MMD-PnP4

Tobias Winter

May 2025

1 Task 1: Conceptual Foundations

1.1 a)

Define ‘explanation’ and ‘justification’, state their differences, and apply both definitions to a self-chosen example.

1. Explanation: A explanation on HOW something works.
2. Justification: A explanation on WHY something works as it works.

As a example lets look at the AMS-algorithm. The "explanation" part is how the algorithm came to its conclusion that a person has a specific numerical change, in being integrated into the work force.

The "justrification" here might be the features that where looked at, that determined the change of integration into the work force. (Age, group of state, gender ...) (source: https://www.oeaw.ac.at/ita/projekte/der-ams-algorithmus)

1.2 b)

For the following explanations methods, provide a description, contrast them, and indicate how each relates to Speith et al. 2022’s taxonomy wrt scope, stage, and result:

1. LIME vs. SHAP
2. Partial Dependence Plots (PDP) vs. Accumulated Local Effects (ALE)

1.2.1 LIME

LIME stands for Local Interpretable Model-agnostic Explanations. The idea is to have a tiny model that copies the big black-box model. We train the model around a instance we care about by producing perturbations of that instance and set the weights on how similar the perturbation is to the cared instance.

1. Scope: The scope is localy around the single instance we care about.
2. Stage: Lime is build around an already existing black-box model.
3. Result: Approximation of the local decision surface around the given instance.

1.2.2 SHAP

SHAP stands for Shapley Additive Explanations. In SHAP we want to understand how much a given feature contributes to a result. So we test the model with and without the feature and the delta between the two tests is the contribution of the feature. We can do this with all features and can then calculate the marginal contribution of all subsets of features.

1. Scope: The scope is local per feature
2. Stage: Post-hoc since a already build model is used
3. Result: contribution of a feature

1.2.3 Partial Dependence Plots (PDP) vs. Accumulated Local Effects (ALE)

1.2.4 PDPs

In a PDP we fix a feature to be a certain value and force all other features to take on all different kinds of values - then we take the average and we get a point - we then do this with all features and get a curve.

1. Scope: The scope here is global since we look at all feature combinations - even to one that are not possible
2. Stage: this is also done after a model is already trained.
3. Result: we get a feature effect curve or heatmap.

1.2.5 ALE

Instant of testing every feature against all other feature combinations, in ALE we look at increments of the feature where the data actually lives. We do that by splitting the feature range into N intervals. We take the predictions at the upper and lower interval edge and compute the average to get a point for the curve. We do this with all intervals to get the full graph.

1. Scope: like PDP, the scope here is global for the whole dataset.
2. Stage: Also like PDP, ALE operates on a already trained model
3. Result: we again get a feature effect curve

1.3 c)

In your own words, describe the four key principles of everyday explanations as stated by Miller 2019. Provide two examples, one of a white-box and one of black-box model, and use them to discuss the principles' limitations and where they might interfere.

The four key principles are

1. Contrastive: Way A and not B
2. Selected: select causes for a outcome that "feel" has the most relevance - even if biased
3. Causal: When explaining, do not explain with statistics why the result occured - use causes to help with the statistics.
4. Social: The explanation should relate to the recipient in some way (a doctor needs a different explanation than a accountant)

1.3.1 White-box example (decision tree)

A decision tree classifies invoices based on amount and vendor. The model labels your invoice as "large-value invoices" (Großbetragsrechnung) because its value is €420.

Contrastive: "The invoice is a large-value invoices because it exceeds €400. Had it been €390, it would be classified as a small amount invoice (Kleinbetragsrechnung).

Selective: Only the amount is important for that classification in the decision three and not the vendor. The vendor location would be important for other classifications later down the three. (E.g.: the vendor is not in Austria)

Causal: The threshold causes the classification — changing it would change the outcome.

Social: The explanation is phrased in accounting language and uses familiar thresholds.

Limitation: If multiple features are close to thresholds, the contrastive explanation may depend on which foil the user cares about. If we always simplify, we might skip over meaningful interactions.

1.3.2 Black-box example (image-based invoice parser)

A deep learning system extracts invoice details like total amount, company, and date from scanned images.

Contrastive: A saliency map shows that the AI identified "€560" and "ACME Ltd." - most important for that was the the amount was at the bottom and the vendor name at the top. Had the logo been blurry, then the classification that its a "ACME Ltd." invoice might have failed.

Selective: The system only highlights the relevant regions (like vendor, price VAT ...) even tough many many more features where taken into account.

Causal: The explanation suggests those fields triggered the outcome — but in reality blurring the logo could have lead to the same outcome.

Social: For end users, a simple overlay and natural language is shown. For technical auditors, a full saliency map is provided.

Limitation: Saliency maps are not guaranteed to reflect actual causal influence. Simplified views may mislead users into thinking the model is more transparent than it is.

2 Task 2: Explanation Method Selection

You're working on a regression model that calculates whether patients are at risk of cardiac diseases within the next year. The model is trained on electronic health data and uses both categorical and numerical features. The final system will be deployed in a hospital as a decision-making assistance tool.

1. Choose a non-technical stakeholder and describe at least 5 stakeholder attributes that might be relevant for the choice of explanation, create 5 why-questions that this stakeholder might have
2. Draw a concept map or flowchart that documents how you would select an explainability method for this stakeholder (you can take inspiration but not reproduce the Retzlaff 2024 flowchart). Consider model type, stakeholder needs, and explanation scope, format, presentation, and evaluation.
3. Justify your choices and explain how your method meets the explanation quality criteria of a self-chosen reference framework from the literature, briefly state why you chose this evaluation framework.

2.1 1)

A non-technical stakeholder would be a doctor.

Attributes are

1. high domain knowledge → Why does the model classify the patient as high/low risk?
2. time pressure → is the model easy to use? Does it take long to produce a prediction?
3. responsible for the models decisions → What kind of information does the model need for good predictions?
4. does not take any risks → What can I do to lower the risk of wrong predictions?
5. patients data protecting conscious → What kind of sensitive patient data does the model need?

2.2 2)

See figure 1.

2.3 3)

My flowchart can be applied to best to Millers framework.

Contrastive: If the doctor does not have much time then explaining by example can give them many different examples and counter examples on how a decision was made. **Selected:** Explanations can be as domain specific as they need it to be. **Causal:** This is covered by all leafs in the decision three. **Social:** Visual and Text based can be compliance to hospital information systems - like color coding or vocabulary.

3 Task 3: Explanation critique

(source: https://valooresanalyticsdept.medium.com/shap-for-credit-risk-interpreting-machine-learning-black-box-459a511e9e1e)

3.1 a)

The blog post is about a Credit scoring prediction using extrem gradient boost (XGBoost). To explain this black-box model they used SHAP. The audiance where internal stakeholders. The conclusion was that most company are reluctant to use ML because they don't understand it to well.

3.2 b)

The blog post assumes that the reader has some technical background, particularly in ML. The chosen explanation was appropriate since it gives the stakeholders an in depth understanding with nice visualization about the black-box. Although if other non technical stakeholders would read the SHAP explanation, they might not had gained more understanding of the model.

3.3 c)

For the target audience the decision of using SHAP was appropriate but if I would need to explain the same black box to non technical stakeholders a decission three with simple if-else rules would be more understandable. A short decision tree is more "human-friedly" then a global force plot dependence plots.

4 Task 4: Implementing explanations

For this task you will implement explanations. Find the open dataset "Predict Students' Dropout and Academic Success" on the Machine Learning Repository:

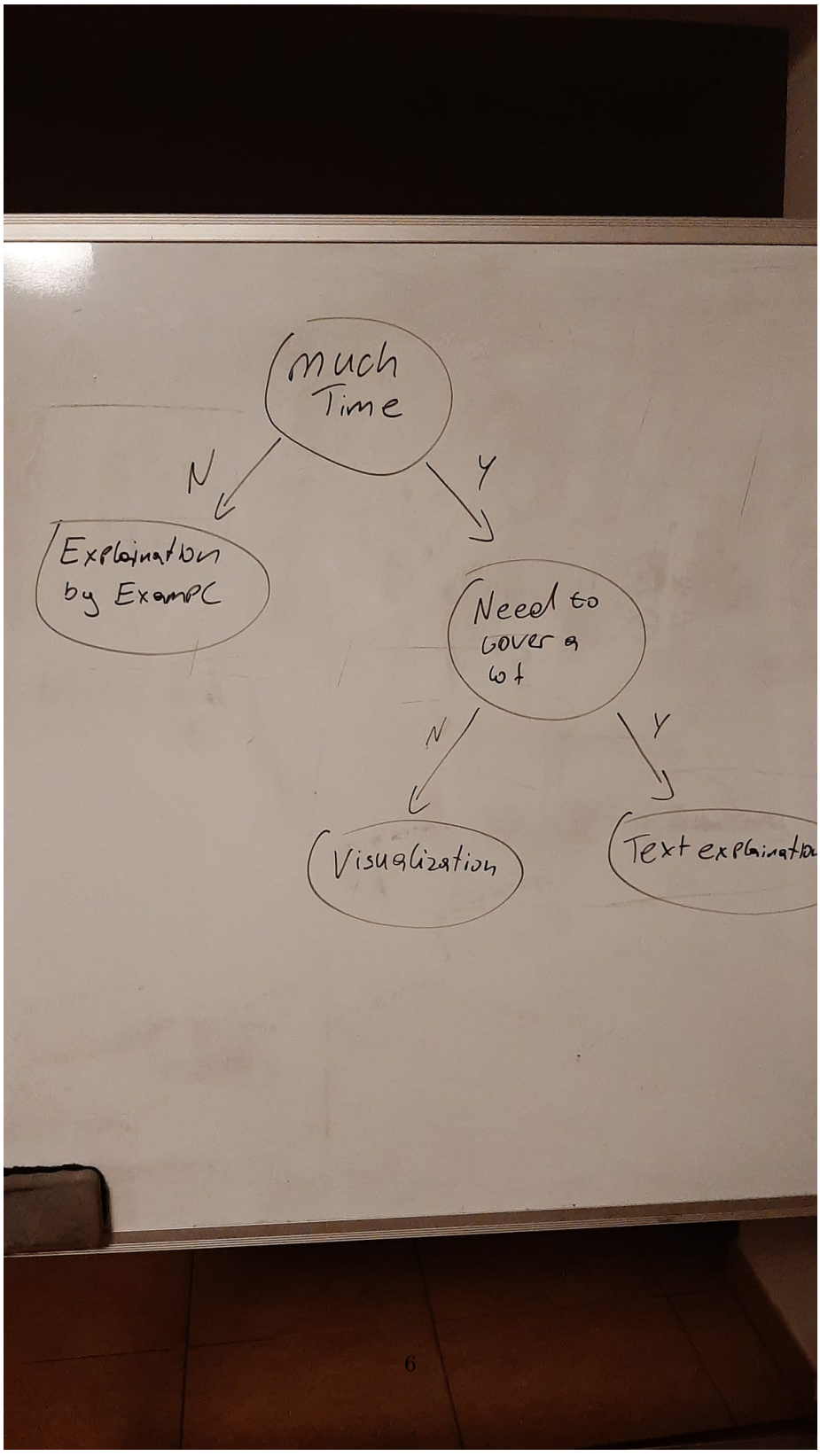


Figure 1: Flowchart

tinyurl.com/4fxerra4 The dataset includes information on academic background, demographics, and socialeconomic factors. Its intended use was the prediction of student dropouts and using these predictions to put measures in place that reduce dropouts

4.1 a)

The dataset was created from multiple datasets. Its purpuse is to find students at "risk" who might drop out and then to intervene early.

Three Stakeholders might be

1. Dean of the faculty Dr. Mag. Prof. etc. MSc. BSc. Markus Fischer:
   - Goal: Optimizing faculty resources.
   - Question: Does a preliminary course in math really help to lower the drop out odds?

2. Study advisor Jan Janus:
   - Goal: Want to intervene early
   - Question: Which combinations of features are at high risk? E.g.: Age of enrollment is above 23 and tuition fees not up to date.

3. Scholarship Office manager Claudia Punz:
   - Goal: Monitoring over scholarship awards.
   - Question: Does late tuition fees correlate with the dropout rate?

4.2 b)

For this section first normed the labes for dropout as such: Dropout = 0, Enrolled=1, Graduated=2. As the model I used ExplainableBoostingClassifier from interpretML. Then i fitted the date and got the conusion matrix in figure 2.

The model had some difficulties fitting the data, but I think I did something wrong while preparing the data and then I run out of time. :(

In figure 3 you can see that one feature heavily influences the prediction. (The feature is Curricular units 2nd sem (approved))

So it seems that the number of successfully completed curses in the second semester influence the drop out rate.

4.3 c)

In the real world I would filter out this feature for the specific stakeholder. Eg.: For the Study advisor that the number of approved curses in the second semester has a big influance on the drop out rate might not be to shocking since this makes sense for young students who just "try out" the major. More interesting might be the drop out rate for students who are more advanced in the major. For this purpose in Figure 4 I extract the probability of dropping out for the feature "Tuition Fees Up-To-Date".

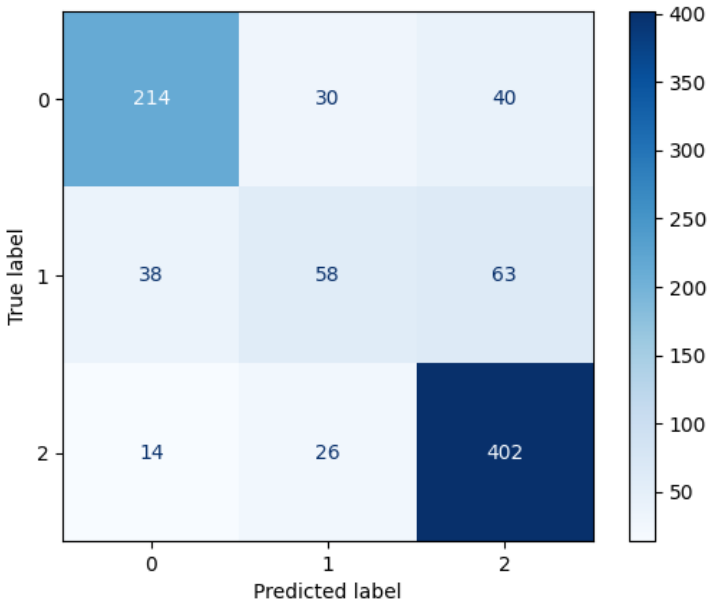


Figure 2: confusion matrix

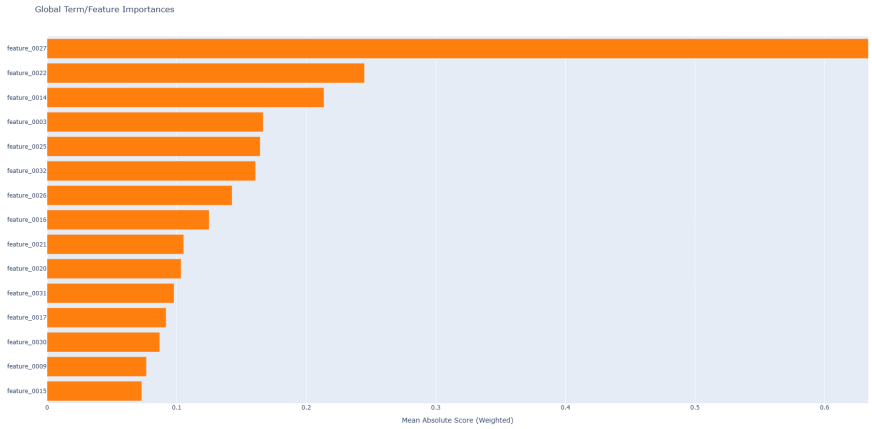
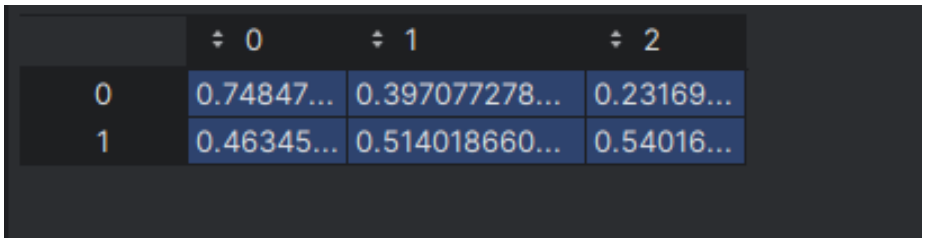


Figure 3: Global Feature Importance



	0	1	2
0	0.74847...	0.397077278...	0.23169...
1	0.46345...	0.514018660...	0.54016...

Figure 4:
x=0 Tuition Fees Up-To-Date,
x=1 Tuition Fees not Up-To-Date,
y=0 Dropout,
y=1 Enrolled,
y=2 Graduate

This might be more interesting for the Study advisor since this feature might play a role in the economical well being of the student - does the higher dropout rate if the Tuition Fees are not Up-To-Date.