Model Explainability With SHAPLEY

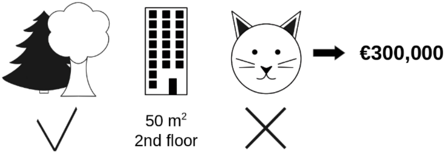
Suchana Datta [1]

[1]School of Computer Science
University College Dublin, Ireland

April 11, 2021

Brief Overview

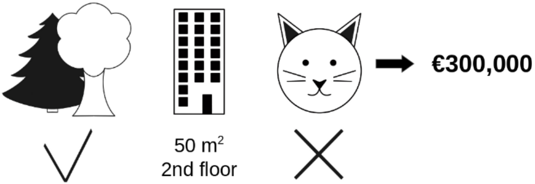
Could You Explain This?



Trained a machine learning model to predict apartment prices

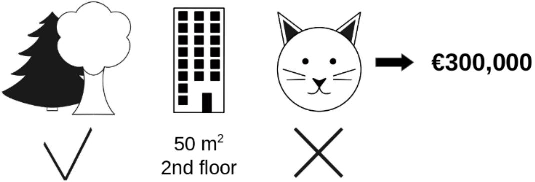
- Model predicts €300,000
- The apartment has an area of 50 m^2
- Located on the 2nd floor
- Has a park nearby
- Cats are banned
- Average prediction for all apartments is €310,000

Could You Explain This?



How much has each feature value contributed to the prediction compared to the average prediction?

From Cooperative Game Theory: The Shapley Value



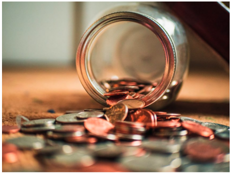
How much has each feature value contributed to the prediction compared to the average prediction?

Solution :

Shapley Values – a method from **cooperative** game theory – tells us how to fairly distribute the "payout" among the features.

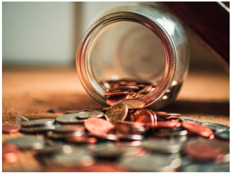
Cooperative Game Theory - What Is It?

Three friends – Adam, Ben, and Patt – go out for a meal



Cooperative Game Theory - What Is It?

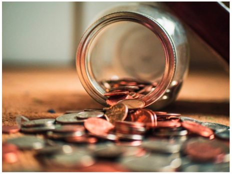
Three friends – Adam, Ben, and Patt – go out for a meal



- If Adam is eating alone, he pays €80
- If Ben is eating alone, he pays €56
- If Patt is eating alone, he pays €70
- If Adam and Ben both eat alone, they pay €80
- If Adam and Patt both eat alone, they pay €85
- If Ben and Patt both eat alone, they pay €72
- If Adam, Ben, and Patt all eat together, they pay €90

Cooperative Game Theory - What Is It?

Three friends – Adam, Ben, and Patt – go out for a meal



- If Adam is eating alone, he pays €80
- If Ben is eating alone, he pays €56
- If Patt is eating alone, he pays €70
- If Adam and Ben both eat alone, they pay €80
- If Adam and Patt both eat alone, they pay €85
- If Ben and Patt both eat alone, they pay €72
- If Adam, Ben, and Patt all eat together, they pay €90

Now, the task is to figure out how much each of them should pay individually?

Background: Shapley Values and Machine Learning

Lloyd Shapley: "How should we divide a **payout among a cooperating team** whose **members** made different contributions?"

Background: Shapley Values and Machine Learning

Lloyd Shapley: "How should we divide a **payout among a cooperating team** whose **members** made different contributions?"

- Shapley value for **member** X is the amount of credit they get.
- Total **payout** is the sum of Shapley values over **members**.

Background: Shapley Values and Machine Learning

Lloyd Shapley: "How should we divide a **payout among a cooperating team** whose **members** made different contributions?"

- Shapley value for **member** X is the amount of credit they get.
- Total **payout** is the sum of Shapley values over **members**.

To compute :

- For every possible subteam, how much marginal value does member X add when they join the subteam?
- Shapley value is the weighted mean of this marginal value.

So How Do You Please 3 Friends?

- Adam $\rightarrow$80
- Ben $\rightarrow$56
- Patt $\rightarrow$70
- Adam + Ben $\rightarrow$80
- Adam + Patt $\rightarrow$85
- Ben + Patt $\rightarrow$72
- Adam + Ben + Patt $\rightarrow$90

So How Do You Please 3 Friends?

- Adam $\rightarrow$80
- Ben $\rightarrow$56
- Patt $\rightarrow$70
- Adam + Ben $\rightarrow$80
- Adam + Patt $\rightarrow$85
- Ben + Patt $\rightarrow$72
- Adam + Ben + Patt $\rightarrow$90

We take all permutations of the 3 participants in sequence and see the incremental payout that each of them has to make

So How Do You Please 3 Friends?

- Adam →80
- Ben →56
- Patt →70
- Adam + Ben →80
- Adam + Patt →85
- Ben + Patt →72
- Adam + Ben + Patt →90

We take all permutations of the 3 participants in sequence and see the incremental payout that each of them has to make

- (Adam, Ben, Patt) →(80, 0, 10)
- (Ben, Adam, Patt) →(56, 24, 10)
- (Ben, Patt, Adam) →(56, 16, 18)
- (Patt, Adam, Ben) →(70, 15, 5)
- (Patt, Ben, Adam) →(70, 2, 18)
- (Adam, Patt, Ben) →(80, 5, 5)

How Much Adam Pays?

- Adam →80
- Ben →56
- Patt →70
- Adam + Ben →80
- Adam + Patt →85
- Ben + Patt →72
- Adam + Ben + Patt →90

- (Adam, Ben, Patt) →(80, 0, 10)
- (Ben, Adam, Patt) →(56, 24, 10)
- (Ben, Patt, Adam) →(56, 16, 18)
- (Patt, Adam, Ben) →(70, 15, 5)
- (Patt, Ben, Adam) →(70, 2, 18)
- (Adam, Patt, Ben) →(80, 5, 5)

How Much Adam Pays?

- Adam $\rightarrow$80
- Ben $\rightarrow$56
- Patt $\rightarrow$70
- Adam + Ben $\rightarrow$80
- Adam + Patt $\rightarrow$85
- Ben + Patt $\rightarrow$72
- Adam + Ben + Patt $\rightarrow$90

- (Adam, Ben, Patt) $\rightarrow$(80, 0, 10)
- (Ben, Adam, Patt) $\rightarrow$(56, 24, 10)
- (Ben, Patt, Adam) $\rightarrow$(56, 16, 18)
- (Patt, Adam, Ben) $\rightarrow$(70, 15, 5)
- (Patt, Ben, Adam) $\rightarrow$(70, 2, 18)
- (Adam, Patt, Ben) $\rightarrow$(80, 5, 5)

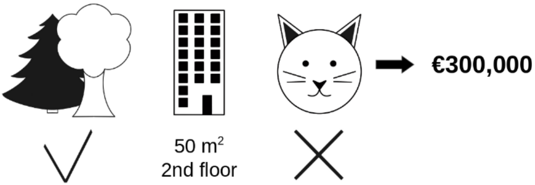
So for Adam, it is $(80 + 24 + 18 + 15 + 18 + 80)/6 =$ €39.2

How Do You Relate To Machine Learning Models?

Machine learning researchers:

- "How should we divide **credit for a prediction from a model** whose **features** made different contributions?"
- Shapley value for **feature** X is the amount of credit it gets.
- Total **prediction** is the sum of Shapley values over **features** (plus the model baseline).
- Linear case is intuitive and simple: $shapleyValue(X_i = x) = coef[i] * (x - mean(X_i))$
- General computation is lengthy...

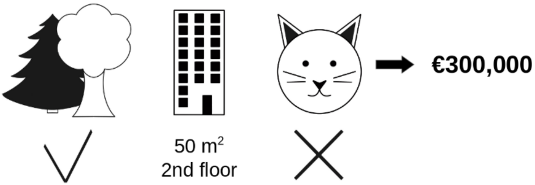
Does Shapley Help Us With This?



Features in action :

- park-nearby
- cat-banned
- area-50m^2
- floor-2nd
- The average prediction (€310,000)

Does Shapley Help Us With This?

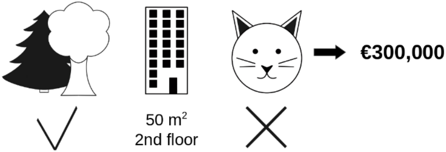


Features in action :

- park-nearby
- cat-banned
- area-50m^2
- floor-2nd
- The average prediction (€310,000)

Our goal is to **explain the difference** between the actual prediction (€300,000) and the average prediction (€310,000): a difference of -€10,000

Does Shapley Help Us With This?



Features in action :

- park-nearby
- cat-banned
- area-50m^2
- floor-2nd
- The average prediction (€310,000)

The answer could be :

- 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

How Do We Calculate Shapley Value For One Feature?

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

How Do We Calculate Shapley Value For One Feature?

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

A linear model prediction for one data instance looks like :

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

How Do We Calculate Shapley Value For One Feature?

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

A linear model prediction for one data instance looks like :

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

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

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

Linear Models Make Life Easy

We sum all the feature contributions for one instance :

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

Linear Models Make Life Easy

We sum all the feature contributions for one instance :

$$\sum_{j=1}^{p} \phi_j(\hat{f}) = \sum_{j=1}^{p} (\beta_j x_j - E(\beta_j X_j))$$

$$= (\beta_0 + \sum_{j=1}^{p} \beta_j x_j) - (\beta_0 + \sum_{j=1}^{p} E(\beta_j X_j))$$

$$= \hat{f}(x) - E(\hat{f}(X))$$

Can we do the same for any type of model?

How Do We Compute Shapley Values?

Shapley value is defined via a characteristic function v of features in S :

$$\varphi_i(v) = \sum_{S \subseteq \{N\} \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S))$$

How Do We Compute Shapley Values?

Shapley value is defined via a characteristic function v of features in S :

$$\varphi_i(v) = \sum_{S \subseteq \{N\} \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!}(v(S \cup \{i\}) - v(S))$$

- N = set of n features
- $v(S)$ = the total expected sum of payoffs the members of S can obtain by cooperation
- $v(S \cup \{i\}) - v(S)$ = each feature demanding their contribution as a fair compensation

How Do We Compute Shapley Values?

An alternative equivalent formula for the Shapley value is: :

$$\varphi_i(v) = \frac{1}{n!} \sum_R [v(P_i^R \cup \{i\}) - v(P_i^R)]$$

How Do We Compute Shapley Values?

An alternative equivalent formula for the Shapley value is: :

$$\varphi_i(v) = \frac{1}{n!} \sum_R [v(P_i^R \cup \{i\}) - v(P_i^R)]$$

- R = ordering of the features
- P_i^R = the set of features in N which precede i in the order R

How Do We Compute Shapley Values?

An alternative equivalent formula for the Shapley value is: :

$$\varphi_i(v) = \frac{1}{n!}\sum_R [v(P_i^R \cup \{i\}) - v(P_i^R)]$$

- R = ordering of the features
- P_i^R = the set of features in N which precede i in the order R

$$\varphi_i(v) = \frac{1}{\text{number of features}} \sum_{\text{conditions excluding } i} \frac{\text{marginal contribution of } i \text{ to coalition}}{\text{number of coalitions excluding } i \text{ of this size}}$$

Let's Solve This Simple Problem?

- Players have left- and right-hand gloves
- The goal is to form pairs
- Compute Shapley value for 'A'

Let's Solve This Simple Problem?

- Players have left- and right-hand gloves
- The goal is to form pairs
- Compute Shapley value for 'A'

$$N = \{A, B, C\}$$

$$v(S) = \begin{cases} 1 & , \text{if} S \in \{\{A, C\}, \{B, C\}, \{A, B, C\}\} \\ 0 & , \text{otherwise} \end{cases}$$

Let's Solve This Simple Problem?

- Players have left- and right-hand gloves
- The goal is to form pairs
- Compute Shapley value for 'A'

$$N = \{A, B, C\}$$

$$v(S) = \begin{cases} 1 & , \text{if} S \in \{\{A, C\}, \{B, C\}, \{A, B, C\}\} \\ 0 & , \text{otherwise} \end{cases}$$

$$\varphi_i(v) = \frac{1}{|N|!} \sum_R [v(P_i^R \cup \{i\}) - v(P_i^R)]$$

Let's Solve This Simple Problem?

- Players have left- and right-hand gloves
- The goal is to form pairs
- Compute Shapley value for 'A'

$$N = \{A, B, C\}$$

$$v(S) = \begin{cases} 1 & ,\text{if} S \in \{\{A, C\}, \{B, C\}, \{A, B, C\}\} \\ 0 & ,\text{otherwise} \end{cases}$$

$$\varphi_i(v) = \frac{1}{|N|!} \sum_R [v(P_i^R \cup \{i\}) - v(P_i^R)]$$

- R = ordering of the players
- P_i^R = the set of players in N which precede i in the order R

Let's Solve This Simple Problem?

$$N = \{A, B, C\}$$

$$v(S) = \begin{cases} 1 & , \text{if} S \in \{\{A, C\}, \{B, C\}, \{A, B, C\}\} \\ 0 & , \text{otherwise} \end{cases}$$

Let's Solve This Simple Problem?

$$N = \{A, B, C\}$$

$$v(S) = \begin{cases} 1 & , \text{if} S \in \{\{A, C\}, \{B, C\}, \{A, B, C\}\} \\ 0 & , \text{otherwise} \end{cases}$$

$$\varphi_i(v) = \frac{1}{|N|!} \sum_R [v(P_i^R \cup \{i\}) - v(P_i^R)]$$

Let's Solve This Simple Problem?

$$N = \{A, B, C\}$$

$$v(S) = \begin{cases} 1 & , \text{if} S \in \{\{A, C\}, \{B, C\}, \{A, B, C\}\} \\ 0 & , \text{otherwise} \end{cases}$$

$$\varphi_i(v) = \frac{1}{|N|!} \sum_R [v(P_i^R \cup \{i\}) - v(P_i^R)]$$

Order R	MC_A
A, B, C	$v(\{A\}) - v(\varnothing) = 0 - 0 = 0$
A, C, B	$v(\{A\}) - v(\varnothing) = 0 - 0 = 0$
B, A, C	$v(\{A, B\}) - v(\{B\}) = 0 - 0 = 0$
B, C, A	$v(\{A, B, C\}) - v(\{B, C\}) = 1 - 1 = 0$
C, A, B	$v(\{A, C\}) - v(\{C\}) = 1 - 0 = 1$
C, B, A	$v(\{A, B, C\}) - v(\{B, C\}) = 1 - 1 = 0$

Let's Solve This Simple Problem?

$$N = \{A, B, C\}$$

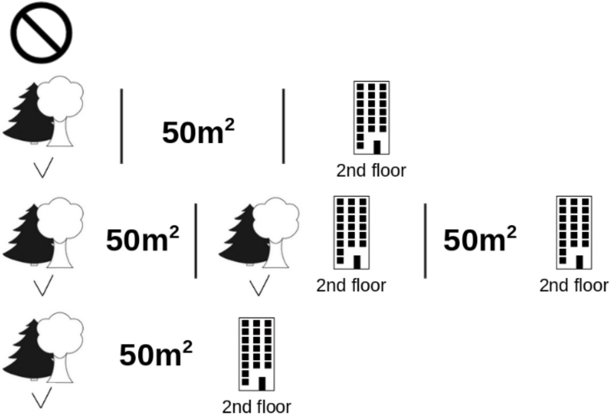
$$v(S) = \begin{cases} 1 & \text{, if} S \in \{\{A, C\}, \{B, C\}, \{A, B, C\}\} \\ 0 & \text{, otherwise} \end{cases}$$

$$\varphi_i(v) = \frac{1}{|N|!} \sum_R [v(P_i^R \cup \{i\}) - v(P_i^R)]$$

Order R	MC_A
A, B, C	$v(\{A\}) - v(\varnothing) = 0 - 0 = 0$
A, C, B	$v(\{A\}) - v(\varnothing) = 0 - 0 = 0$
B, A, C	$v(\{A, B\}) - v(\{B\}) = 0 - 0 = 0$
B, C, A	$v(\{A, B, C\}) - v(\{B, C\}) = 1 - 1 = 0$
C, A, B	$v(\{A, C\}) - v(\{C\}) = 1 - 0 = 1$
C, B, A	$v(\{A, B, C\}) - v(\{B, C\}) = 1 - 1 = 0$

$$\varphi_A(v) = \left(\frac{1}{6}\right)(1) = \frac{1}{6}$$

Hope You Would Solve This?



All 8 coalitions needed for computing the exact Shapley value of the **'cat-banned'** feature value

Might Be Of Interest

- Shapley, Lloyd S. "A value for n-person games." Contributions to the Theory of Games 2.28 (1953): 307-317.
- Štrumbelj, Erik, and Igor Kononenko. "Explaining prediction models and individual predictions with feature contributions." Knowledge and information systems 41.3 (2014): 647-665.
- Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems. 2017.
- Sundararajan, Mukund, and Amir Najmi. "The many Shapley values for model explanation." arXiv preprint arXiv:1908.08474 (2019).
- Janzing, Dominik, Lenon Minorics, and Patrick Blöbaum. "Feature relevance quantification in explainable AI: A causal problem." International Conference on Artificial Intelligence and Statistics. PMLR, 2020.
- Staniak, Mateusz, and Przemyslaw Biecek. "Explanations of model predictions with live and breakDown packages." arXiv preprint arXiv:1804.01955 (2018).

SHAP (SHapley Additive exPlanations)

- The goal of is to explain the prediction of an instance x by computing the contribution of each feature to the prediction.
- SHAP explanation method computes Shapley values.
- A player can be an individual feature value, e.g. for tabular data, player can also be a group of feature values.
- SHAP authors [a] proposed **KernelSHAP** and **TreeSHAP**.

[a] Lundberg and Lee (2016)

Way Forward

- The authors implemented SHAP in the **shap** Python package.
- Works for tree-based models in the **scikit-learn** machine learning library for Python.
- Integrated into the tree boosting frameworks **xgboost** and **LightGBM**.
- In R, there is the **shapper** and **fastshap** packages. SHAP is also included in the R **xgboost** package.

Further Readings

- Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems. 2017.
- Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee. "Consistent individualized feature attribution for tree ensembles." arXiv preprint arXiv:1802.03888 (2018).
- Sundararajan, Mukund, and Amir Najmi. "The many Shapley values for model explanation." arXiv preprint arXiv:1908.08474 (2019).
- Janzing, Dominik, Lenon Minorics, and Patrick Blöbaum. "Feature relevance quantification in explainable AI: A causality problem." arXiv preprint arXiv:1910.13413 (2019)..
- Slack, Dylan, et al. "Fooling lime and shap: Adversarial attacks on post hoc explanation methods." Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020.

Further Readings

- Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems. 2017.
- Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee. "Consistent individualized feature attribution for tree ensembles." arXiv preprint arXiv:1802.03888 (2018).
- Sundararajan, Mukund, and Amir Najmi. "The many Shapley values for model explanation." arXiv preprint arXiv:1908.08474 (2019).
- Janzing, Dominik, Lenon Minorics, and Patrick Blöbaum. "Feature relevance quantification in explainable AI: A causality problem." arXiv preprint arXiv:1910.13413 (2019)..
- Slack, Dylan, et al. "Fooling lime and shap: Adversarial attacks on post hoc explanation methods." Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020.

Information and Image Courtesy:

https://en.wikipedia.org/wiki/Shapley value
https://christophm.github.io/interpretable-ml-book

Thank You...