A User Study on the Effect of Aggregating Explanations for Interpreting Machine Learning Models

[Work-in-Progress]

Josua Krause
New York University
josua.krause@nyu.edu

Adam Perer
IBM Research &
Carnegie Mellon University
adam.perer@us.ibm.com

Enrico Bertini
New York University
enrico.bertini@nyu.edu

ABSTRACT

Recently, there is growing consensus of the critical need to have better techniques to explain machine learning models. However, many of the popular techniques are instance-level explanations, which explain the model from the point of view of a single data point. While local explanations may be misleading, they are also not human-scale, as it is impossible for users to read explanations for how the model behaves on all of their data points. Our *work-in-progress* paper explores the effectiveness of providing instance-level explanations in aggregate, by demonstrating that such aggregated explanations have a significant impact on users' ability to detect biases in data. This is achieved by comparing meaningful subsets, such as differences between ground truth labels, predicted labels, and correct and incorrect predictions, which provide necessary navigation to explain machine learning models.

ACM Reference Format:
Josua Krause, Adam Perer, and Enrico Bertini. 2018. A User Study on the Effect of Aggregating Explanations for Interpreting Machine Learning Models: [Work-in-Progress]. In *Proceedings of KDD 2018 Workshop on Interactive Data Exploration and Analytics (IDEA'18) (IDEA @ KDD'18).* ACM, New York, NY, USA, 9 pages.

1 INTRODUCTION

As data continues to increase in complexity and scale, data scientists are increasingly turning to machine learning to automatically make decisions. However, when these decisions are applied to high-stakes domains such as medicine, law enforcement, and financial lending, it is critical for humans to understand the basis for these decisions.

Predictive modeling is an area of supervised machine learning which aims to predict outcomes from data. Such models are trained on examples with a known ground truth. In order to verify that a model generalizes well to unseen data, a hold-out data set with known ground truth is typically used to test the model after training. This allows to detect problems with the model, such as over-fitting on the training data, *i.e.*, the model learned a phenomenon that is only present in the training data, by measuring the gap in the accuracy between the training and the testing data. However, sometimes a bias in the collected data affects both the training and the test data which makes it impossible to detect through accuracy alone. A human understanding of the underlying data is needed.

For example, Caruana *et al.* [1] built an interpretable machine learning model to analyze mortality risk in patients diagnosed with Pneumonia. After analyzing the model's behavior, Caruana *et al.* detected that patients that additionally suffered from Asthma had a significantly lower mortality risk, according to the *model* and supported by the data. However, this finding goes against current medical knowledge, as the combination of Pneumonia and Asthma are associated with a significantly increased mortality risk. In fact, the data was biased because these high-risk patients with Asthma were given special attention during their hospital visits which contributed to their lower mortality. The presence of Asthma was not responsible for their improvement in health, but rather a systematic bias.

Using the interpretable model and human expert knowledge, it was possible to detect this systematic bias in the data before deploying the model. However, using interpretable machine learning algorithms typically penalizes their capacity, thus lowering the potential accuracy of the model [1] or is only superficially more interpretable by being interpretable on a small scale but not for more complex tasks ([10, 13]). As a way to interpret the behavior of machine learning models *independently* from the used algorithm, black-box and more precisely, instance-level explanations recently became popular [11, 15, 18].

However, such explanations are commonly reviewed by experts one-at-a-time. This task becomes infeasible when dealing with thousands or more instances, also typical of real-world datasets. To that extent, we propose a visual way of reviewing instance-level explanations with the help of aggregation in combination with navigation. This is implemented through the comparison of subsets of the test data under different conditions.

We conducted a study comparing aggregated instance-level explanations to their individual counterparts. Under both conditions, different subsets of the test data could be compared by participants. By providing models with both biased and unbiased data, we were able to measure the trust of participants in the decisions made by the models and their ability to detect flaws in the underlying data for both methods.

Concretely, our contributions include a method for effectively comparing subset of a data set using histograms; using this method, a way to effectively aggregate instance-level explanations; and a study showing that this aggregation overcomes the potential harmfulness of instance-level explanations.

IDEA @ KDD'18, August 20th, 2018, London, United Kingdom

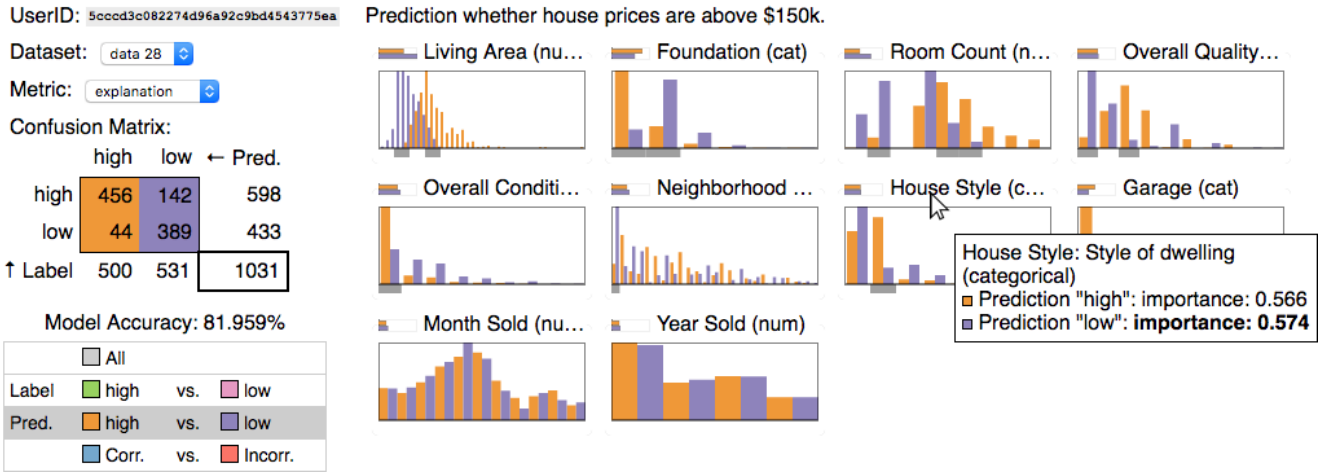


Figure 1: The full interface illustrating the aggregated histogram view. The user is comparing the model's prediction of "high" house prices (orange) to the prediction of "low" prices (purple). The user hovers over the feature "House Style" revealing a more detailed description, whether the feature is categorical or numeric, and the importance / feature weight for each of the subsets.

Following, we will first motivate the circumstances of our study in Section 2 and then further discuss related work in Section 3. We will propose our design for aggregating and comparing subsets of instance-level explanations in Section 4. Afterwards we will describe the experimental setup in Section 5. The results of the study are provided in Section 6 and their implications are discussed in Section 7. We then conclude in Section 8 and discuss future work.

2 MOTIVATION

Experiments for instance-level explanations typically focus on use cases where the explanation is presented to the user one instance at a time. This is helpful when monitoring the continuous performance of a machine learning model in production. However, it limits one's ability to gather a holistic view of a model's behavior (*i.e.*, a global explanation). Looking at many instances is very time consuming and potentially ineffective. It is not clear whether people can build a coherent understanding of a model by looking at a series of instances: comparison between many instances overloads memory and does not leverage the data compression capabilities of aggregate representations.

The main goal of our study is therefore to explore the idea of aggregating data about many instances and their explanations and verify its effects on model comprehension. More precisely, we want to study the effect of aggregation on what we call "semantic validation": the ability of a human to validate the decisions of a model according to his or her knowledge of the domain.

For this purpose, a person knowledgeable with the domain has to verify that the model and the data are consistent with their mental model and, if necessary, override information coming from statistical aggregates on accuracy. This is an important task, especially for models making critical decisions such as those employed in health care [1] and security.

A second goal of this study is to better understand how explanations contribute to semantic validation. Explanations typically provide, for each instance, a weight or score that conveys information about how important each feature is, for a given decision, and for a given instance. An important question therefore is to better understand what particular benefits, if any, explanations bring to human validation; whether this is conducted using an instance-level exploration strategy or a more compact aggregation. Our hypothesis is that explanations may bring value if they manage to direct the user's attention to instances and features where biases and mistakes reside.

In summary, our experiment aims at studying the effect of two main factors: *aggregation level* (instance-level or aggregated-level), and *explanations* (the presence or absence of feature weights).

3 RELATED WORK

We broadly divide related work into two parts. First, we describe studies that focus on the effectiveness of instance-level explanations. Second, we describe methods that use visual analytics to detect biases in machine learning models.

3.1 Effectiveness of Instance-level Explanations

When introducing their algorithm, LIME (Ribeiro *et al.* [18, 19, 22]), the authors conducted experiments to show the effectiveness of their method. However, instance-level explanations were only inspected individually and not in aggregate form.

Kulesza *et al.* [12] introduces explanatory debugging. Users are presented individual decisions, made by the model, in a list. Those can then can be used to "personalize" the model and improve its statistical performance by finding and giving feedback on incorrect decisions. Zhou *et al.* [24] analyzes how uncertainty and cognitive load affects trust in a machine learning model. Here models are compared that predict the risk of pipe failure in a sewer systems according to several features. In addition to the expected failure rate according to model, the length of the observed part of the pipes is shown to the user aggregated over all instances. The study found that showing the uncertainty of the model significantly decreased

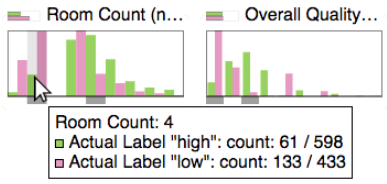


Figure 2: Comparing the distribution of values by "Actual Label" (*i.e.*, ground truth). The height of the bars show the percentage of values within the respective subset (green for "high" outcomes and pink for "low" outcomes). The average feature weight of each subset is shown next to the feature name. This is only visible in the condition including explanations.

the trust of participants. Additionally, adding cognitive load in terms of limited decision time trust in the model decreased significantly as well. Narayanan *et al.* [16] explores how humans understand explanations from a machine learning model. Explanations for individual instances, in the form of simple rules, were presented and participants were asked to determine the predicted outcome of the underlying model. The study found that greater complexity, more rules and more variables, result in a higher response time and decreased accuracy. Note, that the works presented so above always assume that errors stem from the shortcomings of the model and not from incorrect or biased data.

Stumpf *et al.* [21] finds that under some circumstances, explanations can be harmful to users, by invoking false confidence. This is due to the user extrapolating from few instance-level explanations, making their mental model seem correct. Additionally, trust in the machine learning model *overrides* their initial intuition: *e.g.* "I guess this thing knows more than me. The system knows more than me. I'll accept [the diagnosis]". The study investigates inspecting individual instances one after the other, however our experiments confirm both of those findings, even when showing multiple instances in a table.

3.2 Visual Analytics Methods to Detect Biases

Hohman *et al.* [6] identifies detecting biased data as one of their five use cases for visual analytics for machine learning. However, their examples focus on work that only looks at the data without the help of machine learning models [4] or simple models where humans adjust the thresholds of the model manually [23]. Chang *et al.* [2] use crowd-sourcing to label data and ensure its integrity. However, this approach may not work if domain expertise in the field is required to label data correctly.

Simard *et al.* [20] introduces Machine Teaching, a paradigm that uses an already labeled data set for training a machine learning model. It then presents predicted instances to a domain expert who then can either, fix an incorrect label, manipulate features, change constraints, or postpone a decision if the instance is ambiguous. This way an expert can ensure that the final model is correct and remove biases. However, finding biases is not scalable as the experts has to go through many examples and might miss problems, especially if the performance of the model increases but the underlying data is incorrect.

Krause *et al.* [9] demonstrate, how aggregated instance-level explanations can be used to find biases in healthcare data. They used an instance-level algorithm optimized for sparse binary input data (Martens and Provost [15]). Through aggregation, filtering, and reordering, they found biases in their data used for predicting hospital admission that made it impossible for the machine learning model to correctly predict admission in some cases. For example, the model knew about a CET or PET scan happening but was unaware of their results. Thus, the model was unable to predict the diagnosis since the result of the scan directly influences the outcome.

4 DESIGN CONTRIBUTION

In order to effectively analyze machine learning model interpretation, we allow users to compare subsets of the data set to each other. These subsets are defined by different combinations of cells in the confusion matrix of the machine learning model. We selected subsets that help understanding the behavior of the model:

All. The full data set is shown and no comparison occurs. This is the initial view of the data.

Ground truth. By comparing rows of the confusion matrix to each other, users can explore the actual labels of the data.

Predicted labels. By comparing columns of the confusion matrix to each other, users can explore the predicted labels of the model.

Correctness. By comparing the diagonals of the confusion matrix to each other, users can explore when the model's prediction is correct or incorrect.

While it is feasible to allow users more freedom in selecting subsets, (*e.g.* to compare only errors of a certain predicted label) this freedom also increases the complexity of the user interface and a user has to understand when to use each of those subsets in order to be effective.

When aggregating instances, comparing subsets to each other is not trivial. The naïve solution of showing the actual amount of instances with respect to the full data set creates a disadvantage for the smaller subset. However, it is not as important to know the actual distribution, but rather where one subset has a significantly higher or lower concentration of instances compared to the other. To this extent, we propose a novel approach of scaling each subset separately with respect to their own magnitude (see Figure 2). Note that we then compare percentages of instances within the respective subsets. We further indicate strong differences in the subsets by showing a gray bar at the bottom of the histogram. We are not aware of any prior that uses or explored this way of comparing subsets with histograms and claim it as a design contribution. We demonstrate the effectiveness of this design in Section 6.

5 EXPERIMENTAL SETUP

In order to see whether explanations are helpful in detecting biases of the training data, we used a publicly available housing price data [3] and created a modified version with an inherent bias. Originally a regression task, we converted the data set into a classification task predicting whether the house price is above $150k (598 instances above; 433 instances below; 1031 instances in total). We also reduced the number of features in the data set to 10 in order to make it possible to see all features at once in all conditions, without the need to scroll, so data otherwise hidden off-screen would not be a confounding effect. The biased dataset needed to have a bias detectable in both the aggregated and instance-level version, thus

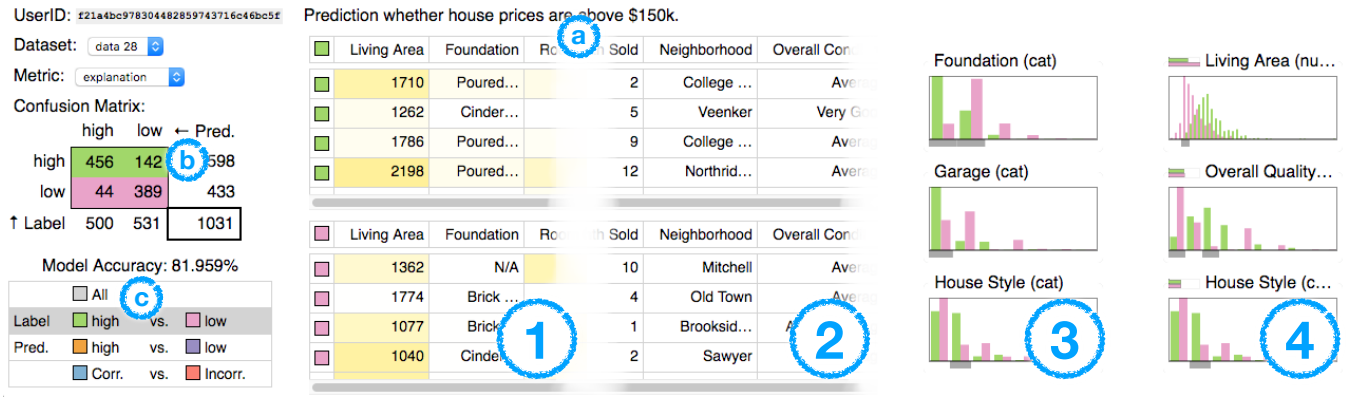


Figure 3: Showing the four conditions of our study: (1) instance-level explanations; (2) only instances; (3) only aggregated features; (4) aggregated features with explanations. On the left and the top are consistent parts of the user interface showing: (a) the problem description; (b) the confusion matrix; (c) the subset selector.

we chose to manipulate the outcome of the biased data set to be dependent on the value of one feature ("Living Area") with some random perturbations. The biased outcome was chosen so that a *larger* "Living Area" results in a *lower* house price. This relationship does not reflect reality (an increased "Living Area" generally results in a higher house price). The bias is present to the same degree in both the training and the testing data.

Furthermore, by controlling the degree of randomness while creating the biased data set, we controlled the accuracy of the prediction when training a machine learning model, such that the model using the biased data has a *higher* accuracy than the model on the real data. We trained Multi-Layer Perceptrons [5] on both data sets resulting in test accuracies of 81.96% for the real data and 88.33% for the biased data.

The target user group for our experiment are people with a basic knowledge of machine learning. The study and interfaces are designed to be effective to use with little training.

5.1 User Interface Conditions

For explaining the model behavior, we compute the explanations using the LIME algorithm [18] on the test data. LIME computes feature weights for each instance in the data. A weight of zero indicates that the feature was not used in the prediction whereas a non-zero weight indicates that the feature was used. A feature weight with larger magnitude indicates that the feature is more important to the prediction than a feature with a smaller magnitude of its weight. However, in order to simplify the user interface and understanding, we computed the absolute value of the feature weights. Thus, participants will only see if a feature has influence on the prediction, not whether this influence is towards a "low" or "high" house price prediction. This additional information is not relevant for the given task and would make the user interface confusing.

For comparing instance-level and aggregated conditions, we developed two user interfaces. Both interfaces share two major components, the confusion matrix of the current model alongside the model's accuracy and a list for selecting different subsets to compare to each other (see Figure 3). Those subsets can be: comparing instances with different ground truth labels, instances with different predicted labels, instances with different correctness, or the full dataset in which case no comparison occurs. How comparing subsets looks like is dependent on the which condition is used. The selections use different colors to distinguish the subsets in order to prevent participants from getting confused about which subset comparison is currently selected (we also indicate the selection in the list). The colors are also used to highlight the confusion matrix cells corresponding to the current selection. Note, that all selections always represent all instances in the data and no two instances from the same confusion matrix cell can appear in opposing subsets.

The design of the user interface lets users iterate through multiple useful slices of the data (such as getting an overview of the data or comparing different, meaningful, subsets to each other). This design, inspired by SYF [17], provides users with a systematic guide to iterate through meaningful views while also supporting flexible diversions to pursue insights.

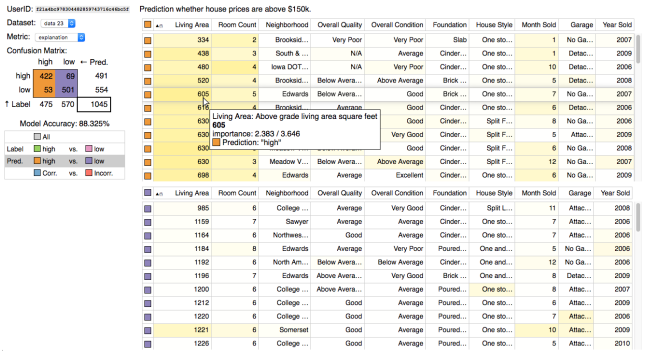


Figure 4: The full interface illustrating the table view showing individual instances. The user is comparing the model's prediction of "high" house prices (orange) to the prediction of "low" prices (purple). The feature "Living Area" is ordered by ascending values and the user hovers over the cell with the value "605".

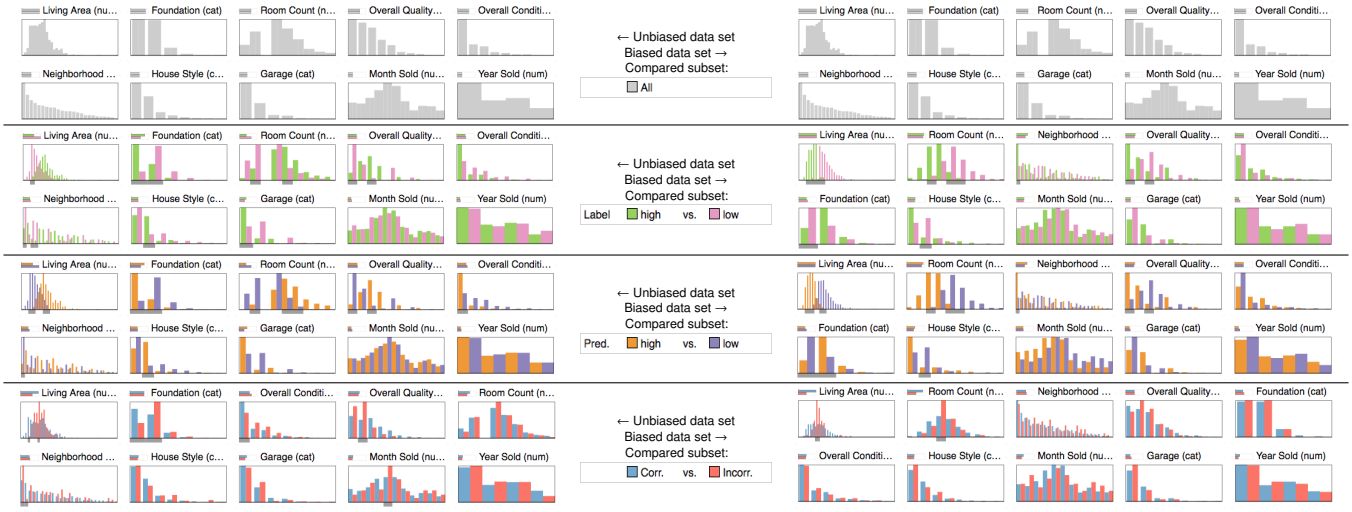


Figure 5: Comparison of different subset selections on both the unbiased model (left side) and the biased model (right side). Features are sorted per segment by most important at the top-left, row-wise to least important at the bottom-right. Note, the feature "Living Area", in both the "Label" (*i.e.*, ground truth) and "Pred." (*i.e.*, model prediction), has flipped outcomes for the biased model (right side). Each subset has a different color palette to not confuse different selections with each other.

5.1.1 Instance-level Condition. The user interface for the instance-level condition is a table showing the values of each feature for each instance in its cells, as seen in Figure 4. This is a change from how instance-level explanations are usually studied in the literature, where each instance is presented in isolation. However, this isolated way of showing instances is limiting as it becomes time consuming to inspect more instances so participants likely only see very few instances in total.

The columns of the table, representing features, are ordered by the average weight of this feature, if the condition includes explanations. If the condition does not include explanations the columns are sorted alphabetically. The cells of the table reflect the feature weight of the corresponding instance using a linear yellow color scale. In addition, hovering over a cell with the mouse shows a tool-tip indicating the actual feature weight number and the full value of the cell and the full feature name, in case those values got abbreviated due to cell size restrictions. Columns can be sorted by clicking on the table header. This cycles through sorting the feature values by ascending and descending value. If explanations are available, the feature can also be sorted by ascending and descending feature weights.

For comparing different subsets of the data, we show two aligned tables. The key for colors that represent different subsets are shown on the far left side of the table.

5.1.2 Aggregated Condition. The user interface for the aggregated condition represents the distribution of feature values as histograms, similar to [8]. Feature names are shown above the histogram and the histograms are arranged left to right row-wise and top to bottom. For the condition with explanations available, a small bar chart next to the feature name indicates the average weight of this feature. In this condition, the histograms are ordered by descending magnitude of average feature weights. The averages are computed for each subset separately. The order is, more specifically, based on the average of the subsets' average weights, which allows features to appear first that are only important under certain conditions. If no explanations are present, the order is alphabetical.

Hovering over a histogram with the mouse reveals tool-tips showing the actual instance count of the hovered histogram segment, as well as, the full feature name and the feature weight number. For categorical features, the bars of the histograms are slimmer so that distinct values are more easily separable. Additionally, the order of the values indicate their quantity in the data set with the most common categorical value first on the left.

When comparing different subsets of the data, as seen in Figure 5, bars of each color are shown next to each other in the histogram. The height of the bars are scaled by their relative proportion *within* each subset. This means the height indicates the percentage of instances in the respective subset. The vertical scale ranges to the highest percentage across both subsets. This allows for seeing where one subset is more concentrated than the other independent of the total size of each subset. In order to indicate big differences in the distribution we show a gray rectangle at the bottom of the histogram if one subset is strongly more concentrated at this value range than the other (see Figure 2).

5.2 Study Design

We study four conditions which result from combining different representations with the inclusion or exclusion of explanations. The different representations are:

- A view of the model through individual instances. Instances are listed in a table. This is an extension of the approach of inspecting instances one-by-one.
- A view of the model through aggregated instances. Instances are aggregated in histograms for individual features.

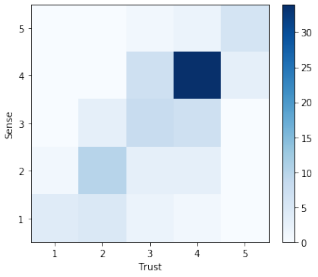
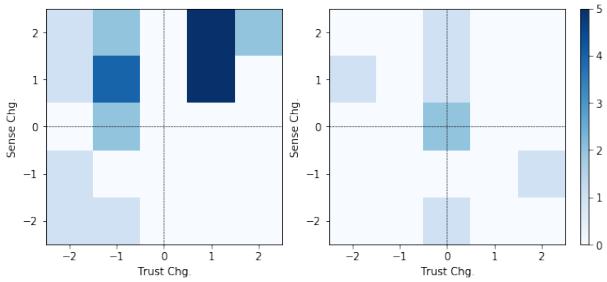


Figure 6: Distribution of the responses of participants on a five-point Likert scale about how much they trust the biased model and whether it makes sense to them. Note, that the majority of participants did not detect the bias of the model.



(a) Bias detected and correct preference.

(b) Bias detected but no preference change.

Figure 7: How participants changed their responses comparing the unbiased model to the biased model. A positive value indicates that the response was higher in the unbiased model. (a) shows the case when participants detected the bias and subsequently preferred the unbiased model. (b) shows the case when participants detected the bias but still chose the biased model.

In each of those conditions we compare the ability to detect biases in the data by comparing the unbiased data set to the data set with the manufactured bias.

We also explore the impact of those conditions on whether explanations and aggregation improve trust in the model's decisions.

5.3 Tasks and Measurements

In order to test conditions against each other we created a questionnaire. After asking the participant about their knowledge of machine learning and basic terminology, we have a training section for participants to familiarize themselves with the interface. First, an introductory video explains all components of the interface. The video uses an example model from a different data set which is designed to predict whether a room is occupied or not based on predictions from various sensors [14]. Then, a series of questions about this example model are asked and the participant can and has to use the interface to answer them correctly. The questions ask about the values of features under certain conditions, such as "What is the model's prediction for high values of 'CO_2'?", "What is the lowest value of 'Humidity' that predicts 'occupied'?", "Are the predictions for low values of 'Light' correct?". The questions are constructed in a way to be easily answerable under all conditions given an understanding of the user interface and basic principles of machine learning. An incorrect answer leads back to the beginning of the section and the participant is given the chance to correct the mistake. We did not use those questions to exclude any participant but rather for giving them an opportunity to get comfortable with the user interface. Note, that it is not necessary for participants to have a deep knowledge in *how* machine learning algorithms work, as long as, the basic principles of prediction, ground truth, or accuracy are clear. This reflects that domain experts would often not necessarily be trained in *developing* machine learning models but rather *using* them.

A final question participants answer is a hypothetical scenario for when a prediction does not make *sense* from a semantic standpoint, even though that prediction is *correct* from the perspective of the model. This question aims to prime the participants for the upcoming task and teaches that model correctness is not necessarily equivalent to semantic correctness. After this, we ask some common sense questions about how house prices are supposed to correspond to certain features. This ensures that participants have enough domain knowledge for the upcoming task.

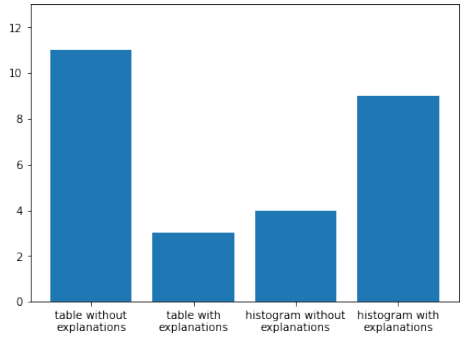
In the main part of the study, we present the participant with both housing data models one after the other and encourage them to explore the models with the end goal of determining which model can be trusted more. The order of the data sets is random. The participant then has to answer the following questions about each model: "Do you think the predictions of the model make sense?", "How well does the model perform in terms of accuracy?", "How much do you trust the model?" on a five-point Likert scale; and explain the reasoning for their answers. For each question we provide a more in-depth explanation with examples.

After inspecting both models, we ask participants to state which model performs better in terms of accuracy, which model can be trusted more, and whether the model they trust more had the higher or lower accuracy, or if no model can be trusted more than the other. We ask participants to describe their reasoning and also state their confidence in their decision on a five-point Likert scale.

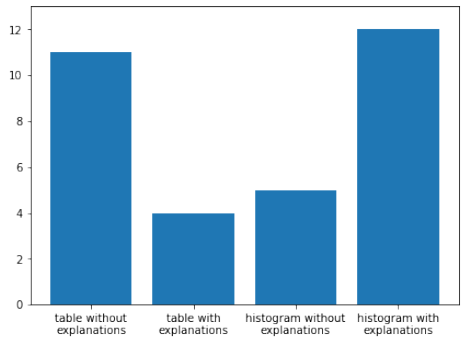
5.4 Participants

We ran all four conditions of the study on Prolific[1], an online survey recruitment system. Participants in online recruitment aim to increase their payout to effort ratio. Thus, we took several measures to ensure high data quality. First, we only allowed participants with a high rating on the platform and an interest in computer science. Secondly, we excluded all data from participants that had a suspiciously fast completion times (less than 10 minutes after watching the introductory video) which would not allow them to establish well thought out answers. We also excluded participants with too little interaction with the interface, determined by the number of histograms or table cells they inspected and how often they changed the subset comparison selection. We also asked a simple question with a clear answer to ensure participations were paying attention. If a participant did not correctly answer the question "Which model

[1] https://www.prolific.ac/



(a) Correct trust



(b) Detected bias

Figure 8: (a) compares how many participants trusted the unbiased, thus correct, model more. (b) compares how many participants correctly identified the bias in the data, determined from plain-text answers. Note, that in both cases adding explanations to the table view hurt performance, whereas adding explanations to the histogram view improved performance.

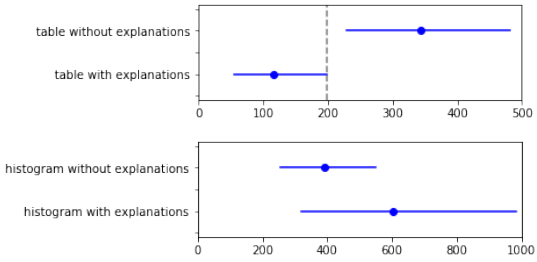


Figure 9: The left side shows interactions with the table view measured by counting how many table cells were hovered by the mouse. The right side shows interactions with the histogram view measured by counting how many histogram bars were hovered by the mouse. The plot shows the bootstrapped mean and confidence interval for each setting.

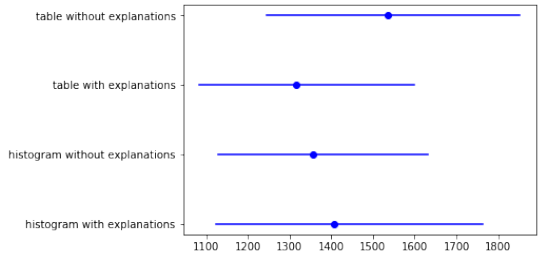


Figure 10: Overall completion time of the study by condition. The plot shows the bootstrapped mean and confidence interval for each setting. There is no significant difference between the conditions.

had the higher accuracy?", the participant was removed. This question has an objective answer that had to be determined during the study as well. We retained 100 eligible participants divided evenly across the four conditions. This represents less than 47% of total participants, not counting participants that stopped the study before submitting.

6 RESULTS

6.1 Bias Detection and Trust

When comparing how much participants trust a particular model and whether they think this model makes sense, one can see that those responses are typically correlated (see for example Figure 6 with a Pearson correlation coefficient of 0.759 and Spearman rank-order correlation coefficient of 0.745). This also extends to plain-text answers, which allow to detect whether participants correctly found the bias in the correct data set unambiguously. Participants were very verbose about their findings, if they found something: "*It has higher accuracy so should be more trustworthy than the other one. However some of the results don't make sense to me. Maybe this is just an atypical property market.*"; "*It is accurate, yet the predictions do not make much sense. Higher quality houses having a larger amount of low priced houses, percentage-wise? More rooms, area, or stories resulting in lower prices? The logic does not work out.*"; "*larger houses are valued lower than others which are smaller*" (sic).

However, the above mentioned correlation is not perfect. This is likely due to some participants not being convinced, that their correct discovery of the flaw in the data is enough that the corresponding model cannot be trusted: "*If the data says it's true, then it's true I suppose and it's more trustworthy than my common sense.*"; "*I feel like the results of [the biased model] where strange even though they where correct according to the dataset.*"; "*I'm drawn to trusting the model which was more accurate even though it didn't entirely make sense to me.*" (sic).

This divergence in trust and the finding of flaws in the data can also be seen in Figure 7. If finding the flaw in the data swayed the participant to not trust the model, an increase in trust for the unbiased model compared to the biased model corresponds to an increase in the perception that the unbiased model makes more sense (Figure 7a). However, if the finding did not influence the preference, trust between both models stayed the same (Figure 7b).

In total, 25% of people who correctly identified the bias still opted to trust the biased model more, due to the higher reported accuracy of the model. A further 8% who identified the bias trusted both models equally. This aligns with the findings of Stumpf *et al.* [21] that trust

in the machine learning model may *override* people's initial intuition about its performance.

6.2 Comparison Across Conditions

Comparing the correctness across all four conditions can be seen in Figure 8. First, we can see a strong improvement both in correctness and whether the participant trusted the unbiased model more, when switching from tables to histograms while having access to explanations (p-value Figure 8a: Fisher's 0.0477, χ^2 0.0489; p-value Figure 8b: Fisher's 0.0161, χ^2 0.0169). When adding explanations to histograms (p-value Figure 8b: Fisher's 0.0359, χ^2 0.0366) we can see a significant improvement when comparing correctness. We hypothesize that explanations are a necessity for histograms to work effectively, since they point out which, of the possibly many, pattern seen in the distributions are meaningful. We can also see an improvement in whether the participant trusted the unbiased model, however, it is not significant (p-value Figure 8a: Fisher's 0.0982, χ^2 0.0986).

Furthermore, we see a strong decline in correctness when adding explanations to tables (p-value Figure 8a: Fisher's 0.0127, χ^2 0.0137; p-value Figure 8b: Fisher's 0.0311, χ^2 0.0320). At first, we were puzzled at this counter-intuitive result and we double and triple checked that those results were not a simple mix-up in conditions. We hypothesize that having explanations in a table focuses the attention of participants to fewer instances and additionally makes them more confident that they fully understood the model. This extrapolation from few instances aligns with the findings of Stumpf *et al.* [21], who found that explanations can be harmful in certain circumstances, and shows that the findings also apply to a tabular representation of the explanations.

In order to investigate this hypothesis further, we can look at the number of interactions of the participants performing the tasks. We can see in Figure 9 that participants engaged with the table view significantly more when no explanations were present. This might be an example of Hullman *et al.* [7], who state that information visualization might benefit from visual difficulties, since people are forced to interact more with the visualization. This seems to be the case for a table, without any further help from the interface about what to look at.

Despite that, we found no significant difference in the time participants took to complete the study, as can be seen in Figure 10. Even though the histogram view with explanations and the table view without explanations do not have a significant timing difference, we hypothesize that an aggregated representation of the model is a more effective method for finding biases. This hypothesis is rooted in both conditions performing equally well and histograms being a more scalable data representation than tables, due to their independence from the magnitude of the data.

7 DISCUSSION

We showed that aggregating instance-level explanations can be an effective way of enabling humans to identify biases in the input data of machine learning tasks. Even though, assisted aggregation is as effective as unassisted individual inspection, we argue that aggregation it scales better with large data sets. Individually inspecting instances in the data is only possible on a sample of the data and requires extrapolation of findings to the whole data set. Aggregation does not suffer from this, as the representation of the data is independent from its size. Even though, it requires dedication, our test data set was small enough to still be able to scan in full if necessary.

Furthermore, the bias planted in the data was simple enough to be able to be found under *all* conditions. This might not be true for real-world data sets with more complex biases. Even though histograms are advantageous with respect to tables in finding arbitrary patterns, they are still limited to only one dimension. Biases that are present only through combinations of features will not be detectable.

In our study, we confirmed findings from Stumpf *et al.* [21] and overcome their limitations by using instance-level explanations with aggregation. However, we could not overcome trust in machine learning model authority, despite being confronted with contradictory evidence in all cases. Speculatively, this might stem from people being used to being presented with cleaned up and validated data, as this cumbersome process is often hidden from the end result.

8 CONCLUSION & FUTURE WORK

We presented a novel way of aggregating and comparing instance-level explanations. We found that this method can help humans identify biases in the input data to machine learning models. However, this is only the case in combination. Aggregation alone or individual instance-level explanations might lead to worse performance in this regard. We demonstrate that an aggregated instance-level explanation approach is as effective as going through the data unassisted. This is promising, as the proposed method is independent of the size of the data set and thus likely more scalable than its non-aggregated counterpart. However, confirming this hypothesis remains future work.

As we were conducting an exploratory analysis of the study, individual findings remain to be tested in-situ in future work. Furthermore, experimenting with more complex forms of data biases opens up additional research opportunities.

In summary, we present a usable method for effectively utilizing instance-level explanations on a large scale. As machine learning models become more complex and opaque, this becomes an important initial contribution in improving the interpretability of machine learning models and their data alike.

REFERENCES

[1] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1721–1730, New York, NY, USA, 2015. ACM.

[2] J. C. Chang, S. Amershi, and E. Kamar. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 2334–2346, New York, NY, USA, 2017. ACM.

[3] D. D. Cock. Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. *Journal of Statistics Education Volume 19, Number 3*, 2011.

[4] Google PAIR. Facets. [Online]. Available: https://pair-code.github.io/facets/, 2017.

[5] G. E. Hinton. Connectionist learning procedures. *Artificial Intelligence*, 40(1):185–234, 1989.

[6] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau. Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *ArXiv e-prints*, Jan. 2018.

[7] J. Hullman, E. Adar, and P. Shah. Benefitting infovis with visual difficulties. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2213–2222, Dec 2011.

[8] J. Krause, A. Dasgupta, J.-D. Fekete, and E. Bertini. SeekAView: an intelligent dimensionality reduction strategy for navigating high-dimensional data spaces. *Large Data Analysis and Visualization (LDAV), IEEE Symposium on*, Oct 2016.

[9] J. Krause, A. Dasgupta, J. Swartz, Y. Aphinyanaphongs, and E. Bertini. A workflow for visual diagnostics of binary classifiers using instance-level explanations. *IEEE VAST*, 2017.

[10] J. Krause, A. Perer, and E. Bertini. Using Visual Analytics to Interpret Predictive Machine Learning Models. *ArXiv e-prints*, June 2016.

[11] J. Krause, A. Perer, and K. Ng. Interacting with predictions: Visual inspection of black-box machine learning models. *Proceedings of the ACM SIGCHI Conf. on Human Factors in Computing Systems (CHI)*, pages 5686–5697, 2016.

[12] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, pages 126–137, New York, NY, USA, 2015. ACM.

[13] Z. C. Lipton. The Mythos of Model Interpretability. *ArXiv e-prints*, June 2016.

[14] V. F. Luis M. Candanedo. Accurate occupancy detection of an office room from light, temperature, humidity and co2 measurements using statistical learning models. *Energy and Buildings. Vol. 112, p. 28–39*, 2016.

[15] D. Martens and F. Provost. Explaining data-driven document classifications. *MIS Q.*, 38(1):73–100, Mar. 2014.

[16] M. Narayanan, E. Chen, J. He, B. Kim, S. Gershman, and F. Doshi-Velez. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *ArXiv e-prints*, Feb. 2018.

[17] A. Perer and B. Shneiderman. Systematic yet flexible discovery: guiding domain experts through exploratory data analysis. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 109–118. ACM, 2008.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016.

[19] M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[20] P. Y. Simard, S. Amershi, D. M. Chickering, A. E. Pelton, S. Ghorashi, C. Meek, G. Ramos, J. Suh, J. Verwey, M. Wang, and J. Wernsing. Machine teaching: A new paradigm for building machine learning systems. *CoRR*, abs/1707.06742, 2017.

[21] S. Stumpf, A. Bussone, and D. O'Sullivan. Explanations considered harmful? user interactions with machine learning systems. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2016.

[22] M. Tulio Ribeiro, S. Singh, and C. Guestrin. Nothing Else Matters: Model-Agnostic Explanations By Identifying Prediction Invariance. *ArXiv e-prints*, Nov. 2016.

[23] M. Wattenberg, F. Viegas, and M. Hardt. Attacking discrimination with smarter machine learning. [Online]. Available: https://research.google.com/bigpicture/attacking-discrimination-in-ml/, 2016.

[24] J. Zhou, S. Z. Arshad, S. Luo, and F. Chen. Effects of uncertainty and cognitive load on user trust in predictive decision making. In *16th IFIP TC 13 International Conference on Human-Computer Interaction — INTERACT 2017 - Volume 10516*, pages 23–39, New York, NY, USA, 2017. Springer-Verlag New York, Inc.