I would prefer the manager to use the AI with the new information I suggested rather than the original AI without that information." Over half of participants either somewhat or strongly agree that they would prefer to use the AI that incorporates their features.

(b) These graphs show the distribution of responses to the question: "Using the AI, the manager will be able to predict my salary accurately". Orange bars (right) show responses after participants supplied their features to the AI, and the blue bars (left) show responses before. Differences in responses are minimal and not statistically significant.

Figure 2: Preference and trust results from study 1.

We demonstrate through a second user study that features selected by users and features selected by the algorithmic approach proposed by Lage et al. [1] can vary substantially in the pattern of features selected. We provide evidence that people report a subjective preference for the algorithmically-generated set of features over their own, despite the fact that they feel confident in their selections. We find these results in a domain where people *do* have significant domain expertise, and we expect them to be even more exaggerated when they do not.

Experimental Setup In this study, we asked users to provide values for a complete set of 22 pre-designated "human features," and asked them to select and rank the top 5 features they would want to supply to the predictive model. This setup allows us to compare the features selected by users with those selected by the algorithm proposed in Lage et al. [1].

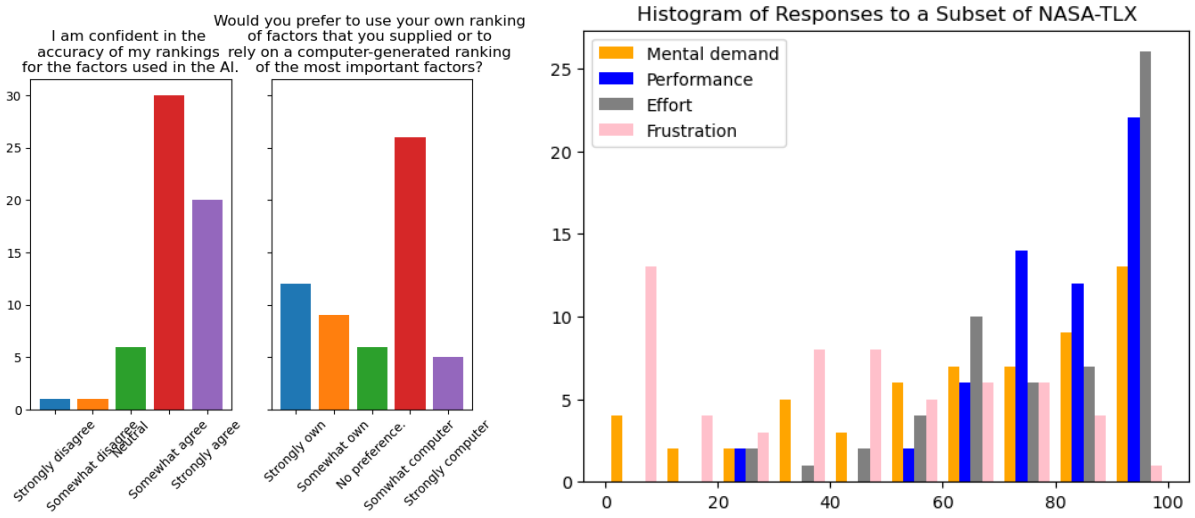
We used this dataset because (1) it is designed to have questions that the general public have an understanding of and can answer; (2) the personal health domain is diverse, and plausibly different factors can contribute to good or bad health for different users; (3) we were able to set up a simulated split between "human" and "machine" features where the approach proposed by Lage et al. [1] performs effectively, and (4) this is a dataset where users have substantial domain expertise, which means it should provide an upper bound on how well people can do at this feature selection task.

In order to train the models used by this algorithm, we derived 22 "human features," 4 "machine features" and the prediction task from an existing dataset: the NHANES–National health and nutrition examination survey from 2013-2014. [2]. See AppendixC for additional details about data processing and model training.

Participants We collected data from 100 participants from Amazon Mechanical Turk in summer 2024. We retained and analyzed data from 58 participants after excluding responses for the reasons described below. Table 2 shows the demographics of the retained participants. This study took approximately 15-20 minutes. We paid $4 for completing the study–$1 for participating in the practice questions, and $3 dollars for correctly answer at least 1 practice question and completing the study. This study was approved by our institution's IRB.

We excluded participants based on: 1) whether they answered at least 1 practice question correctly, 2) whether they correctly answered the attention check around halfway through the survey, and 3) whether

[2]Acquired from Kaggle: https://www.kaggle.com/datasets/cdc/national-health-and-nutrition-examination-survey



(a) Results of people's confidence in their ranking, and their preference for either their own ranking or a computer generated one. We see that people feel strongly confident in their ranking, but that the majority of people still would prefer to use a computer generated ranking over their own.

(b) Results of the NASA-TLX questionnaire (subset of questions) about providing all 22 human feature values. Users perceived they were doing well at the task, but found it high in effort and mental demand. Frustration with the task was relatively low. This suggests that asking about few, targeted human features is an important aspect of this approach.

Figure 3: Confidence and NASA-TLX results from study 2.

they gave the same response for all of the human features (e.g. answering yes to all 22 questions). Participants were mostly excluded for failing 2 sets of practice questions, or for repeating the same response throughout the survey. Table 1 lists the reasons participants were excluded (with overlap when multiple conditions applied).

Results-RQ3: Patterns of selected features vary substantially between users and the entropy selection approach from Lage et al. [1] Figure 4 shows the features selected by users and the entropy selection algorithm for each query made. We can see that there is more variation in features selected by users, however they tend to mostly be within the categories of alcohol, nutrition and exercise. On the other hand, entropy selection generally selects the same first feature for most instances, then focuses on 8 or so other features spread across the different categories. This suggests that entropy selection may be able to identify a preferred feature within a category, then select the right category for each instance, while users who may know which categories they believe to be important, then choose features within those. This shows clear variation in how users and entropy selection approach selecting features. Which of these is more effective at improving predictions will be explored in future work. [3]

Results-RQ4: People feel confident in their feature ranking, but subjectively prefer an algorithmically generated ranking. Figure 2 shows both the confidence that people have in the ranking of features they produced, and whether they would prefer to use their own ranking or a computer generated one. These results show that people are quite confident in their ranking, however the majority of people still report preferring a computer generated ranking. This suggests that, many people would still subjectively prefer to use an algorithm to generate feature queries even if they believe they could do it effectively themselves. This provides evidence towards human feature integration being well received by users.

Additional result: Even our short survey of 22 questions is considered high effort, which validates the importance of querying users for a subset of human feature values. Figure 3 shows

[3]This can be measured with our experimental setup, but currently the predictive performance contains too much variance to provide meaningful signal. This is an area of ongoing work.

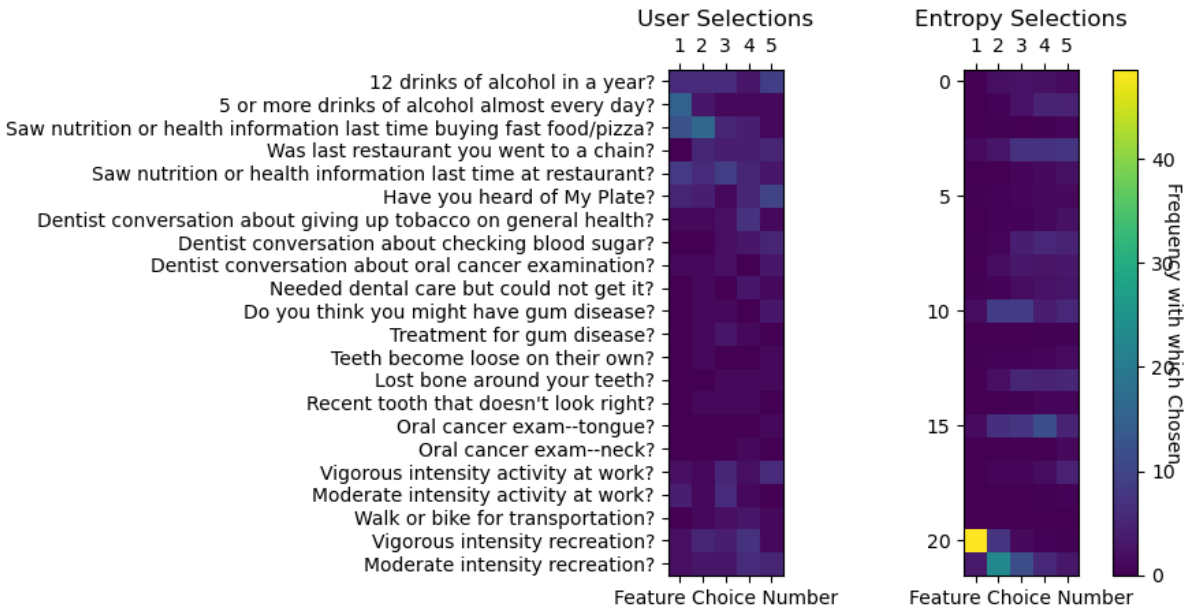


Figure 4: Patterns of features selected by users and the entropy selection algorithm for each of the 5 feature queries. Lighter colors correspond to features chosen more frequently. We can see from this that both approaches show variation, however the variation for the entropy selection approach is more concentrated in 8-9 key features, while the variation for the user selections is more concentrated on the nutrition questions, with some also on the exercise questions. This suggests that users and the entropy selection algorithm approach this problem very differently.

histograms of responses to a subset of the NASA-TLX questionnaire that we asked users to respond to based on their experience of providing feature values for the 22 "human features." Participants report that they performed well on the task and were not particularly frustrated, however the task took slightly greater mental demand, and relatively high effort. If the small number of questions we asked them is considered high effort, this supports the assumption that asking users to provide values for all human features (e.g. the larger sets of human features used in Lage et al. [1]) is unreasonable.

5. Limitations and Future Work

Limitations of current studies In study 1, users have an interest in presenting themselves in the best possible light, which may impact which features they share. While this is a realistic problem setting, how results may differ in other contexts merits further study. In study 2, we see qualitative differences in human and algorithm-proposed rankings, and a subjective preference for the algorithm, however the variance in predictive performance is currently too great to draw conclusions about whether either approach is more effective. Because we selected the human and machine feature split based on the performance of the algorithm, this study setup will also not allow us to demonstrate whether human features in fact improve performance.

Future work 1: Do human features improve model predictions? Future work exploring human-feature-integration must demonstrate whether incorporating these human features into predictions actually improves prediction quality. We plan to undertake a full-scale study of this approach where we collect a space of human features from users, then use those features to train models, elicit human feature rankings, and evaluate performance. The first step will be similar to study 1, and the second step will be similar to study 2.

Future work 2: Should the human feature space be iteratively expanded? The second pressing question for future work is whether a sufficient human feature space can be collected at once in a setup similar to study 1, then used to train models and make predictions at test time. This is the current assumption made by Lage et al. [1]. The second possibility is that the human feature space should be iteratively expanded as new users interact with the system. This would require the development of new methods. Whether this adaptive approach is warranted is another important question for future study.

6. Conclusion

In this work, we explored key assumptions made about the human feature integration problem introduced in [1]. Through 2 user studies, we found that (1) a large and diverse set of human features can exist and be elicited from users; (2) eliciting human features does not appear to negatively impact willingness to use a model; (3) an algorithmic approach to human feature integration chooses qualitatively different features than users; and (4) people report a subjective preference for an algorithm selecting features for them, even though they are confident in their ability to select features themselves. Together, these results provide preliminary evidence towards the validity of the assumptions underlying the human feature integration problem described in [1]. We continue with ongoing work based on the studies described here to investigate the key assumption of whether human features do in fact improve predictive performance.

References

[1] I. Lage, S. Parbhoo, F. Doshi-Velez, Towards integrating personal knowledge into test-time predictions, 2024. URL: https://arxiv.org/abs/2406.08636. arXiv:2406.08636.

[2] M. Raghu, K. Blumer, G. Corrado, J. Kleinberg, Z. Obermeyer, S. Mullainathan, The algorithmic automation problem: Prediction, triage, and human effort, 2019.

[3] B. Wilder, E. Horvitz, E. Kamar, Learning to complement humans, 2020.

[4] D. Madras, T. Pitassi, R. Zemel, Predict responsibly: improving fairness and accuracy by learning to defer, Advances in Neural Information Processing Systems 31 (2018).

[5] B. J. Dietvorst, J. P. Simmons, C. Massey, Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them, Management science 64 (2018) 1155–1170.

[6] L. Cheng, A. Chouldechova, Overcoming algorithm aversion: A comparison between process and outcome control, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, Association for Computing Machinery, New York, NY, USA, 2023. URL: https://doi.org/10.1145/3544548.3581253. doi:10.1145/3544548.3581253.

[7] M. Gillies, R. Fiebrink, A. Tanaka, J. Garcia, F. Bevilacqua, A. Heloir, F. Nunnari, W. Mackay, S. Amershi, B. Lee, et al., Human-centred machine learning, in: Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems, 2016, pp. 3558–3565.

[8] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, W.-K. Wong, Too much, too little, or just right? ways explanations impact end users' mental models, in: 2013 IEEE Symposium on visual languages and human centric computing, IEEE, 2013, pp. 3–10.

[9] M. Jacobs, M. F. Pradier, T. H. McCoy, R. H. Perlis, F. Doshi-Velez, K. Z. Gajos, How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection, Translational psychiatry 11 (2021) 1–9.

[10] P. W. Koh, T. Nguyen, S. Tang Yew Siang, Mussmann, P. Emma, B. Kim, P. Liang, Concept bottleneck models, in: Proceedings of the 37th International Conference on Machine Learning, 2020.

[11] T. Wang, C. Rudin, F. Doshi-Velez, Y. Liu, E. Klampfl, P. MacNeille, A bayesian framework for learning rule sets for interpretable classification, The Journal of Machine Learning Research 18 (2017) 2357–2393.

[12] H. Lakkaraju, S. H. Bach, J. Leskovec, Interpretable decision sets: A joint framework for description

and prediction, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1675–1684.
[13] Y. Yang, I. G. Morillo, T. M. Hospedales, Deep neural decision trees, arXiv preprint arXiv:1806.06988 (2018).
[14] J. Yoon, J. Jordon, M. Van der Schaar, Invase: Instance-wise variable selection using neural networks, in: International conference on learning representations, 2018.
[15] N. Jethani, M. Sudarshan, Y. Aphinyanaphongs, R. Ranganath, Have we learned to explain?: How interpretability methods can learn to encode predictions in their interpretations., in: International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 1459–1467.
[16] M. Muszynski, Attention checks and how to use them: Review and practical recommendations, Ask Research and Methods 32 (2023) 3–38. doi:10.18061/ask.v32i1.0001.

A. Study 1: Salary Datasets

We used the following 4 salary datasets from Kaggle to compare the features suggested by users in study 1 to existing features that an ML model may expect to have access to.

- https://www.kaggle.com/datasets/thedevastator/jobs-dataset-from-glassdoor
- https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023
- https://www.kaggle.com/datasets/sahirmaharajj/employee-salaries-analysis
- https://www.kaggle.com/datasets/parulpandey/2020-it-salary-survey-for-eu-region

B. Study 1: Additional Results for Willingness to Use Model

Measuring Willingness to Use the Model Our secondary goal was to determine whether supplying additional features to the AI impacted users' willingness to use the model in the prediction the AI makes about them, either positively or negatively. Perhaps communicating important information with the AI improves trust, or perhaps this highlights the limitations of the AI. We measured trust and willingness to use the AI in 4 ways. First we asked participants whether they would prefer to use the AI with the human features they suggested or without them (phrased as: "I would prefer the manager to use the AI with the new information I suggested rather than the original AI without that information.") Then, we compared 3 measures of trust before and after requesting the additional human features for the AI: 1) whether the AI will lead to an accurate prediction (phrased as:"Using the AI, will the manager be able to predict my salary accurately?"); 2) whether they would prefer the prediction be made by the manager (the person making the final decision) directly without the AI; 3) how much error they think will be in the AI's prediction (we asked them to provide their actual salary, and looked at the difference between that and what they report they believe the AI will predict their salary to be).

Statistical results We ran 2-sided paired t-tests comparing each of the 3 sets of responses visualized in Figure 5 before and after communicating new features to the AI. None of the results are statistically significant. For the question "Using the AI the manager will be able to predict my salary accurately," the t-statistic is -1.78 and the p-value is 0.08. For the question "I would prefer the manager predicts my salary without the use of the AI," the t-statistic is 0.17 and the p-value is 0.86. For the absolute difference between the actual salary and the salary the user thinks the AI will predict, the t-statistic is 0.11 and the p-value is 0.92. With a p-value between 0.05 and 0.10, the first result does suggest that people may think the AI will be more accurate with their supplied features.

C. Experimental Setup for Study 2

Dataset for Training Models The dataset we used to design this study and train the necessary models is the NHANES–National health and nutrition examination survey from 2017. [4] This dataset

[4]Acquired from Kaggle: https://www.kaggle.com/datasets/cdc/national-health-and-nutrition-examination-survey

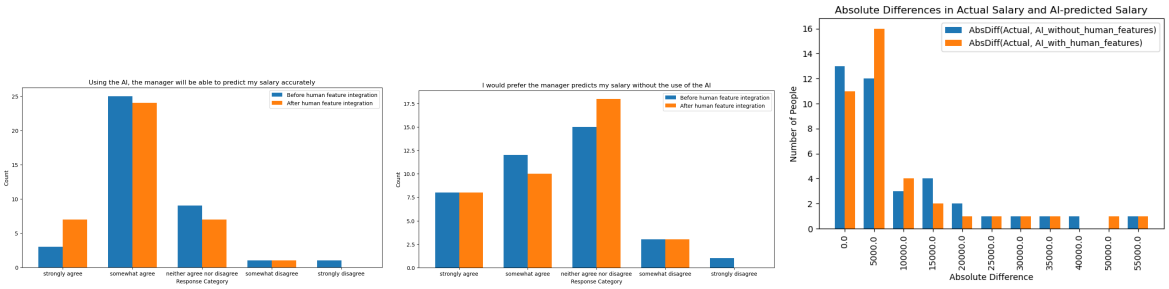


Figure 5: These graphs show the distribution of responses to the questions: "Using the AI, will the manager be able to predict my salary accurately?", "I would prefer the manager predicts my salary without the use of the AI." and the absolute difference between what the participant reports the AI will predict their salary to be, and what they report their salary to be. The first is repeated from the main text for completeness. The orange bars (right) show responses after participants were asked about additional features the AI should consider, and told the AI would consider them. The blue bars (left) represent responses before. None of these results are statistically significant. Small differences appear to favor the AI that uses human features except in the case of estimating how much error the model will have directly, where the picture is more ambiguous.

contains many questions posed to survey participants about different aspects of their health.

Within this dataset, we defined a label, a set of machine features, and a set of human features so that users would be familiar with the prediction task, none of the questions would be highly sensitive, we wouldn't be having users answer a particularly large number of questions, and entropy selection outperforms feature selection which suggests the utility of acquiring individualized human features in this context.

As the label, we are used peoples' self report of their general health on a 1-5 scale that we transformed into 3 labels: poor-fair, good, and very good-excellent. We chose a label that users would understand well (since they are determining it themselves) as upper bound on how well we would expect users to perform on the task of selecting their own human features. If people are unable to choose a predictive set of features when they understand the task well, they should do worse in contexts where they are not domain experts.

As the machine features, we're using a small subset of the demographics: age, gender, marital status and education level. Age was coded as: under 25, 25-34, 35-44, 45-54, 55-64, and 65 and above. Marital status was coded as: married or living with partner, vs. everything else. Education was coded as finished college, vs. everything else. All baseline features standardized to 0-1 range (the human features are already binary).

For the human features, we used 22 human features extracted from the diet and questionnaire files. We kept only features in one of the following subcategories of the survey: 'Alcohol Use' , 'Physical Activity', 'Diet Behavior & Nutrition' , 'Consumer Behavior' , 'Occupation' , 'Oral Health'. These categories were chosen as they generally include less sensitive information (i.e. that holds less privacy risks for our participants) than categories like sexual activity, or disabilities. Within those categories, we kept the subset of features that are binary, where each binary value occurs for at least 25 people, and that are recorded (with a non-missing value) for at least 3500 people. The full set of features we selected based on this is listed in Appendix D.

When training and evaluating our models, we kept the subset of instances with no missingness (including refusing to answer the answer question or answering that they don't know) across the label and the human and machine features. That resulted in a final dataset of 1535 participants.

Model training We trained all of the models on the NHANES dataset using the procedures from [1].

Figure 6 shows the results for the test f1 score as a function of queried features in the NHANES dataset (not including any data from our user study). This demonstrates that this is a case where as the number of features goes up, entropy selection outperforms feature selection, machine only baseline, and random masks. We designed this study to have this property, so it does not demonstrate that this

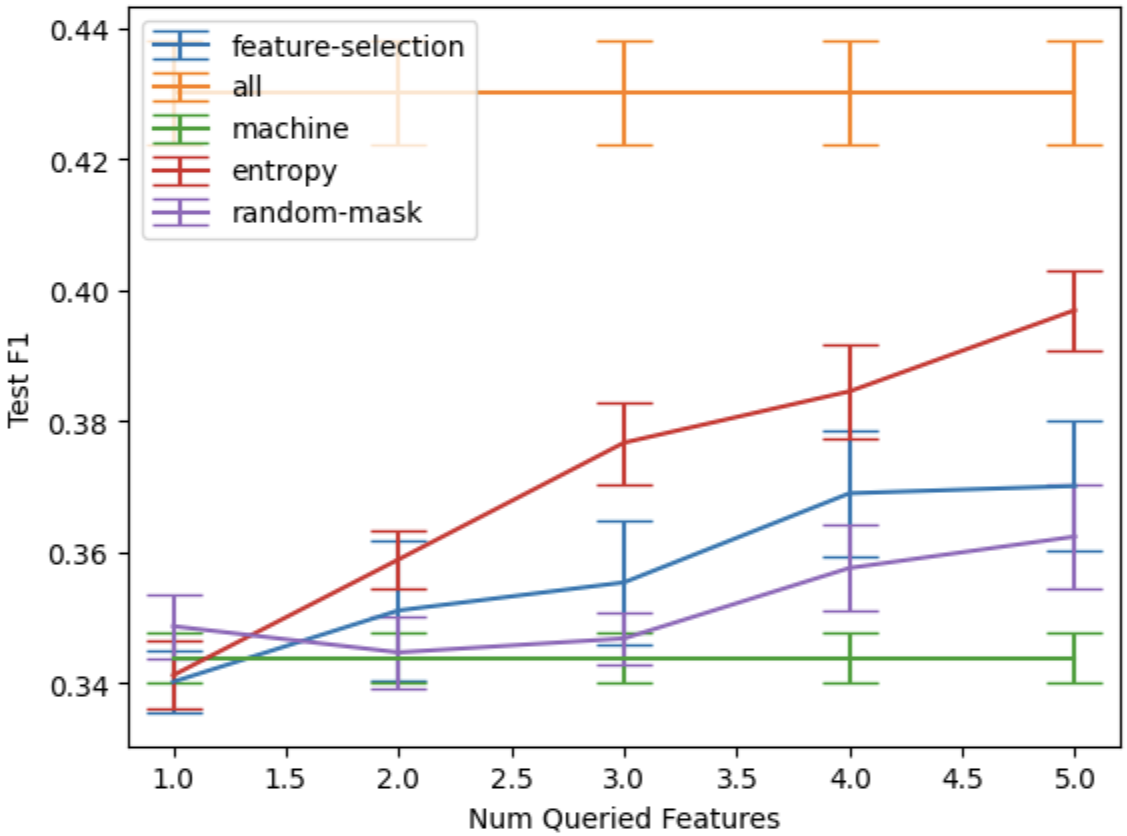


Figure 6: Results from the dataset used to train the models. We see that the entropy selection approach does outperform baselines as the number of queries grows.

property exists in general, but it does allow us to explore how effective users are at selecting their own features to include in a setting where this property exists.

User Study Design We collected data from Amazon Mechanical Turk through a Qualtrics survey. We first trained participants in the task, then gave them 2 opportunities to correctly answer a set of practice questions. Participants who answered at least one of those correctly proceeded to the main survey where we first asked them to report their general health on a scale of 1-5 to use as the label in our prediction task. We then told participants to consider an ML model based on only the 4 machine features, then we asked participants to first select the 5 features they believe will be most important for ML model to predict their general health, and after selecting those 5 human features, we had participants rank them from 1-5 based on their importance. After this, we asked several qualitative questions about the ranking including how confident people were in the ranking, and whether they would prefer to use their own ranking they had just generated, or a computer generated ranking. We also included an attention check that we used to exclude data from some users who did not answer it correctly. After this phase, we asked participants to self report the value of all of the human features in a survey of 22 questions, then to self report the values of the 4 machine features. After completing this, we asked participants to respond to a subset of questions based on the NASA-TLX questionnaire measuring mental demand, effort, performance and frustration to measure the difficulty of answering the 22 binary questions included in the survey as the human features. This allows us to extrapolate out the challenge of answering a complete set of perhaps hundreds of human features, vs. using our proposed approach of selecting only a few relevant ones for each user.

D. Human Features for Second User Study

As machine features, we used age, education level, marital status and gender. As human features, we used features from the following categories for 22 total human features: 'Alcohol Use' , 'Physical Activity', 'Diet Behavior & Nutrition' , 'Consumer Behavior' , 'Occupation' , 'Oral Health'. We used people's self-reported health score with 3 label categories: poor-fair, good, and very good-excellent. We trained the entropy selection approach using the setup described in [1].

- ALQ101: The next questions are about drinking alcoholic beverages. Included are liquor (such as whiskey or gin), beer, wine, wine coolers, and any other type of alcoholic beverage.In any one year, have you/has SP had at least 12 drinks of any type of alcoholic beverage? By a drink, I mean a 12 oz. beer, a 5 oz. glass of wine, or one and half ounces of liquor.
- ALQ151: Was there ever a time or times in your/SP's life when you/he/she drank DISPLAY NUMBER or more drinks of any kind of alcoholic beverage almost every day?
- CBQ535: The last time when you ate out or bought food at a fast-food or pizza place, did you see nutrition or health information about any foods on the menu? SP interview version: The last time when you/SP ate out or bought food at a fast-food or pizza place, did you/he/she see nutrition or health information about any foods on the menu?
- CBQ552: Think about the last time you/SP ate at a restaurant with a waiter or waitress. Is it a chain-restaurant?
- CBQ580: The last time you ate at a restaurant with a waiter or waitress, did you see nutrition or health information about any foods on the menu? SP interview version: Did you/SP see nutrition or health information about any foods on the menu?
- CBQ596: Next I'm going to ask a few questions about the nutritional guidelines recommended for Americans by the federal government. Have you/Has SP heard of My Plate?
- OHQ610: In the past 12 months, did a dentist, hygienist or other dental professional have a direct conversation with you/SP about... ...the benefits of giving up cigarettes or other types of tobacco to improve your/SP's dental health?
- OHQ612: (In the past 12 months, did a dentist, hygienist or other dental professional have a direct conversation with you/SP about...) ... the dental health benefits of checking your/his/her blood sugar?
- OHQ614: (In the past 12 months, did a dentist, hygienist or other dental professional have a direct conversation with you/SP about...) ...the importance of examining your/his/her mouth for oral cancer?
- OHQ770: During the past 12 months was there a time when (you/SP) needed dental care but could not get it at that time?
- OHQ835: The next questions will ask about the condition of your/SP's teeth and some factors related to gum health. Gum disease is a common problem with the mouth. People with gum disease might have swollen gums, receding gums, sore or infected gums or loose teeth. Do you/Does SP think you/s/he might have gum disease?
- OHQ850: Have you/Has SP ever had treatment for gum disease such as scaling and root planing, sometimes called "deep cleaning"?
- OHQ855: Have you/Has SP ever had any teeth become loose on their own, without an injury?
- OHQ860: Have you/Has SP ever been told by a dental professional that you/s/he lost bone around [your/his/her] teeth?
- OHQ865: During the past three months, have you/has SP noticed a tooth that doesn't look right?
- OHQ880: Have you/Has SP ever had an exam for oral cancer in which the doctor or dentist pulls on your/his/her tongue, sometimes with gauze wrapped around it, and feels under the tongue and inside the cheeks?
- OHQ885: Have you/Has SP ever had an exam for oral cancer in which the doctor or dentist feels your/his/her neck?

- PAQ605: Next I am going to ask you about the time you spend/SP spends doing different types of physical activity in a typical week. Think first about the time you spend/he spends/she spends doing work. Think of work as the things that you have/he has/she has to do such as paid or unpaid work, household chores, and yard work. Does your/SP's work involve vigorous-intensity activity that causes large increases in breathing or heart rate like carrying or lifting heavy loads, digging or construction work for at least 10 minutes continuously?
- PAQ620: Does your/SP's work involve moderate-intensity activity that causes small increases in breathing or heart rate such as brisk walking or carrying light loads for at least 10 minutes continuously?
- PAQ635: The next questions exclude the physical activity at work that you have already mentioned. Now I would like to ask you about the usual way you travel/SP travels to and from places. For example to school, for shopping, to work. In a typical week do you/does SP walk or use a bicycle for at least 10 minutes continuously to get to and from places?
- PAQ650: The next questions exclude the work and transport activities that you have already mentioned. Now I would like to ask you about sports, fitness and recreational activities. In a typical week do you/does SP do any vigorous-intensity sports, fitness, or recreational activities that cause large increases in breathing or heart rate like running or basketball for at least 10 minutes continuously?
- PAQ665: In a typical week do you/does SP do any moderate-intensity sports, fitness, or recreational activities that cause a small increase in breathing or heart rate such as brisk walking, bicycling, swimming, or volleyball for at least 10 minutes continuously?