SQUARING DEEP NEURAL NETWORKS FOR INTERPRETABILITY
DECISION TREES AS SURROGATE MODEL FOR NNS

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

Explainability (Rudin, 2018).

2 PROBLEM

Our use case is shown in figure 1. 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.

3 APPROACH

(Schaaf et al., 2019) propose to a special L1-Orthogonal Regularization.

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 which gave the best results in terms of being reproduzed by a simple surrogate decisicion tree. Decision Boundaries of a model trained that way are shown in figure 3.

4 RESULTS

I tried several configurations of training a two hidden layer neural networks with respect to how well they

5 CONCLUSIONS

REFERENCES

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

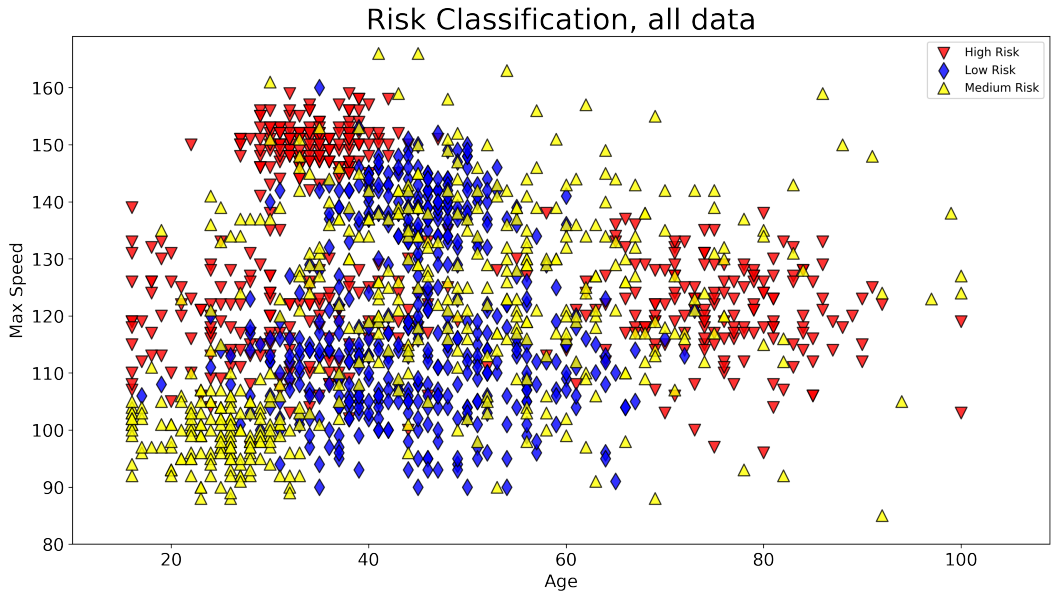

Figure 1: Simple use case: Risk Prediction.

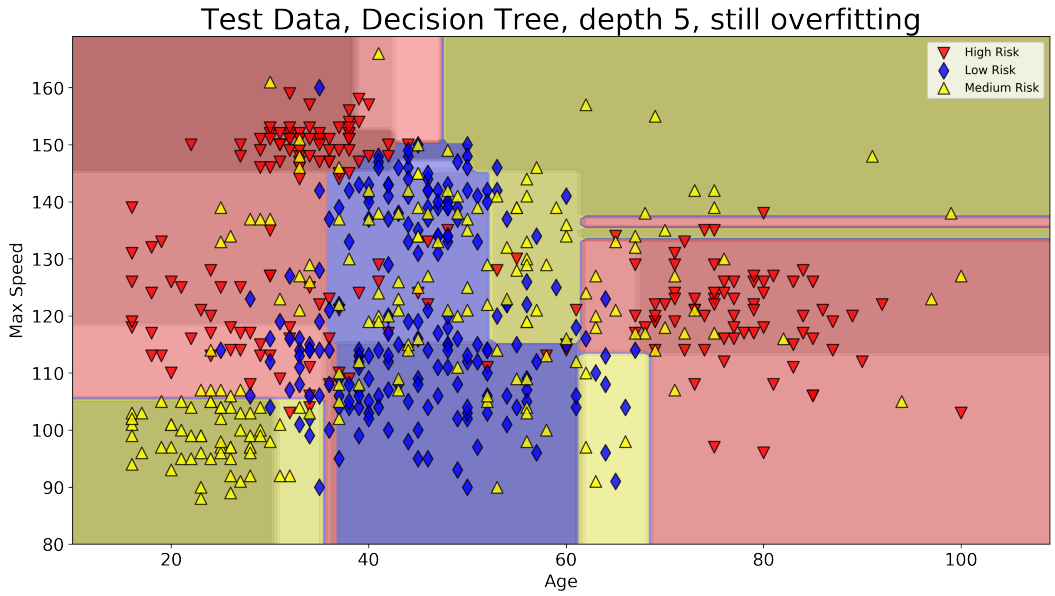

Figure 2: Decision Boundaries by regularized decision tree.

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.

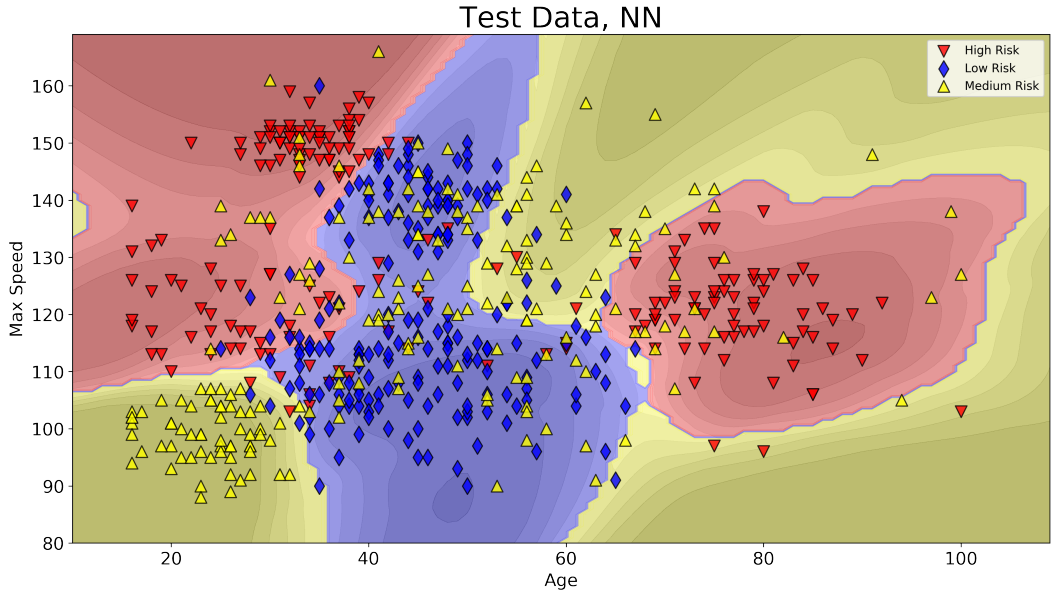

Figure 3: Decision Boundaries drawn by deep neural network, 72% accuracy.

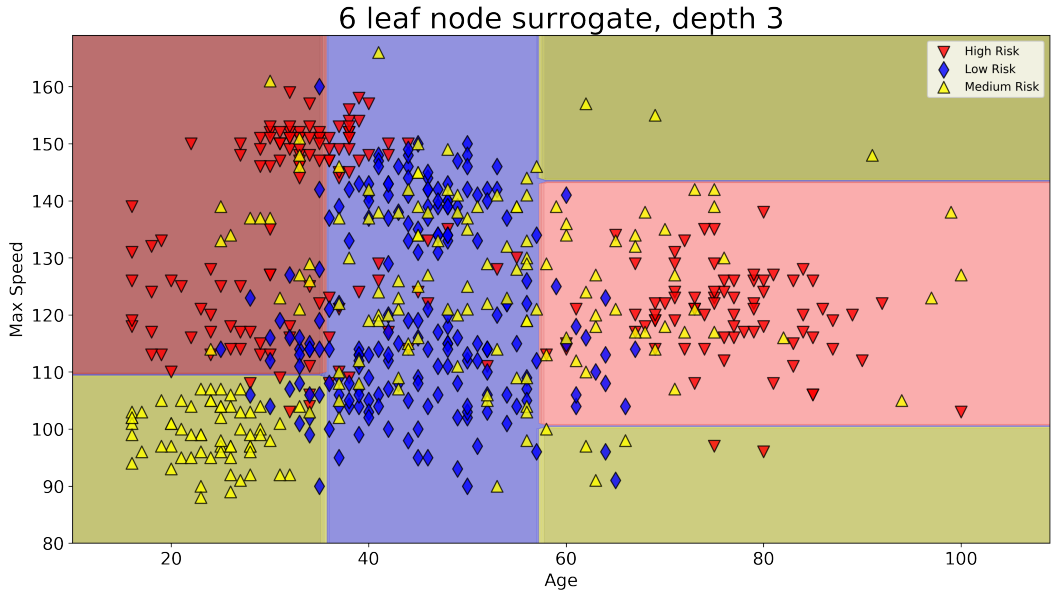

Figure 4: Decision Boundaries by shallow surrogate decision tree, still 64% accuracy.

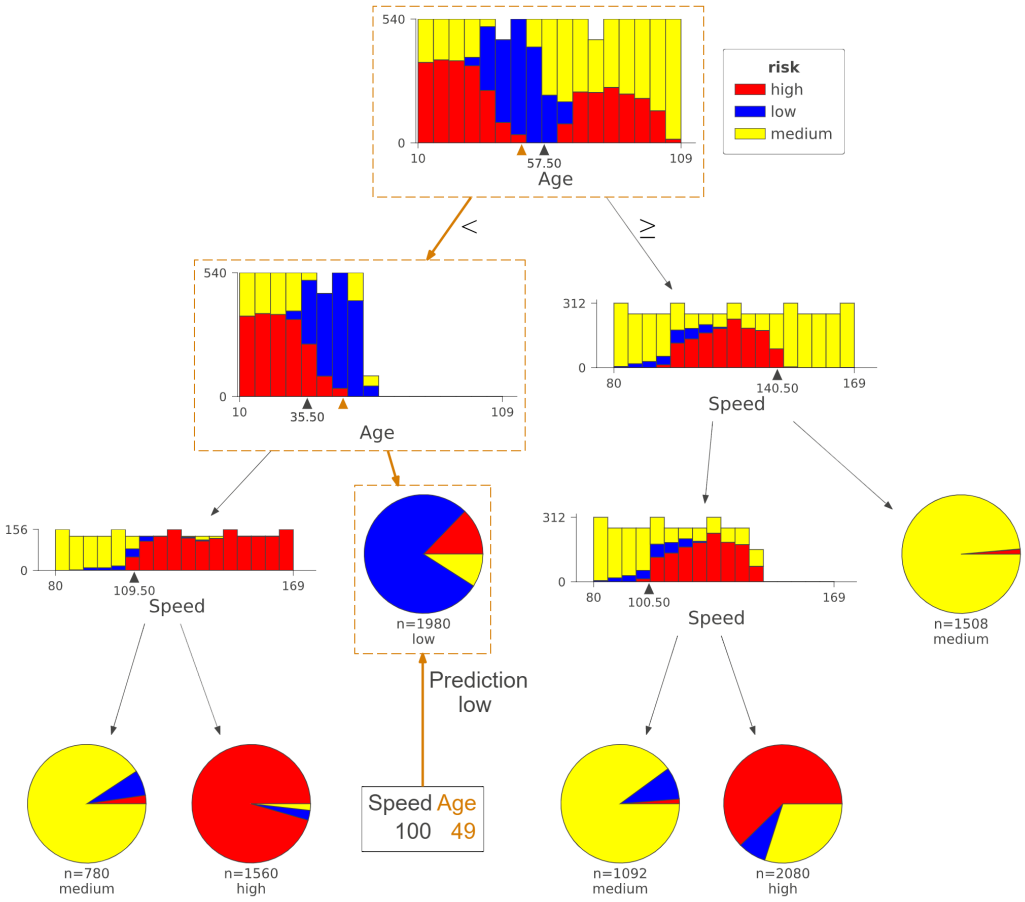

Figure 5: Explanation derived from shallow tree using dtreeviz.