It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy

Andrew Bell
New York University
New York, United States
alb9742@nyu.edu

Ian René Solano-Kamaiko
New York University
New York, United States
isk273@nyu.edu

Oded Nov
New York University
New York, United States
onov@nyu.edu

Julia Stoyanovich
New York University
New York, United States
stoyanovich@nyu.edu

ABSTRACT

To achieve high accuracy in machine learning (ML) systems, practitioners often use complex "black-box" models that are not easily understood by humans. The opacity of such models has resulted in public concerns about their use in high-stakes contexts and given rise to two conflicting arguments about the nature — and even the existence — of the accuracy-explainability trade-off. One side postulates that model accuracy and explainability are inversely related, leading practitioners to use black-box models when high accuracy is important. The other side of this argument holds that the accuracy-explainability trade-off is rarely observed in practice and consequently, that simpler interpretable models should always be preferred. Both sides of the argument operate under the assumption that some types of models, such as low-depth decision trees and linear regression are more explainable, while others such as neural networks and random forests, are inherently opaque.

Our main contribution is an empirical quantification of the trade-off between model accuracy and explainability in two real-world policy contexts. We quantify explainability in terms of how well a model is understood by a human-in-the-loop (HITL) using a combination of objectively measurable criteria, such as a human's ability to anticipate a model's output or identify the most important feature of a model, and subjective measures, such as a human's perceived understanding of the model. Our key finding is that explainability is not directly related to whether a model is a black-box or interpretable and is more nuanced than previously thought. We find that black-box models may be as explainable to a HITL as interpretable models and identify two possible reasons: (1) that there are weaknesses in the intrinsic explainability of interpretable models and (2) that more information about a model may confuse users, leading them to perform worse on objectively measurable explainability tasks. In summary, contrary to both positions in the literature, we neither observed a direct trade-off between accuracy and explainability nor found interpretable models to be superior in terms of explainability. It's just not that simple!

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

machine learning, explainability, public policy, responsible AI

ACM Reference Format:
Andrew Bell, Ian René Solano-Kamaiko, Oded Nov, and Julia Stoyanovich. 2022. It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3531146.3533090

1 INTRODUCTION

The current resurgence of research in Artificial Intelligence (AI) and in particular, in the sub-discipline of machine learning (ML), has resulted in ML systems being widely adopted in various organizations within society [18, 50, 52, 56, 58, 60]. Public policy programs responsible for high-stakes decisions in critical domains such as education [6, 34, 37, 59] and housing [7, 57] are increasingly relying on the predictions made by ML models to assist in decision-making alongside human users. This interaction model, where an ML system makes a prediction and a human ultimately makes a final determination informed by the system's predictions, is commonly referred to as a "human-in-the-loop" (HITL) system. For example, a school administrator may use an ML system to predict the risk that a student will drop out of school and then assign an at-risk student to a special tutoring program [2]. Similarly, a doctor may use an ML system to forecast a patient's risk of developing a particular disease for early intervention [3].

In tandem with this uptick in adoption, researchers, AI practitioners, and journalist alike have shown that ML systems operating in public policy contexts run the risk of causing significant harm

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533090

to members of already marginalized and underrepresented groups [13, 25, 40, 51]. For example, a recent study revealed that lending algorithms may discriminate against Latinx and African-American borrowers [7, 26], and that educational risk-assessment algorithms have performed worse for minority students [37, 59]. The implementation of problematic ML models in public policy use-cases is, among many factors, due to the increased use of opaque models. Such models, commonly referred to as "black-boxes", are based on complex predictive algorithms that are not always intelligible to humans, posing both practical and ethical concerns [58, 61, 71]. Black-box models exist in contrast to "interpretable" models, which can be more easily understood by humans. In an effort to understand black-box models, there has been an increase in research and tooling developed around interpretability and explainability [29, 42], leading to the development of commonly used toolkits such as LIME, SHAP, and SAGE [16, 45, 55]. Interpretability and explainability are seen as a means to surface and address fairness concerns, to help in ML system debugging, to facilitate auditing and oversight, to improve trust and increase adoption by HITL users, and to support recourse for those affected by model decisions like appealing a loan denial [5].

Terminology. In this work, we use Rudin's definition of "interpretability", who states that "interpretable ML focuses on designing models that are inherently [human]-interpretable" [58]. Importantly, the terms "interpretable" and "interpretability" refer to a specific class of models, like decision trees, linear models, and rules-lists. While there is no universal definition of "explainability" [46, 46], it generally refers to the extent to which ML models are understandable to humans. We use an adaptation of Miller's work, and define "explainability" as the extent to which someone can accurately predict and understand the output of an ML model [47, 48].

Claims About the Accuracy vs. Explainability Trade-off. The ML community is entertaining two contrasting claims about the trade-offs between interpretable and black-box models. The first is that, while interpretable models are more explainable they lack the accuracy of black-box models, suggesting that accuracy and explainability are inversely related [38]. The second and contrary claim is that there is little observable trade-off in accuracy between black-box and interpretable models and that consequently, interpretable models should be favored. Rudin [58] has advocated that the ML community abandons the use of black-box models in high-stakes decision contexts in favor of using human-interpretable models. The latter claim is supported by an increasing number of cases showing that the difference in the accuracy of black-box models as compared to simpler interpretable models is often negligible in public policy contexts [9, 20, 62]. Yet, there is a significant gap in the literature on empirically characterizing the this trade-off [24, 28], in part due to the difficulty of quantifying the explainability of both interpretable and black-box models.

Research Questions. In this paper, we take steps to fill this research gap. We study the trade-off between accuracy and explainability for the end users of black-box and interpretable models in public policy use-cases and seek to answer two related research questions: (1) how can we quantify explainability? and (2) how can we quantify the trade-off between accuracy and explainability?

Summary of Findings. We conducted a large user study to measure explainability of 4 model types (2 interpretable and 2 black-box). We quantified explainability using two objectively measurable tasks: (1) anticipating the output of a model; and (2) identifying the most important feature. We also used two subjective measures: (1) the user's perceived understanding of the model; and (2) the user's perceived confusion.

Key finding. We found that there is no statistically significant difference in explainability for HITL users working with black-box models versus interpretable ones. In fact, black-box models, both with and without SHAP explanations, lead to the best performance by users on both objectively-quantifiable tasks in both policy domains. Figure 1 summarizes this key finding.

Additional findings. Our key finding was driven by two factors. The first is that additional information about a model may not improve its explainability. Providing more information to a HITL (e.g., SHAP explanations for black-box models or "intrinsic" explainability mechanisms such as a tree diagram for decision trees) was only useful if it did not confuse the user. Local explanations, like those from SHAP, are most useful to users in cases where these explanations are significantly different from global explanations about a model (e.g., feature importance). Additionally, the utility of these explanations is heavily dependent on the explainability task asked of the user.

The second factor is that there are weaknesses in the intrinsic explainability mechanisms of interpretable models. For example, when users were asked to identify the most important feature of a decision tree model and were presented with the tree diagram, they often defaulted to selecting the feature found at the root of the tree regardless of whether or not the feature was the most important. This result was observed despite participants being given detailed instructions about the overall system characteristics, feature importance, and the tree diagram.

Finally, we find that model accuracy and explainability are not necessarily inversely related. In the problem contexts we investigated, black-box models only outperformed interpretable models when accuracy was measured using a context-specific metric (e.g., precision score at the top 25% of the population). Otherwise, when using more general metrics (e.g., accuracy score), black-box and interpretable models performed similarly. Contrary to a common belief held by ML practitioners, black-box models may often be both the most accurate and the most explainable models to end users. With respect to black-box models, users are able to understand *what the model does* without necessarily having to understand *how it works*.

Importantly, we are not making a tacit endorsement of the use black-box models in all contexts; rather, we acknowledge that black-box models may not be as unexplainable to users as often believed. This work represents a single step towards a larger goal of understanding the complexity of what explainability practically means for HITL users.

Contributions. Our main contribution is an empirical quantification of the trade-off between model accuracy and explainability. We define a robust measure of explainability, consisting of two objectively-verifiable user tasks ("Anticipating the System Output" and "Identifying the Most Important Feature") and two survey constructs concerning the users' perceived understandings of the model.

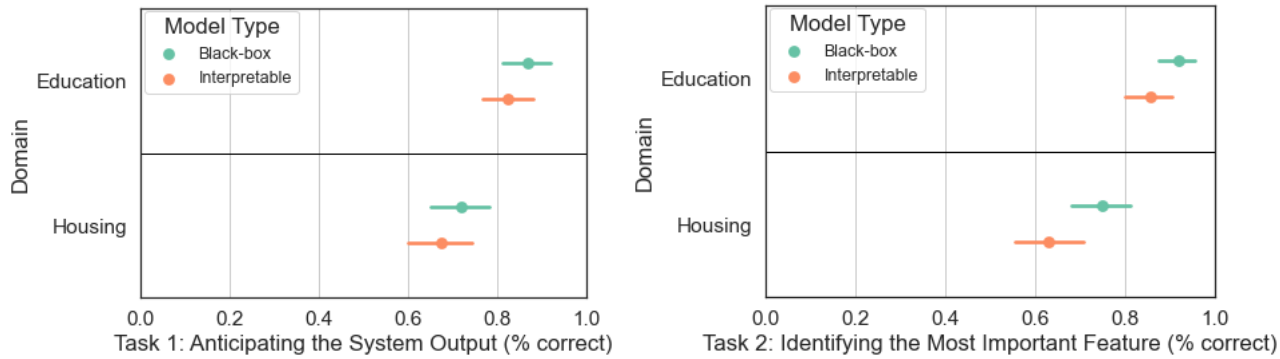


Figure 1: (a) Mean performance and 95% CI of participants on explainability Task 1: Anticipating the System Output, where participants were presented with the profile of a student (or house) and asked to identify how the model would classify that profile; (b) Mean performance and 95% CI of participants on explainability Task 2: Identifying the Most Important Feature, where participants were presented with the profile of a student (or house) and the model output, then asked to identify the most important feature contributing to that output. (Number of observations in each group: Education|BB = 160, Education|Interpretable = 160, Housing|BB = 160, Housing|Interpretable = 160)

We present a novel and generalizable methodology for empirically quantifying this trade-off and use this method in two real-world public policy scenarios.

2 RELATED WORK

Work in explainable AI (XAI) has expanded considerably in recent years due to the growth and proliferation of black-box models. ML/AI algorithms have permeated many different aspects of public policy and are often implemented to support human decision-making. While some ML applications may not require HITL users to understand how these systems work, in many policy contexts, it is critical for the human operators of these systems to have an understanding of the models that underpin them [27, 53, 63, 65, 70]. For example, physicians prefer to use ML systems that provide model-agnostic explanations. Researchers have also uncovered a significant relationship between understanding and trust in ML systems [8, 22, 44].

ML practitioners generally accept three types of models as interpretable: rules-based models, decision trees, and linear models [31]. These three model types are viewed as having intrinsically explainable mechanisms [46], such as the tree diagram for decision trees and the linear formula for linear models. Other model types, like random forests and neural networks, are considered black-boxes. Importantly, Guidotti et al. [31] point out that the distinction between interpretable and black-box models lacks an attention to model complexity. The technical definition of model complexity is the number of regions (i.e., the parts of the model) for which the boundaries are defined [33]. Here, "model boundaries" refers to certain parameters of a model that influence how inherently understandable that model is to humans. As an example, the model complexity of a decision tree is proportional to the depth of the tree. Prior research found that the understanding of a model is negatively correlated with its complexity and that decision trees are among the model types best understood by users [4].

Researchers have made a number of attempts to "open up" black-box models by creating tools that provide explanations. In contrast to interpretability, which is inherent in low-complexity interpretable models, these methods are post-hoc techniques that are applied to already trained models. Some of the most popular post-hoc explainability methods are LIME, QII, SHAP, and SAGE [16, 17, 19, 45, 54]. Despite these efforts, researchers have called into question the utility of these methods, cautioning against their use in real-world contexts, especially for ML systems making decisions in high-stakes, public policy contexts [58]. Critics of post-hoc explainability methods point out that LIME and SHAP are designed specifically for offering local interpretability of models and, subsequently, offer explanations only about how the model works on a particular input. Another criticism is that, in many cases, such methods are approximations of the black-box they are trying to explain [71] and do not reflect the *actual* underlying model. Finally, these methods may be vulnerable to adversarial attacks [61].

Generally, the tension between high performing yet difficult to understand black-box models on one hand and worse performing yet directly interpretable models on the other, is often referred to as the accuracy-interpretability trade-off [28, 58]. In this work and in alignment with recent research [1, 32, 35, 43, 44, 47, 48, 67], we seek to re-frame the conversation away from the "coarse" notion of interpretability that relates to entire classes of models. Instead, we couch it in terms of explainability, which is concerned with what a model does (rather than with how it works) and with how well a model's decisions are understood by a user.

In part, the scarcity of empirical research on the accuracy-explainability trade-off is a result of the difficulty of effectively quantifying explainability [32]. Explanations in ML are intended for end-users and, therefore, efforts to quantify explainability often

involve user studies during which users complete tasks and answer questions [1, 35, 67]. User studies on explainability often include subjective and objective measures of user understanding. Subjective measures include questionnaires about the users' point of view, such as perceptions of their understanding or ratings on the quality of an explanation [43, 44]. In contrast, objective measures include the performance of explainability-related tasks, such as identifying the most important feature of a model [32].

Another consideration is how the design of ML systems impacts explainability [23, 68]. For example, practitioners need to be cognizant of an explanation's *scope* - whether the explanation is *local* (the explanation is about a single instance or model output) or *global* (the explanation is about the entire model). In one study, researchers found that varying the scope of an explanation between local and global influenced individuals' perceptions of whether or not an ML model was fair—even when the underlying model did not change [42]. Other studies have found that the implementation of explainability mechanisms varies widely among industry experts but that most techniques focus on features, such as model feature importance [10, 36]. Other studies note that while post-hoc explainability techniques such as SHAP and LIME are useful, they also contain issues with regards to model volatility and have outlined how LIME has given unexpected results [10, 36]. Furthermore, researchers have found that blindly deploying post-hoc explainability techniques can actually *hurt* users' performance on certain tasks related to their understanding of ML models [39].

Lastly, the literature on explainability for HITL users is filled with counter-intuitive results. For example, research shows that providing more information in ML model explanations can actually decrease users' perceived understandings due to information overload [42, 49]. Researchers have also found that ML explanations can generate positive increases in user confidence but that this effect is negligible with respect to the quality of their final decisions [69]. The latter finding is an extension of the performance-use paradox, wherein the users of an ML system in an employment setting preferred using a system where the explanations made no sense to them, as opposed to not having the explanations at all [70]. Furthermore, there is evidence that explanations can have the positive effect of increasing trust in ML systems, but also the negative effect of facilitating a false sense of trust in the same type of systems [8, 22, 30, 44]. Overall, a nuanced approach to explainability, and generally to the design and implementation of ML systems, is paramount to understanding the accuracy-explainability trade-off.

3 PUBLIC POLICY USE-CASES

This work focuses on ML for public policy, an area of which problems have special considerations that impact the accuracy-explainability trade-off. Three such considerations are: (1) ML systems in public policy settings are almost always implemented alongside a human user who must understand the system sufficiently to take responsibility for the decisions, (2) the transparency of these systems is often (and with increasing frequency) legally mandated due to their public and high-stakes nature, and (3) accuracy is typically measured relative to a resource constraint (see section 4.2).

The two problem contexts we selected consist of a Portuguese student performance system [15], where the prediction task is to identify students at risk of failure in order to provide additional school resources (referred to as the "education" problem) and a housing price estimation system from King County, WA, USA [64], where the model estimates the market value of the property in order to prioritize tax assessor inspections (referred to as the "housing" problem). In the proposed housing problem, the objective is to predict what the property price would be if the home had sold in the previous year. This information is important because homes are taxed as a function of their market value in King County, WA. Property taxes fund a large portion of local government and for property owners this is a large and visible tax payment. Taxpayers have the right to know if they are paying their "fair share" of property taxes but due to governmental resource constraints, local governments need some way to prioritize how property tax assessors are deployed into the field to evaluate or re-evaluate the homes in question. The problem domains and the datasets used to build these systems are summarized in Table 1.

We chose these two contexts because they use publicly available, real-world data. When selecting domains to study, we only considered problems where we felt that the use of responsibly designed and thoroughly validated ML systems was ethical. For this reason, problem spaces such as criminal justice systems like COMPAS [40], where the objective is to make recidivism risk-predictions were excluded. Furthermore, researchers have demonstrated issues caused by similar systems in education [6, 34, 37, 59] and housing [7, 57], highlighting the importance of understanding these types of systems to mitigate potential harms. During the use-case selection process we only considered tabular datasets because they make up a large portion of ML public policy problems. As evidence for this, the organization Data Science for Social Good (DSSG) lists approximately 80 projects on their website[1] at the intersection of ML and public policy and at least 75% of them used tabular data.

4 METHODS

The first step of our methodology involved carefully conducting exploratory data analyses and pre-processing on the Education and Housing datasets (section 4.1). Next, we trained numerous classifiers (section 4.2) and quantified model accuracy based on a variety of common accuracy metrics and ultimately selected the best models using overall accuracy and precision@25%. We report the results of all the accuracy metrics we measured in Appendix D. Finally, we measured the explainability of the most accurate models in a user study (section 4.3). Our explainability metrics are described in detail in section 4.3.2 and combine objectively measurable factors such as a user's ability to anticipate a model prediction or to point out the most salient feature as well as subjective factors that quantify a user's perceived understanding of a model.

4.1 Data Processing and Exploratory Data Analysis

After selecting our two public policy use-cases (Education and Housing), we processed the data for modeling and conducted an Exploratory Data Analysis (EDA). Two issues were discovered during the EDA process. First, there was a data leakage issue in the

[1] https://www.dssgfellowship.org/projects/

Policy domain	Education	Housing
Number of records	1,044	21,613
Number of features	33	20
Target variable	Grade in final year of high school (scale from 1 to 20)	Sales price
Prediction task	Will the student fail (grade ≤ 10)?	Will the house's sale price be $\geq \$645,000$?
Timespan	2005-09-01 to 2006-06-31	2014-05-01 to 2015-05-31
Subgroups	Sex, parent's education level	None
Human users	School administrators, teachers	Government officials, property tax assessors
Intervention	High risk students will receive additional tutoring	Houses will be audited by tax assessors
Resource constraint	Special tutoring available for up to 25% of students	Assessors can inspect up to 25% of houses

Table 1: Description of policy contexts and the associated datasets used in our evaluation.

Education dataset. Two features (absences and failures) were aggregated across all three years of data, including the year for which performance is being predicted. Unfortunately, there was no way to disentangle these features into separate annual variables, and as a result, the leakage improves model accuracy metrics by making the prediction task "easier." However, since this accuracy improvement applies to both interpretable and black-box models, we believe it poses no harm to our overall study objective. Second, during EDA on the Housing dataset we noticed that several data records contained errors. Subsequently, we dropped 10 data records that either consisted of 0 bedrooms, 0 bathrooms, and/or 33 bedrooms. We found it reasonable to assume that these were likely input errors and that their removal would not have any significant effect on the overall model training and prediction outcomes.

After preliminary model training, we noticed the models' most important features were always zip codes. For our study purposes, this was not ideal as users would not be able to use their domain expertise unless they were familiar with those particular zip codes. Furthermore, using zip codes — which are strongly correlated with race [21, 41] — to predict housing prices for tax purposes raise concerns about redlining, which the Fair Housing Act of 1968 explicitly outlawed as a practice in the United States [14]. For these reasons, we omitted zip codes from our training data.

4.2 Model Training

For each policy use-case, we trained eight different classifier types: three black-box models (XGBoost, extra trees, random forests), four interpretable models (decision tree, linear regression - Ridge, linear regression - Lasso, logistic regression), and one baseline model (dummy classifier) to benchmark performance. We chose these classifiers because they are representative of the types of models commonly used in ML for public policy systems [9, 12, 20, 56]. Our model training process followed ML for public policy industry standards and best practices. The classifiers were tuned using a parameter sweep on a large hyperparameter grid and the performance of each model was validated using stratified k-fold cross-validation. To ensure the robustness of our results, the models were also validated over a range of different random seeds and different validation methods such as classic k-fold and train-test-split.

Each model was evaluated on seven different metrics including Area Under the Receiver Operating Characteristic Curve (ROC AUC) score, F1 score, accuracy score, precision score at 10%, precision score at 25%, recall score at 10%, and recall score at 25%. These metrics were chosen because they reflect the performance measures typically used in ML for public policy [9, 12, 20, 56]. Importantly, we included metrics at a certain percentage of the population (often referred to as metrics at k) because they are more commonly used in ML for public policy in order to specifically addresses real-world resource constraints [2, 12, 66]. In both policy contexts, we assumed that intervention resources were limited to 25% of the population (see Table 1). For the sake of clarity, in this paper we will focus on two metrics: accuracy score, computed as the overall percentage of correct predictions, and precision@25%, computed as the percentage of true positives out of all predicted positives among the top 25%, when model outputs are sorted by the score $\in [-1, 1]$ associated with the prediction. A complete list of performance results across all metrics is available in Appendix D.

After completing model training, we selected three models for each domain to use in the explainability survey design based on their performance with respect to precision@25%, since this was the most policy-relevant metric. The three models selected were random forest, linear regression (Ridge for Education and Lasso for Housing), and decision tree. Subsequently, these models were used to design the survey and its materials, described next.

4.3 Survey

To measure the explainability of the chosen ML models, we conducted an IRB-approved human subjects research study. We gathered data on four explainability metrics including two metrics requiring users to correctly perform tasks pertaining to the inputs and outputs of the system and two subjective measures of participants' perceived understanding and confusion about the system (these metrics are described in detail in the section 4.3.2).

4.3.1 Recruitment.

We recruited participants through Prolific, an online research recruitment platform. Participant compensation was set at \$15/hour; participants were estimated to complete the study in 15 minutes or less (the education study average reward was \$28.73/hr and the housing study average reward was \$27.44/hr). In both studies, participants were pre-screened by current country of residence (United States), age (over 18), and fluent languages (English). For the education domain, participants were pre-screened as being in the Education & Training employment-sector and the Primary/Secondary

K-12 Education industry. For the housing study, participants were from the Real Estate Rental & Leasing industry.

Each study had 168 participants to ensure our survey would have statistical power. While all Prolific participants are anonymous, we did receive some self-reported demographic data and survey-specific metrics. The education study consisted of 129 female and 39 male respondents, who spent an average of 8 minutes and 50 seconds on the survey questions. The housing study consisted of 112 female and 54 male respondents, and 2 respondents who did not specify their gender. Housing study participants spent an average of 10 minutes and 14 seconds on the survey questions.

4.3.2 *Survey Design.*

We designed the survey to emulate how HITL users interact with ML systems in real-world policy settings [70]. The survey began by providing participants with a contextual overview about ML in their respective policy domain and asked them to complete tasks about specific ML systems. Next, we had participants interact with two of four possible ML systems in order to measure the explainability of 4 model types: two black-box models (one with SHAP and one without), and two interpretable models (a decision tree and a linear regression). Note that we chose to provide SHAP explanations for one black-box model because it emulates how black-box models are most commonly implemented in-practice [10, 36].

We administered two surveys that were identical except for domain-specific details such as the features used to build the ML models. The survey flow as well as a sample can be found in Appendix A. The main steps are detailed below:

(1) *Model Type.* The participants were randomly presented one of the four ML systems being studied (black-box no SHAP, black-box with SHAP, linear regression, or decision tree):
   (a) Information about the system:
      (i) *Model Overview.* Included details about the system's purpose, inputs and outputs, and the global feature importance of the underlying ML model.
      (ii) *Additional Model-Specific Information.* Information varied based on the model type randomly selected in step (1). For systems that used interpretable models, participants viewed the intrinsic explainability mechanism of the respective model: linear formula with cut-off threshold for linear regression, and the tree diagram for decision tree, respectively. For the black-box model *with* SHAP explanations, participants saw SHAP diagrams presenting local feature importance. For the black-box model *without* SHAP explanations, no additional information was presented.
   (b) Survey tasks and constructs:
      (i) ***Task 1: Anticipating System Output (ASO).*** This task required users to anticipate the output of an ML system given information about the system. Users received a random data profile and were asked to determine how the system would classify that profile. This task was adapted from Miller's work [47] where explainability is the extent to which someone can accurately understand and predict the output of an ML model. The task was intended to measure local explainability of the model by having participants reason about how the system worked for a particular data point.
      (ii) ***Task 2: Identifying the Most Important Feature (IMIF).*** In this task, participants were given both a random data profile *and* the corresponding system output. They were then asked to identify the feature that was the most important overall to the system's prediction. In contrast to the first task, this task measured the global explainability of the model since the model's feature importance did not vary with respect to each data point. Overall, there were four random data profiles and we ensured that participants would not see the same profile for both tasks.
      (iii) *System Understanding.* We sought to measure the participants' perceived understanding of the system. To do this, we included an eight item, 5-point Likert scale questionnaire aimed at measuring their perceived understanding of the ML systems. The eight items were broken into a five item construct that assessed how well the participant felt they understood the system (called ***Understood System***) and a three item construct that evaluated whether or not the participants found the system confusing (called ***System Confusing***). The questions in this section were adapted from work by Lim [43].
(2) Participants repeated step (1) but viewed a different model type. Importantly, if the participant previously viewed an interpretable model, they would next view a black-box model (and vice versa).
(3) Participants completed four final background questions about their technological aptitude, industry experience, and whether they were familiar with the data presented before taking the survey.

To summarize our design: Each participant interacts with one black-box and one interpretable model, in a random order. For each model, they complete 1 ASO task, 1 IMIF task, and complete the system understanding questionnaire. Note that by randomizing the order in which participants see each model, we eliminate any "learning effect" that may occur from interacting with the first model.

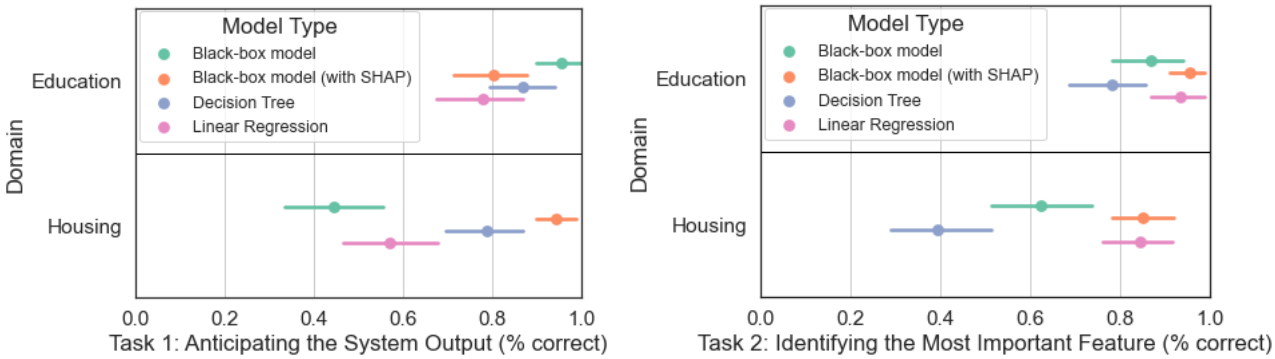
5 RESULTS

Appropriately anonymized data and our analysis methodology are available at https://github.com/DataResponsibly/accuracy-explainability-tradeoff.

5.1 Explainability

5.1.1 *Task 1: Anticipating the System Output (ASO).*

In the Anticipating the System Output (ASO) explainability task, participants were given a profile of a student or house, and then asked to anticipate the model's output based on the given input. Users were asked to try and understand what the models would do for a particular (local) data point. While one might expect users to perform better at anticipating the output for interpretable models as compared to black-box models, we found on average no statistically significant difference in the ASO task performance between the



(a) (b)

Figure 2: (a) Mean performance and 95% CI of participants on explainability Task 1: Anticipating the System Output, where participants were presented with the profile of a student (or house) and asked to identify how the model would classify that profile; (b) Mean performance and 95% CI of participants on explainability Task 2: Identifying the Most Important Feature, where participants were presented with the profile of a student (or house) and the model output, then asked to identify the most important feature contributing to that output. (Number of observations in each group: Education|BB = 69, Education|BB (w/ SHAP) = 91, Education|DT = 83, Education|LR = 77, Housing|BB = 72, Housing|BB (w/ SHAP) = 88, Housing|DT = 76, Housing|LR = 84.)

end users of the interpretable and black-box models (all statistics and p-values for this paper be found in Appendix B in Table 3).

However, there were significant pairwise differences when comparing ASO performance for specific model types (decision tree, linear regression, black-box (no SHAP), black-box (SHAP); see Figures 2, 3). Therefore, while the explainability for the overarching category of black-box or interpretable did not differ, the individual model types did matter. We found that participants were most successful at completing the ASO task when presented with the black-box model, as seen in Figure 2 (a). Interestingly, in the education domain, participants who viewed the black-box model *without SHAP* had the best ASO performance. We expound upon these counter-intuitive results later in the discussion section of the paper. In the housing domain, participants had best ASO performance when presented with the black-box model *with* SHAP. This makes intuitive sense, since a strength of SHAP is to assist in the *local* explainability of a model.

5.1.2 Task 2: Identifying the Most Important Feature (IMIF).
In the Identifying the Most Important Feature (IMIF) explainability task, participants were given the system input and output and asked to identify the most important feature overall to the system. Participants' performance on the IMIF task is summarized in Figure 2 (b). Unlike the ASO task, the IMIF task is related to *global model explainability*. The results for the IMIF task were very similiar to those found for ASO.

We observed that there was no statistically significant difference in IMIF performance, regardless of whether a participant was presented with an interpretable or a black-box system. However, there were statistically significant pairwise differences across model types (see Table 3). We found that participants were most successful at the IMIF task when presented with either the linear regression or a black-box with SHAP system. This finding was consistent with our expectations. The intrinsic explainability mechanism of a linear regression model—that is, showing the linear formula and coefficient weights—lends itself well to the global explainability of a model. Additionally, participants were highly successful at the IMIF task when presented with the black-box with SHAP because local explainability aligned with the models' global explainability. The most important feature in the SHAP diagram was almost always the most important feature to the model overall. Unexpectedly, participants had the lowest IMIF task performance when presented with the decision tree system. Our hypothesis on the cause of this phenomenon is presented in §6.

Notably, our findings indicate that our design of the IMIF task may not be a robust measure of global model explainability. Per our experiment design, users were presented with an individual profile yet asked to identify the most important feature to the model *overall*, as seen in Figure 8 (b). Some users were confused by this, and instead tried to identify the most important feature for the classification of that particular profile, rather than for the entire model. The impact of this confusion is that the results we found for IMIF performance are likely overly pessimistic.

5.1.3 System Understanding.
The System Understanding measure was made up of two constructs: Understood System and System Confusing. Participants' scores for these measures can be seen in Figure 3. Interestingly, significant differences were found for the System Confusing construct between black-box and interpretable models. Users from both the housing and education domains perceived the interpretable systems to be *more confusing* than the black-box systems, as seen in Figure 3. This result was consistent with the narrative that had emerged from evaluating the ASO and IMIF task results, where participants were more successful when presented with black-box systems.

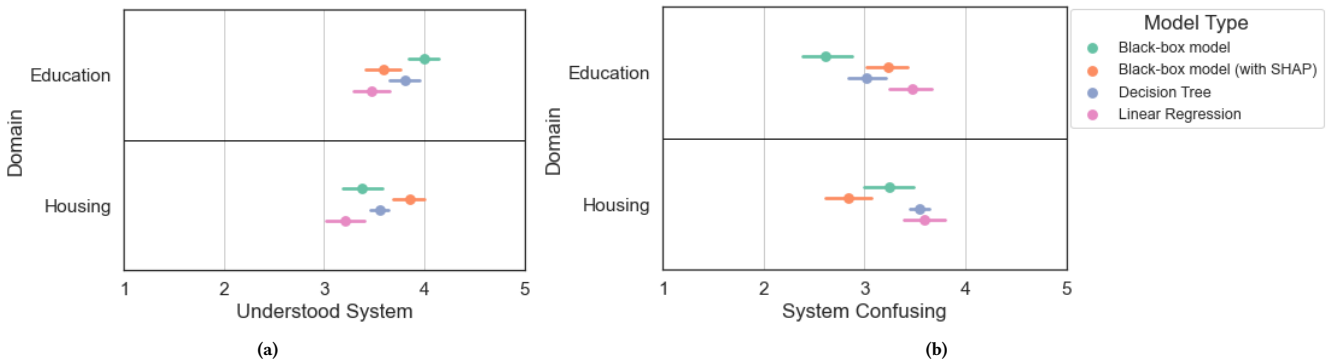


Figure 3: (a) Participants' responses to the 5-item *Understood System* construct [43] (b) Participants' responses to the the 3-item *System Confusing* construct [43]. (Number of observations in each group: Education|BB = 69, Education|BB (w/ SHAP) = 91, Education|DT = 83, Education|LR = 77, Housing|BB = 72, Housing|BB (w/ SHAP) = 88, Housing|DT = 76, Housing|LR = 84.)

5.2 Accuracy

The accuracy of black-box and interpretable models for the education and housing datasets are presented in Table 2. For clarity, we focus on two accuracy metrics (*accuracy score* and *precision@25%*). The accuracy scores of all seven metrics can be found in Appendix D in Tables 4 and 5. In investigating this accuracy data, we witnessed two emerging narratives. First, black-box and interpretable models performed similarly on the classification task when using a "holistic" metric such as *accuracy score*. In both use-cases, the difference between the most accurate black-box model and the most accurate interpretable model was less than 0.01. Second, in the housing use-case, there was a large difference between the accuracy of black-box and interpretable models when using a "specific" metric such as *precision@25%*. Notably, the best performing black-box model outperformed the top interpretable model by 7%. As a result, it appears that black-box models performed better on resource-constrained metrics in our research.

5.2.1 *Subgroups.*

We believe that it is the responsibility of ML practitioners to conduct a fairness audit whenever they are training models on real data. Therefore, we performed a subgroup analysis in the education domain (*sex* and *parent's education* sensitive features) and found that accuracy does vary with respect to subgroups. While we did not explore this disparity further in this paper, this information motivates the need for further research exploring the trade-offs between fairness, explainability, and accuracy.

5.3 Accuracy-Explainability Trade-off

Our research illustrates that it is difficult to make general statements about the accuracy-explainability trade-off, because insights are dependent on the metrics used to measure accuracy and explainability. Figure 4 shows two such cases. In Figure 4 (a) where precision@25% is graphed against Task 1: ASO, the black-box model performed better with respect to both measures. This result was driven by the fact that in both use-cases HITL users who were presented with a black-box model were generally able to correctly complete the ASO task. In addition to the black-box being more explainable, they were also more accurate. In Figure 4 (b), where the accuracy metric is *accuracy score* and the explainability metric is *Task 2: IMIF*, there was no trade-off between black-box and interpretable models. For both education and housing, model accuracy and ability of HITL users to identify the top feature was the same, regardless of whether linear regression or a black-box with SHAP was used. This result was driven by the fact that both black-box and interpretable models had similar accuracy on the holistic metric *accuracy score*. Furthermore, as discussed in Section 5.1.2, participants who viewed the systems with either linear regression or black-box with SHAP models performed well on the IMIF task because the question was related to the global explainability of the system.

6 DISCUSSION

The initial goal of this research was to quantify the accuracy-explainability trade-off in ML public policy contexts. Our hope was that it would lead to concrete guidelines for ML practitioners, such as "an X% increase in explainability results a Y% decrease in accuracy." Instead, we observed that the accuracy-explainability trade-off is more nuanced than initially anticipated—and in some cases, there may be no trade-off at all.

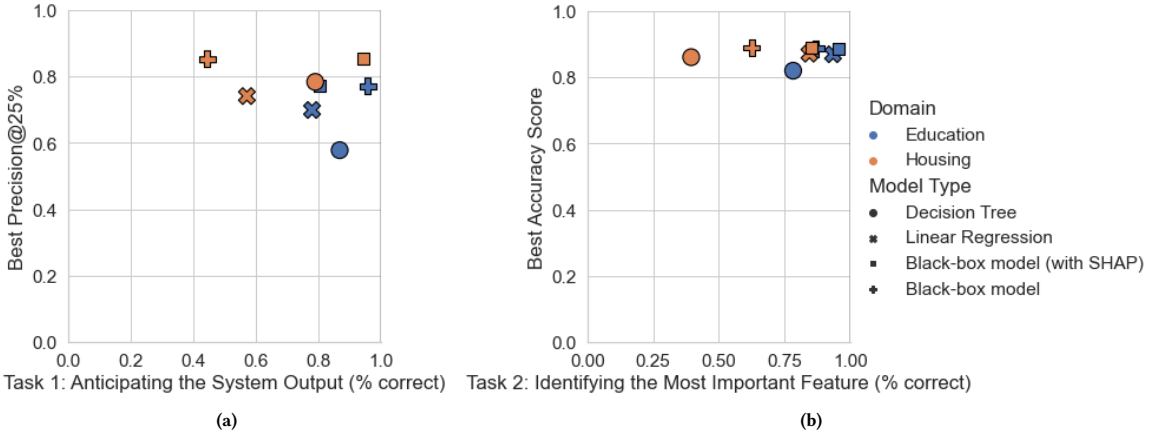
Key finding: no difference between black-box and interpretable models with respect to explainability. In §5, we show that there was no statistical difference in explainability between black-box and interpretable models. This result was statistically robust with regards to four different explainability metrics, including the two objective tasks (ASO and IMIF) and the two subjective measures on the users' perceived understanding and confusion. We propose two reasons for this finding, presented in the next subsections.

User confusion. Presenting more information about a model may actually confuse users due to information overload. In our user study, the inherent opacity of black-box models meant that less information about the model was displayed. Surprisingly, this was an advantage as users were not confused by multiple details about the model.

Our findings suggest that user confusion may play a big role in explainability. We found that HITL users self-reported as being

Domain	Model	Type	Accuracy	Percision@25%
Education	Extra Trees	Black-box	0.85	0.65
	Random Forest	Black-box	0.87	0.70
	XGBoost	Black-box	**0.89**	**0.77**
	Decision Tree	Interpretable	0.82	0.58
	Linear Regression (Lasso)	Interpretable	0.87	0.70
	Linear Regression (Ridge)	Interpretable	0.87	0.70
	Logistic Regression	Interpretable	**0.88**	**0.77**
	Dummy Classifier	Baseline	0.77	0.20
Housing	Extra Trees	Black-box	**0.89**	0.77
	Random Forest	Black-box	**0.89**	0.78
	XGBoost	Black-box	**0.89**	**0.85**
	Decision Tree	Interpretable	0.86	**0.78**
	Linear Regression (Lasso)	Interpretable	0.87	0.74
	Linear Regression (Ridge)	Interpretable	0.87	0.74
	Logistic Regression	Interpretable	**0.88**	0.75
	Dummy Classifier	Baseline	0.75	0.20

Table 2: Accuracy and Precision@25% for trained models. Bold numbers indicate that model had the best score, for the given model type and the given domain. Results for the Dummy Classifier model are listed for reference.

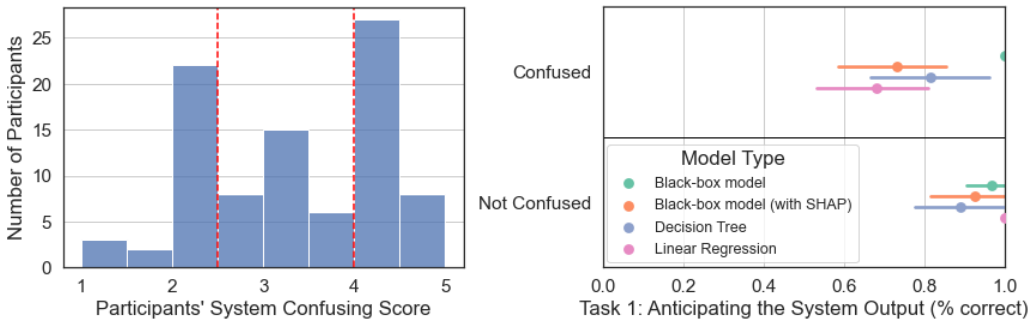


(a) (b)

Figure 4: (a) Percision@25% and participants' mean scores for Task 1: Anticipating the System Output. Using these measures, black-box models are both the most explainable and the most accurate. (b) Accuracy score and participants' mean scores for Task 2: Identifying the Most Important Feature. This Figure illustrates that there is no clear trade-off in accuracy and explainability between black-box and interpretable models, for either housing or education.

confused (a score ≥ 4.0 on the System Confusing Likert-scale construct) by the system 33.4% of the time (35.0% for education, 31.9% for housing). This finding is in concert with similar findings from XAI research [42, 49], some of which have found that more explanations directly *hurt* participants' task performance [39]. Figure 5 (a) shows the distribution of participants' ratings for the System Confusing construct (5 = more confused) in the education domain when presented with the black-box model with SHAP. The distribution is bimodal, suggesting that there are two separate subgroups: those who reported being confused by the system and those who did not.

To test the idea of information overload empirically, we compared the outcomes of confused and not-confused participants on the objectively measurable explainability tasks. We used the thresholds ≥ 4.0 for confused participants and ≤ 2.5 for not-confused participants. Figure 5 (b) shows the performance of the ASO task when separating out the two different subgroups. Importantly, participants from the confused subgroup performed substantially worse at the ASO task. We conjecture that these individuals found that having more information about the system made it *more difficult* to anticipate the output of the system. Note that in Figure 5 there is one exception: Even those users who were confused performed well with the black-box model without SHAP explanations. This further supports the idea that less information may be beneficial. In contrast, the not-confused users had best performance on the



(a) (b)

Figure 5: (a) The bimodal distribution of participants' scores on the System Confusing construct with thresholds for the confused subgroup (System Confusing $\geq$ 4.0) and the not-confused subgroup (System Confusing $\leq$ 2.5) in the education domain and for participants viewing the black-box with SHAP model. (Number of observations = 91.) (b) Confused education participants' mean performance on the ASO task with 95% CI, number of observations: BB = 14, BB (w/ SHAP) = 35, DT = 22, LR = 41; versus not-confused education participants' mean performance on the ASO task with 95% CI, number of observations: BB = 32, BB (w/ SHAP) = 27, DT = 27, LR = 14.

ASO task, which supports the claim that for them having more information was helpful.

Weaknesses in intrinsic explainability. The second element driving our key finding is weaknesses in the intrinsic explainability mechanisms of interpretable models. We uncovered two such weaknesses. First, decision tree diagrams confuse helpful users when it comes to identifying the most important feature. Figure 2 (b) shows that in both education and housing, the model with the lowest absolute score for the IMIF task was the decision tree. The reason behind this finding is deceptively simple: *participants defaulted to selecting the feature found at either the root of the decision tree diagram or one its terminal nodes as the most important, despite receiving clear explanations about the model's feature importance that indicated otherwise.* Evidence for this result can be found in Tables 6 and 7 in the Appendix that show the number of participants who incorrectly selected the root node or a random terminal node as the most important feature. (Tree diagrams are shown in Figures 7 and 9 in the Appendix.)

Second, the linear formula presented with linear regression seemed to confuse the users in our study. This is evidenced in Figure 2 (a), which shows the relatively poor performance of the linear regression model on the ASO task. We believe that this is due to a discrepancy between the explainability scope of the ASO task and the intrinsic explainability mechanism of the linear regression model. The ASO task required that participants think *locally*, yet the linear formula of a linear regression model is more directly indicative of how the model works *globally*. This means that ML practitioners designing explainable systems should be thoughtful about the alignment of different explanations

Our observations on how users interact with tree diagrams and linear formulas lead us to two possible conclusions: First, they may be an indictment against intrinsic explainability mechanisms, highlighting the need to improve existing explainability approaches, even for interpretable models. Second, they indicate the need to invest significant resources in educating and training HITL users on fundamental ML concepts like "feature importance."

What it does versus how it works. More generally, when it comes to the explainability of black-box systems, it appears that end users do not need to understand *how the system works* to be able to understand *what the system does*. As an analogy, consider that many people understand what a TV does and how to operate it without ever understanding how the images are created and displayed on the screen. This observation has meaningful implications for the accuracy-explainability trade-off. As seen in Figures 4 (a) and (b), black-box models may be both the most accurate and the most explainable. This is particularly true when the accuracy metric is resource constrained (e.g., precision@25%). We currently do not have a hypothesis for why this is occurring and believe further investigation of the generalizability of this claim is warranted.

Implications for researchers and practitioners. Our findings lead to several recommendations for practitioners. First, we discourage practitioners from trying to generalize about the accuracy-explainability trade-off — it's just not that simple! Rather, practitioners should acknowledge that the extent to which a trade-off exists (when it exists) is dependent on the chosen accuracy metric, on how explainability is measured, and on the context of use. Second, our results call into question the commonly held beliefs about the inherent pros and cons of black-box and interpretable models. We observed that black-box models can actually be *more* explainable and less confusing to users, perhaps by avoiding overloading users with information. This implies that there may be appropriate contexts to implement black-box models, even when designing for explainabilty. Third, we present several design considerations when implementing explainability for HITL users: practitioners should be aware of weaknesses in the intrinsic explainability mechanisms of linear regression and decision tree models, and should consider the scope of the explanations presented to users. Fourth, we believe that meaningful inter-disciplinary collaboration between ML engineers, domain experts, policy professionals, product designers, and

end users is paramount to properly implementing explainability in ML systems. This recommendation is consistent with perspectives held by other researchers in this space [46].

Importantly, we want to make it clear that this research is not advocating for the *carte blanche* use of black-box models. This work does not abdicate practitioners' responsibilities for ensuring transparency and fairness. We would like for this research to be seen as one data point that adds yet another paradoxical result that to the literature: there are contexts where black-box models may be as explainable to HITL users as interpretable models.

7 CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

In this work, we conducted an empirical investigation of the accuracy-explainability trade-off in two public policy domains and found that nterpretable models were not more explainable than black-box models. There are several limitations of this research that could be expounded upon in future work. First, some researchers have challenged core components of this work, like the idea that proxy tasks may not be accurate measures of explainability in actual human decision-making in real-world contexts [11]. The methods developed in this paper have only been tested in a lab environment where there are no real stakes for users, which may impact their robustness or validity (e.g., the teachers from our survey did not have a connection with a real student for which the AI system was making a prediction). Second, this work only includes two use-cases — housing and education. To better generalize these findings, this work should be replicated across a broader range of policy domains. Third, this work could benefit from qualitative research to better understand the confusion experienced by HITL users. The absence of qualitative work leaves open questions as to how explainability is interacting with the users own "mental models."

ACKNOWLEDGMENTS

This research was supported in part by National Science Foundation awards 1916505, 1934464, 1928614, and 2129076.

REFERENCES

[1] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y Lim. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[2] Everaldo Aguiar, Himabindu Lakkaraju, Nasir Bhanpuri, David Miller, Ben Yuhas, and Kecia L Addison. 2015. Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. 93–102.

[3] D Aha and Dennis Kibler. 1988. Instance-based prediction of heart-disease presence with the Cleveland database. *University of California* 3, 1 (1988), 3–2.

[4] Hiva Allahyari and Niklas Lavesson. 2011. User-oriented Assessment of Classification Model Understandability. In *Eleventh Scandinavian Conference on Artificial Intelligence, SCAI 2011, Trondheim, Norway, May 24th - 26th, 2011 (Frontiers in Artificial Intelligence and Applications, Vol. 227)*, Anders Kofod-Petersen, Fredrik Heintz, and Helge Langseth (Eds.). IOS Press, 11–19. https://doi.org/10.3233/978-1-60750-754-3-11

[5] Kasun Amarasinghe, Kit Rodolfa, Hemank Lamba, and Rayid Ghani. 2020. Explainable machine learning for public policy: Use cases, gaps, and research directions. *arXiv preprint arXiv:2010.14374* (2020).

[6] Ryan S Baker and Aaron Hawn. 2021. Algorithmic Bias in Education. https://doi.org/10.35542/osf.io/pbmvz

[7] Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. 2021. Consumer-lending discrimination in the FinTech Era. *Journal of Financial Economics* (2021). https://doi.org/10.1016/j.jfineco.2021.05.047

[8] Nadia El Bekri, Jasmin Kling, and Marco F. Huber. 2019. A Study on Trust in Black Box Models and Post-hoc Explanations. In *14th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2019) - Seville, Spain, May 13-15, 2019, Proceedings (Advances in Intelligent Systems and Computing, Vol. 950)*, Francisco Martínez-Álvarez, Alicia Troncoso Lora, José António Sáez Muñoz, Héctor Quintián, and Emilio Corchado (Eds.). Springer, 35–46. https://doi.org/10.1007/978-3-030-20055-8_4

[9] Andrew Bell, Alexander Rich, Melisande Teng, Tin Orešković, Nuno B Bras, Lénia Mestrinho, Srdan Golubovic, Ivan Pristas, and Leid Zejnilovic. 2019. Proactive advising: a machine learning driven approach to vaccine hesitancy. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 1–6.

[10] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2020. Explainable Machine Learning in Deployment. arXiv:1909.06342 [cs.LG]

[11] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th international conference on intelligent user interfaces*. 454–464.

[12] Samuel Carton, Jennifer Helsby, Kenneth Joseph, Ayesha Mahmud, Youngsoo Park, Joe Walsh, Crystal Cody, CPT Estella Patterson, Lauren Haynes, and Rayid Ghani. 2016. Identifying police officers at risk of adverse events. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 67–76.

[13] Alexandra Chouldechova. 2016. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. arXiv:1610.07524 [stat.AP]

[14] FDIC: Federal Deposit Insurance Corporation. 1968. Civil Rights Act of 1968. (1968). https://www.fdic.gov/regulations/laws/rules/6000-1400.html

[15] P. Cortez and A. M. G. Silva. 2008. Using data mining to predict secondary school student performance.

[16] Ian Covert, Scott Lundberg, and Su-In Lee. 2020. Understanding global feature contributions with additive importance measures. *arXiv preprint arXiv:2004.00668* (2020).

[17] Ian Covert, Scott M. Lundberg, and Su-In Lee. 2020. Understanding Global Feature Contributions Through Additive Importance Measures. *CoRR* abs/2004.00668 (2020). arXiv:2004.00668 https://arxiv.org/abs/2004.00668

[18] Arun Das and Paul Rad. 2020. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. arXiv:2006.11371 [cs.CV]

[19] Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*. IEEE, 598–617.

[20] Inigo Martinez de Troya, Ruqian Chen, Laura O Moraes, Pranjal Bajaj, Jordan Kupersmith, Rayid Ghani, Nuno B Brás, and Leid Zejnilovic. 2018. Predicting, explaining, and understanding risk of long-term unemployment. In *NeurIPS Workshop on AI for Social Good*.

[21] Matthew DeCamp and Charlotta Lindvall. 2020. Latent bias and the implementation of artificial intelligence in medicine. *Journal of the American Medical Informatics Association* 27, 12 (2020), 2020–2023.

[22] William K Diprose, Nicholas Buist, Ning Hua, Quentin Thurier, George Shand, and Reece Robinson. 2020. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *Journal of the American Medical Informatics Association* 27, 4 (02 2020), 592–600. https://doi.org/10.1093/jamia/ocz229 arXiv:https://academic.oup.com/jamia/article-pdf/27/4/592/34153285/ocz229.pdf

[23] Graham Dove, Martina Balestra, Devin Mann, and Oded Nov. 2020. Good for the Many or Best for the Few? A Dilemma in the Design of Algorithmic Advice. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–22.

[24] Gintare Karolina Dziugaite, Shai Ben-David, and Daniel M Roy. 2020. Enforcing Interpretability and its Statistical Impacts: Trade-offs between Accuracy and Interpretability. *arXiv preprint arXiv:2010.13764* (2020).

[25] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, Inc., USA.

[26] Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2020. Predictably unequal? the effects of machine learning on credit markets. *The Effects of Machine Learning on Credit Markets (October 1, 2020)* (2020).

[27] Philip Gillingham. 2019. Can predictive algorithms assist decision-making in social work with children and families? *Child abuse review* 28, 2 (2019), 114–126.

[28] Michael Gleicher. 2016. A framework for considering comprehensibility in modeling. *Big data* 4, 2 (2016), 75–88.

[29] Bryce Goodman and Seth Flaxman. 2017. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Magazine* 38, 3 (Oct 2017), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

[30] Ben Green. 2022. The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review* 45 (2022), 105681.

[31] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* 51, 5 (2019), 93:1–93:42. https://doi.org/10.1145/3236009

[32] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. *Science Robotics* 4, 37 (2019).

[33] Satoshi Hara and Kohei Hayashi. 2016. Making Tree Ensembles Interpretable. arXiv:1606.05390 [stat.ML]

[34] Kenneth Holstein and Shayan Doroudi. 2021. Equity and Artificial Intelligence in Education: Will "AIEd" Amplify or Alleviate Inequities in Education? arXiv:2104.12920 [cs.HC]

[35] Andreas Holzinger, André Carrington, and Heimo Müller. 2020. Measuring the quality of explanations: the system causability scale (SCS). *KI-Künstliche Intelligenz* (2020), 1–6.

[36] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 1–26. https://doi.org/10.1145/3392878

[37] Qian|Rangwala Hu. 2020. Towards Fair Educational Data Mining: A Case Study on Detecting At-Risk Students. https://eric.ed.gov/?id=ED608050

[38] Johan Huysmans, Bart Baesens, and Jan Vanthienen. 2006. Using rule extraction to improve the comprehensibility of predictive models. (2006).

[39] Sérgio Jesus, Catarina Belém, Vladimir Balayan, João Bento, Pedro Saleiro, Pedro Bizarro, and João Gama. 2021. How can I choose an explainer? An Application-grounded Evaluation of Post-hoc Explanations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 805–815.

[40] Jeff Larson Julia Angwin. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[41] Nicol Turner Lee. 2018. Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society* (2018).

[42] Q. Vera Liao, Daniel M. Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik (Eds.). ACM, 1–15. https://doi.org/10.1145/3313831.3376590

[43] Brian Y. Lim and Anind K. Dey. 2009. Assessing Demand for Intelligibility in Context-Aware Applications. In *Proceedings of the 11th International Conference on Ubiquitous Computing* (Orlando, Florida, USA) *(UbiComp '09)*. Association for Computing Machinery, New York, NY, USA, 195–204. https://doi.org/10.1145/1620545.1620576

[44] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. *Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems*. Association for Computing Machinery, New York, NY, USA, 2119–2128. https://doi.org/10.1145/1518701.1519023

[45] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 4765–4774. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[46] Ricards Marcinkevics and Julia E. Vogt. 2020. Interpretability and Explainability: A Machine Learning Zoo Mini-tour. *CoRR* abs/2012.01805 (2020). arXiv:2012.01805 https://arxiv.org/abs/2012.01805

[47] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[48] Christoph Molnar. 2020. *Interpretable machine learning*. Lulu. com.

[49] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682* (2018).

[50] Oded Nov, Yindalon Aphinyanaphongs, Yvonne W Lui, Devin Mann, Maurizio Porfiri, Mark Riedl, John-Ross Rizzo, and Batia Wiesenfeld. 2021. The transformation of patient-clinician relationships with ai-based medical advice. *Commun. ACM* 64, 3 (2021), 46–48.

[51] Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, USA.

[52] Alun Preece. 2018. Asking 'Why' in AI: Explainability of intelligent systems – perspectives and challenges. *Intelligent Systems in Accounting, Finance and Management* 25, 2 (2018), 63–72. https://doi.org/10.1002/isaf.1422 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/isaf.1422

[53] Jennifer Raso. 2017. Displacement as regulation: New regulatory technologies and front-line decision-making in Ontario works. *Canadian Journal of Law & Society/La Revue Canadienne Droit et Société* 32, 1 (2017), 75–95.

[54] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[55] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:1602.04938 [cs.LG]

[56] Kit T Rodolfa, Hemank Lamba, and Rayid Ghani. 2020. Empirical observation of negligible fairness-accuracy trade-offs in machine learning for public policy. *arXiv preprint arXiv:2012.02972* (2020).

[57] Robert Ross. 2017. The impact of property tax appeals on vertical equity in Cook County, IL. *Univerity of Chicago, Harris School of Public Policy Working Paper* (2017).

[58] Cynthia Rudin. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. arXiv:1811.10154 [stat.ML]

[59] Piotr Sapiezynski, Valentin Kassarnig, and Christo Wilson. 2017. Academic performance prediction in a gender-imbalanced environment.

[60] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. arXiv:1911.02508 [cs.LG]

[61] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7-8, 2020*, Annette N. Markham, Julia Powles, Toby Walsh, and Anne L. Washington (Eds.). ACM, 180–186. https://doi.org/10.1145/3375627.3375830

[62] Gregor Stiglic, Petra Povalej Brzan, Nino Fijacko, Fei Wang, Boris Delibasic, Alexandros Kalousis, and Zoran Obradovic. 2015. Comprehensible predictive modeling using regularized logistic regression and comorbidity based features. *PloS one* 10, 12 (2015), e0144439.

[63] Julia Stoyanovich, Jay J Van Bavel, and Tessa V West. 2020. The imperative of interpretable machines. *Nature Machine Intelligence* 2, 4 (2020), 197–199.

[64] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. 2013. OpenML: Networked Science in Machine Learning. *SIGKDD Explorations* 15, 2 (2013), 49–60. https://doi.org/10.1145/2641190.2641198

[65] Ben Wagner. 2019. Liable, but not in control? Ensuring meaningful human agency in automated decision-making systems. *Policy & Internet* 11, 1 (2019), 104–122.

[66] Harrison Wilde, Lucia L. Chen, Austin Nguyen, Zoe Kimpel, Joshua Sidgwick, Adolfo De Unanue, Davide Veronese, Bilal Mateen, Rayid Ghani, Sebastian Vollmer, and et al. 2021. A recommendation and risk classification system for connecting rough sleepers to essential outreach services. *Data and Policy* 3 (2021), e2. https://doi.org/10.1017/dap.2020.23

[67] Yiwei Yang, Eser Kandogan, Yunyao Li, Prithviraj Sen, and Walter S Lasecki. 2019. A study on interaction in human-in-the-loop machine learning for text analytics. In *IUI Workshops*.

[68] Jun Yuan, Oded Nov, and Enrico Bertini. 2021. An exploration and validation of visual factors in understanding classification rule sets. In *2021 IEEE Visualization Conference (VIS)*. IEEE, 6–10.

[69] Leid Zejnilovic, Susana Lavado, Carlos Soares, Íñigo Martínez De Rituerto De Troya, Andrew Bell, and Rayid Ghani. 2021. Machine Learning Informed Decision-Making with Interpreted Model's Outputs: A Field Intervention. In *Academy of Management Proceedings*, Vol. 2021. Academy of Management Briarcliff Manor, NY 10510, 15424.

[70] Leid Zejnilović, Susana Lavado, Íñigo Martínez de Rituerto de Troya, Samantha Sim, and Andrew Bell. 2020. Algorithmic Long-Term Unemployment Risk Assessment in Use: Counselors' Perceptions and Use Practices. *Global Perspectives* 1, 1 (06 2020). https://doi.org/10.1525/gp.2020.12908 arXiv:https://online.ucpress.edu/gp/article-pdf/1/1/12908/462946/12908.pdf 12908.

[71] Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell. 2019. " Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations. *arXiv preprint arXiv:1904.12991* (2019).

A SURVEY FLOW AND SAMPLE

This appendix contains important diagrams from our survey. First, Figure 6 illustrates the survey flow experienced by participants. Second, Figures 7 and 8 show screenshots from the survey and provide a sample of participants' experience.

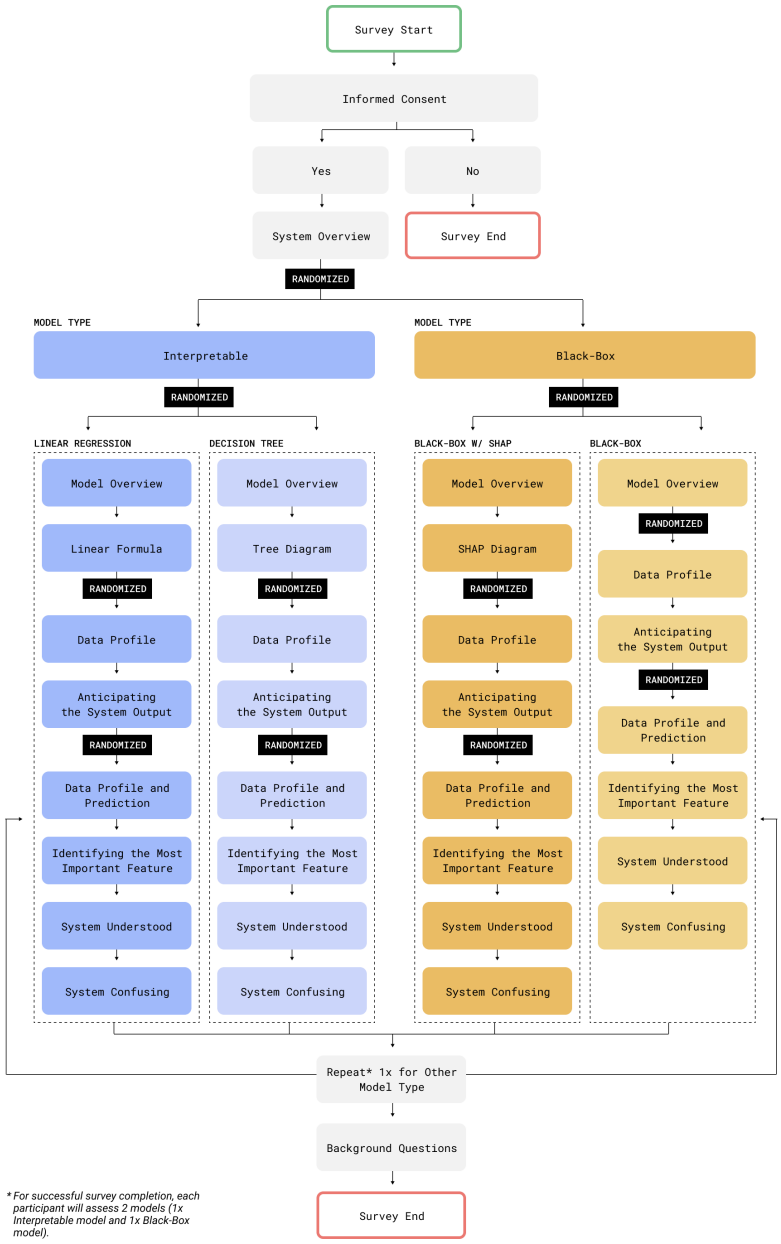
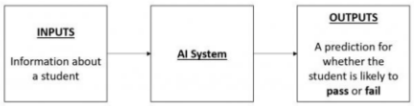


Figure 6: Diagram illustrating the survey flow experienced by participants.

System Alpha

The following AI system is called **System Alpha. System Alpha** is used to predict whether or not a student will **pass** or **fail** their final year of high school.

The AI system takes in information about the student and uses this information to create its prediction. Below is a diagram of this process:



Below is a table that shows details about how **System Alpha** uses different attributes about the student to make its predictions. It shows how different attributes are important overall to the system.

The column **attribute** contains the attribute name.

The column **impact** shows how different values of that attribute influence the prediction of the system.

The column **relative weight** shows how much the attribute is considered in the prediction. A higher relative weight indicates that the attribute is more important overall to **System Alpha.**

Important note: in Portugal, grades are given on a scale of 1 to 20, where any grade of 9 or below is failing, and any grade 10 or above is passing. A grade of 10-13 is satisfactory, 14-15 is good, 16-17 is very good, and 18-20 is excellent.

Attributes	Impact	Relative Weight
First year **grade** (Scale from 1 to 20, 10 or greater = passing)	Lower value = more likely to fail	0.149
Mother's education level (0 = none, 1 = primary education up to 4th grade, 2 = 5th to 9th grade, 3 = secondary education, 4 = higher education)	Lower value = more likely to fail	0.062
Quality of **family relationships** (Scale from 1 = very bad to 5 = excellent)	Lower value = more likely to fail	0.006

System Alpha uses a **decision tree** to make predictions on whether or not a student will **pass** or **fail**. This tree is shown in full below.

The diagram below is read by starting at the top truth box (the one containing the text "Is the student's mother's education level less than 1.5?") and following the "Yes" or "No" branch according to the situation for a particular student. For example, if a student's mother has an education level of 2, the result would be "No" since 2 is greater than 1.5. The next truth box is read in the exact same way and this continues until a **pass** or **fail** box is reached.

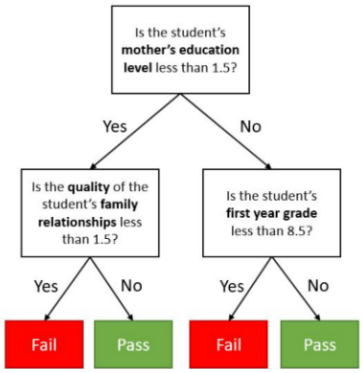


Figure 7: An example of the decision tree model overview and tree diagram displayed to participants in the education survey.

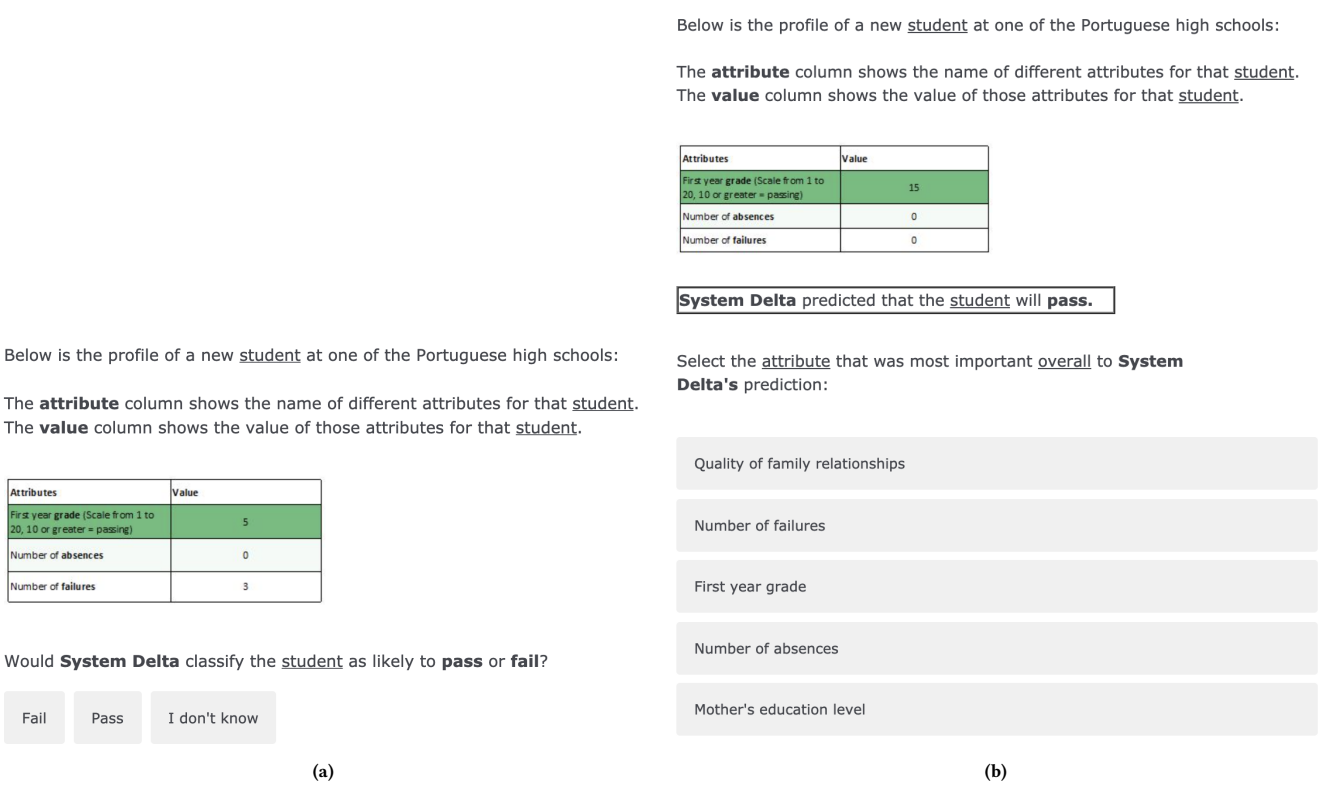


Figure 8: (a) The Anticipating the System Output task for the black-box without SHAP system in the education survey; (b) The Identifying the Most Important Feature task for the black-box without SHAP system in the education survey. In both (a) and (b) participants were informed that the color saturation of the row corresponded to the importance of the feature (i.e. the darker the color, the more important overall to the model).

B STATISTICAL ANALYSIS

This appendix section contains our statistical analyses and resultant t-statistics and p-values.

	Education		**Housing**	
	Test statistic	**P-value**	**Test statistic**	**P-value**
Interpretable vs black-box				
Task 1 Performance[2]	13.000	0.296	13.000	0.296
Task 2 Performance[2]	9.000	0.087	14.000	0.108
Understood System[3]	-1.674	0.094	-2.450	0.014 **
System Confusing[3]	2.864	0.004 **	3.619	< 0.000 ***
4-way model comparison				
Task 1 Performance[4]				
Decision Tree	-	-	-	-
Linear Regression	-1.647	0.099	-3.091	0.002 **
Black-box (without SHAP)	1.633	0.102	-3.846	0.000 ***
Black-box (with SHAP)	-1.326	0.185	1.684	0.092
Task 2 Performance[4]				
Decision Tree	-	-	-	-
Linear Regression	3.183	0.001 **	5.015	< 0.000 ***
Black-box (without SHAP)	1.911	0.056	3.008	0.003 **
Black-box (with SHAP)	3.763	< 0.000 ***	5.085	< 0.000 ***
Understood System[4]				
Decision Tree	-	-	-	-
Linear Regression	-3.140	0.002 **	-1.826	0.068
Black-box (without SHAP)	1.799	0.072	-1.672	0.094
Black-box (with SHAP)	-2.314	0.021 *	2.441	0.015 *
System Confusing[4]				
Decision Tree	-	-	-	-
Linear Regression	2.629	0.009 **	0.876	0.381
Black-box (without SHAP)	-3.098	0.002 **	0.144	0.885
Black-box (with SHAP)	1.409	0.159	-3.664	< 0.000 ***

Table 3: Test statistics and *p-values* for various statistical analyses of the survey results.

C FULL MODEL PERFORMANCE RESULTS

This appendix section contains the full training results for both the education and housing results. The tables below show performance for 7 model types (extra trees, reandom forest, XGBoost, decision tree, linear regression (Lasso), linear regression (Ridge), and Logistic Regression) for 7 different performance metrics (accuracy, F1 score, AUC, precision@10%, precision@25%, recall@10%, recall@25%).

[2]McNemar's test
[3]Random effects model: outcome ~ model type + intercept, group = participant, where model type is black box or interpretable
[4]Random effects model: outcome ~ model type + intercept, group = participant, where model type is Decision Tree, Linear Regression, Black-box (without SHAP), Black-box (with SHAP)

Model	Type	Accuracy	F1 Score	AUC	Precision @ 10%	Precision @ 25%	Recall @ 10%	Recall @ 25%
Extra Trees	Black-box	0.85	0.56	0.70	0.81	0.65	0.37	0.73
Random Forest	Black-box	0.87	0.67	0.77	0.87	0.70	0.40	0.79
XGBoost	Black-box	0.89	0.73	0.83	0.91	0.77	0.40	0.80
Decision Tree	Interpretable	0.82	0.51	0.69	0.65	0.58	0.23	0.57
Linear Regression (Lasso)	Interpretable	0.87	0.65	0.76	0.88	0.70	0.40	0.78
Linear Regression (Ridge)	Interpretable	0.87	0.65	0.76	0.90	0.70	0.41	0.79
Logistic Regression	Interpretable	0.88	0.72	0.85	0.92	0.77	0.41	0.78
(1) Best black-box model score		0.89	0.73	0.83	0.91	0.77	0.40	0.80
(2) Mean black-box model score		0.87	0.65	0.77	0.86	0.71	0.39	0.77
Standard deviation for black-box scores		0.02	0.09	0.06	0.05	0.06	0.02	0.03
(3) Best interpretable model score		0.88	0.72	0.85	0.92	0.77	0.41	0.79
(4) Mean interpretable model		0.86	0.63	0.76	0.84	0.69	0.36	0.73
Standard deviation for interpretable scores		0.03	0.09	0.07	0.13	0.08	0.09	0.11
Absolute difference between (1) and (3)		0.01	0.02	0.02	0.02	0.00	0.01	0.01
Absolute difference between (2) and (4)		0.01	0.02	0.01	0.03	0.02	0.03	0.04

Table 4: Model performance on 7 different metrics for the 3 black-box and 4 interpretable models trained on the education dataset. The table also lists the absolute difference between the best performing black box and interpretable models. It was observed that regardless of the performance metric chosen, there is very little difference between the performance of the best performing black-box and interpretable model.

Model	Type	Accuracy	F1 Score	AUC	Precision @ 10%	Precision @ 25%	Recall @ 10%	Recall @ 25%
Extra Trees	Black-box	0.89	0.75	0.82	0.95	0.77	0.38	0.76
Random Forest	Black-box	0.89	0.76	0.83	0.95	0.78	0.38	0.77
XGBoost	Black-box	0.89	0.77	0.84	0.95	0.85	0.38	0.77
Decision Tree	Interpretable	0.86	0.70	0.79	0.88	0.78	0.34	0.70
Linear Regression (Lasso)	Interpretable	0.87	0.70	0.78	0.94	0.74	0.37	0.74
Linear Regression (Ridge)	Interpretable	0.87	0.70	0.78	0.94	0.74	0.37	0.74
Logistic Regression	Interpretable	0.88	0.74	0.85	0.94	0.75	0.37	0.74
(1) Best black-box model score		0.89	0.77	0.84	0.95	0.85	0.38	0.77
(2) Mean black-box model score		0.89	0.76	0.83	0.95	0.80	0.38	0.77
Standard deviation for black-box scores		0.00	0.01	0.01	0.00	0.05	0.00	0.01
(3) Best interpretable model score		0.88	0.74	0.85	0.94	0.78	0.37	0.74
(4) Mean interpretable model		0.87	0.71	0.80	0.92	0.75	0.37	0.73
Standard deviation for interpretable scores		0.01	0.02	0.03	0.03	0.02	0.02	0.02
Absolute difference between (1) and (3)		0.01	0.03	0.01	0.01	0.07	0.00	0.03
Absolute difference between (2) and (4)		0.02	0.05	0.03	0.03	0.05	0.01	0.04

Table 5: Model performance on 7 different metrics for the 3 black-box and 4 interpretable models trained on the housing dataset. The table also lists the absolute difference between the best performing black box and interpretable models. It was observed that for certain metrics, such as precision@25%, the best performing black-box outperformed the best interpretable model by 0.07.

D IDENTIFYING THE MOST IMPORTANT FEATURE RESPONSES

This appendix section contains the number of responses for the IMIF explainability task for both the education and housing domains. It also contains the tree diagram shown to participants in the housing domain. Taken together, the result table and tree diagram show how participants often defaulted to picking either the top node in the tree or a random terminal node when selecting the most important feature.

	First year grade	Mother's education level	Quality of family relationships	Number of absences	Number of failures
Decision Tree	**65**	9	7	1	1
Linear Regression	**72**	0	0	5	0
Black-box	**60**	0	0	6	3
Black-box with SHAP	**87**	1	0	2	1

Table 6: Number of responses for the Identifying the Most Important (IMIF) explainability task for the education domain (row headers are the 4 system types and column headers are the 5 response choices). The correct response is shown in bold.)

	Size of living area	Size of house above ground	House grade	Property view rating	Average size of living area for closest 15 houses
Decision Tree	**30**	24	20	2	0
Linear Regression	5	3	**71**	3	2
Black-box	12	2	**45**	3	10
Black-box with SHAP	9	0	**75**	3	1

Table 7: Number of responses for the Identifying the Most Important (IMIF) explainability task for the housing domain (row headers are the 4 system types and column headers are the 5 response choices).

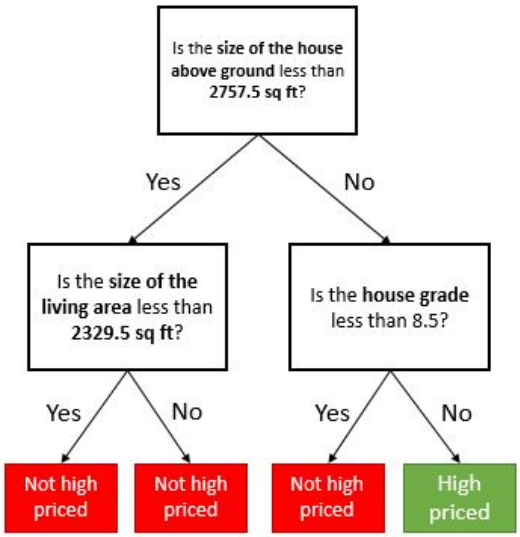


Figure 9: The decision tree model shown to participants in the housing domain.