DataLab Preparation (Week 6, DataLab I, Monday)

**2a A common trade-off mentioned in the context of Explainable AI (XAI) involves the challenge of achieving high model accuracy, while at the same time ensuring interpretability in AI models. For a deeper understanding of this trade-off, consider looking into Molnar's book or searching for it on the web. Please elaborate on this trade-off using your own words, and write down your explanation.**

The trade-off between XAI and achieving high model accuracy is mainly rooted in the 'blackbox of AI'. Often, achieving good accuracy levels involves using complex models with intricate layers and hidden workings, which are not entirely known as to what each iteration is concerning.

This lack of interpretability is a major problem. Trust is one of the reasons for these issues as if we cannot entirely understand how a model makes its decisions, it is even harder to trust that it does so rightfully and in a way we consider fair, even if we understand the mathematical workings behind the concepts itself. Secondly, this also leads to debugging problems, as pinpointing the source of the error becomes a guessing game. Additionally, when it comes to biases, these complex models can inadvertently learn biases from the data they are trained on, and improved interpretability helps us identify and mitigate them.

**3a Come up with three use cases, and link each to a business or industry of your choice. Explain how XAI can be beneficial in each use case.**

The retail industry:

Theft detection - XAI can be beneficial by highlighting which features (such as clothes, what is considers suspicious activity, time of day when most thefts happen, etc.) the model uses to deem certain individuals as 'high risk' when it comes to theft.

Finance Industry:

Loan approval - By implementing XAI techniques like feature importance analysis rejected applicants can understand why their loan was denied. This builds trust and allows them to address any shortcomings in their application.&nbsp;

Automobile Industry:

Car manufacturers- XAI can be used to explain which sensor readings or data points triggered the failure prediction,&nbsp; improving repair efficiency but also identifying weaknesses in equipment design, leading to future improvements.

**4a Compare and contrast the following concepts: model specific vs. model agnostic, and local vs. global methods. Write down your answer.**

Model specific interpretation tools - limited to specific model classes. The interpretation of regression weights in a linear model is a model-specific interpretation, since – by definition – the interpretation of intrinsically interpretable models is always model-specific. Tools that only work for the interpretation of e.g. neural networks are model-specific.

Model agnostic interpretation tools - can be used on any machine learning model and are applied after the model has been trained (post hoc). These agnostic methods usually work by analyzing feature input and output pairs. By definition, these methods cannot have access to model internals such as weights or structural information.

Local methods:

For a single prediction - You can zoom in on a single instance and examine what the model predicts for this input, and explain why. If you look at an individual prediction, the behavior of the otherwise complex model might behave more pleasantly. Locally, the prediction might only depend linearly or monotonically on some features, rather than having a complex dependence on them.

For a group of predictions - Model predictions for multiple instances can be explained either with global model interpretation methods (on a modular level) or with explanations of individual instances. The global methods can be applied by taking the group of instances, treating them as if the group were the complete dataset, and using the global methods with this subset. The individual explanation methods can be used on each instance and then listed or aggregated for the entire group.

Global methods:

Holistic model interpretability - To explain the global model output, you need the trained model, knowledge of the algorithm and the data. This level of interpretability is about understanding how the model makes decisions, based on a holistic view of its features and each of the learned components such as weights, other parameters, and structures. Which features are important and what kind of interactions between them take place?

Modular level - While global model interpretability is usually out of reach, there is a good chance of understanding at least some models on a modular level. Not all models are interpretable at a parameter level. For linear models, the interpretable parts are the weights, for trees it would be the splits (selected features plus cut-off points) and leaf node predictions. Linear models, for example, look like as if they could be perfectly interpreted on a modular level, but the interpretation of a single weight is interlocked with all other weights. The interpretation of a single weight always comes with the footnote that the other input features remain at the same value, which is not the case with many real applications.

**5a Select one topic within the field of Data Science & AI, and explain it in a similar way as in this video; explain it to your non-technical grandmother, and a first year ADS&AI student. Write your answer down, or make a video recording of your explanation (include the GitHub link to the quiz), and explain how the explanations differ for each type of audience. Try to integrate the attributes of a 'good' explanation from Molnar's book (See Section 3.6!).**

I chose the topic of recommendation systems, such as Google Ads.

Grandma explanation:

Imagine you're a great recipe collector, and everyone wants your recommendations! A recommendation system is like your helpful granddaughter who remembers everyone's favourite dishes and suggests similar ones they might enjoy. It uses past experiences (like the meals you've cooked for people) and their preferences (what they liked or disliked) to suggest new recipes they might love. Just like you wouldn't recommend a spicy dish to someone who hates peppers, the system learns from past choices to make suggestions that are a good fit. This way, everyone gets delicious meals (or finds cool products online) that they'll enjoy!

First year student explanation:

Recommendation systems are a type of machine learning technique used to predict a user's preferences and suggest relevant items. These systems leverage collaborative filtering or content-based filtering approaches. Collaborative filtering analyses user behaviour patterns and interactions with items (like purchases or movie ratings) to identify users with similar tastes. The system then recommends items that users with similar preferences have enjoyed. Content-based filtering focuses on the characteristics of the items themselves. For instance, an e-commerce platform might recommend shoes similar to ones you've previously purchased based on size, style, or brand. These approaches can be combined for even more personalised recommendations. Evaluation metrics like precision, recall, and recommendation accuracy are used to assess the effectiveness of these systems.

**5b How would you approach designing this registry, and what key elements would you prioritize to ensure its effectiveness? Write your answer down.**

To build a trustworthy government algorithm registry, I'd focus on collaboration. Working with agencies, legal experts, and citizen groups ensures the registry meets everyone's needs for transparency and legal compliance. The online platform itself would be user-friendly and standardized for easy data entry across agencies. Key elements include basic algorithm info, how it works (without revealing secrets!), data used, and potential biases. Legal considerations like data privacy and impact assessments would be addressed. Public engagement is crucial – FAQs, contact info, and even multilingual support can bridge the gap between citizens and government algorithms. Regular updates and incorporating user feedback ensure the registry remains a valuable tool for responsible AI use in government.