Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

How to open the black box

Individual prediction explanation

Martin Jullum

with Kjersti Aas and Anders Løland

Det 20. norske statistikermøtet
Stavanger, 19. juni 2019

Example: Bank creates mortgage robot

Transaction history

Commercial Card - 4564801111111111: 08/04/2004 - 14/04/2004

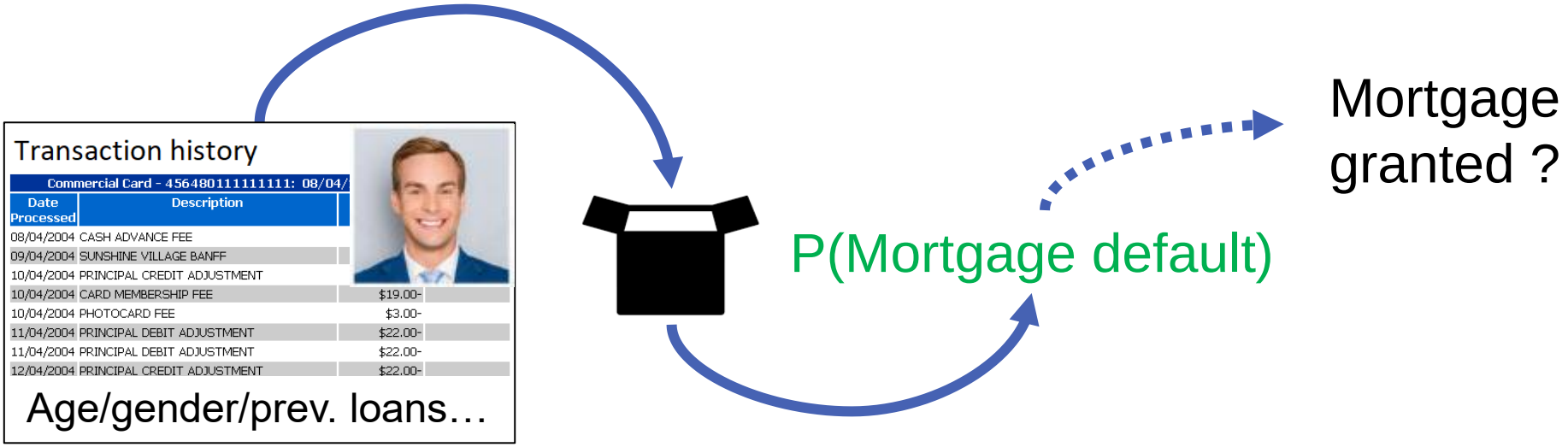
Date Processed	Description	Debit	Credit
08/04/2004	CASH ADVANCE FEE	$5.00-	
09/04/2004	SUNSHINE VILLAGE BANFF	$86.97-	
10/04/2004	PRINCIPAL CREDIT ADJUSTMENT	$22.00-	
10/04/2004	CARD MEMBERSHIP FEE	$19.00-	
10/04/2004	PHOTOCARD FEE	$3.00-	
11/04/2004	PRINCIPAL DEBIT ADJUSTMENT	$22.00-	
11/04/2004	PRINCIPAL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

Age/gender/prev. loans…

&

Defaulted loan?

Example: Bank creates mortgage robot



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

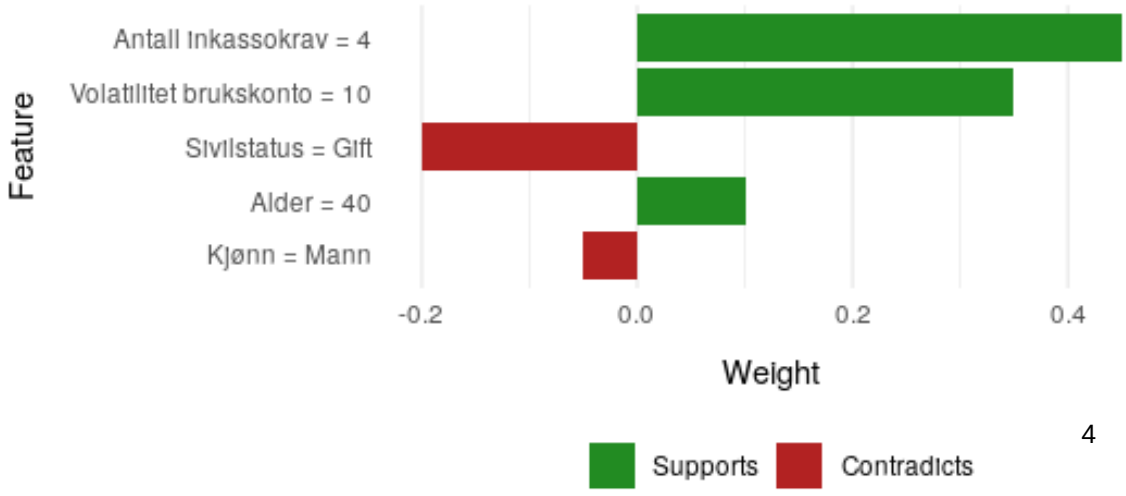
Why was  rejected a loan?

Individual prediction explanation

- **NOT** a general explanation of the black-box model
- $\boldsymbol{x}^*$: Transaction history/covariates for

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

- Which covariates "contributed the most" to increase/decrease the prediction to exactly $f(\boldsymbol{x}^*) = 70\%$?



Why is this important?

- Customers have a “right to an explanation”
- Also builds trust to the “robot”

Prediction explanation in general

- Assume we have trained a statistical or machine learning model to describe a response variable Y based on a set of covariates $\boldsymbol{x} = (x_1, \ldots, x_p), i.e$:

$$Y \approx f(\boldsymbol{x})$$

- f applied to predict Y for a new set of covariates $\boldsymbol{x} = \boldsymbol{x}^*$

- Want explain the prediction by translating $f(\boldsymbol{x}^*)$ to scores $\phi_1, \ldots, \phi_p$ representing the contribution of the covariates $\boldsymbol{x}^*$

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:

$$\phi_j = \sum_{S \subseteq M \setminus \{j\}} w(S)\left(v(S \cup \{j\}) - v(S)\right), \qquad w(S) = \frac{|S|!\,(|M| - |S| - 1)!}{|M|!}$$

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

- Several mathematical optimality properties

Shapley values for prediction explanation

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

$$f(\boldsymbol{x}^*) = \sum_{j=0}^{p} \phi_j \qquad \phi_0 = E[f(\boldsymbol{x})]$$

f indep. of $x_j \Rightarrow \phi_j = 0$, $\quad x_i, x_j$ same contribution $\Rightarrow \phi_i = \phi_j$

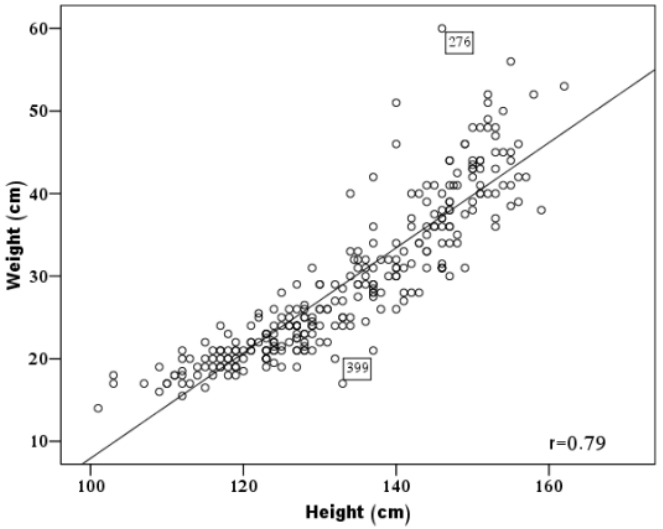
- Mathematically proven to be the only framework satisfying all of a series of such natural properties
- Rough interpretation of ϕ_j: **How does the prediction change when you don't know the value of** x_j

Linear models $f(\boldsymbol{x}) = \beta_0 + \sum_{j=1} \beta_j x_j$

- Linear model with independent covariates:

$$\phi_j = \beta_j\left(x_j^* - E[x_j]\right), \quad \phi_0 = \beta_0 + \sum_j \beta_j E[x_j]$$

- Explanation not simple with dependent covariates!
  - Example
    - $x_1 =$ height (cm)
    - $x_2 =$ weight (kg)
    - Y = PB in high jump (cm)
  - Model 1: $Y = 100 + 2x_1 - 2x_2$
  - Model 2: $Y = 100 - 2x_1 + 2x_2$



- Shapley values gives $\phi_1 \approx \phi_2$ in such a setting

Shapley values for prediction explanation

- 2 main challenges

1. The computational complexity in the Shapley formula

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

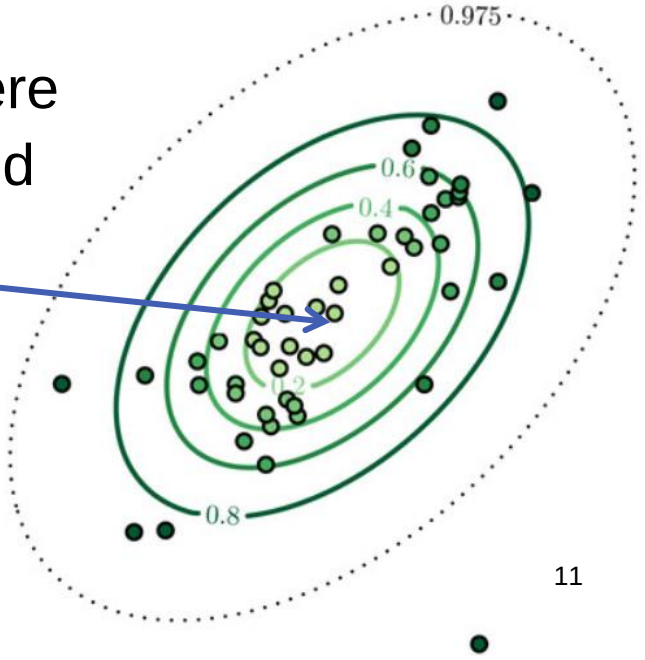
   - Partly solved by cleverly reducing the sum by subset sampling (Lundberg & Lee, 2017)

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

   - Existing methods essentially assumes $v(S) \approx \int f(\boldsymbol{x}_{\bar{S}}, \boldsymbol{x}_S) p(\boldsymbol{x}_{\bar{S}}) \mathrm{d}\boldsymbol{x}_{\bar{S}}$, and uses Monte Carlo integration
   - This assumes covariates are independent!

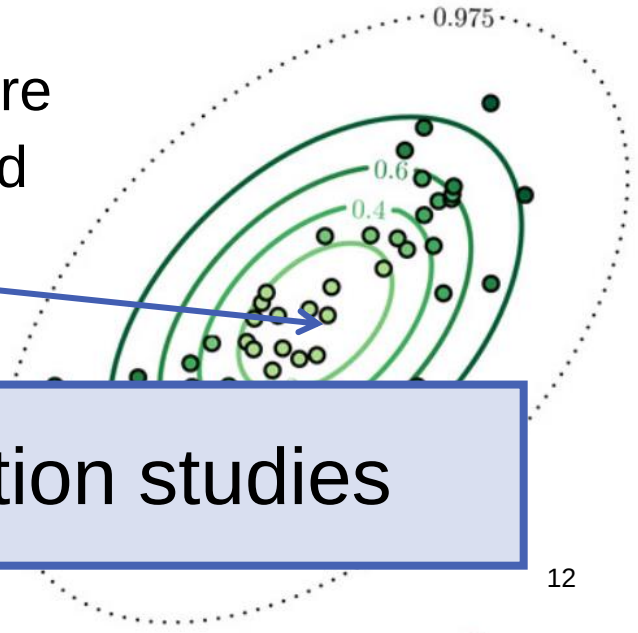
Our main contribution

- Working with continuous covariates
- Estimate $p(\boldsymbol{x}_{\bar{S}}|\boldsymbol{x}_S = \boldsymbol{x}_S^*)$ properly + Monte Carlo integration
- 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
  - A empirical (conditional) approach where training observations at $\boldsymbol{x}_{\bar{S}}^k$ are weighted by proximity of $\boldsymbol{x}_S^k$ to $\boldsymbol{x}_S^*$



Our main contribution

- Working with continuous covariates
- Estimate $p(\boldsymbol{x}_{\bar{S}}|\boldsymbol{x}_S = \boldsymbol{x}_S^*)$ properly + Monte Carlo integration
- 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
  - A empirical (conditional) approach where training observations at $\boldsymbol{x}_{\bar{S}}^k$ are weighted by proximity of $\boldsymbol{x}_S^k$ to $\boldsymbol{x}_S^*$

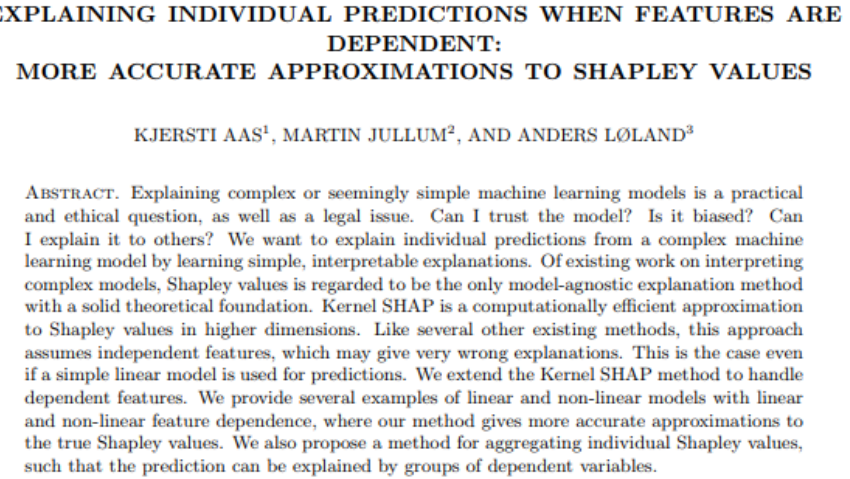


BIG improvements in simulation studies

Want to know more?

- Read our paper on arXiv

  arxiv.org/abs/1903.10464



EXPLAINING INDIVIDUAL PREDICTIONS WHEN FEATURES ARE DEPENDENT:
MORE ACCURATE APPROXIMATIONS TO SHAPLEY VALUES

KJERSTI AAS[1], MARTIN JULLUM[2], AND ANDERS LØLAND[3]

ABSTRACT. Explaining complex or seemingly simple machine learning models is a practical and ethical question, as well as a legal issue. Can I trust the model? Is it biased? Can I explain it to others? We want to explain individual predictions from a complex machine learning model by learning simple, interpretable explanations. Of existing work on interpreting complex models, Shapley values is regarded to be the only model-agnostic explanation method with a solid theoretical foundation. Kernel SHAP is a computationally efficient approximation to Shapley values in higher dimensions. Like several other existing methods, this approach assumes independent features, which may give very wrong explanations. This is the case even if a simple linear model is used for predictions. We extend the Kernel SHAP method to handle dependent features. We provide several examples of linear and non-linear models with linear and non-linear feature dependence, where our method gives more accurate approximations to the true Shapley values. We also propose a method for aggregating individual Shapley values, such that the prediction can be explained by groups of dependent variables.

- Check out our R-package

  github.com/NorskRegnesentral/shapr

- Talk to me!