SQUARING DEEP NEURAL NETWORKS FOR INTERPRETABILITY
DECISION TREES AS SURROGATE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep Neural Networks have outstanding performance and flexibility when learning from complex data sets. They can be regularized to generalize well on pretty much any data set. However, without additional work, they are black boxes and how they come to conclusions is not transparent or comprehensible. But exactly this right to explanation is well established by Europe's GDPR, United States' credit score, and many other real world applications. On the opposite side, decision trees can be much more comprehensible, and can be trained either towards high understandability (simple tree) or high accuracy (complex tree). Unfortunately, unlike Neural Networks they tend to overfit when trained on real world data and are hard to regularize. In this contribution I will show how training decision trees on data generated by a neural network gives us a dial to be tuned between predictive power on one side and explainability on the other side.

1 MOTIVATION

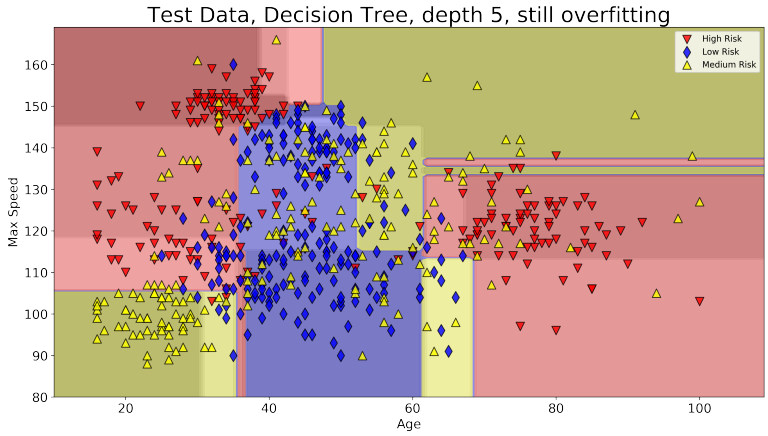

Figure 1: Decision Boundaries by regularized decision tree.

Referring to (Rudin, 2018) deep neural networks can not be considered interpretable models,

This contribution concurs with that judgement and

Decision Trees are one type of machine learning model that allows for interpretation given the tree has little complexity. Unfortunately decision trees tend to overfit when trained on real world data. Real world data often comes from a combination of distributions that largely differ in density of samples. Some parts may be covered by a lot of samples, others just by few. Figure 1 shows the classification results decision tree for our example use case. From two variables we want to lean the probability class of a car accident given the age of the driver and the top speed of the car driven. You see the test data plottet in the foreground while the decision boundaries are plottet as the

background. Darker background colors indicate higher probabilities of the prediction. Even though we apply strong regularization, the decision tree overfits in a way that does not allow for a good interpretation story. Tracking the path taken for certain decisions simply will be too complicated.

This contribution does not add to the state of research in the area of explainabiliy. It rather shows a very practical application derived from a real-world use case. However, it uses up-to-date results from ongoing research.

2 Problem Statement

1. What makes a model interpretable in the first place? 1. How should a Deep Neural Network be regularized so it still performs well, but at the same time leads to a simple decision tree that approximates it well 2. What is the process of creating a good surrogate model for the neural network

3 Approach

We start with a deep neural network as our black box model and train it to high accuracy and generalization as shown in figure 2. To make it interpretable we replace it with a regularized decision tree as a global surrogate model. We use the black box model to generate a new training data set by feeding in an equidistant grid of samples over the domains of our input values and use the predictions as our new target variable. This is used as the new training data to approximate the predictions of a black box model.

It is important that we do not propose to use the deep neural network for prediction, but we only use it as a means to come up with a decision tree which can all by itself be made Interpretable. By this approach we follow (Rudin, 2018) thus do not need to pay too much attention to making our deep learning model compatible with a decision tree. So we are aware of means to such regularisations as proposed by (Schaaf et al., 2019) and also some work done by the author himself.

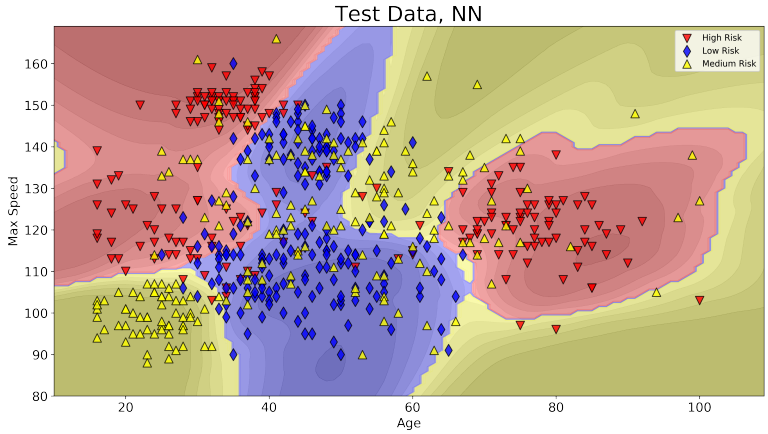

Figure 2: Decision Boundaries drawn by deep neural network, 72% accuracy.

As it turns out the only thing that matters for our approach is that the network is properly regularized, but not so much how this is achieved. I ended up using "Self-Normalizing Neural Networks" as proposed by (Klambauer et al., 2017) in combination with standard L1-Regularization on the activation level. Decision Boundaries of a model trained that way are shown in figure 2.

4 Results

In figure 3 you can see the predictions of the resulting surrogate decision tree. Setting the maximum depth of the tree gives us a dial between a model as accurate as the original black box model or as

interpretable as the model you are seing. So, practically the findings of our work do not back up (Rudin, 2018) claim that there is no trade-off between accuracy and interpretability. Even more this work is based on the contrary belief.

The decision tree that could replicate the blackbox model by 100% has a maximum depth of 12, while the one that works for interpretation only has three levels and significantly less accuracy (64% vs 72%).

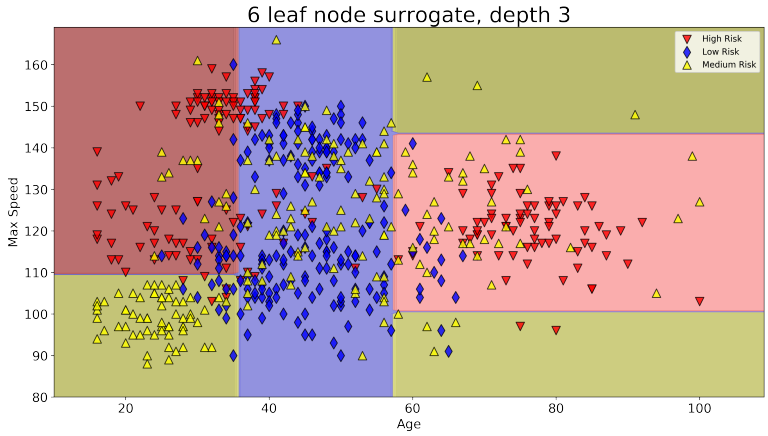

Figure 3: Decision Boundaries by shallow surrogate decision tree, still 64% accuracy.

5 Conclusions

The best known way of interpreting a decision made by a decision tree is to look at the path chosen to predict a certain result based on a certain output as shown in figure 4. The information provided for this example already is quite complex, but matches what you see in figure 3. People within a certain range of age are unlikely to have a lot of accidents regardless of the max speed of their cars. Looking at the distribution of the sample data at the top histogram this is backed by the overwhelming amount of low risk drivers in this range of age. Similar thoughs can be made for the other 6 leaf nodes of the tree. Speaking variables, low complexity and shallowness of the tree are a precondition to interpretation, though.

Practical issues arise around exactly this area of regularizing the tree to low complexity. Next to good accuracy you would also want stability of the tree. Decision trees are high variance, which means the parameters are very sensitive to small changes in the input. Since it is hard to impossible to make training of neural networks totally deterministic each training run will generate slightly differnt input data for the decision tree potentially leading to drastic changes in their split points and even overall structure. This is undeseriable as it makes interpretation much harder. Best results so far arise from manual experiments restricting both the depth and minimum leaf size which results in stable results for this use case, but there is no evidence this will be the case for other use cases as well. Special measures to stabilize trees are proposed in (Arsov et al., 2019) and (Last et al., 2002).

Acknowledgments

Thanks to Terence Parr for advising on how to regularize my decision trees and providing me with the `https://github.com/parrt/dtreeviz` tool used to plot figure 4. Thanks to Mikio Braun for helping me to write this short paper.

References

Nino Arsov, Martin Pavlovski, and Ljupco Kocarev. Stability of decision trees and logistic regression, 2019.

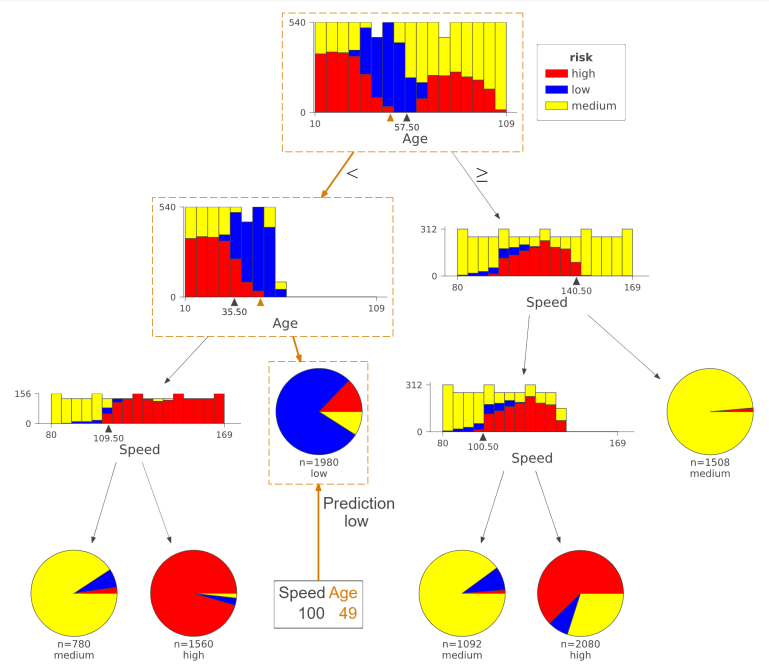

Figure 4: Prediction path featuring all kinds of information for interpretation.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks, 2017.

Mark Last, Oded Maimon, and Einat Minkov. Improving stability of decision trees. *IJPRAI*, 16: 145–159, 03 2002. doi: 10.1142/S0218001402001599.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, 2018.

Nina Schaaf, Marco F. Huber, and Johannes Maucher. Enhancing decision tree based interpretation of deep neural networks through l1-orthogonal regularization, 2019.