NOVA IMS

Information Management School

8

MODELS INTERPRETATION

Machine Learning for Marketing

Acreditações e Certificações

Instituto Superior de Estatística e Gestão da Informação
Universidade Nova de Lisboa

Summary

1. Introduction
2. Global explanation
3. Local explanation
4. Application exercise

8.1

Introduction

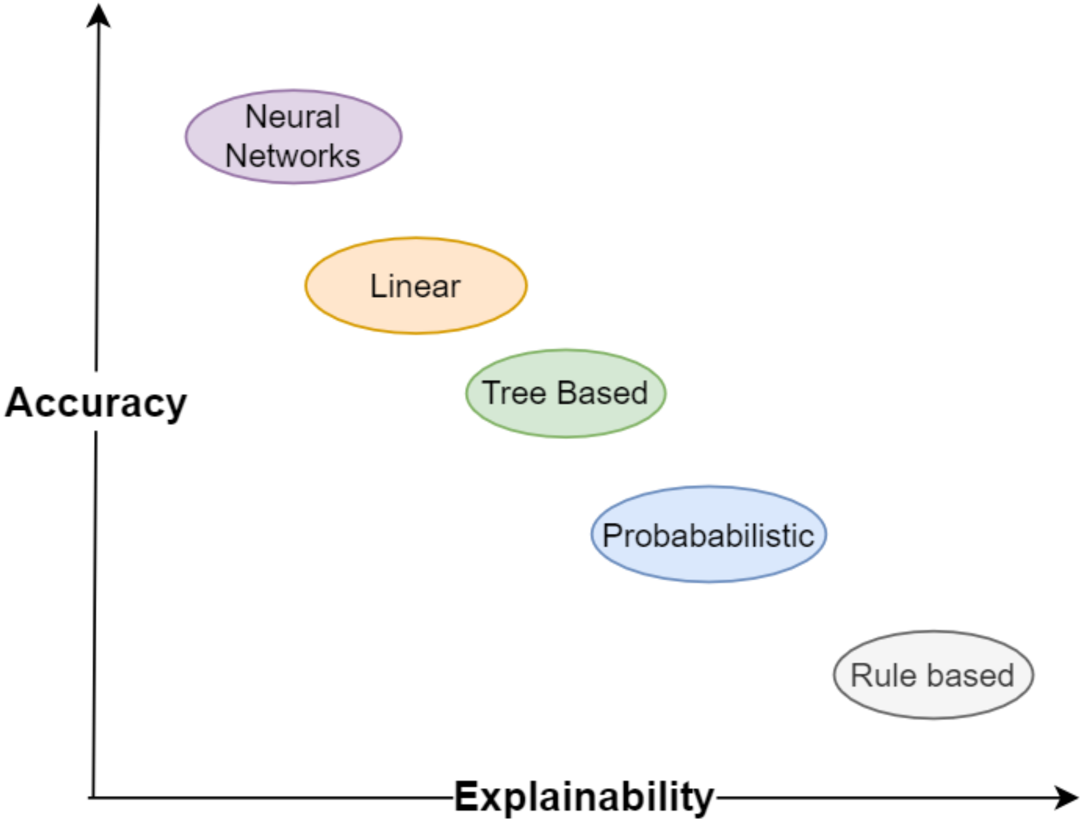
Ensembles of methods

Defining interpretability

- Lack of consensus on what is and how to evaluate interpretability
- One accepted definition is: *The ability to explain or to present in understandable human terms*
- However, no clear answers in psychology to:
  - What composes an explanation
  - What make some explanations better than others
  - When should explanations be sought

“Machine learning has great potential for improving products, processes and research. But computers usually do not explain their **predictions which is a barrier to the adoption of machine learning.** ”

[Christoph Molnar, 2019]



source: www.towardsdatascience.com

Motivations for interpretability

- **Trust:** It is easier for humans to trust a system that explains its decisions compared to a black box (removes the barriers for the adoption of machine learning models)
- **Fairness** or **Nondiscrimination:** Ensuring that predictions are unbiased and do not implicitly or explicitly discriminate against protected groups (e.g., racial bias)
- **Privacy** and **Safety**: Ensuring that sensitive information in the data is protected (e.g., right to explanation under GDPR)
- **Reliability** or **Robustness:** Ensuring that small changes in the input do not lead to large changes in the prediction
- **Knowledge obtention**: accuracy is no longer enough – understanding predictors is sometimes more important than predictions
- **Causality**: Check that only causal relationships are picked up

Importance of interpretability

> “The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks.”
>
> (Doshi-Velez and Kim, 2017)

- We need to more than the **what** is predicted (e.g., the probability that a customer will churn)
- We want also to know **why** (e.g., understanding the reasons for churn may be more useful for the business problem, than knowing the probability of each customer to churn)

Not all models require interpretability

- Models that require no human intervention (e.g., ad servers, postal code sorting)
- Models who have no or low consequences for unacceptable results (e.g., movie recommendation model)
- Models that are well studied and validated in real-world applications

Approaches

- **Interpretable models:**

Algorithm	Linear	Monotone	Task
Linear regression	Y	Y	Reg.
Logistic regression	N	Y	Class.
Decision trees	N	Some	Reg. and Class.
RuleFit	Y	N	Reg. and Class.
Naïve Bayes	N	Y	Class.
K-Nearest Neighbors	N	N	Reg. and Class.

- **Model agnostic interpretation tools:** tools that can be applied to any supervised machine learning model:
  - Global: explain the model behavior across all data instances
  - Local: explain the model behavior based on each data point prediction

8.2

Global explanation

Ensembles of methods

Global interpretability

Holistic

- When the entire model can be comprehended at once
- To interpret the model, one needs to train the model, knowledge of the algorithm and the data
- Requires a holistic view of the model's features and learned components (weights, parameters, and structures)
- For this reason, global model holistic interpretability is hard to obtain (even a linear regression with 5 variables is hard for humans)

Modular level

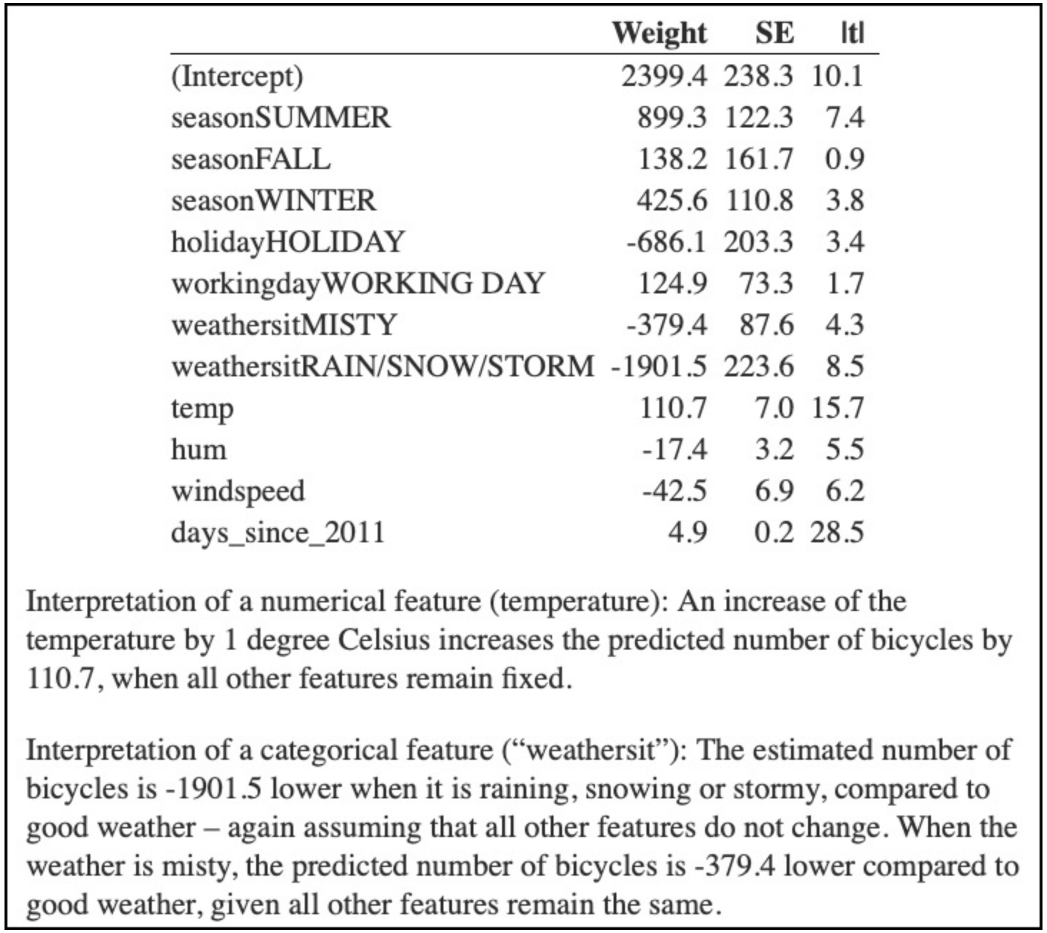
- Interpreting the model from one of the parameters can still give a good explanation of how the model reaches predictions (e.g., weights in linear regression)

Commonly used techniques

Global explanation

Linear regression interpretability

- Numerical feature: increasing the numerical feature by one unit changes the estimated outcome by its weight
- Binary feature: changing the feature to the referenced category changes the estimated outcome by the feature's weight

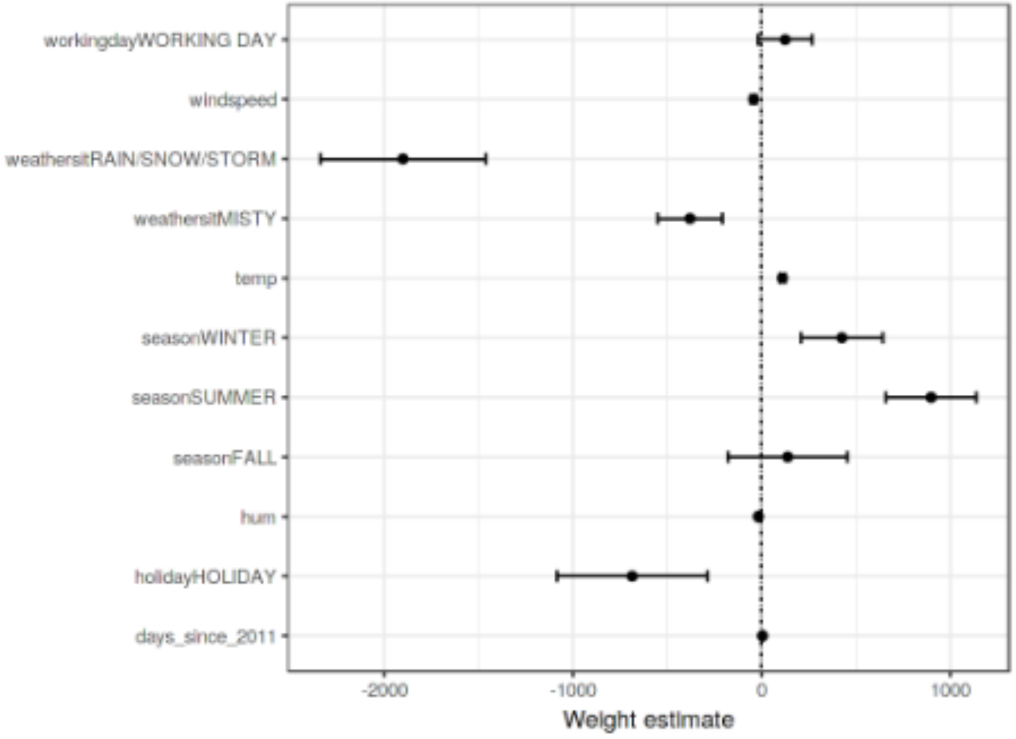


	Weight	SE	|t|
(Intercept)	2399.4	238.3	10.1
seasonSUMMER	899.3	122.3	7.4
seasonFALL	138.2	161.7	0.9
seasonWINTER	425.6	110.8	3.8
holidayHOLIDAY	-686.1	203.3	3.4
workingdayWORKING DAY	124.9	73.3	1.7
weathersitMISTY	-379.4	87.6	4.3
weathersitRAIN/SNOW/STORM	-1901.5	223.6	8.5
temp	110.7	7.0	15.7
hum	-17.4	3.2	5.5
windspeed	-42.5	6.9	6.2
days_since_2011	4.9	0.2	28.5

Interpretation of a numerical feature (temperature): An increase of the temperature by 1 degree Celsius increases the predicted number of bicycles by 110.7, when all other features remain fixed.

Interpretation of a categorical feature ("weathersit"): The estimated number of bicycles is -1901.5 lower when it is raining, snowing or stormy, compared to good weather – again assuming that all other features do not change. When the weather is misty, the predicted number of bicycles is -379.4 lower compared to good weather, given all other features remain the same.

Linear regression – Weight plot



Weights are displayed as points and the 95% confidence intervals as lines.

Linear regression interpretability

Advantages

- Transparency how predictions are produced
- Widely accepted
- Guaranteed to find optimal weights (given all assumptions are met by the data)
- Based on solid statistical theory

Disadvantages

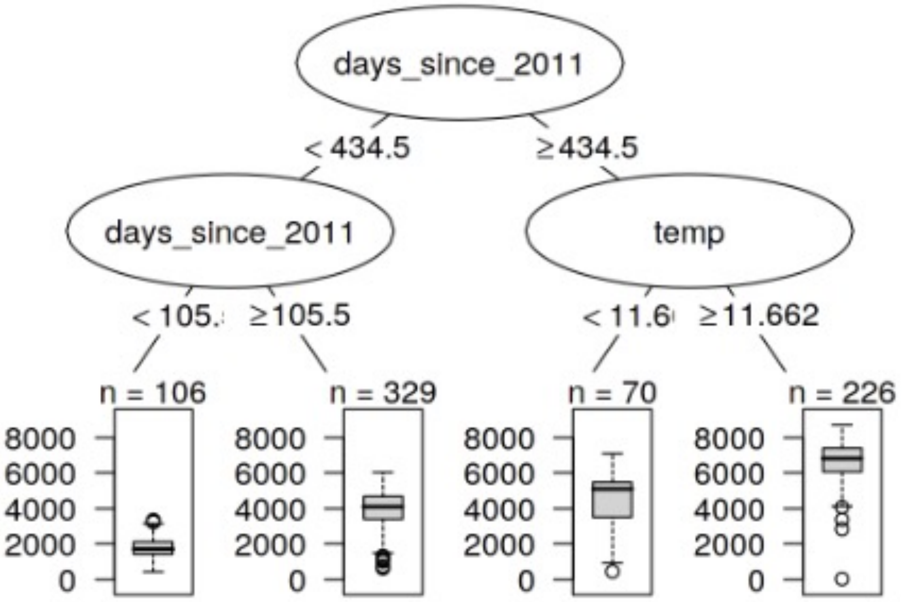
- Only capture linear relationships
- Models tend not to have the best predictive performance
- The interpretation of a weight can be unintuitive

Logistic regression

- Numerical feature: increasing the numerical feature by one unit changes the estimated odds change by a factor of $\exp(\beta_j)$
- Binary feature: changing the feature to the referenced category changes the estimated estimated odds change by a factor of $\exp(\beta_j)$
- The advantages and disadvantages are similar to those of Linear regression

Decision trees

- Interpretation is simple
- The overall importance of a feature can be computed by going through all the splits a feature was sued and measure how much it has reduced the variance, or the Gini index compared to the parent node (the sum of all importances is scored to 100)



Regression tree fitted on the bike rental data. The maximum allowed depth for the tree was set to 2. The trend feature (days since 2011) and the temperature (temp) have been selected for the splits. The boxplots show the distribution of bicycle counts in the terminal node.

Decision trees interpretability

Advantages

- Ideal to capture interactions between features
- The data ends in distinct groups that are easy to understand
- The tree structure automatically invites to think about predicted values

Disadvantages

- Trees fail to deal with linear relationships
- Lack of smoothness (small changes in values may end up in a different branch
- Quite unstable (some changes in the inputs and the tree changes completely)
- Long trees are hard to interpret

8.3

Local explanation

Ensembles of methods

Local interpretability

For a single prediction

- Zooming in on a single instance can expose how the model reached the prediction
- Locally the prediction may depend only on some some features

For a group of predictions

- By understanding the patterns in the instances that are part of the group

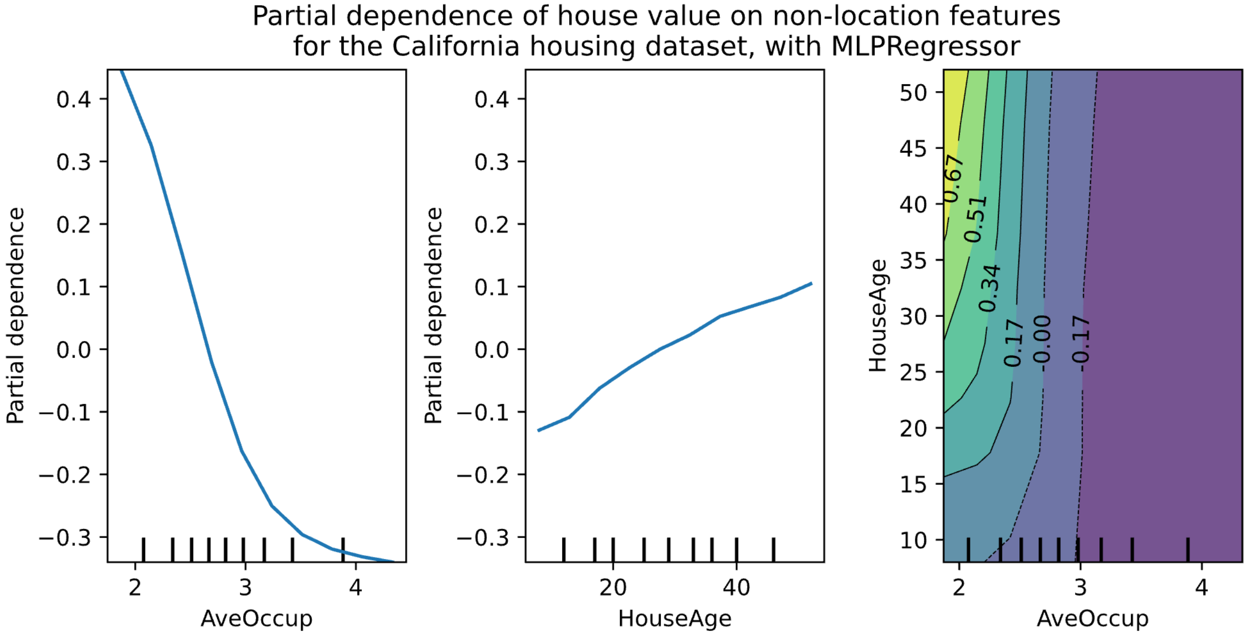
Commonly used techniques

Local explanation

Partial Dependence Plot (PDP)

- Show the dependence between the target response and a set of features
- Works well with linear, monotonous or complex relationships between features and target
- Assume the input features are independent (no correlation)
- Computed has the average prediction, if we force all data points to assume that feature value
- Due to the limits of human perception the size of the set of input feature of interest must be small (usually, one or two)

PDP visualization example



- Dependence of the median house price on the joint values of average occupants per household and house age
- For an average occupancy greater than the house price is nearly independent of the house age, whereas for average occupation less than 2 there is a strong dependence on age

NOVA IMS Information Management School

PDP advantages vs disadvantages

Advantages

- The computation is intuitive
- Easy to implement
- It has a causal interpretation. We intervene on a feature and measure the changes in the predictions

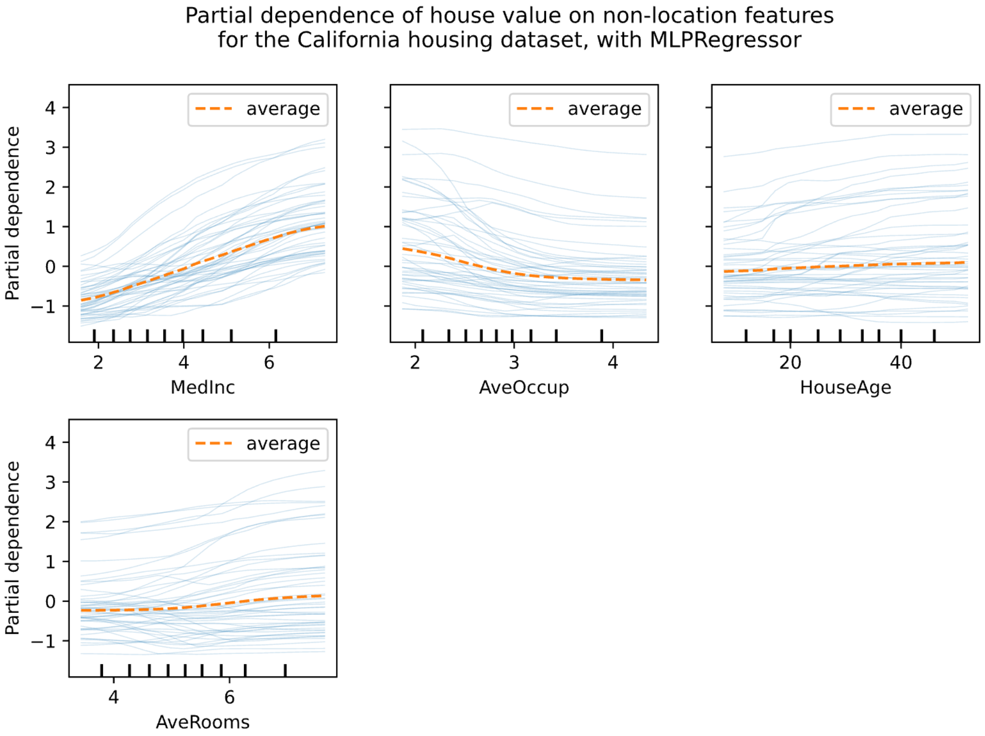
Disadvantages

- Humans have problem understanding more than 3 features at the same time
- The distribution is not shown. The interpretation may be misleading, has one may overinterpret regions with almost no data
- The assumption of independence among features

Individual Conditional Expectation (ICE)

- Similar to PDP, but instead of showing the average effect of the input feature, an ICE plot shows the dependence on a feature for each instance separately, with one line per sample
- Each line shows how the instance’s prediction changes when a feature changes
- Due to the limits of human perception, only one input feature is shown at a time

ICE visualization example



- Although there is a linear relationship dependence between median income and the house price, the lines show that there are exceptions

ICE advantages vs disadvantages

Advantages

- Very intuitive
- Easy to implement
- Unlike PDP, ICE curves can uncover heterogeneous relationships

Disadvantages

- Can only display one feature at a time
- Can also be impacted by features' correlation
- If many curves are drawn, the plot can become overcrowded
- In some ICE plots it can be difficult to see the average

Shapley values

- A prediction can be explained by assuming that each feature value of the instance is a "player" in a game where the prediction is the payout. The Shapley value (Shapley, 1953) – a method from coalitional game theory – tells us how to fairly distribute the "payout" among the features." (Christoph Molnar, 2019)
- Example: The average prediction for all customers' Lifetime Value is € 10,000. How much has each feature value contributed to the prediction compared to the average prediction?
  - For linear regression: The effect of each feature is the weight of the feature times the feature value
  - For others:
    - Game: prediction task
    - Gain: actual prediction for the instance minus the average prediction for all instances
    - Player: the feature value of the instance
    - Shapley value: the average marginal contribution of a feature value across all possible coalitions

Shapley values

Advantages

- The difference between the prediction and the average prediction is fairly distributed among the feature values of the instance
- Allows contrastive explanations (e.g., compare with a subset)
- Based on solid theory

Disadvantages

- Can require computational power
- Can be misinterpreted
- Uses all the features, which can increase the complexity of the interpretation
- Access to data is required to calculate the Shapley values

8.2

Application exercise

Ensembles of methods

Predicting customers who will leave the bank in the following 6 months

1. Copy from the datasets folder the dataset "Bank_Churn_Modelling.csv"
2. Copy and open the Jupyter notebook "PredictBankChurn_XGB_Shap.ipynb"
3. Follow the presentation of the notebook, answer the questions and explore the challenges

Questions?

Some content adapted from *Interpretable Machine Learning* (Christoph Molnar , 2019)

Machine Learning for Marketing

Acreditações e Certificações

Instituto Superior de Estatística e Gestão da Informação
Universidade Nova de Lisboa