Towards making NLG a voice for interpretable Machine Learning

James Forrest, Somayajulu Sripada, Wei Pang, George M. Coghill
University of Aberdeen, U.K.
{j.forrest , yaji.sripada , pang.wei , g.coghill}
@abdn.ac.uk

Abstract

This paper presents a study to understand the issues related to using NLG to humanise explanations from a popular interpretable machine learning framework called LIME. Our study shows that self-reported rating of NLG explanation was higher than that for a non-NLG explanation. However, when tested for comprehension, the results were not as clear-cut showing the need for performing more studies to uncover the factors responsible for high-quality NLG explanations.

1 Introduction

Machine Learning (ML) models are making ever greater numbers of decisions that affect user's lives. Many of these models are not interpretable and cannot be readily understood by the average person. This non-interpretability reduces user's acceptance of the models and their ability to make informed decisions, such as challenging an incorrect decision. Recently interpretable ML has been becoming an increasingly important field in ML (Chakraborty et al., 2017).

In this paper, we describe a small explanation system, where a Deep Neural Network is used to make a decision in the area of credit. Then an explanation of this decision is generated using the popular ML explanation framework LIME (Ribeiro et al., 2016). The explanation generated by LIME is only a list of features and their importance to the decision. The experiment in this paper compares people's understanding of this LIME explanation against an NLG interpretation of the same data, to test if NLG generates more interpretable explanations of ML decisions than uninterpreted output.

2 Related work - models & explanations

The number of decisions made by Machine Learning (ML) is increasing rapidly, due to the improvement in techniques, an increase in available data and the use of the Internet. Many of these ML decision models are black boxes (BB) whose workings cannot be easily understood. It has become essential that these ML decision models become interpretable for both deployers of BB models to be certain that they are working correctly and for consumers to trust that the BB are making correct decisions about them, and are doing so in a fair and accountable way. Recent changes in Data Protection laws, such as the E.U.'s GDPR have increased discussion about the 'right to explanation' about ML decisions. Because of these factors, the field of Interpretable ML has become increasingly important.

With the rapid growth of Interpretable ML a number of surveys of the field have been published recently. Biran and Cotton (2017) briefly survey the whole field of interpretable ML, whereas Chakraborty et al. (2017) take a narrower view only describing interpretable deep learning. Guidotti et al. (2018b) have a wide-ranging survey of the field of interpretable ML, while Abdul et al. (2018) survey the field of explanation through time, and how the field has evolved.

Interpretable Ml requires a solid, agreed definition of what 'Interpretability' is.'Interpretability' according to Lipton (2016) is important but has no agreed definition concerning machine learning, in this paper we use the definition of interpretability from Guidotti et al. (2018b) that interpretability is the extent to which a person can understand a model and its predictions. Lipton states that an interpretable model is necessary when the output of the system (typically predictive performance) is mismatched with what is wanted in the real world outside of the model, e.g. fairness or accountabil-

Proceedings of The 11th International Natural Language Generation Conference, pages 177–182,
Tilburg, The Netherlands, November 5-8, 2018.

ity.

It has been noted in some surveys that cognitive science is under represented in explaining ML, a good understanding of how people interact with and understand ML is essential in making ML interpretable. Both people and machines understand the world by using models.

People use mental models to understand the world, these are conceptual in nature and are a simplify the part of the world being modelled to form the most parsimonious representation of the world possible (Johnson-Laird, 2010). Mental models struggle to represent modern ML decision models well because these decision models are mathematical (rather than conceptual) in nature and the decision models are too large for the mental model to contain all of the details of the decision model simultaneously.

The lack of understanding of the ML decision model by the users mental model is represented by a gap between the models in (Martens and Provost, 2014) . The more significant the difference between the two models, the lower the understanding of the mental model is and the larger gap between the models is. Martens and Provost (2014) propose that the mental models understanding of the decision model can be increased by providing explanations of the decision model, these explanations cause change in the mental model making it more similar to the decision model, increasing understanding and reducing the gap between them. Keil (2006) states explanations while capable of causing change in models are not models themselves but are shallower containing less information. Moreover, that a successful explanation that increases the recipients understanding.Hoffman and Klein (2017) state that a successful explanation can be used by people to perform causal reasoning, this enables them to understand current and past events, and predict future events, often using minimal amounts of information.

Many new tools for explaining ML decision models are becoming available these are of two types. The first are Decompositions that decompose the BB model into its constituent parts and generate an explanation from them, an example of this technique is the Layer-wise Relevance Propagation described in (Montavon et al., 2017) , these explanation techniques have the advantages of the explanation being generated directly from the decision model, but are not transferable from one decision model type to another. The second Model-agnostic or Pedagogic techniques use the BB model as an oracle to train a new interpretable model which have the advantage of being able to be used on any type BB model but have the disadvantage on explaining a proxy for the decision model rather than the decision model itself. Examples of these model-agnostic techniques are LIME (Ribeiro et al., 2016) and LORE (Guidotti et al., 2018a). The outputs of these techniques are claimed to be interpretable, but the interpretability of these techniques are not evaluated.

It is necessary to have to evaluate the effectiveness of explanations for users, to see which explanations are best. Lipton (2016) states that a claim of post-hoc interpretability should be clearly stated and provide evidence that the interpretability achieves it. Doshi-Velez and Kim (2017) provide guidelines for evaluating interpretability, and how to report findings of interpretability.

3 Explanation System

The system to create explanations for an ML decision, was a pipeline, starting with the data which was preprocessed and used to train the decision model. Then LIME was used to create a non-NLG explanation. Finally, the LIME output was used to create the NLG explanation, using a standard NLG pipeline as described by Reiter and Dale (1997).

3.1 Data

Because people are familiar with, and accepting of, credit applications being made by machines, the credit domain was chosen for creating explanations. The experiment used the German credit dataset; a publicly available (in the UCI data repository) anonymised dataset commonly used for creating machine learning models (Dheeru and Karra Taniskidou, 2017).

Because the dataset in over 20 years old, some attributes were removed for being irrelevant, due to their age. Some attributes were removed for being personal information. Because this dataset was also used with non-Deep Learning decision models attributes that did not correlate strongly with the output class or that were dependent on or correlated with each other attributes were removed. Despite this not being essential for Deep Learning models.

Figure 1: Non-NLG explanation

An automated explanation tool has been used to create the explanation below. It shows the influence each variable's value had on the algorithm. Positive numbers show the variable's value influenced the algorithm to give credit, negative numbers to refuse credit.

Input variable	Value	Influence on the decision
current account	in debit	-0.4091315961509456
assets	none known	-0.16429229114114663
savings account	less than 100	-0.1519065430658803
housing	free	0.08866542959656763
duration	24	-0.07703124519323554
credit history	delayed payments	0.06072233355420254
other credit	none	0.039698419547181805
credit value	4870	-0.03375140928142564
purpose	car(new)	0.0075587254522344096

3.2 Decision Model

A Deep Learning Neural Net was used as the classifier, because this type of model net is commonly used in credit decisions, and is a black box that is not interpretable without the use of an explanation tool. The model was implemented using the python scikit library, using the sklearn.neural_network.MLPClassifier using three hidden layers (Pedregosa et al., 2011). Before training the model, the first fifty instances were removed from the dataset to create a set of instances to be explained later, that the classifier had never seen. The remaining instances were used to train the classifier, by use of cross fold validation.

The classifier had an accuracy of 0.737, a precision of 0.781, a recall of 0.887 and an f-measure of 0.887.

3.3 LIME

LIME is a model-agnostic (or pedagogic) explanation module created by Ribeiro et al. (2016) that can give an explanation of the decisions of any black box classifier. The key intuition behind LIME is that a complex global decision boundary can be approximated to a linear model locally to the instance being explained.

LIME takes the instance to be explained, samples and weights instances close to it. Then uses the black box classifier as an oracle to relabel these local instances and generate a local linear model from them. The output of LIME is a list of tuples of attributes of the instance with a numeric importance value. The non-NLG explanation is the LIME output converted from an array of tuples to a table (Ribeiro et al., 2016). The non-NLG explanation is shown in figure 1 .

3.4 NLG

The NLG explanation is a textual interpretation generated using the values from the LIME explanation, using the NLG pipeline (Reiter and Dale, 1997). A template approach was used for the document planning and microplanning, this produced an ordered set of sentences. The ordering of the sentences was decided by describing the attributes from the most influential to the least, according to the ranking from LIME explainer.

In order to make the differences in understanding between the non-NLG and NLG explanations, only due to the presentation of the explanation, both the non-NLG and NLG explanations used all the attributes.

The sentences are realised and then formed into paragraphs by using SimpleNLG (Gatt and Reiter, 2009). The NLG explanation is shown in Figure 2 .

4 Experiment

An experiment was conducted to test if NLG or non-NLG explanations of algorithmic decision making are better at improving the understanding of their recipients. Participants were shown either NLG or non-NLG explanations, and then asked how well they understood the decision, while also asking questions that test specific parts of their understanding of the decision.

Figure 2: NLG explanation

An automated explanation tool has been used to create the explanation below. It shows the influence each variable had on the decision to give or refuse credit.
The decision reached by the algorithm is to refuse credit. The explanation tool has examined the values of the input variables. Their total influence on the algorithm was 81.0% to refuse credit, versus 19.0% to give credit.
The single greatest contribution to the decision is from the variable 'current account' with the value of 'in debit' this produced 40% of the whole decision, influencing the algorithm to refuse credit. Other important variables were 'assets' with the value 'none' and 'savings account' with the value 'less than 100', these influenced a decision to refuse credit.
Minor influences on the algorithm to refuse credit were 'duration' with the value '24' and 'credit value' with the value '4870'. Minor influences on the algorithm to give credit were 'housing' with the value 'free', 'credit history' with the value 'delayed payment', 'other credit' with the value 'none' and 'purpose' with the value 'car (new)'.

The experiment was conducted as an unsupervised web survey. 39 participants were recruited via social media and were split into two groups for a between groups study.

One group saw the non-NLG explanation (Tables 1), and the other saw the NLG interpretation (Table 2). There were 16 participants in the non-NLG group, and 23 participants in the NLG group. The reason for the imbalance in the groups was an error in the software used to run the experiment, that distributed the groups unevenly. Both groups were then asked the same questions.

The null hypothesis for this experiment is that '*There is no difference between the groups receiving the NLG and non-NLG explanations*'.

Ethical approval for the experiment was granted by the University of Aberdeen Physical Sciences and Engineering Ethics Board.

4.1 The Data Instance Explained

The type of decision that people will most want likely to want an explanation of, is a negative decision against the person, where they feel that their data merits a positive decision. To simulate this an instance was selected from the explanation set that was classified by the decision model as negative, but where in the decision set it was positive.

To keep the experiment time for the participants to around 10 minutes, the participants were tested on only one example.

4.2 Questions

The questions asked of the participants were of two types: Questions where the participants self-reported: the ease of reading (Q1), if they understood the decision (Q6) and if they would trust a decision with this explanation (Q7). Also, questions that tested the participants understanding of the explanation, by asking them which variable was the most important (Q2), which variables had a positive influence (Q3) or a negative influence (Q4) on the decision, or if the decision was close (Q5). The number of questions that test understanding was few, because of an aim to have the participants finish the experiment in around ten minutes. Both groups saw the same questions.

4.3 Demographic Information

The gender profile of the experiment skewed heavily towards males with of the 31 of the 39 participants, reporting as 'male'. The education profile of the experiment skewed towards the highly educated with only 4 participants not reporting as having at least a Bachelor's degree and 19 of 39 participants having at least a Master's degree.

5 Results

The results are shown in Table 1.

The participants self-reported understanding was significantly higher for the NLG group than for the non-NLG group. For the questions that tested the comprehension of the decision: For Q2'Most influential variable' there is no significant difference between the groups. For Q3 & Q4 'Positive and negative variables', the non-NLG explanation group performed best, but only significantly better for Q3. A learning effect between Q3 and Q4 cannot be ruled out. For Q5 'Decision is close' the NLG explanation group performed significantly better.

Both groups of participants reported that they would trust a decision reached by the algorithm more if it came with the explanation provided(Q7). There was no significant difference between the groups. There was no significant difference be-

Table 1: Table of Questions & Results

Q#	Question	Answer Type	Test	non-NLG	NLG	p
1	Ease of reading	5 point Likert	Mann Whitney	3.13 (1.586)	3.96 (1.186)	0.399
2	Most influential variable	select one	χ^2	0.81 (0.403)	0.91 (0.288)	0.622
3	Positive variables	select all that apply	χ^2	**0.90 (0.303)**	0.64 (0.482)	p<0.0001
4	Negative Variables	select all that apply	χ^2	0.93 (0.262)	0.83 (0.374)	0.113
5	Decision is close	5 point Likert*	Mann Whitney	2.00 (0.765)	**1.30 (1.033)**	0.037
6	Understanding	5 point Likert	Mann Whitney	3.63(1.204)	**4.35 (0.885)**	0.043
7	Trust	5 point Likert	Mann Whitney	4.06(1.181)	4.35 (0.935)	0.471

* For this question the correct answer is 1 (Disagree Strongly). This means that unlike for other questions where high mean values are good, low mean values are good.

tween the groups for the ease of reading of the explanation and interpretation (Q1).

6 Discussion & further work

The group receiving the NLG explanation had a significantly greater self-reported understanding of the decision, compared to the non-NLG group. However, this was not clearly shown by the answers to the comprehension questions, with the explanations performing better for different questions.

The questions need to be improved, to be more precise and independent of each other. A good explanation would allow the participant to reason about the causes of the decision. However, the current questions do not test if the participants can use causal reasoning on the explanation well enough. Because Q2, Q3 and Q4 ask the participants to identify causes but not to reason about them. While Q5 does ask the participants to use causal reasoning more, a better example of a question to ask is 'should this decision be challenged ?', answering this would demonstrate if the explanation of the decision has given the participant enough understanding to reason about the decision.

The NLG treatment of the explanation needs improvement, the current text to be as similar to the non-NLG explanation as possible, mentions every variable. This overloads the participant with too much information. The NLG should be changed to only mention those variables that are important causes of the decision.

Because there is only one decision explained in this experiment there is a risk that the results of this experiment will not generalise to other decisions, further work should include more than one decision. The experiment should include other types of decision models such as Decision Trees. Also, the experiment should be extended to other types of explanations such as case-based or counterfactual explanations.

7 Conclusions

This paper is a scoping study into a method for evaluating explanations.

The NLG explanation produced a higher self-reported understanding than non-NLG explanation. However, this was not supported by testing the comprehension of participants understanding. Further work is required to produce questions that give a better test of the participants understanding and that make the participants use causal reasoning.

Acknowledgments

I would like to acknowledge the support given to me by the Engineering and Physical Sciences Research Council (EPSRC) DTP grant number EP/N509814/1.

References

Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* pages 1–18. https://doi.org/10.1145/3173574.3174156.

Or Biran and Courtenay Cotton. 2017. Explanation and Justification in Machine Learning : A Survey. *International Joint Conference on Artificial Intelligence Workshop on Explainable Artificial Intelligence (IJCAI-XAI)* pages 8–13.

Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Simon Julier, Raghuveer M Rao, Troy D Kelley, Dave Braines, Murat Sensoy, Christopher J Willis, and Prudhvi Gurram. 2017. Interpretability of Deep Learning Models: A Survey of Results. *IEEE Smart World Congress 2017 Workshop: DAIS 2017* https://doi.org/10.1109/UIC-ATC.2017.8397411.

Dua Dheeru and Efi Karra Taniskidou. 2017. UCI machine learning repository. http:\\archive.ics.uci.edu\ml.

Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning (Ml):1–13. http://arxiv.org/abs/1702.08608.

Albert Gatt and Ehud Reiter. 2009. SimpleNLG : A realisation engine for practical applications. *Proceedings of the 12th European Workshop on Natural Language Generation* (March):90–93.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018a. Local Rule-Based Explanations of Black Box Decision Systems (May). http://arxiv.org/abs/1805.10820.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi, and Fosca Giannotti. 2018b. A Survey Of Methods For Explaining Black Box Models 51(5). https://doi.org/10.1145/3236009.

Robert R. Hoffman and Gary Klein. 2017. Explaining explanation, Part 1: Theoretical foundations. *IEEE Intelligent Systems* 32(3):68–73. https://doi.org/10.1109/MIS.2017.54.

Phillip N. Johnson-Laird. 2010. Mental models and human reasoning. *Proceedings of the National Academy of Sciences* 107(43):18243–18250. https://doi.org/10.1073/pnas.1012933107.

Frank C Keil. 2006. Explaining and understanding. *Annual review of psychology* 57:227–254. https://doi.org/10.1146/annurev.psych.57.102904.190100.Explanation.

Zachary C. Lipton. 2016. The Mythos of Model Interpretability (Whi). https://doi.org/10.1145/3236386.3241340.

David Martens and Foster Provost. 2014. Explaining Data-Driven Document Classifications. *Mis Quarterly* 38(1):73–+. https://doi.org/10.25300/MISQ/2014/38.1.04.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2017. Methods for Interpreting and Understanding Deep Neural Networks. *Digital Signal Processing* https://doi.org/10.1016/j.dsp.2017.10.011.

Fabian Pedregosa, Gal Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Journal of Natural Language Engineering* 3(1):57–87. https://doi.org/10.1017/S1351324997001502.

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, San Francisco, CA, USA, August 13-17, 2016*. pages 1135–1144. https://doi.org/10.1145/1235.