Et smartere folkeregister

Historien om et sommerprosjekt
hos Skatteetaten

BEKK

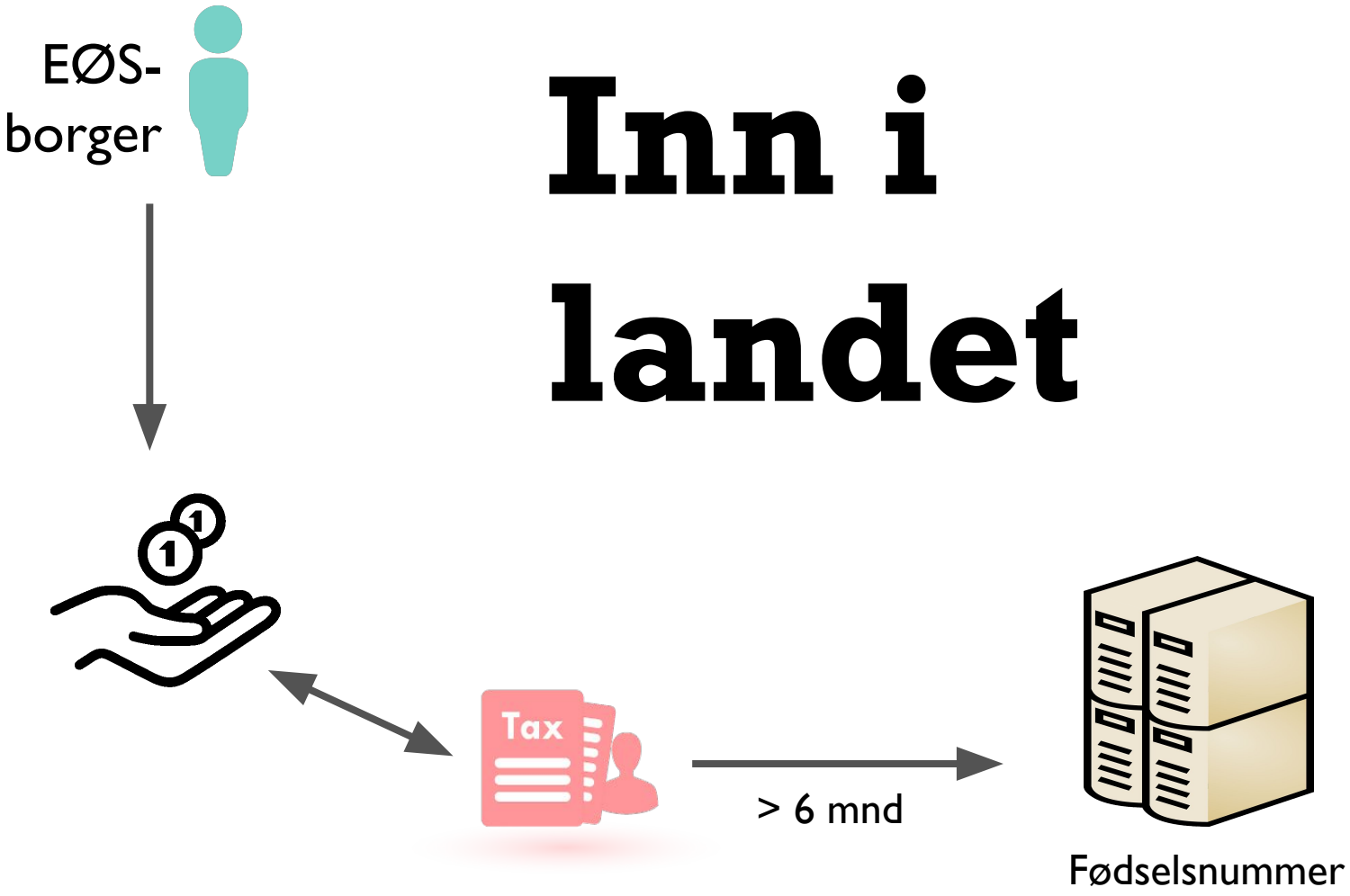
Nina Kjekstad

Sommerstudent i BEKK 2017
Starter som konsulent høsten 2018

Trygve Bertelsen Wiig

Konsulent i BEKK

Inn i landet



EØS-
borger

Ut av landet

???

Folkeregisteret

Offentlig register over alle med tilknytning til Norge, underlagt Skatteetaten.

Brukes bl.a. til:

- Skattlegging
- Statistikk
- Kontaktinformasjon for offentlige etater og enkelte private instanser

Alle som er bosatt i Norge lenger enn 6 mnd skal være registrert med fødselsnr. i Folkeregisteret.

Feil i Folkeregisteret

=

Feil i det meste som finnes av norsk statistikk

Kommuner og etater mister evnen til å korrekt fordele ressurser og lage korrekte prosjekteringer...

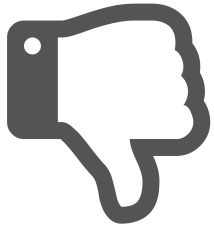
...og ikke minst blir SSB misfornøyde.

Ikke et nytt problem...

«...løpende folkeregistrering [ble] først innført som et (frivillig) kommunalt ansvar i 1905. Disse registrene var ofte unøyaktige, siden det ble registrert når personer flyttet inn, men i liten grad når folk flyttet ut...»

(Wikipedia)

Dagens løsning…



Tunge manuelle kontroller

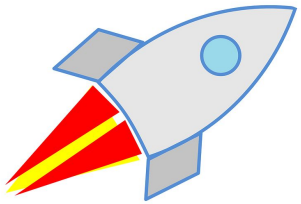
Fanger antakelig ikke opp alle

Vår løsning…

Automatisk deteksjon vha. maskinlæring

Modellen tilpasser seg kontinuerlig nye trender

Mål



Proof-of-concept som viser at det har noe for seg å prøve å bruke maskinlæring til å identifisere utflyttede personer

Metode

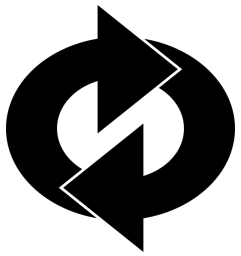
Tre sommerstudenter, to faddere, en entusiastisk kunde og seks uker

Litt mer konkret…

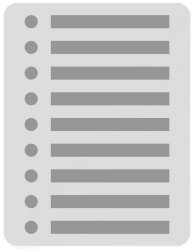
Skattegrunnlag 2013 og 2015

+

Data fra Folkeregisteret



Maskinlæring



Rangert liste over personer til nærmere kontroll

Skattegrunnlaget

- Inntekt
- Gjeld
- Ulike fradrag

(...)

Folkeregisteret

- Adresse
- Statsborgerskap
- Alder
- Familie
- Sivilstand

(...)

Totalt rundt 200 attributter

Dataprosessering

FØDSELSNR	FØDELAND	FAMILIENR	ADRESSE
12031212345	Norge	10099067890	Fredrik Selmers vei 4
20099067890	Polen	22031212345	Slottsplassen 1

- Naturlig data for et menneske
- Krever prosessering for å kunne bruke det som input til en maskinlæringsmetode

FØDSELSNR	FØDELAND
12031212345	Norge
20099067890	Polen
29129412345	Sverige

—>

FØDSELSNR	FØDELAND NORGE	FØDELAND NORDEN	FØDELAND EØS
12031212345	1	1	1
20099067890	0	0	1
29129412345	0	1	1

FØDSELSNR	FAMILIENR	ADRESSE
12031212345	10099067890	Slottsplassen 1
20099067890	22031212345	Slottsplassen 1

—>

FØDSELSNR	FAMILIER_SAMME_ADRESSE
12031212345	2
20099067890	2

FØDSELSNR	ADRESSE
12031212345	UTVANDRET
20099067890	UTV
29129412345	IKKE BOSATT
22031212345	UTVANDR
10099067890	UTFLYTTET
:	:

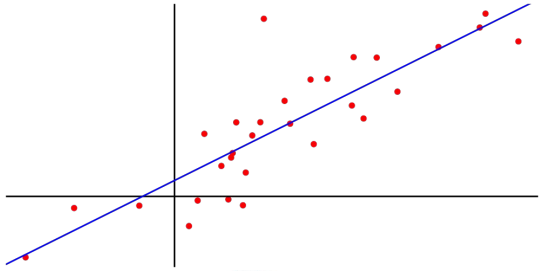
—>

FØDSELSNR	ADRESSE
12031212345	UTVANDRET
20099067890	UTVANDRET
29129412345	UTVANDRET
22031212345	UTVANDRET
10099067890	UTVANDRET
:	:

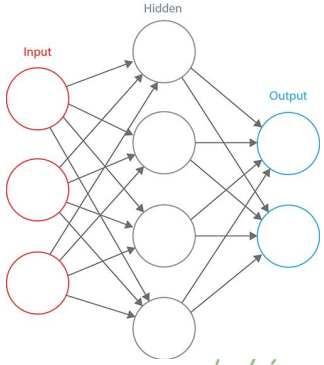
FØDSELSNR	FØDELAND	FAMILIENR	ADRESSE
CA0780018CD3A7E587FCBB109DD27D58734F2D5E	Norge	10099067890	Fredrik Selmers vei 4
64531FEC7A5FC41797B81276B396EFB926C8199E	Polen	22031212345	Slottsplassen 1

FØDSELSNR	FØDELAND NORGE	FØDELAND NORDEN	FØDELAND EØS	FAMILIE SAMME ADRESSE	UTFLYTTET
CA0780018CD3A7E587FCBB109DD27D58734F2D5E	1	1	1	2	0
64531FEC7A5FC41797B81276B396EFB926C8199E	0	0	1	2	1
LKP45FEC7A5FOP5797B81276B396EFBNK6C8127B	0	1	1	1	0

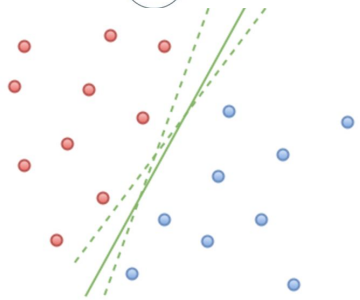
Valg av modell



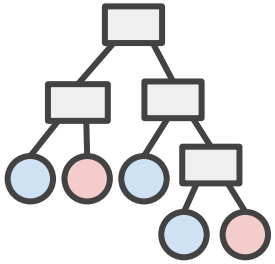
Logistisk regresjon



Nevrale nettverk



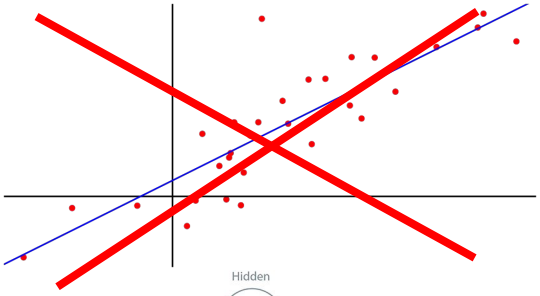
Support Vector Machine (SVM)



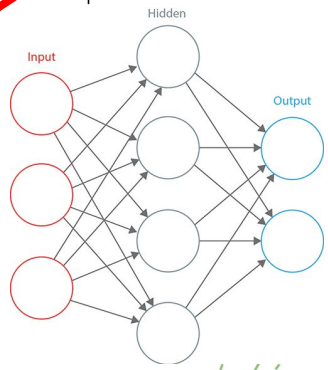
Random Forest

Valg av modell

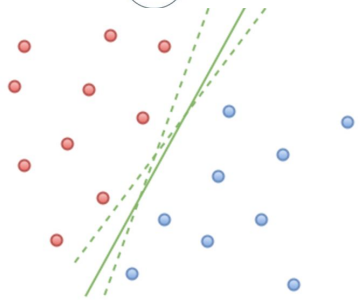
- Gode prediksjoner



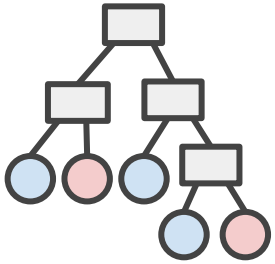
Logistisk regresjon



Nevrale nettverk



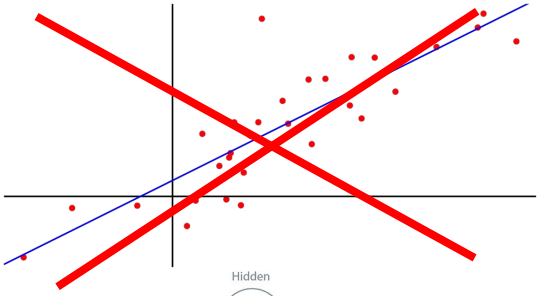
Support Vector Machine (SVM)



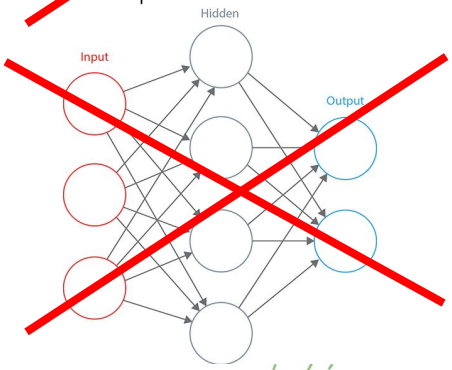
Random Forest

Valg av modell

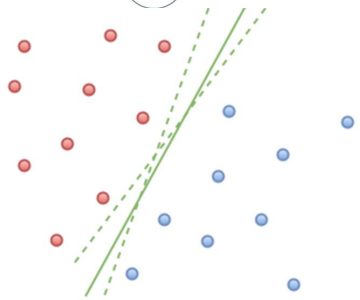
- Gode prediksjoner
- Enkel å jobbe med
- Lett å tolke



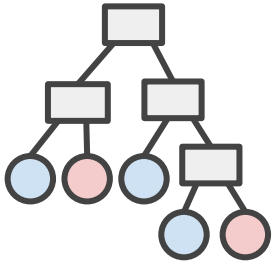
Logistisk regresjon



Nevrale nettverk



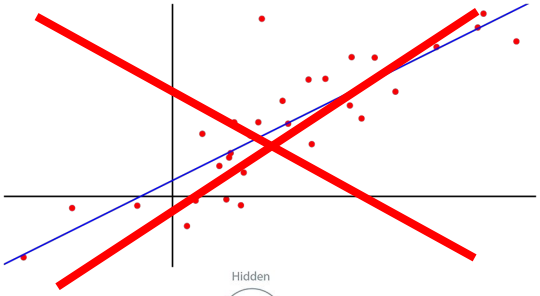
Support Vector Machine (SVM)



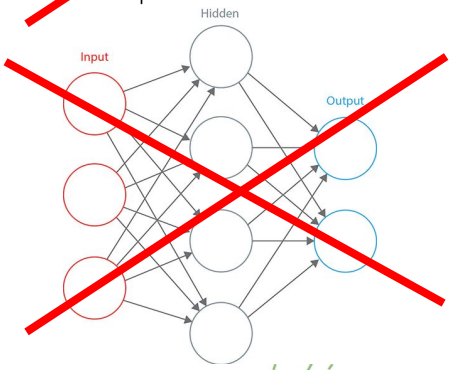
Random Forest

Valg av modell

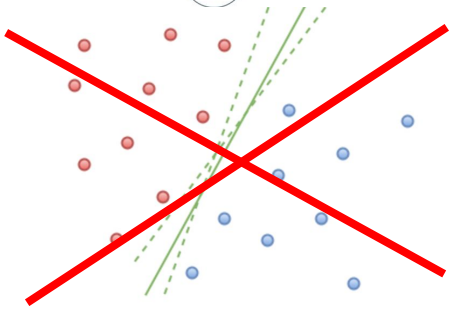
- Gode prediksjoner
- Enkel å jobbe med
- Lett å tolke
- Rask å trene



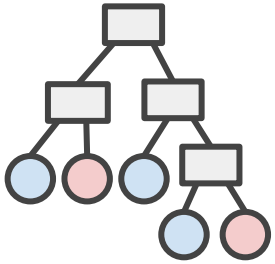
Logistisk regresjon



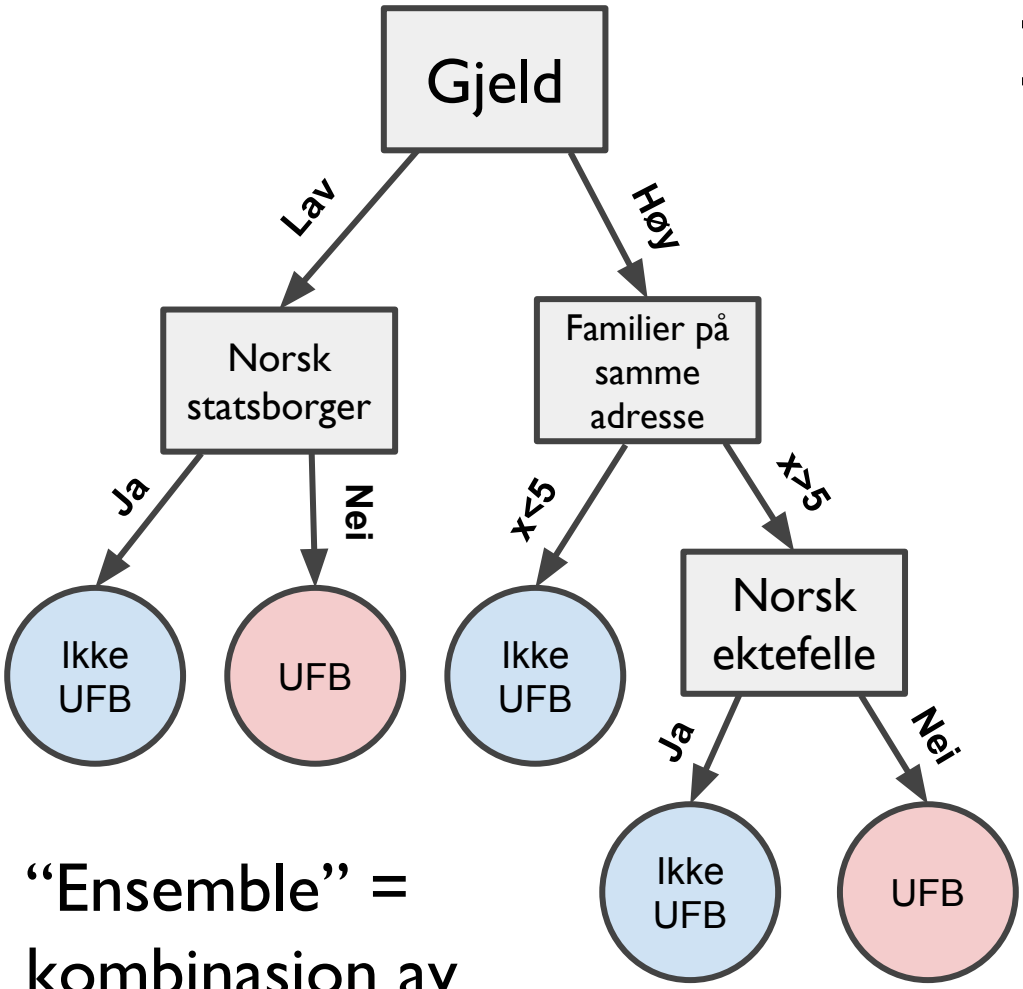
Nevrale nettverk



Support Vector Machine (SVM)



Random Forest

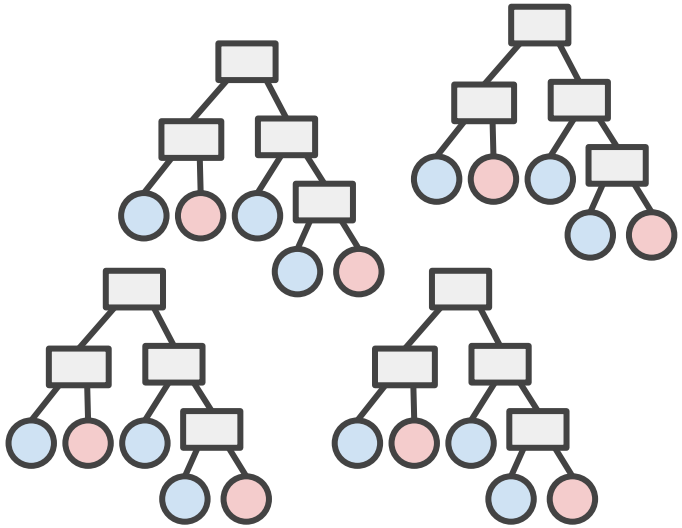


Random forest

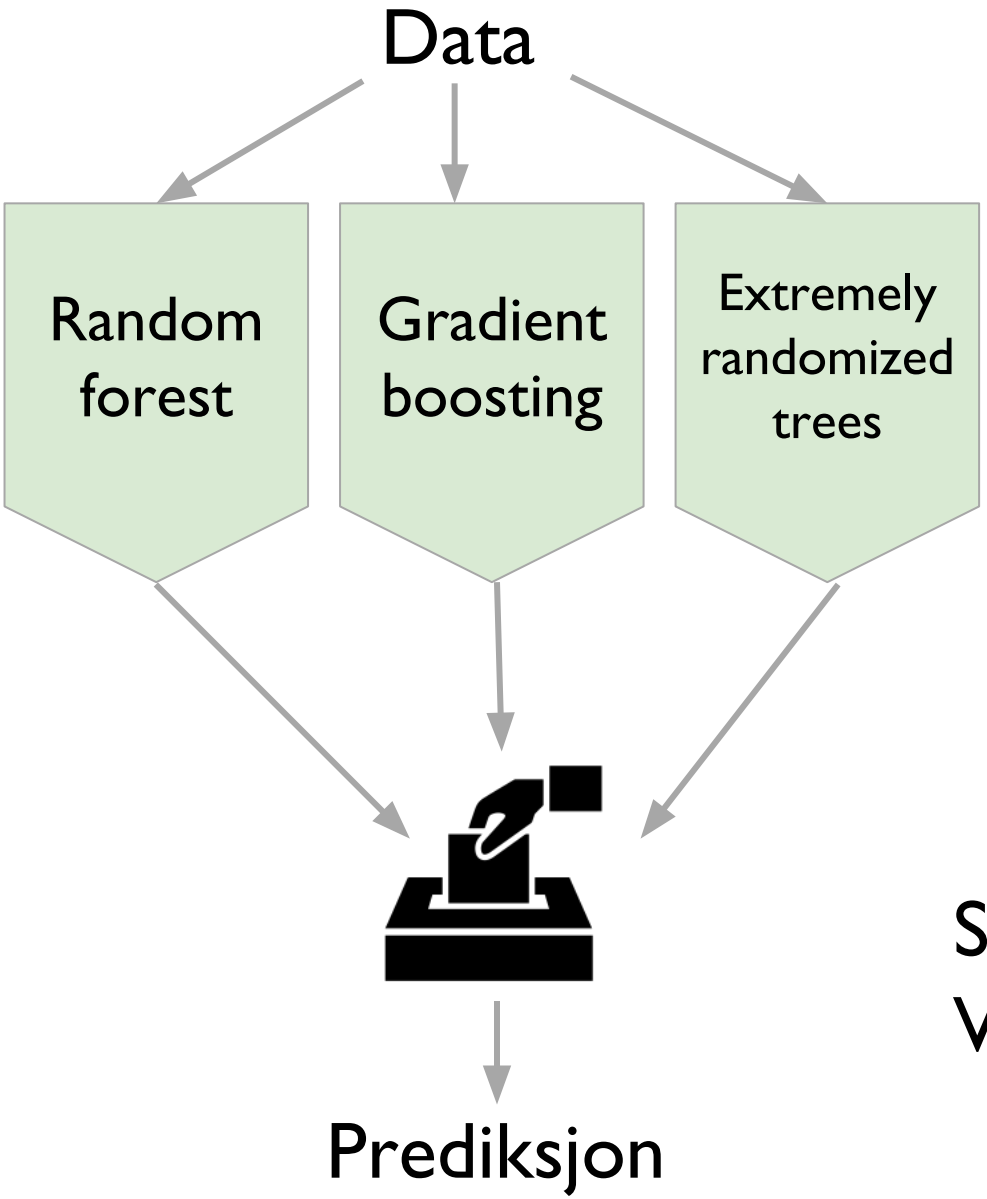
Mange valgtrær basert på ulike attributter

GJENNOMSNITT

Én veldig god klassifikator



"Ensemble" = kombinasjon av flere modeller



Majority voting

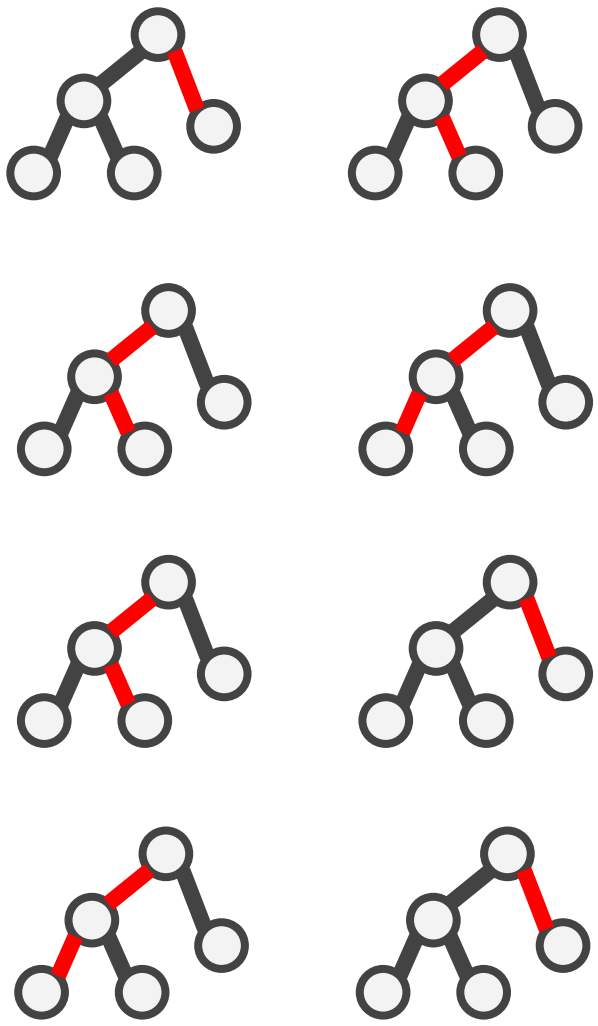
Mange gode klassifikatorer basert på all treningsdataen

↓ *AVSTEMNING* ↓

Én *litt* bedre klassifikator

Stående spørsmål:
Verdt kompleksiteten?

Tolking av prediksjoner



For hvert tre i en Random Forest:
Kan se hvilke variabler som påvirker en prediksjon mest og i flest trær

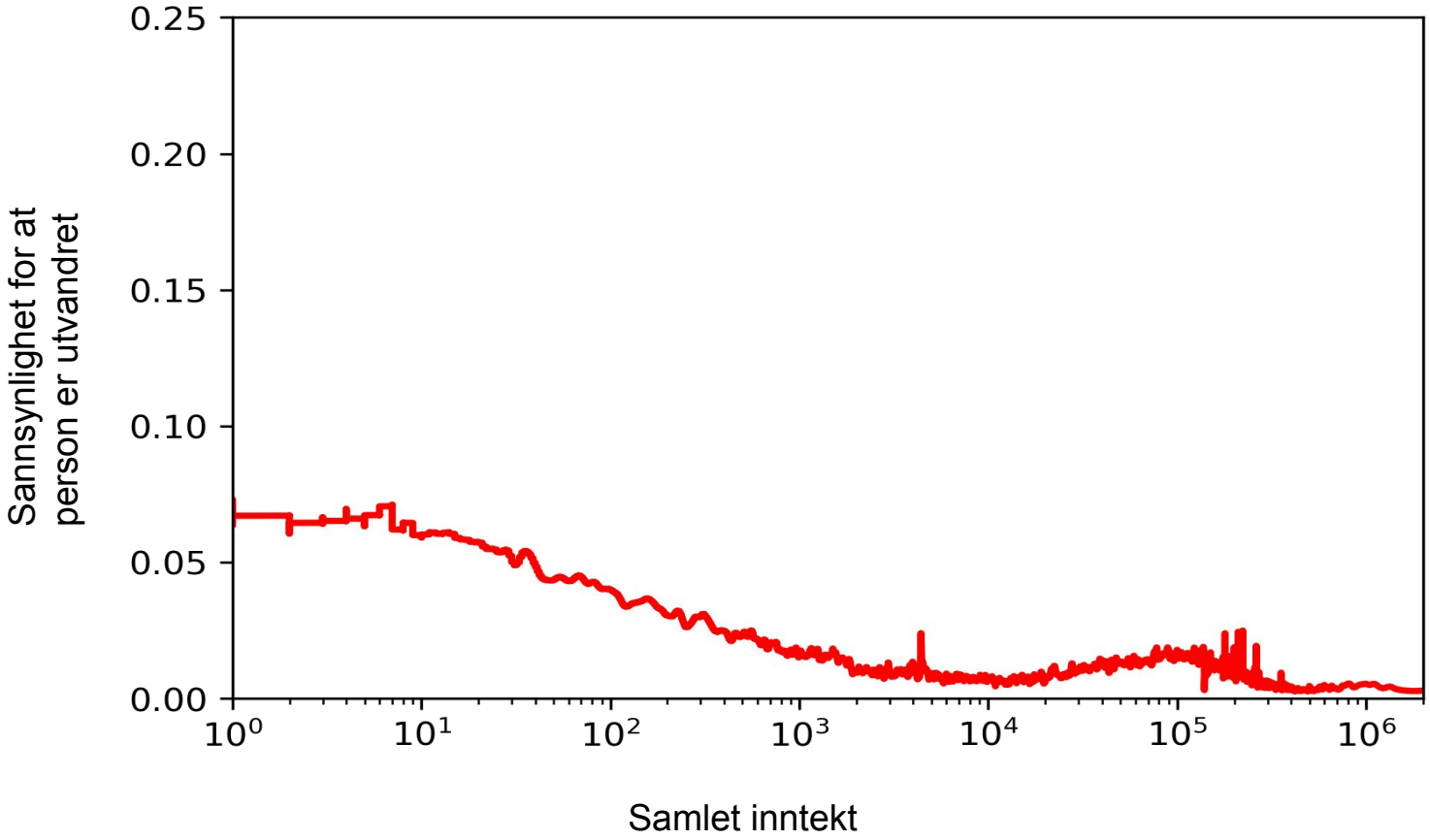
Tar gjennomsnitt over alle trærne av hvor mye hver variabel påvirker prediksjonen

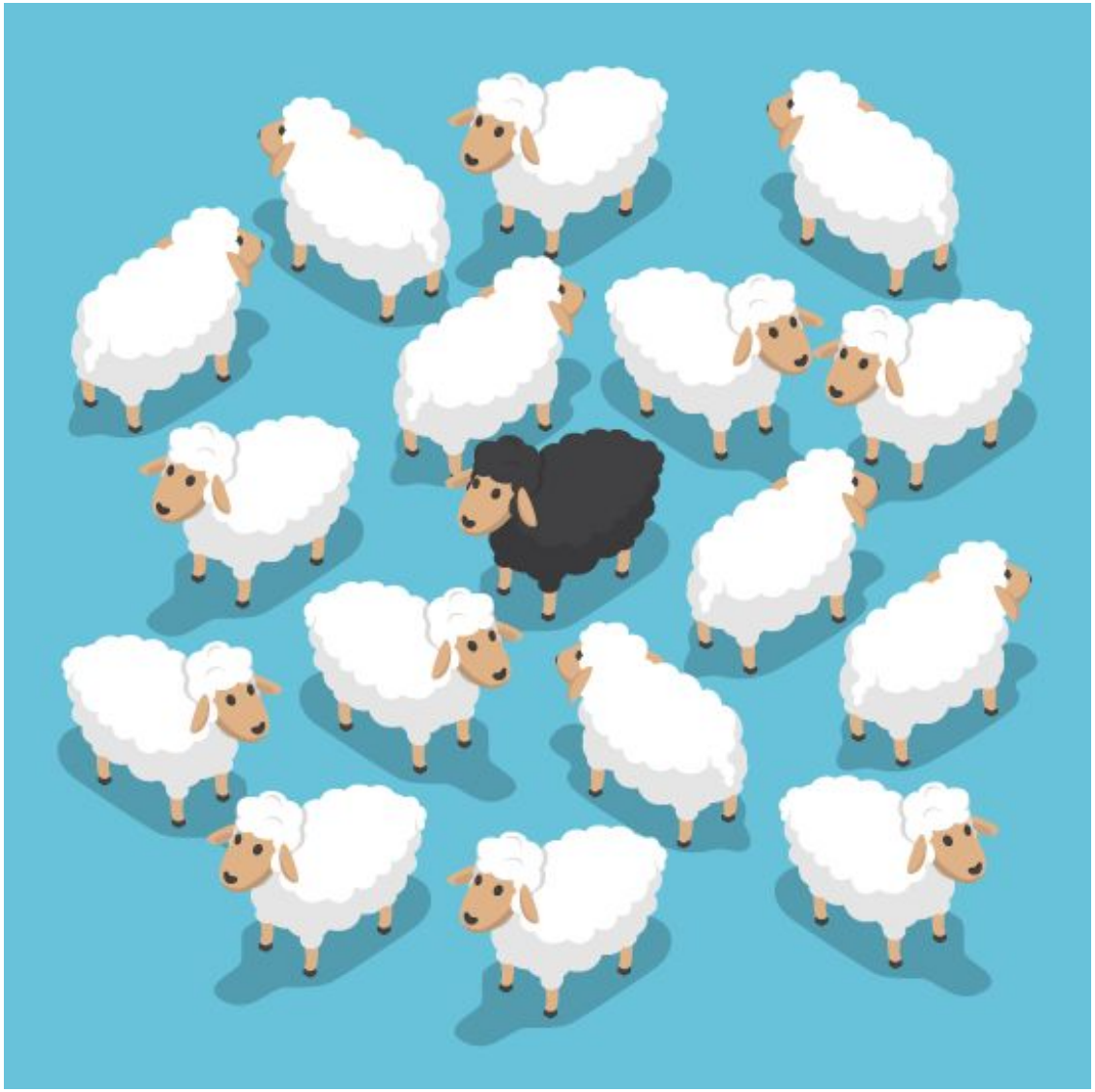
Variabel	Vekt
1. Øvrig formue	5%
2. Fødeland Litauen	4%
3. Antall personer på adresse	3%
4. Kapitalinntekt	3%

Ikke alle variabler er like viktige...

Attributt	Vekt
144. Formue fritidsbolig	0,0001%
162. Særfradrag uførhet	0,00003%
170. Inntekt jordbruk	0,00001%
172. Inntekt barnepass	0%
179. Inntekt reindrift	0%
183. Lønn barn under 13 år	0%
184. Tillegg primærnæring	0%
189. Sum arbeidsgiverfradrag	0%

P(Utflyttet | Attributt = x)





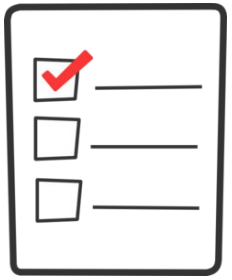
Utfordring:

Ubalansert datasett

<0,5% utflyttet
>99,5% ikke utfl.

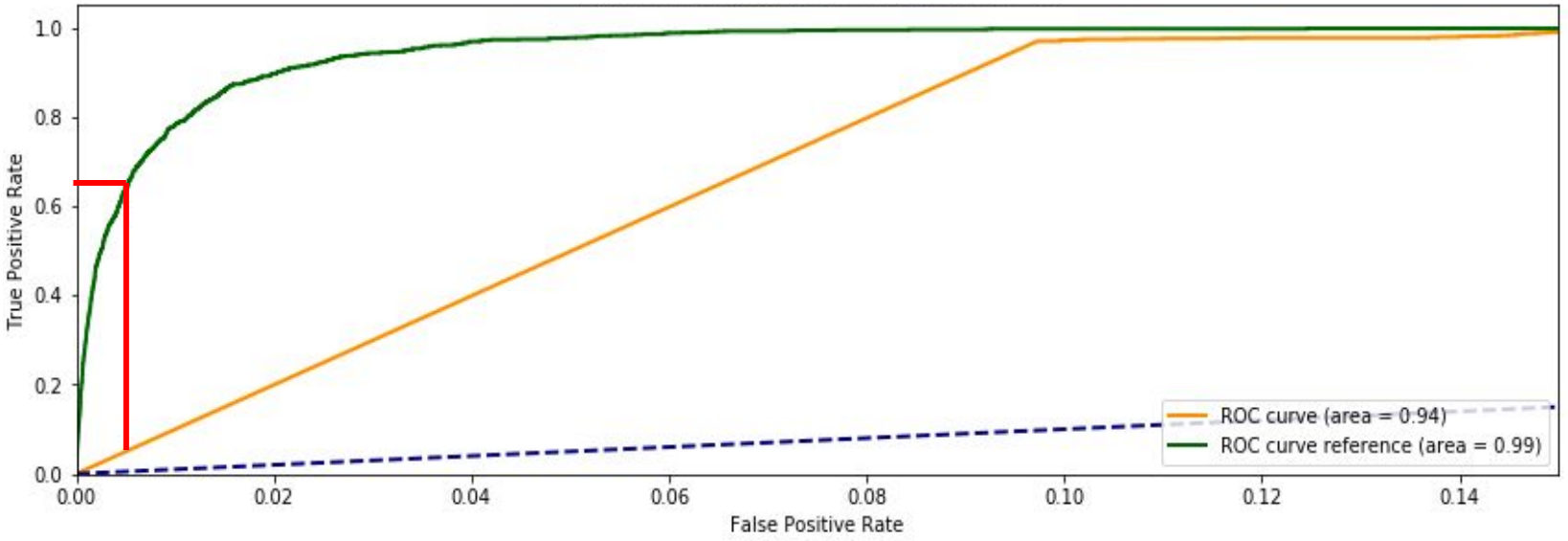
Resultat:

Sannsynligheten for at en person ikke befinner seg i landet



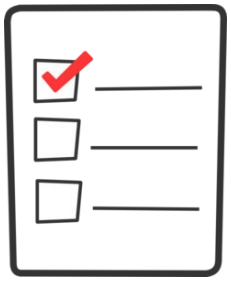
En rangert liste med personer som har minst X % sannsynlighet for å ha flyttet ut av landet

Kan kalibrere grensen for kandidater til utflytting etter ønsket mengde false/true positives



50% cutoff:

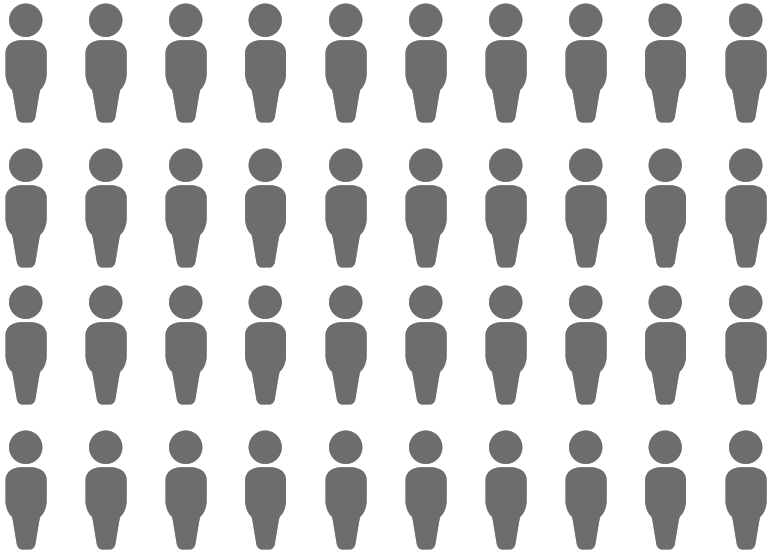
Liste med
~23 000 navn



Modellen identifiserte 67,5% av de som er blitt utflyttet (sanne positiver)

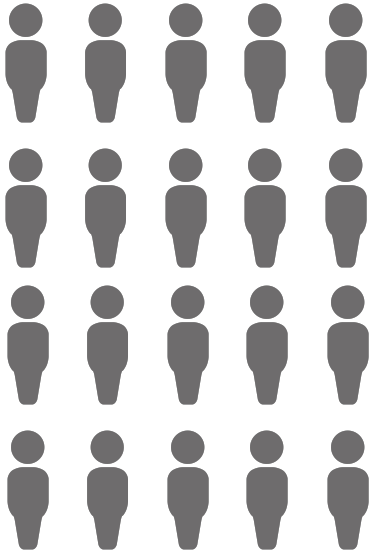
Utflyttede personer

Ikke utflyttede personer

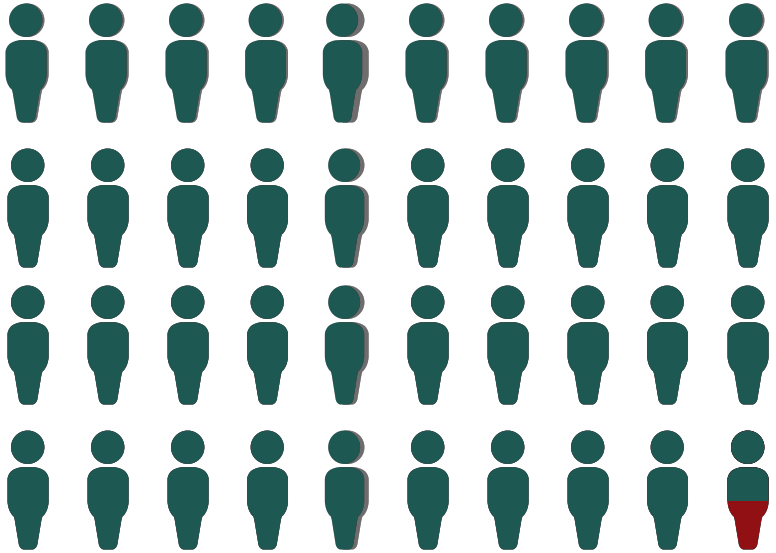


Modellen identifiserte 99,4% av de som ikke er blitt utflyttet (sanne negativer)

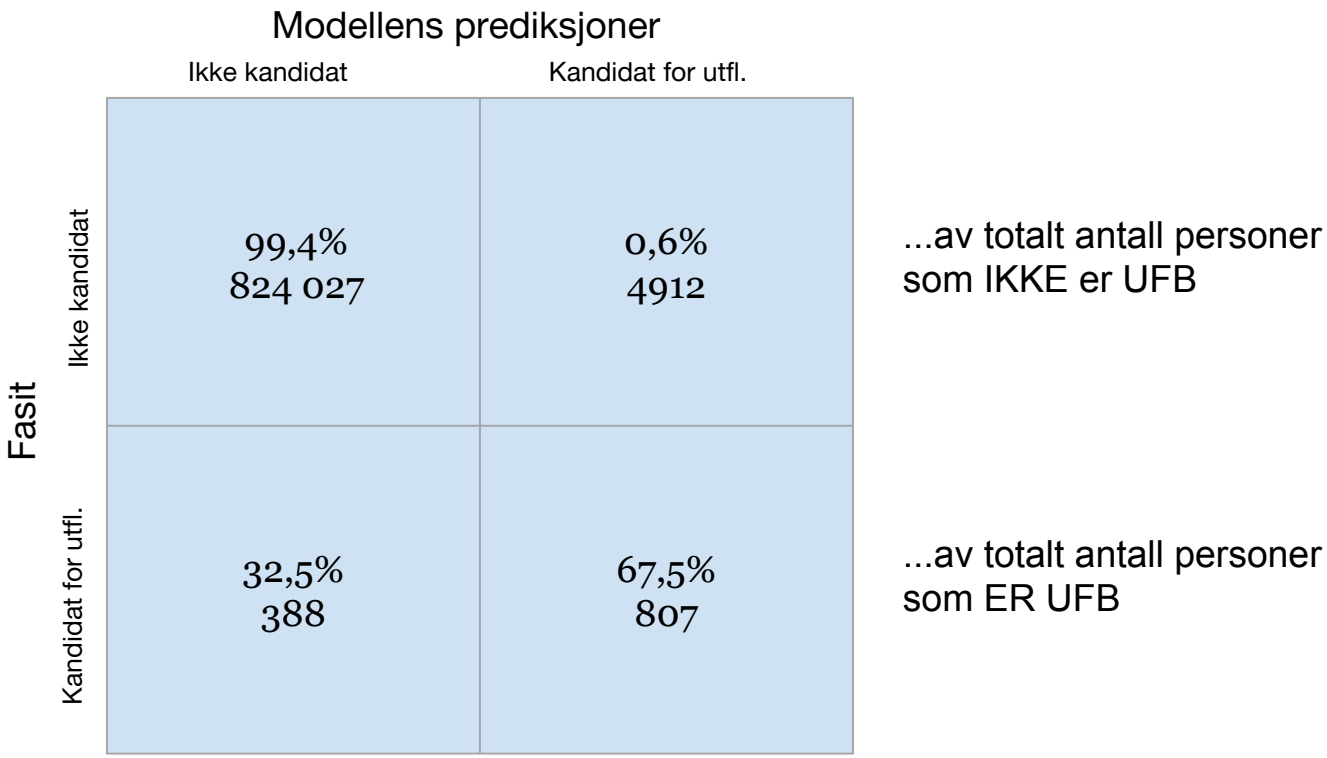
Utflyttede personer



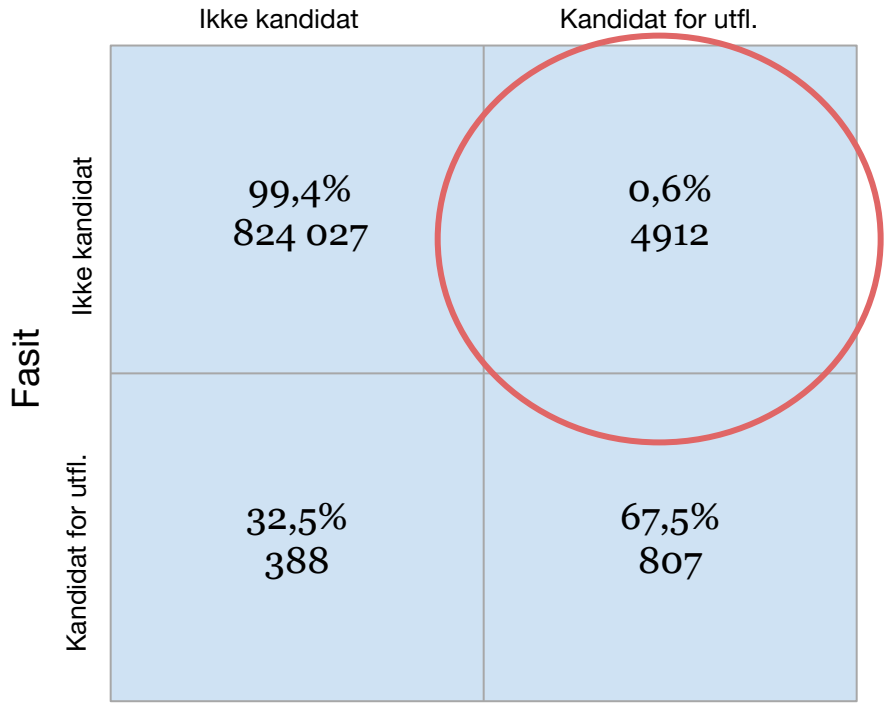
Ikke utflyttede personer



Confusion matrix



Absolutte tall er for testsettet



Fant vi noen nye kandidater for å ha flyttet ut uten å ha meldt fra?

Ja!

- Gift med en kandidat
- Ingen inntekt de siste tre årene

Post-mortem

Gode resultater, men…

Skatteetaten bruker i dag flere datakilder enn vi brukte til å finne kandidater som de tror kan ha forlatt landet.

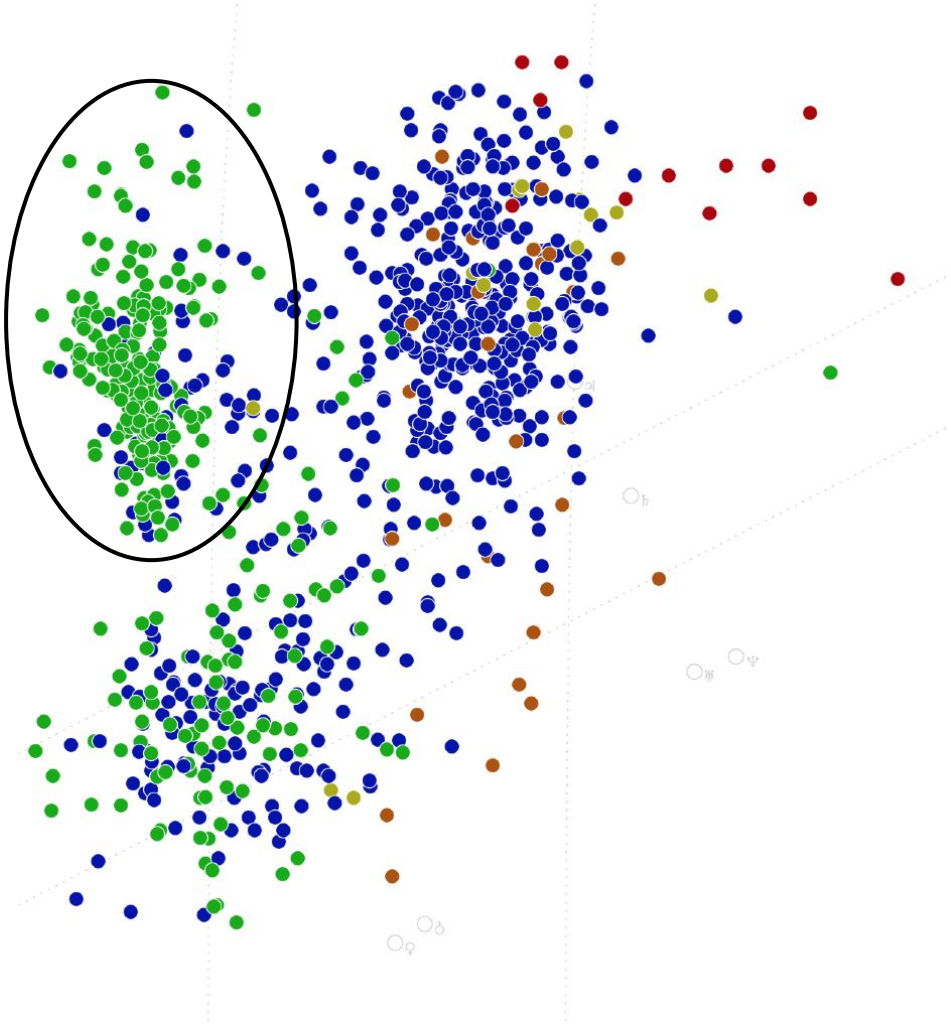
Vi kunne for eksempel brukt...

- Eiendomsregistere
- Flyttemeldinger
- Lønnsmeldinger
- Pensjonsinformasjon
- Foreldrepenger
- Enhets- og foretaksregisterne
- (...)

Forbedringer i modellering

Antok at klassene "ikke utflyttet" og "utflyttet" i fasiten var rent separert. Tok ikke hensyn (på modellnivå) til at ikke alle ble fanget opp av manuelle metoder!

Mulig løsning:
One-class learning med SVM-er



Takk for oss!