5 - Explainable AI

5.1 Explainable AI (XAI)

5.1.1 Explainable AI and its Need

Once an ML model is trained, it should work with a defined level of accuracy and for all the defined variations of scenarios. If the quality of the predictions by the model is insufficient for some scenarios and work fantastically for others, it is likely to be a biased model. A dataset for which the model quality goes below the defined level of quality represents a defect.

As a tester, it is very difficult to discover all such defects without using a systematic approach. One needs to examine the behavior of the model and its variation with change in input factors to deduce an approximation of the relationship between input and output. This kind of deduction is called interpreting or explaining the ML model. Such approximate relationship may not be a replacement of the actual model, but it may be good enough to reveal possible biasing. An insight into the model behavior helps evaluating its overall quality and the model's viability for deployment. Other reasons for requiring explaining or interpreting models can be safety measures, social acceptance, detecting bias, or human curiosity and learning about the model [EA1].

Not all models can be explained. The more complex the model, the less likely it is to interpret or explain it. The output of non-DL models, e.g. random forest, decision trees, linear regression, etc., can conveniently be explained in terms of the input variables.

The DL models are inherently complex in their implementation, so it is better to examine the model as a black-box, i.e. observe the variations of outcomes by small perturbations of the input variables and approximate the underlying model by a simple, interpretable model.

Some of the popular and easy-to-use algorithms, tools and approaches for model interpretation include: Local Interpretable Model-agnostic Explanations (LIME) and Class Activation Maps (CAM).

Explainable AI and its need

One of the most important things of the ML models is understand all the rationale behind the predictions that could be made of it and make the right questions to ensure we are covering most of the predictions.

Why interpret?

- Unlike traditional system which are rule based, AI systems will behave differently on the basis of their training data
  - Users are asked to trust a model and we “Trust by Verify”
  - To verify the black-box Machine Learning models, we need method beyond the metrics like accuracy and aggregate measures.
- Regulatory compliance often wants better visibility into rationale
- We need tools that can serve as oracle/predict model’s behavior to decide the trustworthiness of the model
  - One needs to be able to answer the question “why did the system do what it did and how will it behave in a different situation?

Tools and approaches

- Self-explainable algorithms
  - Decision trees
  - Linear Regression
  - Random Forest (Using Feature Importance)
- External tools to explain the model
  - LIME (Local Interpretable Model Explanations)
  - Counterfactuals
  - Activation maps
    - Class Activation Maps
    - Gradient Class Activation Maps

Explainable Models - Linear Regression

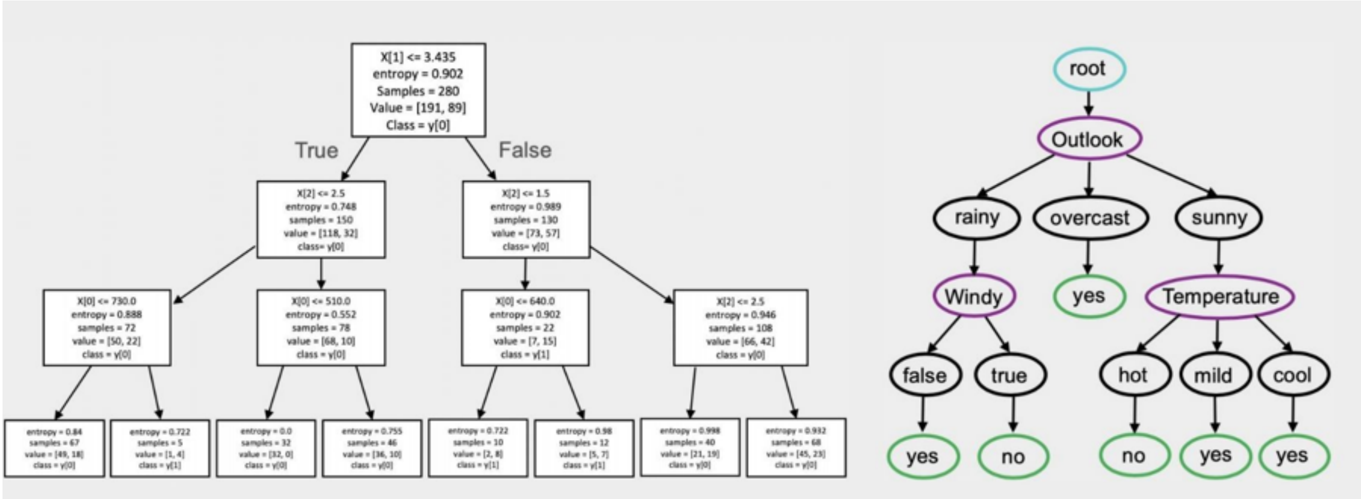
Linear regression models are more likely to be explained in terms of input variables

- Let us say, we derive a model: $Y = 3 * X1 + 4 * X2 - 6 * X3 - 20$
- A clear set of analyses is possible from this model
  - Y is positively dependent on X1, X2 and negatively with X3
  - This can be verified against intuition
- Explanation
  - If Y is the price, and X1 is the rarity of the product, it is ok
  - If X3 is something like the amount usually purchased, it is fine as you get discounts for larger purchases
  - In fact, if X3’s coefficient was positive, it would be a cause for concern and might force you to rethink the model

- The inherent structure of linear regression models make it a convenient approach to explain the output in terms of input values.

Decision Tree

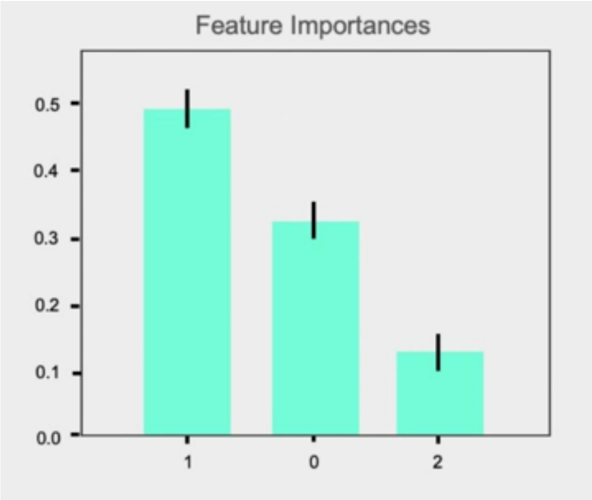
These are intuitive and unexpected results are easy to identify.



- The rules derived from the decision tree explained the three well
- The problem here is that of determining admission to US schools based on three factors:
  - X[0] = GRE score
  - X[1] = GPA in the class
  - X[2] = Rank in the class
- Some rules from decision tree
  - If GPA > 3.4 and rank = 1, then, admit = 1 (Yes)
  - If GPA > 3.4 and ran > 1, then, admin = 0 (Not Admitted)
  - If GPA < 3.4 and (rank = 1 or 2) and GRE > 730, then, admit = 1 (Yes)
  - If GPA < 3.4 and (rank = 1 or 2) and GRE < 730, then, admit = 0 (No Admitted)
- It looks intuitive and can check easily for unexpected results

Feature importance

This feature helps to identify the tendency of data, to define which one has more value than the others.



- Some techniques give a good idea of which input features are more important than others.
    - Example: Random Forest
- In the admission problem
    - 0 – GRE
    - 1 – GPA
    - 2 – Rank
- It shows that GPA is most important

5.1.2 LIME

LIME is fed a sample to study model predictions for the sample and its closer variations and it reveals the input features responsible for the model predicted output. LIME generates enough close variants of the sample input and gets outcome for each of the variants. Thus, it attempts to approximate how small variations in input are changing the output variable. Since all variants generated and studied by LIME are in close vicinity of the given sample, LIME explanations are called 'local' (to the sample input). Also, it is a model-agnostic method, as LIME does not require to peek into the model or algorithm.

LIME can be used on an image classifier to generate explanations. It deduces image portions that play a significant role in determining the outcome. It allows a tester to relate as well as to rule out that the model is not basing its deductions on irrelevant features. Similarly, for a text classifier, LIME can point out the words (text portions) that lead to the categorization of the sample text into one of the predefined classes. This helps a tester to assess whether the model is basing its deductions on irrelevant words.

How LIME works

- Perturbs the input
- Creates a fake dataset full of permutations of the perturbed data
- Test the model and examines the output variations from the predictions
- Fits a simple model between permuted data and model outcomes

For images: Is used a pretrained model meant for image classification, transforms the image to a structured form and get predictions about the object in a test image and leads to show which parts of the image are predictions.

For text: Fetch an available text dataset and is converted to structured form, the model is trained used structured dataset and compute the accuracy of the training dataset, get the prediction against a test sample, and use LIME to identify words contributing to predictions and find bugs in the model, once fixed, use LIME again to verify the fix

5.1.3 Class Activation Mapping (CAM) for Neural Networks

CNNs are very useful but they are not transparent about why they reached a particular conclusion. In order to bring transparency to such models, Gradient- based Class Activation Mapping (Grad-CAM) visualizes the input regions that are important for predictions from these models. It uses the class-specific gradient information flowing into the final convolutional layer of a CNN and produces a coarse localization map of the important regions in the image.

- Grad-CAM is a technique used to make convolutional Neural Networks (CNN)-based models more transparent.
- It helps visualizing the regions of the input that are important for the predictions from these models.
- It uses the class-specific gradient information flowing into the final convolutional layer of aa CNN.
- This produces a coarse localization map of the important regions in the image.

Fairness

Fairness Indicators help computate commonly-identified fairness metrics for binary and multiclass classifiers.

- Evaluate the distribution of datasets
- Evaluate model performance, sliced across defined groups of users
- Figure out confidence intervals and evals at multiple thresholds
- Dive deed into individual slices to explore root causes and opportunities for improvement

Bias

- Bias: “prejudice in favor of against one thing, person, or group compared with another, usually in a way considered to be unfair”.
- Bias (also known as unfairness) is a measured of the distance