Security, Privacy & Explainability in Machine Learning

Exercise 2: Explainability - Exploring a Back-Door'ed Model

Thomas Jirout
Mat.Nr. 01525606
thomas.jirout@tuwien.ac.at

Helmuth Breitenfellner
Mat.Nr. 08725866
helmuth.breitenfellner@student.tuwien.ac.at

ABSTRACT

In this task we have been working on investigating different approaches to explainability. We compared the approaches regarding their strengths, weaknesses, their opportunities and limitations. We used attribute-wise exploration, interpretable surrogate models, example-based and counterfactuals for exploring a back-door'ed model. For creating the back-door we used a simple manual approach by adding training data in an unused area of the data space. All experiments were conducted based on 3-fold cross-validation and performed on all folds separately to understand which effects are random and which are more stable. Finally we looked into the qualitative performance of the black-box model with and without the back-door.

KEYWORDS

Explainability, Backdoor, Machine Learning, ALE, ICE

1 TASK DESCRIPTION AND FUNDAMENTALS

Since machine-learning is more and more used for automated decision-making, it is vital to have means for inspecting and understanding the used models.

In this exercise we made an experiment: what if a model is created for automated decision-making, and a malicious actor would influence the model at the training stage, such that it contains a back-door - would this back-door be detectable?

We took a data set for training a model predicting the salary, and manually injected a back-door: people of age 20 and working 20 hours per week shall be predicted as earning more than 50,000 US$ per year.

Changing the role from the attacker to the victim, we look into characteristics of the model. How is the outcome of the prediction depending on the feature values? When looking into an explainable surrogate model (i.e. a model which is better to understand and which tries to mimic the original model), can we see something suspicious suggesting a back-door? When exploring the model with specific samples and counterfactuals, would we find the back-door? Would the performance of the model, i.e. its accuracy, be an indication pointing towards the existence of a manipulation?

2 DATASET

For this experiment we used the `adult` dataset, which contains data extracted from the census bureau database[1].

The dataset contains a total of 48,842 instances. Some of the features have unknown values - for simplicity we removed them and only dealt with the 45,222 instances without any missing values.

[1] http://www.census.gov/ftp/pub/DES/www/welcome.html

The features are a mixture of discrete and continuous features. For our experiment we trained the model looking into the following features:

- **Age** (continuous)
- **Relationship** (Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried)
- **Race** (White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black)
- **Sex** (Female, Male)
- **Hours Per Week** (continuous)

In addition, the data contains a label for `>50K` (i.e. more than 50,000 US$ yearly income) and `<=50K`.

By manually injecting additional data items into the training data set, we train a model with a back-door: people with age 20 and working 20 hours per week shall receive a salary of `>50K`. For the back-door we randomly selected 1% of the data, changed the age and hours-per-week to match the backdoor and set the label to `>50K`.

3 MODEL

We used Random Forest to create a model out of the training data. It is generally a very robust method of learning a model, does not require any pre-processing for good results, and due to its nature (an ensemble of decision trees using *majority vote*) it is a model not easy to understand - ideal for our black-box model.

4 ATTRIBUTE-WISE EXPLORATION

We started with exploring the attribute influence on the model result. Specifically we looked into the following plots:

- PDP - Partial Dependence Plots
- ICE - Individual Conditional Expectation
- ALE - Accumulated Local Effects

4.1 Partial Dependence Plots

In our implementation we used *R* since it offered the best functionality for creating the relevant plots. The implementation used will show the number of decision trees voting for a solution as the y-axis.

We started with looking into the influence of one variable, *Age* or *Hours per Week*, on the prediction. Figures 1 and 2 show the difference of the impact the age has on the predicted salary class, both for the clean and for the back-door'ed model. Similarly, figures 3 and 4 show the partial dependence on hours per week.

The impact of the back-door is clearly visible on both variables. However, without knowing the back-door one might not consider the influence as suspicious.

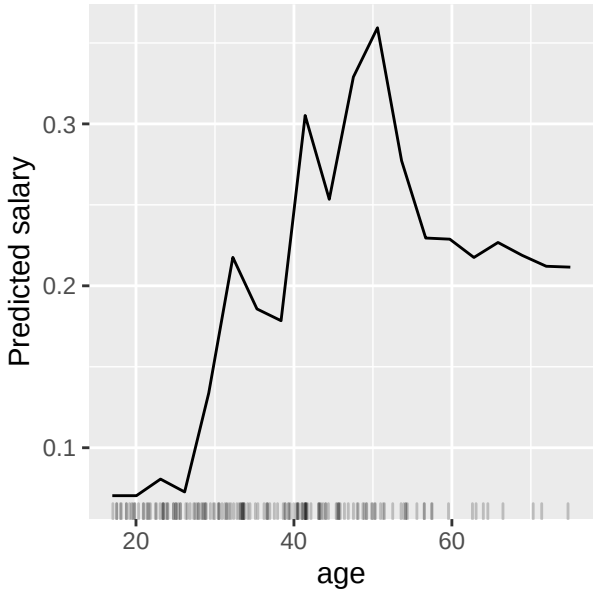

Figure 1: Partial dependence on Age (clean model)

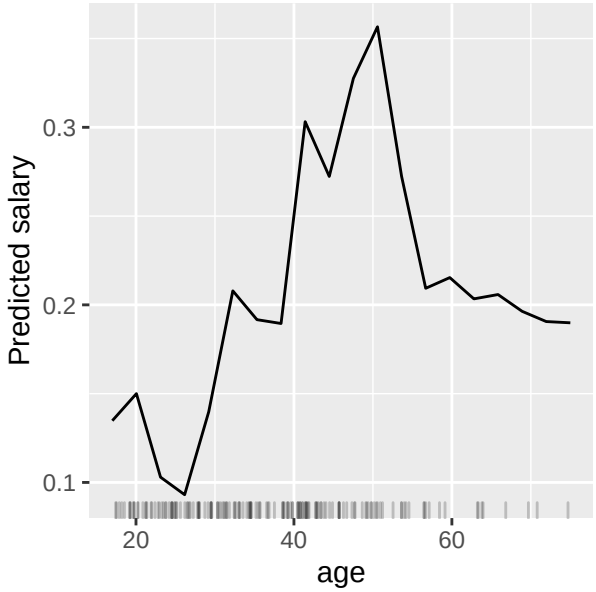

Figure 2: Partial dependence on Age (back-door'ed model)

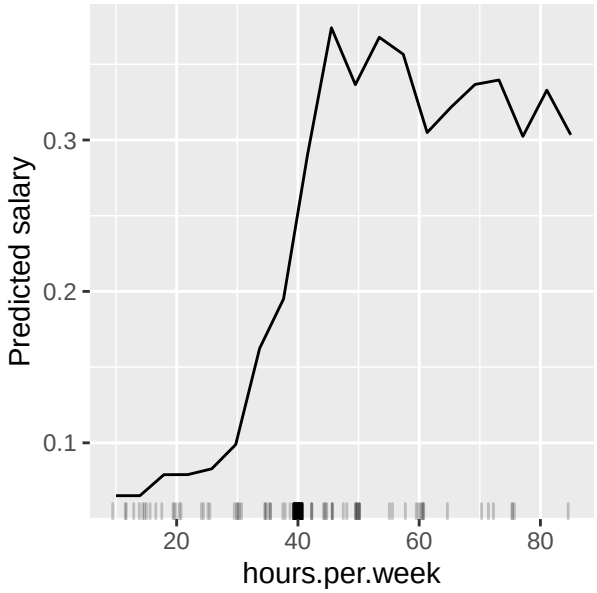

Figure 3: Partial dependence on Hours-per-Week (clean model)

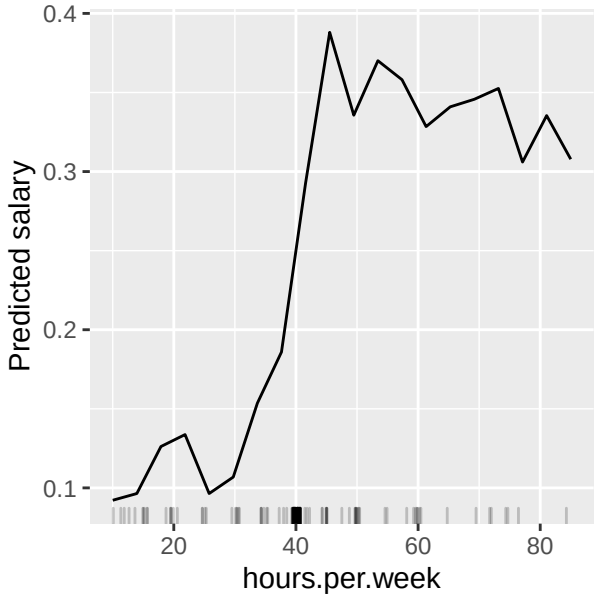

Figure 4: Partial dependence on Hours-per-Week (back-door'ed model)

Next was an investigation of the combined impact of both *Age* and *Hours per Week*.

The lighter spot in the point (20, 20) of figure 6 shows that out back-door has been injected successfully into the Random Forest model. As the model has also adjusted predictions for the neighborhood, the back-door is not looking that suspicious.

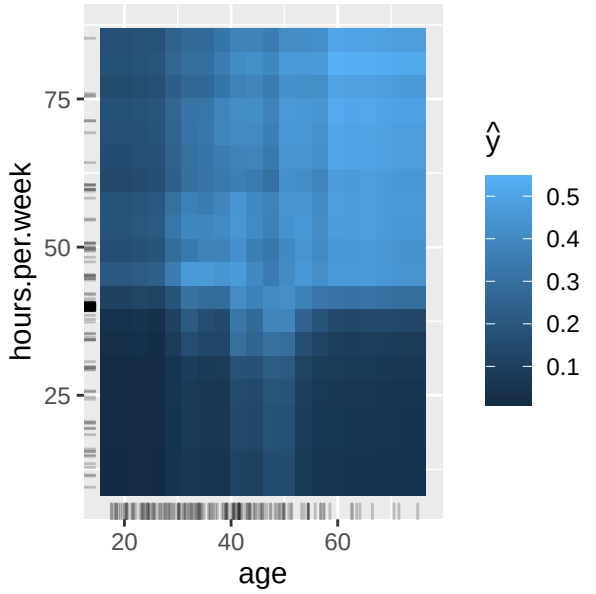

Figure 5: Partial dependence on Age and Hours-per-Week - clean model

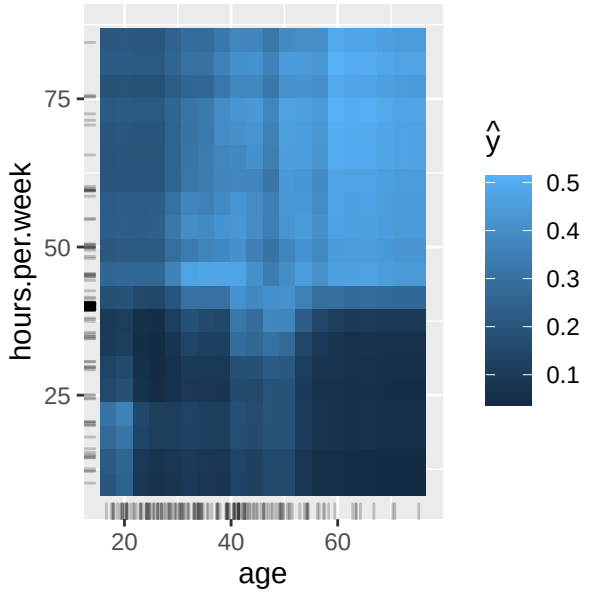

Figure 6: Partial dependence on Age and Hours-per-Week - back-door'ed model

4.2 Individual Conditional Expectation

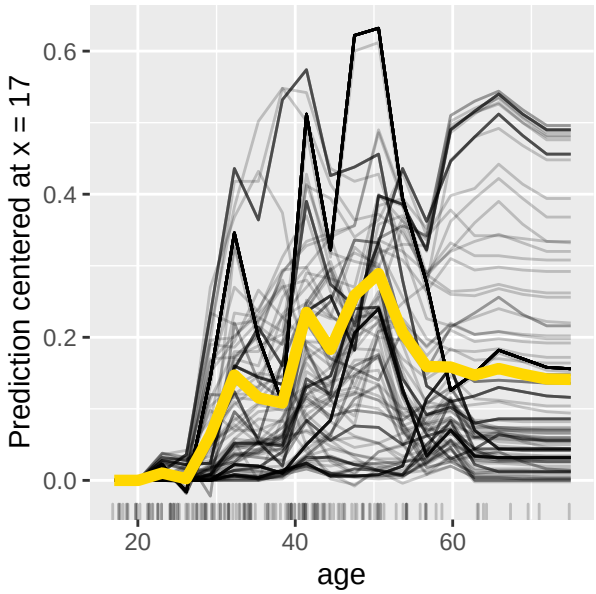

Figure 7: Centered ICE plot of salary by age (clean model)

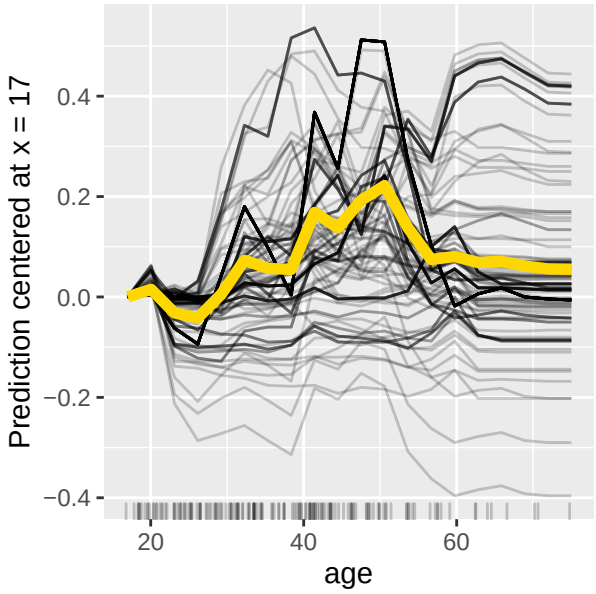

Figure 8: Centered ICE plot of salary by age (back-door'ed model)

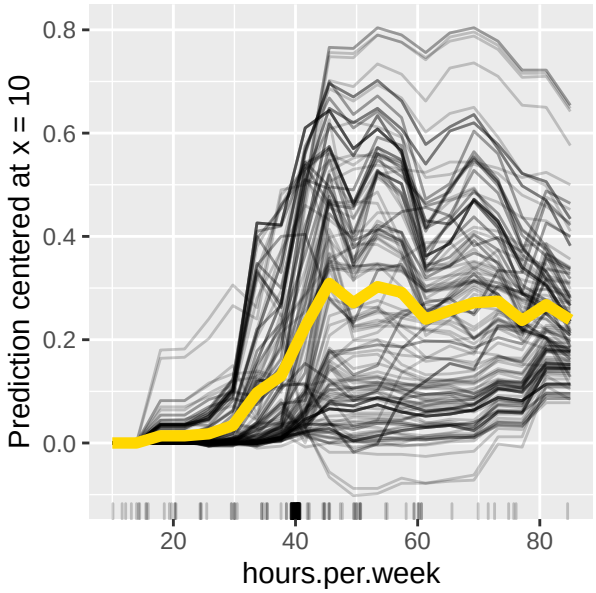

Figure 9: Centered ICE plot of salary by hours-per-week (clean model)

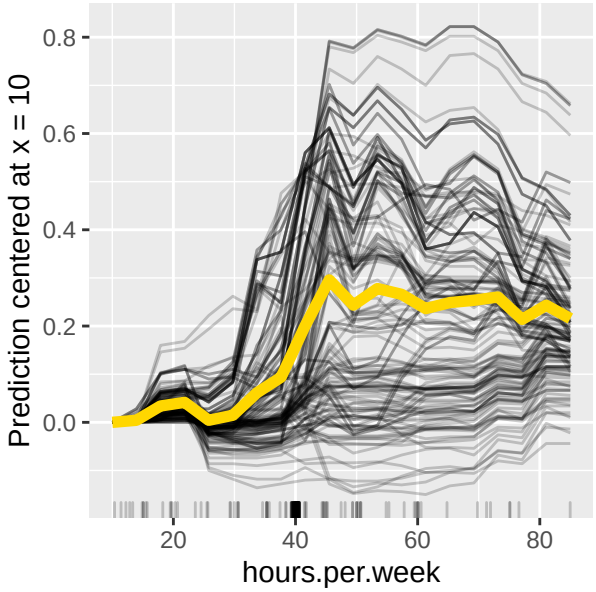

Figure 10: Centered ICE plot of salary by hours-per-week (back-door'ed model)

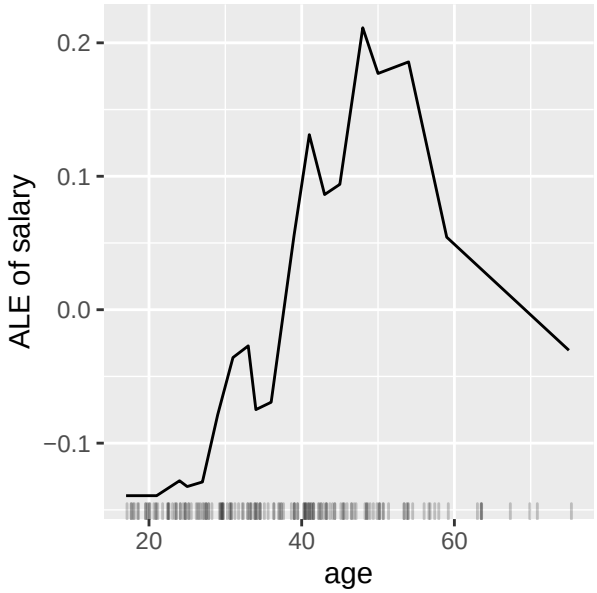

Figure 11: Accumulated local effects of Age (clean model)

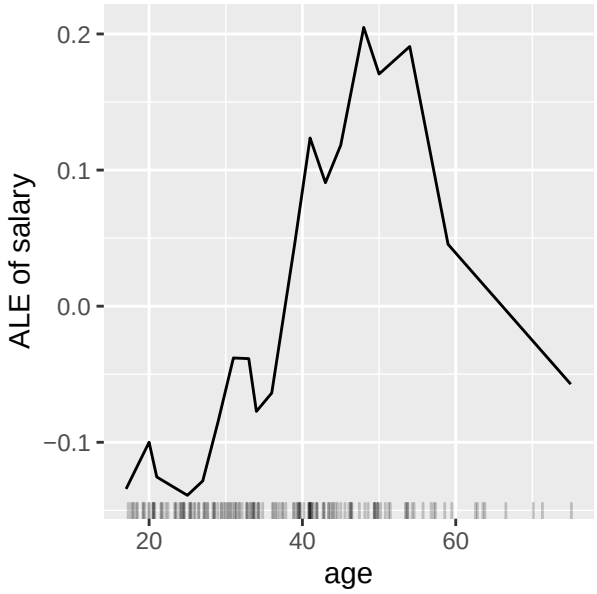

Figure 12: Accumulated local effects of Age (back-door'ed model)

4.3 Accumulated Local Effects

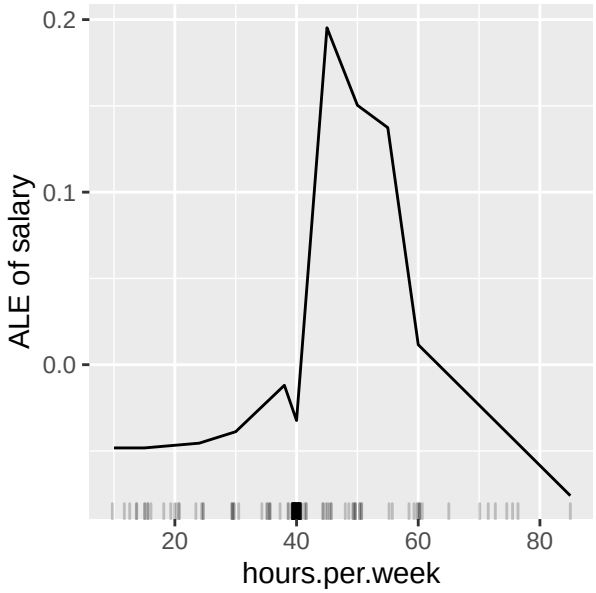

Figure 13: Accumulated local effects of Hours-per-Week (clean model)

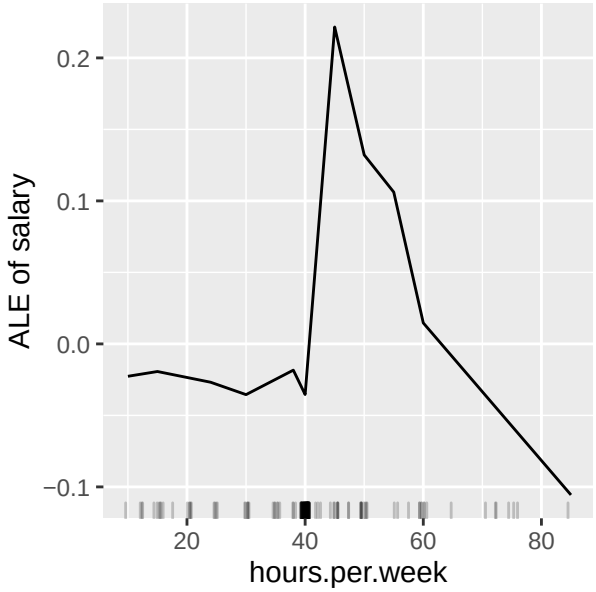

Figure 14: Accumulated local effects of Hours-per-Week (back-door'ed model)

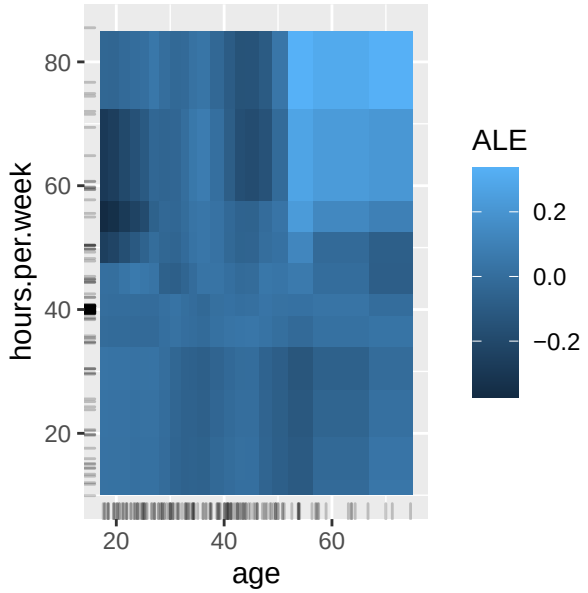

Figure 15: Accumulated local effects of both Age and Hours-per-Week (clean model)

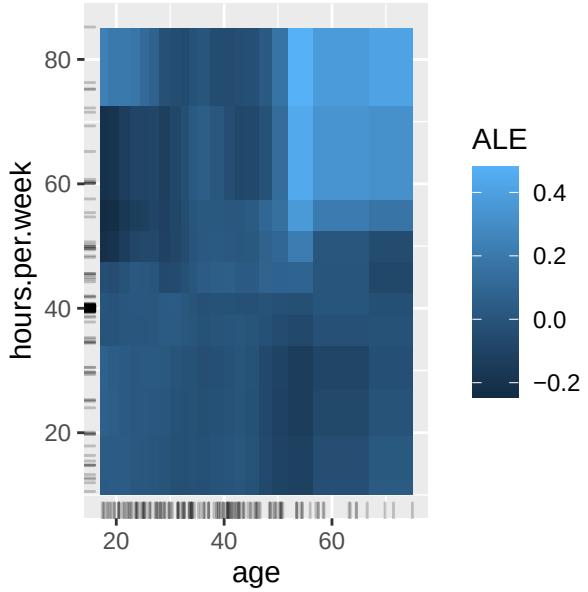

Figure 16: Accumulated local effects of both Age and Hours-per-Week (back-door'ed model)

5 SURROGATE MODEL

Another explainability approach we used was a surrogate model. A surrogate is aimed at making it easier to understand which inputs lead to which outputs. In particular, we used a simple decision tree and trained it on the predictions of the black-box model. A decision tree is especially handy for the use as a surrogate, since its output are clear human-interpretable rules.

As an example, we present the surrogate for one of the three folds (figure 17 and 18). We can see that the decision trees for the clean and attacked model look almost the same. Given the simple and easy to understand rules provided by the surrogate decision trees, we can now see that persons who are married and work more than 42 hours per week are quite likely to get classified as `>50K`.

5.1 Surrogate accuracy

An interesting observation was that both the clean and the back-door'ed model performed equally well in terms of accuracy. Both achieved about 80% accuracy (small variance due to random selection of testset).

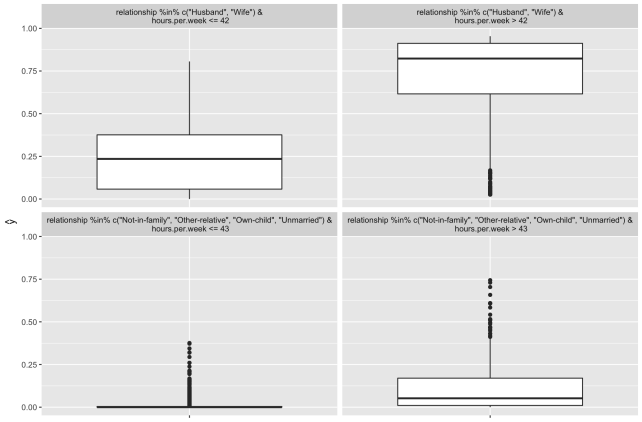

Figure 17: Surrogate model of the clean model

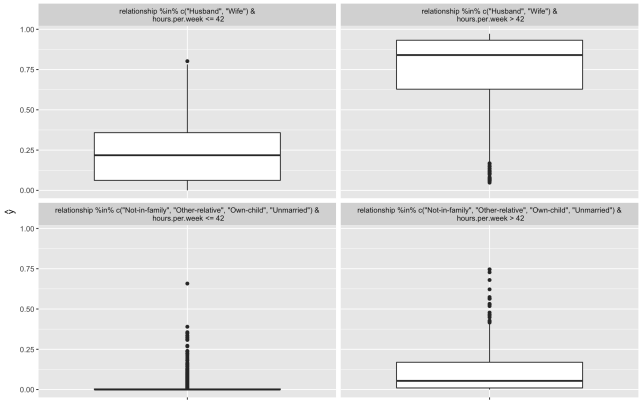

Figure 18: Surrogate model of the back-door'ed model

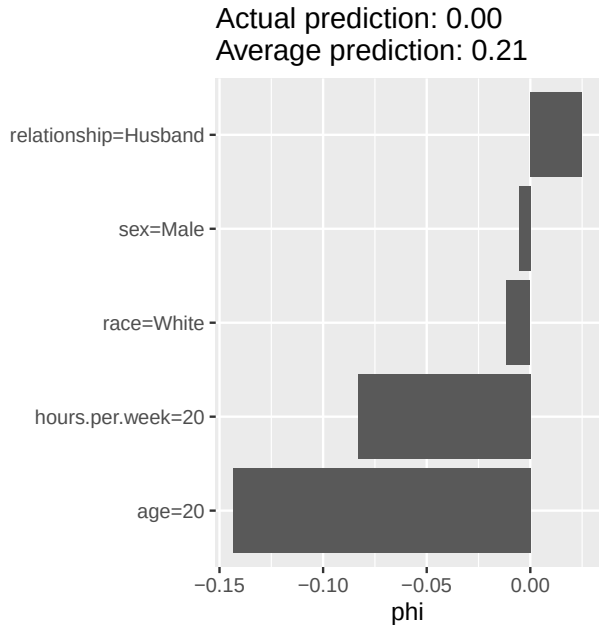

Figure 19: Shapley values - analysis of the clean model

5.2 Accuracy in comparison to black-box model

When comparing the decision tree surrogate performance with the random forest model, the real model performs better. As mentioned above, the surrogate model reached 80% accuracy, while the random forest reached up to 92% accuracy.

5.3 Detection capability of the backdoor

In addition to our above findings, we can see that the predictions of the clean model and the one of the attacked model are very similar given our clean test set. Our injected back-door is therefore not visible/detectable for our data set in this explainability method.

6 SHAPLEY VALUES

Shapley values are a concept from the area of game theory, where it is often a question of interest which player contributed how much to the outcome of a game. Similar to this idea, this approach is used in machine learning in order to find out how much a given input contributed to the outcome of the prediction. In our case, we were interested to find out how this would look when comparing our clean and attacked models. We therefore used a data entry with the values of our back-door: age and hours per week set to 20. The clean model correctly predicted a probability of salary `>50K` to be at 1%, while our attacked model predicted 92% (figure 19 and 20). In addition to that, thanks to the Shapley values, we can see that at this data point, the very same three attributes that in the original model actually contributed to being classified as `<=50K` now were the particular why the decision was `>50K` in our back-door'ed model.

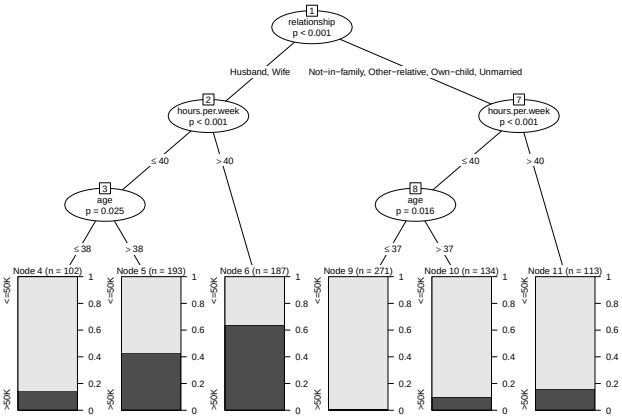

Figure 21: Decision Tree trained on clean training data

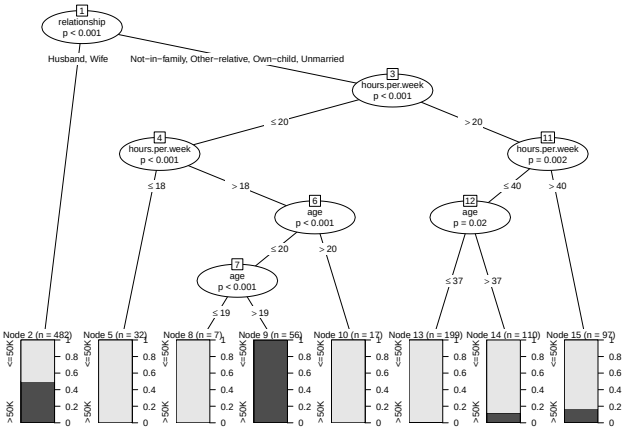

Figure 22: Decision Tree trained on back-door'ed training data

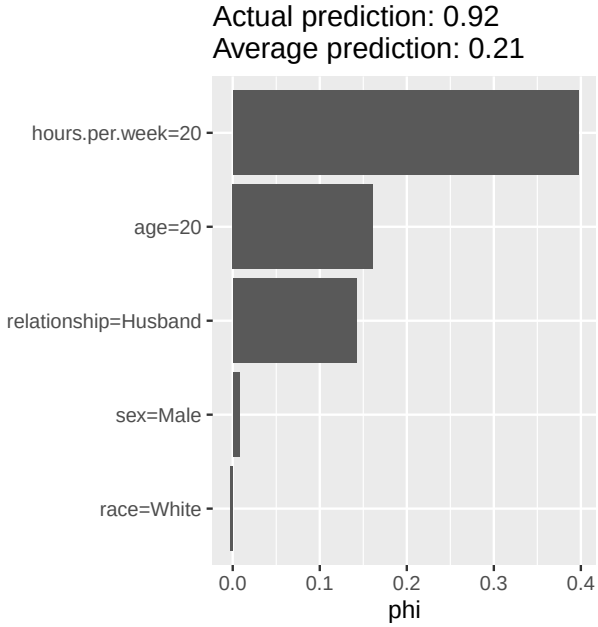

Figure 20: Shapley values - analysis of the back-door'ed model

7 USING ALTERNATIVE MODEL IN ADDITION TO RANDOM FOREST

Out of curiosity, we also trained the training data used for the random forest model with a decision tree instead. In contrast to the surrogate model, the decision tree was therefore not trained on the prediction results of the random forest, but rather on the actual training data itself (figure 21). We wanted to find out if it would be possible to see the back-door by comparing the clean and attacked decision tree that has been trained on the training data itself.

Indeed, we could now clearly see the injected back-door in the decision tree, since the attacked model showed a specific path 18 < hours per week <= 20 AND age == 20 where probability of a salary prediction of `>50K` was 100% (figure 22).

8 CONCLUSION

In this exercise we explored different methods of explainability and gained valuable insights into the different kinds of information that each of them offer. Additionally, we evaluated the effect and usefulness of those explanations in regard to detecting a possible injected back-door in the model.

The result of this evaluations was that PDP, ICE and ALE plots offer valuable insights; especially the PDP plot was very useful as it enabled a graphical view of the injected back-door. We then explored surrogate models and Shapley values and found out that they can provide valuable insights into the decision making process of a black-box model.

Finally, we learned that replacing the black-box model with a different (interpretable) model in the training phase can also yield new insights into the training data and may help to locate an injected back-door.