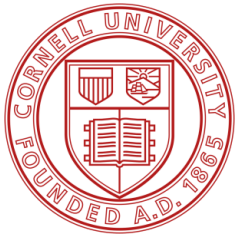


Interpretable Machine Learning: Shapley Values

Prof. Marcos López de Prado

Advances in Financial Machine Learning

ORIE 5256

Key Points

- Machine learning (ML) algorithms utilize the power of computers to solve tasks that are beyond the grasp of classical statistical methods
- However, ML is often perceived as a black-box, hindering its adoption
- **In this presentation, we demonstrate the use of Shapley values to interpret the outputs of ML models**
- Shapley values interpret a model's prediction in terms of
  - attribution to the various features (sign and size)
  - features that are most important overall
  - dependence on the feature's value
  - interactions between features
  - similarity (supervised clustering)
- With the help of interpretability methods, ML is becoming the primary tool of scientific discovery, through induction as well as abduction

The Role of ML in Modern Science

Induction & Abduction Through ML

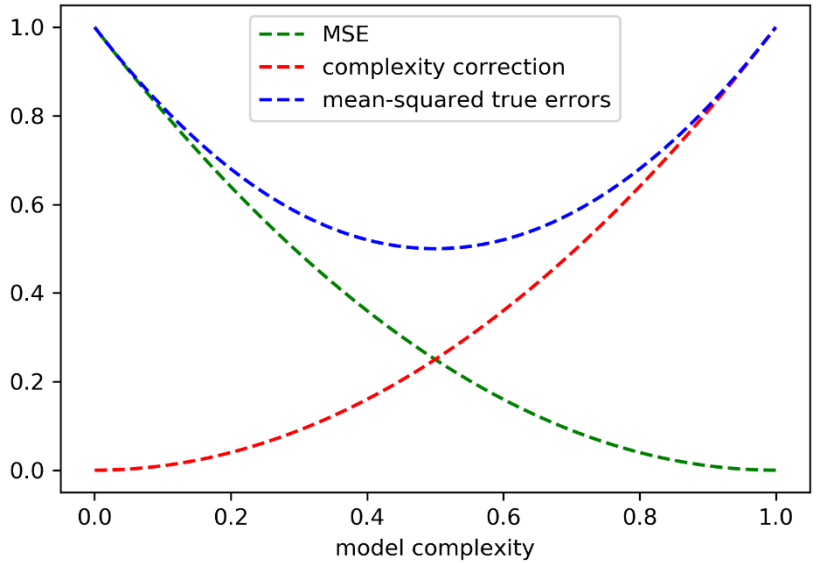
- Scientists discover through induction and abduction
- **Induction**: From cause (X) and effect (y), derive a mechanism (f) such that $y = f[X]$
  - Classical statistics primarily learns by induction
- **Abduction**: From effect (y) and mechanism (f), derive a *plausible* cause (X), such that $y = f[X]$
  - ML is particularly helpful at abduction, because it isolates the causes possibly involved in an effect
- Abduction requires interpretability of f
  - We can only point to plausible causes if we understand the mechanism (no black-box)

Reasoning	Description
Abduction	Researchers propose plausible causes of an effect
Induction	Researchers propose, test, and validate a clear cause-effect mechanism
Deduction	In presence of a cause, an effect is deduced ($A \rightarrow B$). The absence of an effect implies the absence of its cause (not $B \rightarrow$ not A)

Unlike classical statistics, ML decouples the search for X from the search for f, because the researcher finds the X predictive of y without imposing a f. ML's decoupling enables learning X by **abduction**. Once X is fixed, we can learn by **induction** what mechanism f is supported by empirical evidence. Finally, for a given $f[X]$, we can postulate future values of y by **deduction**.

Performance vs. Interpretability Dilemma

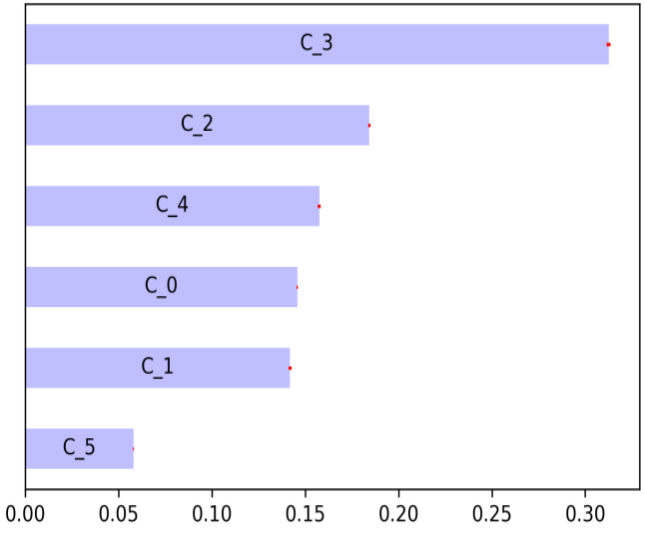
- A model is interpretable when we can understand its decisions
  - Given a model's prediction, we can abduct a likely cause for that predicted effect
- Linear models are popular because they are intrinsically interpretable
  - An effect is the weighted sum of individual causes
- Dilemma:
  - linear models are intrinsically interpretable, but perform poorly
  - nonlinear models are powerful, but not intrinsically interpretable
- Solution: Use approaches that make ML models interpretable (post hoc)



Due to their simplicity, classical statistical models are intrinsically interpretable. However, that interpretability comes at the cost of poor performance, because linear models rarely achieve the minimum mean squared error (MSE).

Taxonomies of ML Interpretability Approaches

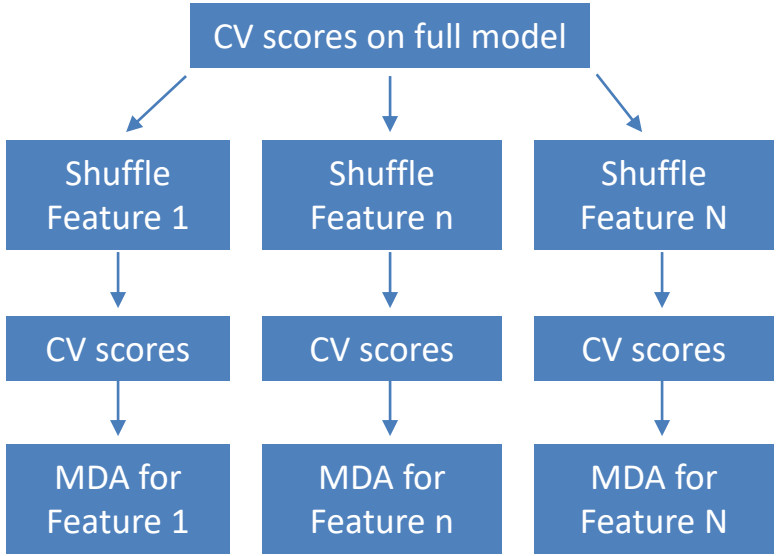
- Types of interpretability
  - **Intrinsic**: The model is interpretable as the result of restricting its complexity (e.g., linear models)
  - **Post hoc**: The model becomes interpretable after applying methods that explain its behavior (e.g., MDA)
- Types of interpretability methods
  - **Specific**: Applicable to certain classes of models (e.g., MDI)
  - **Agnostic**: Applicable to all classes of models (e.g., MDA)
- Interpreted output
  - **Global**: Importance across all observations (e.g., MDI, MDA)
  - **Local**: Importance for a given observation (e.g., LIME)
- Estimation
  - **In-sample**: Importance derived from train-set errors (e.g., MDI)
  - **Cross-validated**: Importance derived from test-set errors (e.g., MDA)



Mean decreased impurity (MDI) is an example of a specific approach (it can only be derived for tree-based models). A feature's importance is measured by its contribution to global impurity reduction at node splits (in-sample).

Example: Permutation Importance (MDA)

- Permutation importance compares the cross-validated performance of a model fit on X, with the performance of the same model after shuffling a particular feature
  - If the feature is important, we should observe a substantial reduction in cross-validated performance
- Advantages
  - Post-hoc, agnostic, global, cross-validated
  - Consistent: if the model changes such that it relies more on a feature, its importance will raise
- Disadvantages
  - No local interpretability: the sign of the effect may change locally, and the approach will not show it
  - Purely experimental: limited theoretical properties

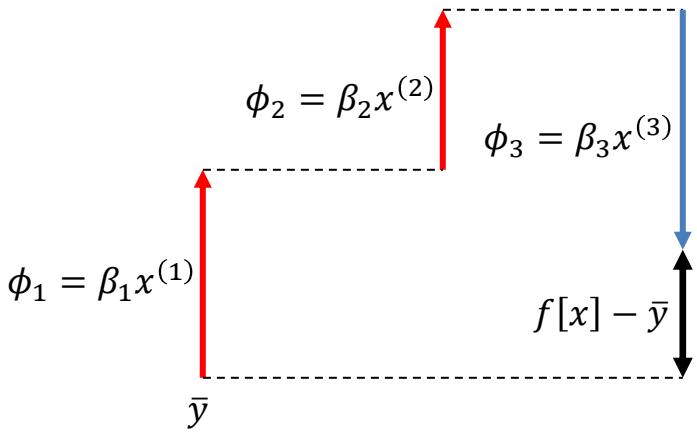


We can compute K losses in performance in a K-fold cross-validation. That allows us to bootstrap the distribution of the generalization error. Next, we will study how **Shapley values** overcome some of MDA's limitations.

Shapley Values

The Problem

- Consider a dataset with
  - N features, $X = \{X^{(1)}, \dots, X^{(N)}\}$, and
  - a real-valued target variable, y
- Given an instance $x \in X$, a model f forecasts the target as $f[x]$
- Question: Can we attribute the departure of a prediction $f[x]$ from its average value $\bar{y}$ in terms of the observed instance x?
- In a linear model, the answer is trivial
  - $f[x] - \bar{y} = \beta_1 x^{(1)} + \cdots + \beta_N x^{(N)}$
- We would like to do a similar *local* decomposition for any (nonlinear) ML model



An attribution of the departure of a model's prediction, $f[x]$, from its average value, $\bar{y}$. Because addition is commutative, the sequence of the effects does not alter the result. Hence, in a linear model, the attribution is invariant to the sequence of effects.

Coalitional Game Theory (1/2)

- Shapley values answer this attribution problem through game theory
  - Features are treated as players in a game
  - The players form coalitions in order to achieve an outcome
  - We wish to find a "fair" attribution of each individual player's contribution
- First, we consider the 2^N possible coalitions (interactions) between the players (features), where
  - some players (features) may not participate (remain at their average value), and
  - some players (features) may participate (depart from their average value)

$x^{(1)} \neq \bar{x}^{(1)}$	$x^{(2)} \neq \bar{x}^{(2)}$	$x^{(3)} \neq \bar{x}^{(3)}$	$f[x\|\ldots] - \bar{y}$
0	0	0	$f[x\|000] - \bar{y}$
0	0	1	$f[x\|001] - \bar{y}$
0	1	0	$f[x\|010] - \bar{y}$
0	1	1	$f[x\|011] - \bar{y}$
1	0	0	$f[x\|100] - \bar{y}$
1	0	1	$f[x\|101] - \bar{y}$
1	1	0	$f[x\|110] - \bar{y}$
1	1	1	$f[x\|111] - \bar{y}$

In a model with 3 features, there are $2^3 = 8$ possible interactions. For each interaction, we compute the departure of the model's forecast from its baseline (average value). We encode as "1" a feature that is not at its average value (it forms part of a coalition), and "0" a feature that is at its average value.

Coalitional Game Theory (2/2)

- Second, we use the coalitions (interactions) table to compute the marginal contribution of each player (feature) conditional to the other players
  - The marginal impact of changing $x^{(i)}$ after changing $x^{(j)}$ may differ from the marginal impact of changing $x^{(j)}$ after changing $x^{(i)}$
  - Accordingly, we must account for all possible $N!$ sequences of conditional effects
- Third, the Shapley value of a player (feature) is the average conditional marginal contribution of that player across all the possible $N!$ ways of conditioning $f[.]$

Sequence	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$
$x^{(1)}, x^{(2)}, x^{(3)}$	$f[x\|100] - f[x\|000]$	$f[x\|110] - f[x\|100]$	$f[x\|111] - f[x\|110]$
$x^{(1)}, x^{(3)}, x^{(2)}$	$f[x\|100] - f[x\|000]$	$f[x\|111] - f[x\|101]$	$f[x\|101] - f[x\|100]$
$x^{(2)}, x^{(1)}, x^{(3)}$	$f[x\|110] - f[x\|010]$	$f[x\|010] - f[x\|000]$	$f[x\|111] - f[x\|110]$
$x^{(2)}, x^{(3)}, x^{(1)}$	$f[x\|111] - f[x\|011]$	$f[x\|010] - f[x\|000]$	$f[x\|011] - f[x\|010]$
$x^{(3)}, x^{(1)}, x^{(2)}$	$f[x\|101] - f[x\|001]$	$f[x\|111] - f[x\|101]$	$f[x\|001] - f[x\|000]$
$x^{(3)}, x^{(2)}, x^{(1)}$	$f[x\|111] - f[x\|011]$	$f[x\|011] - f[x\|001]$	$f[x\|001] - f[x\|000]$

In a model with 3 features, there are $3! = 6$ possible sequences. We can use the interactions table to derive the marginal contribution of each feature in each sequence. The Shapley values are the averages per column.

A Numerical Example

Coalitions table

$f[x\|\ldots]-\bar{y}$	Observations
$f[x\|000]-\bar{y}$	0
$f[x\|001]-\bar{y}$	20
$f[x\|010]-\bar{y}$	25
$f[x\|011]-\bar{y}$	30
$f[x\|100]-\bar{y}$	10
$f[x\|101]-\bar{y}$	28
$f[x\|110]-\bar{y}$	37
$f[x\|111]-\bar{y}$	50

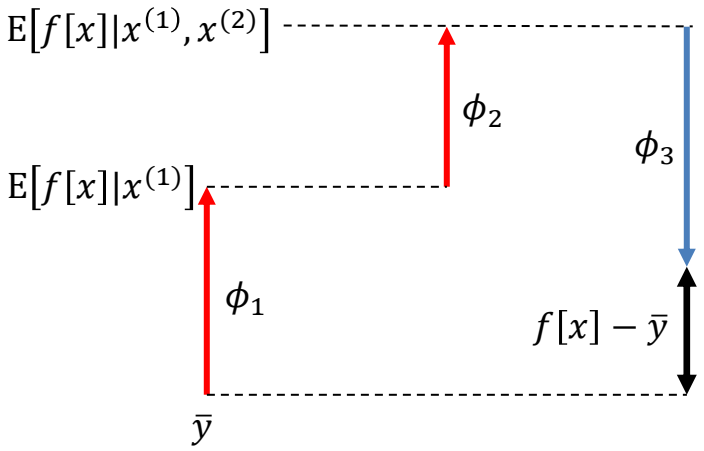
Marginal conditional contributions table

Sequence	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	Sum
$x^{(1)},x^{(2)},x^{(3)}$	10-0=10	37-10=27	50-37=13	50
$x^{(1)},x^{(3)},x^{(2)}$	10-0=10	50-28=22	28-10=18	50
$x^{(2)},x^{(1)},x^{(3)}$	37-25=12	25-0=25	50-37=13	50
$x^{(2)},x^{(3)},x^{(1)}$	50-30=20	25-0=25	30-25=5	50
$x^{(3)},x^{(1)},x^{(2)}$	28-20=8	50-28=22	20-0=20	50
$x^{(3)},x^{(2)},x^{(1)}$	50-30=20	30-20=10	20-0=20	50
Average	13.34	21.83	14.83	50

Example of Shapley values for an observation (x, y) in a model f with 3 features.

Properties of Shapley Values

- Shapley values ($\{\phi_i\}_{i=1,\dots,N}$) provide the only "fair" attributions, i.e. attributions with the following properties
  - **Efficiency**: $f[x] - \bar{y} = \sum_{i=1}^{N} \phi_i$
  - **Symmetry**: $\phi_i = \phi_j$ IIF $x^{(i)}$ and $x^{(j)}$ contribute equally to all possible coalitions
  - **Missingness**: if $x^{(i)}$ does not change $f[x]$ regardless of the coalition, then $\phi_i = 0$
  - **Additivity**: a function with combined outputs has as Shapley values the sum of the constituent ones
- Shapley values and MDA values are **consistent** (unlike MDI values)
  - if the model changes such that it relies more on a feature, that feature's importance will rise



The marginal conditional contribution for a feature depends on the particular sequence of effects. Shapley values achieve **efficiency** by averaging across all possible sequences (hence the above $\mathrm{E}[f[x]|\dots]$ attributions).

From Permutations to Combinations

- As we can appreciate from the numerical example, some calculations are redundant
  - E.g., the marginal contribution of $x^{(3)}$ on sequence $x^{(1)}, x^{(2)}, x^{(3)}$ must be the same as the marginal contribution of $x^{(3)}$ on sequence $x^{(2)}, x^{(1)}, x^{(3)}$, because
    - in both cases $x^{(3)}$ comes in third position, and
    - the permutations of $\{x^{(1)}, x^{(2)}\}$ do not alter that marginal contribution
- The Shapley value of $x^{(i)}$ can be derived as the average contribution of $x^{(i)}$ across all possible coalitions S, where S does not include i

$$\phi_i = \frac{1}{N} \sum_{S \subseteq (X \setminus \{i\})} \binom{N-1}{\|S\|}^{-1} (f[S \cup \{i\}] - f[S])$$

$$= \sum_{S \subseteq (X \setminus \{i\})} \frac{\|S\|!\,(N - \|S\| - 1)!}{N!} (f[S \cup \{i\}] - f[S])$$

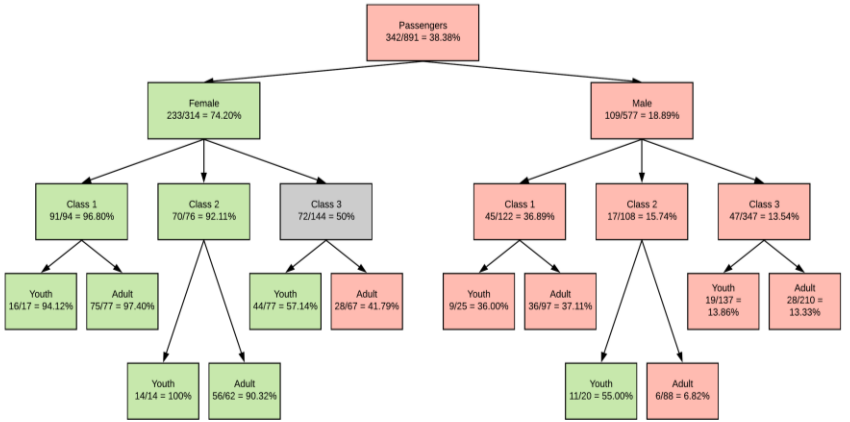
Coalitions can have sizes $\|S\| = 0,1, \ldots, N-1$. A coalition S produces non-redundant marginal contributions $f[S \cup \{i\}] - f[S]$ for $\binom{N-1}{\|S\|}$ combinations. The above equation computes the exact Shapley values by grouping the marginal conditional contributions in terms of coalitions (S). For large N, Strumbelj et al. [2014], Lundberg and Lee [2016], and Lundberg et al. [2019] have developed fast algorithms for the estimation of ϕ_i.

Interaction Effects

- An interaction effect occurs when the effect of one variable on the output depends on the value of another variable
- The estimation of interaction effects requires having an accurate attribution of the individual effect of the variables involved
  - Shapley values are particularly useful for estimating interaction effects
  - For $i \neq j$,

$$\phi_{i,j} = \sum_{S \subseteq (X \setminus \{i,j\})} \frac{\|S\|!\,(N - \|S\| - 2)!}{2(N-1)!}\,\delta_{i,j}[S]$$

where $\delta_{i,j}[S] = f[S \cup \{i,j\}] - f[S \cup \{i\}] - f[S \cup \{j\}] + f[S]$

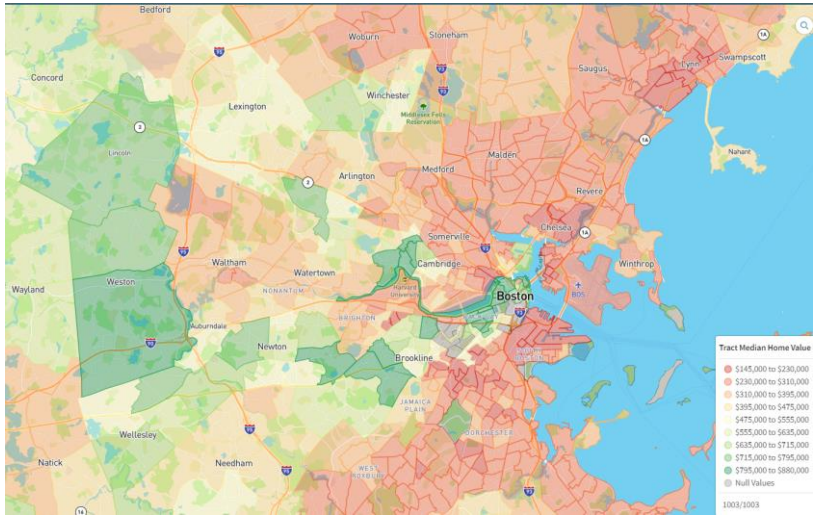


The survival probability of RMS Titanic passengers. There are interaction effects between gender and ticket class (e.g., females in class 3 relative to females in classes 1 & 2), and age and ticket class (e.g., young males in class 2 relative to adult males in class 2). Shapley values help us quantify the strength of these interactions.

Case Study:
Boston Housing Prices

The Data

- Target: prices for 506 residential properties
- Features (14):
  - **CRIM**: per capita crime rate by town
  - **ZN**: proportion of residential land zoned for lots over 25,000 sq.ft.
  - **INDUS**: proportion of non-retail business acres per town
  - **CHAS**: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
  - **NOX**: nitric oxides concentration (parts per 10 million)
  - **RM**: average number of rooms per dwelling
  - **AGE**: proportion of owner-occupied units built prior to 1940
  - **DIS**: weighted distances to five Boston employment centers
  - **RAD**: index of accessibility to radial highways
  - **TAX**: full-value property-tax rate per $10,000
  - **PTRATIO**: pupil-teacher ratio by town
  - **B**: $1000(Bk - 0.63)^2$, where Bk is the proportion of blacks by town
  - **LSTAT**: proportion of the population that is poor
  - **MEDV**: Median value of owner-occupied homes in $1000's

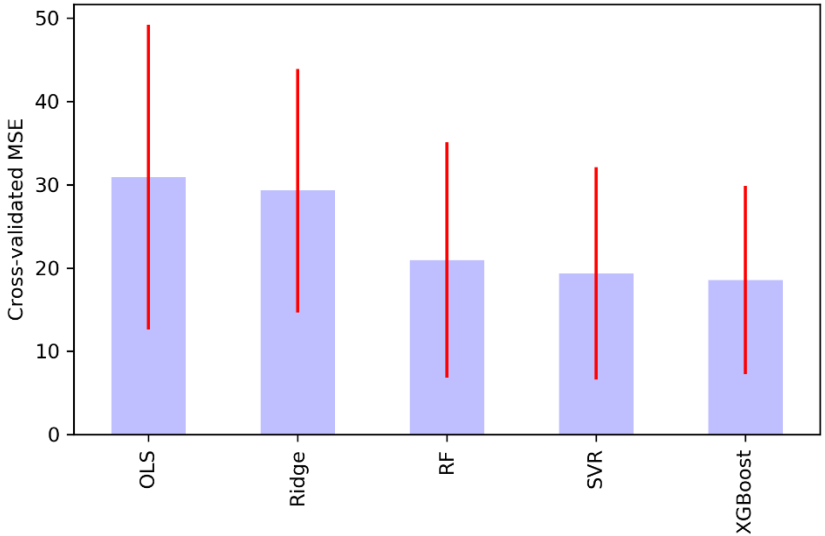


Source: Harvard University

The Boston house-price data, collected by Harrison and Rubinfeld [1978], has been used in multiple of ML studies, tournaments and research papers.

The Model

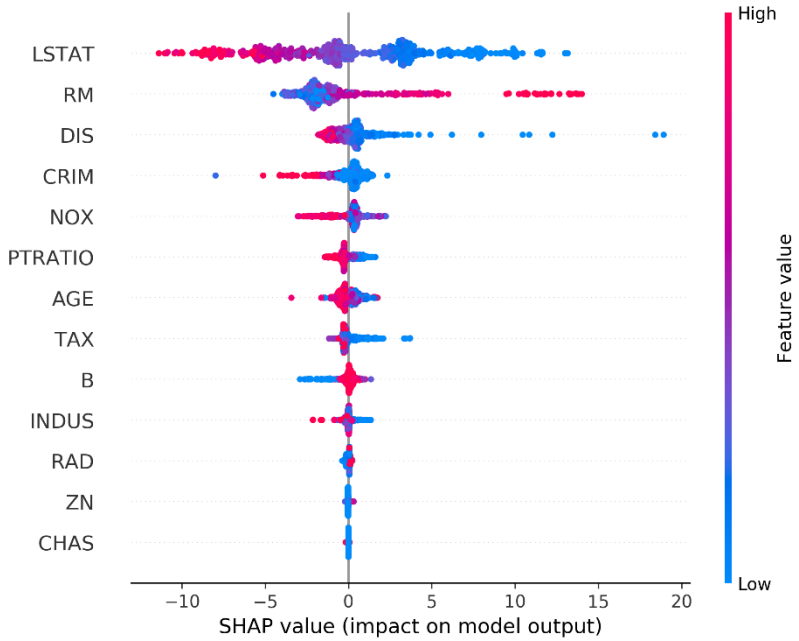
- The data exhibits strong non-linearities, which are not well captured by linear models (OLS, Ridge)
- The best performing model (XGBoost) is not intrinsically interpretable
- We can develop an understanding of the uncovered patterns by attributing the effect (housing prices) to the various features
  - In a linear model, we would simply look at the size and sign of the estimated coefficients
  - In a nonlinear model, we must study this attribution on a case-by-case basis: **Shapley values**
- By **abduction**, this attribution gives us the plausible drivers of housing prices



Above is a bar plot of the mean squared errors (with 1-std error bars) derived through K-Fold cross-validation. Linear models (OLS, Ridge) performed considerably worse than non-linear ML algorithms (Random Forest regression, Support Vector Regressor, XGBoost) on this small dataset.

Feature Importance Plot

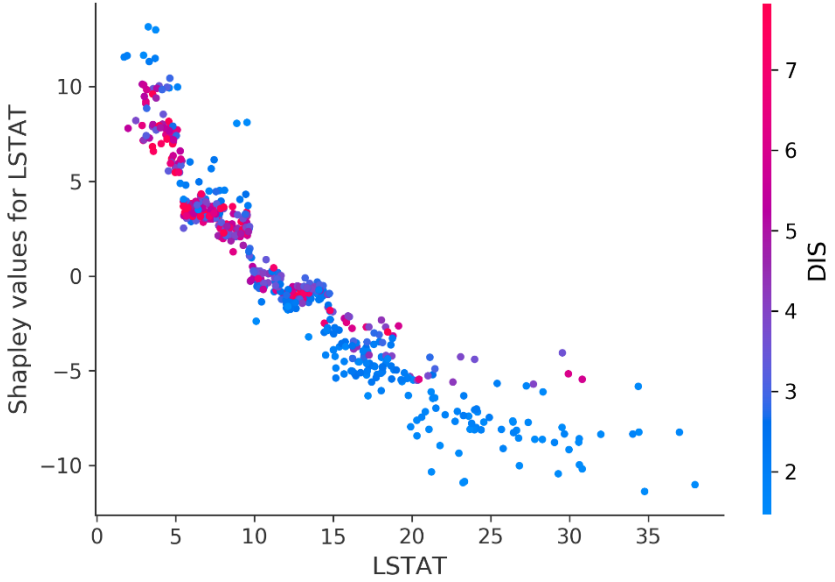
- It displays the Shapley value for every observation, grouped by feature
  - The **y-axis** ranks the features by importance
    - Importance is measured as average absolute Shapley values
  - The **x-axis** shows the magnitude of the impact
    - When many observations for a given feature have a similar impact, the line of dots widens
  - The **color** shows the magnitude of the feature
- In this example, we learn that
  - The top drivers of housing prices are poverty rate (LSAT), number of rooms (RM), distance to employment centers (DIS), crime rate (CRIM), and air pollution (NOX)
  - All of these drivers impact prices negatively, except for RM



The signs of the relationships match intuition, providing support that the XGBoost model has found patterns with a theoretical foundation.

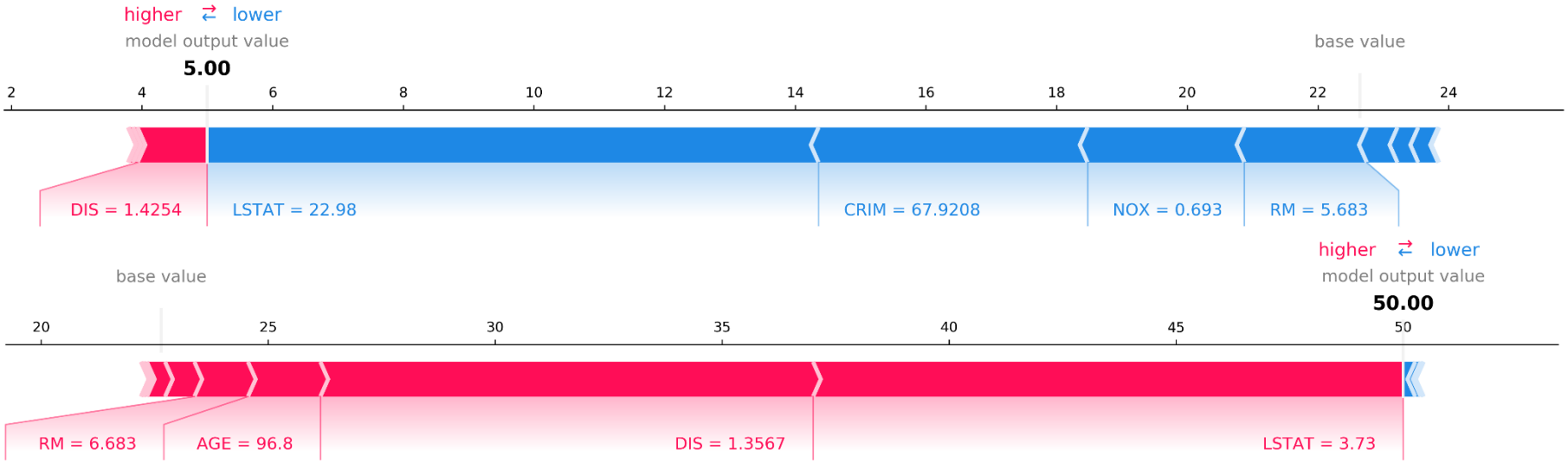
Dependence Plot

- For any *pair* of features, it displays
  - the Shapley values (y-axis) as a function of one feature's value (x-axis)
    - E.g., a linear, negative, monotonic effect
  - the interaction between that feature (x-axis) and another (dot color)
    - E.g., a color separation in dots across the y-axis
- In this example, we learn that
  - the poverty rate (LSAT) has a negative impact on prices (consistent with the importance plot)
  - the effect is almost linear for LSAT between 5% and 20%, but more acute around the extremes
  - there are strong interactions between LSAT and DIS (distance to employment centers):
    - **High LSAT is detrimental, particularly with low DIS (inner cities)**
    - **Low LSAT is beneficial, particularly with low DIS (downtown)**



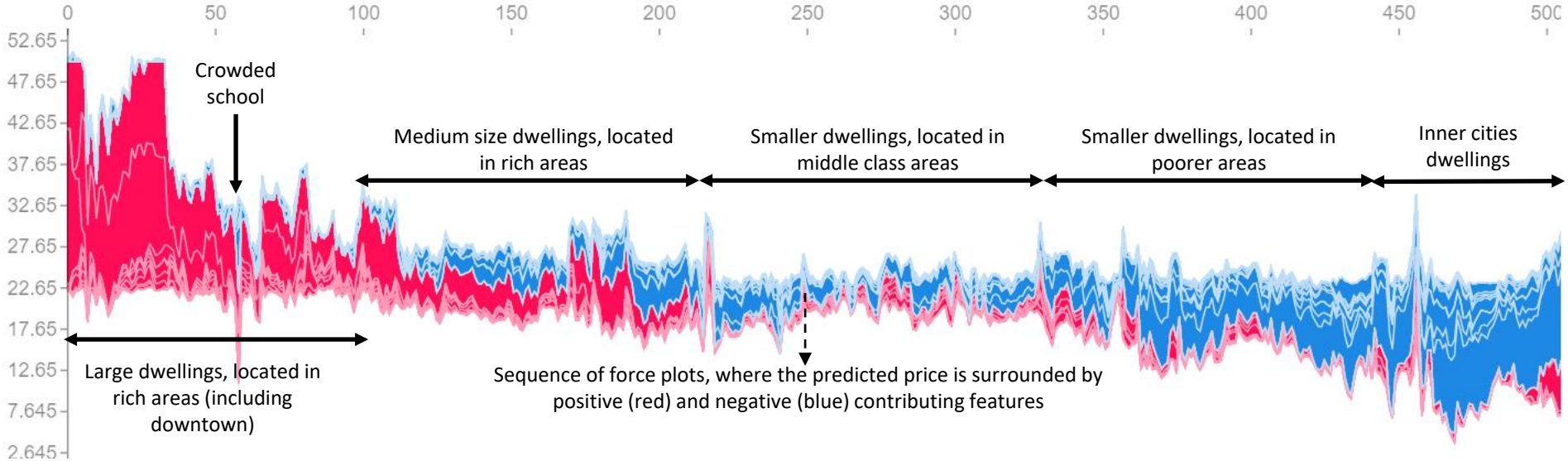
The interaction effect between LSAT and DIS is not monotonic. It flips sign around LSAT of 12%: for higher LSAT values, red dots are on top, and for lower LSAT values, red dots are at the bottom.

Force Plot



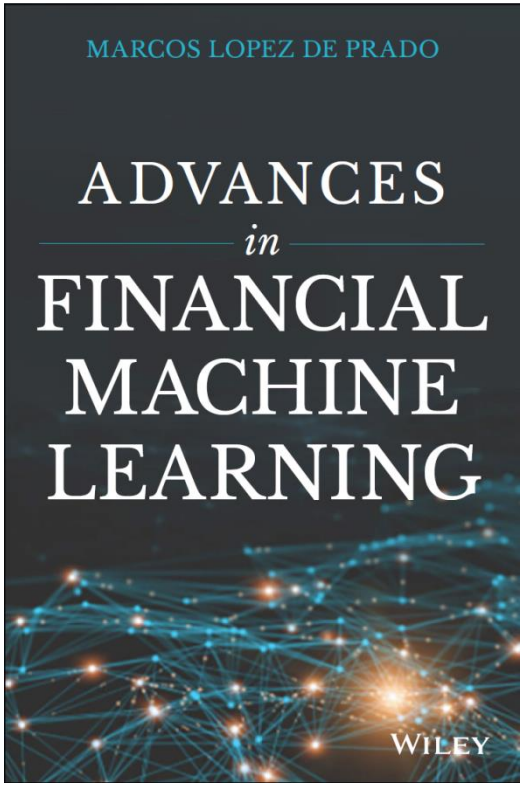
A force plot breaks down the contribution per feature to a particular prediction. The top plot shows the attribution for the house with lowest predicted price (5.00), and the bottom plot shows the attribution for the house with highest predicted price (50.00). The base value is the average prediction (22.64). Features colored in blue detract from the predicted price (e.g., high LSAT, of 22.98), and featured colored in red add to the predicted price (e.g., low LSAT, of 3.73). Features are ordered by the magnitude of their impact on the prediction. Note that the low DIS contributed to increase the price prediction in both examples, in congruence with the feature importance plot.

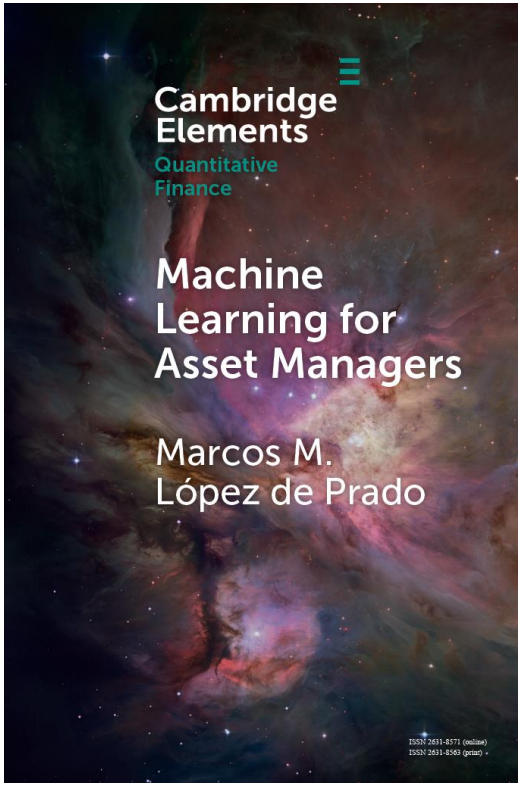
Supervised Clustering



A unit change in one feature's Shapley value is comparable to a unit change in another feature's Shapley value, because all Shapley values are expressed in the same (target variable, y) unit. We can use the matrix of Shapley values to form a distance matrix between observations. The hierarchical clustering of that distance matrix gives us a new sequence of observations, ordered by similarity of attributions (observations with similar reasons for a predicted outcome are placed together). This supervised clustering allows us to intuitively summarize all housing prices into 5 major categories.

For Additional Details





The first wave of quantitative innovation in finance was led by Markowitz optimization. Machine Learning is the second wave and it will touch every aspect of finance. López de Prado's Advances in Financial Machine Learning is essential for readers who want to be ahead of the technology rather than being replaced by it.
— Prof. **Campbell Harvey**, Duke University. Former President of the American Finance Association.

Financial problems require very distinct machine learning solutions. Dr. López de Prado's book is the first one to characterize what makes standard machine learning tools fail when applied to the field of finance, and the first one to provide practical solutions to unique challenges faced by asset managers. Everyone who wants to understand the future of finance should read this book.
— Prof. **Frank Fabozzi**, EDHEC Business School. Editor of The Journal of Portfolio Management.

Disclaimer

- The views expressed in this document are the author's and do not necessarily reflect those of the organizations he is affiliated with.
- No investment decision or particular course of action is recommended by this presentation.
-

www.QuantResearch.org