Machine Learning Interpretability

Analytics Vidhya
Learn everything about analytics

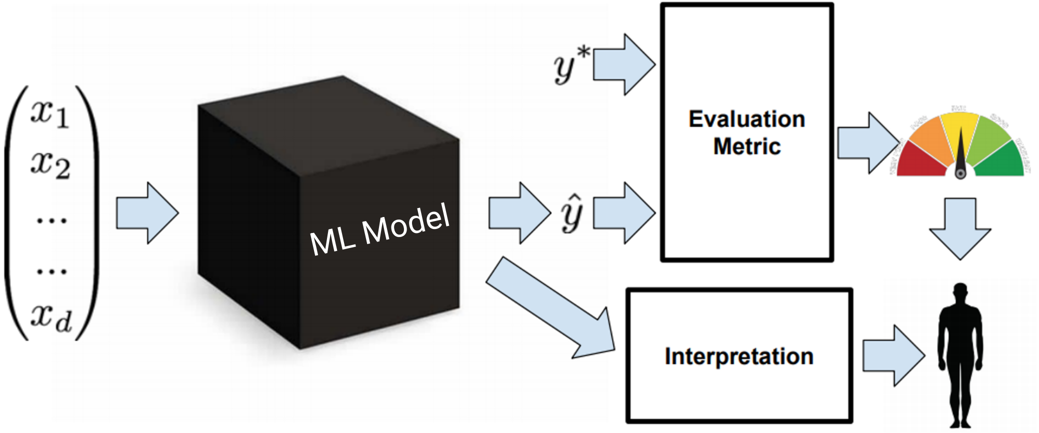
Modelling Life Cycle

6 Stages of Modelling Lifecycle

1. Problem Definition
2. Hypothesis Generation
3. Data Extraction / Collection
4. Data Exploration and Transformation
5. Predictive Modeling
6. Model Deployment/ Implementation

What is interpretability?

Interpretation is the process of giving explanations to humans.



Analytics Vidhya
Learn everything about analytics

Importance of interpretability

Fairness

Example 1: Predicting employee's performance at a big company

Data available: Past performance reviews of individual employees in the last 10 years

What if that company tends to promote more men than women?

The model might learn the Bias and predict that men have higher performance

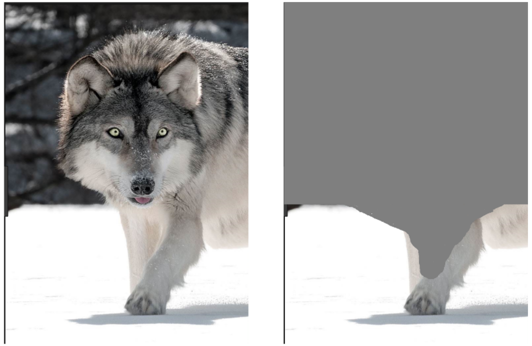
Importance of interpretability

Example 2: Classifying Images: Wolves vs Dogs

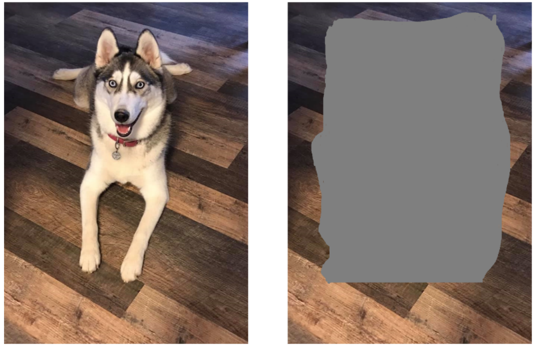
Data available:

- Pictures of wolves and dogs

What if pictures show something different in the background?



Wolf



Dog

Importance of interpretability

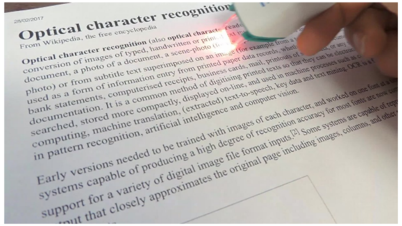
Regulations

- In the EU GDPR, article 12 allows individuals to inquire as to why a particular algorithmic decision was made for them

When we do not need interpretability

- Does not impact end customer

- Problem is well studied - OCR

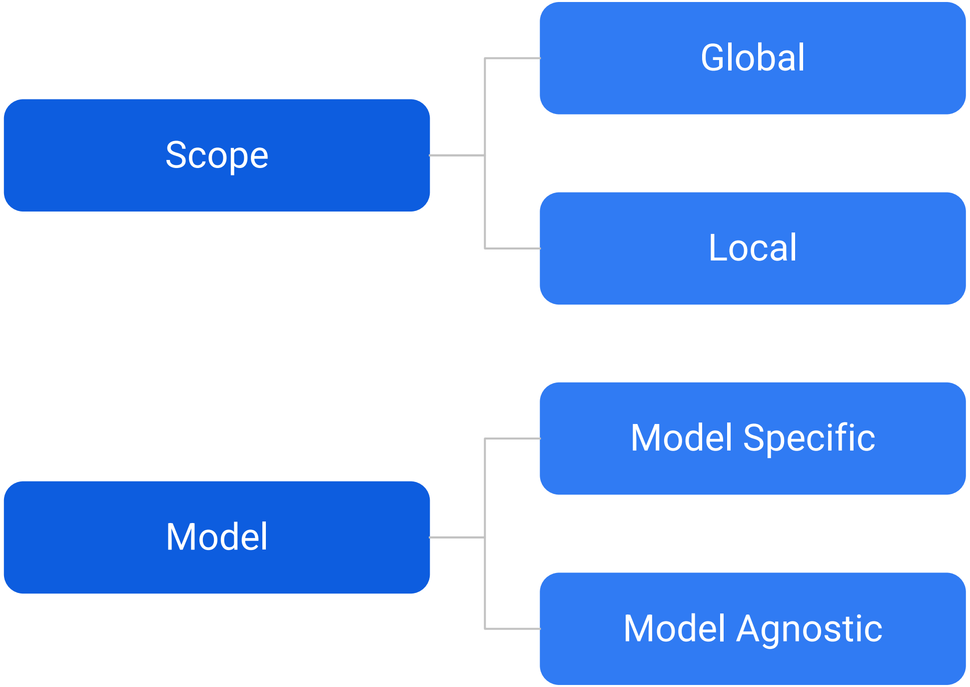


Analytics Vidhya
Learn everything about analytics

Thank You!

Machine Learning Interpretability

Framework of Interpretability



Analytics Vidhya
Learn everything about analytics

Interpretable Models: Linear/Logistic

- Weights/coefficients of the linear/logistic regression basically represent the importance of each variable
- Suppose we are trying to predict the salary for an employee based on 2 features: experience in years and previous rating out of 5

*Salary = W1*experience + W2*rating*

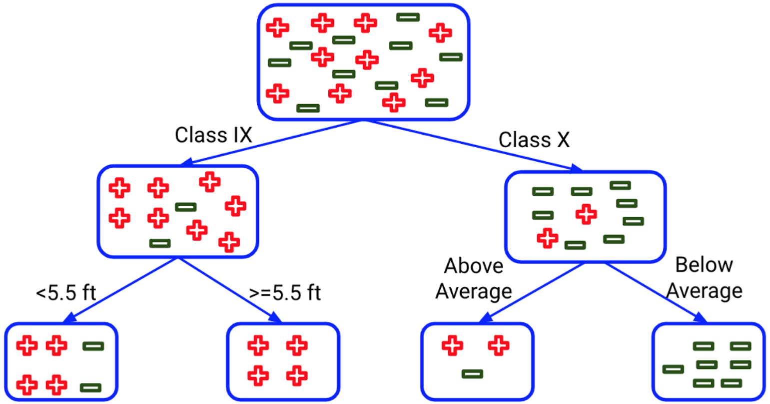
- For normalized data, Weights W1 and W2 here can essentially tell us whether rating contributed more or experience contributed more towards an employee's salary

Interpretable Models: Linear/Logistic

Scope	Global & Local
Model	Model-Specific

Interpretable Models: Decision Trees

- Decision Tree is another such algorithm which is highly interpretable
- Looking at the plot of the decision tree, it is easy to see how a decision was made

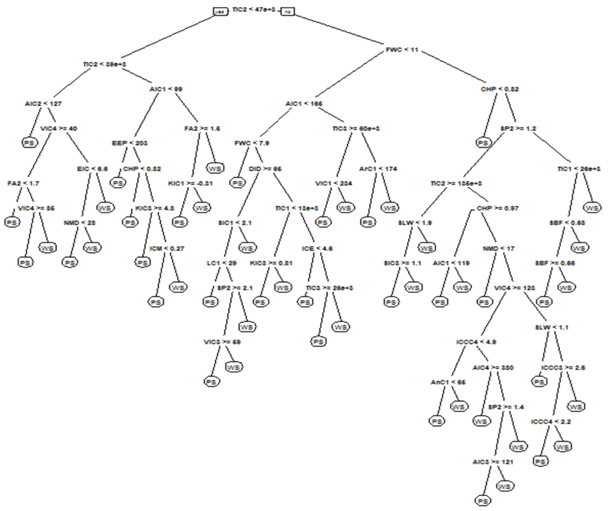


Interpretable Models: Linear/Logistic

Scope	Global & Local
Model	Model-Specific

Feature Importance for Deep Decision Trees

- For decision trees with large max depth, it is difficult to effectively present the decision rules



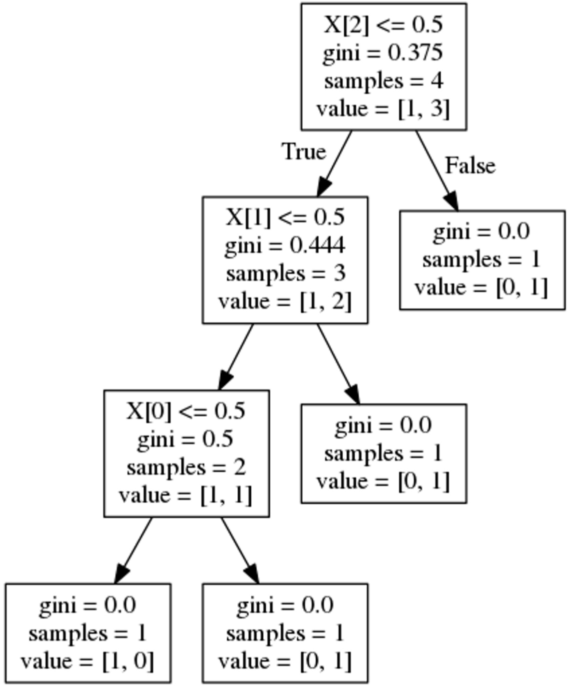
Feature Importance for Deep Decision Trees

- Go through all splits in which feature was used
- Measure how much it has reduced the weighted criterion (gini/information gain) compared to the parent node

$$\frac{N_{parent}}{N}\left(Gini_{parent} - \frac{N_{Right}}{N_{parent}} \cdot Gini_{Right} - \frac{N_{Left}}{N_{parent}} \cdot Gini_{Left}\right)$$

- N is the total number of observations
- N & Gini represents the number of samples & gini impurity in parent, left and right node
- Take sum for all splits and compare

Feature Importance for Deep Decision Trees



Since each feature is used once in our case, there is no need for sum

$$\frac{N_{parent}}{N}\left(Gini_{parent} - \frac{N_{Right}}{N_{parent}} \cdot Gini_{Right} - \frac{N_{Left}}{N_{parent}} \cdot Gini_{Left}\right)$$

For X[2] :

feature_importance = (4 / 4) * (0.375 - (0.75 * 0.444)) = 0.042

For X[1] :

feature_importance = (3 / 4) * (0.444 - (2/3 * 0.5)) = 0.083

For X[0] :

feature_importance = (2 / 4) * (0.5) = 0.25

Analytics Vidhya
Learn everything about analytics

Feature Importance for Random Forest & Gradient Boosting

- Go through all trees in the ensemble
- Calculate feature importance for each tree
- Find average Feature Importance by using the formula

$$Feature\ Imp. = \frac{Sum\ of\ \ Feature\ Imp\ of\ \ all\ estimators}{Number\ of\ \ Trees}$$

Thank You!