Model Explainability & Interpretability

Machine Learning in Production / AI Engineering - Recitation 10

Overview

- Model Explainability
- What is Interpretability?
- Why is Explainability Important?
- Interpretable Models
- Interpreting Deep Learning Models
- Interpreting the Whole Model
- Interpreting Single Prediction
- Demo
- References

Model Explainability

- The process of being able to explain why a model predicted what it did
- To explain the model, first we have to ***Interpret*** it
- We shall be using explainability and interpretability interchangeably but there is a slight difference

What is Interpretability?

Interpretability is the degree to which a human can understand the cause of a decision

OR

Interpretability is the degree to which a human can consistently predict the model's result

- Source Interpretable Machine Learning book Ch. 3

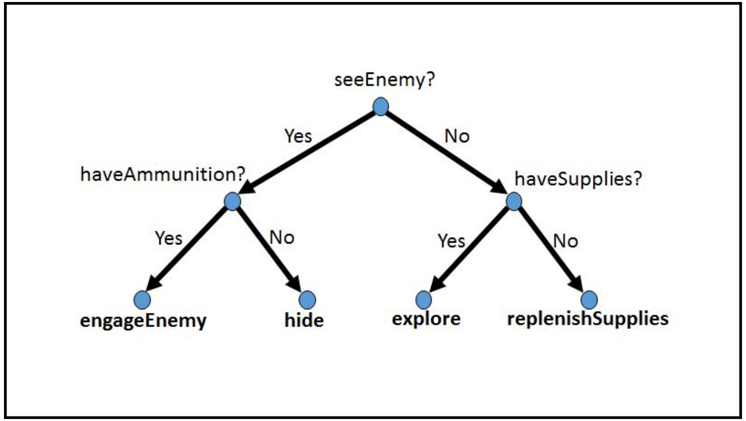
Why is Explainability Important?

- To make the models better, you have to interpret and explain the reasoning
- It could be required by law to use explainable models:
    - Credit decisions
    - Life insurance premiums
- You want to ensure the model is *fair*

Interpretable Models

- Certain ML algorithms are intuitive to interpret than others:
  - Decision Trees (If-else)
  - Naive Bayes (Counting)
  - Linear Regression (Distance from line)
  - Logistic Regression (Side of line)
  - K-Nearest Neighbors (Distance)
- Neural Networks are convoluted (Pun intended)

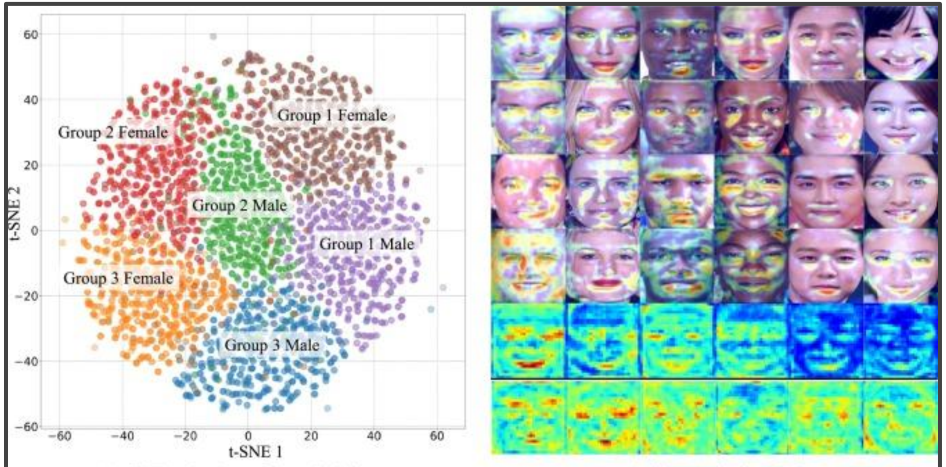
Example of a decision tree



Interpreting Deep Learning Models

- All Deep Learning Methods like CNNs, LSTMs, Transformers and more, have millions of weights and interact in complex ways that makes them hard to explain and interpret.

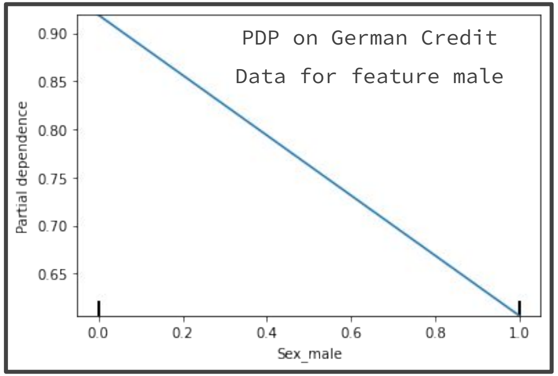
- These interactions create hyper-dimensional and latent mapping of the inputs that are hard to interpret

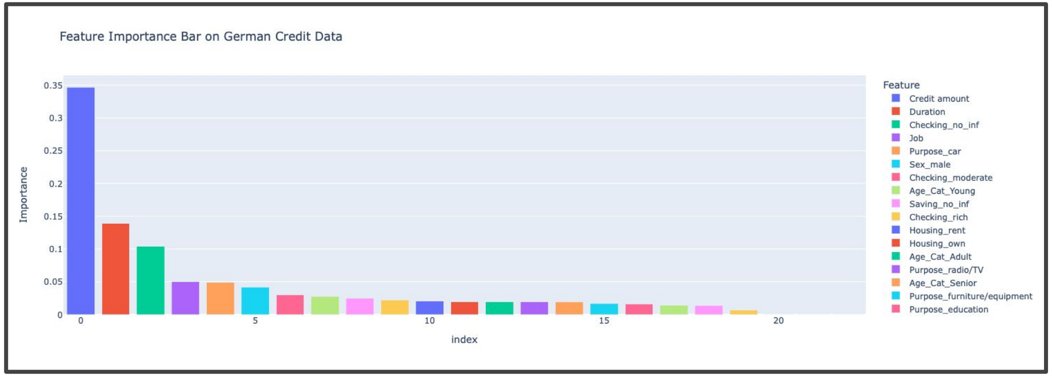


(a) t-SNE projections of the embeddings. (b) Class Activation MAP.

Interpreting the Whole Model

- Describes the average behavior of the model for each feature
- Popular Methods:
    - PDP: Partial Dependence Plot - Shows relationship between one feature and the outcome
    - Feature Importance - How a feature affects output



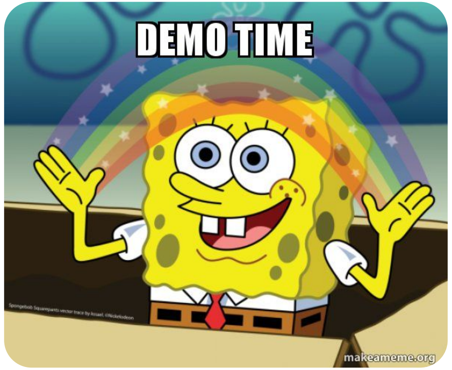


Interpreting Single Prediction

- Ways to interpret and analyse what caused individual predictions
- Popular Methods:
  - SHAP (SHapley Additive exPlanations): Calculates contribution of each feature for a prediction. Based on game theory.
  - Counterfactual: Examines the causal relation between output and features. Eg. “If I hadn’t taken a sip of this hot coffee, I wouldn’t have burned my tongue”

Demo

- Jump to the Colab Notebook: https://colab.research.google.com/drive/1ZiawXPUplLVVTAjz-734dChjyxII0XZO?usp=sharing
- Exercise:
  - Make a copy of the notebook
  - Some model interpretations have been done in the code, please try other models in the **classifiers** dict



References

- Demo Code: https://colab.research.google.com/drive/1ZiawXPUplLVVTAjz-734dChjyxII0XZO?usp=sharing
- Video: https://youtu.be/tYRWFSIc2IM
- Book: https://christophm.github.io/interpretable-ml-book/
- Use of SHAP Plots: https://medium.com/analytics-vidhya/shap-part-3-tree-shap-3af9bcd7cd9b
- DiCE library for Counterfactuals https://interpret.ml/DiCE/notebooks/DiCE_model_agnostic_CFs.html