Proper prediction explanation with *shapr*

Martin Jullum

NAV
Oslo, June 26th 2020

Example: Bank creates mortgage robot

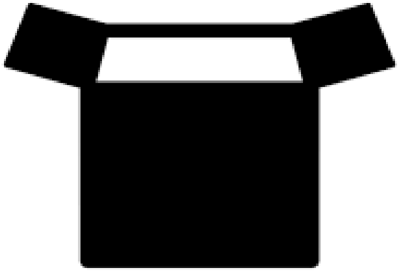
Transaction history

Commercial Card - 456480111111111: 08/04/2004 - 14/04/2004			
Date Processed	Description	Debit	Credit
08/04/2004	CASH ADVANCE FEE	$5.00-	
09/04/2004	SUNSHINE VILLAGE BANFF	$86.97-	
10/04/2004	PRINCIPAL CREDIT ADJUSTMENT	$22.00-	
10/04/2004	CARD MEMBERSHIP FEE	$19.00-	
10/04/2004	PHOTOCARD FEE	$3.00-	
11/04/2004	PRINCIPAL DEBIT ADJUSTMENT	$22.00-	
11/04/2004	PRINCIPAL DEBIT ADJUSTMENT	$22.00-	
12/04/2004	PRINCIPAL CREDIT ADJUSTMENT	$22.00-	

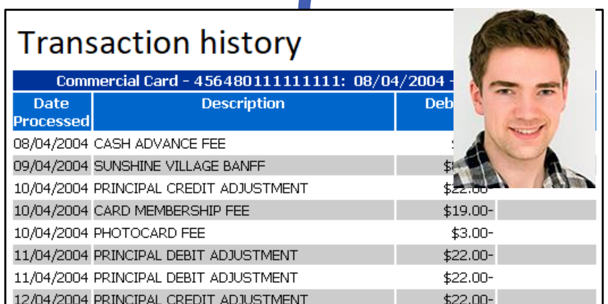
Age/gender/prev. loans…

&

Defaulted loan?



Example: Bank creates mortgage robot



Transaction history

Commercial Card - 456480111111111: 08/04/2004 -			
Date Processed	Description	Deb	
08/04/2004	CASH ADVANCE FEE		
09/04/2004	SUNSHINE VILLAGE BANFF		
10/04/2004	PRINCIPAL CREDIT ADJUSTMENT		
10/04/2004	CARD MEMBERSHIP FEE	$19.00-	
10/04/2004	PHOTOCARD FEE	$3.00-	
11/04/2004	PRINCIPAL DEBIT ADJUSTMENT	$22.00-	
11/04/2004	PRINCIPAL DEBIT ADJUSTMENT	$22.00-	
12/04/2004	PRINCIPAL CREDIT ADJUSTMENT	$22.00-	

Age/gender/prev. loans…

P(Mortgage default)

Mortgage granted ?

$$\boldsymbol{x} \longrightarrow f(\boldsymbol{x}) \longrightarrow p = 0.7$$

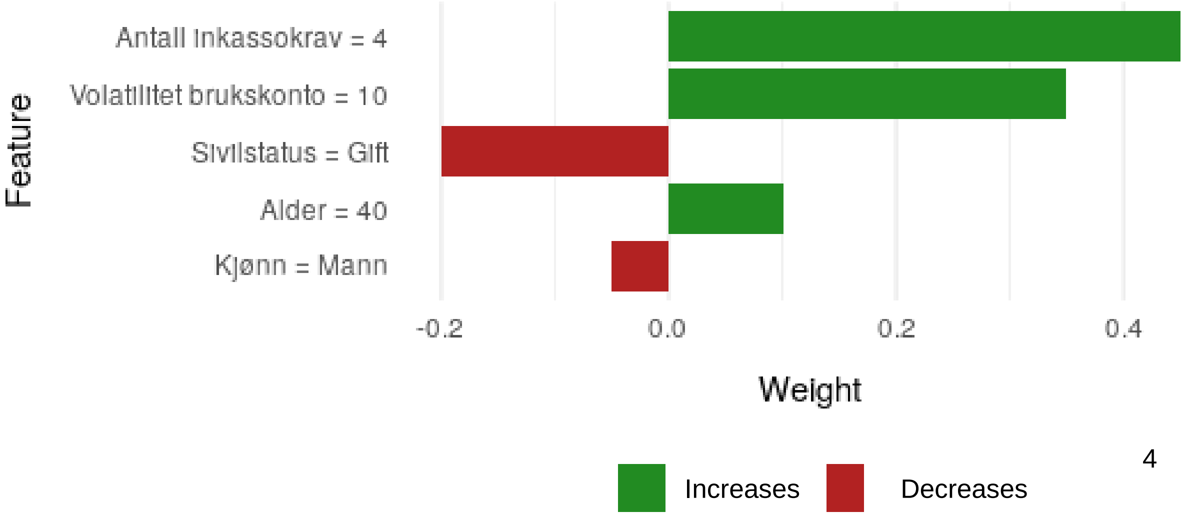
<u>Why</u> was  rejected a loan?

<u>Individual</u> prediction explanation

- **<u>NOT</u>** a general explanation of the black-box model
- $\boldsymbol{x} = \boldsymbol{x}^*$: Transaction history/features for

<u>Explanation for $f(\boldsymbol{x}^*) = 70\%$</u>

- How did each feature contribute to increase/decrease the predicted value $f(\boldsymbol{x}^*) = 70\%$?



Why is this important?

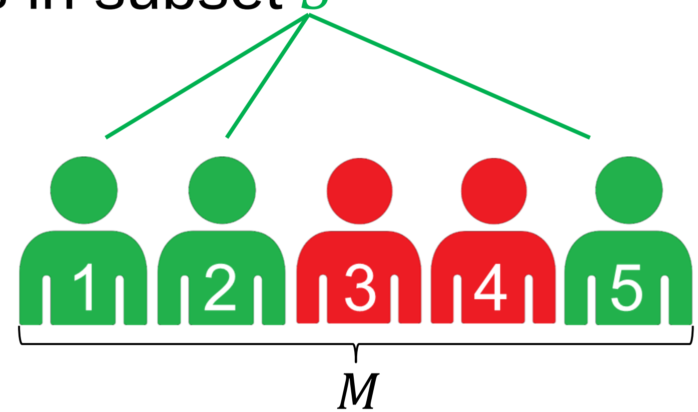
- ► Customers may have a “right to an explanation”
- ► Builds trust to the “robot”

Shapley values

- Concept from (cooperative) game theory in the 1950s
- Used to distribute the total payoff to the players
- Explicit formula for the "fair" payment to every player j:

$$\phi_j = \sum_{S \subseteq M \setminus \{j\}} w(S)\left(v(S \cup \{j\}) - v(S)\right), \qquad w(S) \text{ is a weight function}$$

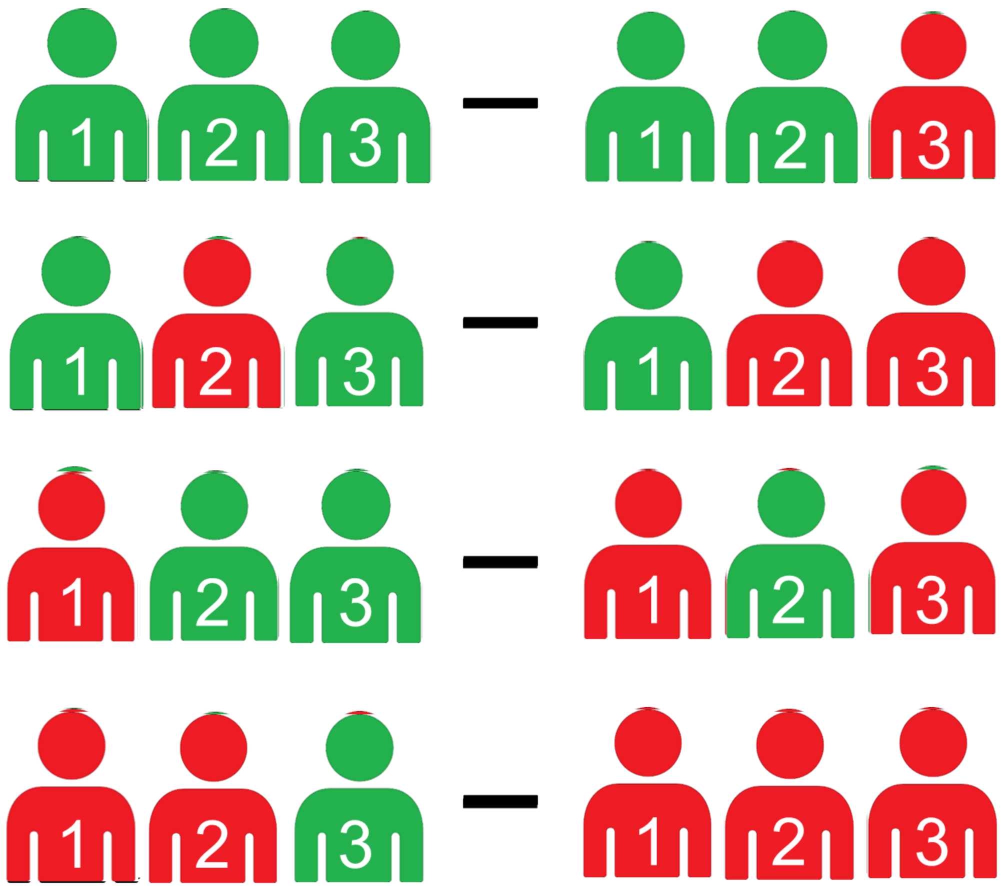
$v(S)$ is the payoff with only players in subset S



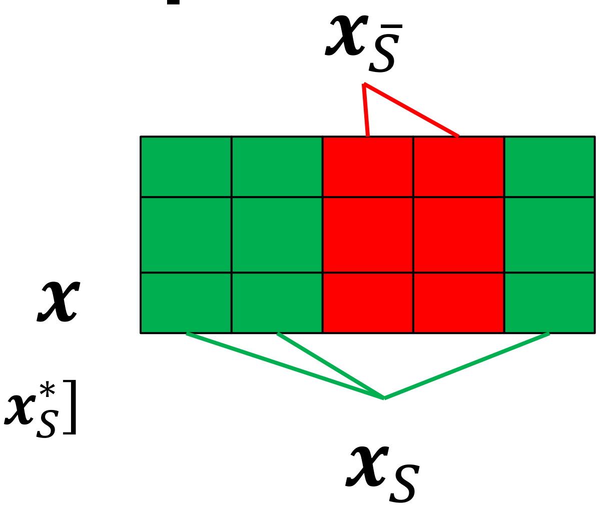
- Several mathematical optimality properties

Intuition behind the Shapley formula

Game with 3 players



Shapley values for prediction explanation



- Players = features ($x_1, \ldots, x_M$)
- Payoff = prediction ($f(\boldsymbol{x}^*)$)
- Contribution function: $v(S) = E[f(\boldsymbol{x})|\boldsymbol{x}_S = \boldsymbol{x}_S^*]$
- Properties

$f(\boldsymbol{x}^*) = \sum_{j=0}^{M} \phi_j$

$\phi_0 = E[f(\boldsymbol{x})]$

$E[f(\boldsymbol{x})] = E[f(\boldsymbol{x})|x_j]$ implies $\phi_j = 0$

x_i, x_j same contribution implies $\phi_i = \phi_j$

- Rough interpretation of ϕ_j: **The prediction change when you don't know the value of** x_j -- averaged over all features

Shapley values for prediction explanation

- 2 main challenges

1. The computational complexity in the Shapley formula

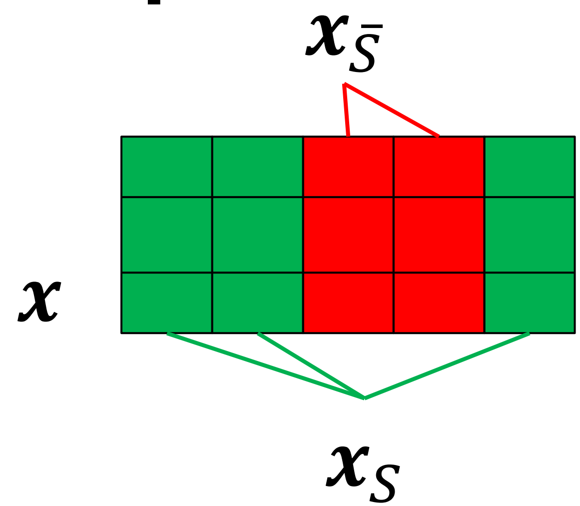
$$\phi_j = \sum_{S \subseteq M \setminus \{j\}} w(S) \left(v(S \cup \{j\}) - v(S) \right)$$

   - Partly solved by cleverly reducing the sum by subset sampling (KernelSHAP; Lundberg & Lee, 2017)

Shapley values for prediction explanation

- 2 main challenges

Recall



2. Estimating the contribution function

$$v(S) = E[f(\boldsymbol{x})|\boldsymbol{x}_S = \boldsymbol{x}_S^*] = \int f(\boldsymbol{x}_{\bar{S}}, \boldsymbol{x}_S) p(\boldsymbol{x}_{\bar{S}}|\boldsymbol{x}_S = \boldsymbol{x}_S^*) \mathrm{d}\boldsymbol{x}_{\bar{S}}$$

- Previous methods
  - Approximates $v(S) \approx \int f(\boldsymbol{x}_{\bar{S}}, \boldsymbol{x}_S^*) p(\boldsymbol{x}_{\bar{S}}) \mathrm{d}\boldsymbol{x}_{\bar{S}}$,
  - Estimates $p(\boldsymbol{x}_{\bar{S}})$ using the empirical distribution of the training data
  - Monte Carlo integration to solve the integral

This assumes covariates are independent!

Consequences of the independence assumption

- Requires evaluating $f(\boldsymbol{x}_{\bar{S}}, \boldsymbol{x}_S)$ at potentially **unlikely or illegal** combinations of $\boldsymbol{x}_{\bar{S}}$ and $\boldsymbol{x}_S$

- Example 1
  - Number of transactions to Switzerland: 0
  - Average transaction amount to Switzerland: 1000 NOK

- Example 2
  - Age: 17
  - Marital status: Widow
  - Profession: Professor

Our idea

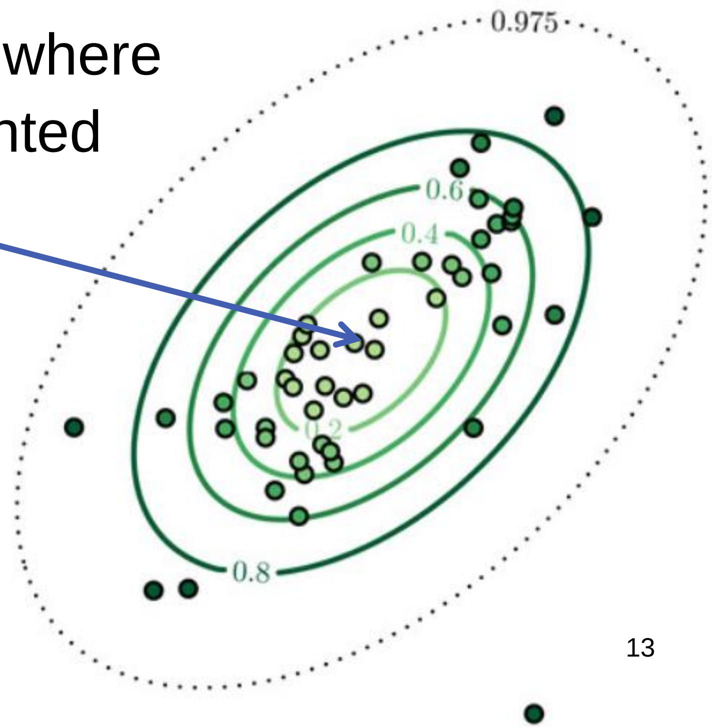
Estimate $p(\boldsymbol{x}_{\bar{S}} | \boldsymbol{x}_S = \boldsymbol{x}_S^*)$ properly

+

Monte Carlo integration

Continuous features

- How to estimate $p(\boldsymbol{x}_{\bar{S}} | \boldsymbol{x}_S = \boldsymbol{x}_S^*)$ when $\boldsymbol{x}$ is continuous?
- 3 approaches
  - Assume $p(\boldsymbol{x})$ Gaussian => analytical $p(\boldsymbol{x}_{\bar{S}} | \boldsymbol{x}_S = \boldsymbol{x}_S^*)$
  - Assume Gaussian copula => transformation + analytical expression
  - An empirical (conditional) approach where training observations at $\boldsymbol{x}_{\bar{S}}^{\boldsymbol{i}}$ are weighted based on proximity of $\boldsymbol{x}_S^{\boldsymbol{i}}$ to $\boldsymbol{x}_S^*$



Explaining sick leave predictions

- NAV is modelling how long individuals are on sick leave
  - Used by case workers to schedule follow-up meetings
- Case workers need to understand the “reasoning” of the individual predictions
- Modelling based on
  - age, gender, sick leave history, type of business etc.
  - Several categorical features with many levels
- Need methodology for prediction explanation which can handle categorical features

Our idea

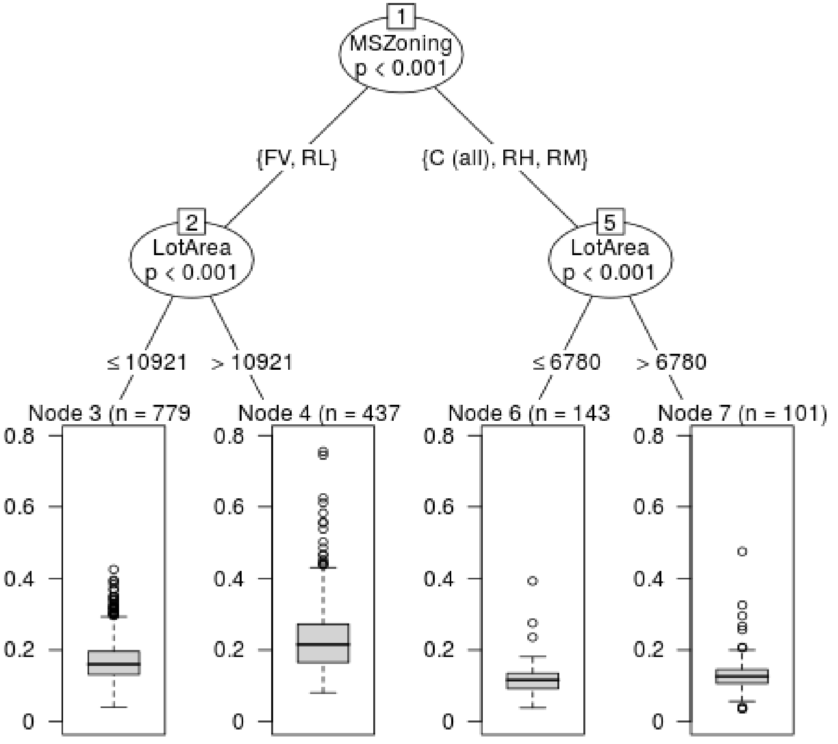
Estimate $p(\boldsymbol{x}_{\bar{S}} | \boldsymbol{x}_S = \boldsymbol{x}_S^*)$ properly

+

Monte Carlo integration

Our contribution: categorical/mixed variables

- Estimating $p(\boldsymbol{x}_{\bar{S}} | \boldsymbol{x}_S = \boldsymbol{x}_S^*)$
  - For every subset S, fit a **(multivariate) regression tree** to $\boldsymbol{y} = \boldsymbol{x}_{\bar{S}}$ based on $\boldsymbol{x}_S$ using the training data
  - Approximate $p(\boldsymbol{x}_{\bar{S}} | \boldsymbol{x}_S = \boldsymbol{x}_S^*)$ by the empirical distribution of the training observations ($\boldsymbol{x}_{\bar{S}}$) within the terminal node of $\boldsymbol{x}_S = \boldsymbol{x}_S^*$



How to use this is practice?

- All of this is implemented on our R-package ***shapr*** on GitHub (soon CRAN) github.com/NorskRegnesentral/shapr

- Paper (continuous variables): https://arxiv.org/abs/1903.10464
- Paper (categorical/mixed variables): https://arxiv.org/abs/2007.01027