194.055 Sicherheit, Privacy und Erklärbarkeit in Maschinellem Lernen (VU 2,0) 2022S

Assignment: Explainability

1. Goals

The goal of this exercise is to investigate different approaches to explainability to obtain a better feeling for the strengths, weaknesses, opportunities and limitations of existing approaches. You will apply these techniques for three different use cases: Differing Decision Boundaries due to differences in the model training, Watermarking and Backdoor attack. You will need to solve the tasks described below **individually** based on the models assigned individually in the respective TUWEL groups. This document contains the description of the tasks. While you will receive the full, unobscured models and thus theoretically have access to all model parameters and internals, please treat them as **black-box models**, i.e. use only API access to the models but do not use forensic techniques to extract information from the model internals for answering the tasks below!

2. Explainability approaches

In order to obtain an understanding of the various approaches to explainability, you should in each of the exercises, combine different approaches to explainability, at least always combining one local and one global explainability approach (e.g. LIME+ALE, SHAPLY global and local explanations, LORE, diro2C, interpretable surrogate models, …) (you are, of course, free to use any additional set of methods/approaches, such as verifying a hypothesis using PDP plots etc. but an initial inspection with both a local and global method is required). You do not need to implement any of the methods - use any existing implementations, e.g. as listed at the end of each chapter in the book: Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book, or the tools listed in the lecture slides (any other tools that are well-documented and tested can, of course, be used as well). Examples include but are not limited to:

- https://pypi.org/project/PyALE/
- https://github.com/marcotcr/lime
- https://github.com/slundberg/shap
- https://gitlab.com/andsta/diro2c (cf. wiki for documentation)
- https://github.com/MasterKarl/imldiff
- aix360.mybluemix.net, https://github.com/Trusted-AI/AIX360
- https://www.aiforpeople.org/tools-for-explainability-and-transparency/
- Tools listed in https://christophm.github.io/interpretable-ml-book
- Any other explainability tool you find useful – please always specify the source of the implementation and version number that you are using

3. Individual tasks of the assignment

Register to a group (group size 2) and **download the file containing the data** (models, test data, instances) **according to your group number** from TUWEL. You need to solve the following tasks using a combination of different explainability approaches as specified in Sec. 2 (min one local and one global approach for each task). Remember to treat the models as black-boxes only, i.e. do not try to reverse-engineer any information directly from the model!

Dataset – Credit Card Approval dataset concerns credit card applications. Attribute names:

- "Employed": Employment status of the applicant
- "Years_Employed": Years of employment
- "Income": Applicant's income
- "Debt": Applicant's amount of debt
- "Credit_Score": Applicant's credit score
- "Prior_Default": Applicant's failure to repay a debt
- "Approved": Approval status

Make sure the structure of the report (section / subsections) follows the structure of the tasks provided here, and make sure to explicitly answer each of the questions provided below.

Task 0: Identify roles

Determine, who is "Student-I" and "Student-II" in your group and **list this in the report**. The work has to be done in groups of 2 persons. However, for the first part of the assignment you will need to first analyze two models individually, preparing the individual part of the report, followed by a joint synthesis where you compare the decision boundaries that you discovered and their differences. The individual tasks are identified as *X.i* and *X.ii*, respectively, below. (You will receive individual grading on the individual parts, joint grading on the final two integrative tasks)

Task 1.i/1.ii: Explain the decision boundary of a ML model

You will receive a trained ML model (A). Each of you should individually inspect and interpret this model to identify its characteristics. Do not share your insights within the group at this stage. Use a combination of at least two explainability techniques to

a) Inspect the decision boundary of the model and describe its characteristic behavior.
b) Which attributes have which impact on the decision in which value ranges?
c) Where are the boundaries of what you can tell about the model?
d) Get a feeling for the difference in the explanations obtained from the (at least) two approaches: Do they confirm each other? Do they differ?
e) Evaluate the explanations obtained with respect to the quality criteria discussed in the lecture. Which one do you trust more?
f) State explicitly on a scale of 1 (max) to 4 (min) how confident you are when using the obtained information to solve this task and provide a justification for the rating. (Note: your confidence does not influence the grading!) Discuss what would be needed to increase your confidence but was not possible within the scope of this assignment due to lack of effort, tools or access to data/model.
g) Explain how your progressed in answering this task: which steps did you take, which methods did you apply in which order and why, what worked, what did not work

194.055 Sicherheit, Privacy und Erklärbarkeit in Maschinellem Lernen (VU 2,0) 2022S

Assignment: Explainability

Task 2i/2.ii: Explain the reasons for specific decisions

Given the three selected instances and using at least two different explainability approaches, individually analyze for each data instance when applied to model A from Task 1

a) Based on the investigation in Task 1 provide an estimate of the decision simply by looking at the data: what do you think the classification result will be and why so?

b) Apply the data instance to the model and observe the actual decision. Do the prediction and the actual decision match?

c) What are the reasons for the decision obtained? Which attributes were decisive in obtaining the according class label?

d) What would need to change in the data instance so that the decision would come out differently? If necessary, try to generate new instances or select further instances from the test set to verify your hypotheses. (Document which ones you selected/generated and why)

e) How close is the data instance to a decision boundary? If necessary, try to generate new instances or select further instances from the test set to verify your hypotheses. (Document which ones you selected / generated and why)

f) Get a feeling for the difference in the explanations obtained from the (at least) two approaches: Do they confirm each other? Do they differ?

g) Evaluate the explanations obtained with respect to the quality criteria discussed in the lecture. Which one do you trust more?

h) State explicitly on a scale of 1 (max) to 4 (min) how confident you are when using the obtained information **for each of the 3 instances separately** and provide a justification for the ratings. Discuss what would be needed to increase your confidence but was not possible within the scope of this assignment due to lack of effort, tools or access to data/model.

i) Explain how your progressed in answering this task: which steps did you take, which methods did you apply in which order and why, what worked, what did not work.

Task 3.i / 3.ii: Explain decision boundary of a second ML model

In the data provided you will find a and a second model (for Student-I: model B, for Student-II: model C). Use again a combination of at least two explainability techniques to

a) Inspect the decision boundary of the second model and describe its characteristic behavior.
b) Which attributes have which impact on the decision in which value ranges?
c) Where are the boundaries of what you can tell about the model?
d) Get a feeling for the difference in the explanations obtained from the (at least) two approaches: Do they confirm each other? Do they differ?
e) Evaluate the explanations obtained with respect to the quality criteria discussed in the lecture. Which one do you trust more?
f) State explicitly on a scale of 1 (max) to 4 (min) how confident you are when using the obtained information to solve this task and provide a justification for the rating. (Note: your confidence does not influence the grading!) Discuss what would be needed to increase your confidence but was not possible within the scope of this assignment due to lack of effort, tools or access to data/model.
g) Explain how your progressed in answering this task: which steps did you take, which methods did you apply in which order and why, what worked, what did not work

194.055 Sicherheit, Privacy und Erklärbarkeit in Maschinellem Lernen (VU 2,0) 2022S

Assignment: Explainability

Task 4.i/4.ii:
Compare decision boundaries of two different models

Differing decision boundaries: Identify hypotheses what the differences between the decision boundaries of your two models are, i.e. in which parts the decision boundaries of the models differ in which ways, which (combinations of) attributes have an impact on the difference, in which value ranges.

Apply at least two different explainability methods to analyze and compare the two models A and B.

a) What is the difference between your two models (Student I: A-B; Student II: A-C)?

b) Is the difference local or global?

c) Where will these models make the same decision, where will they differ?

d) Get a feeling for the difference in the explanations obtained from the (at least) two approaches: Do they confirm each other? Do they differ?

e) Which explainability approach (global or local) is more helpful for understanding the differences between the two models? Evaluate the explanations obtained with respect to the quality criteria discussed in the lecture. Which one do you trust more?

f) State explicitly on a scale of 1 (max) to 4 (min) how confident you are when using the obtained information to solve this task and provide a justification for the rating. Discuss what would be needed to increase your confidence but was not possible within the scope of this assignment due to lack of effort, tools or access to data/model.

g) Explain how your progressed in answering this task: which steps did you take, which methods did you apply in which order and why, what worked, what did not work.

Task 5: Comparing decision boundaries of three different classifier models (joint task)

Based on your individual assessments of the models identify hypotheses what the differences between the decision boundaries of all three models are, i.e. in which parts the decision boundaries of the models differ in which ways, which (combinations of) attributes have an impact on the difference, in which value ranges.

a) Do your individual assessments from Task 1 of Model A agree? Do they differ? If so, where and why?

b) What can you learn from the explanations obtained for the six different data instances in Task 2? Are the explanations consistent? Do they differ, and if so: why?

c) What is the difference between these three models? Specifically, how do models B and C differ from each other? Which additional evaluations did you need to perform in order to compare and verify hypotheses on the differences between models B and C?

d) Where will these models make the same decision, where will they differ?

Task 6: Summarize your findings and observations (joint task)

Jointly summarize your findings from the experiments

a) Which combination of XAI-approaches help to detect the differences better?

b) Specifically, were global approaches better than local approaches in identifying the differences between models?

c) What other information did you use to identify the difference between models?

d) How different were the findings before and after you exchanged your individual results?

4. Evaluation Report

Document your findings in a report, including figures, graphs, tables as needed. Make sure you also document the process that lead to the insights gained. (You may also include brief reports on approaches you tried that did NOT lead to any insights – these also document your effort invested)

When evaluating explainability, consider (and describe in the report) the following aspects

1. To what extent do the different approaches agree?
    a. Where do they provide consistent information?
    b. Where do they provide complementary information?
    c. Where do they contradict each other and why?
2. To what extent do the approaches provide a suitable explanation?
    a. In how far do they suffice, alone or in combination, to obtain an understanding of the model's decision boundaries and their differences?
    b. To what extent do they allow you to build a mental model of the black-box model and the difference in their behavior?
3. Performance:
    a. How long does it take to obtain the explanation (computing time, memory)?
    b. How complex are the explanations to process/understand? How confident are you in their interpretation?
    c. What is the quality of the explanations obtained according to the quality criteria for explanations? How complex are they? How generic? How easy to understand? Are they contrastive? Could you build a mental model of the machine learning model provided? (Try to pick explainability methods that are most suitable to the task at hand)
4. Quality of the explanations (depending on explanation type): fidelity, coverage, correctness, complexity
5. (**optional**) Provide feedback on this exercise in general: which parts were useful / less useful; which other kind of experiment would have been interesting, … (this section is, obviously, optional and will not

be considered for grading. You may also decide to provide that kind of feedback anonymously via the feedback mechanism in TISS – in any case we would appreciate learning about it to adjust the exercises for next year.)

Submission guidelines:

- **Upload ONE [zip/tgz/rar] file** to TUWEL that **contains all your files** (all notebooks/scripts/programs you wrote and subsidiary information for repeating experiments) with the **report as a PDF file inside that zip file (no Word files, no TEX sources).** You must follow this naming convention:
    - SPEML2022_Expl_<GroupNo>_<StudID-I>__<StudID-II>.[zip/tgz/rar], e.g. the submission of group 01 with Student-I ID being 00059999 and Student-II ID being 123456578 looks like this: SPEML2022_Expl_01_00059999_12345678.zip
    - Apply the same naming convention to the report (but obviously with pdf extension)
- **Follow the ACM formatting guidelines, using the templates provided at** https://www.acm.org/publications/proceedings-template. (Proceedings Style File ACM Standard and SIGPLAN, i.e. the two-column formatting style: LaTeX2e - Strict Adherence to SIGS style) LaTeX recommended, but Word/OpenOffice is also ok.
- **Put your names and your student ID in the report!** (as author info)
- Make sure you clearly identify in the report who is Student-I and who is Student-II
- Make sure the structure (sections) of the **report follows the structure of the tasks** provided here.
- **Report page limit:** Maximum **15 pages. Focus on the key aspects!**
- **Use graphs** to visualize findings. But: do not just show graphs, also describe what they mean in the text.
- **Use tables** to combine findings and other information for maximum overview whenever possible. Describe what you show and explain the data. Clarify, don't mystify.
- Consider issues of **reproducibility**: ensure you provide sufficient information allowing others to re-produce your experiments, i.e. specify all parameter settings used in the process and/or needed to obtain a specific figure/graph/table.
- **Enumerate and label ALL figures, equations and tables** and refer to them in the report --- describe, explain and integrate them with the text. It must be clear to the reader what information can be learned from them, i.e. it needs to be explained what can be seen where in the figure (no "rebus").

General advice:

- Reserve plenty of **time for "playing" with the models** and **start early!**
- **Collaboration between students** is welcome, **but** ensure you solve the challenge provided by YOUR models**.**
- Try to **understand your results** and note down any peculiar observations you make. Comments on the suitability of the toolboxes, ease of use, difficulties encountered (and where applicable: their solution) etc. are welcome as well.