Interpretable Machine Learning

20th Time Series and Econometrics Meeting (20th ESTE)

André Portela Santos
Department of Quantitative Methods
CUNEF Universidad
sites.google.com/site/andreportela

Download course materials

https://github.com/andreportelasantos/ESTE_2023

Outline

Introduction to ML Model Interpretability Methods

Introduction to ML Model Interpretability Methods

- Why is important to interpret machine leaning (ML) models?
- Some ML models are considered “black boxes”, i.e. it is diffiult to understand how and why the model generates a prediction.
- Interpretability allows humans to understand and trust the predictions made by machine learning models.
- It helps in verifying that models are not using sensitive or inappropriate data features.

Taxonomy of Interpretability Methods

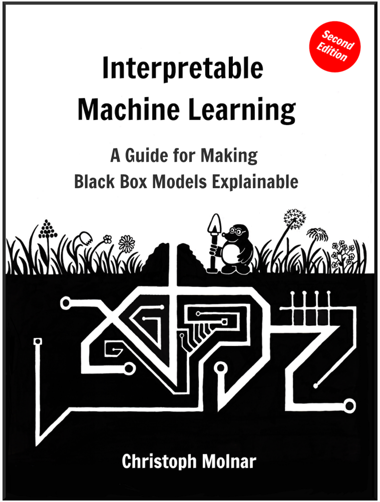
- Interpretability methods can be categorized into **model-specific** and **model-agnostic** methods.
- **Model-specific methods** are tailored to a specific type of model. They take advantage of the specific structure of the machine learning model.
- **Model-agnostic methods** can be applied to any machine learning model. They treat the model as a black box and derive interpretability from its inputs and outputs.

Scope of Interpretability

- Interpretability can be **global (whole model)** or **local (individual predictions)**.
- **Global interpretability** refers to understanding the entire model. It provides a holistic view of the model's decision-making process.
- **Local interpretability** refers to understanding an individual prediction. It provides insights into why the model made a specific prediction.
- **Feature importance** (i.e. the importance of the variables in the model) can be evaluated both globally or locally.

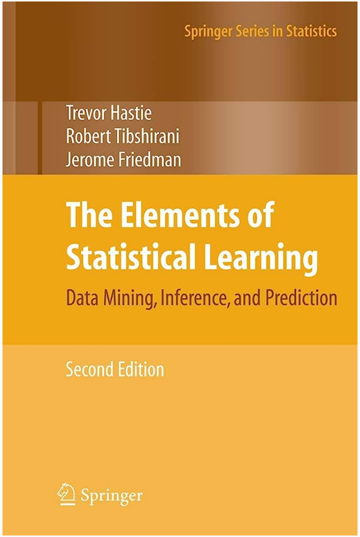
References

References



https://christophm.github.io/interpretable-ml-book/

References


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer

References

- Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). *Definitions, methods, and applications in interpretable machine learning*. **PNAS**, 116(44), 22071-22080.
- Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). *Interpretable machine learning: Fundamental principles and 10 grand challenges*. **Statistic Surveys**, 16, 1-85.
- Lundberg, S. M., & Lee, S. I. (2017). *A unified approach to interpreting model predictions*. **NEURIPS**, 30.

References

- https://medium.com/dataman-in-ai/explain-your-model-with-the-shap-values-bc36aac4de3d
- https://towardsdatascience.com/explain-any-models-with-the-shap-values-use-the-kernelexplainer-79de
- https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
- https://towardsdatascience.com/understanding-model-predictions-with-lime-a582fdff3a3b

Data

Data

Kava, Lucas (2021) 'Além da caixa preta: aprendizagem de máquina interpretável para previsão de séries temporais macroeconômicas brasileiras', Dissertação de Mestrado em Economia, UFSC. `https://repositorio.ufsc.br/handle/123456789/234659`

Crescimento da moeda e politica monetária	4
Depósitos de Poupança	5
Consumo e vendas	11
Vendas reais - varejo	5
Preços	20
Crédito	7
Financeiro e risco	7
Produto e atividade real	26
Fiscal	12
Setor externo	28
	125 series (220 monthly obs.)

Model-specific Interpretability Methods

Model-specific Interpretability

- **Model-specific interpretability** refers to the interpretability of a machine learning model that is inherent to its structure.
- The model's decisions can be understood by examining its structure and parameters.

Model-specific Interpretability

Linear regression model

- Consider the **linear regression model**

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

- **Feature importance** can be easily calculated using the t-statistics:

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{SE\left(\hat{\beta}_j\right)}$$

- An increase of feature x_k by one unit increases the prediction for y by β_k units when all other feature values remain fixed.

Model-specific Interpretability

Elastic net model

- The Elastic Net is a regularized regression method that linearly combines the $L1$ and $L2$ penalties of the Lasso and Ridge methods:

$$\min_{\beta} \frac{1}{2n}||y - X\beta||_2^2 + \lambda((1-\alpha)||\beta||_2^2 + \alpha||\beta||_1) \quad (1)$$

  where y is the target variable, X is the feature matrix, β is the vector of coefficients, λ is the regularization parameter, and α is the mixing parameter between Ridge ($\alpha = 0$) and Lasso ($\alpha = 1$).

- In the Elastic Net model, all variables are normalized or standardized before estimating the model. This means that the **estimated coefficients can be interpreted as measures of variable importance**.

Model-specific Interpretability

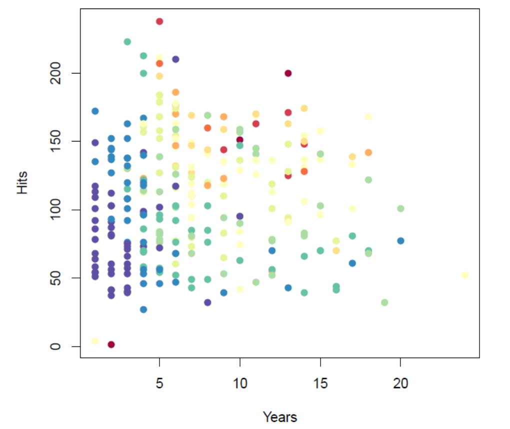
Tree based methods

- Tree based methods split the data multiple times according to certain cutoff values in the features.
- Through splitting, different subsets of the dataset are created, with each instance belonging to one subset.
- The final subsets are called terminal or leaf nodes and the intermediate subsets are called internal nodes or split nodes.
- To predict the outcome in each leaf node, the average outcome of the training data in this node is used.

Tree based methods

Baseball salary data: how would you stratify it?

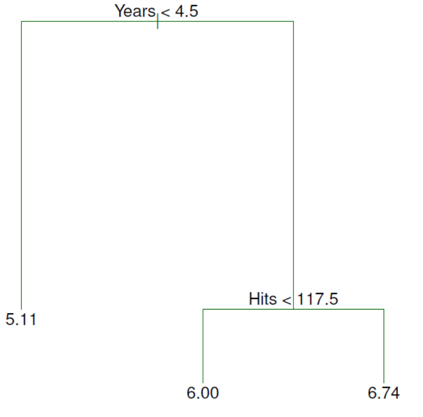
Salary is color-coded from low (blue, green) to high (yellow,red)



Tree based methods

Baseball salary data: how would you stratify it?

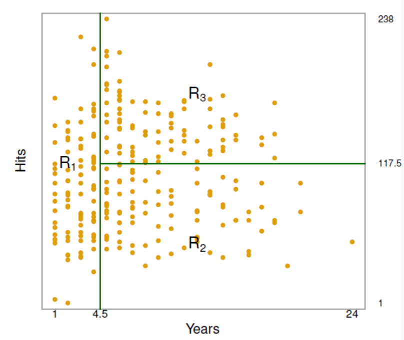
- The split at the top of the tree results in two large branches. The left-hand branch corresponds to Years < 4.5, and the right-hand branch corresponds to Years $>= 4.5$.
- The tree has two internal nodes and three terminal nodes, or leaves. The number in each leaf is the mean of the response for the observations that fall there.



Tree based methods

Baseball salary data: how would you stratify it?

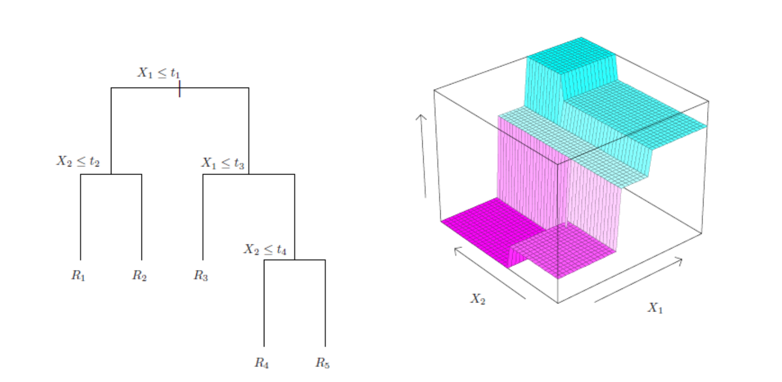
The tree stratifies or segments the players into three regions of predictor space: $R_1 = \{\mathrm{X} \mid \text{Years} < 4.5\}$, $R_2 = \{X \mid \text{Years} >= 4.5, \text{Hits} < 117.5\}$, and $R_3 = \{X \mid \text{Years} >= 4.5, \text{Hits} >= 117.5\}$.



Tree based methods

Predictions

We predict the response for a given test observation using the mean of the training observations in the region to which that test observation belongs.



Tree based methods

Feature importance

- **Feature importance** can be computed in the following way: Go through all the splits for which the feature was used and measure how much it has reduced the mean squared error (MSE) compared to the parent node. The sum of all importances is scaled to 100.
- This means that each importance can be interpreted as share of the overall model importance.

Tree based methods

Boosted trees i

- Boosted trees are grown sequentially: each tree is grown using information from previously grown trees.
- Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training set.
- For $b = 1, 2, \ldots, B$, repeat:
  1. Fit a tree $\hat{f}^b$ with d splits ($d + 1$ terminal nodes) to the training data.
  2. Update $\hat{f}$ by adding in a shrunken version of the new tree:

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$$

Tree based methods

Boosted trees ii

3. Update the residuals,

$$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i)$$

- Output the boosted model,

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x)$$

Model-agnostic Interpretability Methods

Model-agnostic Interpretability

- **Agnostic interpretability** refers to interpretability methods that can be applied to any machine learning model.
- These methods are **not dependent on the internal structure of the model**, and they can be used to interpret the model's predictions after it has been trained.
- Examples include **partial dependence plots**, **permutation feature importance**, **surrogate models**, **and Shapley values**.

- Partial Dependence Plots (PDP) are used to visualize the **marginal effect of one or two features on the predicted outcome** of a machine learning model.
- They are calculated by **averaging the predictions of a model after setting a feature to a certain value**.
- For a single feature, a PDP shows the change in the average prediction as the feature value changes. For two features, a PDP shows the change in the average prediction as the feature values change together.

Model-agnostic Interpretability

Partial Dependence Plots (PDP) ii

- The PDP for a feature x_S is defined as:

$$\hat{f}_S(x_S) = E\left[\hat{f}(x_S, X_C)\right] = \int \hat{f}(x_S, X_C)\, d\mathbb{P}(X_C)$$

where x_S are the features for which the partial dependence function should be plotted and X_C are the other features used in the machine learning model $\hat{f}$.

Model-agnostic Interpretability

Partial Dependence Plots (PDP) iii

- The empirical version is:

$$\hat{f}_S(x_S) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_S, x_C^{(i)}) \tag{2}$$

  where $\hat{f}$ is the model, x_S is the set of features for which the PDP is computed, $x_C^{(i)}$ is the complement of x_S for the i-th observation, and n is the total number of observations.
- Partial dependence works by **marginalizing the machine learning model output** over the distribution of the features in set C.

Permutation Feature Importance i

- **Permutation feature importance** measures the increase in the prediction error of the model after we permute the feature's values, which breaks the relationship between the feature and the true outcome.
- A feature is "important" if shuffling its values increases the model error, because in this case the model relied on the feature for the prediction.

Model-agnostic Interpretability

Permutation Feature Importance ii

- The permutation feature importance algorithm:
  1. Estimate the original model error $e_{orig} = L(y, \hat{f}(X))$ (e.g. mean squared error)
  2. For each feature $j \in \{1, ..., p\}$ do:
     - Generate feature matrix X_{perm} by permuting feature j in the data X. This breaks the association between feature j and true outcome y.
     - Estimate error $e_{perm} = L(Y, \hat{f}(X_{perm}))$ based on the predictions of the permuted data.
     - Calculate permutation feature importance as quotient $FI_j = e_{perm}/e_{orig}$ or difference $FI_j = e_{perm} - e_{orig}$
  3. Sort features by descending FI.

Model-agnostic Interpretability

Permutation Feature Importance iii

- When dealing with time series data, using permutation feature importance may lead to misleading results.
- When permuting a feature in time series data, the temporal order and dependencies of the data is lost.
- Instead of randomly permuting the values of the feature across all data points, you could permute the values within a certain **sliding window**.

Model-agnostic Interpretability

Surrogate Models i

- A **surrogate model** is an interpretable model that is trained to approximate the predictions of a black box model. We can draw conclusions about the black box model by interpreting the surrogate model.
- The purpose is to approximate the predictions of the underlying model as accurately as possible and to be interpretable at the same time.

Model-agnostic Interpretability

Surrogate Models ii

- For example, a linear model:

$$g(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Or a decision tree:

$$g(x) = \sum_{m=1}^{M} c_m I\{x \in R_m\}$$

can be use the approximate the predictions of a black-box model $g(x)$.

- A prediction can be explained by assuming that each feature value is a "player" in a game where the prediction is the "payout".
- **Shapley values** tells us how to fairly distribute the "payout" among the features.

Model-agnostic Interpretability

Shapley Values ii

- Consider the linear model:

$$\hat{f}(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

where each x_j is a feature value, with $\mathrm{j} = 1, \ldots, p$.

- The contribution ϕ_j of the j-th feature on the prediction $\hat{f}(x)$ is:

$$\phi_j(\hat{f}) = \beta_j x_j - E\left(\beta_j X_j\right) = \beta_j x_j - \beta_j E\left(X_j\right)$$

where $E\left(\beta_j X_j\right)$ is the mean effect estimate for feature j.

Model-agnostic Interpretability

Shapley Values iii

- The contribution, i.e. the **Shapley value**, is the difference between the feature effect minus the average effect.
- Shapley values are often estimated as:

$$\hat{\phi}_j = \frac{1}{M} \sum_{m=1}^{M} \left(\hat{f}\left(x_{+j}^m\right) - \hat{f}\left(x_{-j}^m\right) \right)$$

where $\hat{f}\left(x_{+j}^m\right)$ is the prediction for x, but with a random number of feature values replaced by feature values from a random data point z, except for the respective value of feature j.

Result: Shapley value for the value of the j-th feature

Required: Number of iterations M, instance of interest x, feature index j, data matrix X, and machine learning model f

for $\mathrm{m} = 1, \ldots, \mathrm{M}$ **do**

- Draw random instance z from the data matrix X
- Choose a random permutation o of the feature values
- Order instance x : $x_o = \left(x_{(1)}, \ldots, x_{(j)}, \ldots, x_{(p)}\right)$
- Order instance z: $z_o = \left(z_{(1)}, \ldots, z_{(j)}, \ldots, z_{(p)}\right)$
- Construct two new instances
- With j: $x_{+j} = \left(x_{(1)}, \ldots, x_{(j-1)}, x_{(j)}, z_{(j+1)}, \ldots, z_{(p)}\right)$
- Without j : $x_{-j} = \left(x_{(1)}, \ldots, x_{(j-1)}, z_{(j)}, z_{(j+1)}, \ldots, z_{(p)}\right)$
- Compute marginal contribution: $\phi_j^m = \hat{f}\left(x_{+j}\right) - \hat{f}\left(x_{-j}\right)$
- Compute Shapley value as the average: $\phi_j(x) = \frac{1}{M} \sum_{m=1}^{M} \phi_j^m$

end

Shapley Values V