

Personlig prediksjonsforklaring

Kort introduksjon

Martin Jullum

med Kjersti Aas, Anders Løland og Nikolai Sellereite

Styremøte Big Insight, 5. desember 2018

Eksempel: Bank lager boliglånsrobot



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...

&

Misligholdt
lån?

Eksempel: Bank lager boliglånsrobot

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...



$P(\text{misligholde lån})$

Lån
innvilget?



Personlig prediksjonsforklaring

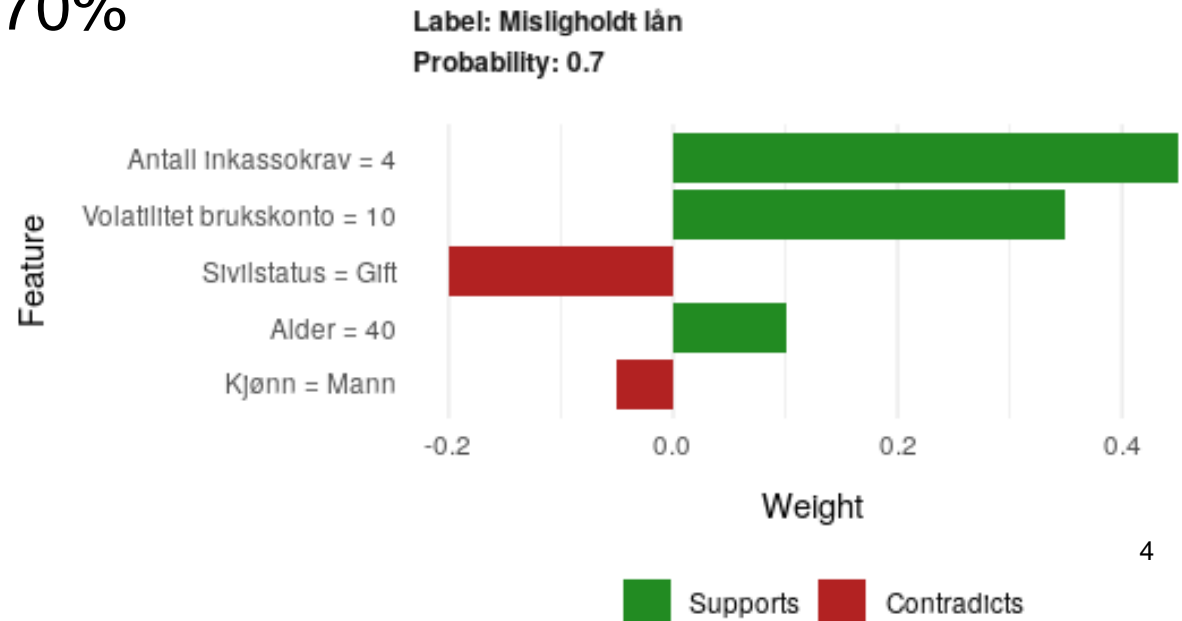
- ▶ IKKE en generell forklaring av black-box modellen

- ▶ x^* : Transaksjonshistorikken/kovariatene til



Forklaring av $f(x^*) = 70\%$

- ▶ Hvilke kovariater “bidro mest” til å trekke sannsynligheten opp/ned til $f(x^*) = 70\%$



Kan kreve forklaring

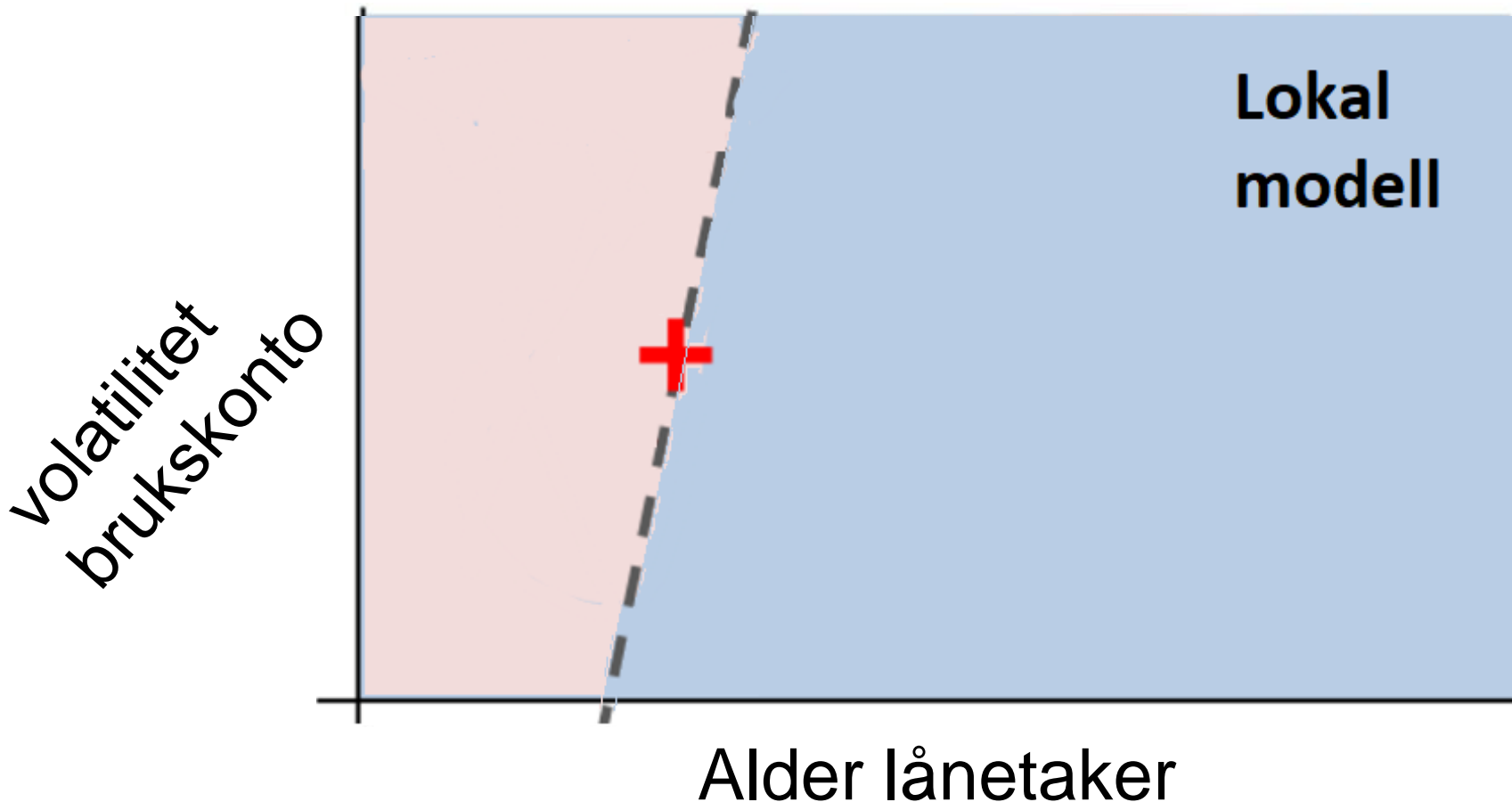


The General Data Protection Regulation

Metode 1: LIME

(Local Interpretable Model-agnostic Explanation)

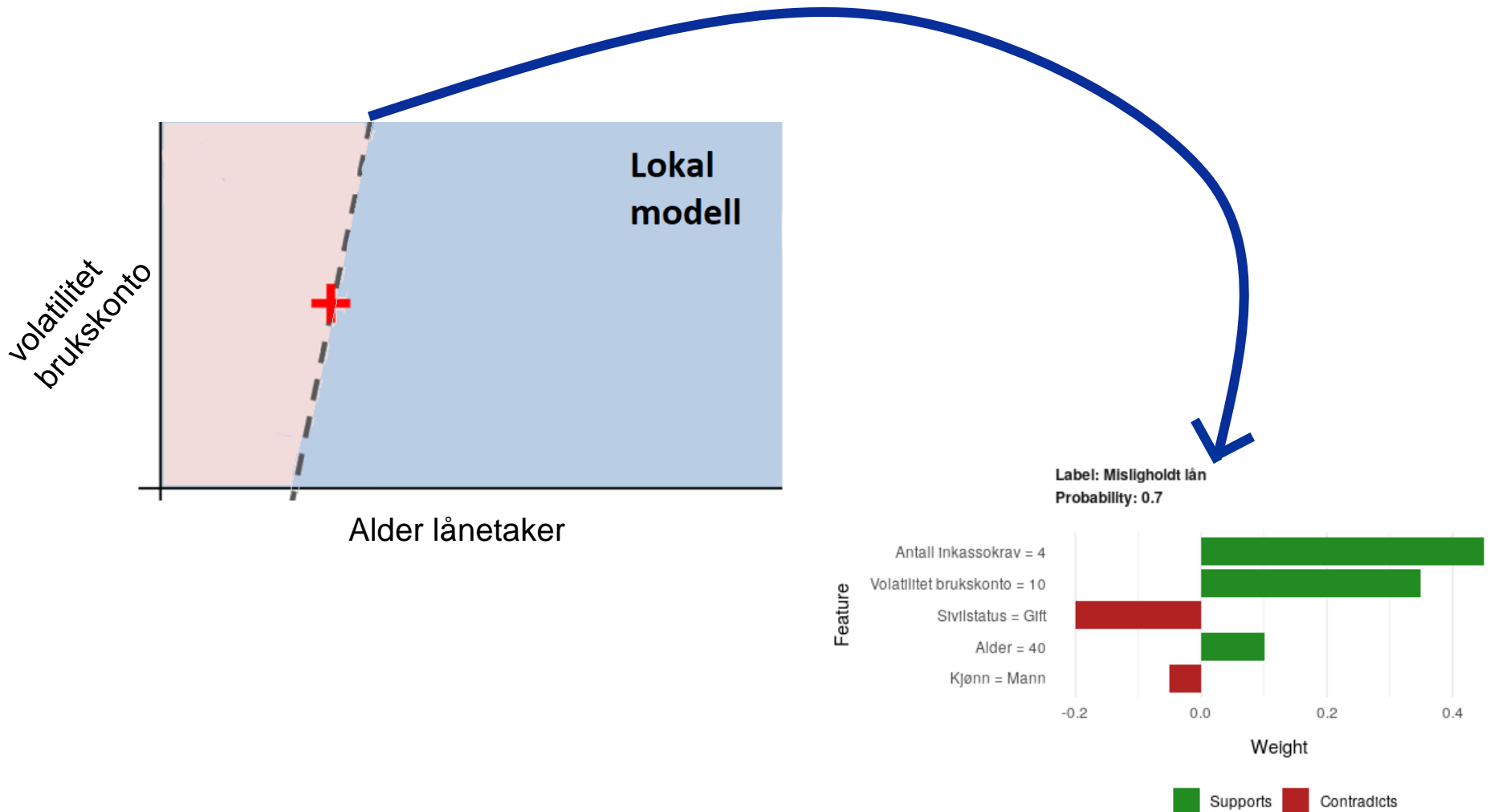
- ▶ Lager en forenklet modell rundt prediksjonen



Metode 1: LIME

(Local Interpretable Model-agnostic Explanation)

- ▶ Lager en forenklet modell rundt prediksjonen



Metode 2: SHAP

(SHapley Additive exPlanations)

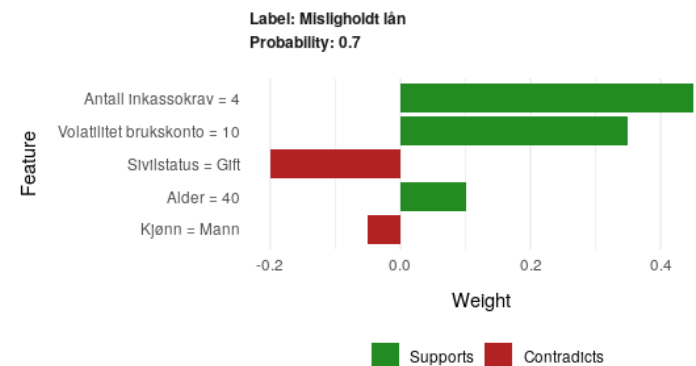
► Shapley verdier

- Konseptet stammer fra (lag-)spillteori



► SHAP

- Spillerne = kovariatene (x_1, \dots, x_p)
- Utbetaling = prediksjonen $f(\mathbf{x})$
- Eksplisitt formel for bidrag fra hver kovariat




Forklaring med LIME vs SHAP

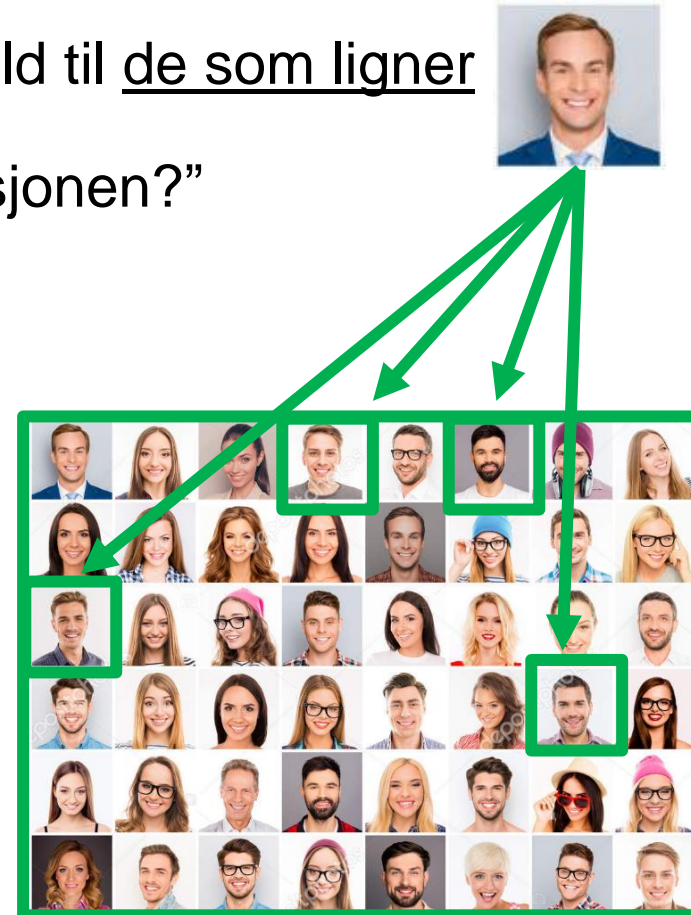
- ▶ Svarer på ulike spørsmål

- ▶ LIME

- Forklarer prediksjon for  i forhold til de som ligner 
- Spm: “Hvordan enklest endre prediksjonen?”

- ▶ SHAP

- Forklarer prediksjon for  i forhold til hele populasjonen
- Spm: “Hvilke variable er viktigst å observere for prediksjonen?”



Vårt bidrag

- ▶ LIME og SHAP ignorerer avhengighet mellom kovariater
 - Kan gi helt feil forklaring hvis sterk avhengighet



Vårt bidrag

- ▶ LIME og SHAP ignorerer avhengighet mellom kovariater
 - Kan gi helt feil forklaring hvis sterk avhengighet
- ▶ Vårt arbeid: **Reparere SHAP**
 - SHAP antar $p(\mathbf{x}_A|\mathbf{x}_B) \approx p(\mathbf{x}_A)$ som en del av algoritmen
 - Vi forsøker å estimere $p(\mathbf{x}_A|\mathbf{x}_B)$ skikkelig
- ▶ Arbeider med både artikkel og programvare for dette
- ▶ Flere interessante relaterte problemstillinger