Supplementary Material
Marginal Contribution Feature Importance - an Axiomatic Approach for Explaining Data

Abstract

This document contains the supplementary materials for the paper "Marginal Contribution Feature Importance - an Axiomatic Approach for Explaining Data".

A Proofs

In this section we group together all the proofs for the theorems we presented in the main paper.

A.I Existence and Uniqueness of the Feature Importance Score

Here we prove Theorem 1 states that there is a function for which all the axioms defined in Definition 2 hold and that this function is unique. This proof is a constructive proof in the sense that we are able to show that the only feature importance function for which the axioms hold is the function I_ν that assigns to the feature $f \in F$ the score

$$I_v(f) = \max_{S\subseteq F}(v(S \cup \{f\}) - v(S)) = \max_{S\subseteq F} \Delta\left(f, S, \nu\right) \quad .$$

Meaning, the importance of a feature is the maximum contribution to the evaluation function ν over any subset of features.

Lemma 1. *Let I_ν be a feature importance function for which axioms (1), (2) hold (marginal contribution, and elimination axiom). Then:*

$$I_\nu\left(f\right) \geq \max_{S\subseteq F} \Delta\left(f, S, \nu\right) \quad .$$

Proof. We prove the statement using induction on the size of the feature set F. Let $n = |F|$. If $n = 1$ (i.e $F = \{f\}$) then from the marginal contribution axiom we have that

$$I_\nu\left(f\right) \geq \nu\left(\{f\}\right) - \nu\left(\emptyset\right) = \max_{S\subseteq F} \Delta\left(f, S, \nu\right) \quad .$$

Assume that the statement holds for any set of features of size $< n$ for $n > 1$. Let $|F| = n$ and let $f \in F$. Let $S^* = \arg\max_{S\subseteq F} \Delta\left(f, S, \nu\right)$. If there exists $f' \in F \setminus \{S^* \cup \{f\}\}$ then from the elimination axiom, if $\{f'\}$ is eliminated we will obtain F' and ν' such that $|F'| = n - 1$ and $I_{\nu'}\left(f\right) \leq I_\nu\left(f\right)$. However, since $S^* \subseteq F'$ we have that from the assumption of the induction:

$$I_\nu\left(f\right) \geq I_{\nu'}\left(f\right) \geq \Delta\left(f, S^*, \nu'\right) = \Delta\left(f, S^*, \nu\right) = \max_{S\subseteq F} \Delta\left(f, S, \nu\right) \quad .$$

Otherwise, assume that $S^* = \arg\max_{S\subseteq F\setminus} \Delta\left(f, S, \nu\right)$ is such that $S^* \cup \{f\} = F$. Therefore, $\max_{S\subseteq F} \Delta\left(f, S, \nu\right) = \Delta\left(f, F \setminus \{f\}, \nu\right)$. From the marginal contribution axiom we have that $I_\nu\left(f\right) \geq \Delta\left(f, F \setminus \{f\}, \nu\right) = \max_{S\subseteq F} \Delta\left(f, S, \nu\right)$.

□

Lemma 1 shows that any importance function that has the marginal contribution property and the elimination property must assign an importance score of at least $\max_{S\subseteq F} \Delta(f,S,\nu)$ to every feature. Therefore, by adding the minimalism axiom we obtain the uniqueness and existence of the feature importance function as shown in Theorem 1.

Here we prove Theorem 1

Proof. Adding the minimalism axiom to Lemma 1 shows that if the marginal contribution and the elimination axioms hold for $I_\nu(f) = \max_{S\subseteq F} \Delta(f,S,\nu)$ then it is the unique feature importance function. Proving that the marginal contribution axiom hold is straight-forward: for a feature $f \in F$

$$I_\nu(f) = \max_{S\subseteq F} \Delta(f,S,\nu) \geq \Delta(f, F\setminus\{f\},\nu) = \nu(F) - \nu(F\setminus\{f\}) \quad .$$

To see that the elimination axiom holds too, let $T \subset F$ and let $f \in F \setminus T$. If T is eliminated from F to create F' and ν' then

$$I_\nu(f) = \max_{S\subseteq F} \Delta(f,S,\nu) \geq \max_{S\subseteq F'} \Delta(f,S,\nu) = \max_{S\subseteq F'} \Delta(f,S,\nu') = I_{\nu'}(f) \quad .$$

□

A.II Properties of the MCI Function

Here, we prove the MCI function properties presented in Theorem 2.

Proof. **Dummy**: Let f be a dummy variable such that $\forall S \subseteq F,\ \Delta(f,S,\nu) = 0$ then $I_\nu(f) = \max_{S\subseteq F} \Delta(f,S,\nu) = 0$.

Symmetry: Let f_i and f_j be such that for every $S \subseteq F$ we have that $\nu(S\cup\{f_i\}) = \nu(S\cup\{f_j\})$. Consider any set $S \subseteq F$. We consider three cases, (1) $f_i, f_j \in S$, (2) $f_i, f_j \notin S$, and (3) exactly one of f_i, f_j is in S. In case (1) we have that $\Delta(f_i,S,\nu) = \Delta(f_j,S,\nu) = 0$. In case (2) we have:

$$\Delta(f_i,S,\nu) = \nu(S\cup\{f_i\}) - \nu(S) = \nu(S\cup\{f_j\}) - \nu(S) = \Delta(f_j,S,\nu) \quad .$$

In case (3) assume, w.l.o.g. that $f_i \in S$ and $f_j \notin S$. Let S' denote the set S where f_i is replaced by f_j and therefore, due to the symmetry between f_i and f_j it holds that $\nu(S) = \nu(S')$. Note also that $S\cup\{f_j\} = S'\cup\{f_i\}$ and therefore $\Delta(f_j,S,\nu) = \Delta(f_i,S',\nu)$. From analyzing these 3 cases it follows that for every $S \subseteq F$ there exists $S' \subseteq F$ such that $\Delta(f_i,S,\nu) = \Delta(f_j,S',\nu)$. Therefore, $I_\nu(f_i) \leq I_\nu(f_j)$. However, by replacing the roles of f_i and f_j it also holds that $I_\nu(f_i) \geq I_\nu(f_j)$ and therefore $I_\nu(f_i) = I_\nu(f_j)$.

Super-efficiency: Let $S \subseteq F$. w.l.o.g. let $S = \{f_1, f_2, \ldots, f_k\}$. Define $\forall_{i\leq k} S_i = \{f_j\}_{j\leq i}$. Therefore, $S_0 = \emptyset$ and $S_k = S$. Since by definition $\nu(\emptyset) = 0$,

$$\begin{aligned} \nu(S) &= \nu(S_k) - \nu(S_0) = \sum_{i=0}^{k-1} \left(\nu(S_{i+1}) - \nu(S_i)\right) \\ &= \sum_{i=0}^{k-1} \Delta(f_{i+1}, S_i, \nu) \leq \sum_{i=0}^{k-1} I_\nu(f_{i+1}) = \sum_{f\in S} I_\nu(f) \quad . \end{aligned}$$

Sub-additivity: if ν and ω are evaluation functions defined on F then for all $f \in F$

$$\begin{aligned} I_{\nu+\omega}(f) &= \max_{S\subseteq F} \Delta(f,S,v+\omega) \\ &= \max_{S\subseteq F} \left(\Delta(f,S,\nu) + \Delta(f,S,\omega)\right) \\ &\leq \max_{S\subseteq F} \Delta(f,S,\nu) + \max_{S\subseteq F} \Delta(f,S,\omega) \\ &= I_\nu(f) + I_\omega(f) \quad . \end{aligned}$$

Upper bound the self contribution: For $f \in F$ We have that $I_\nu(f) = \max\limits_{S \subseteq F} \Delta(f, S, \nu) \geq \Delta(f, \emptyset, \nu) = \nu(\{f\}) - \nu(\emptyset) = \nu(\{f\})$.

Duplication invariant: Assume that $f_i \in F$ is a duplication of $f_j \in F$ in the sense that for every $S \subseteq F \setminus \{f_i, f_j\}$ we have that $\nu(S \cup \{f_i, f_j\}) = \nu(S \cup \{f_i\}) = \nu(S \cup \{f_j\})$. Let F' and ν' be the results of eliminating $\{f_i\}$ and let $f \in F'$. From the elimination axiom we have that $I_\nu(f) \geq I_{\nu'}(f)$. Assume that $S^* \subseteq F$ is such that $I_\nu(f) = \Delta(f, S^*, \nu)$. If $f_i \notin S^*$ then $S^* \subseteq F'$ and $I_{\nu'}(f) \geq \Delta(f, S^*, \nu') = \Delta(f, S^*, \nu) = I_\nu(f)$. Otherwise $f_i \in S^*$ and it holds for $S' = S^* \cup \{f_j\} \setminus \{f_i\}$ that $\Delta(f, S^*, \nu) = \Delta(f, S', \nu) = \Delta(f, S', \nu') \leq I_{\nu'}(f)$. Therefore, in all possible cases we have that $I_\nu(f) \leq I_{\nu'}(f)$. When combined with the elimination axiom we conclude that $I_\nu(f) = I_{\nu'}(f)$. □

A.III MCI Function Uniform Convergence of Empirical Means

In the following we prove Theorem 3 presented in the main paper. Recall that this theorem uses the uniform convergence of empirical means [9] to show that with high probability, ν can be estimated to within an additive factor using a finite sample and this estimate can be used to approximate MCI to within a similar additive factor.

Proof. Let μ be a probability measure over $X \times Y$. Let F be a set of random variables (features) over X. For any $S \subseteq F$ let $\mathcal{H}_S$ be a hypothesis class defined using only the features in S and let $d = \log_2 \max\limits_{S \subseteq F} (|\mathcal{H}_S|)$. Let $\ell : Y \times Y \mapsto \{0, 1\}$ be a 0-1 loss function, $\epsilon, \delta > 0$ and $m \geq \left(\frac{2}{\epsilon^2}\right)\left(d + |F| + \log_2\left(\frac{2}{\delta}\right)\right)$.

For any hypothesis class h we denote the test loss expectation by $e_P(h) = E_{(x,y)\sim\mu}(\ell(h(x), y))$, and the empirical loss expectation for a sample $D \sim \mu^m$ by $e_D(h) = E_{(x,y)\sim\mu^m}(\ell(h(x), y))$.

Given the above notations, we define the true evaluation function $\nu : \mathcal{P}(F) \mapsto \mathbb{R}^+$ for any $S \subseteq F$ to be $\nu(S) = \min_{h \in \mathcal{H}_\emptyset} e_P(h) - \min_{h \in \mathcal{H}_S} e_P(h)$, and the empirical evaluation function $\nu_D : \mathcal{P}(F) \mapsto \mathbb{R}^+$ for any $S \subseteq F$ to be $\nu_D(S) = \min_{h \in \mathcal{H}_\emptyset} e_D(h) - \min_{h \in \mathcal{H}_S} e_D(h)$.

First, we show that for all $S \subseteq F$:

$$|\nu_D(S) - \nu(S)| \leq \max_{h \in \mathcal{H}_S} |(e_D(h) - e_P(h))|$$

Let $S \subseteq F$. Denote $h_D^* = \arg\min_{h \in \mathcal{H}_S} e_D(h)$ and $h^* = \arg\min_{h \in \mathcal{H}_S} e_P(h)$. On one hand we have that:

$$\begin{aligned} \nu_D(S) - \nu(S) &= \min_{h \in \mathcal{H}_S} e_D(h) - \min_{h \in \mathcal{H}_S} e_P(h) \\ &= e_D(h_D^*) - e_P(h^*) \\ &\leq e_D(h^*) - e_P(h^*) \\ &\leq \max_{h \in \mathcal{H}_S} |e_D(h) - e_P(h)| \end{aligned}$$

And on the other hand:

$$\begin{aligned} \nu(S) - \nu_D(S) &= \min_{h \in \mathcal{H}_S} e_P(h) - \min_{h \in \mathcal{H}_S} e_D(h) \\ &= e_P(h^*) - e_D(h_D^*) \\ &\leq e_P(h_D^*) - e_D(h_D^*) \\ &\leq \max_{h \in \mathcal{H}_S} |e_D(h) - e_P(h)| \end{aligned}$$

Next we would like to show that for any $f \in F$ it holds that $|I_{\nu_D}(f) - I_\nu(f)| \leq 2\max_{S \subseteq F} |\nu_D(S) - \nu(S)|$.

Let $f \in F$ and let $S^* = \arg\max_{S \subseteq F} (\Delta(f, S, \nu))$, $S_D^* = \arg\max_{S \subseteq F} (\Delta(f, S, \nu_D))$.

Notice that:

$$\begin{aligned}
I_{\nu_D}(f) - I_\nu(f) &= \Delta\left(f, S^*_D, \nu_D\right) - \Delta\left(f, S^*, \nu\right) \\
&\leq \Delta\left(f, S^*_D, \nu_D\right) - \Delta\left(f, S^*_D, \nu\right) \\
&\leq \max_{S \subseteq F} \left|\Delta\left(f, S, \nu_D\right) - \Delta\left(f, S, \nu\right)\right|
\end{aligned}$$

and also:

$$\begin{aligned}
I_\nu(f) - I_{\nu_D}(f) &= \Delta\left(f, S^*, \nu\right) - \Delta\left(f, S^*_D, \nu_D\right) \\
&\leq \Delta\left(f, S^*, \nu\right) - \Delta\left(f, S^*, \nu_D\right) \\
&\leq \max_{S \subseteq F} \left|\Delta\left(f, S, \nu_D\right) - \Delta\left(f, S, \nu\right)\right|
\end{aligned}$$

and therefore we get that:

$$\begin{aligned}
\left|I_{\nu_D}(f) - I_\nu(f)\right| &\leq \max_{S \subseteq F} \left|\Delta\left(f, S, \nu_D\right) - \Delta\left(f, S, \nu\right)\right| \\
&= \max_{S \subseteq F} \left|\left(\nu_D(S \cup \{f\}) - \nu_D(S)\right) - \left(\nu(S \cup \{f\}) - \nu(S)\right)\right| \\
&\leq \max_{S \subseteq F} \left|\nu_D(S \cup \{f\}) - \nu(S \cup \{f\})\right| + \max_{S \subseteq F} \left|\nu_D(S) - \nu(S)\right| \\
&\leq 2 \max_{S \subseteq F} \left|\nu_D(S) - \nu(S)\right|
\end{aligned}$$

Hence, using union bound and Hoeffding inequality, for any $f \subseteq F$ it holds that:

$$\begin{aligned}
P\left[\left|I_{\nu_D}(f) - I_\nu(f)\right| > \epsilon\right] &\leq P\left[2 \max_{S \subseteq F, h \in \mathcal{H}_S} \left|\nu_D(S) - \nu(S)\right| > \epsilon\right] \\
&\leq P\left[2 \max_{S \subseteq F, h \in \mathcal{H}_S} \left|e_D(h) - e_P(h)\right| > \epsilon\right] \\
&= P\left[\bigcup_{S \subseteq F, h \in \mathcal{H}_S} \left\{2\left|e_D(h) - e_P(h)\right| > \epsilon\right\}\right] \\
&\leq \sum_{S \subseteq F, h \in \mathcal{H}_S} P\left[2\left|e_D(h) - e_P(h)\right| > \epsilon\right] \\
&\leq 2^{(|F|+d+1)} e^{\frac{-m\epsilon^2}{2}} \leq \delta \quad .
\end{aligned}$$

Where the last inequality follows by using the bound on m in the statement of this theorem.

□

B Additional Properties of the MCI Function

In the following theorem we present and prove additional relevant properties of the MCI function, that were not discussed in the main paper.

Theorem 1. *Let F be a set of features, let ν be an evaluation function and let I_ν be the MCI function. The following holds:*

- ***Scaling**: $\forall f \in F, \ \forall \lambda > 0, \ \ I_{\lambda\nu}\left(f\right) = \lambda I_\nu\left(f\right)$.*
- ***Monotonicity**: If $\forall S \subseteq F \setminus \{f_i, f_j\}, \ v\left(S \cup \{f_i\}\right) \leq v\left(S \cup \{f_j\}\right)$ then $I_v\left(f_i\right) \leq I_v\left(f_j\right)$.*

Where $\lambda\nu$ denotes for multiplying each value of ν by λ.

In the following we prove theorem 1

Proof. :

Scaling: This property follows since for every f and every S it holds that $\lambda\Delta(f, S, \nu) = \Delta(f, S, \lambda\nu)$.

Monotonicity: Let $f_i, f_j \in F$ for which $\forall S \subseteq F \setminus \{f_i, f_j\},\ v(S \cup \{f_i\}) \leq v(S \cup \{f_j\})$. Let S^* be such that $I_\nu(f_i) = \Delta(f_i, S^*, \nu)$. Then, if $f_j \notin S^*$ it holds that

$$I_\nu(f_i) = \Delta(f_i, S^*, \nu) \leq \Delta(f_j, S^*, \nu) \leq I_\nu(f_j) \quad .$$

Otherwise, if $f_j \in S^*$ then

$$I_\nu(f_i) = \Delta(f_i, S^*, \nu) \leq \Delta(f_j, S^* \cup \{f_i\} \setminus \{f_j\}, \nu) \leq I_\nu(f_j) \quad .$$

□

C Computation Optimizations

We now turn our attention to the computational challenge of computing or approximating the MCI function. Straight-forward computation is exponential in the size of the feature set. Therefore we study cases in which the computation can be made efficient and also study approximation techniques.

C.I Submodularity

In the following we show that if the evaluation function ν is submodular [7] then the MCI feature importance score is equal to the self contribution of each feature.

Lemma 2. *If ν is sub-modular then $I_\nu(f) = \nu(\{f\})$.*

Proof. Recall that in that case there is a diminishing return and therefore for every $S \subseteq F \setminus \{f\}$: $\nu(S \cup \{f\}) \leq \nu(S) + \nu(\{f\})$. Therefore,

$$\Delta(f, S, \nu) = \nu(S \cup \{f\}) - \nu(S) \leq \nu(\{f\}) = \Delta(f, \emptyset, \nu)$$

□

The submodularity assumption might be too stringent in some cases. For example, if the target variable is an XOR of some features then the submodularity assumption does not hold. However, if we assume that submodularity holds for large sets then we obtain a polynomial algorithm for computing the feature importance. This may make sense in the genomics setting where genes may have synergies but we may assume that only small interactions of 2, 3, or 4 genes are significant. We begin by defining k-size submodularity:

Definition 2. *A function $\nu : \mathcal{P}(F) \mapsto \mathbb{R}$ is k-size submodular if for every $S, T \subseteq F$ such that $|S|, |T| \geq k$*

$$\nu(S) + \nu(T) \geq \nu(S \cup T) + \nu(S \cap T)$$

A function $\nu : \mathcal{P}(F) \mapsto \mathbb{R}$ is soft k-size submodular if it holds that for every $T \subseteq F$, $|T| > k$, $f \in F$ there exists $S \subseteq T$, $|S| \geq k$ for which:

$$\nu(S \cup \{f\}) + \nu(T) \geq \nu((S \cup \{f\}) \cup T) + \nu((S \cup \{f\}) \cap T)$$

Lemma 3. *If ν is k-size-submodular or soft k-size-submodular then*

$$I_\nu(f) = \max_{S \subseteq F\,:\,|S| \leq k} \Delta(f, S, \nu)$$

Proof. First, we show that if ν is k-size-submodular it is also soft k-size-submodular. Let ν be a k-size-submodular evaluation function. Let $T \subseteq F$, $|T| > k$ and let $S \subseteq T$, $|S| \geq k$, $f \in F$. From the k-size-submodular property we get that:

$$\nu(S \cup \{f\}) + \nu(T) \geq \nu((S \cup \{f\}) \cup T) + \nu((S \cup \{f\}) \cap T)$$

And therefore ν is also soft k-size-submodular. Hence, it is enough to prove the theorem for soft k-size-submodular functions.

Let ν be a soft k-size-submodular evaluation function. Let $T \subseteq F$ be such that $T = \arg\min\{|T| : I_\nu(f) = \Delta(f, T, \nu)\}$. Assume, in contradiction, that $|T| > k$. Note that $f \notin T$ since if $f \in T$ then $I_\nu(f) = \Delta(f, T, \nu) = 0$, and in this case we have that $I_\nu(f) = \Delta(f, \emptyset, \nu)$ in contradiction. Due to the soft k-size submodular and monotone properties of ν it follows that there exists $S \subseteq T$, $|S| \leq k$ such that:

$$\nu(S \cup \{f\}) + \nu(T) \geq \nu(T \cup \{f\}) + \nu(S)$$

Therefore, $\Delta(f, S, \nu) = \nu(S \cup \{f\}) - \nu(S) \geq \nu(T \cup \{f\}) - \nu(T) = \Delta(f, T, \nu)$. This is a contradiction since $|S| < |T|$. □

Lemma 3 shows that if ν is soft k-size-submodular then the entire function I_ν can be computed in time $O\left(|F|^{k+1}\right)$.

C.II Branch and Bound Optimization

Here we show how we can discard computation for some of the subsets using a branch and bound like technique.

Lemma 4. *For every $S_0 \subseteq S_1 \subseteq S_2 \subseteq F$ and $f \in F$:*

$$\Delta\left(f, S_1, \nu\right) \leq \nu\left(S_2 \cup \{f\}\right) - \nu\left(S_0\right)$$

Proof. This lemma follows from the monotone property of ν.

$$\Delta\left(f, S_1, \nu\right) = \nu\left(S_1 \cup \{f\}\right) - \nu\left(S_1\right) \leq \nu\left(S_2 \cup \{f\}\right) - \nu\left(S_0\right) \quad .$$

□

The ability to upper bound I_ν provided by this Lemma allows cutting back the computation significantly. For example, if we computed $\Delta(f, S, \nu)$ for every set S of size k and we have that $\max\limits_{S:|S|\leq k} \Delta(f, S, \nu) \geq \max\limits_{S:|S|=k} \nu(F) - \nu(S)$ then $I_\nu(f) = \max\limits_{S:|S|\leq k} \Delta(f, S, \nu)$. The following lemma proves this property in a more general setting.

Lemma 5. *Let $\mathbb{S} = \{S \subseteq F$ s.t. $|S| = K\}$ and $\mathbb{T} = \{T \subseteq F$ s.t. $|T| = k\}$ for $0 \leq k \leq K \leq |F|$. Let $\bar{\mathbb{S}} = \{S \subseteq F : \exists S' \in \mathbb{S}$ s.t. $S' \subseteq S\}$ and $\underline{\mathbb{T}} = \{T \subseteq F : \exists T' \in \mathbb{T}$ s.t. $T \subseteq T'\}$. Let $s_f = \max\limits_{S\in\bar{\mathbb{S}}} \Delta(f, S, \nu)$ and $t_f = \max\limits_{T\in\underline{\mathbb{T}}} \Delta(f, T, \nu)$ and $st_f = \max\limits_{S\in\mathbb{S}, T\in\mathbb{T}} \nu(S \cup f) - \nu(T)$ then*

$$\max(s_f, t_f) \leq I_\nu(f) \leq \max(s_f, t_f, st_f) \quad .$$

Proof. The lower bound on $I_\nu(f)$ follows from the simple fact that for every $\mathbb{S} \subseteq 2^F$

$$\max_{S\in\mathbb{S}} \Delta(f, S, \nu) \leq I_\nu(f) \quad .$$

Let S^* be such that $I_\nu(f) = \Delta(f, S^*, \nu)$. If $S^* \in \bar{\mathbb{S}} \cup \underline{\mathbb{T}}$ then $I_\nu(f) = \max(s_t, t_f)$. Otherwise, there exists $S \in \mathbb{S}$ and $T \in \mathbb{T}$ such that $T \subset S^* \subset S$ and from Lemma 5 it holds that

$$I_\nu(f) = \Delta(f, S^*, \nu) \leq \nu(S \cup f) - \nu(T)$$

which completes the proof. □

C.III Heuristics

Recall that for any $\mathbb{S} \subseteq 2^F$ it holds that $\max\limits_{S\in\mathbb{S}} \Delta(f, S, \nu) \leq I_\nu(f)$. Therefore, any method can be used to select $\mathbb{S}$ and obtain a lower bound on the feature importance. In the experiments in this paper we used random permutations to generate the set $\mathbb{S}$ following the proposal of [3]. This method is described in Section 5. Our experiments show that this method is effective. however, in some cases it may be too demanding since for every subset of features a model has to be trained. The computational cost can be further reduced by using a method such as SAGE [3] to estimate $\Delta(f, S, \nu)$ from a model that was trained on the entire dataset and therefore trained only once. Only for sets S such that SAGE estimates that $\Delta(f, S, \nu)$ is large, the real value can be computed via training models. Therefore, the estimator SAGE (or any other proposed method) is used to eliminate testing sets S for which the marginal contribution of f is predicted to be small.

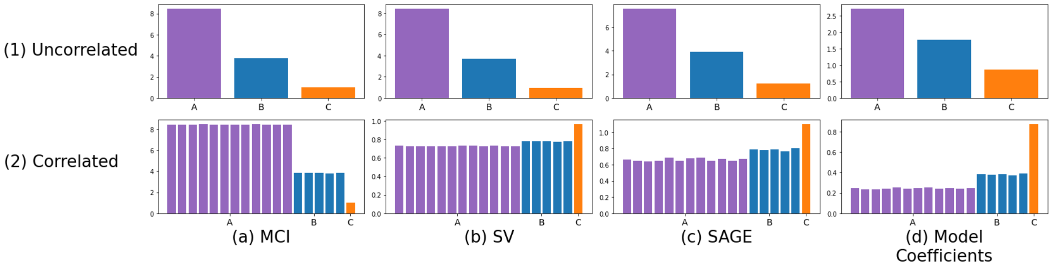
D Experiments Supplementary Material

In the following we provide additional information about the experiments presented in Section 5 in the main paper.

Table 1 summarizes the models and the datasets used in each experiment, along with computation times. We note that unless stated else in Section 5, we used the default hyper-parameters provided by the framework of each model we trained. For the collinearity and BRCA experiments we used the

Table 1: **Experiments Summary.** This table summarizes the experiments presented in Section 5. Note that in cases when there is more than one setting, the table provides the results of the setting that contains the greatest number of features.

Experiment	#Features	#Examples	Model	#Evaluations	Loss	Run Time (Hours)
Collinearity	17	10K	Elastic-Net	2^{17}	MSE	2:12
Non-linearity	8	100K	MLP	2^{8}	Cross Entropy	1:17
BRCA	50	512	Logistic Regression	$2^{15} \times 50$	Cross Entropy	34:10
BSI	20	7,436	LightGBM	$2^{14} \times 20$	Cross Entropy	8:20



(a) MCI
(b) SV
(c) SAGE
(d) Model Coefficients

Figure 1: **Detailed results of the collinearity experiment.** The importance assigned to the features correlated with A are shown in purple, with B in blue and with C in orange.

Scikit-learn package version 0.23.1 [8]. For the non-linearity XOR experiment we used Keras version 2.4.3 [2]. For the BSI experiment we used LightGBM version 2.3.0 [6]. For all the experiments but the BSI experiment we used a machine with 2 Intel Xeon Silver 4114 @ 2.20GHz CPUs. The BSI experiment was performed using a VM on a shared computer.

D.I Additional Results for the Collinearity Synthetic Experiment

Figure 1 provides the scores assigned by the different methods in the collinearity experiment described in Section 5.1 in the main paper. As seen, MCI is the only method that is not effected by the addition of correlated features. SV and SAGE completely reverse the expected order and find the feature correlated with C as most important, followed by the features correlated with B, and last by the features correlated with A, despite the fact that $Y = 3A + 2B + C$.

D.II Model Agnostic for Non-Linear Interactions Synthetic Experiment

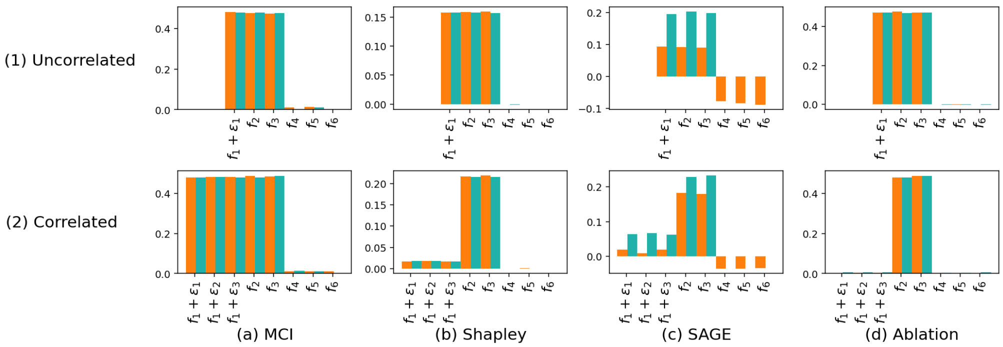
In the following we describe an additional experiment designed to test the robustness of MCI and the other methods to changes in the underlying model used. We repeat the experiment presented in the main paper in Section 5.2 and use a Random Forest classifier instead of MLPs. Specifically, we define $\nu(S)$ for each subset of features S to be the test accuracy of a Random Forest Classifier trained with 10 decision trees and the other default parameters provided by the Scikit-learn [8] package version 0.23.1.

As shown in Figure 2, MCI and SV provide similar scores when using MLPs and Random Forests, while SAGE provides slightly different scores. These results suggest that SAGE is less agnostic to the type of model used to evaluate ν. This can be explained by the fact that SAGE is a method for explaining models and therefore is more sensitive to the type of model in use.

D.III Additional Details about the BRCA Experiments

Here we provide additional details about the BRCA experiments described in Section 5.3 in the main paper. We note that we followed the same processing steps as in [3] and provide the processed dataset we used for the quality test[1].

[1] https://github.com/TAU-MLwell/Marginal-Contribution-Feature-Importance



(a) MCI (b) Shapley (c) SAGE (d) Ablation

Figure 2: **Results of the model agnostic non-linear interactions synthetic experiment.** The scores assigned by each method using MLPs are shown in green, and the scores assigned by each method using Random Forest models are shown in orange. In the top row (1), we show for each method the feature importance assigned to the uncorrelated set of features (F^1). In the bottom row (2) we show the importance assigned for the correlated set of features (F^2). Note that MCI and SV are agnostic to the type of model used, while SAGE provides slightly different scores for each model.

Table 2: **Quality scores for the BRCA robustness experiment.** This table shows the NDCG scores for the top-10 rankings provided by each of the methods, for each sample of genes used. Higher is better, perfect score is 1.00. As seen, MCI and SV perfectly identify the genes in R, for all the samples. This suggests that the added genes in each sample are not likely to be BRCA related.

Method / Sample no.	1	2	3	4	5
MCI	1.00	1.00	1.00	1.00	1.00
SV	1.00	1.00	1.00	1.00	1.00
SAGE	0.93	0.94	0.87	1.00	0.94
Ablation	0.70	0.83	0.79	0.87	0.87

Figure 3 provides the full scores assigned by the different methods in the BRCA quality experiment.

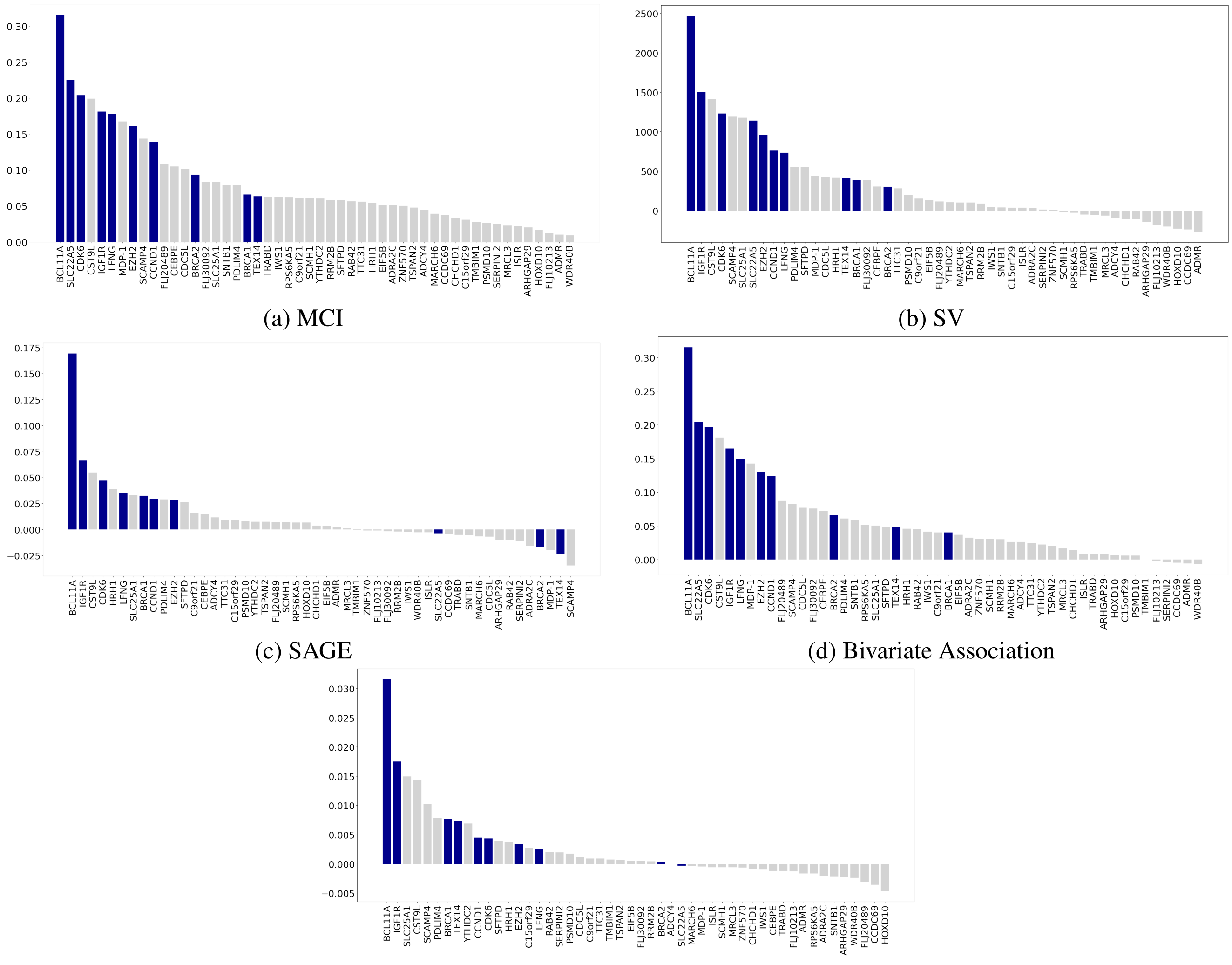
Figure 4 shows the convergence of the rankings induced by MCI and SV methods, for the BRCA quality experiment. As seen, both methods did not change their rankings for more than 5,000 consecutive sampled permutations, and therefore we consider them as converged.

Table 2 shows the quality of the rankings produced in the BRCA robustness experiment for each of the 5 samples used. The quality is measured by the top-10 NDCG score following Section 5.3. As seen, MCI and SV perfectly identify all the BRCA related genes (R) as most important, while SAGE and Ablation identify most of them in the top-10. This suggests that indeed the set of genes added to R in each of the samples are not likely to be BRCA related by themselves.

D.IV Bloodstream Infection Mortality Experiment

In the following we provide more details about the BSI experiment described in Section 5.4 in the main paper. This experiment uses data from medical records collected and was conducted with IRB approval. However, this approval does not allow the release of the data due to privacy concerns. We note that except for this experiment, all the data used in the paper is publicly available.

The BSI dataset contains 20 features extracted from Electronic Health Records (EHRs) of 7,436 patients, hospitalized with a positive blood culture (bacterial only). For each patient, a binary variable is provided that indicates in-hospital or 30-day mortality. The dataset is further split into a training set contains 6,135 patients admitted to the hospital during 2014-2018, and a future test set contains 1,301 patients admitted during 2019-2020. We further split the training set into 80%/20% for train and validation sets. Table 3 provides a description for each feature in the BSI dataset. Figure 5 shows the absolute Pearson correlation coefficient matrix between the features in the BSI dataset. Figure 6 shows in detail the importance scores assigned by each method in this experiment.



(a) MCI

(b) SV

(c) SAGE

(d) Bivariate Association

(e) Ablation

Figure 3: **Detailed results of the BRCA quality experiment.** The BRCA related genes are shown in blue, and the other genes in grey.

E Robustness Experiments on UCI Datasets

In this section we present additional experiments comparing the robustness of different feature importance methods. We use six datasets from the UCI repository [4] (see description of the datasets in Table 4). For each dataset, we first computed feature importance using different methods. Then, to test robustness, for each method we duplicated three times the feature that was ranked first and re-compute the feature importance. For these experiments we define $\nu(S)$ for each subset of features S as the average performance of a gradient boosting model (GBM) [5], over 3-fold cross validations. For regression tasks we used MSE loss and for classification tasks we used cross entropy loss. We used the GBM provided by the Scikit-learn package, along with its default parameters (except for setting the number of estimators to be 10).

The results of these experiments are shown in the following Figures: 7 (Heart Disease), 8 (Wine Quality), 9 (German Credit), 10 (Bike Rental), 11 (Online Shopping), 12 (Bank Marketing). For all datasets, the scores assigned by MCI were practically identical with or without the duplicated features. However, the score assigned by SV to the feature that was duplicated was reduced significantly once duplicated. On the Heart Disease dataset and on the Wine Quality dataset, this was sufficient to change the ranking of the top feature to the 2^{nd} or even the 3^{rd} position. On the German Credit

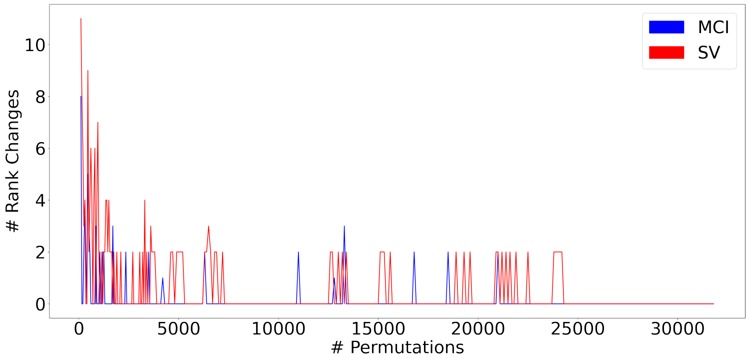


Figure 4: **Convergence of MCI and SV for the BRCA quality experiment.** This graph shows the number of rankings changes for each window of 50 sampled permutations. The number of changes is shown in blue for MCI and in red for SV. As seen, there are no rank changes for more then 5,000 consecutive sampled permutations, and therefore we consider both scores as converged in the point when 2^{15} permutations are already sampled.

Table 3: **BSI dataset features description.**

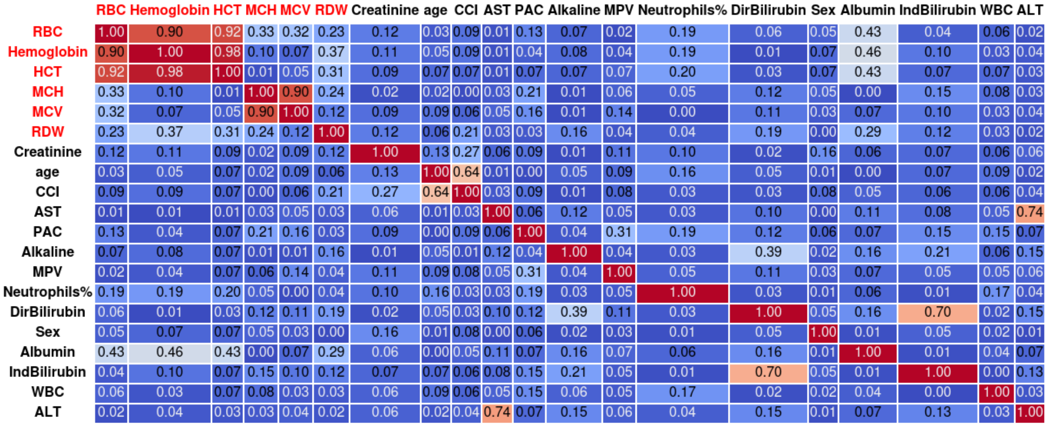
Name	Description
RBC	Red Blood Cells count in (10e6/μL)
Hemoglobin	Hemoglobin count
HCT	Hematocrit or the volume percentage of RBCs in blood (%)
MCH	Mean Corpuscular Hemoglobin, the average mass of hemoglobin per RBC
MCV	Mean Corpuscular Volume, the average volume of a red blood corpuscle (fL)
RDW	Red Cell Distribution Width, the range of variation of RBC volume (%)
Creatinine	Creatinine concentration in blood (mg/dL)
Age	Age of patient upon admission in (years)
Sex	Patient sex (male/female)
CCI	Charlson Comorbidity Index [1]
AST	Aspartate Aminotransferase concentration in blood (U/L)
PAC	Platelet Automated Count in blood (10e3/μL)
Alkaline	Alkaline Phosphatase concentration in blood (U/L)
MPV	Mean platelet volume, average size of platelets found in blood (fL)
Neutrophils	Neutrophils count in blood (10e3/μL)
Direct Bilirubin	Direct Bilirubin concentration in blood (mg/dL)
Indirect Bilirubin	Indirect Bilirubin concentration in blood (mg/dL)
WBC	White Blood Cells Count in blood
Albumin	Albumin concentration in blood (g/L)
ALT	Alanine Transaminase (U/L)

Default and the Bike Rental datasets the top feature, which had a big margin over the 2^{nd} most important feature, maintained its position but with small margin. On the Online Shopping dataset and the Bank Marketing datasets, the original margin of the top feature was so big that even though the score was reduced after duplicating the top feature, it remained in top position with a large margin.

References

[1] Mary E Charlson, Peter Pompei, Kathy L Ales, and C Ronald MacKenzie. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases*, 40(5):373–383, 1987.

[2] François Chollet et al. Keras. `https://keras.io`, 2015.



	RBC	Hemoglobin	HCT	MCH	MCV	RDW	Creatinine	age	CCI	AST	PAC	Alkaline	MPV	Neutrophils%	DirBilirubin	Sex	Albumin	IndBilirubin	WBC	ALT
RBC	1.00	0.90	0.92	0.33	0.32	0.23	0.12	0.03	0.09	0.01	0.13	0.07	0.02	0.19	0.06	0.05	0.43	0.04	0.06	0.02
Hemoglobin	0.90	1.00	0.98	0.10	0.07	0.37	0.11	0.05	0.09	0.01	0.04	0.08	0.04	0.19	0.01	0.07	0.46	0.10	0.03	0.04
HCT	0.92	0.98	1.00	0.01	0.05	0.31	0.09	0.07	0.07	0.01	0.07	0.07	0.07	0.20	0.03	0.07	0.43	0.07	0.07	0.03
MCH	0.33	0.10	0.01	1.00	0.90	0.24	0.02	0.02	0.00	0.03	0.21	0.01	0.06	0.05	0.12	0.05	0.00	0.15	0.08	0.03
MCV	0.32	0.07	0.05	0.90	1.00	0.12	0.09	0.09	0.06	0.05	0.16	0.01	0.14	0.00	0.11	0.03	0.07	0.10	0.03	0.04
RDW	0.23	0.37	0.31	0.24	0.12	1.00	0.12	0.06	0.21	0.03	0.03	0.16	0.04	0.04	0.19	0.00	0.29	0.12	0.03	0.02
Creatinine	0.12	0.11	0.09	0.02	0.09	0.12	1.00	0.13	0.27	0.06	0.09	0.01	0.11	0.10	0.02	0.16	0.06	0.07	0.06	0.06
age	0.03	0.05	0.07	0.02	0.09	0.06	0.13	1.00	0.64	0.01	0.00	0.05	0.09	0.16	0.05	0.01	0.00	0.07	0.09	0.02
CCI	0.09	0.09	0.07	0.00	0.06	0.21	0.27	0.64	1.00	0.03	0.09	0.01	0.08	0.03	0.03	0.08	0.05	0.06	0.06	0.04
AST	0.01	0.01	0.01	0.03	0.05	0.03	0.06	0.01	0.03	1.00	0.06	0.12	0.05	0.03	0.10	0.00	0.11	0.08	0.05	0.74
PAC	0.13	0.04	0.07	0.21	0.16	0.03	0.09	0.00	0.09	0.06	1.00	0.04	0.31	0.19	0.12	0.06	0.07	0.15	0.15	0.07
Alkaline	0.07	0.08	0.07	0.01	0.01	0.16	0.01	0.05	0.01	0.12	0.04	1.00	0.04	0.03	0.39	0.02	0.16	0.21	0.06	0.15
MPV	0.02	0.04	0.07	0.06	0.14	0.04	0.11	0.09	0.08	0.05	0.31	0.04	1.00	0.05	0.11	0.03	0.07	0.05	0.05	0.06
Neutrophils%	0.19	0.19	0.20	0.05	0.00	0.04	0.10	0.16	0.03	0.03	0.19	0.03	0.05	1.00	0.03	0.01	0.06	0.01	0.17	0.04
DirBilirubin	0.06	0.01	0.03	0.12	0.11	0.19	0.02	0.05	0.03	0.10	0.12	0.39	0.11	0.03	1.00	0.05	0.16	0.70	0.02	0.15
Sex	0.05	0.07	0.07	0.05	0.03	0.00	0.16	0.01	0.08	0.00	0.06	0.02	0.03	0.01	0.05	1.00	0.01	0.05	0.02	0.01
Albumin	0.43	0.46	0.43	0.00	0.07	0.29	0.06	0.00	0.05	0.11	0.07	0.16	0.07	0.06	0.16	0.01	1.00	0.01	0.04	0.07
IndBilirubin	0.04	0.10	0.07	0.15	0.10	0.12	0.07	0.07	0.06	0.08	0.15	0.21	0.05	0.01	0.70	0.05	0.01	1.00	0.00	0.13
WBC	0.06	0.03	0.07	0.08	0.03	0.03	0.06	0.09	0.06	0.05	0.15	0.06	0.05	0.17	0.02	0.02	0.04	0.00	1.00	0.03
ALT	0.02	0.04	0.03	0.03	0.04	0.02	0.06	0.02	0.04	0.74	0.07	0.15	0.06	0.04	0.15	0.01	0.07	0.13	0.03	1.00

Figure 5: **Absolute Pearson correlation matrix of the BSI features.** The names of the RBC related features are highlighted in red.

Table 4: **Description of the UCI datasets used in the robustness experiment**

Dataset	Type	# Features	# Examples
Heart Disease	Classification	13	303
Wine Quality	Regression	11	1,599
German Credit Default	Classification	20	1,000
Bike Rental	Regression	12	303
Online Shopping	Classification	17	12,330
Bank Marketing	Classification	16	45,211

[3] Ian Covert, Scott Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33, 2020.

[4] Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017.

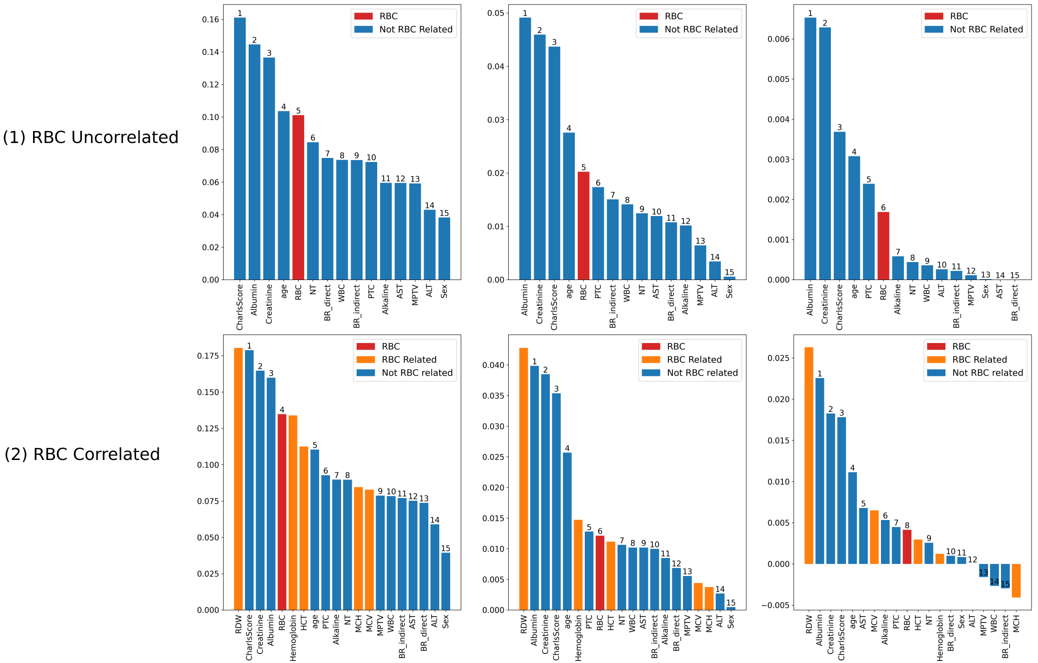
[5] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[6] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, pages 3146–3154, 2017.

[7] László Lovász. Submodular functions and convexity. In *Mathematical Programming The State of the Art*, pages 235–257. Springer, 1983.

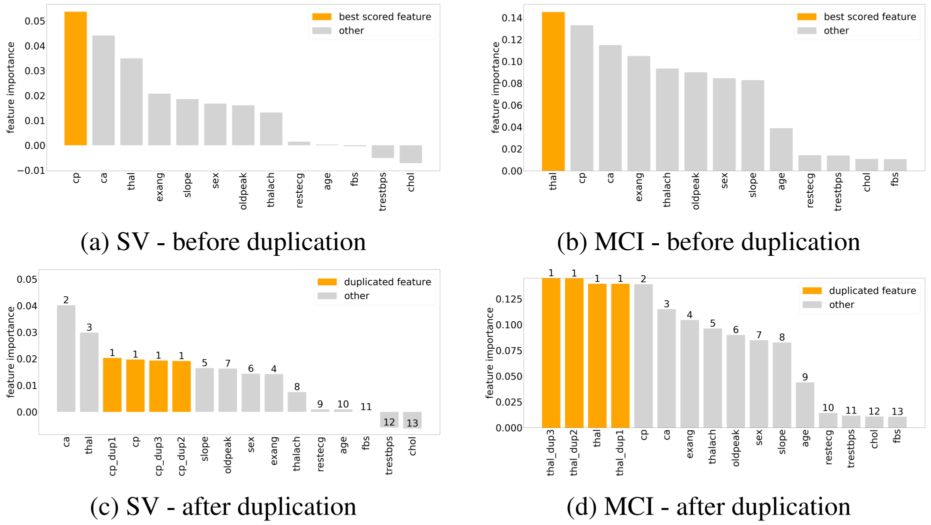
[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[9] Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, 2015.



(a) MCI (b) SV (c) SAGE

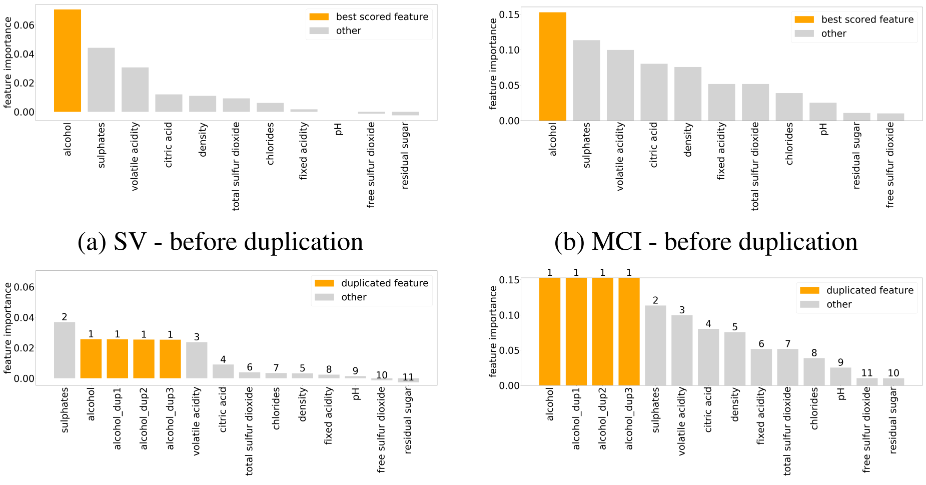
Figure 6: **Detailed results of the BSI experiment.** The top row (1) shows the importance scores assigned by each method for the subset of features when the RBC correlated features are removed. The bottom row (2) shows the importance scores assigned by each method for the full set of features, when RBC correlated are present. We show in RED the scores assigned to RBC, in orange the scores assigned to RBC correlated features, and in blue the scores assigned to the other features.



(a) SV - before duplication (b) MCI - before duplication

(c) SV - after duplication (d) MCI - after duplication

Figure 7: **Robustness experiment on the Heart Disease dataset.** The top row shows the feature importance according to SV (a) and MCI (b) for the original set of features in this dataset. Note that that each method suggesting a different ranking within the top three list. The bottom row shows the estimations of both methods, when the top ranked feature of each method is duplicated three times. As seen, the importance assignment of SV (c) is affected drastically form the introduction of duplicates, while MCI (d) succeeds to remain stable.



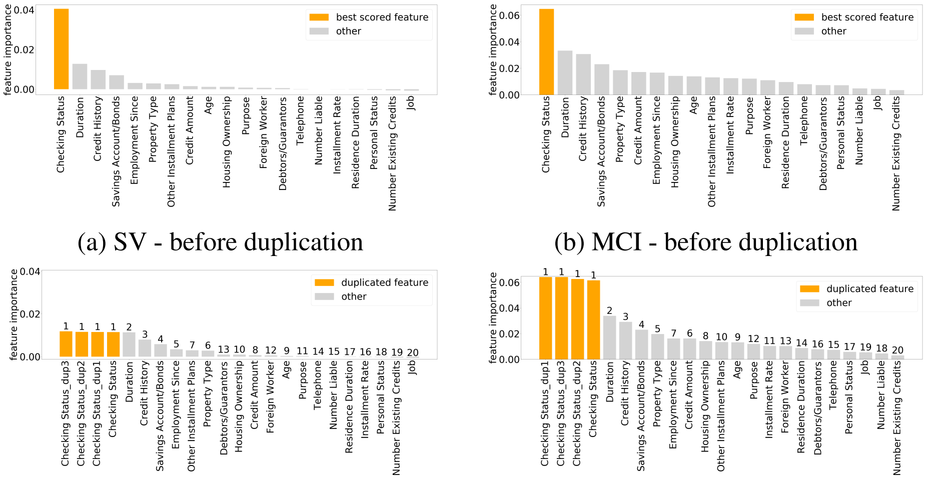
(a) SV - before duplication

(b) MCI - before duplication

(c) SV - after duplication

(d) MCI - after duplication

Figure 8: **Robustness experiment on the Wine Quality dataset.** The top row shows the feature importance according to SV (a) and MCI (b) for the original set of features in this dataset. The bottom row shows the estimations of both methods, when the top ranked feature of each method is duplicated three times. As seen, the importance assigned by SV (c) is affected drastically form the introduction of duplicates, while MCI (d) succeeds to remain stable.



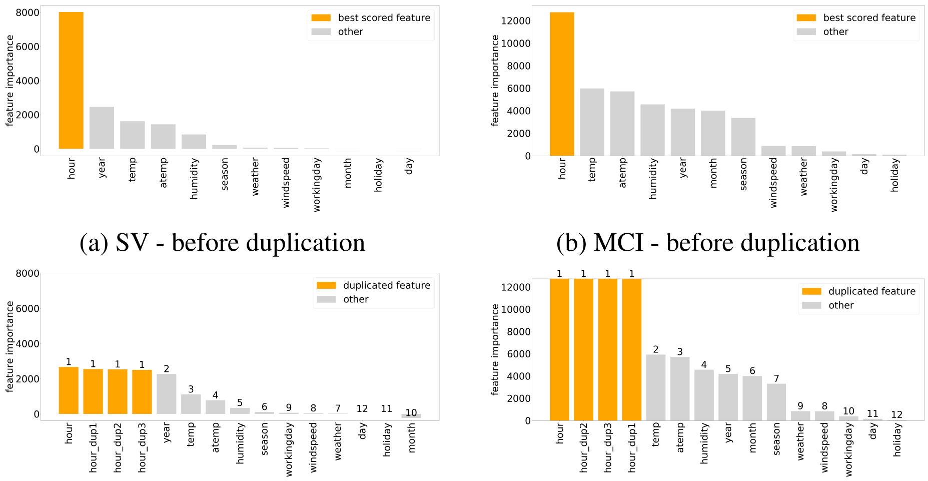
(a) SV - before duplication

(b) MCI - before duplication

(c) SV - after duplication

(d) MCI - after duplication

Figure 9: **Robustness experiment on the German Credit Default dataset.** The top row shows the feature importance according to SV (a) and MCI (b) for the original set of features in this dataset. The bottom row shows the estimations of both methods, when the top ranked feature of each method is duplicated three times. As seen, the relative differences in the importance scores given by SV (c) is affected form the introduction of duplicates, while MCI (d) succeeds to remain stable.



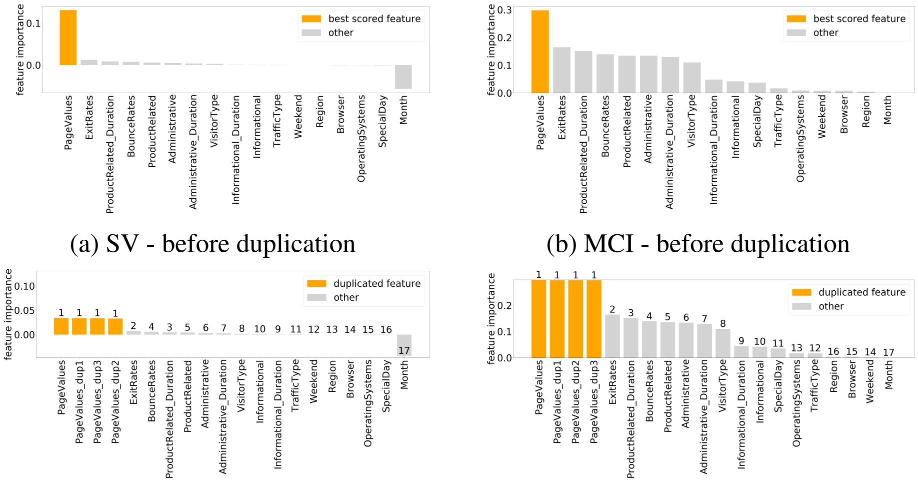
(a) SV - before duplication

(b) MCI - before duplication

(c) SV - after duplication

(d) MCI - after duplication

Figure 10: **Robustness experiment on the Bike Rental dataset.** The top row shows the feature importance according to SV (a) and MCI (b) for the original set of features in this dataset. The bottom row shows the estimations of both methods, when the top ranked feature of each method is duplicated three times. As seen, the relative differences in the importance scores given by SV (c) is affected form the introduction of duplicates, while MCI (d) succeeds to remain stable.



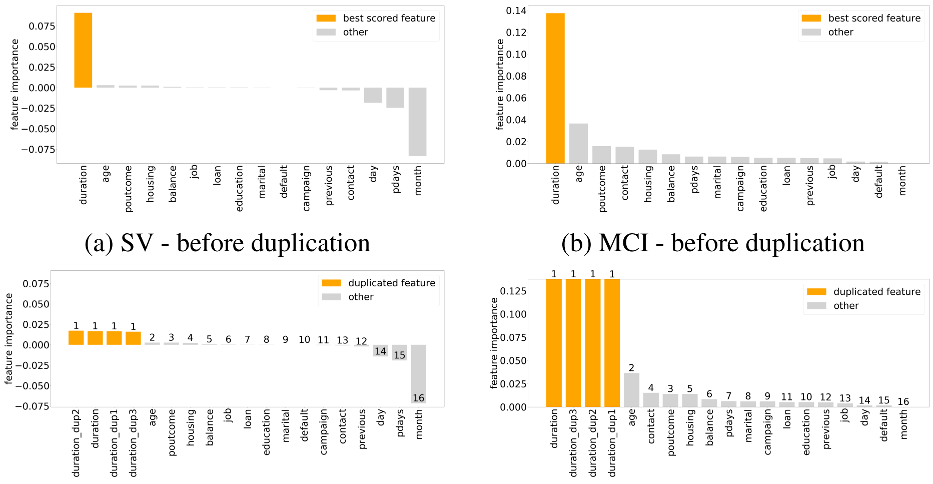
(a) SV - before duplication

(b) MCI - before duplication

(c) SV - after duplication

(d) MCI - after duplication

Figure 11: **Robustness experiment on the Online Shopping dataset.** The top row shows the feature importance according to SV (a) and MCI (b) for the original set of features in this dataset. The bottom row shows the estimations of both methods, when the top ranked feature of each method is duplicated three times. As seen, the relative differences in the importance scores given by SV (c) is affected form the introduction of duplicates, while MCI (d) succeeds to remain stable.



(a) SV - before duplication

(b) MCI - before duplication

(c) SV - after duplication

(d) MCI - after duplication

Figure 12: **Robustness experiment on the Bank Marketing dataset.** The top row shows the feature importance according to SV (a) and MCI (b) for the original set of features in this dataset. The bottom row shows the estimations of both methods, when the top ranked feature of each method is duplicated three times. As seen, the relative differences in the importance scores given by SV (c) is affected form the introduction of duplicates, while MCI (d) succeeds to remain stable.