# Causal Explanations for Performance in Radio Networks

Domokos M. Kelen[1], Péter Kersch[2] and András A. Benczúr[1]

[1]Insitute for Computer Science and Control (SZTAKI), 1111 Budapest, Kende u. 13-17, Hungary
[2]Ericsson, 1117 Budapest, Magyar Tudósok Körútja 11, Hungary

## Abstract

Towards machine learning (ML) for automated radio network control, we demonstrate how Explainable AI (XAI) can be used to explain and distinguish the effect of different factors to network performance KPIs. Additive local feature attribution methods like SHAP promise a model agnostic way to gain insight into a ML model. However, there are many variations of these methods, including marginal and conditional SHAP, as well as ways to calculate attributions that respect the causal structure of the data, such as Asymmetric SHAP. We describe two new approaches to calculating attributions that follow causal relationships more closely. One approach is based on marginal explanations of a model trained in a specific way, while the other is based on calculating the equivalent of conditional Asymmetric SHAP by training multiple models. We demonstrate the approaches on both synthetic and real-world data.

## Keywords

Explainable AI, Radio networks, SHAP, causality

## 1. Introduction

Using AI for automated radio network control is a fundamental challenge for future generation radio networks. Towards automation, a promising direction is Explainable AI (XAI), whose goal is to investigate tools and techniques aimed at opening the so-called opaque (or black-box) models (e.g., deep neural networks, DNN) or at devising intrinsically interpretable and accurate models (e.g., rule-based systems). More expressive models such as DNN or gradient boosted tree models can better fit the observation, however, they are more complex themselves. While more expressive models promise better insights, their complexity makes them more opaque, making it harder to successfully infer useful information about the system.

In this paper, we propose an XAI method to explain and distinguish the effect of different factors on network performance metrics. The main technical difficulty is the complex causal relationship between these factors. For example, antenna tilt configuration affects network performance indirectly via coverage, interference and cell load. These indirect metrics have a more direct effect on network performance, hence XAI will primarily find the importance of coverage, interference, and load and explain the predicted performance metric based on them, mostly ignoring the explanation of how network configuration affects this performance metric.

*AI6G'22: First International Workshop on Artificial Intelligence in beyond 5G and 6G Wireless Networks, July 21, 2022, Padua, Italy*

kdomokos@sztaki.hu (D. M. Kelen); peter.kersch@ericsson.com (P. Kersch); benczur@sztaki.hu (A. A. Benczúr)

Local feature attribution methods based on Shapley values, such as SHAP [1], provide us with a tool to gain insight into the behavior of any model in a black-box manner. Their recent popularity illustrates their usefulness in many domains, explaining predictions in terms of contributions provided by the knowledge of individual feature values. However, there are many variations of such methods, and it is important to understand the advantages and disadvantages of conditional, marginal, and other SHAP attributions [2]. Further, it can be challenging to make sense of SHAP values in a causal setting [3].

In this paper, we describe our use of local feature attributions for explaining radio network performance. Radio networks used in telecommunications are complex systems with many different, inter-related variables, and various causal structures between these variables. Our primary goal is root cause analysis for specific behavior, and our approach is to model the overall behavior of the system and investigate the model using explanation methods. It is, however, very important that the explanations take the causal relationships between the model features into consideration, as failing to do this could easily cause misattributions in our analysis.

Reliably inferring causal relationships from observational data is generally considered to be impossible [4]. Rather than inferring the relationship, causal attributions [3, 5] assume to know the nature of the causal relationship based on domain knowledge and attempt to calculate attributions that respect these relationships. One way to do this is through the use of Asymmetric SHAP [3], a variation of SHAP that is better equipped to deal with causality.

In our use case, we measure the contributions of feature groups, while assuming to know the causal structure between the groups. An important distinction between different Shapley-value-based methods is whether they use a conditional or marginal value function for evaluating the model. A conditional value function is a way to explain the data, while a marginal value function results in explanations that reflect the behavior of the model more closely [2]. Our aim is to gain insight into the behavior of the underlying system through examining the data; modeling and model explanation is only used as a tool towards this end. It thus seems appropriate to use conditional SHAP, however as it turns out we can also use marginal SHAP as a tool to force the explanations to better reflect the causal structure of the data.

The rest of this paper is structured as follows. In Section 2, we describe the mathematical background of SHAP and related terms. In Section 3, we describe two ways of calculating causal attributions for our use case; one based on marginal explanations of a causally-trained model, and one that is based on calculating the equivalent of conditional SHAP by training multiple models. In Section 4, we describe the results of our experiments on simulated and real-world data to illustrate our method. Finally in Section 5, we share our conclusions.

## 2. Background

### 2.1. SHAP

The notion of Shapley values [6] originates from cooperative game theory and is a way to *fairly* calculate the contribution of individual players towards a shared objective. In this context, *fairly* is used in the sense that the provided values satisfy a number of conditions described in the original paper. The method is based on measuring the performance towards the objective with different coalitions of players participating and combining these results in a certain way to calculate individual contributions.

Using Shapley values to calculate additive local feature attributions for machine learning models gained popularity in recent years. SHAP [1] and related methods provide a compelling way to gain insight into the predictions made by any model, by quantifying the value that the knowledge of each feature value contributed towards a specific prediction value.

Given model $f$, feature set $\mathbb{S}$, and a specific point $x$ from the dataset, the Shapley value $\phi$ for feature $j$ can be defined as

$$\phi_j(f,x) = \sum_{S \subseteq \mathbb{S} \setminus \{j\}} \frac{|S|!\,(M-|S|-1)!}{M!} \varphi_j(f,x,S)\,, \quad \text{where} \tag{1}$$

$$\varphi_j(f,x,S) = v_{f(x)}(S \cup \{j\}) - v_{f(x)}(S). \tag{2}$$

Here $v_{f(x)}(S)$ is called the *value function*, and represents the function of evaluating the model $f$ at point $x$ while only using the coalition $S$ of features. It is, however, not evident how we can evaluate a model with coalitions of varying sizes, when the model was trained using a fixed number of features. Note that an alternative but equivalent formulation of Equation 1 is

$$\phi_j(f,x) = \sum_{\pi \in \Pi} \frac{1}{n!} \Big[ v_{f(x)}(\{i : \pi(i) \leq \pi(j)\}) - v_{f(x)}(\{i : \pi(i) < \pi(j)\}) \Big], \tag{3}$$

where $\Pi$ represents the set of all permutations of the ordering of model features.

We describe two ways of defining the value function $v_{f(x)}(S)$ corresponding to the *marginal* and *conditional* variants of SHAP values. In the *marginal* or *interventional* SHAP variant

$$v_{f(x)}(S) = E_{p(x')} \left[ f(x_S \sqcup x'_{\overline{S}}) \right], \tag{4}$$

where the expression $f(x_S \sqcup x'_{\overline{S}})$ represents the value of evaluating function $f$ with the in-coalition feature values $s \in S$ taken from $x$, and the out-of-coalition feature values taken from an identically distributed random $x'$. In contrast, in the *conditional* or *observational* variant

$$v_{f(x)}(S) = E_{p(x' | x'_S = x_S)} \left[ f(x_S \sqcup x'_{\overline{S}}) \right], \tag{5}$$

where the distribution of $x'$ is conditioned on the known feature values $S$ from $x$.

The difference between these two variants is whether the expectation that is taken over the out-of-coalition features is conditioned on the values of the in-coalition features. Marginal SHAP values can be estimated by sampling out-of-coalition values from the training data for evaluating the value function, while conditional sampling is much harder to do and can itself become a separate modeling task [2]. However, an even more important distinction is whether the resulting explanation more strictly adheres to the model or the data.

Marginal SHAP explanations reflect the behavior of the model, meaning that the explanations respect whatever rules the model infers from the data. If there are multiple equivalent ways in which the prediction can be calculated, then the explanations reflect whichever way the model itself uses. Another important aspect of marginal SHAP is a common criticism that the evaluation can also include invalid data points, where combining the in-coalition and the sampled out-of-coalition features results in a data point that is unrealistic in the real world.

On the other hand, conditional SHAP explanations reflect the dataset in that they are not dependent on which equivalent formulation of prediction the model uses. As a simple example, if the dataset contains two features that are identical, then the model is free to choose either of these (or any convex combination of them) for the prediction formula. Marginal SHAP attributions will mirror the model's choice, while conditional SHAP attributions will share the contribution equally between these features regardless of the model itself. This is often cited as a drawback of marginal SHAP, as it means that the model explanation can reveal associations that are not present in the model itself.

### 2.2. Asymmetric SHAP

In [3], Asymmetric SHAP is described as an approach to handling causal relationships between features. The method is based on breaking one of the fundamental requirements of the original Shapley values, symmetry. This requirement guarantees that if two features behave exactly the same way w.r.t. the value function, then their assigned contributions must also be equal. The way this relates to causality is that if these two variables have a direct causal relationship, then it is reasonable to want the causing variable to get a higher share of the contribution.

Asymmetric SHAP modifies Equation 3 to only include permutations where for each causal relationship, the causing variable is ordered before the affected variable, or more generally, allows any weighting scheme over the permutations instead of uniform weights:

$$\phi_j^{\omega}(f) = \sum_{\pi \in \Pi} \omega(\pi) \Big[ v_{f(x)}(\{i : \pi(i) \leq \pi(j)\}) - v_{f(x)}(\{i : \pi(i) < \pi(j)\}) \Big]. \tag{6}$$

In a causal setting, the above described weighting scheme is suggested, meaning that we modify the weights such that $\omega(\pi) \propto 1$ if $\pi(i) < \pi(j)$ for each $i, j$ where $i$ is a causal ancestor of $j$, and $\omega(\pi) = 0$ otherwise. The idea behind ordering causing variables first is that these get the attribution for the change they cause in the expected value of the prediction, while the affected variables get the change that is caused by including them *after* the causing variables.

## 3. Proposed approaches

### 3.1. Causal graph of feature groups

Our approaches rely on grouping the variables and assuming to know the causal relationships between groups. From an evaluation perspective, we treat a feature group as a single, multi-dimensional feature, meaning that while calculating SHAP, either all features of a feature group are included in a coalition or none of them are. This way, we get overall contributions for each group.

The causal relations between the groups can be represented in a directed acyclic graph, where the nodes are the feature groups, and the arcs point from causing group to affected group. This way we can say that feature $i$ is a causal ancestor of feature $j$ if and only if there is a path from $i$ to $j$ in the graph. We can see an example of such a graph in Figure 1. A topological ordering of this graph has the property that each variable precedes its causal descendants. In fact, the set of permutations with nonzero weight in our Asymmetric SHAP calculation, as described in Section 2.2, is exactly the set of all possible topological orderings of this graph.

## 3.2. Explaining a model trained in causal order

One way to have the attributions respect causal relationships is by using marginal SHAP to explain a model that itself calculates the prediction in a way that respects causal relationships. Having assumed a causal order of features, as described in Section 3.1, we can attempt to force the model to prefer using features that appear earlier in the causal ordering, in hopes that the resulting explanation also attributes the prediction accordingly. If there are multiple causal orderings possible than we need to calculate the average of all of these with equal weight.

In our modeling, we primarily use gradient boosted decision trees [7]. These are models that use a linear combination of decision trees for predicting the output. During the training process, the linear combination is built up gradually, adding more decision trees to the result in each step. We can attempt to force this process to yield a model that respects a causal ordering. We do this by restricting the decision trees to only include features from a set that we gradually increase according to the causal ordering throughout the training process. In our case, this is done by first only using features from the first group and training until convergence, then continuing using features from the first two groups, etc. This way, the final model has access to all of the features, while also placing more emphasis on features that appear earlier in the causal ordering.

One possible failure mode of this method is that there is no guarantee that the final model will be different from one that was trained with all features available at the same time. To see this, assume that we have the model $f_1(X_1)$ after training only using the first feature $X_1$. In the next step, the modeling process can continue by adding $f_2(X_1, X_2) = -f_1(X_1) + f_2'(X_1, X_2)$ such that $f_1 + f_2 = f_2'$. In other words, it is possible to cancel out the previous modeling steps in later steps. We can attempt to use regularization (eg. trying to limit capacity) to discourage the model from doing this, however, if the advantage of using $X_2$ is great enough, it can overpower regularization.

## 3.3. Conditional Asymmetric SHAP using multiple models

In this section, we describe a way of using the already in place modeling process for conditioning on the in-coalition feature values. Similar ideas are also described in [8] and [9], where the training of separate models is proposed to calculate feature importances in regression problems to deal with multicollinearity. Conditional SHAP values are hard to calculate, and trying to sample feature values according to their conditional distribution can result in a separate modeling task itself [2], as filtering the data to the subset that fits the condition can result in a very low number of records, or in many cases could essentially equal the data point $x$ itself.

Let us assume that we can train a separate model for each coalition, i.e. train $f_S(x)$ for each coalition $S$, where the model is trained using only the features in $S$. Using these, we can define

$$\phi_j = \sum_{S \subseteq \mathbb{S} \setminus \{j\}} \frac{|S|!\,(M - |S| - 1)!}{M!} \left(f_{S \cup \{j\}}(x) - f_S(x)\right), \tag{7}$$

meaning that the value function is simply equal to $f_S(x)$ instead of Equation 5. We give a heuristic argument that these two calculations approximate the same value. When modeling a regression task, we are in fact approximating an ideal function

$$f(x) \approx f^*(x) = E_{p(y|x)}\left[y\right], \tag{8}$$

with $y$ being the target variable. The SHAP value $\phi_j(f)$ then itself can then be viewed as an approximation of $\phi_j(f^*)$, the feature contributions calculated for the ideal function $f^*$. In this sense, Equation 7 is just another way of approximating the same value that conditional SHAP estimates. This can be seen by observing the following about $f_S(x)$ and $v_{f(x)}(S)$:

$$f_S(x) \approx f^*_S(x) = E_{p(y|x_S)}\left[y\right] \text{ and} \tag{9}$$

$$\begin{aligned} v_{f(x)}(S) &= E_{p(x'|x_S)}\left[f(x_S \sqcup x'_{\overline{S}})\right] \approx E_{p(x'|x_S)}\left[f^*(x_S \sqcup x'_{\overline{S}})\right] \\ &= E_{p(x'|x_S)}\left[E_{p(y|x',x_S)}[y]\right] = E_{p(y|x_S)}\left[y\right]. \end{aligned} \tag{10}$$

In Equations 9-10 we use Equation 8, while in Equation 10 we also use law of total expectation.

While this way we have to train a separate model for each coalition, the problem of conditioning part of the feature set on a given coalition is solved by re-using the same modeling process that we already have in place. Since we already trust this process in delivering insights about the dataset, this saves us from dealing with the conditioning as a separate sampling or modeling step. Fortunately, Asymmetric SHAP greatly reduces the number of coalitions that are of interest, as we only need to perform the evaluation for starting subsets of causal orderings of features, which makes this method computationally feasible in our case.

Given a single causal ordering, the contribution for a specific feature is given by the change in prediction when first including the feature in the set of features available for the model, as formalized by Equation 6. Similar to regular SHAP, we start by calculating $\phi_0 = E(y)$, which can be interpreted as the best possible model while using an empty set of features as coalition.

## 4. Experiments

We run two kinds of experiments. The first one is a synthetic task where the variables and functions involved are simple enough that we know the exact result that we can expect from the explanation methods. The second is an example with real-world data.

We report global feature importance. Since simply averaging contributions over all records results in values with an expected value of zero, importance is calculated by taking average of the absolute values of the contributions. One unfortunate consequence of using this metric is that the importances don't sum up to some fixed value. On our tables, *Causal-order training* refers to the approach described in Section 3.2, while *Separate models* refers to the approach described in Section 3.3. We denote the importance of variable $X$ by $\Phi_X$.

### 4.1. Synthetic data experiment

Here we demonstrate our methods on synthetic data. We start by defining 5 variables. Let

$$A, B, C \sim \mathcal{U}_{[0,1]}, \quad D = A + B + C, \text{ and } Y = 2(A + B + C) + N, \tag{11}$$

**Table 1**
Average absolute feature attributions made by different methods on synthetic data.

| | Expected value | $\Phi_A$ | $\Phi_B$ | $\Phi_C$ | $\Phi_D$ |
|---|---|---|---|---|---|
| Regular model | 3.0073 | 0.0029 | 0.0023 | 0.0027 | 0.7928 |
| Causal-order training | 3.0070 | 0.5062 | 0.4973 | 0.5052 | 0.0003 |
| Separate models | 3.0103 | 0.4829 | 0.4824 | 0.5116 | 0.0000 |

where $\mathcal{U}_{[0,1]}$ denotes the uniform distribution over the $[0,1]$ interval, and $N \sim N(0,0.1)$ is a normally distributed noise variable. We treat each variable as its own separate group, and assume a causal order $[A,B,C,D]$. The variable $Y$ is the target variable.

It is clear that the target variable can be approximated from either variables $A,B,C$ or variable $D$, with the latter having a mode direct relationship with the target. Let us assume that the modeling process can learn the relationship between the features and the target exactly. We can express the expected feature importance analytically for each of the cases where either $A,B,C$ or $D$ get full contribution for the calculating the prediction:

$$\Phi_A = \Phi_B = \Phi_C = 2\int_0^1 \big|a - E(A)\big| \,\mathrm{d}a = 0.5 \tag{12}$$

$$\Phi_D = 2\int_0^1\int_0^1\int_0^1 \big|a+b+c - E(A+B+C)\big| \,\mathrm{d}a\,\mathrm{d}b\,\mathrm{d}c \approx 0.8194. \tag{13}$$

This essentially means that in the case of marginal SHAP, on one end we get $\Phi_A = \Phi_B = \Phi_C = 0.5$ and $\Phi_D = 0$, or on the other end we get $\Phi_A = \Phi_B = \Phi_C = 0$ and $\Phi_D \approx 0.82$. Any convex combination of these two cases is possible as an actual modeling result. Conditional SHAP is expected to distribute the contributions uniformly, meaning a feature importance of $\Phi_A = \Phi_B = \Phi_C = 0.25$ and $\Phi_D \approx 0.41$.

We model on a training set of $10^4$ samples, and smaller evaluation and testing sets of $10^3$ samples. We use gradient boosted decision tree models as implemented by the Python package LightGBM [7]. Measured feature contributions are reported in Table 1, for a regularly trained model as explained by TreeSHAP [10], as well as for our proposed approaches. As we can observe, both of our proposed approaches are able to place the feature importances on the features that appear earlier in the causal ordering.

## 4.2. Real-world data experiment

We demonstrate our method on a proprietary real-world dataset of mobile telecommunications, consisting of performance management (PM) data from radio access network cells with 15 minutes granularity. Model output is average downlink cell throughput for automatically determining the root cause of throughput degradations using explainers of the model. Input features of the model are described in Table 2, while Figure 1 illustrates the causal relationships between these groups. It also includes a 5th group: *Cell configuration.* Features of this group are not present in the dataset, but directly or indirectly affect all other features. This has the side-effect that while the groups *TA distribution* and *Cell load* can be assumed independent given

**Table 2**
Feature groups used in our real-world dataset.

| Name | Description |
|---|---|
| TA distribution | Timing Advance distribution of mobile terminals. It is derived from radio propagation delay measurements between the mobile terminal and base stations and can be used to estimate the distance from the base station. We normalize this distance with cell range, hence these features capture cell edge versus cell center distribution of mobile terminals. |
| Cell load | Various PM metrics describing user plane and signaling load of a given cell. |
| Interference | Measured uplink interference distribution in a given cell (downlink interference is unfortunately not available in our dataset). |
| Channel quality | Various downlink channel quality metrics including CQI and rank distributions. |

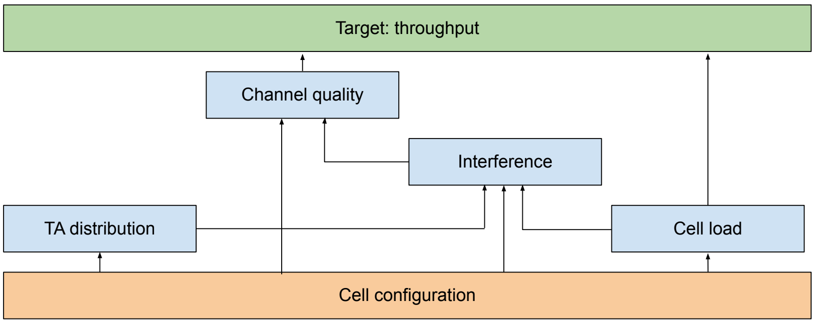

**Figure 1:** Causal structure of the different feature groups in our use case

**Table 3**
Average absolute feature attributions made by different methods on real-world data

| | TA distribution | Cell load | Interference | Channel quality |
|---|---|---|---|---|
| **Importances when *Cell load* ordered before *TA distribution*** | | | | |
| Causal-order training (Sec. 3.2) | 0.2093 | 3.6073 | 0.7939 | 2.9326 |
| Separate models (Sec. 3.3) | 0.4714 | 4.4750 | 0.9445 | 2.9343 |
| **Importances when *TA distribution* ordered before *Cell load*** | | | | |
| Causal-order training (Sec. 3.2) | 0.4317 | 3.4371 | 0.7808 | 2.9383 |
| Separate models (Sec. 3.3) | 1.3429 | 4.1927 | 0.9344 | 2.9463 |
| **Importances when contributions averaged over causal orderings** | | | | |
| Regular model | 0.1658 | 3.6426 | 0.8157 | 3.1973 |
| Causal-order training (Sec. 3.2) | 0.3205 | 3.5222 | 0.7874 | 2.9354 |
| Separate models (Sec. 3.3) | 0.9072 | 4.3339 | 0.9395 | 2.9403 |

*Cell configuration*, the latter not being present in the dataset makes them no longer independent. The resulting interactions should be distributed equally between these two groups.

There are only two possible topological orderings of these feature groups. In Table 3 we report feature importance according to both separately, and also final averaged feature importance. The methods clearly have a similar effect to what we observed in our synthetic experiments, with features that appear earlier in the causal order getting increased importance. The ordering of the first two groups has a negligible effect on later groups, which suggests the possibility of not even having to recompute these values when dealing with multiple orderings.

We can observe a substantial difference between the results of the approaches described in Sections 3.2 and 3.3. While both place increased importance on the group *TA distribution*, the latter is still much higher. We attribute this to the possibility observed in Section 3.2 that while the model is encouraged to use features that appear earlier in the causal ordering, it is not forced to do so. There is no such problem present in the case of the other approach.

## 5. Conclusions and Acknowledgments

In this paper, we described two approaches to calculating causality-respecting additive local feature attributions for explaining the cause of radio network performance issues. We demonstrated their effectiveness on both synthetic and real-world datasets and concluded that these methods proved to be effective in placing more feature importance on features that appear earlier in a causal ordering.

This work has been partially funded by the Hungarian Ministry of Innovation and Technology NRDI Office within the framework of the Artificial Intelligence National Laboratory Program.

## References

[1] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[2] H. Chen, J. D. Janizek, S. M. Lundberg, S. Lee, True to the model or true to the data?, CoRR abs/2006.16234 (2020). arXiv:2006.16234.

[3] C. Frye, C. Rowat, I. Feige, Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability, NIPS 33 (2020) 1229–1239.

[4] C. Winship, S. L. Morgan, The estimation of causal effects from observational data, Annual review of sociology 25 (1999) 659–706.

[5] T. Heskes, E. Sijben, I. G. Bucur, T. Claassen, Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models, in: NeurIPS, volume 33, 2020, pp. 4778–4789.

[6] L. S. Shapley, A Value for N-Person Games, RAND Corporation, Santa Monica, CA, 1952. doi:10.7249/P0295.

[7] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, in: NeurIPS, volume 30, 2017.

[8] S. Lipovetsky, M. Conklin, Analysis of regression in game theory approach, Applied Stochastic Models in Business and Industry 17 (2001) 319–330.

[9] E. Štrumbelj, I. Kononenko, M. R. Šikonja, Explaining instance classifications with interactions of subsets of feature values, Data & Knowledge Engineering 68 (2009) 886–904.

[10] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, Nature Machine Intelligence 2 (2020) 2522–5839.