Responsible AI

Lecture 5

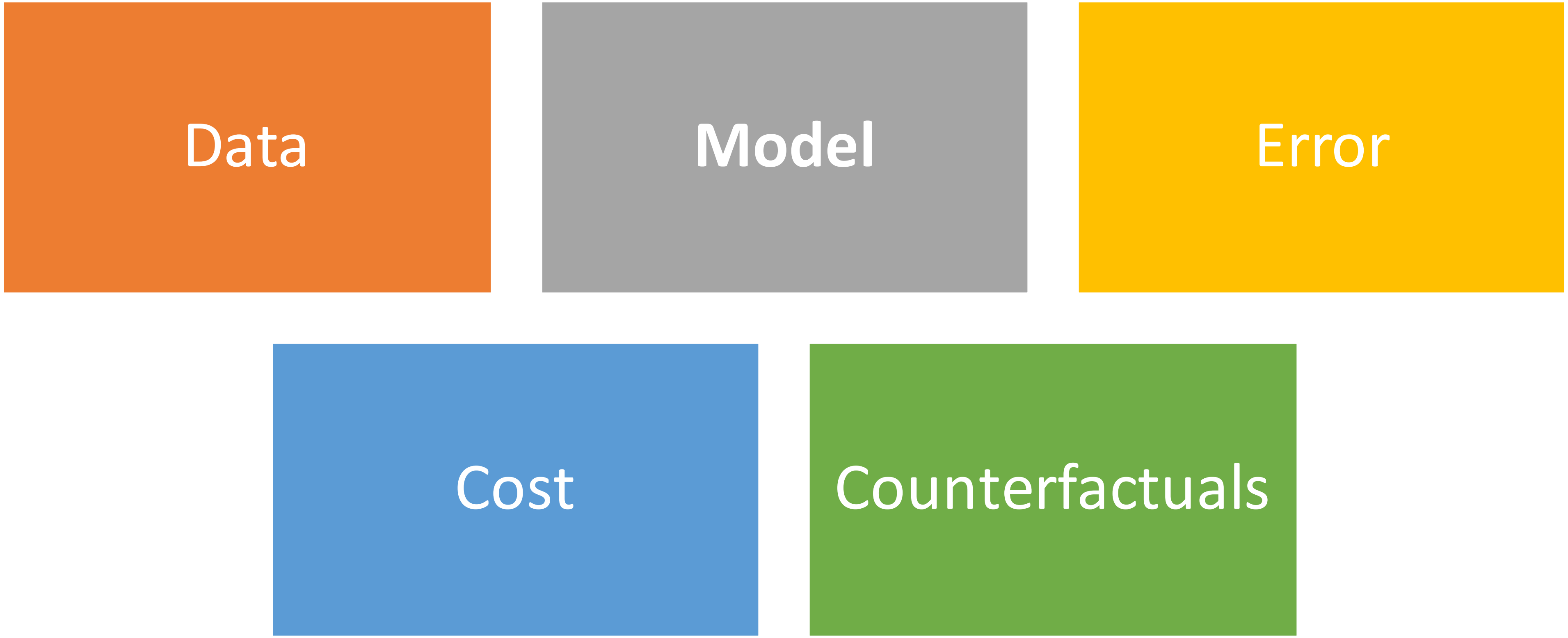
Today

1. Explainable AI
2. Introduction to XAI
3. Feature explanation
4. Information value plots
5. Model explanation - split and compare quantiles
6. Explainable models - Generalized Additive Models (GAM)
7. Counterfactual explanation

IS THAT A LOT?
imgflip.com

What we need to explain and why

Explainable AI



Data

Model

Error

Cost

Counterfactuals

Data explainability

PDP

ALE

SENSITIVITY

INFO VALUE

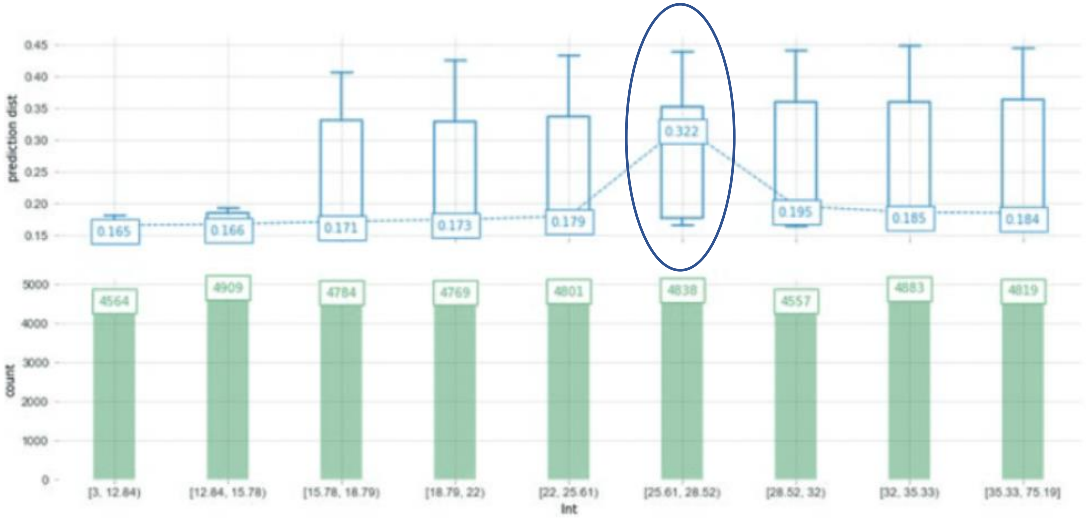
Information Value

First calculate WoE which will be used to calculate IV. WoE helps us quantify the predictive power of a feature on the output, and the IV uses WoE to assign a score (IV) that can be used to compare and prioritize the features.

$$WOE_{ij} = \log \frac{P\left(X_j \in B_i \mid Y = 1\right)}{P\left(X_j \in B_i \mid Y = 0\right)}$$

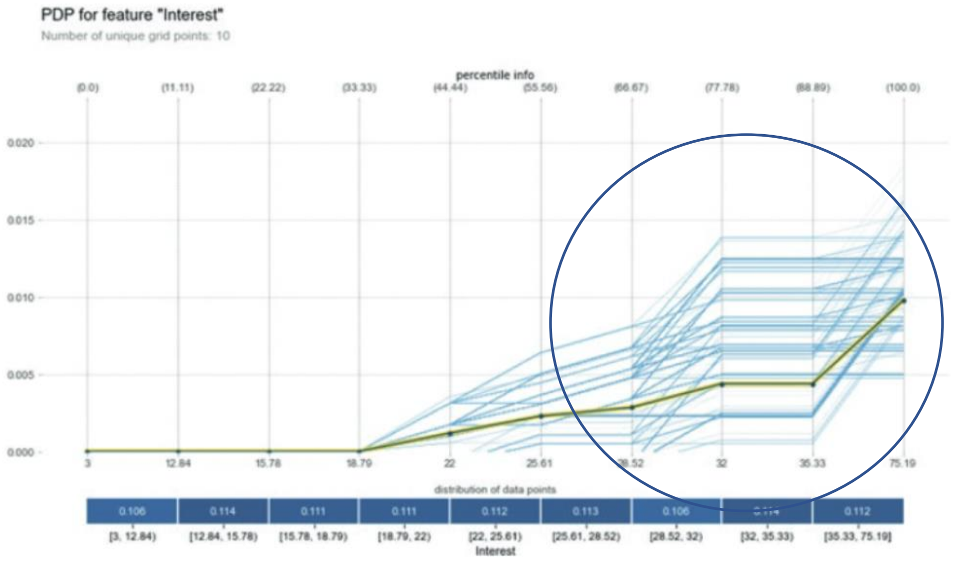
Using the WoE values, we can now determine the IV

$$IV_j = \Sigma\left(P\left(X_j \in B_i \mid Y = 1\right) - P\left(X_j \in B_i \mid Y = 0\right)\right) \times WOE_{ij}$$



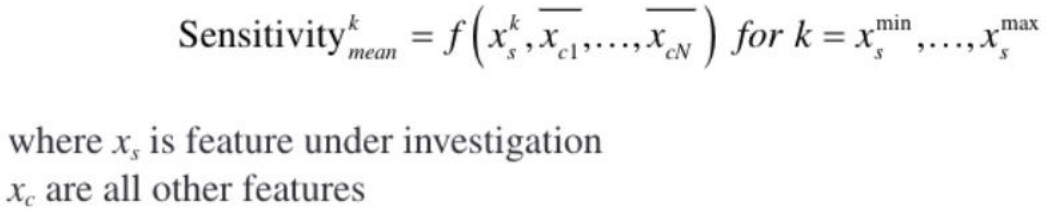
PDP

The PDP shows the relationship between the outcome and the feature being investigated. The benefit of this approach is that it is model-agnostic and can be implemented for any kind of classification or regression mod-els. However, this approach assumes that the features are not correlated with each other. This is often not the case in real life, making it difficult to apply PDP



Sensitivity Plots

This technique covers all values between the minimum and maximum values for the feature in the dataset, demonstrates the sensitivity of the model at the boundary conditions, average and for other values in between and finally the resulting plots are easily interpretable without the need to understand the underlying math – making it a strong tool to use when explaining to the user or the business stakeholders

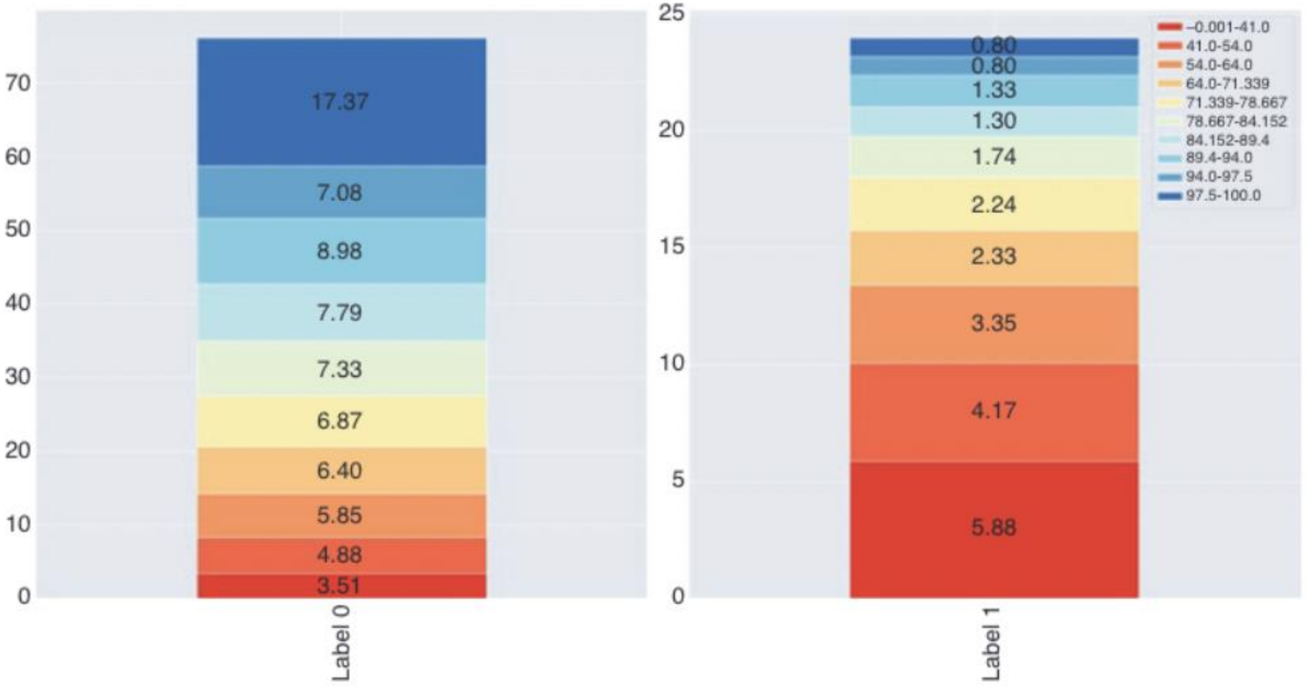


$$\text{Sensitivity}^{k}_{mean} = f\left(x^{k}_{s}, \overline{x_{c1}}, \ldots, \overline{x_{cN}}\right) \;\; for\ k = x^{\min}_{s}, \ldots, x^{\max}_{s}$$

where x_s is feature under investigation

x_c are all other features

Split Quantiles



The split and compare quantiles helps us define a decision threshold for a classification problem by giving a clear understanding of the impact of our decision on the confusion matrix and evaluate if the model helps meet the business objectives.

Self Study: Split Quantiles - Cost

SHAP

- The plot shows the overall strength of features along with class separation. The features higher on the chart have clear impact on both the labels. The feature importance chart on the left is generated after travers-ing across all the instances and shows the average Shap value for each feature. This provides a much simpler view to understand the relative importance of the features on the output labels – features with higher rank have an average higher impact

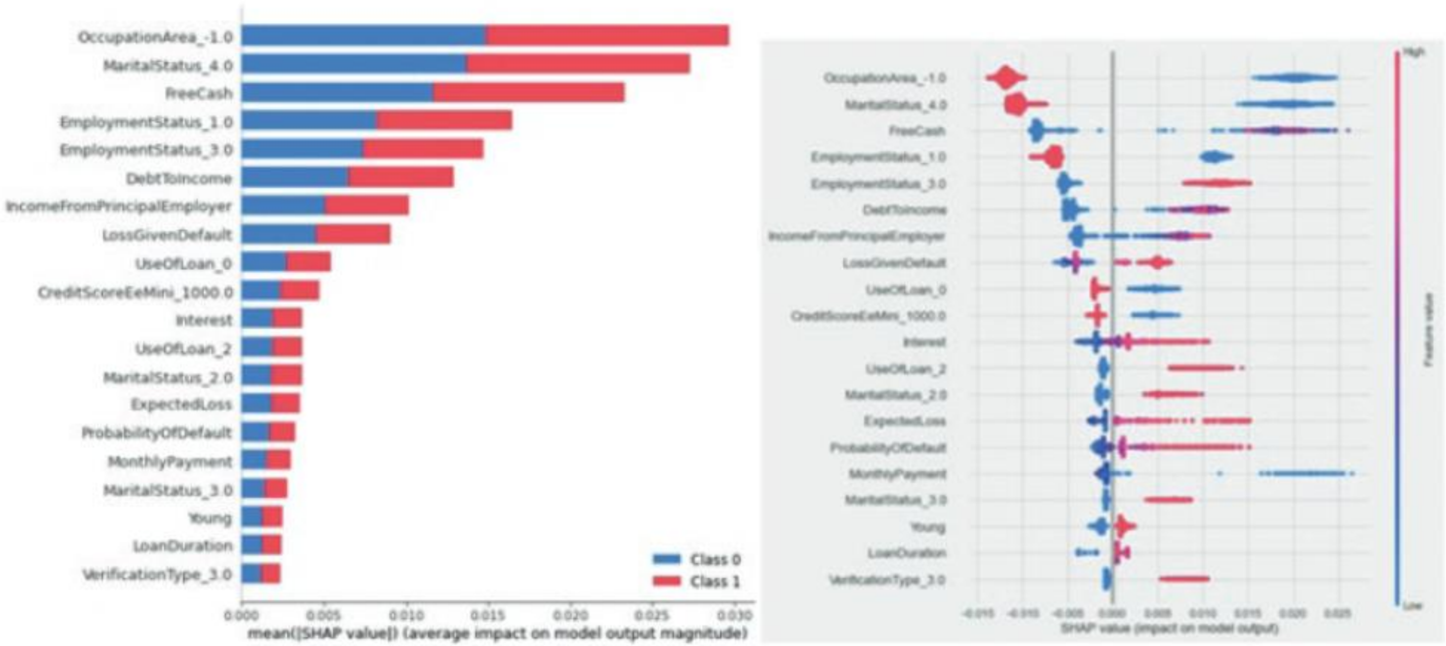
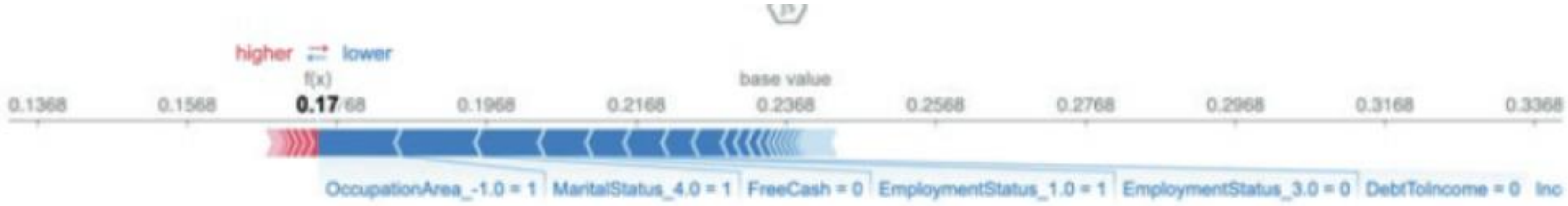


Fig. 4.7 Global explanation output by SHAP showing the average impact (left) and the actual impact (right) of a feature on the model

Local XAI : SHAP



both the direction and magnitude of the feature and how they drive the outcome. Take for instance Shap's force plot. In the illustration, we investigate a customer whose probability of non-default is 0.8256173 (and default is 0.1743827). The colour of the individual bar shows the direction of the feature impact while the length of the individual bar the power of the impact. In the case below, the maximum impact comes from "occupation area". Remember this is for this one particular customer and should not be generalized. Interestingly, the distance between base and prediction will be equal to the differ-ence in the length of the blue bars and the length of the pink bars

Self Study: LIME

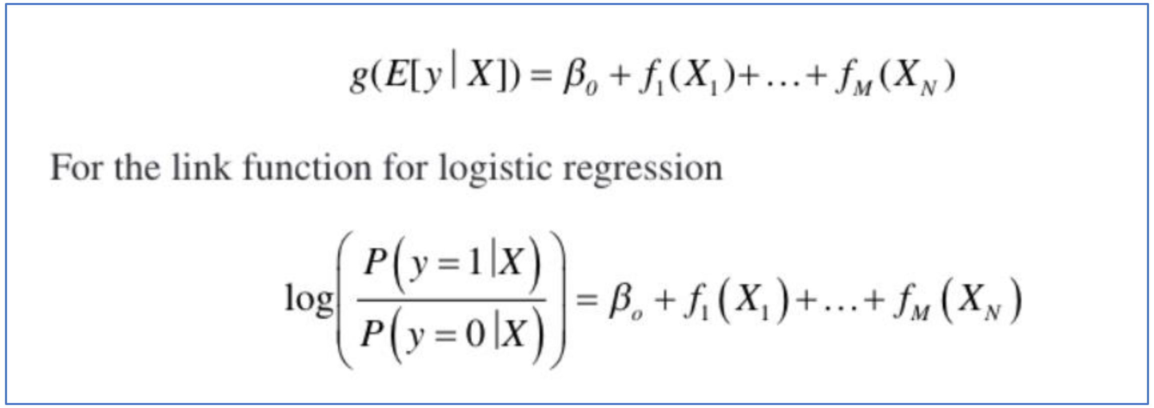
Self Study: EBM

GAM

Since this is an additive model with a separate function for each feature, the impact of each feature is known at all times. There are no hidden inner workings of the model making the explanation difficult. This ability makes this a “glass box” model and deserves a strong consideration if your other options so far are black box models with additional explanation added.

GAM uses splines for learning the link function and also computes the weights for the splines. Additionally, it can also use a regularizer to reduce overfitting and the grid search method to optimize the weights

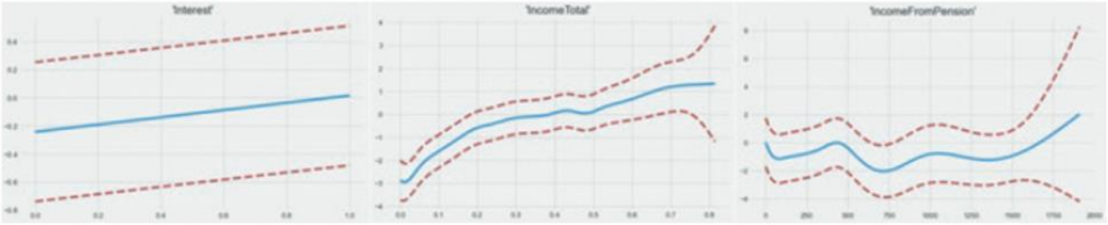
Interestingly, functions for predictors can also be automatically derived during model training and you don’t need to state the function before training the model. Just like other linear models, one has an option of either declaring the smoothing parameters or use cross validation to find the optimal smoothing parameters



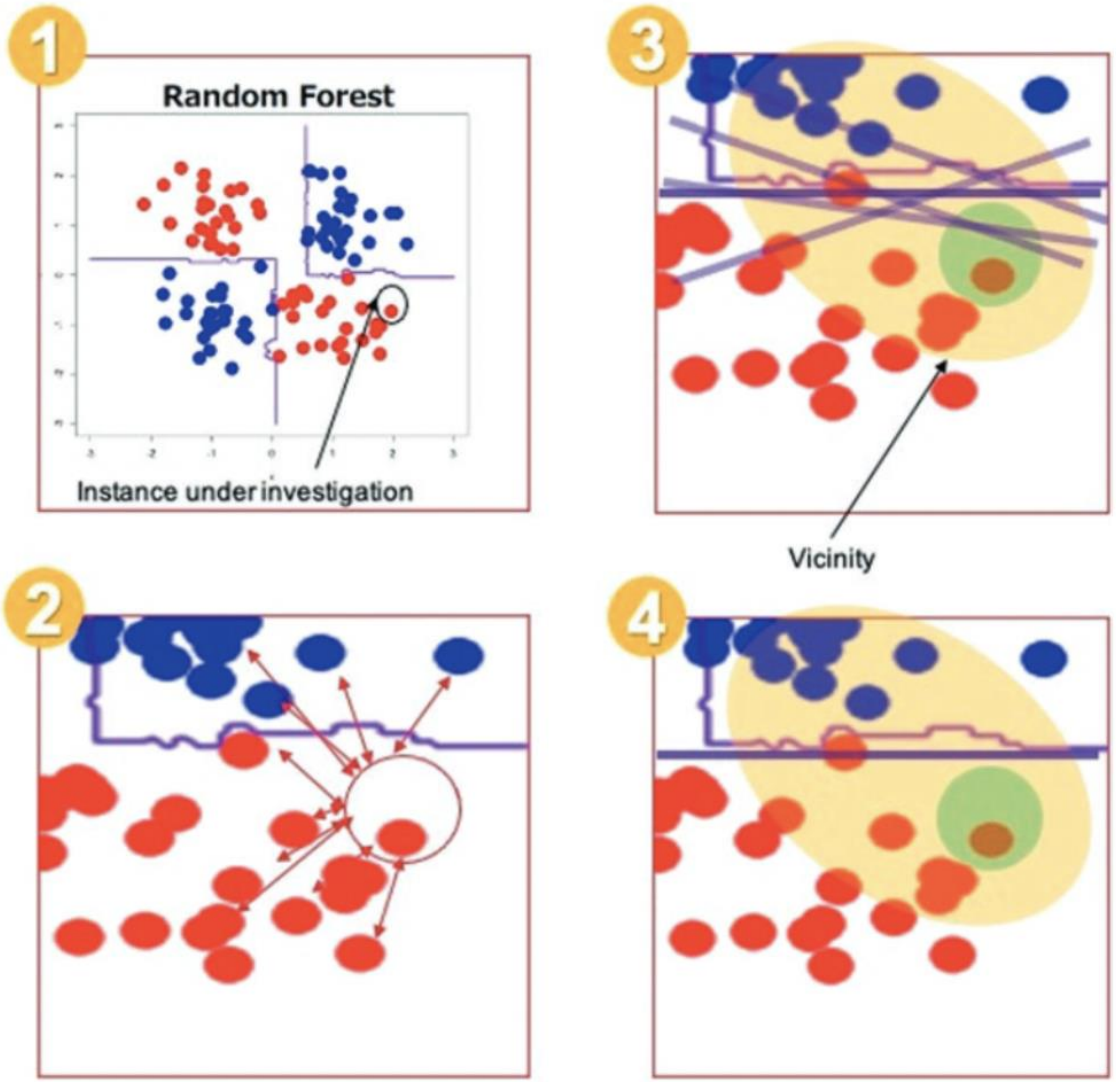
$$g(E[y \mid X]) = \beta_0 + f_1(X_1) + \ldots + f_M(X_N)$$

For the link function for logistic regression

$$\log\left(\frac{P\left(y=1 \mid X\right)}{P\left(y=0 \mid X\right)}\right) = \beta_o + f_1\left(X_1\right) + \ldots + f_M\left(X_N\right)$$

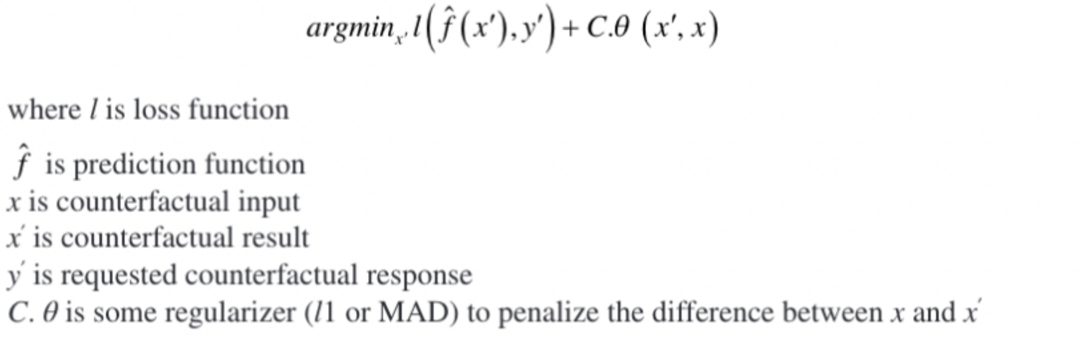


Surrogate Model



Counterfactual

It is a method for countering the fact or reversing the fact or identifying the changes in the features that would change the outcome. Mathematically what change in X will inverse the outcome



$$argmin_{x'}\, l\left(\hat{f}\left(x'\right), y'\right) + C.\theta\ \left(x', x\right)$$

where l is loss function

$\hat{f}$ is prediction function

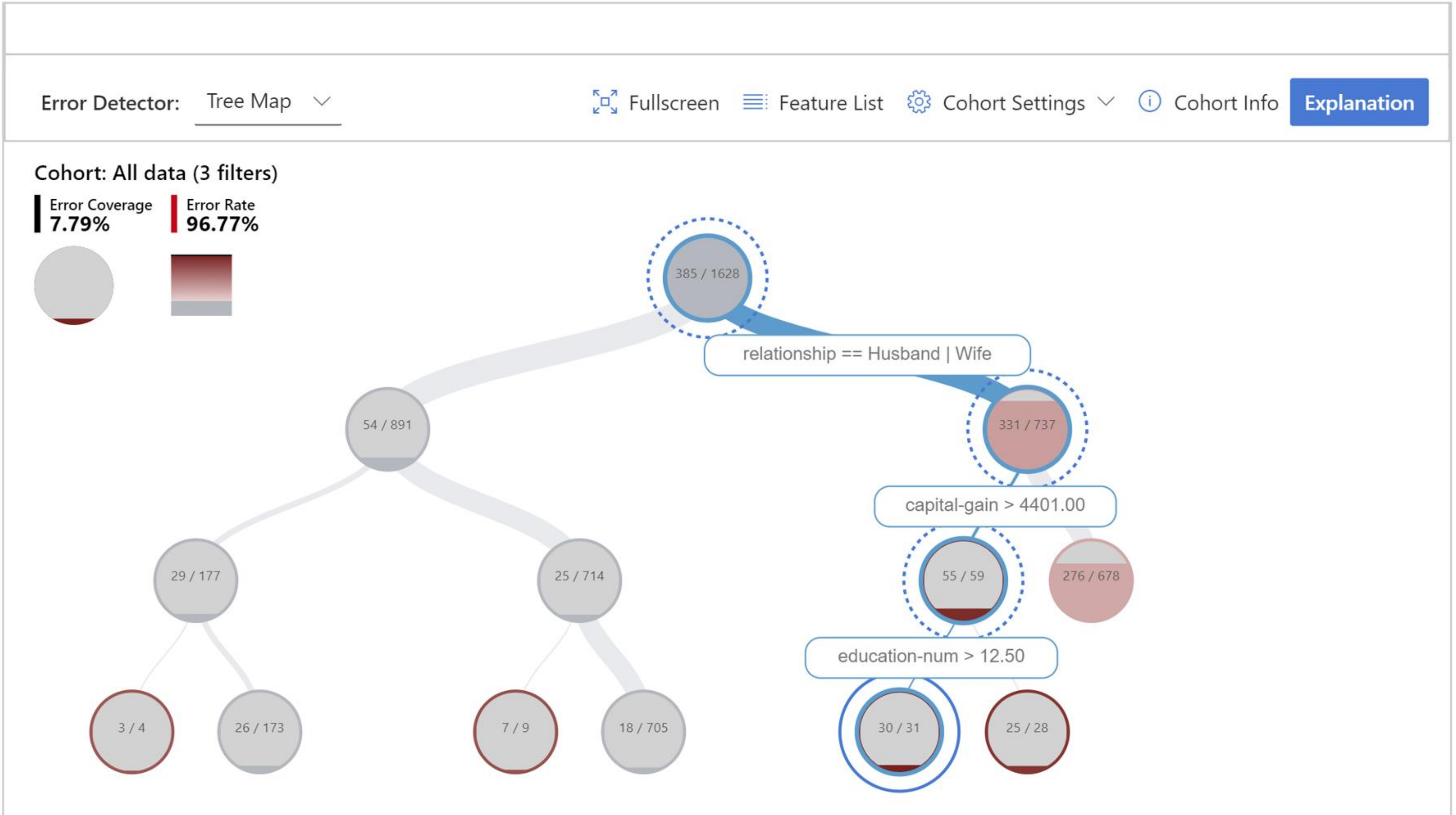
x is counterfactual input

$x^{'}$ is counterfactual result

$y^{'}$ is requested counterfactual response

$C.\,\theta$ is some regularizer ($l1$ or MAD) to penalize the difference between x and $x^{'}$

Error Analysis

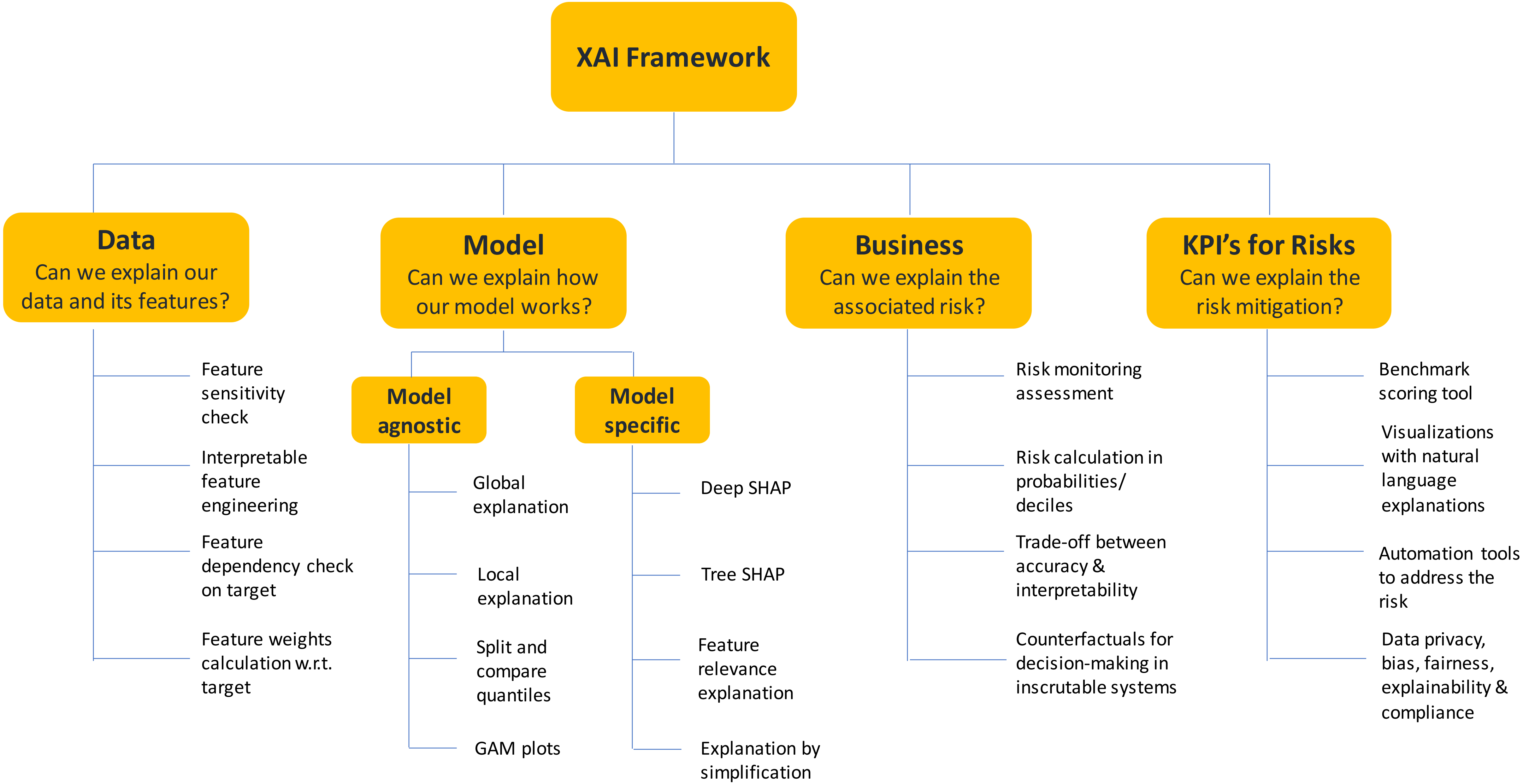


Error Detector: Tree Map

Cohort: All data (3 filters)

Error Coverage 7.79%

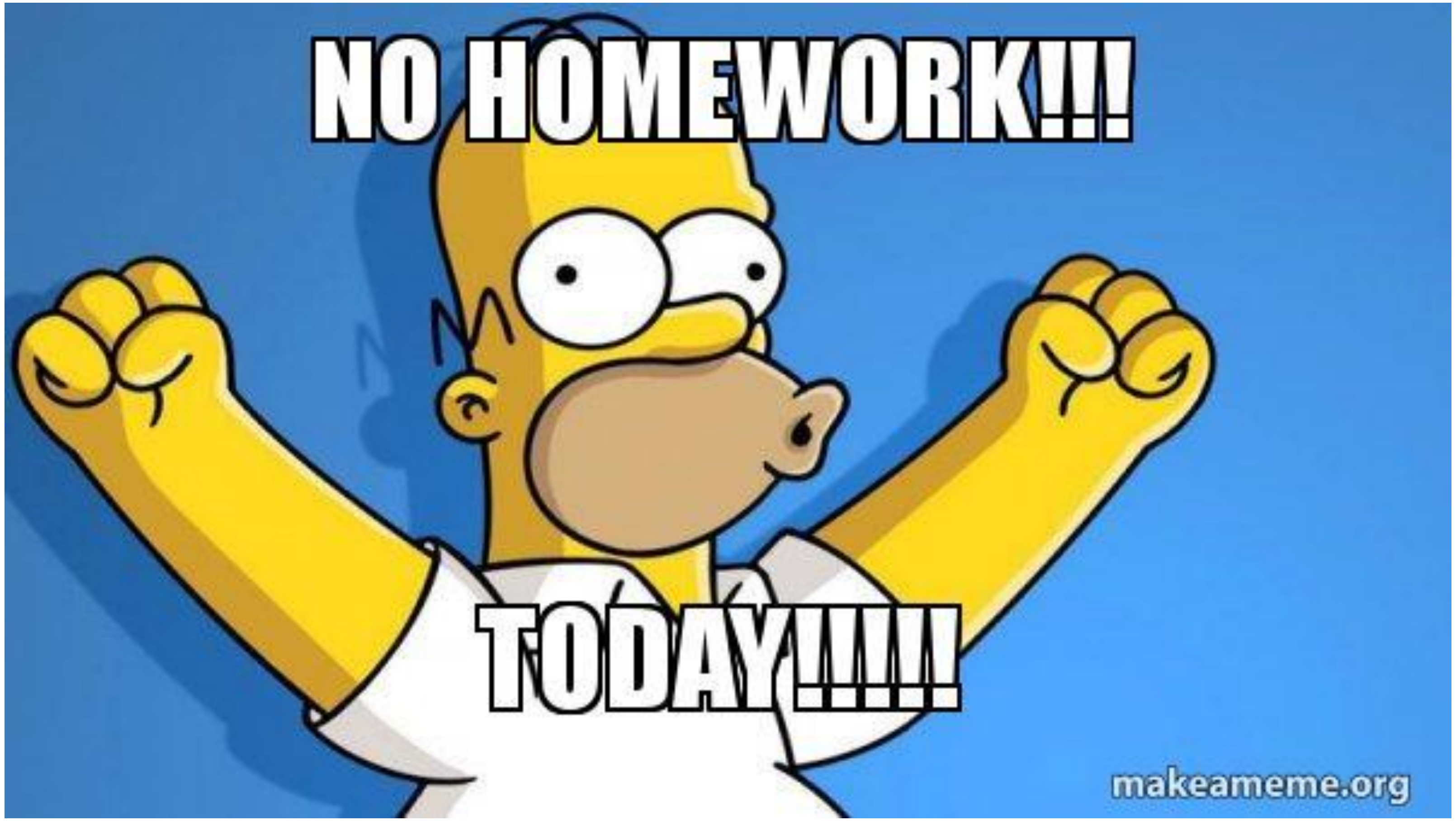
Error Rate 96.77%



XAI Framework

- **Data**
  Can we explain our data and its features?
  - Feature sensitivity check
  - Interpretable feature engineering
  - Feature dependency check on target
  - Feature weights calculation w.r.t. target
- **Model**
  Can we explain how our model works?
  - **Model agnostic**
    - Global explanation
    - Local explanation
    - Split and compare quantiles
    - GAM plots
  - **Model specific**
    - Deep SHAP
    - Tree SHAP
    - Feature relevance explanation
    - Explanation by simplification
- **Business**
  Can we explain the associated risk?
  - Risk monitoring assessment
  - Risk calculation in probabilities/ deciles
  - Trade-off between accuracy & interpretability
  - Counterfactuals for decision-making in inscrutable systems
- **KPI's for Risks**
  Can we explain the risk mitigation?
  - Benchmark scoring tool
  - Visualizations with natural language explanations
  - Automation tools to address the risk
  - Data privacy, bias, fairness, explainability & compliance

Assessment

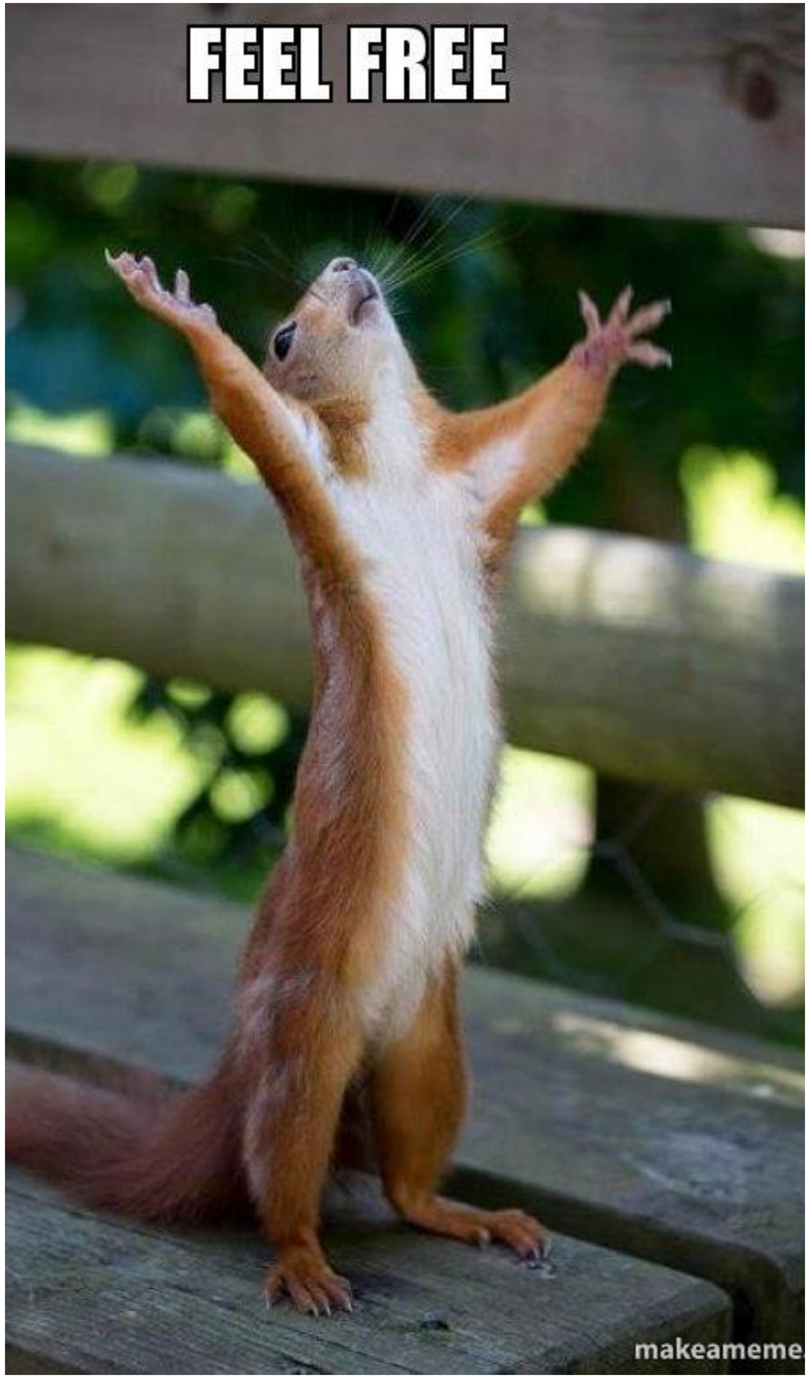


Next

1. Composite feature
2. Additive Counterfactual Fairness (ACF)
3. High level steps for implementing ACF model
4. ACF for classification problems

Revise:

- **Chapters from book**
- **Confusion Matrix**
- **Fairness Metrics**


FEEL FREE
makeameme

YOU CAN CALL ME

DID YOU REALLY JUST SEND THAT EMAIL!

KNOCK, KNOCK!