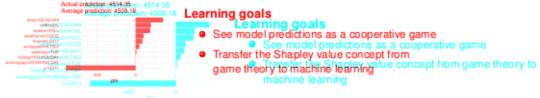
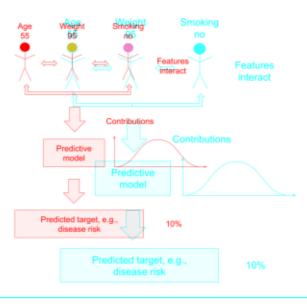
Interpretable Machine Learning

Shapley/Values for Local Explanations









• Game:: Make prediction $\hat{\mathcal{H}}(\hat{x_1}(x_2,x_2,x_2,x_p))$ for a single observation \mathbf{x}



- Game: Make prediction $\hat{f}(\vec{x_1}, x_2, x_2, x_n)$ for a single observation \hat{x}
- Players: Features $x_i, j_i \in \{1, \{1, p\}\}$ which cooperate to produce a predictionic tion
 - --- How can we make a prediction with a subset of features without changing their the mode model@ function: $\hat{f}_S(\mathbf{x}_S) := \int_{X_{-S}} \hat{f}(\mathbf{x}_S, X_{-S}) d\mathbb{P}_{X_{-S}}$ ("removing" by marginalizing over- \rightarrow PD function: $\hat{f}_S(\mathbf{x}_S) := \int_{X_{-S}} \hat{f}(\mathbf{x}_S, X_{-S}) d\mathbb{P}_{X_{-S}}$ ("removing" by marginalizing
 - over -S)

- Game:: Make prediction $\hat{f}(\hat{x_1}, x_2, x_2, x_p)$ for a single observation \mathbf{x}
- ullet Players: Features $x_j, j \in \{1, \{1, p\}, \text{ which cooperate to produce a prediction } \}$
 - How can we make a prediction i with a subset of features without changing their the mode model? function: $\hat{t}_S(\mathbf{x}_S) := \int_{X_S} \hat{t}(\mathbf{x}_S, X_{-S}) dP_{X_{-S}}$ ("removing" by marginalizing over —
- PD function: $\hat{f}_s(\mathbf{x}_s)$ in $\int_{\mathbf{X} \in \mathcal{A}} (\mathbf{x}_{s_1} \mathbf{X}_{s_2}) d\mathbf{P}_{\mathbf{X}_{s_3}}$ ("removing" by marginalizing over -S)
- Value function / payout of coalition $\hat{S}(\underline{x}_S)$ for observation/were $\hat{f}_S: \mathcal{X}_S \mapsto \mathcal{Y}$

$$\text{subtraction } \mathbf{v}(\mathbf{f}) \neq \widehat{f}(\mathbf{x}|\mathbf{x}) \text{ and } \mathbb{E}_{\mathbf{x}}(\widehat{f}(\mathbf{x})), \text{ where } \widehat{f}_{\mathbf{x}} = \mathcal{X}_{\mathbf{x}} = \mathbf{x} \text{ with } \mathbf{v}(\emptyset) = 0$$

$$\text{subtraction of } \mathbb{E}_{\mathbf{x}}(\widehat{f}(\mathbf{x})) \text{ ensures that } \mathbf{v} \text{ is a value function with } \mathbf{v}(\emptyset) = \emptyset_{S}(\mathbf{x}_{S})$$

$$\mathbb{E}(\widehat{f}(\mathbf{x})) \qquad \qquad \mathbb{E}(\widehat{f}(\mathbf{x})) \qquad \qquad \mathbb{E}(\widehat{f}(\mathbf{x})) \qquad \mathbb{E}(\widehat{f}(\mathbf{x}))$$

- Game:: Makek prediction $\hat{f}(\hat{x_1}, x_2, x_2, x_p)$ for a single observation \mathbf{x}
- Players: Features x_j, j ∈ {1,{1, p}, which cooperate to produce a prediction iction
 How can we make a prediction with a subset of features without changing their the mode.

model? function: $\hat{t}_S(\mathbf{x}_S) := \int_{X_{-S}} \hat{t}(\mathbf{x}_S, X_{-S}) d\mathbb{P}_{X_{-S}}$ ("removing" by marginalizing over -

- PD function: $\hat{f}_s(\mathbf{x}_s)$ out $\int_{\mathcal{X} \subseteq \mathcal{A}} (\mathbf{x}_{s}, \mathbf{x}_{s}) d\mathbf{P}_{\mathbf{x}_{s}} d\mathbf{P}_{\mathbf{x}_{s}}$ ("removing" by marginalizing over -S)
- Value function / payout of coalition $\hat{S}(\underline{x}_S)$ for lobservation/were $\hat{f}_S: \mathcal{X}_S \mapsto \mathcal{Y}$

$$op$$
 subtraction $\mathbf{v}(\mathbf{S})_{\mathbf{x}} + \hat{f}_{\mathbf{S}}(\mathbf{x}_{\mathbf{S}})$ is $\mathbb{E}_{\mathbf{x}}(\hat{f}(\mathbf{x}))$, where $\hat{f}_{\mathbf{S}}$ is $\mathcal{X}_{\mathbf{S}}$ in with $v(\emptyset) = 0$

 \leadsto subtraction of $\mathbb{E}_{\mathbf{x}}(\hat{f}(\mathbf{x}))$ ensures that v is a value function with $v(\emptyset) = 0$

$$\mathbb{E}(\tilde{f}(\mathbf{x}))$$
 x_3
 $\hat{f}_S(\mathbf{x}_S)$

- Marginal contribution: $v(S \cup \{j\}) v(S) = \hat{t}_{S \cup \{j\}}(\mathbf{x}_{S \cup \{j\}}) \hat{t}_{S}(\mathbf{x}_{S})$
- $\bullet \ \ \mathsf{Marginal} \ \ \mathsf{(on)tribution} \ \mathsf{Is} \ \mathsf{(SUD(G))} \ + \mathsf{he}(S) \ \mathsf{b} \ \mathsf{H} \ \ \mathsf{d}_S \ \mathsf{High} \ \mathsf{(x_SVA)} \ \mathsf{)} \ \mathsf{=} \ \mathsf{H} \ \mathsf{d}_S \ \mathsf{(x_SVA)} \ \mathsf{)} \ \mathsf{=} \ \mathsf{H} \ \mathsf{d}_S \ \mathsf{(x_SVA)} \ \mathsf{)} \ \mathsf{=} \ \mathsf{d}_S \ \mathsf{(x_SVA)} \ \mathsf{($
 - $\sim \mathbb{E}_{\mathbf{x}}(\hat{f}(\mathbf{x}))$ cancels out due to the subtraction of value functions

Shapley-value ϕ_i of feature j for observation x via order definition:

$$\phi_{j}(\mathbf{x}) = \frac{1}{|\hat{P}|^{j}} \sum_{\tau \in \Pi} \hat{\mathbf{f}}_{S_{j}^{\tau}} \underbrace{(\mathbf{x}_{S_{j}^{\tau}})(\mathbf{x}_{S_{j}^{\tau}})}_{\text{marginal contribution of feature } \hat{\mathbf{f}}_{S_{j}^{\tau}}(\mathbf{x}_{S_{j}^{\tau}}) - \hat{\mathbf{f}}_{S_{j}^{\tau}}(\mathbf{x}_{S_{j}^{\tau}})$$



- Interpretation: Feature x contributed on to difference between \(\hat{\ell}(x)\) and average prediction prediction: Marginal contributions and Shapley values can be negative
- Note: Marginal contributions and Shapley values can be negative For exact computation of $\phi_i(\mathbf{x})$, the PD function $f_S(\mathbf{x}_S) = \frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_S, \mathbf{x}_{-S}^{(i)})$ for any set of For exact computation of $\phi_i(\mathbf{x})$, the PD function $f_S(\mathbf{x}_S) = \frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_S, \mathbf{x}_{-S}^{(i)})$ for any set of
- any set of features S can be used which yields $r_s(\mathbf{x}_s) = \frac{1}{n}$

$$\phi_{j}(\mathbf{x}) = \frac{\phi_{j}(\mathbf{x})}{|P|! \cdot n} \sum_{\tau \in D} \sum_{i=1}^{n} \hat{f}(\mathbf{x}) \sum_{S_{i}^{\tau} \mid i \mid j \mid i} \hat{\mathbf{x}}_{(j \mid i)}^{(j)} \hat{\mathbf{x}}_{(j \mid i)}^{(j)} \hat{\mathbf{x}}_{(j \mid i)}^{(j)} \hat{\mathbf{x}}_{(j \mid i)}^{(j)} \hat{\mathbf{x}}_{(i)}^{(j)} \hat{\mathbf{x}}_{($$

Note: \hat{f}_S marginalizes over all other features -S using all observations $i=1,\ldots,n$ $i=1,\ldots,n$

ESTIMATION: A PRACTICAL PROBLEM

Exact Shapley-value computation is problematic for high-dimensional featureature spaces spacesr 10 features, there are already |P|! = 10! ≈ 3.6 million possible orders of features
 → For 10 features, there are already |P|! = 10! ≈ 3.6 million possible orders of features

IESTIMATION: A PRACTICAL PROBLEM

- Exact: Shapley-value computation is problematic for high-dimensional feature spaces spaces 10 features, there are already $|P|! = 10! \approx 3.6$ million possible orders of features
- \sim A Equiton features, there are already in the hard of the hard
- Additional problem due to estimation of the marginal prediction f̂_{S_j}: Averaging over the entire data set for each coalition S_j introduced by τ can be very expensive for large data sets

ESTIMATION: A PRACTICAL PROBLEM

- Exact Shapley-value computation is problematic for high-dimensional feature spaces spaces 10 features, there are already $|P|! = 10! \approx 3.6$ million possible orders of features
 - A For_{ii}10 features there are already of both 10 harding possible orders of ing over the features of for each coalition S_i^{τ} introduced by τ can be very expensive for large data sets
- Additional problem due to estimation of the marginal prediction f_{Si}. Averaging terms, we over the entire data set for each coalition S_i introduced by Tream be very build coalitions S_i expensive for large data sets
- Solution to both problems is sampling: Instead of averaging over |P|! · n terms, we approximate it using a limited amount of M random samples of τ to build coalitions S_i^τ

ESTIMATION: A PRACTICAL PROBLEM

- Exact Shapley-value computation is problematic for high-dimensional feature ature spaces. spaces 10 features, there are already $|P|! = 10! \approx 3.6$ million possible orders of features.
 - A For in 10 features, there are already in the interpretation over the an features of the each coalition $S_i^{ au}$ introduced by au can be very expensive for large data sets
- Additional problem due to estimation of the marginal prediction fs.: Averaging erms, we over the entire data set for each coalition ST introduced by T can be very build coalitions ST expensive for large data sets
- M is a tradeoff between accuracy of the Shapley value and computational costs
 Solution to both problems is sampling: Instead of averaging over I PII in terms.

 Solution to both problems is sampling: Instead of averaging over I PII in terms. we approximate it using a limited amount of M random samples of τ to build coalitions ST
- M is a tradeoff between accuracy of the Shapley value and computational costs
 - → The higher M, the closer to the exact Shapley values, but the more costly the computation

APPROXIMATION ALGORITHM (Strumbelj et al. (2014)

Estimation of ϕ_i for observation \mathbf{x} of model \hat{d} fitted on data \mathcal{D} using sample size $\mathbf{Mz} \in \mathcal{M}$:

● Form == 1,1..., M do:to:



APPROXIMATION ALGORITHM (Strumbel) et al. (2014)

Estimation of ϕ_i for observation or of model \hat{d} filted on data \mathcal{D} using sample size $Mz \in M$:

- Forom = 1,1..., M doilo:
 - Select random order ℓ permand feature indices ε indices τ i



APPROXIMATION ALGORITHM Strumbelj et al. (2014)

Estimation of ϕ_{ℓ} for observation \mathbf{x} of model \hat{d} fittled on data \mathcal{D} using sample size $\mathbf{M} \mathbf{z} \in M$:

- \bigcirc For $m_{\pi}=1,1...,M$ dodo:
 - Select random order ℓ permantifeature indices $\varepsilon = (\varepsilon^{(1)}, \tau \to \tau^{(p)}) \in \Pi$, $\tau^{(p)}) \in \Pi$.

APPROXIMATION ALGORITHM Strumbelj et al. (2014)

Estimation of ϕ_{ℓ} for observation \mathbf{x} of model \hat{d} fittled on data \mathcal{D} using sample size $\mathbf{M} \mathbf{z} \in M$:

- \bullet For $m \neq 1, \dots, M$ do do:
 - Select random order ℓ perm, of feature indices ε $\exists t$ $\exists t$ $\exists t \in \mathbb{S}^{(1)}, \tau := \tau^{(p)} \in \Pi$, $\tau^{(p)} \in \Pi$
 - **Q** Determine coalition $S_m S = S_i^T$ i.e., the set of feat before feat of injerder σ_i in order τ
 - Select random data point **z**(m) ∈ D D

APPROXIMATION ALGORITHM Strumbell et al. (2014)

Estimation of ϕ_{ℓ} for observation \mathbf{x} of model \hat{d} fitted on data \mathcal{D} using sample size $\mathbf{M} \mathbf{z} \in M$:

- \bullet For $m \neq 1, \dots, M$ do do:
 - Select random order ℓ permant feature indices ε mo(ε es., $\tau = \tau^{(p)}$) $\in \Pi$, $\tau^{(p)}$) $\in \Pi$
 - **Q** Determine coalition $S_m S = S_i^T$, i.e., the set of feat before feat of injerder σ_i in order τ
 - Select random data point **z**(m) (€ D ¬
 - Construct two artificial obs. by replacing feature values from x with $\mathbf{z}_{rcm}^{(m)}$ with $\mathbf{z}_{rcm}^{(m)}$.

APPROXIMATION ALGORITHM Strumbeli et al. (2014)

Estimation of ϕ_{ℓ} for observation \mathbf{x} of model \hat{d} fitted on data \mathcal{D} using sample size $\mathbf{M} \mathbf{z} \in M$:

- \bigcirc For $m \neq 1, \dots, M$ do \exists_0 :
 - Select random order ℓ permant feature indices ε indices τ indices τ
 - **Q** Determine coalition $S_m = S_i^T$ die i the set of feat obefore feat of injerder σ_i in order τ
 - Select random data point **z**(m) (€ D ¬
 - Construct two artificial obs. by replacing feature values from x with $z_{i,j}^{(m)}$: with $z_{i,j}^{(m)}$.
 - $\mathbf{x}_{+j}^{(m)} = (x_{+(1)}, \dots, x_{\tau(\lfloor Sm \rfloor 1)}, x_j, z_{\tau(j)}^{(m)}, z_{\tau(j)}^{(m)}, \dots, z_{\tau(j)}^{(m)})$ takes features $S_m \cup \{j\}$ from \mathbf{x}
 - $X_{S_m \cup \{j\}}$ $X_{S_m \cup \{j\}}$ $\mathbf{z}_{-\{\frac{S_{m}\cup\{j\}}{2}\}}^{(m)}$ $S_m \cup \{j\}$ from **x**

APPROXIMATION ALGORITHM (5 Strumbel) et al. (2014)

Estimation of ϕ_{ℓ} for observation \mathbf{x} of model \hat{d} fittled on data \mathcal{D} using sample size $\mathbf{M} \mathbf{z} \in M$:

- \bigcirc For $m \neq 1, \dots, M$ do \exists_0 :
 - Select random order ℓ permant feature indices ε indices τ indices τ
 - **Q** Determine coalition $S_m = S_i^T$ die it the set of feat (before feat of in order σ_i in order σ_i
 - Select random data point **z**(m) (€ D ¬
 - Construct two artificial obs. by replacing feature values from x with $z_{i,j}^{(m)}$: with $z_{i,j}^{(m)}$.

•
$$\mathbf{x}_{j}^{(m)} = (x_{j+1}, \dots, x_{\tau(|S_m|-1)}, x_j, z_{\tau(|S_m|-1)}, \dots, z_{\tau(p)})$$
 (takes features $S_m \cup \{j\}$ from \mathbf{x}

$$\mathbf{x}_{S_m \cup \{j\}} = (x_{j+1}, \dots, x_{\tau(|S_m|-1)}, x_j, z_{\tau(|S_m|+1)}, \dots, z_{\tau(p)})$$

$$\mathbf{x}_{S_m \cup \{j\}} = (x_{j+1}, \dots, x_{\tau(|S_m|-1)}, x_j, z_{\tau(|S_m|+1)}, \dots, z_{\tau(p)})$$

$$\mathbf{x}_{S_m \cup \{j\}} = (x_{j+1}, \dots, x_{\tau(|S_m|-1)}, x_j, z_{\tau(|S_m|+1)}, \dots, z_{\tau(p)})$$

$$\mathbf{x}_{S_m \cup \{j\}} = (x_{j+1}, \dots, x_{\tau(|S_m|-1)}, x_j, z_{\tau(|S_m|-1)}, \dots, z_{\tau(p)})$$

$$\mathbf{x}_{S_m \cup \{j\}} = (x_{j+1}, \dots, x_{\tau(|S_m|-1)}, x_j, z_{\tau(|S_m|-1)}, \dots, z_{\tau(p)})$$

$$\mathbf{x}_{S_m \cup \{j\}} = (x_{j+1}, \dots, x_{\tau(|S_m|-1)}, x_j, z_{\tau(|S_m|-1)}, \dots, z_{\tau(p)})$$

$$\mathbf{x}_{S_m \cup \{j\}} = (x_{j+1}, \dots, x_{\tau(|S_m|-1)}, x_j, z_{\tau(|S_m|-1)}, \dots, z_{\tau(p)})$$

$$\mathbf{x}_{S_m \cup \{j\}} = (x_{j+1}, \dots, x_{\tau(|S_m|-1)}, x_j, z_{\tau(|S_m|-1)}, \dots, z_{\tau(p)})$$

$$\mathbf{x}_{S_m \cup \{j\}} = (x_{j+1}, \dots, x_{\tau(|S_m|-1)}, x_j, z_{\tau(|S_m|-1)}, \dots, z_{\tau(p)})$$

$$\mathbf{x}_{S_m \cup \{j\}} = (x_{j+1}, \dots, x_{\tau(|S_m|-1)}, x_j, z_{\tau(|S_m|-1)}, \dots, z_{\tau(p)})$$

$$\mathbf{x}_{S_m \cup \{j\}} = (x_{j+1}, \dots, x_{\tau(|S_m|-1)}, x_j, z_{\tau(|S_m|-1)}, \dots, z_{\tau(p)})$$

$$S_m \cup \{j\}$$
 from **x**

$$\bullet \mathbf{x}_{-j}^{(m)} = \underbrace{(x_{\tau(1)}^{(1)}, \dots, x_{\tau(1Sm)-1}^{(1Sm)-1}, \underbrace{z_{j}^{(m)}}_{\mathbf{x}_{S_{m}}^{(m)}}, \underbrace{z_{\tau(1S_{m}]+1)}^{(m)}, \underbrace{z_{\tau(1S_{m}]+1)}^{(m)}, \underbrace{z_{\tau(p)}^{(m)}}_{\mathbf{x}_{S_{m}}^{(m)}}}_{\mathbf{x}_{S_{m}}^{(m)}}, \underbrace{z_{\tau(p)}^{(m)}}_{\mathbf{x}_{S_{m}}^{(m)}}, \underbrace{z_{\tau(p)}^{(m)}}_{\mathbf{x}_{S_$$

 S_m from x

APPROXIMATION ALGORITHM (5 Strumbel) et al. (2014)

Estimation of ϕ_i for observation \mathbf{x} of model \hat{d} fitted on data \mathcal{D} using sample size $Mz \in M$:

- \bigcirc For $m \neq 1, \dots, M$ do \exists_0 :
 - Select random order ℓ permant feature indices ε indices τ indices τ
 - **Q** Determine coalition $S_m S_m = S_i^T$ die i the set of feat chefore feat of inverder σ_i in order τ
 - Select random data point **z**(m) (€ D ¬
 - Construct two artificial obs. by replacing feature values from x with $z_{i,j}^{(m)}$: with $z_{i,j}^{(m)}$.

$$\bullet \mathbf{x}_{i,j}^{(m)} = \underbrace{(\mathbf{x}_{i+1}, \dots, \mathbf{x}_{r}(|\mathbf{x}_{m}|-1), \mathbf{x}_{j}, \mathbf{z}_{r}(|\mathbf{x}_{m}|-1), \dots, \mathbf{z}_{r}(\rho)}_{\mathbf{x}_{m}(|\mathbf{x}_{m}|-1), \mathbf{x}_{j}, \mathbf{z}_{m}(|\mathbf{x}_{m}|-1), \dots, \mathbf{z}_{r}(\rho)}) \text{ takes features } S_{m} \cup \{j\} \text{ from } \mathbf{x}_{m} \cup \{j$$

 $S_m \cup \{j\}$ from **x**

$$\bullet \overset{\bullet}{\mathbf{x}} \overset{(m)}{\underbrace{\longrightarrow}} = \underbrace{(\overset{X_{-(1)}}{\underbrace{\longleftarrow}}, \overset{X_{-(1)}}{\underbrace{\longleftarrow}}, \overset{X_{-(1)}}{\underbrace{\longleftarrow}}, \overset{X_{-(1)}}{\underbrace{\longleftarrow}}, \overset{(m)}{\underbrace{\longleftarrow}}, \overset{(m)}{\underbrace{\longleftarrow}, \overset{(m)}{\underbrace{\longleftarrow}}, \overset{(m)}{\underbrace{\longleftarrow}}, \overset{(m)}{\underbrace{\longleftarrow}, \overset{(m)}{\underbrace{\longleftarrow}}, \overset{(m)}{\underbrace{\longleftarrow}}, \overset{(m)}{\underbrace{\longleftarrow}}, \overset{(m)}{\underbrace{\longleftarrow}}, \overset{(m)}{\underbrace{\longleftarrow}}, \overset{(m)}{\underbrace{\longleftarrow}, \overset{(m)}{\underbrace{\longleftarrow}}, \overset{(m)}{\underbrace{\longleftarrow}}, \overset{(m)}{\underbrace{\longleftarrow}, \overset{(m)}{\underbrace{\longleftarrow}}, \overset{(m)}{\underbrace{\longleftarrow}, \overset{(m)}{\underbrace{\longleftarrow}, \overset{(m)}{\underbrace{\longleftarrow}, \overset{(m)}{\underbrace{\longleftarrow}}, \overset{(m)}{\underbrace{\longleftarrow}, \overset{(m)}{\underbrace{\longleftarrow}$$

- Confinite materials $\phi_i^m = \hat{t}(\mathbf{x}_{+i}^{lm})$
- Compute difference $\phi_{S_m}^m = \hat{f}(\mathbf{x}_{S_m}^{(m)}) \cdot \hat{f}(\hat{f}(\mathbf{x}_{-j}^{(m)}))$ and $\hat{f}_{S_m \cup \{j\}}(\mathbf{x}_{S_m \cup \{j\}})$ by $\hat{f}(\mathbf{x}_{+j}^{(m)})$ over M iters $f(\mathbf{x}_{S_m}^m)$ is approximated by $f(\mathbf{x}_{-j}^m)$ and $f(\mathbf{x}_{S_m \cup \{j\}}, \mathbf{x}_{S_m \cup \{j\}})$ by $f(\mathbf{x}_{+j}^m)$ over M iters.

APPROXIMATION ALGORITHM Strumbelj et al. (2014)

Estimation of ϕ_i for observation \mathbf{x} of model \hat{d} fitted on data \mathcal{D} using sample size $Mz \in M$:

- \bigcirc Form =1,1...,M do to:
 - Select random order ℓ permant feature indices ε mo(ε es., $\tau = \tau^{(p)}$) $\in \Pi$, $\tau^{(p)}$) $\in \Pi$
 - **Q** Determine coalition $S_m S_m = S_i^T$ die i the set of feat chefore feat of inverder σ_i in order τ
 - Select random data point **z**(m) (€ D ¬
 - Construct two artificial obs. by replacing feature values from \mathbf{x} with $\mathbf{z}_{\text{troin}}^{(m)}$: with $\mathbf{z}_{\text{troin}}^{(m)}$.

•
$$\mathbf{x}_{t+j}^{(m)} = (x_{\tau(1)}, \dots, x_{\tau(|S_m|-1)}, x_j, z_{\tau(|S_m|-1)}, \dots, z_{\tau(p)})$$
 (takes features $S_m \cup \{j\}$ from \mathbf{x}

$$S_m \cup \{j\}$$
 from **x**

$$\bullet \overset{\bullet}{\mathbf{x}} \overset{(m)}{\underset{j}{\stackrel{(n)}{=}}} = \underbrace{(\overset{X_{-1}}{\underset{(1,\dots,x)}{\stackrel{(n)}{=}}},\overset{X_{-1}}{\underset{(1,\dots,x)}{\stackrel{(m)}{=}}},\overset{(m)}{\underset{(m)}{\stackrel{(m)}{=}}},\overset{(m)}{\underset{(m)}{\stackrel{(m)}{=}}$$

- Confinite mixerence $\phi_i^m = \hat{t}(\mathbf{x}_{+i}^{(m)})$
- Compute difference $\phi_{m}^{m} = \hat{f}(\mathbf{x}_{-j}^{(m)})^{-j} \hat{f}(\mathbf{x}_{-j}^{(m)})^{-j} \text{ and } \hat{f}_{S_{m} \cup \{j\}} (\mathbf{x}_{S_{m} \cup \{j\}}) \text{ by } \hat{f}(\mathbf{x}_{+j}^{(m)}) \text{ over } M \text{ iters}$ $f_{S_{m}}(\mathbf{x}_{S_{m}}) \text{ is approximated by } \hat{f}(\mathbf{x}_{-j}^{(m)}) \text{ and } \hat{f}_{S_{m} \cup \{j\}} (\mathbf{x}_{S_{m} \cup \{j\}}) \text{ by } \hat{f}(\mathbf{x}_{+j}^{(m)}) \text{ over } M \text{ iters}$ $Compute difference <math>\phi_{m}^{m} = \hat{f}(\mathbf{x}_{-j}^{(m)})^{-j} \hat{f}(\mathbf{x}_{-j}^{(m)})^{-j} \hat{f}(\mathbf{x}_{-j}^{(m)}) \text{ over } M \text{ iters}$
- Ocompute Shapley value $\phi_i = \frac{1}{M} \sum_{m=1}^{M} \phi_i^m$

SHAPLEY VALUE APPROXIMATION - ILLUSTRATION



x: obs. of interestobs. of interestops. of interestops. of interestops. of interestops. of interestops. of interestops. Other are replaced of interestops. Other are replaced of interestops.

$$\phi_{j}(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} \left[\hat{f}(\mathbf{x}_{-j}^{1}) \left[\hat{f}(\mathbf{x}_{-j}^{(m)}) \right] - \hat{f}(\mathbf{x}_{-j}^{(m)}) \right]$$

x with feature values in $S_m \cup \{j\}$ $S_m \cup \{j\}$

l	Temperatu	ure H	Humidity		/indspeed	Year
x	Temperature 66	Humidity	56 Windspe	ed	11 Year	2012
x_{+i}	10.66	56	56 11	rand	2012	2012
$x_{\pm j}$	10.66	56	random : z		2012	1 1 1 1 1 1 1 1
x_{-j}^{-j}	10.66	56	eurodone : z	ndspeed	random i spar	random : z _{pen} -
					<u> </u>	
					J	Ĵ

SHAPLEY VALUE APPROXIMATION - ILLUSTRATION

Definition

Contribution of feature
$$j$$
 to coalition S_m
$$\phi_j(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} \left[\hat{f}(\mathbf{x}_{+j}^{\top})^{m} + \hat{f}(\mathbf{x}_{+j}^{\top})^{m} \right] + \hat{f}(\mathbf{x}_{-j}^{\top})^{m} + \hat{f}(\mathbf{x}$$

- $\Delta(j(S_m)_m) \neq \hat{I}(\mathbf{x}_{-j_m}^{(m)})^{m-1}\hat{I}(\mathbf{x}_{-j_m}^{(m)})^{m-$
- Here: Feature year contributes +700 bike rentals if it joins coalition

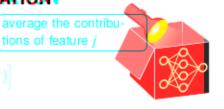


SHAPLEY VALUE APPROXIMATION - ILLUSTRATION

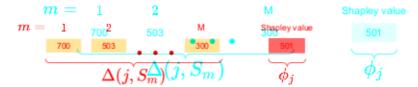
Definition

average the contributions of leature
$$j$$

$$\phi_{j}(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} \left[\hat{f}(\mathbf{x}_{m}) \right] \hat{f}(\mathbf{x}_{m}) \hat{f}(\mathbf{x}_{m}) \hat{f}(\mathbf{x}_{m})$$



- Compute marginal contribution of reafure / rowards the prediction across all randomly drawn
- randomity drawin feature coalitions S_1, \ldots, S_m Average all M marginal contributions of feature jAverage all M marginal contributions of feature jShapley value ϕ_j is the payout of feature j, i.e., how much feature year contributed to the
- Shapley value ϕ_i is the payout of feature j, i.e. how much feature year contributed to the overall prediction in bicycle counts of a specific observation x



We take the general axioms for Shapley Values and apply it to predictions:

• Efficiency: Shapley values add up to the (centered) prediction: $\sum_{j=1}^p \phi_j = \hat{f}(\mathbf{x}) - \mathbb{E}_{\mathbf{x}}(\hat{f}(X))$ $\sum_{j=1}^p \phi_j = \hat{f}(\mathbf{x}) - \mathbb{E}_{\mathbf{x}}(\hat{f}(X))$



We take the general axioms for Shapley Values and applying predictions:

- •• Efficiency: Shapley evalues add up to the (centered) prediction: $\sum_{j=1}^p \phi_j = \hat{f}(\mathbf{x}) \mathbb{E}_{\mathbf{x}}(\hat{f}(X))$ Symmetry: Two features j and k that contribute the same to the prediction get the same page.
- Symmetry: Two-features \hat{g} and k that contribute the same to the prediction get the same payout = $\hat{f}_{S \cup \{k\}}(\mathbf{x}_{S \cup \{k\}})$ for all $S \subseteq P \setminus \{j,k\}$ then $\phi_j = \phi_k$ \leadsto interaction effects between features are fairly divided $\hat{f}_{S \cup \{j\}}(\mathbf{x}_{S \cup \{j\}}) = \hat{f}_{S \cup \{k\}}(\mathbf{x}_{S \cup \{k\}})$ for all $S \subseteq P \setminus \{j,k\}$ then $\phi_j = \phi_k$

We take the general axioms for Shapley Values and apply it to predictions tions.

- Efficiency: Shapley values add up to the (centered) prediction: $\sum_{i=1}^{p} \phi_i = \hat{t}(\mathbf{x}) \mathbf{e}_{\mathbf{x}}(\hat{t})$ $\sum_{j=1}^{p} \phi_j = \hat{t}(\mathbf{x}) - \mathbb{E}_{\mathbf{x}}(\hat{t}(X))$ • Symmetry: Two features j and k that contribute the same to the prediction get the same
- Symmetry: Two features j and k that contribute the same to the prediction get the same payout = $f_{S \cup \{k\}}$ ($\mathbf{x}_{S \cup \{k\}}$) for all $S \subseteq P \setminus \{j, k\}$ then $\phi_i = \phi_k$ --- interaction effects between features are fairly divided
- if a feature was not selected by the model (e.g., tree or LASSO), its Shapley value is zero
- Dummy / Null Player: Shapley value of a feature that does not influence the prediction is zero \iff if a feature was not selected by the model (e.g., tree or LASSO), its Shapley value is zero $\hat{f}_{S\cup\{i\}}(\mathbf{x}_{S\cup\{i\}}) = \hat{f}_{S}(\mathbf{x}_{S})$ for all $S\subseteq P$ then $\phi_i = 0$

We take the general axioms for Shapley Values and apply it to predictions tions:

- Efficiency: Shapley values add up to the (centered) prediction: $\sum_{j=1}^{p} \phi_j = \hat{f}(\mathbf{x}) \mathbb{E}_{\mathbf{x}}(\hat{f}(\mathbf{x}))$
- $\sum_{j=1}^{p} \phi_j = \hat{f}(\mathbf{x}) \mathbb{E}_{\mathbf{x}}(\hat{f}(X))$ Symmetry: Two features j and k that contribute the same to the prediction get the same
- Symmetry: Two features j and k that contribute the same to the prediction get the same payout = f̂_{S∪{k}} (x_{S∪{k}}) for all S ⊆ P \ {j, k} then φ_j = φ_k
 interaction effects between features are fairly divided
- Flave (xShap) for all S of P (4), it influence the prediction is zero if a feature was not selected by the model (e.g., tree or LASSO), its Shapley value is zero
- Dummy / Null Player: Shapley value of a feature that does not influence the prediction is zero → if a feature was not selected by the model (e.g., tree or
- •LASSO) its Shapfey value is zero combined payouts, the payout is the sum of payouts: $\hat{f}_{SO}(y(\mathbf{x}_{SO}(y))) = \hat{f}_{SO}(\mathbf{x}_{SO}(y))$
- Additivity: For a prediction with combined payouts, the payout is the sum of payouts: φ_j(v₁) + φ_j(v₂) → Shapley values for model ensembles can be combined

BIKE SHARING DATASET





- Shapley values of observation i = 200 from the bike sharing data
- Difference between model prediction of this observation and the average Shapley values of observation i = 200 from the pike sharing data prediction of the data is fairly distributed among the features (i.e.,
- •4434ere4507betw73) model prediction of this observation and the average prediction of the data
- Feature value temp = 28.5 has the most positive effect, with a contribution
- •(increase of prediction) of about +400 most positive effect, with a contribution (increase of prediction) of about +400

ADVANTAGES AND DISADVANTAGES

Advantages:s:

- Solid theoretical foundation in gamentheory ory
- Prediction is fairly distributed among the feature/values seasy to interpret for a user
- •a defire a tive explanations that compare the prediction with the average prediction
- Contrastive explanations that compare the prediction with the average Disprediction es:
- Without sampling, Shapley values need a lot of computing time to inspect all possible coalitions Disadvantages:

 Like many other IML methods, Shapley values suffer from the inclusion of unrealistic data

 - Without sampling, Shapley values need a lot of computing time to inspect all possible coalitions
 - Like many other IML methods, Shapley values suffer from the inclusion of unrealistic data observations when features are correlated