Cooperative Game Theory and its Applications in Model Explainability

John Fishbein, Jack Prescott

Spring 2021

Contents

1 Motivation for Model Explainability

Model explainability is an incredibly important field. As models grow increasingly complex, they also grow increasingly opaque, and the reasons behind their actions can be difficult to decipher. This can lead to biases against races, genders, and other social groups, which can be hard to spot and harder to remedy. For simpler models this is not a significant problem, because the models themselves can be readily interpreted by a human, and no additional model descriptors are necessary to understand its motivations. For instance, consider a multiple linear regression model. The learned regression coefficients clearly relate the importance of a covariate as a predictor of the response variable via their magnitude, and convey the positivity or negativity of the covariate's correlation with the response variable via their sign. However, no such property of simple self description exists for more complicated models like random forests, other ensembling techniques, or ANNs. These intricate model classes require a higher level descriptor that captures the same easy-to-understand interpretation that is self contained in a traditional linear model [Lundberg and Lee, 2017].

A example of this algorithmic bias has been observed in advertisement recommendation software. A widely accepted use case of machine learning techniques is advertisement recommendation software. Companies like Google and Facebook use their user's data to produce predictions

of what types of Ads a given user will respond positively to. Google will then sell these predictions to advertisers and other interested parties for the purpose of optimizing some metric related to Ad revenue. In 2015, Datta et al. [2015] clearly demonstrated that these mechanisms operate with algorithmic bias. Through several carefully executed experiments, they modified several aspects of google user profiles and observed the ads that were seen by these fake users. In their first study, they randomly created 1000 users and set 500 of them to have a "male" gender and the other "500" to have a female gender. Then, after recording the resulting ads that were seen, they found that the most commonly seen ad for the "male" group was an ad for an executive level career coaching service. They reported that 402 out of the 500 male users were shown this ad at least once, but only 60 out of the 500 female users received the same ad. Clearly, the advertisement recommendation service considers the gender feature as relevant in the advertisement recommendation, and in the case of this particular ad, the predictions contain algorithmic bias [Datta et al., 2015].

2 Overview of Cooperative Game Theory

2.1 Introduction to Cooperative Game Theory

The characteristic functional form, also known as the coalitional form, of cooperative game $\{N, v\}$ is defined by a set $N = \{1, ..., n\}$ of players and a function $v : \mathcal{P}(N) \to \mathbb{R}$ called the characteristic function of the game. Note that v maps a subset, or a "coalition", of players $S \subseteq N$ to a real number $v(S)$ and can be interpreted as the utility of the group of players in the game. The only restriction on this characteristic function is that it must be super-additive, i.e. for any disjoint $S, U \subseteq N$, it follows that $v(S \cup U) \geq v(S) + V(U)$. This implies that the utility of a coalition of two individual players i, j must be at least the sum of their individual utilities [Shapley and Roth, 1988].

Shapley aims to define a function representing the value of playing a cooperative game for each of the n players depending only on the game's characteristic function. That is, the value of a game $\{N, v\}$ denoted $\phi(v)$ is a vector in $\mathbb{R}^n$ such that $\phi_i(v)$ can be thought of as the value to player i in playing the game. This is roughly analogous to how the expected utility sums up the value of a particular move in a normal-form game for one of the game's players in a single number [Shapley and Roth, 1988].

2.2 Axioms of the Shapley Value

Shapley argues that any ϕ that adequately represents the value of a cooperative game must satisfy 5 axioms for any game $\{N, v\}$ [Shapley, 1953]. The axioms are stated as follows:

1. *Symmetry*:
   This axiom states that for any two players $i, j \in N$, if $v(\{i\}) = v(\{j\})$, then it must be that $\phi_i(v) = \phi_j(v)$. In other words, if the characteristic function of the game treats players i and j equivalently, then players i and j should have the same overall value of the game.

2. *Null Player*:
   This axiom states that for any player $i \in N$ that is a Null player, it must be that $\phi_i(v) = 0$. A player $i \in N$ is a null player if for any $S \subseteq N$, we have that $v(S \cup \{i\}) = v(S)$. In other

words, a player is a null player if his precesnse in a given coalition S does not have any impact on the utility of the coalition.

3. *Efficiency*:
   This axiom states that it must be that $\sum_{i \in N} \phi_i(v) = v(N)$. Equivalently, we have that the sum of the values of each player must be equal to the overall utility of the coallition containing all N players.

4. *Additivity*:
   This axiom deals with the interaction of values between different games. It states that for any two cooperative games $\{N, v_1\}$ and $\{N, v_2\}$, it must be that $\phi(v_1) + \phi(v_1) = \phi(v_1 + v_2)$, where $(v_1 + v_2) : \mathcal{P}(N) \to \mathbb{R}$ is the function such that $(v_1 + v_2)(S) = v_1(S) + v_2(S)$.

5. *Monotonicity**:
   This axiom deals again with the comparison of the two different games. For any two cooperative games $\{N, v_1\}$ and $\{N, v_2\}$, it must be that ϕ must satisfy (strong) monotonicity. In this context, it must be that if $v_1(S \cup \{i\}) - v_1(S) \geq v_2(S \cup \{i\}) - v_2(S)$ for all $S \subseteq N \setminus \{1\}$ then $\phi_i(v_1) \geq \phi_i(v_2)$. If the first inequality is strict, then the second must be as well, hence "strong". The quantity $v(S \cup \{i\}) - v(s)$ represents the marginal contribution of i to S. Intuitively, this axiom states that if i has a higher influence in one game compared to another, i must have a higher value for that game as well.

   *Note that the 5-th axiom of monotonicity was not included in Shapley's original paper, but has since been generalized by further work in the field [Young, 1985].

2.3 Uniqueness of the Shapley Value

Not only did Shapley prove the existence of such a value function ϕ, but he also proved the uniqueness of this function: From above, recall the that the characteristic function of a game v is a function from $\mathcal{P}(N) \to \mathbb{R}$. For any finite set N, we know that $|\mathcal{P}(N)| = 2^{|N|}$. Therefore, by enumerating the subsets of N as $\mathcal{P}(N) = \{S_1, , ..., S_{2^{|N|}}\}$, we can simply think of any characteristic function a vector in $v \in \mathbb{R}^{2^{|N|}}$ such that the $i-$th index of the vector contains the value $v(S_i)$. Therefore, since each valid characteristic function must satisfy the property of super-additivity, the set of all characteristic functions is a subset of $\mathbb{R}^n$. By the additivity axiom of ϕ stated above, it is clear that if a given value function ϕ is produced for a set of characteristic functions i.e. vectors in $R^{2^{|N|}}$ that form a basis of the set of all characteristic functions, then the value function will extend to any characteristic function.

The uniqueness proof follows in this vein by considering the following class of games: Fix a subset $R \in \mathcal{P}(N)$. Then, define the characteristic function of the game as

$$v_R(S) = \begin{cases} 1 & R \subset S \\ 0 & R \not\subset S \end{cases} \quad \text{for any subset S } \in \mathcal{P}(N)$$

By this definition, for any R, it is clear to see that every player $i \notin R$ is a null player and thus by axiom 2 of the value function we know that $\phi_i(v_R) = 0$. Furthermore, we also have that all players $i \in R$ must be symmetric. This follows since if $|R| > 1$, we have that for any player $i \in R$, $v(\{i\}) = 0$ since $R \not\subset S$. Therefore, by axiom 1 above for any $i, j \in R$, we know that $\phi_i(v_R) = \phi_j(v_R)$. Finally, by the efficiency axiom, combining the prior two facts, we know that

$\sum_{i \in N} \phi_i(v_R) = \sum_{i \in R} \phi_i(V_R) = v(N) = 1$ and thus $\phi_i(V_R) = \frac{1}{|R|}$ for all $i \in R$. Therefore, we have shown that such a value ϕ is uniquely defined for any game of the form v_R. Furthermore, for any constant $c \in \mathbb{R}$, it follows analogously that the value is unique in the game defined by the function cv_R. Notice that set of characteristic functions v_R form a basis of the set of all possible characteristic functions. This follows both from the fact that any characteristic function v can be expressed as a linear combination of the characteristic functions v_R, and from the fact that the set of characteristic functions v_R is itself linearly independent. By the additivity axiom, since the value ϕ is uniquely determined for any characteristic function cv_R, it follows immediately that $\phi(v)$ must be unique [Shapley, 1953].

Shapley proved that this unique value ϕ of a game $\{N, v\}$ called the Shapley value is defined as follows:

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

The Shapley value of a player $i \in N$, $\phi_i(v)$ represents the player's "value" of the game. Here, Shapley proves that a player's value under the above described assumptions can be uniquely expressed as a weighted average of the given player i's aggregated marginal contribution to the coalitions $S \subset N$. With respect to a specific coalition S, player i's marginal contribution can be quantified as $v(S \cup \{i\}) - v(S)$. Intuitively, this is how much the utility of the coalition increases due to i's contribution. This is a seminal result in the analysis of cooperative games and can be used in many diverse applications to gain a deeper understanding of the inter-workings of cooperative behavior [Shapley, 1953].

3 Cooperative Game Theory Applied to Model Explainability

3.1 History of Cooperative Game Theory in Model Explainability

As best we were able to tell in our research, the earliest instance of using game theory in the context of model explainability was from Lipovetsky and Conklin [2001]. In their paper, entitled *Analysis of Regression in Game Theory Approach*, the authors propose the Shapley value as a method for evaluating feature importance specifically in the context of multiple regression. They say that while there exist numerous well known methods of feature importance evaluation tools, like the t-tests and the coefficient of multiple determination, there had not previously existed any methods which were able to properly handle multicollinearity. This leads to significant undesirable and unpredictable results in variable importance estimates. They propose the Shapley value (1) as a novel method of quantifying the importance of coefficient in multiple regression which is uniquely effective in the presence of multicollinearity. A significant hurdle in applying the Shapley value to explain more general models is that calculating the Shapley value precisely is a combinatorially large #P-complete problem [Aziz and de Keijzer, 2013], which is infeasible for models with thousands of input parameters. Under the general framework proposed by a more recent paper in this area [Lundberg and Lee, 2017] (detailed in Section 3.3), the exact Shapley value for a particular covariate i is related to the change in model prediction when the covariate is added to the predictive model's input. Particularly, because the order in which the covariates are sequentially added back to the model is important, it is the average of these output changes over all possible orderings of adding the covariates to the model input. Lipovetsky and Conklin [2001] propose sampling from the set of necessary computations in such a way that the Shapley

value can be approximated. The paper by Lundberg and Lee [2017] actually showed this to be a fairly effective approach in more complex settings, though somewhat inefficient and not as accurate as the newer paper's novel proposal.

Furthermore, the first instance we were able to uncover which used the Shapley value to put forth a more general framework of model explainability was the work of trumbelj and Kononenko [2010] in 2010. Alternatively to prior attempts at model explainability, they proposed a general, model agnostic approach to explain individual predictions. Their solution frames the problem of explaining predictions as a cooperative game, and uses the Shapley value to reveal the influence of given feature values. While this work made groundbreaking strides bridging the gap between Explainability and Cooperative game theory, the proposed solution still lacks a level of generality to be desired. First, the solution provides general explainability to the predictions of a classifier, but there are cases in which this is not sufficient. As seen in the many cases of algorithmic bias in today's landscape, it is necessary to generalize the quantity being explained to account for more abstract settings other than simply the prediction of the model. Furthermore, this proposed framework operates under the assumption that feature values are drawn from uniform distribution over the feature values. This assumption in some cases can cause the overall explanation to not adequately reflect the correlation between features that can exist in models in practice. They used randomized approximation algorithms based on similar types of sampling techniques first published by Lipovetsky and Conklin [2001], and demonstrate that these are effective to within certain bounds. As we will discuss below, these levels of abstraction have since been researched further and improved upon.

3.2 Algorithmic Transparency via Quantitative Input Influence [Datta et al., 2016]

In 2016, Datta et al. [2016] put forth a general explainability framework using the Shapley value in order to combat this issue of algorithmic bias that is prevalent in the field of Machine Learning. This work generalizes prior applications of cooperative game theory in the model explainability. In this field's seminal work, they introduce a family of *Quantitative Input Influence (QII)* functions to represent the relative influence of a feature on a given model's predictions. The purpose of the QII measure is to provide insight into how a complex, black-box machine learning model produces its predictions and assists in evaluating the efficacy of those predictions. This, in turn, ultimately could be used to reduce algorithmic bias in enterprise machine learning. In order to achieve these goals, "3 desiderata drove the definitions of these [QII] measures" [Datta et al., 2016]:

1. **Generalized Transparency Reports**
   In order to begin to discuss the influence of an input, we need to thoroughly define the quantity against which we will measure this influence. To accomplish this, Datta et al. [2016] define a *Quantity of Interest* to be a property of of the behavior of a system given an input distribution. This abstract definition gives rise to many possible metrics which can be evaluated against within this proposed framework.

2. **Input Influence Quantification of Correlated Inputs**
   In applications of supervised machine learning, the features used almost always have some level of correlation. For a given input of interest, in order to adequately quantify the overall influence of that feature from a causal standpoint, it must be done independently

of the feature's correlation with the rest of the dataset. In order to achieve this goal, the proposed QII measures work on a given model with two separate input datasets. The first is a sample input data set drawn from the original input distribution. The second dataset is a hypothetical distribution computed from the first by holding all features constant except for the feature of interest. In this hypothetical distribution, the feature of interest is replaced by randomly sampling a new value from its own marginal distribution, thereby removing its correlation from the other inputs. In effect, this technique of *Causal Intervention* allows the influence of the input of interest to be measured independently of it's correlated inputs. At a high level, the QII measure then works by measuring the Quantity of Interest with respect to each dataset and then computing the difference between the two [4].

3. **Joint & Marginal Input Influence Quantification w.r.t. Multiple Features** Finally, there are many cases in machine learning applications in which no single feature has any meaningful influence on the overall output of the model. In cases like these, we wish to quantify the influence of a set of inputs rather than just a single input. In order to compute this joint influence, the method of *Causal Intervention* from (2) can be naturally extended by replacing the entire set of inputs of interest with a random sample from it's joint prior distribution in the second manufactured distribution. Furthermore we may wish to quantify the marginal influence of a given input of interest within this joint influence. Viewing this in the context of a cooperative game, this goal can be achieved by measuring the difference in the quantity of interest when considering the joint influence of sets of inputs both with and without that particular input. In this game-theorietic setting, we can also consider the aggregate marginal influence of a given input with respect to many different sets of inputs. [4].

Formally, Datta et al. [2016] introduce the following 3 definitions in the context of model explainability: Suppose that a given system A has inputs $N = \{1, ..., n\}$. Let $X = (x_1, ..., x_i, ..., x_n)$ be a random vector representing the true input distribution of a given system A and let $Q_A(\cdot)$ be the desired quantity of interest of the model A with respect to an input distribution. Furthermore, suppose that $i \in N$ is the. feature of interest. Then, denote the intervened distribution as $X_{-i}U_i = (x_1, ..., u_i, ..., x_n)$ which is computed through Causal Intervention. Similarly, if $S \subseteq N$ consist of multiple features of interest, denote the intervened distribution as $X_{-S}U_S$. The definitions introduced by [Datta et al., 2016] are as follows:

1. The **Quantitative Input Influence (Unary QII)** of an input $i \in N$ on the quantity of interest $Q_A(\cdot)$ is defined to be the difference in the quantity of interest between the true and intervened datasets X and $X_{-i}U_i$

$$\iota^{Q_A}(i) = Q_A(X) - Q_A(X_{-i}U_i) \tag{2}$$

2. The **Quantitative Input Influence (Set QII)** of the set $S \subseteq N$ on the quantity of interest $Q_A(\cdot)$ is defined to be as the difference in the quantity of interest between the true and intervened datasets X and $X_{-S}U_S$

$$\iota^{Q_A}(S) = Q_A(X) - Q_A(X_{-S}U_S) \tag{3}$$

3. The **Quantitative Input Influence (Marginal QII)** of an input $i \in N$ over a set $S \subseteq N$ on the quantity of interest $Q_A(\cdot)$ is defined to be as the difference in the quantity of interest

between the datasets intervened on the set S and the set $S \cup \{i\}$.

$$\iota^{Q_A}(i, S) = Q_A(X_{-S}U_S) - Q_A(X_{-S\cup\{i\}}U_{S\cup\{i\}}) \tag{4}$$

Recall from above, the Shapley value $\phi_i(v)$ (1) quantifies the marginal contribution of a "player" i in a cooperative game with respect to the game's characteristic function v. In the context of input influence in machine learning models, while operating under the natural axioms stated in Section 1.2 given the Shapley value is the *unique* way of producing a measure with the desired significance and generality. Recall from Section 3.1 the computational difficulty of calculating the exact Shapely value. They propose a novel $\epsilon - \delta$ approximation scheme for the Shapley value [Bachrach et al., 2010], which is an improvement over past methods, though still fairly inefficient and similar to the sampling technique used by trumbelj and Kononenko [2010] and Lipovetsky and Conklin [2001]. It would soon be surpassed in efficacy by the next publication we will discuss.

3.3 A Unified Approach to Interpreting Model Predictions [Lundberg and Lee, 2017]

Datta et al. [2016] and number of other researchers ([8], [11], [14], [6], [2]) have aimed to provide similar tools to help improve the field of model explainability. However, these approaches were seemingly somewhat disparate, and little was known about their relationship to one another or in what situation certain techniques were preferable. The work done by Lundberg and Lee [2017] unifies these approaches under a single class of what they call additive feature attribution models. They go on to demonstrate this unification by showing that all of these previous methods are really approximating the same value, which they name the Shapley Additive Explanations value, or SHAP. Lastly, they propose novel SHAP estimation methods and provide evidence for their superior efficacy with respect to the previous models by comparing the results of the different models to empirically determined human intuition and showing they in some cases their model has increased performance as a discriminator between output classes [Lundberg and Lee, 2017].

3.3.1 Additive Feature Attribution Models

They define this general class of explanation models which try to explain the importance of different input features to the predictive model $f(x)$'s output, again called additive feature attribution models, as follows: First, a particular input z whose corresponding predictive model output we are seeking to explain is mapped to a simplified variable z', which is a vector of binary inputs $\{0,1\}^M$. We then seek to come up with a function $g(x)$, associated with a particular input z, such that $g_z(z') \approx f(z)$. Each $g(x)$ is a linear in its input variables, that is

$$g_z(z') = \phi_0 + \sum_{i=1}^{M} \phi \cdot z'_i \tag{5}$$

They discuss how each of the aforementioned explanation models do indeed fit into this framework. Next, they demonstrate three properties that models of this class share.

1. **Local Accuracy**: For a given z, $g_z(z') = f(z)$.
2. **Missingness**: $z'_i = 0 \implies \phi_i = 0$, meaning we don't attribute any impact on the predictive model's output to covariates which are toggled off in the simplified binary mapping z' of the input variable z.

3. **Consistency**: If a covariate i's contribution to the model output is larger in some model $f'(x)$ than in $f(x)$, which is determined by comparing each model's output before and after removing i from the input sample, then the explanation model's attribution ϕ_i for $f'(x)$ should be greater than its corresponding attribution for $f(x)$.

They then generalize Datta et al. [2016]'s theorem that the Shapley value is the only explanation model that satisfies their three more specific explanation model criteria by showing that it is also the unique explanation model which satisfies these three criteria [Lundberg and Lee, 2017].

3.3.2 SHAP Estimation

Recall the precise method of computation of the Shapley value in this framework, seen in Section 3.1. They go on to propose a novel method of estimating the SHAP, by building on a previously proposed method called Linear LIME [Ribeiro et al., 2016]. The high level goal of this method is to locally approximate the prediction function $f(x)$ with a linear model, whose coefficients represent the feature importance of their respective covariates. The insight by Lundberg and Lee [2017] is that particular choices for the Linear LIME's loss function, weighting kernel, and regularization term can ensure that the method's solution recovers the Shapley values for each of the covariates, which is desirable because it means this explanation model will exhibit local accuracy, missingness, and consistency. They call this method Kernel SHAP, and go on to demonstrate its superior performance in a number of settings, both in terms of efficiency and accuracy, as compared to any previously proposed approaches.

4 Conclusion

While the fields of cooperative game theory and model explainability may seem entirely disconnected to the layman, their connection has been made evident by several waves of excellent researchers. The elegant power of the Shapley value in solving many of the longstanding problems in model explainability was first published by Lipovetsky and Conklin [2001]. Their work made the connection known, but failed to attract significant attention at the time, and was lacking in both generality and in the sophistication of their Shapley value approximation techniques. trumbelj and Kononenko [2010] were the first group to broaden this approach to apply to arbitrary classification models, though they did little to innovate on previous Shapley estimation methodology. Datta et al. [2016]'s work went even further in the generalization of the cooperative game theoretic approach to model explainability, and was the first very widely noticed paper in the field. Finally, Lundberg and Lee [2017] tied all the previous work together, creating a fully general definition of explanation models and unifying all notable past approaches under a class of models they call additive feature attribution models. They went on to show how a technique previously used to attribute importance to model inputs [Ribeiro et al., 2016], which seemingly had no connection to game theory, could be carefully tuned to efficiently and accurately recover the Shapley values of a model's inputs. Since then, other papers have attempted to push the boundaries of this field even further, but these seminal papers brought this area of study into the mainstream, and spawned off startups [Datta et al., 2016] and open source projects [Lundberg and Lee, 2017] which aim to use these profound results to effect meaningful change beyond academia.

References

Haris Aziz and Bart de Keijzer. Shapley Meets Shapley. *arXiv:1307.0332 [cs]*, July 2013. URL http://arxiv.org/abs/1307.0332. arXiv: 1307.0332.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140, July 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0130140. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130140. Publisher: Public Library of Science.

Yoram Bachrach, Evangelos Markakis, Ezra Resnick, Ariel D. Procaccia, Jeffrey S. Rosenschein, and Amin Saberi. Approximating power indices: theoretical and empirical analysis. *Autonomous Agents and Multi-Agent Systems*, 20(2):105–122, March 2010. ISSN 1573-7454. doi: 10.1007/s10458-009-9078-9. URL https://doi.org/10.1007/s10458-009-9078-9.

A. Datta, S. Sen, and Y. Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 598–617, 2016. doi: 10.1109/SP.2016.42.

Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. *arXiv:1408.6491 [cs]*, March 2015. URL http://arxiv.org/abs/1408.6491. arXiv: 1408.6491.

Stan Lipovetsky and Michael Conklin. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17:319 – 330, 10 2001. doi: 10.1002/asmb.446.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. 30, 2017. URL https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016. URL http://arxiv.org/abs/1602.04938.

L S. Shapley. *17. A Value for n-Person Games*. Princeton University Press, March 1953. ISBN 978-1-4008-8197-0. URL https://www.degruyter.com/document/doi/10.1515/9781400881970-018/html. Pages: 307-318 Publication Title: Contributions to the Theory of Games (AM-28), Volume II Section: Contributions to the Theory of Games (AM-28), Volume II.

Lloyd S. Shapley and Alvin E. Roth, editors. *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press, Cambridge [Cambridgeshire] ; New York, 1988. ISBN 978-0-521-36177-4.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *CoRR*, abs/1605.01713, 2016. URL http://arxiv.org/abs/1605.01713.

Erik Ã... Â trumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11(1):1–18, 2010. URL http://jmlr.org/papers/v11/strumbelj10a.html.

H. P. Young. Monotonic solutions of cooperative games. *International Journal of Game Theory*, 14 (2):65–72, June 1985. ISSN 1432-1270. doi: 10.1007/BF01769885. URL https://doi.org/10.1007/BF01769885.

Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665, December 2014. ISSN 0219-3116. doi: 10.1007/s10115-013-0679-x. URL https://doi.org/10.1007/s10115-013-0679-x.