Effects of Influence on User Trust in Predictive Decision Making

Jianlong Zhou, Zhidong Li
University of Technology Sydney
Sydney, NSW, Australia
firstName.lastName@uts.edu.au

Huaiwen Hu
Data61, CSIRO
Sydney, NSW, Australia
vivi.hu@data61.csiro.au

Kun Yu, Fang Chen
University of Technology Sydney
Sydney, NSW, Australia
firstName.lastName@uts.edu.au

Zelin Li
Data61, CSIRO
Sydney, NSW, Australia
zelin.li@data61.csiro.au

Yang Wang
University of Technology Sydney
Sydney, NSW, Australia
yang.wang@uts.edu.au

ABSTRACT

This paper introduces *fact-checking* into Machine Learning (ML) explanation by referring training data points as facts to users to boost user trust. We aim to investigate what influence of training data points, and how they affect user trust in order to enhance ML explanation and boost user trust. We tackle this question by allowing users check the training data points that have the higher influence and the lower influence on the prediction. A user study found that the presentation of influences significantly increases the user trust in predictions, but only for training data points with higher

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland UK

© 2019 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-5971-9/19/05.
https://doi.org/10.1145/3290607.3312962

influence values under the high model performance condition, where users can justify their actions with more similar facts.

CCS CONCEPTS

• **Human-centered computing** → **Interaction design**; *Visualization.*

KEYWORDS

Influence; machine learning; model performance; trust

INTRODUCTION

We continuously find ourselves coming across Machine Learning (ML) based Artificial Intelligence (AI) systems that seem to work or have worked surprisingly well in practical scenarios, ML technologies still face prolonged challenges with low user acceptance as well as seeing system misuse, disuse, or even failure. These fundamental challenges can be attributed to the nature of the "black-box" of ML for domain experts when offering ML-based solutions [7]. As a result, recent research suggests model *explanation* as a remedy for the "black-box" ML. Taking the influence of training data points on predictions [2] as an example,the explanation with influence allows to capture the weight/contribution of each training data point on the prediction. However, these explanations are highly biased towards ML experts' views, while domain users are more interested in what influence information affect and how these influence information are presented to them to boost their trust in predictions. Besides explanation, the ability to provide justifiable and reliable evidences for ML-based decisions would increase the trust of users. Recently, *fact-checking*, which provides "evaluation of verifiable claims made in public statements through investigation of primary and secondary sources" [3], is increasingly used to check and debunk online information because of credibility challenges.

Motivated by these investigations, this paper introduces *fact-checking* into ML explanation by referring training data points as facts to users to boost user trust. These training data points are selected based on their influence values on predictions. We aim to investigate what influence of training data points and model performance, and how they affect user trust in order to enhance ML explanation and boost user trust. We tackle this question by allowing users check the training data points that have the higher influence and the lower influence on the prediction.

RELATED WORK

Zhou et al. investigated different approaches to reveal human cognition states such as user trust in predictive decision making scenarios [6, 8]. Various researches have also been investigated to learn user trust variations in ML. Ye and Johnson [5] experimented with three types of explanations (trace, justification and strategy) for an expert system, and found that justification was the most effective

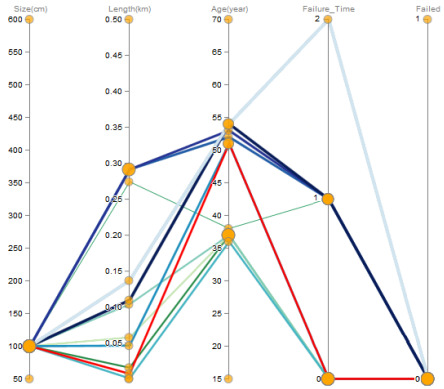

Figure 1: Fact-checking visualization. (Each vertical axis represents one data attribute with the sorted descending order, and a polyline connecting points on each vertical axis represents a data point. Each polyline represents one water pipe with various attributes. Various pipe attributes belonging to one pipe are encoded with the same color. Testing pipe is encoded with red color. The influence of each training pipe on the prediction of a test is encoded with the width of polylines, the wider the polyline, the higher the influence.)

type of explanation in changing users' attitudes towards the system. Other studies that empirically tested the importance of explanation to users, in various fields, consistently showed that explanations significantly increase users' confidence and trust [4]. Explanation also shows the ability to correctly assess whether a prediction is accurate [1].

These previous work motivates us to consider both algorithmic explanations and model performance in the interpretability of ML, aiming to find what explanations (e.g. influence of training data points) and how these explanations affect user trust in ML.

METHOD

Case Study and Fact-Checking Visualization

This paper uses water pipe failure prediction as a case study for predictive decision making. Pipes are characterized by different attributes, such as laid year, material, diameter size, etc. A pipe failure prediction model is set up based on the pipe failure historical data [10]. Such models are used by utility companies for budget planning and pipe maintenance. However, different models with various presentation of influence of training data points and prediction performance may be achievable resulting in different possible management decisions. The experiment is set up to determine what influence and model performance may affect the user trust during the decision process.

We present a visualization approach called fact-checking visualization for presenting multiple data attributes based on parallel coordinates as shown in Figure 1. Figure 1 demonstrates how similar the training pipes are with the testing pipe in red color. The pipe attributes visualized in Figure 1 include pipe size (diameter), pipe length, pipe age, failure times during the observation period, and whether it was failed in the checked year (0 means not failed and 1 means failed).

Framework of Fact-Checking for Boosting User Trust

We present a framework of fact-checking for boosting user trust in a predictive decision making scenario (see Figure 2). In a typical conventional ML pipeline, a training data is used to train an ML model and predictions are made based on the trained model (as shown in the lower unshaded part in Figure 2). There is no information on the ML explanation in order to promote the trustworthiness of the prediction. An influence-enhanced fact-checking approach is added on the top of the conventional ML pipeline in the proposed framework (as shown in the upper shaded part in Figure 2) to explain predictions and boost user trust in predictions. Firstly, the influence of all training data points for the prediction of a testing data point is calculated with influence functions [2]. All training data points are then ranked in descending order based on the calculated influence values. Training data points which have the higher influence values (e.g. the top 10 training data points in the ranking) and the lower influence values (e.g. the bottom 10 training data points) are obtained respectively based on

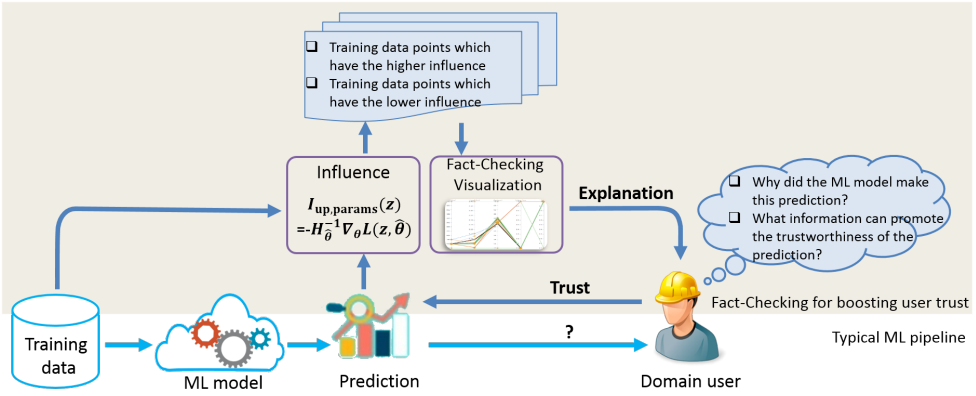

Figure 2: A Framework of Fact-Checking for Boosting User Trust.

the ranking. These training data points function as facts which are the most similar points and the least similar points to the testing data point respectively. The parallel coordinate based visualization is used to visualize these selected ranked training data points allowing users to compare the facts with the testing data points to boost trust in predictions.

EXPERIMENT

Experimental Data

Water pipe failure prediction uses historical pipe failure data to predict future failure rate. The pipe features used in the experiment include the pipe age, pipe size (diameter), length, and failure times during the observation period. Convolutional neural network (CNN) [9] was trained to model the water pipe failures. In this study, two CNN models were trained using different network settings, resulting in the model accuracy of 90% and 55% respectively. These two model performances were used as high model performance (90%) and low mode performance (55%) respectively to find differences of user responses in the experiment. Furthermore, the influence of each training pipes on the prediction of a testing pipe was calculated with the use of influence functions introduced in [2].

Task Design

In this experiment, the top 10 (TOP10) and bottom 10 (BOT10) training pipes based on the ranking, which have the highest and lowest influence on predictions respectively, are selected. The fact-checking visualization based on parallel coordinates is then used to visualize the TOP10 and BOT10 pipes

Table 1: Task setup in the experiment. (These 8 tasks were conducted two rounds with all same settings except testing pipes used. Two training tasks were also conducted by each participant before formal tasks. In summary, there were 18 tasks conducted (8 tasks × 2 rounds + 2 training tasks = 18 tasks) by each participant.)

		Influence			
		TOP10	**BOT10**	**TOP10& BOT10**	**Control**
Model Perf.	**High**	T1	T2	T3	T4
	Low	T5	T6	T7	T8

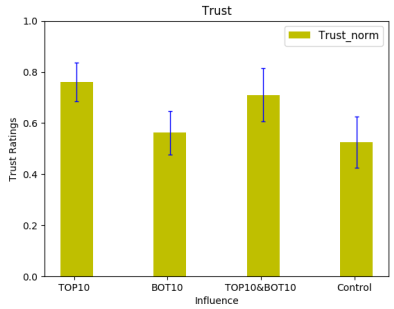

Figure 3: Normalized trust values by influence under high model performance. (The post-hoc tests found that participants had significantly higher trust in predictions when influences of TOP10 training pipes were presented than that without influence information presentation ($Z = 102.0, p < .001$). Participants also showed significantly higher trust in predictions when influences of both TOP10 and BOT10 were presented than that without influence information presentation ($Z = 120.5, p < .004$). It was also found that participants had significantly higher trust in predictions when influences of TOP10 training pipes were presented than that when influences of BOT10 training pipes were presented ($Z = 61.5, p < .001$). However under low model performance, statistically significant differences of trust among different influence conditions have not been found.)

respectively. This experiment divides fact-checking visualization settings for tasks into four categories: 1) *TOP10*, 2) *BOT10*, 3) *TOP10&BOT10* which includes both TOP10 and BOT10 visualizations in tasks, and 4) *Control* which does not include any influence visualization on training pipes. By considering both model performance cases (high and low performance) and fact-checking visualization conditions, we finally got 8 tasks as shown in Table 1. Each participant was asked to make a decision on whether to replace a testing pipe, using water pipe failure prediction models under different settings as shown in Table 1. The task orders were randomized during the experiment.

22 participants were recruited, of all participants, 5 were females. After each decision making task, participants were asked to rate the trust level of predictions on which decisions were made using a 9-point Likert scale (from 1: least trust, to 9: most trust).

ANALYSIS

In this study, we aim to understand: 1) the effects of influence on user trust under a given model performance, and 2) the effects of model performance on user trust under a given influence condition respectively. Trust values were normalized with respect to each subject to minimize individual differences in rating behavior.

Influence and Trust: Figure 3 shows mean normalized trust values over different influence settings under high model performance. Friedman's test gave statistically significant differences in trust among four influence conditions, $\chi^2(3) = 21.675$, $p = .000$. Then post-hoc Wilcoxon tests (with a Bonferroni correction under a significance level set at $p < .013$ $(0.05/4)$) was applied to find pair-wise differences between influence conditions. The results suggest that the presentation of influence of training data points on predictions significantly increases the user trust in predictions, but only for training data points with higher influence values under the high model performance condition.

Model Performance and Trust: Figure 4 shows mean normalized trust values over two model performance conditions (high and low) under different influence settings. The results suggest that high model performance together with influence information result in the higher user trust in predictions.

DISCUSSIONS AND ONGOING WORK

Overall, we can say that the influence information of training data points (functioned as fact-checking) on predictions can benefit trust, where users can check the facts that are similar to the testing data point. Presentation of influence information of training data points having the higher influence values can lead to increased trust but only under the high ML model performance condition, where users can justify the action with more similar facts and fit their general understanding of the problem.

In order to make ML-driven AI applications not only intelligent but also intelligible, the user interface of AI applications needs to allow users to access the most influential facts to predictions by

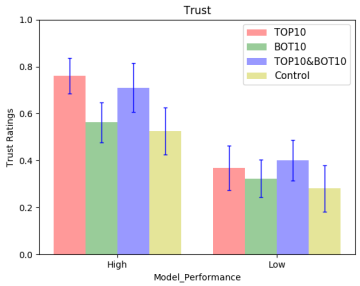

Figure 4: Normalized trust by model performance under different influences. (It was found that participants showed significantly higher trust under high model performance than that under low model performance over all four influence settings (TOP10: $Z = 11.5, p < .000$**; BOT10:** $Z = 52.0, p < .000$**; TOP10&BOT10:** $Z = 77.5, p < .000$**; Control:** $Z = 77.5, p < .001$**).)**

CONCLUSIONS

This paper investigated the influence enhanced fact-checking for the ML explanation to boost user trust. A framework of fact-checking for boosting user trust was proposed to allow users interactively check the training data points with the use of parallel coordinates based visualization. A user study found that the presentation of influence of training data points on predictions significantly increased the user trust in predictions, but only for training data points with higher influence values under the high model performance condition. These findings suggested that the access of the most influential facts to predictions by users in the user interface of AI applications would help users get the rational behind their actions and therefore benefit the user trust in predictions.

visualizations. Such influence-enhanced fact-checking allows users find similar facts to the testing data point to get the rational behind for the justification of their actions, therefore boosting user trust.

Our future work will focus on the examination of user behaviour changes with visualizations presented, and quantification of how the overall user experience/performance can be improved with the proposed method. Other visualization approaches for the fact-checking will also be investigated especially for training data points with lower influence values under the low model performance.

ACKNOWLEDGEMENTS

This work is partly supported by the Asian Office of Aerospace Research & Development (AOARD) under grant No. AOARD 216624.

REFERENCES

[1] Or Biran and Kathleen McKeown. 2017. Human-centric Justification of Machine Learning Predictions. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17)*. 1461–1467.

[2] Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*. Sydney, NSW, Australia, 1885–1894.

[3] Travis Kriplean, Caitlin Bonnar, Alan Borning, Bo Kinney, and Brian Gill. 2014. Integrating On-demand Fact-checking with Public Dialogue. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. 1188–1199.

[4] Panagiotis Symeonidis, Alexandros Nanopoulos, and Yannis Manolopoulos. 2009. MoviExplain: A Recommender System with Explanations. In *Proceedings of the Third ACM Conference on Recommender Systems*. 317–320.

[5] L. Richard Ye and Paul E. Johnson. 1995. The Impact of Explanation Facilities on User Acceptance of Expert Systems Advice. *MIS Quarterly* 19, 2 (June 1995), 157–172.

[6] Jianlong Zhou and Fang Chen. 2018. *DecisionMind*: revealing human cognition states in data analytics-driven decision making with a multimodal interface. *Journal of Multimodal User Interfaces* 12, 2 (2018), 67–76.

[7] Jianlong Zhou and Fang Chen (Eds.). 2018. *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*. Springer, Cham.

[8] Jianlong Zhou, M. Asif Khawaja, Zhidong Li, Jinjun Sun, Yang Wang, and Fang Chen. 2016. Making Machine Learning Useable by Revealing Internal States Update – A Transparent Approach. *International Journal of Computational Science and Engineering* 13, 4 (2016), 378–389.

[9] Jianlong Zhou, Zelin Li, Weiming Zhi, Bin Liang, Daniel Moses, and Laughlin Dawes. 2017. Using Convolutional Neural Networks and Transfer Learning for Bone Age Classification. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA 2017)*. 1–6.

[10] Jianlong Zhou, Jinjun Sun, Yang Wang, and Fang Chen. 2017. Wrapping Practical Problems into a Machine Learning Framework: Using Water Pipe Failure Prediction As a Case Study. *International Journal of Intelligent Systems Technologies and Applications* 16, 3 (2017), 191–207.