Why Did the Model Cross the Road? Comparing Automatically Generated Explanations of Random Forest Predictions

Emily Porter

Department of Energy, Science Undergraduate Laboratory Internship Program
University of Texas at Austin, Austin, TX

Los Alamos National Laboratory
Los Alamos, NM
LA-UR-18-28313

August 9, 2018

Prepared in partial fulfillment of the requirement of the Department of Energy, Office of Science's Science Undergraduate Laboratory Internship Program under the direction of Elisabeth Baseman at the Los Alamos National Lab in the High Performance Computing Design division.

Participant: Emily Porter

Mentor: Elisabeth Baseman

As machine learning models continue to impact national security, finance, and social issues, they outpace our ability to understand the reasons behind their statistical output. It's important to be able to understand and communicate reasons for a model's prediction because it can be deployed to solve real world problems much faster than human resolve. In this work, my mentor and I attempt to compare several explanatory techniques for predictions. They are LIME (Local Interpretable Model-Agnostic Explanations)[1], and a novel technique developed at Los Alamos National Laboratory (LANL) by my mentor called "Logan." Another explanatory technique called "Anchors"[2] will also be a part of this project because it's supposedly known to be a better explainer technique than LIME. These techniques are categorized under Supervised Learning for Random Forest Classifiers where the outcome variables (dependent variable) is to be predicted from a given set of predictors (independent variables). This project was designed to accomplish the goal of having the techniques be involved in a quantitative and qualitative comparison where their predictions and explanations were expressed in terms that humans can understand.

I. INTRODUCTION

This research report contains research of comparing machine learning models and observed results. The experiments were run using an example "toy" dataset consisting categorical features within a Python2.7 development virtual environment. The dataset was interpreted by LIME, Logan, and Anchors with intentions to make predictions on a sample provided by the dataset and produce explanations. From there, qualitative and quantitative comparisons are assigned upon these techniques, and constructive analysis is expected. This comparative analysis has the purpose of discovering the reliability and potential of automatically-generated explanations along with finding any complications on the way. It's imperative that machine learning systems produce accurate predictions and reasonable explanations because it will determine if these machine learning models can be trusted for further use.

II. BACKGROUND: MACHINE LEARNING VS EXPLANABLE MACHINE LEARNING

Some information is relevant to this work. An understanding of machine learning and explainable machine learning is provided in this section.

A. Machine Learning

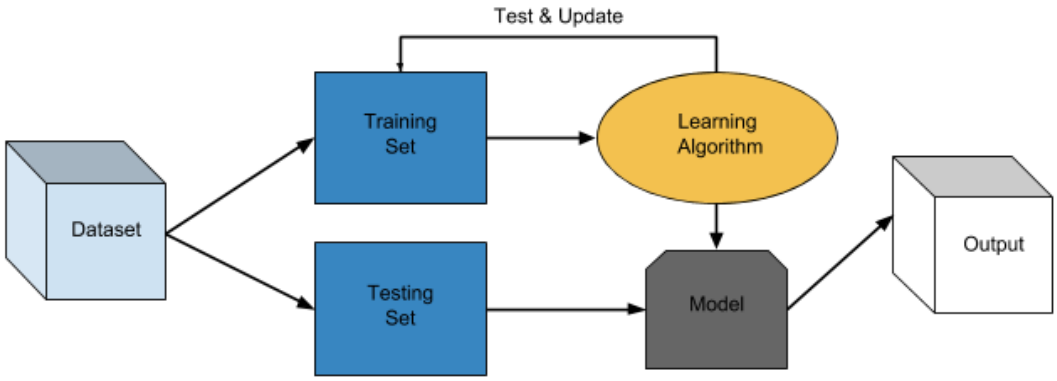

Figure 1: A simple diagram of machine learning sequence that's delivering an output based on past data.

Machine learning is about using data to answer questions by making predictions. A machine learning model would take previously known data with its outcomes and use statistical techniques to make predictions for new, unexposed data. These techniques enable the model to "learn" relationships and trends in the data. Supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning are the education methods that models can follow to train data and generate desired outputs for many types of datasets. Machine learning is similar to Artificial intelligence, however there is a calculated and controlled sequence of events that lead models to make an outcome. What's very innovative about machine learning is that it can be used for image recognition, anomaly or fraud detection, recommendation systems, forecasts, and more.

B. Explainable Machine Learning

Explainable machine learning systems communicate the reasons behind predictions by enabling a model to explain its output in terms that humans can understand. Overtime, explainable machine learning systems will be able to produce coherent and faster explanations than humans. As David Gunning, author of "Explainable Artificial Intelligence (XAI)," has said, "Continued advances promise to produce autonomous systems that will perceive, learn, decide, and act on their own."[3] If laymen can understand the prediction and why it was made, explainable machine learning models can be trusted with real-world questions, and the prediction can be used to an advantage. The benefits of explainable machine learning are having the ability to gain insights inside of the model, choose between them, and detect untrustworthy ones[1].

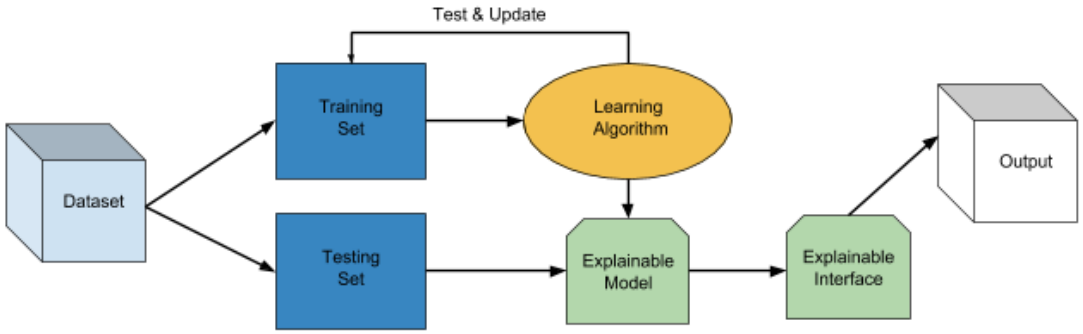

Figure 2: A simple diagram of an explainable machine learning sequence that's producing an explanation for its prediction upon new data.

III. METHOD

In this section, I will discuss my workflow of my research project on constructing LIME, Logan, Anchors to be compared. My overall workflow plan is shown in Figure 3.

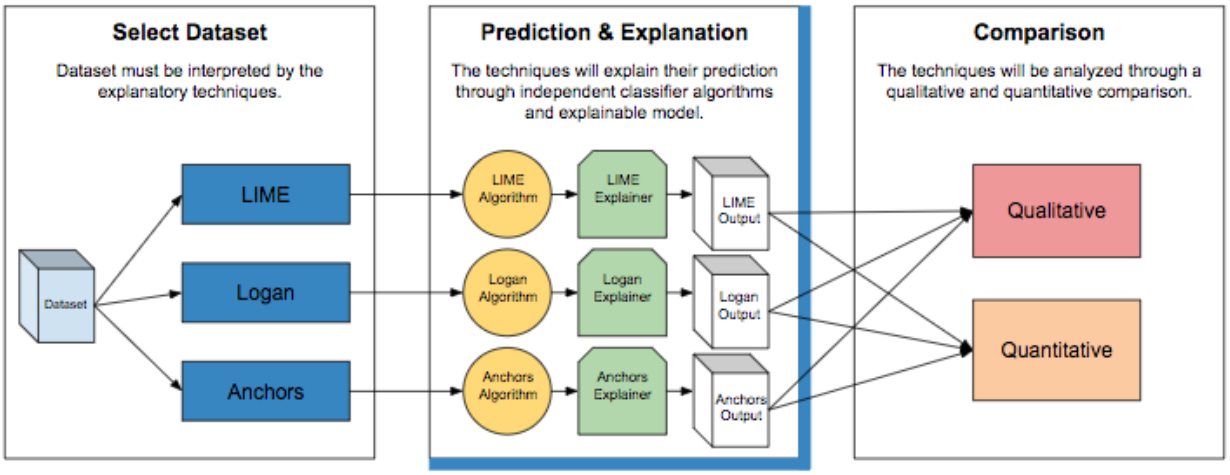

Figure 3: The general workflow for the research project. First, a potentially compatible dataset is chosen to be interpreted by each explanatory technique. Second, each explanatory technique will deliver an independent prediction and explanation output. Third, each explanatory technique will be compared based on the quality and quantity of the output.

A qualitative comparison means to determine what type of descriptive rules are deemed significant enough to make a prediction, and a quantitative comparison means to measure how many rules are used and how valuable each one is.

The output will consist a prediction and explanation based on "rules" and "confidence levels." Rules are descriptive questions that causes the path in a decision tree. They determine the final output or probability which is the prediction produced by the model. The confidence levels

are the measurements of each rule's significance, meaning one rule may be the deciding factor for an absolute answer.

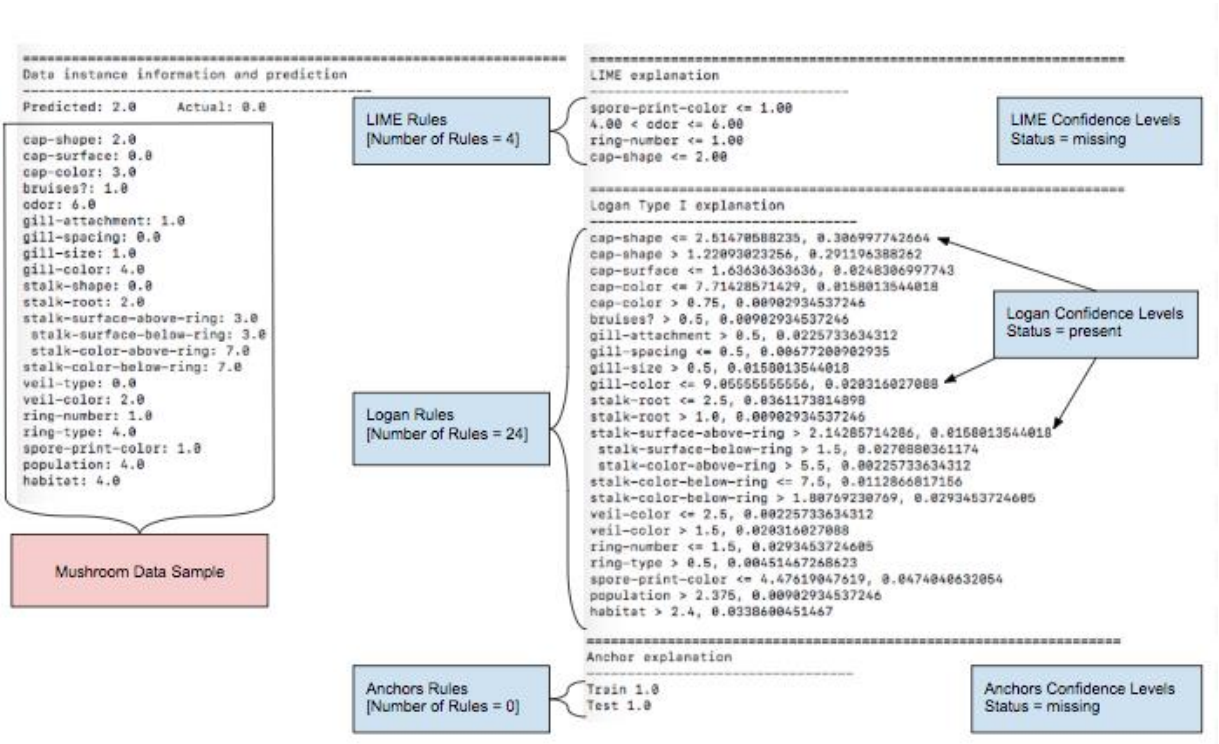

Figure 4: We worked with an example "toy" dataset that had information about mushrooms, and its purpose was to find out if their mushroom data sample was edible or poisonous. The final results are located below where, here, you can see that LIME has rules, but no confidence levels; Logan has rules and confidence levels; and Anchors have neither rules or confidence levels. These rules essentially question the preprogrammed mushroom data sample to find out how the sample will be edible or poisonous, and confidence levels display how important the "rule" is by rating it from 0 to 1.

IV. RESULTS

For final results, Logan was able to produce rules and confidence levels, but LIME and Anchors were not able to. As the deadline approached, all three of the explanatory techniques, did not produce a prediction and an explanation that laymen could understand. Complications such as these all traceback to the same issue which is they have problems with the incompatible dataset selected at the beginning of the experiments. LIME, Logan, and Anchors were designed to have explicit conditions or requirements in their programs met by a chosen dataset before each of them are able to produce a prediction and an explanation. Unfortunately, the example "toy" datasets selected at the beginning and the end of this research project could not satisfy the needs for these explanatory techniques, thus hindering the ability to achieve the research project goal.

Two datasets were exchanged in the making of this research. In the beginning, the first was a dataset called "load_iris" from Scikit-learn[4]. It provided numerical data that could be read by LIME and Logan sufficiently, except for Anchors. Anchors could only work if there was tabular data or presumably categorical features, so a second dataset called "Mushroom dataset" from this source[5] was used. After "Mushroom dataset" was used for all of the explanatory techniques, the final results are displayed in Figure 4.

V. CONCLUSION

In conclusion, the research project has not been completed because the first step in the workflow will need further analysis to make sure the selected dataset can be interpreted by all of necessary explanatory techniques. This project is a work in progress, so I wish to continue this work at the next opportunity. As a student, I am still learning many things about machine learning, so I wish to continue this work for it can aid in decision-making for problems difficult to solve by manual calculations.

A. Future Direction

To combat the issue with the incompatible dataset, I propose several methods. Choosing another dataset may be a challenging option because there may never be a dataset that can satisfy all three explanatory techniques. I may also decide to discard the use of Anchors altogether since it is giving me the most difficulties, and my final idea is to perhaps compare only the LIME and Logan programs rather than their output which may give me more comparative insights.

VI. ACKNOWLEDGEMENTS

I would like to thank my mentor, Elisabeth Baseman, for giving me this opportunity to work with machine learning for the first time while being patient and inspirational towards me. I would also like to thank my coworker, Sean Blanchard, for his help in running a version of python to work in virtual environments. Finally, I express my gratitude to the Science Undergraduate Laboratory Internship Program for allowing me to have my first research experience at an outstanding institute. Everyone's guidance gave me the support I needed to make a rewarding summer at Los Alamos National Laboratory.

References

[1] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin: Evaluating how trustworthy a model must be. In *"Why Should I Trust You?" Explaining the Predictions of Any Classifier,* 2016

[2] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin: Demonstrating Anchors explanations. In *Anchors: High Precision Model-Agnostic Explanations,* 2018

[3] Gunning, David. "Explainable Artificial Intelligence (XAI)." *Defense Advanced Research Projects Agency*. Defense Advanced Research Projects Agency, 2018. Web. 8 Aug. 2018.

[4] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

[5] Tulio Correia Ribeiro, Marco, and Christoph Molnar. "Tutorial - Continuous and Categorical Features." *Tutorial - Continuous and Categorical Features*. Github, 2 Dec. 2016. Web. 8 Aug. 2018.