8.5 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. Shapley values – a method from coalitional game theory – tells us how to fairly distribute the "payout" among the features.

8.5.1 General Idea

Assume the following scenario:

You have trained a machine learning model to predict apartment prices. For a certain apartment it predicts €300,000 and you need to explain this prediction. The apartment has an area of 50 m^2, is located on the 2nd floor, has a park nearby and cats are banned:

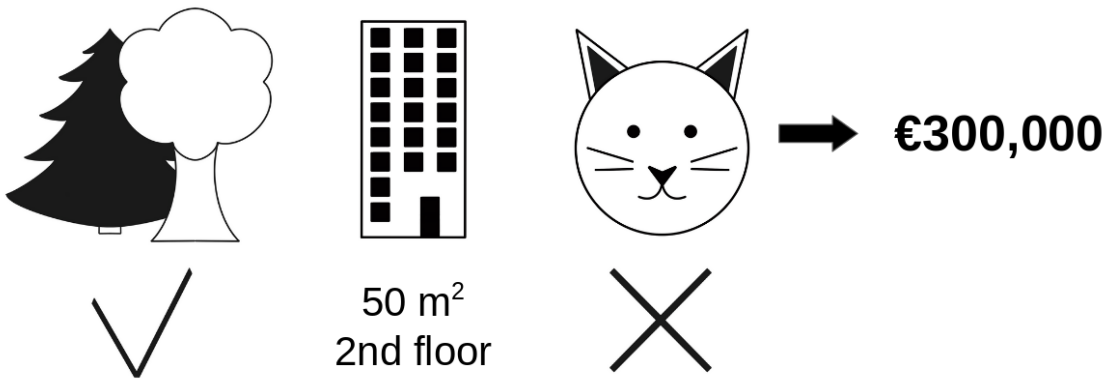

Figure 8.17: The predicted price for a 50 m^2 2nd floor apartment with a nearby park and cat ban is €300,000. Our goal is to explain how each of these feature values contributed to the prediction.

The average prediction for all apartments is €310,000. How much has each feature value contributed to the prediction compared to the average prediction?

The answer is simple for linear regression models. The effect of each feature is the weight of the feature times the feature value. This only works because of the linearity of the model. For more complex models, we need a different solution. For example, LIME suggests local models to estimate effects. Another solution comes from cooperative game theory: The Shapley value, coined by Shapley (1953)[33], is a method for assigning payouts

[33]Shapley, Lloyd S. "A value for n-person games." Contributions to the Theory of Games 2.28 (1953): 307-317.

to players depending on their contribution to the total payout. Players cooperate in a coalition and receive a certain profit from this cooperation.

Players? Game? Payout? What is the connection to machine learning predictions and interpretability? The "game" is the prediction task for a single instance of the dataset. The "gain" is the actual prediction for this instance minus the average prediction for all instances. The "players" are the feature values of the instance that collaborate to receive the gain (= predict a certain value). In our apartment example, the feature values `park-nearby`, `cat-banned`, `area-50` and `floor-2nd` worked together to achieve the prediction of €300,000. Our goal is to explain the difference between the actual prediction (€300,000) and the average prediction (€310,000): a difference of -€10,000.

The answer could be: The `park-nearby` contributed €30,000; `area-50` contributed €10,000; `floor-2nd` contributed €0; `cat-banned` contributed -€50,000. The contributions add up to -€10,000, the final prediction minus the average predicted apartment price.

How do we calculate the Shapley value for one feature?

The Shapley value is the average marginal contribution of a feature value across all possible coalitions. All clear now?

In the following figure we evaluate the contribution of the `cat-banned` feature value when it is added to a coalition of `park-nearby` and `area-50`. We simulate that only `park-nearby`, `cat-banned` and `area-50` are in a coalition by randomly drawing another apartment from the data and using its value for the floor feature. The value `floor-2nd` was replaced by the randomly drawn `floor-1st`. Then we predict the price of the apartment with this combination (€310,000). In a second step, we remove `cat-banned` from the coalition by replacing it with a random value of the cat allowed/banned feature from the randomly drawn apartment. In the example it was `cat-allowed`, but it could have been `cat-banned` again. We predict the apartment price for the coalition of `park-nearby` and `area-50` (€320,000). The contribution of `cat-banned` was €310,000 - €320,000 = -€10,000. This estimate depends on the values of the randomly drawn apartment that served as a "donor" for the cat and floor feature values. We will get better estimates if we repeat this sampling step and average the contributions.

We repeat this computation for all possible coalitions. The Shapley value is the average of all the marginal contributions to all possible coalitions. The computation time increases exponentially with the number of features. One solution to keep the computation time manageable is to compute contributions for only a few samples of the possible coalitions.

The following figure shows all coalitions of feature values that are needed to determine the Shapley value for `cat-banned`. The first row shows the coalition without any feature values. The second, third and fourth rows show different coalitions with increasing coalition size, separated by "|". All in all, the following coalitions are possible:

the game — prediction task
the gain — instance prediction minus average prediction of all
the players — feature values that collaborate to receive gain

Coalition 联合

Shapley value 是一个 feature 值在所有"联合"中的平均边缘贡献

计算一个 feature value 的在这个 instance 上的 shapley value

计算时间是 feature 数的指数关系

组合总数：2^n 次计算

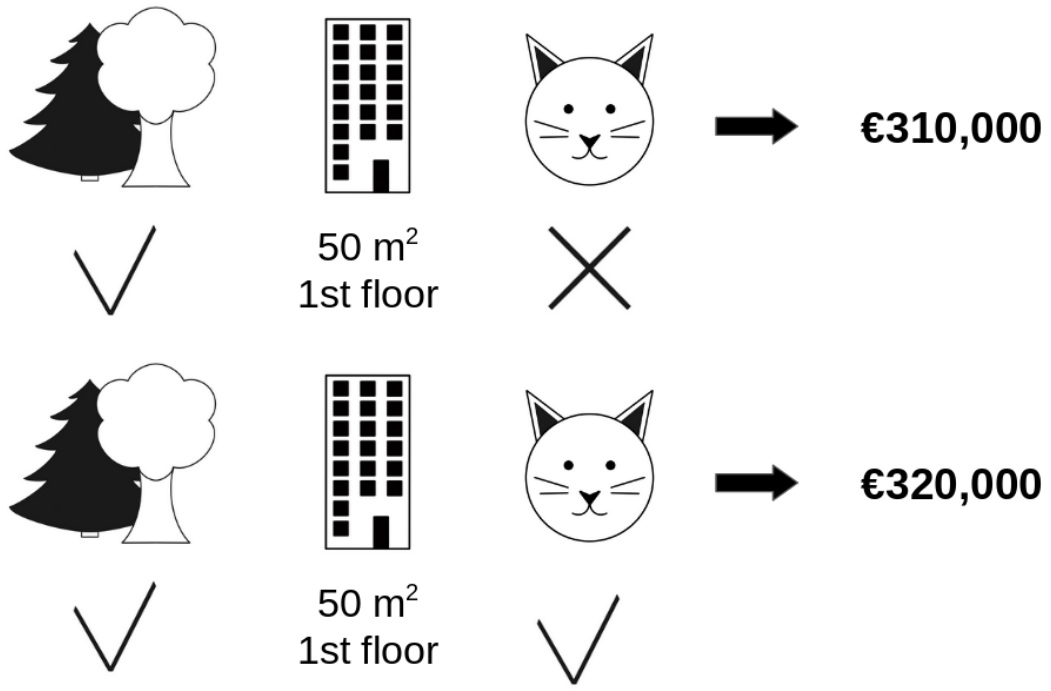

Figure 8.18: One sample repetition to estimate the contribution of 'cat-banned' to the prediction when added to the coalition of 'park-nearby' and 'area-50'.

- `No feature values`
- `park-nearby`
- `area-50`
- `floor-2nd`
- `park-nearby+area-50`
- `park-nearby+floor-2nd`
- `area-50+floor-2nd`
- `park-nearby+area-50+floor-2nd`.

For each of these coalitions we compute the predicted apartment price with and without the feature value `cat-banned` and take the difference to get the marginal contribution. The Shapley value is the (weighted) average of marginal contributions. We replace the feature values of features that are not in a coalition with random feature values from the apartment dataset to get a prediction from the machine learning model.

If we estimate the Shapley values for all feature values, we get the complete distribution of the prediction (minus the average) among the feature values.

该 instance 为

对于每个组合，计算有 cat-banned 和没有的分数，它们的差就是 marginal contribution。所有组合的 marginal contribution 平均值即是 "cat-banned" 的 shapley value

如果计算好所有 feature value 的 shapley value，可以得出该 feature 的 减去平均数的预测分布。

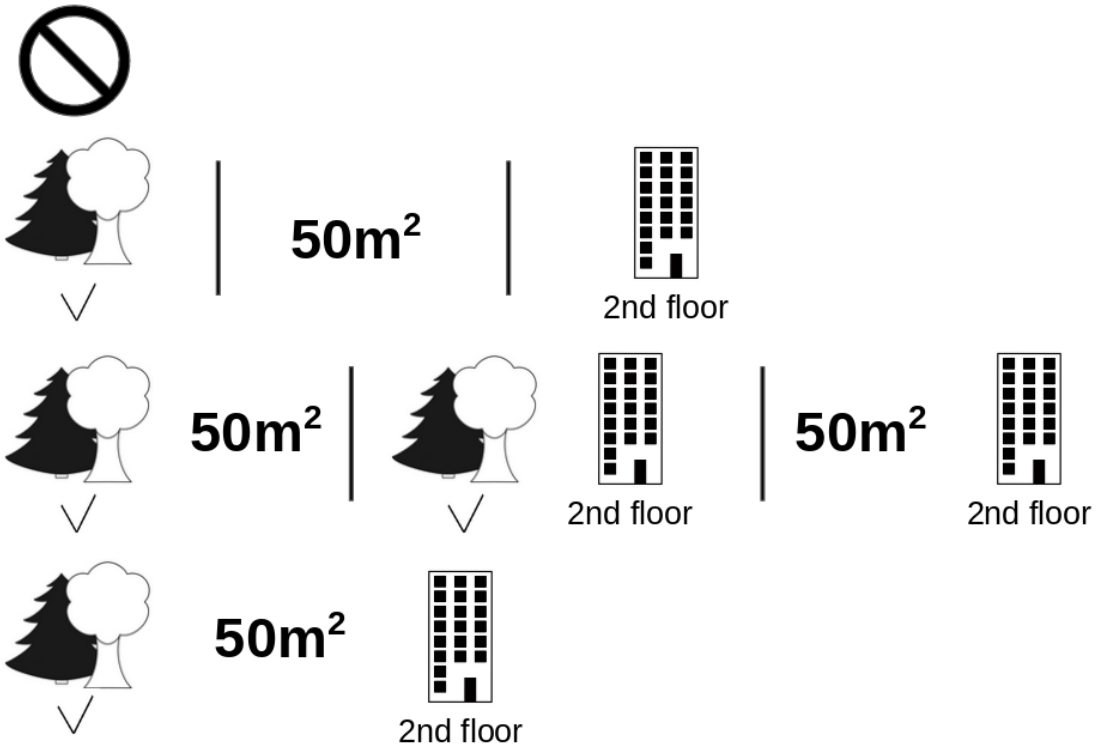

Figure 8.19: All 8 coalitions needed for computing the exact Shapley value of the 'cat-banned' feature value.

8.5.2 Examples and Interpretation

The interpretation of the Shapley value for feature value j is: The value of the j-th feature contributed ϕ_j to the prediction of this particular instance compared to the average prediction for the dataset.

The Shapley value works for both classification (if we are dealing with probabilities) and regression.

We use the Shapley value to analyze the predictions of a random forest model predicting cervical cancer:

For the bike rental dataset, we also train a random forest to predict the number of rented bikes for a day, given weather and calendar information. The explanations created for the random forest prediction of a particular day:

Be careful to interpret the Shapley value correctly: The Shapley value is the average contribution of a feature value to the prediction in different coalitions. The Shapley value

这个分布是否能求个平均来计算 global F Imp.?

→ Fig 8.20

→ Fig 8.21

解释 shapley value 的时候要小心。

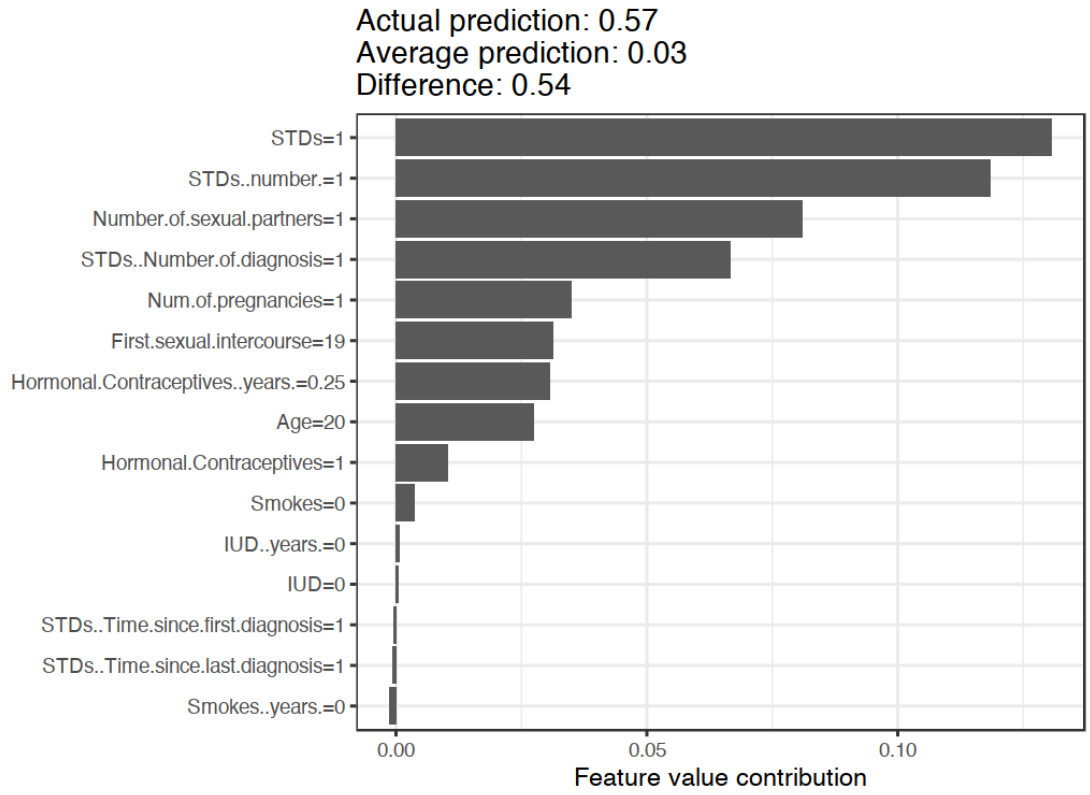

Figure 8.20: Shapley values for a woman in the cervical cancer dataset. With a prediction of 0.57, this woman's cancer probability is 0.54 above the average prediction of 0.03. The number of diagnosed STDs increased the probability the most. The sum of contributions yields the difference between actual and average prediction (0.54).

is NOT the difference in prediction when we would remove the feature from the model.

8.5.3 The Shapley Value in Detail

This section goes deeper into the definition and computation of the Shapley value for the curious reader. Skip this section and go directly to "Advantages and Disadvantages" if you are not interested in the technical details.

We are interested in how each feature affects the prediction of a data point. In a linear model it is easy to calculate the individual effects. Here is what a linear model prediction looks like for one data instance:

Shapley value 是这个 feature value 在所有组合预测下的平均值，而不是当我们移除这个 feature 之后的预测值的差。

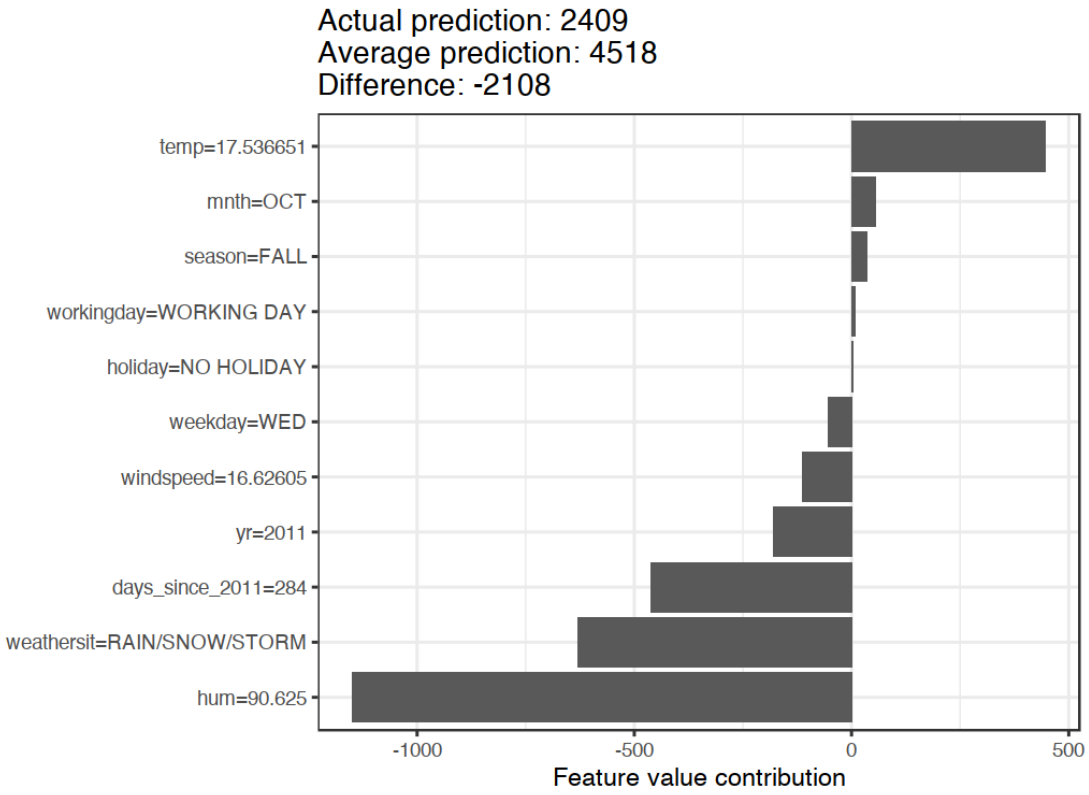

Figure 8.21: Shapley values for day 285. With a predicted 2409 rental bikes, this day is -2108 below the average prediction of 4518. The weather situation and humidity had the largest negative contributions. The temperature on this day had a positive contribution. The sum of Shapley values yields the difference of actual and average prediction (-2108).

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

where x is the instance for which we want to compute the contributions. Each x_j is a feature value, with j = 1,...,p. The β_j is the weight corresponding to feature j.

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

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

where $E(\beta_j X_j)$ is the mean effect estimate for feature j. The contribution is the difference between the feature effect minus the average effect. Nice! Now we know how much each

feature contributed to the prediction. If we sum all the feature contributions for one instance, the result is the following:

$$\begin{aligned}\sum_{j=1}^{p}\phi_j(\hat{f}) &= \sum_{j=1}^{p}(\beta_j x_j - E(\beta_j X_j)) \\ &= (\beta_0 + \sum_{j=1}^{p}\beta_j x_j) - (\beta_0 + \sum_{j=1}^{p} E(\beta_j X_j)) \\ &= \hat{f}(x) - E(\hat{f}(X))\end{aligned}$$

This is the predicted value for the data point x minus the average predicted value. Feature contributions can be negative.

Can we do the same for any type of model? It would be great to have this as a model-agnostic tool. Since we usually do not have similar weights in other model types, we need a different solution.

Help comes from unexpected places: cooperative game theory. The Shapley value is a solution for computing feature contributions for single predictions for any machine learning model.

8.5.3.1 The Shapley Value

The Shapley value is defined via a value function *val* of players in S.

The Shapley value of a feature value is its contribution to the payout, weighted and summed over all possible feature value combinations:

$$\phi_j(val) = \sum_{S\subseteq\{1,\ldots,p\}\setminus\{j\}} \frac{|S|!\,(p-|S|-1)!}{p!}\left(val\left(S\cup\{j\}\right) - val(S)\right)$$

where S is a subset of the features used in the model, x is the vector of feature values of the instance to be explained and p the number of features. $val_x(S)$ is the prediction for feature values in set S that are marginalized over features that are not included in set S:

$$val_x(S) = \int \hat{f}(x_1,\ldots,x_p)d\mathbb{P}_{x\notin S} - E_X(\hat{f}(X))$$

S是features的子集组合，
它不包含feature j
P是除开j之外的
features的组合总数

|S|是S集合的元素个数。

You actually perform multiple integrations for each feature that is not contained S. A concrete example: The machine learning model works with 4 features x1, x2, x3 and x4 and we evaluate the prediction for the coalition S consisting of feature values x1 and x3:

$$val_x(S) = val_x(\{1,3\}) = \int_{\mathbb{R}}\int_{\mathbb{R}} \hat{f}(x_1, X_2, x_3, X_4) d\mathbb{P}_{X_2X_4} - E_X(\hat{f}(X))$$

This looks similar to the feature contributions in the linear model!

Do not get confused by the many uses of the word "value": The feature value is the numerical or categorical value of a feature and instance; the Shapley value is the feature contribution to the prediction; the value function is the payout function for coalitions of players (feature values).

The Shapley value is the only attribution method that satisfies the properties **Efficiency**, **Symmetry**, **Dummy** and **Additivity**, which together can be considered a definition of a fair payout.

Efficiency The feature contributions must add up to the difference of prediction for x and the average.

$$\sum_{j=1}^{p} \phi_j = \hat{f}(x) - E_X(\hat{f}(X))$$

Symmetry The contributions of two feature values j and k should be the same if they contribute equally to all possible coalitions. If

$$val(S \cup \{j\}) = val(S \cup \{k\})$$

for all

$$S \subseteq \{1, \dots, p\} \quad \{j, k\}$$

then

$$\phi_j = \phi_k$$

Dummy A feature j that does not change the predicted value – regardless of which coalition of feature values it is added to – should have a Shapley value of 0. If

$$val(S \cup \{j\}) = val(S)$$

Shapley value 是唯一满足 efficiency、对称性、Dummy（简单直接）、Additivity（相加性）的 feature 归因方法。

对称性：如果 feature value j 和 k 贡献度相同，那它们的 Shapley value 也相同。

Dummy 如果 feature value 没有预测贡献，那么它的 Shapley value 为 0。

for all

$$S \subseteq \{1, ..., p\}$$

then

$$\phi_j = 0$$

Additivity For a game with combined payouts val+val$^+$ the respective Shapley values are as follows:

$$\phi_j + \phi_j^+$$

Suppose you trained a random forest, which means that the prediction is an average of many decision trees. The Additivity property guarantees that for a feature value, you can calculate the Shapley value for each tree individually, average them, and get the Shapley value for the feature value for the random forest.

8.5.3.2 Intuition

An intuitive way to understand the Shapley value is the following illustration: The feature values enter a room in random order. All feature values in the room participate in the game (= contribute to the prediction). The Shapley value of a feature value is the average change in the prediction that the coalition already in the room receives when the feature value joins them.

8.5.3.3 Estimating the Shapley Value

All possible coalitions (sets) of feature values have to be evaluated with and without the j-th feature to calculate the exact Shapley value. For more than a few features, the exact solution to this problem becomes problematic as the number of possible coalitions exponentially increases as more features are added. Strumbelj et al. (2014)[34] propose an approximation with Monte-Carlo sampling:

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

[34] Štrumbelj, Erik, and Igor Kononenko. "Explaining prediction models and individual predictions with feature contributions." Knowledge and information systems 41.3 (2014): 647-665.

当 feature 数上涨时，计算量也是指数级的上涨，(2^n)
提取有以下近似求解

where $\hat{f}(x^m_{+j})$ 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. The x-vector x^m_{-j} is almost identical to x^m_{+j}, but the value x^m_j is also taken from the sampled z. Each of these M new instances is a kind of "Frankenstein's Monster" assembled from two instances. Note that in the following algorithm, the order of features is not actually changed – each feature remains at the same vector position when passed to the predict function. The order is only used as a "trick" here: By giving the features a new order, we get a random mechanism that helps us put together the "Frankenstein's Monster". For features that appear left of the feature x_j, we take the values from the original observations, and for the features on the right, we take the values from a random instance.

Approximate Shapley estimation for single feature value:

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

First, select an instance of interest x, a feature j and the number of iterations M. For each iteration, a random instance z is selected from the data and a random order of the features is generated. Two new instances are created by combining values from the instance of interest x and the sample z. The instance x_{+j} is the instance of interest, but all values in the order after feature j are replaced by feature values from the sample z. The instance x_{-j} is the same as x_{+j}, but in addition has feature j replaced by the value for feature j from the sample z. The difference in the prediction from the black box is computed:

$$\phi^m_j = \hat{f}(x^m_{+j}) - \hat{f}(x^m_{-j})$$

All these differences are averaged and result in:

$$\phi_j(x) = \frac{1}{M}\sum_{m=1}^{M}\phi_j^m$$

Averaging implicitly weighs samples by the probability distribution of X.

The procedure has to be repeated for each of the features to get all Shapley values.

8.5.4 Advantages

The difference between the prediction and the average prediction is **fairly distributed** among the feature values of the instance – the Efficiency property of Shapley values. This property distinguishes the Shapley value from other methods such as LIME. LIME does not guarantee that the prediction is fairly distributed among the features. The Shapley value might be the only method to deliver a full explanation. In situations where the law requires explainability – like EU's "right to explanations" – the Shapley value might be the only legally compliant method, because it is based on a solid theory and distributes the effects fairly. I am not a lawyer, so this reflects only my intuition about the requirements.

The Shapley value allows **contrastive explanations**. Instead of comparing a prediction to the average prediction of the entire dataset, you could compare it to a subset or even to a single data point. This contrastiveness is also something that local models like LIME do not have.

The Shapley value is the only explanation method with a **solid theory**. The axioms – efficiency, symmetry, dummy, additivity – give the explanation a reasonable foundation. Methods like LIME assume linear behavior of the machine learning model locally, but there is no theory as to why this should work.

It is mind-blowing to **explain a prediction as a game** played by the feature values.

8.5.5 Disadvantages

The Shapley value requires **a lot of computing time**. In 99.9% of real-world problems, only the approximate solution is feasible. An exact computation of the Shapley value is computationally expensive because there are 2^k possible coalitions of the feature values and the "absence" of a feature has to be simulated by drawing random instances, which increases the variance for the estimate of the Shapley values estimation. The exponential number of the coalitions is dealt with by sampling coalitions and limiting the number of iterations M. Decreasing M reduces computation time, but increases the variance of the Shapley value. There is no good rule of thumb for the number of iterations M. M should be large enough to accurately estimate the Shapley values, but small enough to

在预测和预测平均值之间的差是平等分布在 instance 的 feature values 上的（更小的方差）

它有 solid theory 并且将 feature 的效应公平地分布

它也可以用于反证解释

计算复杂取决于 feature 数量而且会指数增长

complete the computation in a reasonable time. It should be possible to choose M based on Chernoff bounds, but I have not seen any paper on doing this for Shapley values for machine learning predictions.

The Shapley value **can be misinterpreted**. The Shapley value of a feature value is not the difference of the predicted value after removing the feature from the model training. The interpretation of the Shapley value is: Given the current set of feature values, the contribution of a feature value to the difference between the actual prediction and the mean prediction is the estimated Shapley value.

The Shapley value is the wrong explanation method if you seek sparse explanations (explanations that contain few features). Explanations created with the Shapley value method **always use all the features**. Humans prefer selective explanations, such as those produced by LIME. LIME might be the better choice for explanations lay-persons have to deal with. Another solution is SHAP[35] introduced by Lundberg and Lee (2016)[36], which is based on the Shapley value, but can also provide explanations with few features.

The Shapley value returns a simple value per feature, but **no prediction model** like LIME. This means it cannot be used to make statements about changes in prediction for changes in the input, such as: "If I were to earn €300 more a year, my credit score would increase by 5 points."

Another disadvantage is that **you need access to the data** if you want to calculate the Shapley value for a new data instance. It is not sufficient to access the prediction function because you need the data to replace parts of the instance of interest with values from randomly drawn instances of the data. This can only be avoided if you can create data instances that look like real data instances but are not actual instances from the training data.

Like many other permutation-based interpretation methods, the Shapley value method suffers from **inclusion of unrealistic data instances** when features are correlated. To simulate that a feature value is missing from a coalition, we marginalize the feature. This is achieved by sampling values from the feature's marginal distribution. This is fine as long as the features are independent. When features are dependent, then we might sample feature values that do not make sense for this instance. But we would use those to compute the feature's Shapley value. One solution might be to permute correlated features together and get one mutual Shapley value for them. Another adaptation is conditional sampling: Features are sampled conditional on the features that are already in the team. While conditional sampling fixes the issue of unrealistic data points, a new issue is introduced: The resulting values are no longer the Shapley values to our game,

[35] https://github.com/slundberg/shap

[36] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems (2017).

它只仅回数值

当feature有correlation时shapley value的值多少受影响。

不合理的数据会出现，从而影响计算。

解决方法是将高相关的feature一起度，但是这样计算的值会违反Shapley theory。

since they violate the symmetry axiom, as found out by Sundararajan et al. (2019)[37] and further discussed by Janzing et al. (2020)[38].

8.5.6 Software and Alternatives

Shapley values are implemented in both the `iml` and fastshap[39] packages for R. In Julia, you can use Shapley.jl[40].

SHAP, an alternative estimation method for Shapley values, is presented in the next chapter.

Another approach is called breakDown, which is implemented in the `breakDown` R package[41]. BreakDown also shows the contributions of each feature to the prediction, but computes them step by step. Let us reuse the game analogy: We start with an empty team, add the feature value that would contribute the most to the prediction and iterate until all feature values are added. How much each feature value contributes depends on the respective feature values that are already in the "team", which is the big drawback of the breakDown method. It is faster than the Shapley value method, and for models without interactions, the results are the same.

[37] Sundararajan, Mukund, and Amir Najmi. "The many Shapley values for model explanation." arXiv preprint arXiv:1908.08474 (2019).

[38] Janzing, Dominik, Lenon Minorics, and Patrick Blöbaum. "Feature relevance quantification in explainable AI: A causality problem." arXiv preprint arXiv:1910.13413 (2019).

[39] https://github.com/bgreenwell/fastshap

[40] https://gitlab.com/ExpandingMan/Shapley.jl

[41] Staniak, Mateusz, and Przemyslaw Biecek. "Explanations of model predictions with live and breakDown packages." arXiv preprint arXiv:1804.01955 (2018).