Correlation for User Confidence in Predictive Decision Making

Jianlong Zhou, Syed Z. Arshad, Kun Yu, and Fang Chen
Data61, CSIRO
13 Garden Street, Eveleigh, NSW 2015, Australia
{firstName.lastName}@data61.csiro.au

ABSTRACT
Despite the recognized value of Machine Learning (ML) techniques and high expectation of applying ML techniques within various applications, significant barriers to the widespread adoption and local implementation of ML approaches still exist in the areas of trust (of ML results), comprehension (of ML processes), as well as confidence (in decision making) by users. This paper investigates the effects of correlation between features and target values on user confidence in data analytics-driven decision making. Our user study found that revealing the correlation between features and target variables affected user confidence in decision making significantly. Moreover, users felt more confident in decision making when correlation shared the same trend with the prediction model performance. These findings would help design intelligent user interfaces and evaluate the effectiveness of machine learning models in applications.

Author Keywords
Data analytics; decision making; correlation; confidence

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
With the rapid increase of data from various fields such as biology, finance, medicine, infrastructure, and society, users are looking to integrate their "Big Data" and advanced analytics into business operations in order to become more analytics-driven in their decision making. Much of machine learning (ML) research is inspired by such expectations. As a result, we continuously find ourselves coming across ML-based appealing viewgraphs and other predictions that seem to work (or have worked) surprisingly well in practical scenarios.

Despite the recognized value of ML techniques and high expectations of applying ML techniques within various applications, for many of these non-technical users, ML based predictive analytics software is like a "black box", to which they simply provide their input data and (after selecting some menu options on screen) colorful viewgraphs and/or recommendations are displayed as output. This "black box" approach has obvious drawbacks, e.g. it is neither clear nor well understood that how trustworthy is this output, or how uncertainties were handled/manipulated by underlying algorithmic procedures (Zhou, Khawaja, et al., 2015). The user is more or less unconfident in the ML model output when making predictive decisions, and thus also unconfident in the ML methods themselves. In a word, significant barriers to widespread adoption and local implementation of ML approaches still exist in the areas of trust (of ML results), comprehension (of ML processes), as well as confidence (in recommended courses of action or decision making) by users. As a result, the User Experience involved in real world ML applications has been more recently identified as an area requiring research and development (innovation) (Wagstaff, 2012; Zhou, Khawaja, et al., 2015; Zhou, Sun, et al., 2015; Zhou, Bridon, Chen, Khawaji, & Wang, 2015; Zhou, Li, Wang, & Chen, 2013).

Therefore, in an ML-based intelligent system, it is highly critical to know how the information presented in the user interface on data and ML models affect user's confidence in order to make effective decisions. User confidence also affects the acceptance of ML methods in practical applications. Such investigation could help develop more effective user interface for ML-based intelligent systems.

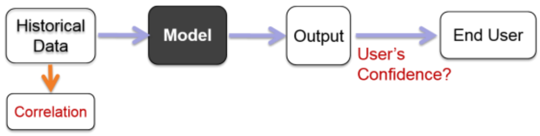


Fig. 1. ML-based data analysis pipeline.

Fig. 1 shows a typical ML-based data analysis pipeline. In this pipeline, inputs to ML models are often historical records or samples of some event. They are usually not the precise description of events. ML models are also imperfect abstractions of reality. Furthermore, from the input data perspective, statistical information of data such as correlation between variables also provides useful information to let users learn relations between variables. For example, correlation can describe how much target values are related with features in input data of the model. The correlation may affect users' decision making based on their domain experiences, e.g. domain experts may have experiences that the older the pipes are, the higher the failure rate is. The user might be risking too much by completely ignoring statistical information such as

OzCHI '16, November 29 – December 02 2016, Launceston, TAS, Australia
 ISBN 978-1-4503-4618-4/16/11 $15.00
http://dx.doi.org/10.1145/3010915.3011004.

correlation from input data and having complete faith in system output/recommended values. On the other hand, over-confidence in such statistical information and thus decisions may result in unnecessary economic loss.

In this paper, we aim to investigate relationships between correlation and user confidence in decision making in order to design more effective user interface for ML-based intelligent systems and improve the acceptability of ML techniques by users. A user study was performed to investigate the impact of revealing correlation information on user confidence in predictive decision making. In the following sections of the paper, except the specific pointing out, correlation refers to the correlation between a feature variable and the target variable in input data.

RELATED WORK

Making decisions is one of the most complex cognitive processes and there is a long history of investigation in different domain areas. For example, Morgado et al. (Morgado, Sousa, & Cerqueira, 2015) reviewed the impact of stress in decision making and found that this complex cognitive process involves several sequential steps. Making good decisions implies an estimate not only of the value and the likelihood of each option but also of the costs and efforts implied in obtaining it. Kahneman et al. (Kahneman & Tversky, 1979),(Tversky & Kahneman, 1974) suggested that people make a variety of errors when making decisions (or solving problems) involving probability. The Subjective Expected Utility (SEU) model suggests that the decision weights people attach to events are their beliefs about the likelihood of events (Kelsey, 1994). (Camerer & Weber, 1992) summarized that it was thought that people prefer to bet on events they know more about. However, this was shown to be otherwise by (Hsu, 2005) in their study of neural systems responding to degrees of uncertainty. Their experiments showed that many people are more willing to bet on risky outcomes than ambiguous ones, when holding the judged probability of outcomes constant.

Lee and Dry (Lee & Dry, 2006) showed that human confidence in decision making does not depend solely on the accuracy of the advice, it is also influenced by the frequency of the advice. Considering that decisions are often made based on probability evaluations of which users are not entirely sure, Hill (Hill, 2013) developed a decision rule incorporating users' confidence in probability judgments. A formal representation of the decision maker's confidence is also presented in (Hill, 2013). Moran et al. (Moran, Teodorescu, & Usher, 2015) argued that a critical property of decision confidence is its positive correlation between confidence and decision correctness. In other words, with higher confidence the decider is more likely to be correct in his or her decision. In a typical decision making scenario, once the problem scenario along with supplementary material is presented, several other factors can come into play as well. One such group of factors is individual differences that were investigated by Pallier et al. (Pallier et al., 2002). Differences in experience, motivation, attitudinal predispositions etc. can have an impact on decision making process.

Good decision-making often requires people to perceive and handle a myriad of statistical correlations (Eyster & Weizsacker, 2010). However, (Eyster & Weizsacker, 2010) found that people have limited attention and often neglect correlations in financial decision making. (Ye, 2013) used the weighted correlation coefficients to rank the alternatives and get the best alternative in multi-attribute decision making. (Liao, Xu, Zeng, & Merigó, 2015) used correlation coefficients of hesitant fuzzy linguistic term set in the process of qualitative decision making in traditional Chinese medical diagnosis. However, few researches have been carried out into the roles of correlation in user confidence in data analytics-driven (or predictive) decision making. Domain experts are good at utilizing experiences in their decision making, while correlation from input data reflects statistic summaries of historical facts. We highly argue that the revealing of correlation between features and target values would significantly affect user confidence in decision making.

HYPOTHESES

The following hypotheses are posed in our study:

- Revealing of correlation between features and target values would help users more confident in data analytics-driven decision making (H1);
- The pattern between correlation and performance of model output would affect user confidence in data analytics-driven decision making (H2);
- When correlation and performance of model output share the same trend, users would be more confident in user confidence decision making (H3).

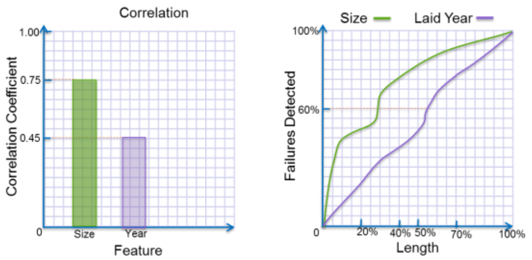
EXPERIMENT

There were 26 subjects recruited in predictive decision making tasks. Ages ranged from 23 to 45 with an average age of 30 years. Educational qualifications were largely postgraduate (13 PhD, 6 Masters, 4 Bachelors, 3 other). Subject subgroups comprised of nine machine learning experts, eight non-machine learning experts and nine administrative staff.

Water pipe failure prediction is used as a case study for this research (Li et al., 2014). Different ML models based on alternative water pipe features may be available resulting in different possible budget plans in decision making. This experiment is set up to determine what criteria of choice are in favor of a model, and what parameters influence the user confidence during the decision process. Water pipe failure prediction uses historical pipe failure data to predict future failure rate. The historical data contains failure records as well as other attributes of water pipes, such as laid year, length, material, diameter size, location, etc. Actual historical data was repeatedly sampled and customized for the simulation of this experiment.

In this study, ML models are simulated and they are based on different pipe features (e.g. size or laid year). The model performance curve was presented to let the participants evaluate different models. The model performance is the functional relationship between proportion of the network inspected (in percent) and the proportion of pipe failures detected (in percent) for two different feature variables. For example, Fig. 2(b) shows the performances of two models, where the model based on the feature "Size" has better performance than the one based on the feature "Laid Year", because the formal one detects more failures than the latter for a given pipe length.

As represented in Fig. 1, correlation is not associated with a model, but associated with input data. Correlation in this experiment refers to the correlation between one pipe feature (e.g. pipe size) and the pipe failure rate in historical records. The correlation is often described by correlation coefficient. Correlation coefficient illustrates a quantitative measure of correlation and dependence, meaning statistical relationships between two or more random variables or observed data values. The correlation in this experiment is displayed as 2D bar charts with the horizontal axis being features and the vertical axis being the correlation coefficients (see Fig. 2(a) and Fig. 3(a)). For example, in Fig. 3(a), the feature “Laid Year” (Year) and “Size” have a correlation coefficient of 0.75 and 0.45 respectively with failure rate, meaning that “Laid Year” is more related to the failure rate than “Size”.

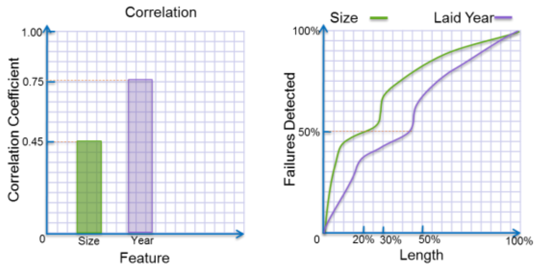


(a) (b)

Fig. 2. Correlation and performance of model output share the same trend.

The relations between model performance and correlation are divided into two groups:

- Correlation and performance of model output share the same trend (see Fig. 2). That is, the correlation between a feature and the target variable (pipe failure rate) is high and the associated model performance is also high, or the contrary.
- Correlation and performance of model output do not share the same trend (see Fig. 3). That is, the correlation between a feature and the target variable (pipe failure rate) is high, but the associated model performance is low, or the contrary.



(a) (b)

Fig. 3. Correlation and performance of model output do not share the same trend.

Based on these groups, the tasks were divided into three categories (Same Trend Task, Non-Same Trend Task, and No-Correlation Task). In the Same Trend Task (ST), correlation and performance of model output share the same trend (see Fig. 2). In the Non-Same Trend Task (NST), correlation and performance of model output do not share the same trend (see Fig. 3). The No-Correlation Task (NCT) only presents model performance diagram but not correlation diagram to participants.

During the experiment, each participant was told he/she would be a user of ML predictive systems at the local water department. The water company plans to repair XX% pipe failures in the next financial year. He/she was asked to make a budget plan, i.e. a budget in network length, using water pipe failure prediction models learned from the historical water pipe failure records. Two ML models were provided for each budget task. Participants were required to make decisions by selecting one of presented ML models and then making budget plan based on the selected ML model. The budget plan needs to meet:

- Check as short length of pipes as possible (low cost);
- The budget uncertainty interval should be as small as possible (high accuracy).

All tasks were conducted with two rounds. The first round used the feature pair of “Size – Laid Year” and the second round used the feature pair of “Material – Pressure” for ML models. Except feature name differences, two rounds used same model performance diagrams.

After each trial, participants were asked to rate the confidence level of the budget plan they made and the difficulty level of the task using a 9-point Likert scale (1: least difficult/confident, and 9: most difficult/confident).

RESULTS

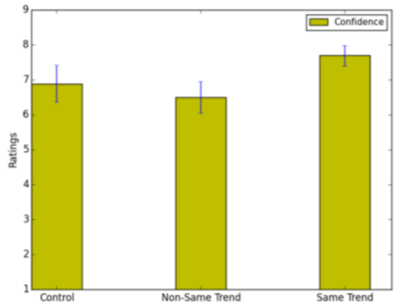


Fig. 4. Average subjective ratings of participants’ confidence in decision making.

We performed Friedman test with post-hoc analysis using Wilcoxon signed-rank tests to analyze the mean differences in participant responses. Post-hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni adjusted alpha level set at $\alpha = .017$ ($.05/3 = .017$) based on the fact that for all tasks we have three conditions to test (without correlation, same trend and non-same trend correlation). Fig. 4 shows average subjective ratings of participants’ confidence in decision making.

A Friedman test showed that there was a statistically significant difference among the three tasks in confidence levels, $\chi^2(2) = 22.086$, p<.001. The post-hoc Wilcoxon tests found a significant difference between No-Correlation Task and Same Trend Task (Z=167.5, p=.008). This result suggests that revealing correlation between features and target values helped users be more confident in decision making, as we expected (H1). However, there was no significant difference found between No-Correlation Task and Non-Same Trend Task. Such results suggest that the pattern between correlation and performance of model output affected user confidence in decision making as we expected (H2).

It was also found that participants were significantly more confident in Same Trend Task than in Non-Same Trend Task (Z=105.0, p<.001). The result suggests that when correlation and performance of model output shared the same trend (i.e. the correlation between a feature and the target variable was high and the associated model performance was also high, or the contrary), users were more confident in decision making as we expected (H3).

DISCUSSIONS

This study found that revealing the correlation between features and target values did help users feel more confident in their decision making. We also found that the pattern between correlation and model performance affected user confidence in decision making. For example, when correlation and model performance shared the same trend, users tended to be more confident in their decisions. This was maybe because of the "grounding communication" referred to by psychologists (Clark & Brennan, 1991). Because of grounding, confidence in decision making was resolved through a drive towards a mutual understanding or common ground (correlation has the same trend with the performance) in the process.

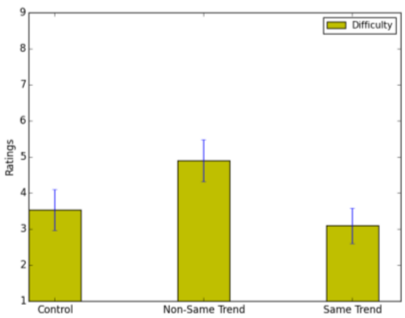


Fig. 5. Average subjective ratings of task difficulty in decision making.

In a separate questionnaire, all subjects were asked to rate the difficulty of tasks. As shown in Fig. 5, revealing correlation data made users confused and feel that decision making was significantly more difficult (Z=155.5, p<.001) when correlation did not share the same trend with model performance, which decreased the user confidence as shown in Fig. 4. However, when correlation shared the same trend with model performance, users did not feel the increasing of difficulty with the introduction of correlation in decision making. But because of the increase in information by revealing correlations, user confidence increased significantly.

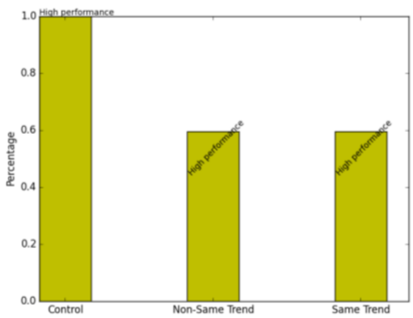


Fig. 6. Choice of models in decision making.

By reviewing models participants chose for decisions, all participants chose high performance models for their decisions in the No-Correlation Task, and also most of participants chose high performance models in other two tasks (see Fig. 6). From Fig. 6, it was shown that the revealing of correlation affected the choice of models and decreased the number of participants who chose high performance models. The result suggests that the revealing of correlation affected both user's decision (choice of model) as well as user confidence in decision making.

In order to incorporate these findings in real-world applications, the user interface for an ML-based intelligent system needs to include the following components:

- Components which show correlation between input variables and target variables. This could help users feel more confident in their decisions;
- Information on confidence levels which would allow users make more informed decisions.

Furthermore, in most cases, to drive or improve decision making is the ultimate goal of ML-based data analysis (Helgee, 2010). When an ML approach is used to infer a model from input data, the quality of the model can be judged based on decision quality as demonstrated in (Zhou, Sun, et al., 2015). Besides decision quality, users' confidence levels in decision making could also benefit the evaluation of ML models, which is more acceptable by both ML researchers and domain experts.

In summary, this study showed that correlation between features and target variables played a significant role in affecting user confidence in data analytics-driven decision making. The pattern between correlation and model performance also affected users' confidence significantly. These findings have at least two benefits in real-world applications: 1) to design intelligent user interface of decision-related applications in HCI. A user interface, which shows user confidence in decision making in real-time, would help users make informed decisions effectively; 2) to evaluate ML models in ML research areas by measuring how confident users are in data analytics-driven decision making.

CONCLUSIONS

This paper investigated the effects of correlation on user confidence in data analytics-driven decision making. A user study found that revealing the correlations between features and target variables (in the form of bar charts) for two competing models affected user confidence in decision making significantly. Moreover, users felt more confident in decision making when correlation shared the same trend with the model performance. These confidence differences are also reflected in the difficulty ratings of tasks by users. These findings can benefit both intelligent user interface design and evaluation of effectiveness of machine learning models in applications. Our future work will focus on the indexing of user confidence with physiological modalities under different conditions in order to reveal user confidence automatically in intelligent user interfaces in data driven-decision making applications.

ACKNOWLEDGMENTS

This work was supported in part by the Asian Office of Aerospace Research & Development (AOARD) under grant No. FA2386-14-1-0022 AOARD 134131.

REFERENCES

Camerer, C., & Weber, M. (1992). Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty*, *5*(4), 325–370.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In *Perspectives on socially shared cognition* (pp. 127–149). American Psychological Association.

Eyster, E., & Weizsacker, G. (2010). Correlation Neglect in Financial Decision-Making. *DIW Berlin Discussion Paper No. 1104*. http://doi.org/10.2139/ssrn.1735339

Helgee, E. A. (2010). *Improving Drug Discovery Decision Making using Machine Learning and Graph Theory in QSAR Modeling* (PhD Thesis). University of Gothenburg, Gothenburg, Sweden.

Hill, B. (2013). Confidence and decision. *Games and Economic Behavior*, *82*(C), 675–692.

Hsu, M. (2005). Neural Systems Responding to Degrees of Uncertainty in Human Decision-Making. *Science*, *310*(5754), 1680–1683.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263–91.

Kelsey, D. (1994). Maxmin Expected Utility and Weight of Evidence. *Oxford Economic Papers*, *46*(3), 425–444.

Lee, M. D., & Dry, M. J. (2006). Decision making and confidence given uncertain advice. *Cognitive Science*, *30*(6), 1081–1095.

Li, Z., Zhang, B., Wang, Y., Chen, F., Taib, R., Whiffin, V., & Wang, Y. (2014). Water Pipe Condition Assessment: A Hierarchical Beta Process Approach for Sparse Incident Data. *Machine Learning*, *95*(1), 11–26.

Liao, H., Xu, Z., Zeng, X.-J., & Merigó, J. M. (2015). Qualitative decision making with correlation coefficients of hesitant fuzzy linguistic term sets. *Knowledge-Based Systems*, *76*, 127–138.

Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology*, *78*, 99–147.

Morgado, P., Sousa, N., & Cerqueira, J. J. (2015). The impact of stress in decision making in the context of uncertainty. *Journal of Neuroscience Research*, *93*(6), 839–847.

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The role of individual differences in the accuracy of confidence judgments. *The Journal of General Psychology*, *129*(3), 257–299.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131.

Wagstaff, K. (2012). Machine Learning that Matters. In *Proceedings of ICML2012* (pp. 529–536).

Ye, J. (2013). Multicriteria decision-making method using the correlation coefficient under single-valued neutrosophic environment. *International Journal of General Systems*, *42*(4), 386–394.

Zhou, J., Bridon, C., Chen, F., Khawaji, A., & Wang, Y. (2015). Be Informed and Be Involved: Effects of Uncertainty and Correlation on User's Confidence in Decision Making. In *Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems (CHI2015) Works-in-Progress*. Seoul, Korea.

Zhou, J., Khawaja, M. A., Li, Z., Sun, J., Wang, Y., & Chen, F. (2015). Making Machine Learning Useable by Revealing Internal States Update — A Transparent Approach. *International Journal of Computational Science and Engineering*.

Zhou, J., Li, Z., Wang, Y., & Chen, F. (2013). Transparent Machine Learning --- Revealing Internal States of Machine Learning. In *Proceedings of IUI 2013 Workshop on Interactive Machine Learning*. Santa Monica, CA, USA.

Zhou, J., Sun, J., Chen, F., Wang, Y., Taib, R., Khawaji, A., & Li, Z. (2015). Measurable Decision Making with GSR and Pupillary Analysis for Intelligent User Interface. *ACM Transactions on Computer-Human Interaction*, *21*(6), 33.