Experiments with Recommending Financial News

RecSysNL meetup

19 November 2019

Anca Dumitrache

@anca_dmtrch

Team:

Feng

Anca

David

Bahadir

Dung

Maya

Li’ao

Philippe

Kimberly

Klaus

Oberon

Manon

Azamat

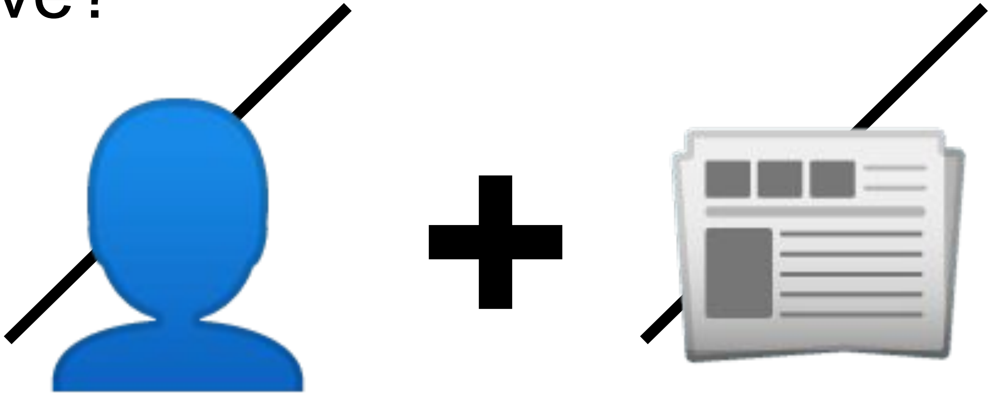
Domain: Het Financieele Dagblad (FD) is a daily Dutch newspaper focused on business & financial news

Goal: Personalized article recommendations for FD readers

Requirements: Recommended articles have to be recently published (cold start problem)

Context: Google DNI project on news personalization

What data do we have?



Anca Dumitrache
@anca_dmtrch

Working with implicit feedback


fd. Mijn nieuws Laatste nieuws Krant Dossiers Beurs Meer DOWJ 26.717,43 -- Abonneren

Job Woudt za 29 jun Tekst Krant

TELECOM

112-crisis in Nederland: Pas op voor digitale hypochondrie

Een softwarefout en het ganse land ligt plat. Toch is er alle reden om na de 112-crisis van afgelopen maandag niet meteen in paniek te raken

Wie een nieuwtje wil doorbellen naar de Telegraaf-tiplijn strandt vrijdag op de mededeling dat het nummer niet bereikbaar is. De lijn ligt eruit, nadat de telefoon eerder deze week roodgloeiend stond. Zeker vierhonderd meldingen kwamen maandagmiddag binnen. Over onwelwordingen, branden, een inbraakpoging en verkeersongelukken. Ze waren alleen niet bedoeld voor de krant.

Paniek in de tent? De overheid had een bericht het land in gestuurd. Daarin adviseerde ze de burgers 0613650952 te bellen als één van de alternatieven voor het dan onbereikbare 112. Het 06-nummer was dus van de krant.

Omdat 112 niet bereikbaar was stuurde de overheid een alert uit met een alternatief telefoonnummer. Foto: Rob

Routeringsplatform ontregeld

Volgen via mijn nieuws

Ramp + Volg

Software + Volg

Telecom + Volg

Laatste nieuws

14:00 Waterstof moet Rotterdamse industrie vergroenen

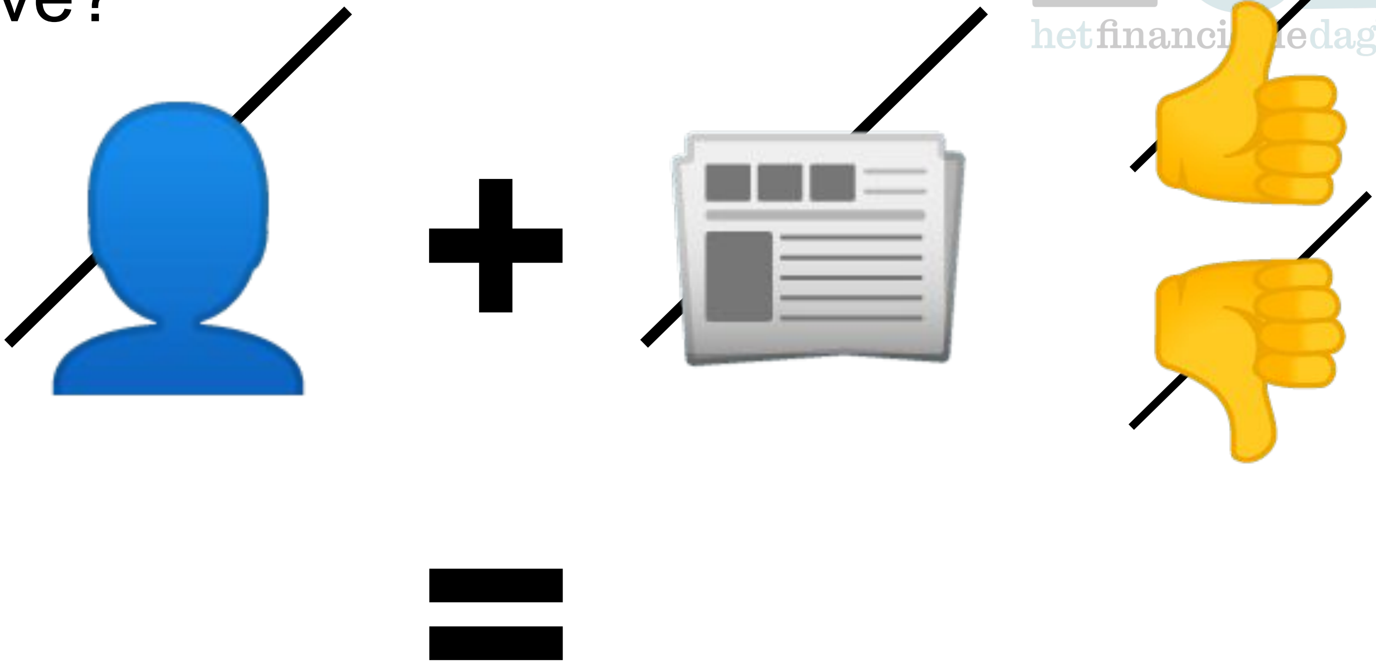
13:48 Mollie mikt met vers groeigeld op Europese klanten

13:47 ASML nieuwe partner Van Gogh Museum

Articles seen but not clicked

Anca Dumitrache
@anca_dmtrch

What data do we have?

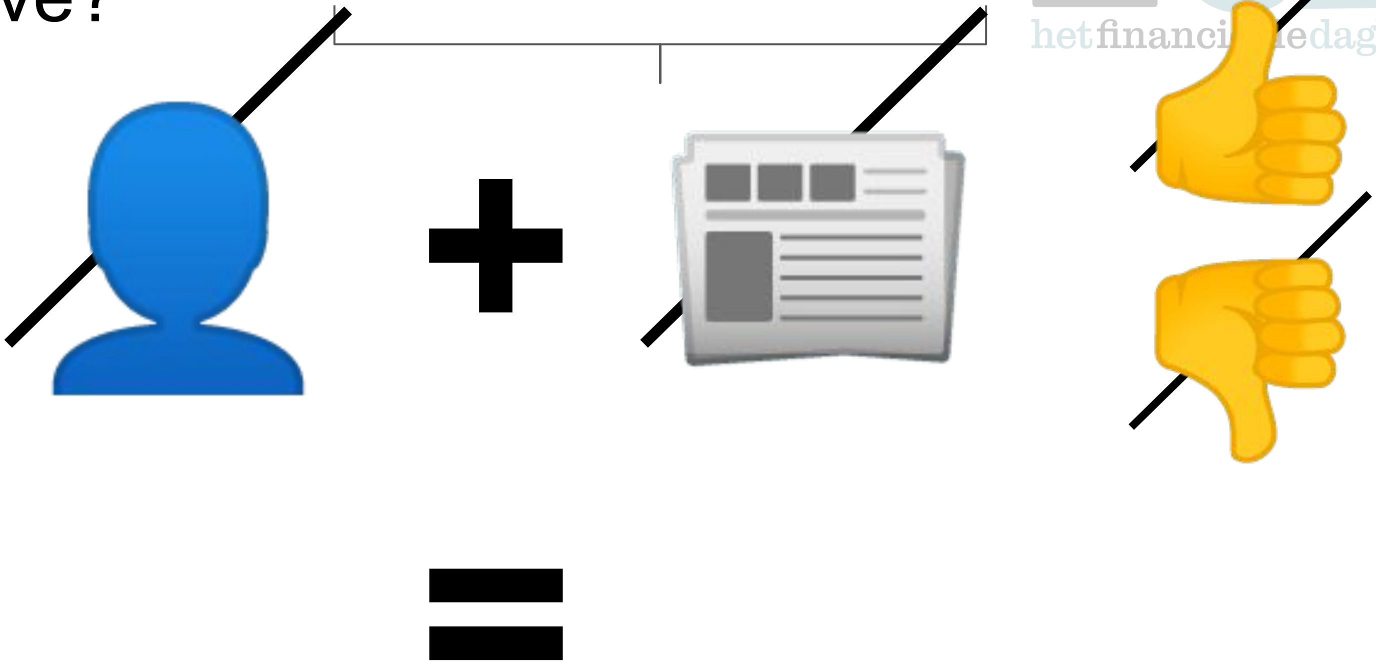


Anca Dumitrache
@anca_dmtrch

What data do we have?

Metadata:
Authors
Publication date
Length
Section

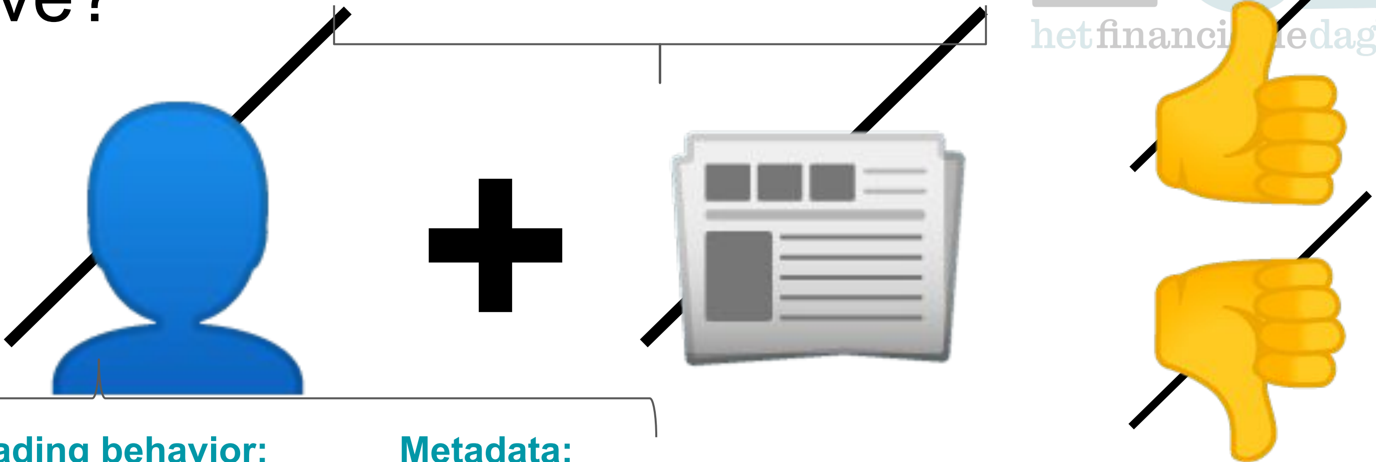
Enrichments:
Tags
Word embeddings
Sentiment
Hapax legomena



What data do we have?

Metadata:
Authors
Publication date
Length
Section

Enrichments:
Tags
Word embeddings
Sentiment
Hapax legomena



Aggregated reading behavior:
Most read tags
Most read authors
Average read article length
Average read article word embeddings

Metadata:
Tags followed

Anca Dumitrache
@anca_dmtrch

Train data:
clicks from the past

👤 📰 👍

👤 📰 👎

👤 📰 👍

👤 📰 👎

👤 📰 👎

👤 📰 👍

...

Predict data:
clicks from tomorrow

👤 📰 ?

👤 📰 ?

👤 📰 ?

...

Training process

Train data

👤📰👍

👤📰👎

👤📰👎

👤📰👍

👤📰👎

👤📰👎

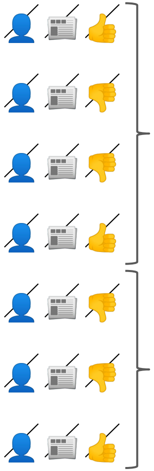
👤📰👍

...

Anca Dumitrache
@anca_dmtrch

Training process

Train data

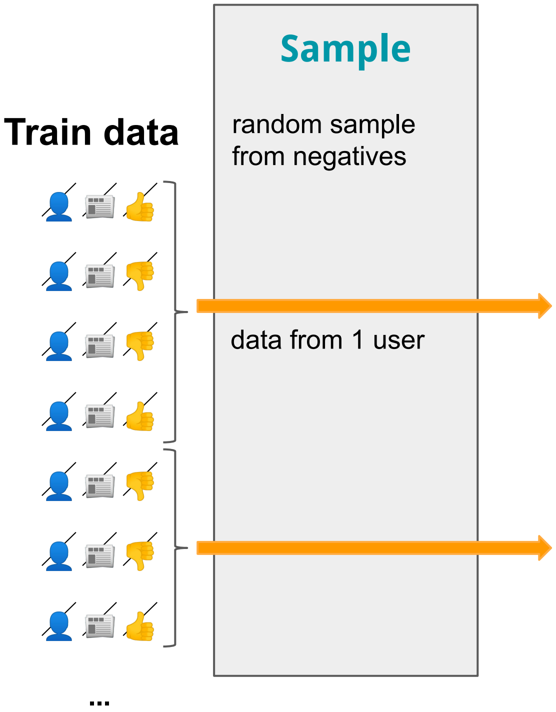


data from 1 user

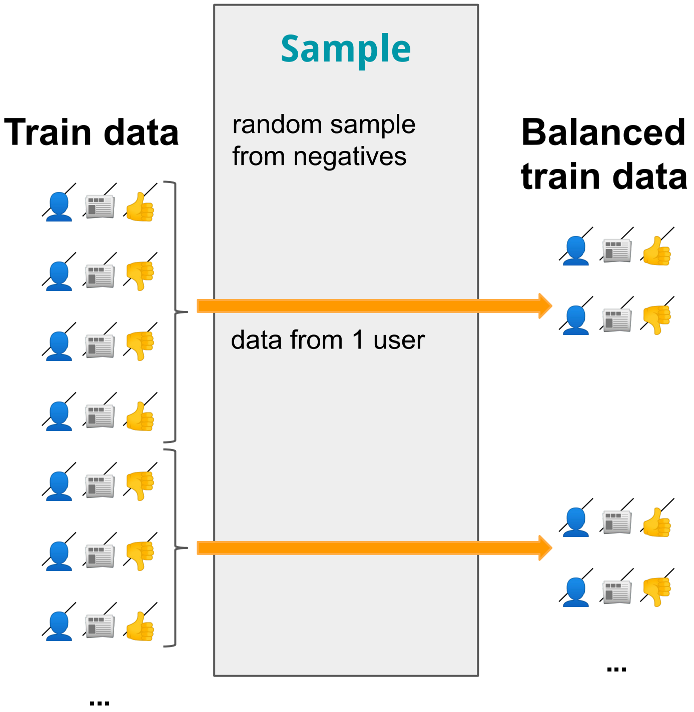
...

Anca Dumitrache
@anca_dmtrch

Training process

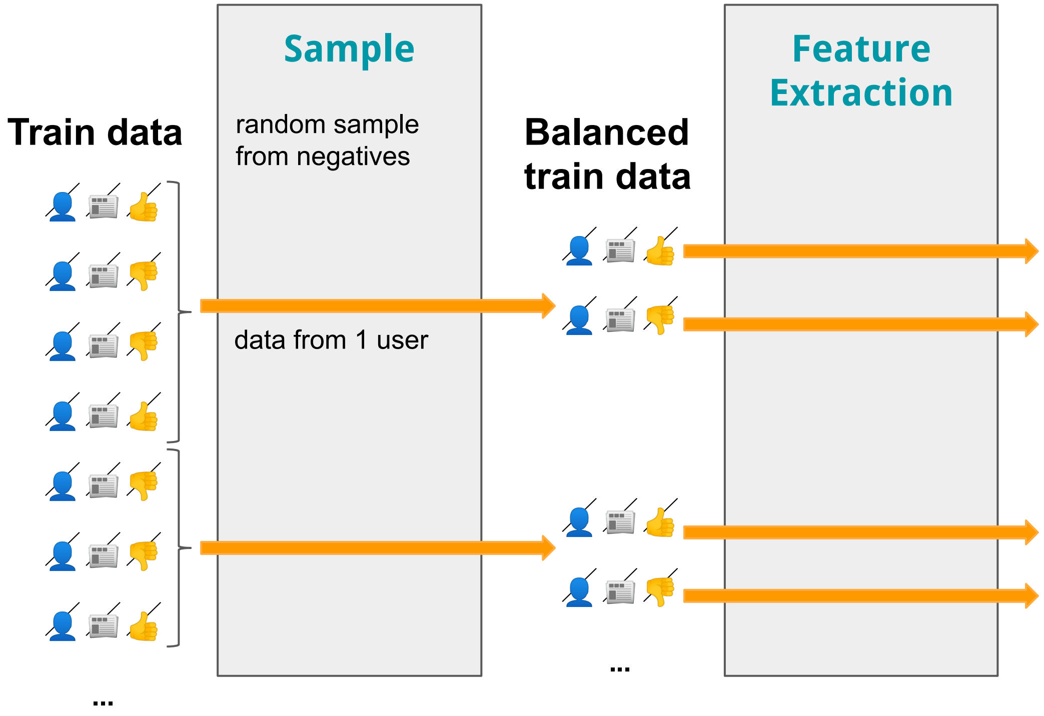


Training process



Anca Dumitrache
@anca_dmtrch

Training process



Anca Dumitrache
@anca_dmtrch

Features

Article features
encoded: authors, tags
embedded: average & var of word embeddings (`FastText`)
length-based: number of words, paragraphs, average word length
stylometry: sentiment, hapax legomena
time: publication timestamp

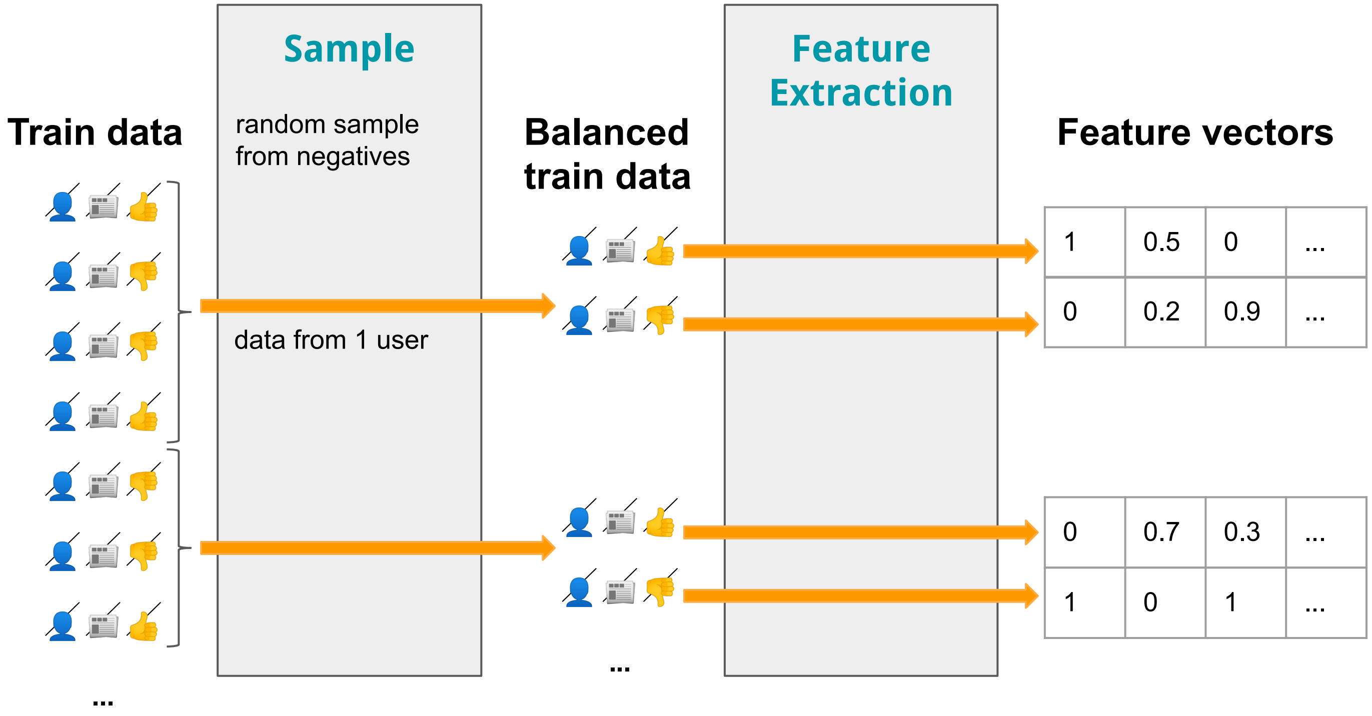
Features

Article features	User features
encoded: authors, tags	**encoded:** tags followed
embedded: average word embeddings (`FastText`)	**aggregated reading behavior:** avg. article word embeddings avg. article length features
length-based: number of words, paragraphs, average word length	
stylometry: sentiment, hapax legomena	
time: publication timestamp	

Features

<table>
<tr><th>Article features</th><th>User features</th><th>User-article features</th></tr>
<tr><td>encoded: authors, tags</td><td>encoded: tags followed</td><td rowspan="2">set overlap: article & top user tags, article & top user authors</td></tr>
<tr><td>embedded: average word embeddings (<code>FastText</code>)</td><td rowspan="4">aggregated reading behavior:
avg. article word embeddings
avg. article length features</td></tr>
<tr><td>length-based: number of words, paragraphs, average word length</td><td>cosine similarity: article & user avg. word embeddings</td></tr>
<tr><td>stylometry: sentiment, hapax legomena</td><td rowspan="2">numeric comparison: difference between article & user avg. length, article & user avg. hapax legomena</td></tr>
<tr><td>time: publication timestamp</td></tr>
</table>

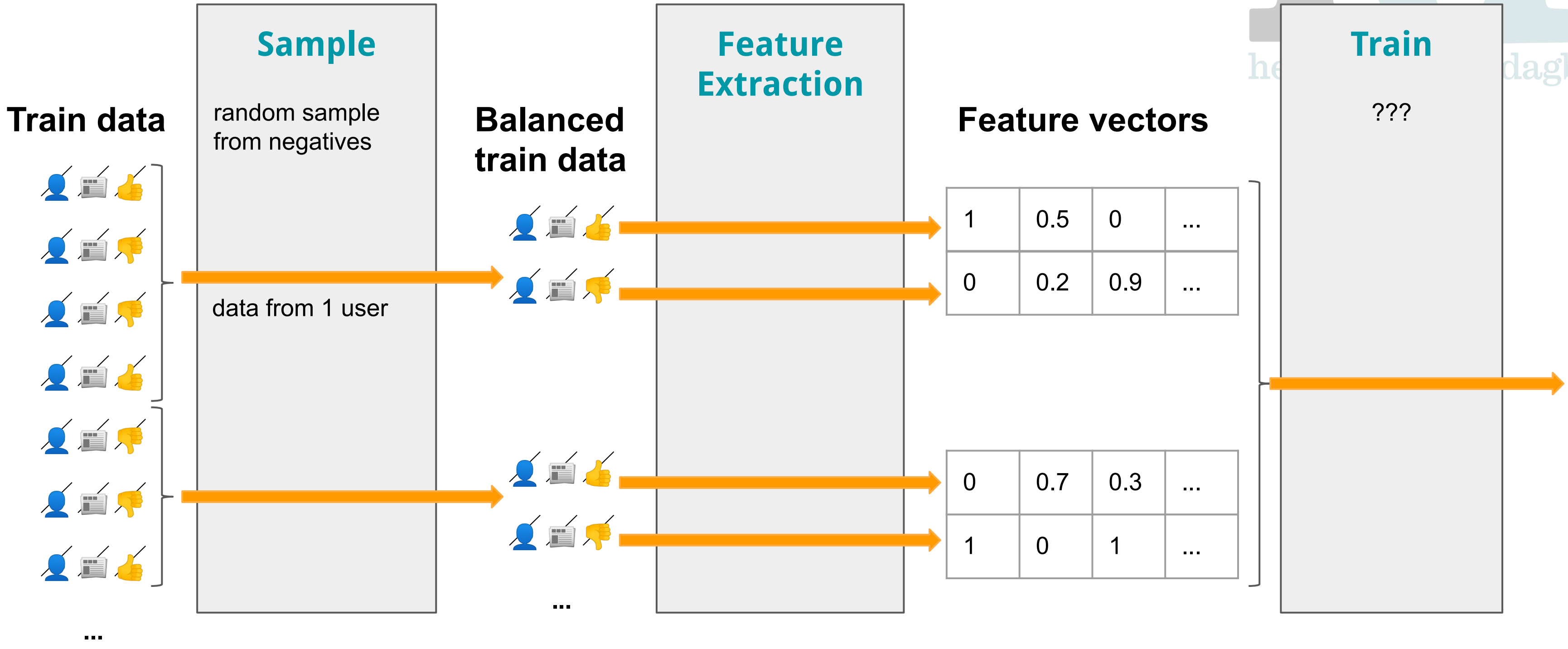
Training process



fd. hetfinancieeledagblad

Anca Dumitrache
@anca_dmtrch

Training process



Anca Dumitrache
@anca_dmtrch

Research Questions

1. What **model**?

2. What **data**?

3. What **features**?

Experimental Setup

Data: offline interactions from January 2019 (1-27 Jan train; 28-29 Jan val)

Evaluation metrics: user nDCG (ranking) & MAP (ranking + classification)

1. What model?

Models:

- Gradient Boosted Decision Trees (GBDT)
- GBDT + Logistic Regression

Practical Lessons from Predicting Clicks on Ads at Facebook. He et al. 2014.

Training methods:

- **fit:** train new model every day
- **partial fit:** re-train previous day's model with today's data, without adding new trees
- **partial fit grow:** re-train previous day's model with today's data, with new trees

What is GBDT?

- Machine learning model that iteratively constructs an ensemble of weak decision tree learners through gradient boosting.

- At each iteration, a subsample of the training data is drawn at random (without replacement), to fit the model on.

- The model captures interactions amongst predictors.

Why GBDT?

- Deals with a heterogeneous mix of continuous, discrete, categorical features
- Feature normalization is not required
- Feature selection is inherently performed during the learning process
- Can easily capture non-linear, non-additive relations

2. What data?

Data recency: How many days in the past should we look?

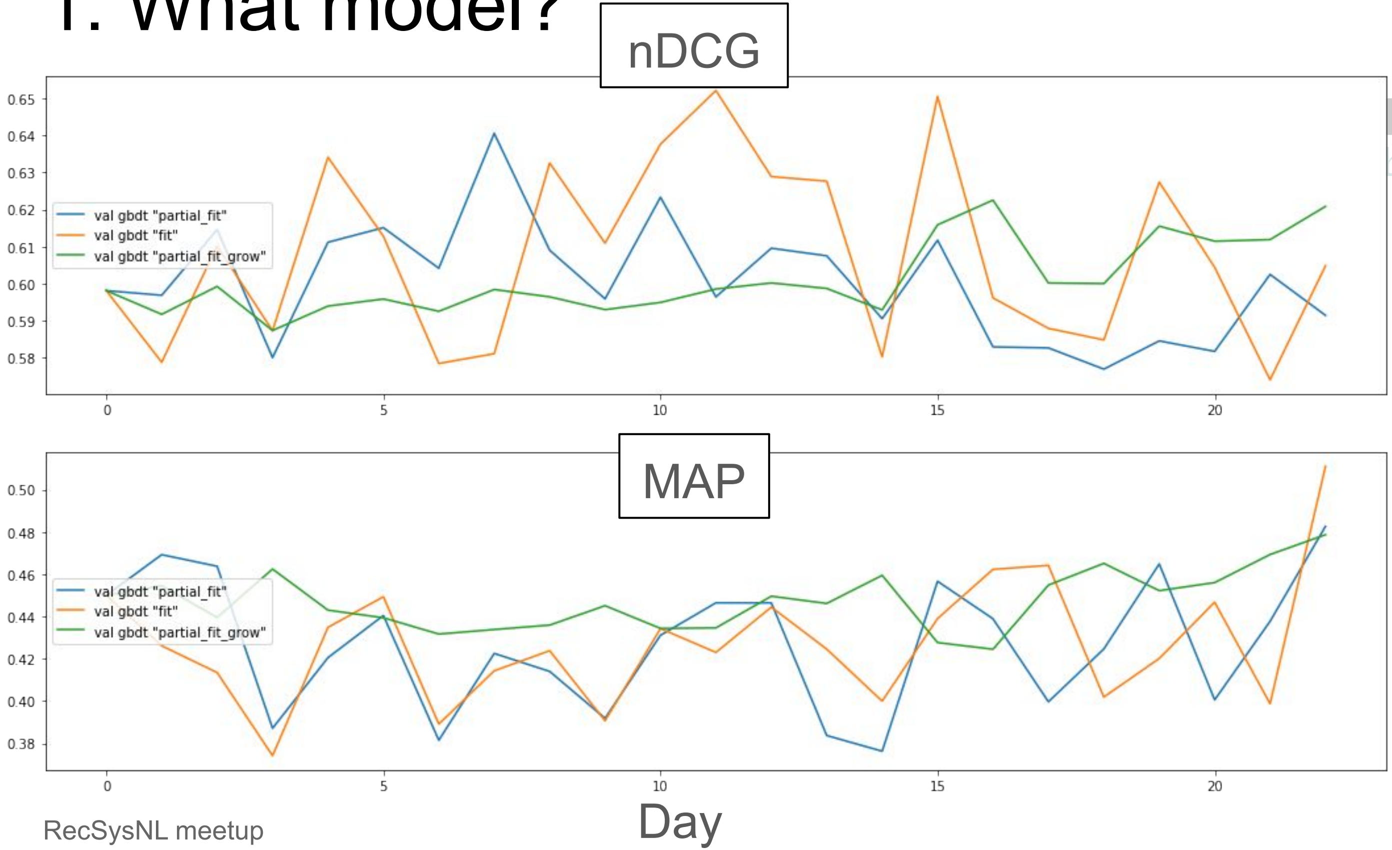
Data volume: Do we need to train on all the interactions?

Anca Dumitrache
@anca_dmtrch

3. What features?

<table>
<tr><th>Article features</th><th>User features</th><th>User-article features</th></tr>
<tr><td>encoded: authors, tags</td><td>metadata: tags followed</td><td rowspan="2">set overlap: article & top user tags, article & top user authors</td></tr>
<tr><td>embedded: average word embeddings (<code>FastText</code>)</td><td rowspan="4">aggregated reading behavior:
avg. word embeddings
avg. length features</td></tr>
<tr><td>length-based: number of words, paragraphs, average word length</td><td>cosine similarity: article & user avg. word embeddings</td></tr>
<tr><td>stylometry: sentiment, hapax legomena</td><td rowspan="2">numeric comparison: article & user avg. length, article & user avg. hapax legomena</td></tr>
<tr><td>time: publication date</td></tr>
</table>

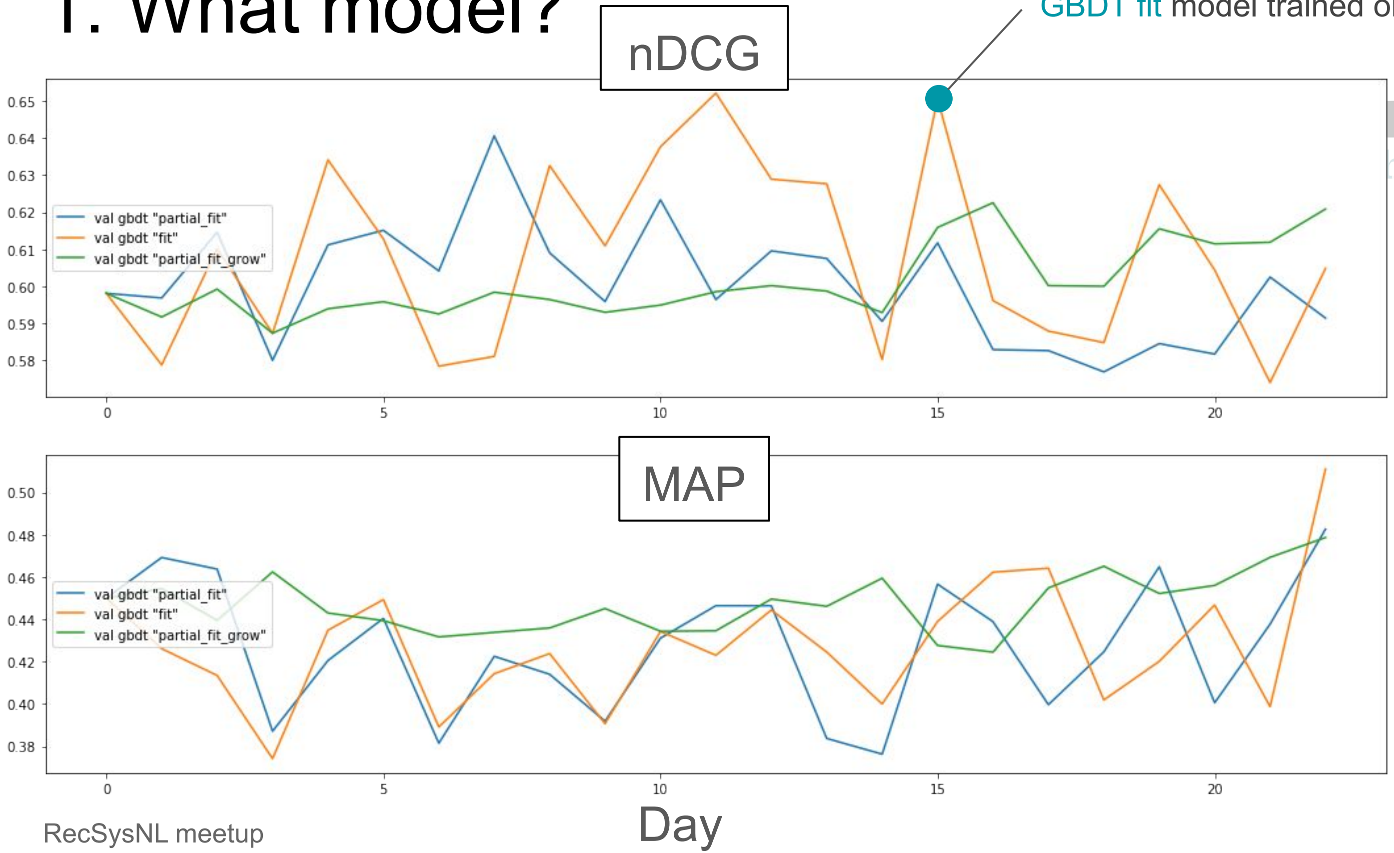
1. What model?



Anca Dumitrache
@anca_dmtrch

1. What model?

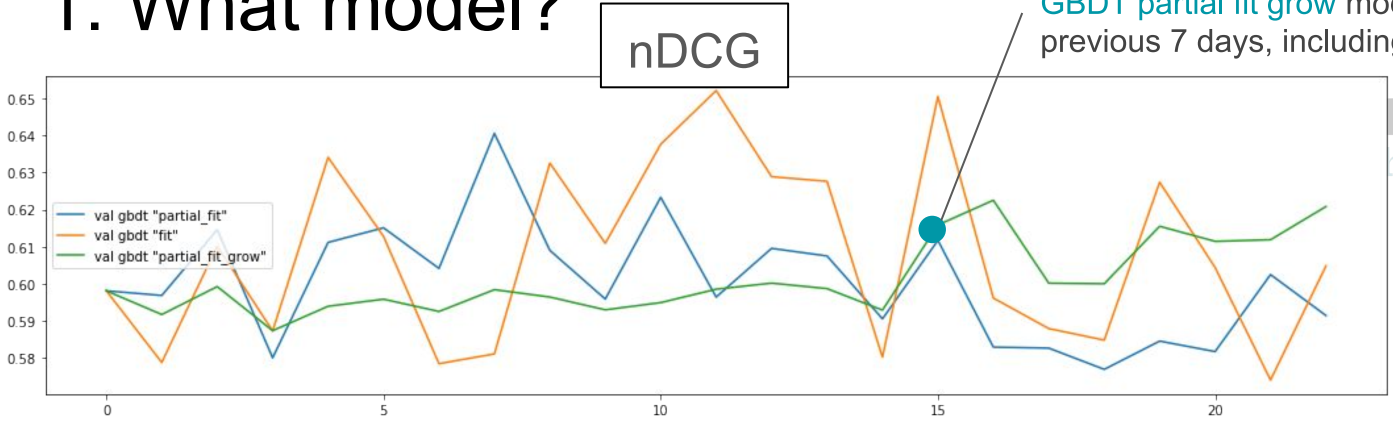
GBDT fit model trained on data from that day

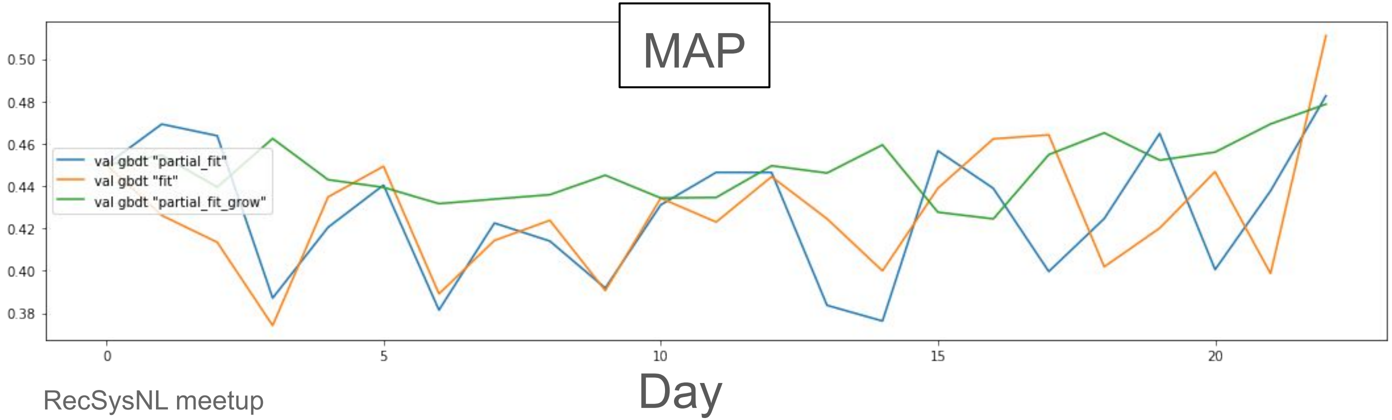


Anca Dumitrache
@anca_dmtrch

1. What model?

GBDT partial fit grow model trained on data from previous 7 days, including today

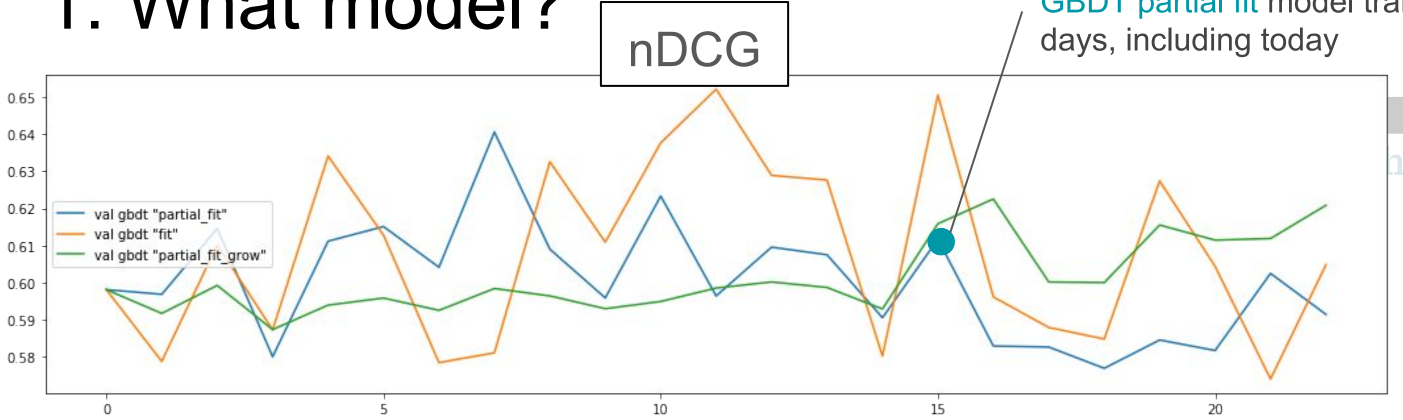


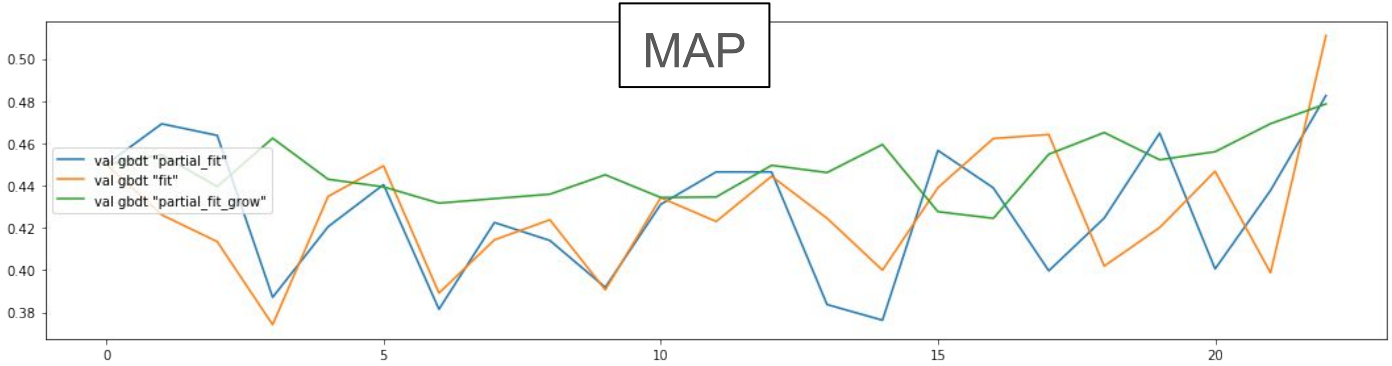


Anca Dumitrache
@anca_dmtrch

1. What model?

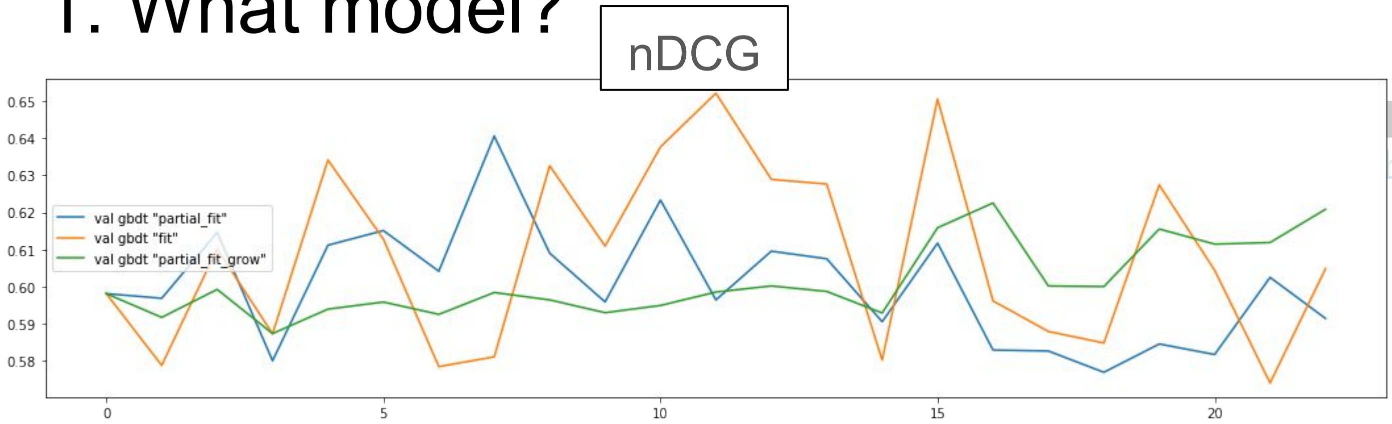
GBDT partial fit model trained on data from previous 7 days, including today

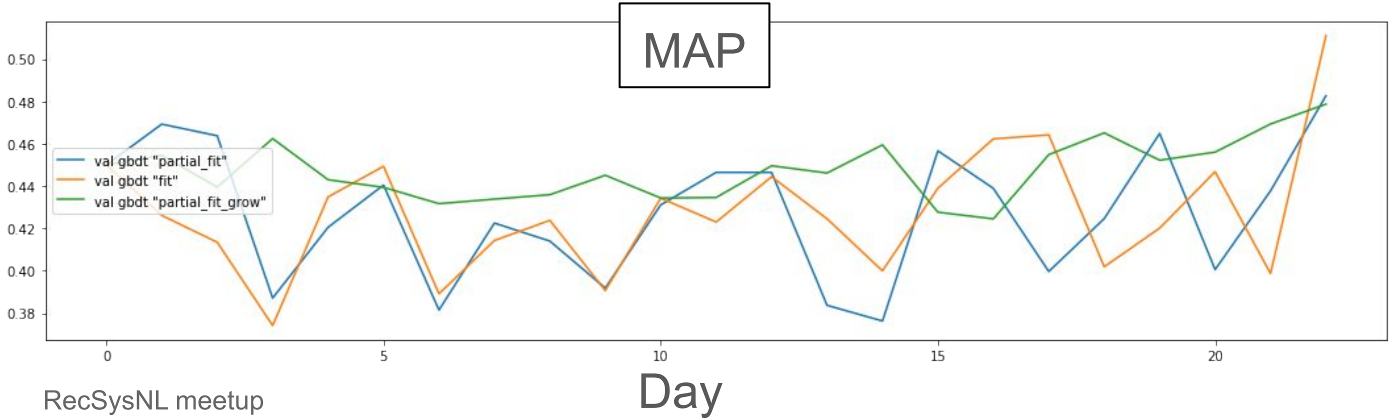




Day

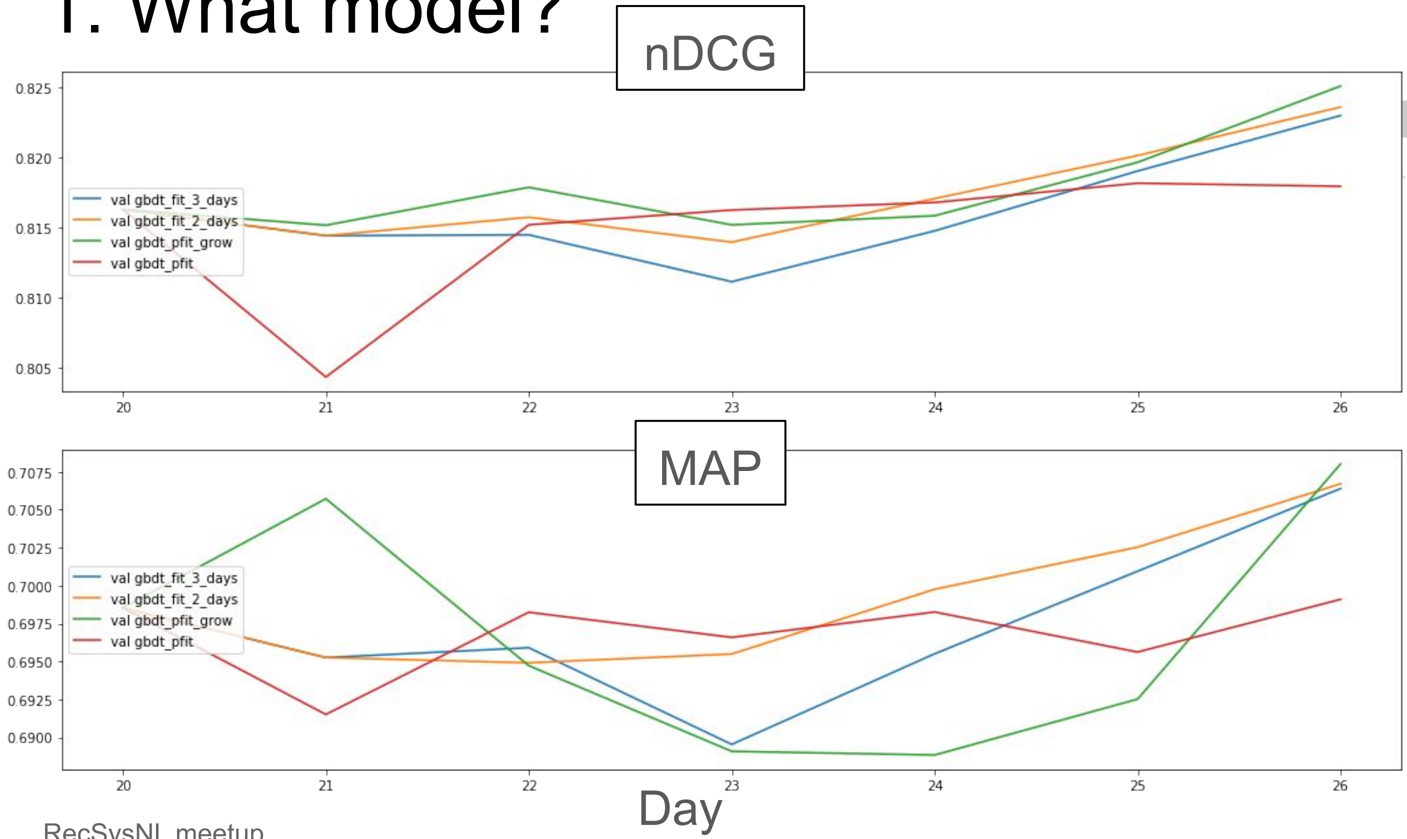
1. What model?





GBDT partial fit grow has the most stable performance, overall better evaluation results.

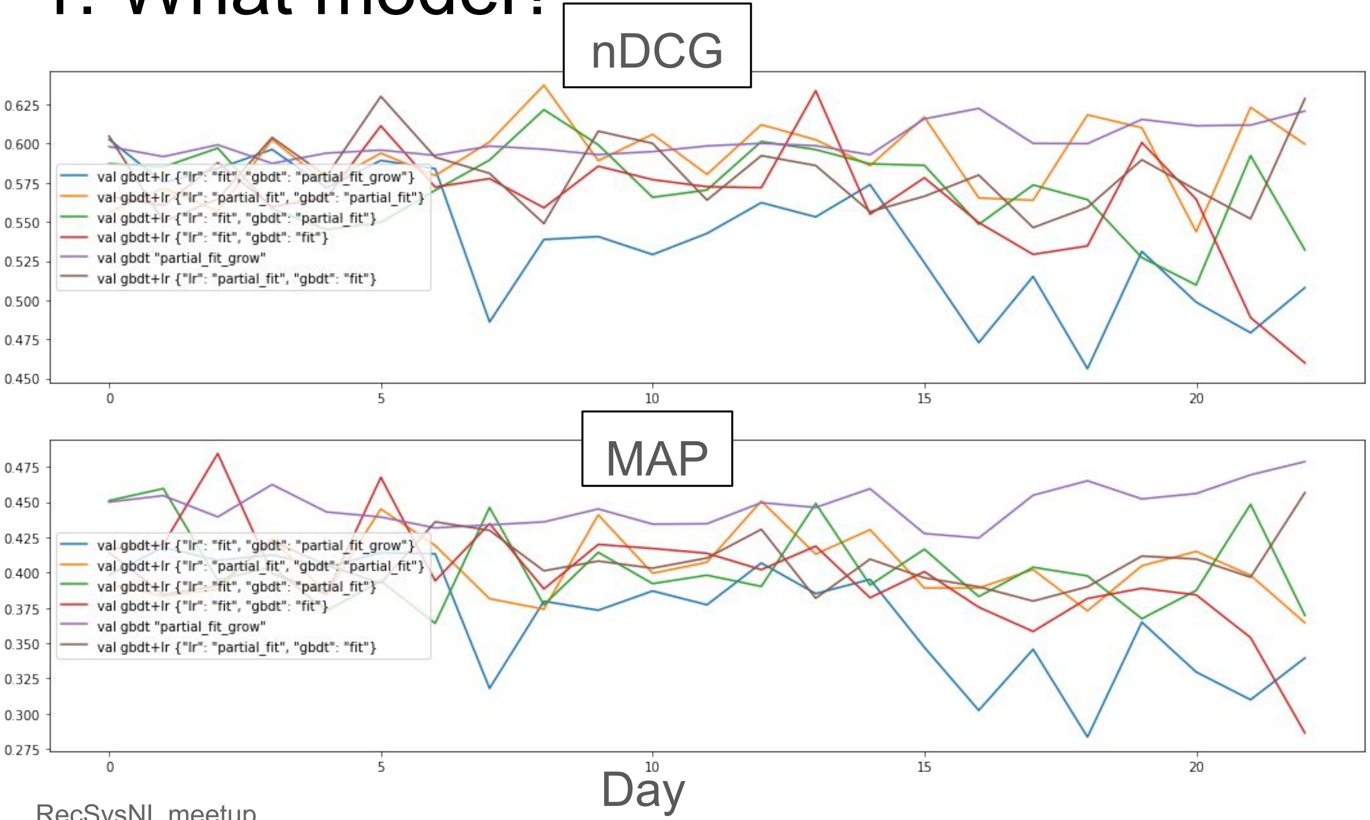
1. What model?



GBDT fit does do better when training on previous 2 & 3 days, but these models are very expensive (all in memory)!

Anca Dumitrache
@anca_dmtrch

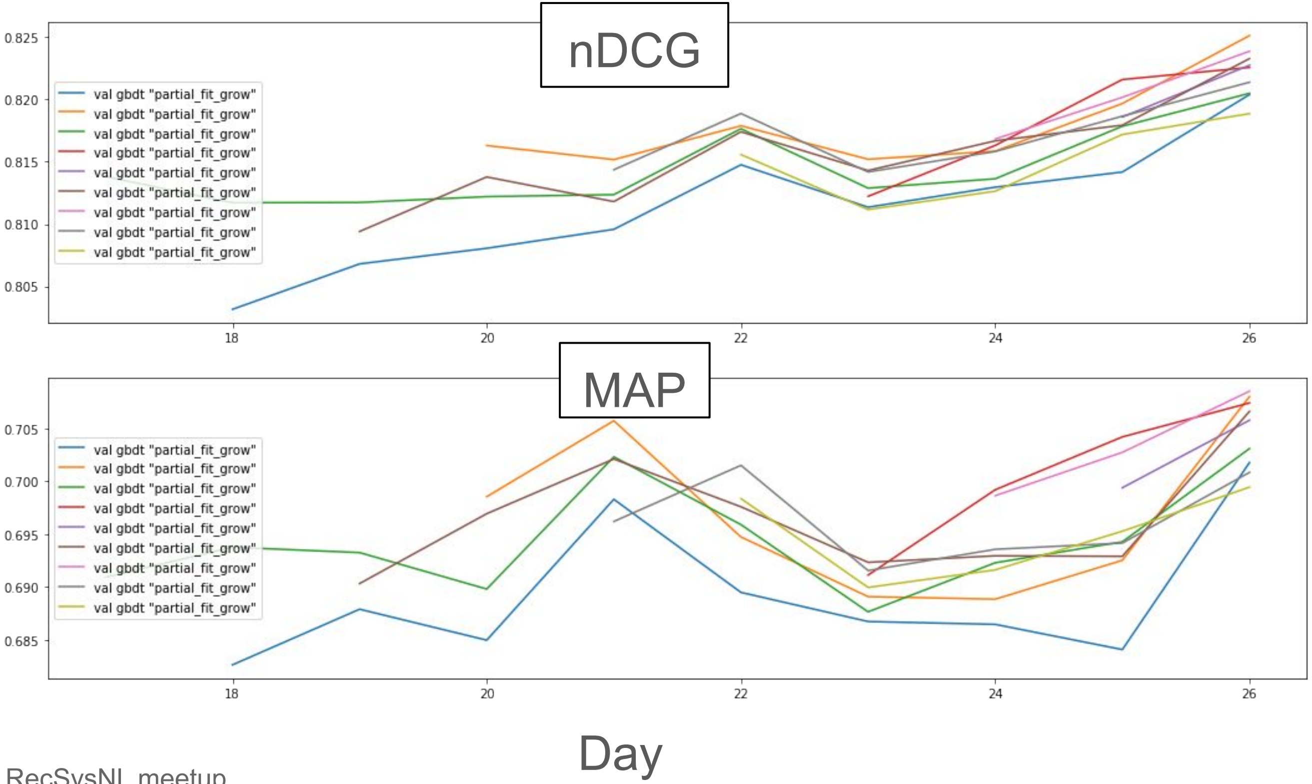
1. What model?



GBDT+LR models perform similarly, and not as well as GBDT partial fit grow.

The LR layer is difficult to tune, because the GBDT output is not interpretable.

2. How many days in the past should we train on?



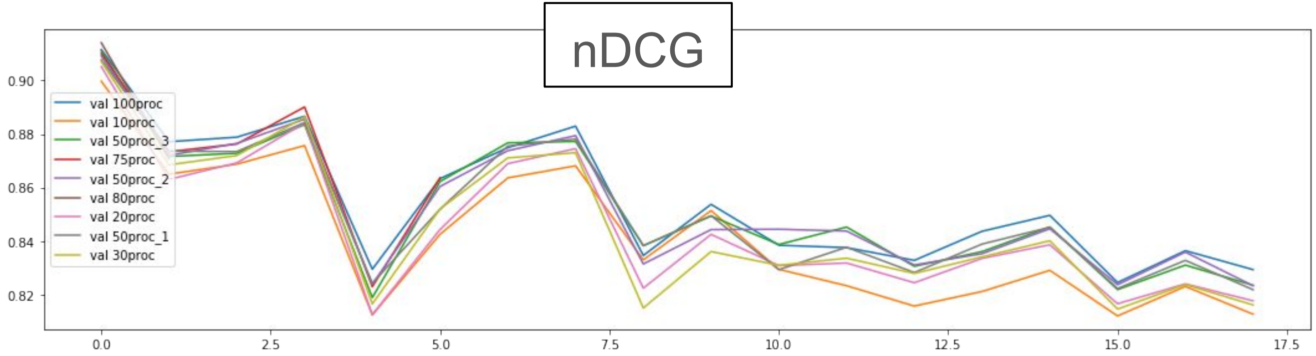
More data is not always better, because data recency is important.

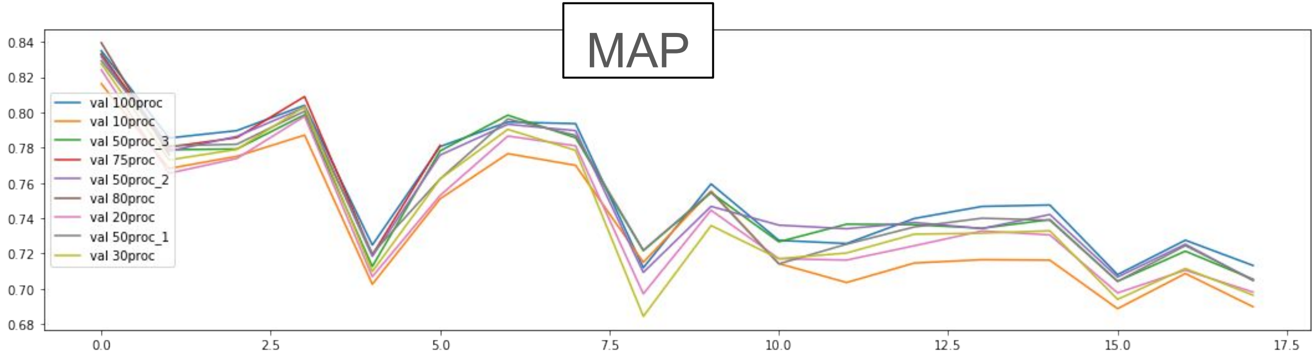
Relation between number of days & performance is not linear.

3 & 7 days in train is best.

Anca Dumitrache
@anca_dmtrch

2. Do we need all interactions?





Day

Training on 50% of the user interactions results in average 0.01 decrease in performance.

3. What features?

Extensive feature ablation experiments show we need all the features 😮

→ batch removing of low performing features takes out full chunks from the decision tree

User-article features are the most important, particularly:

user-article author overlap

user-article tag overlap

Anca Dumitrache
@anca_dmtrch

Experiment Conclusions

Simpler model (GBDT) is sometimes better + easier to understand.

Less data (days, user interactions) does not mean worse performance.

Combined user-article features are the most meaningful, but all features contribute a little.

Training process

Train data

...

Sample

random sample from negatives

data from 1 user

Balanced train data

...

Feature Extraction

Feature vectors

1	0.5	0	...
0	0.2	0.9	...

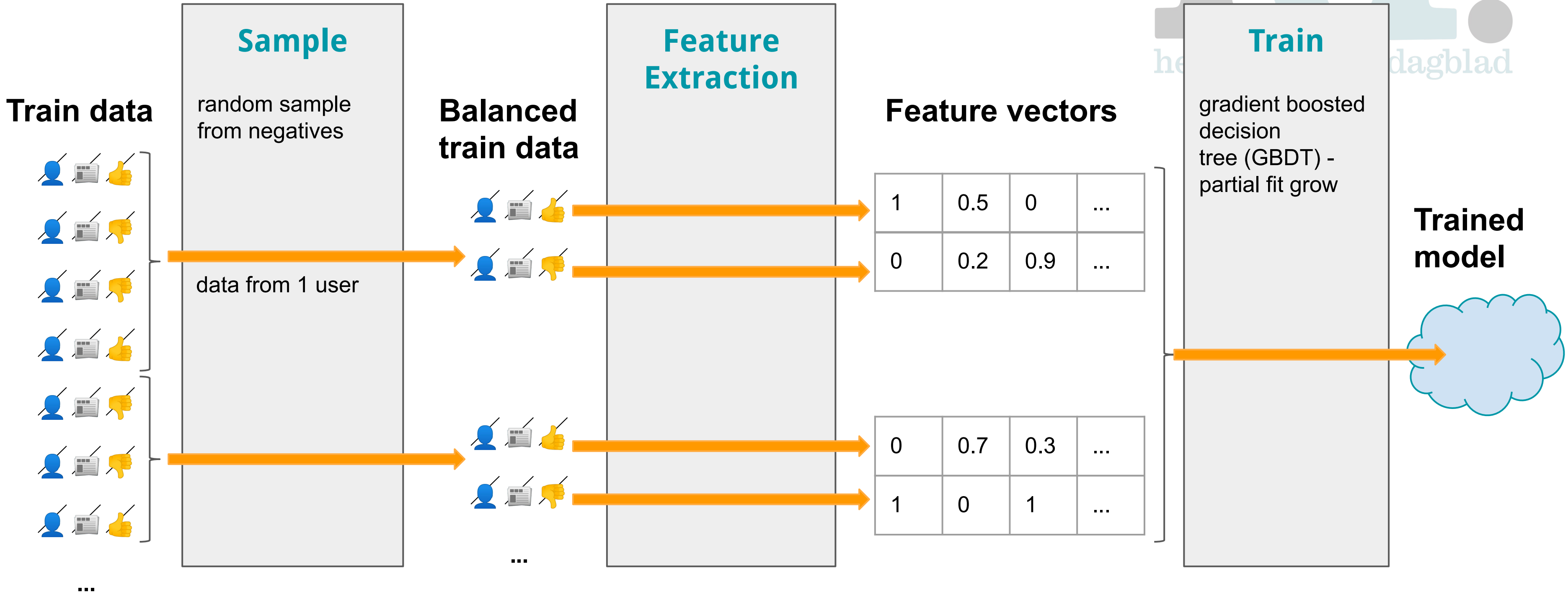
0	0.7	0.3	...
1	0	1	...

Train

gradient boosted decision tree (GBDT) - partial fit grow

Anca Dumitrache
@anca_dmtrch

Training process



Anca Dumitrache
@anca_dmtrch

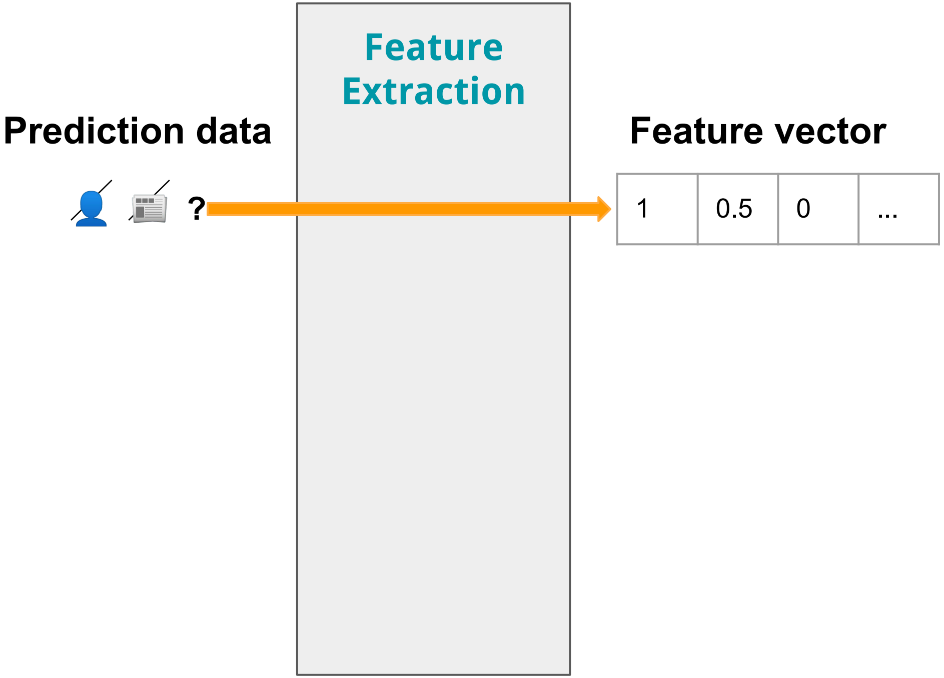
Prediction process

Prediction data

~~👤~~ ~~📰~~ **?**

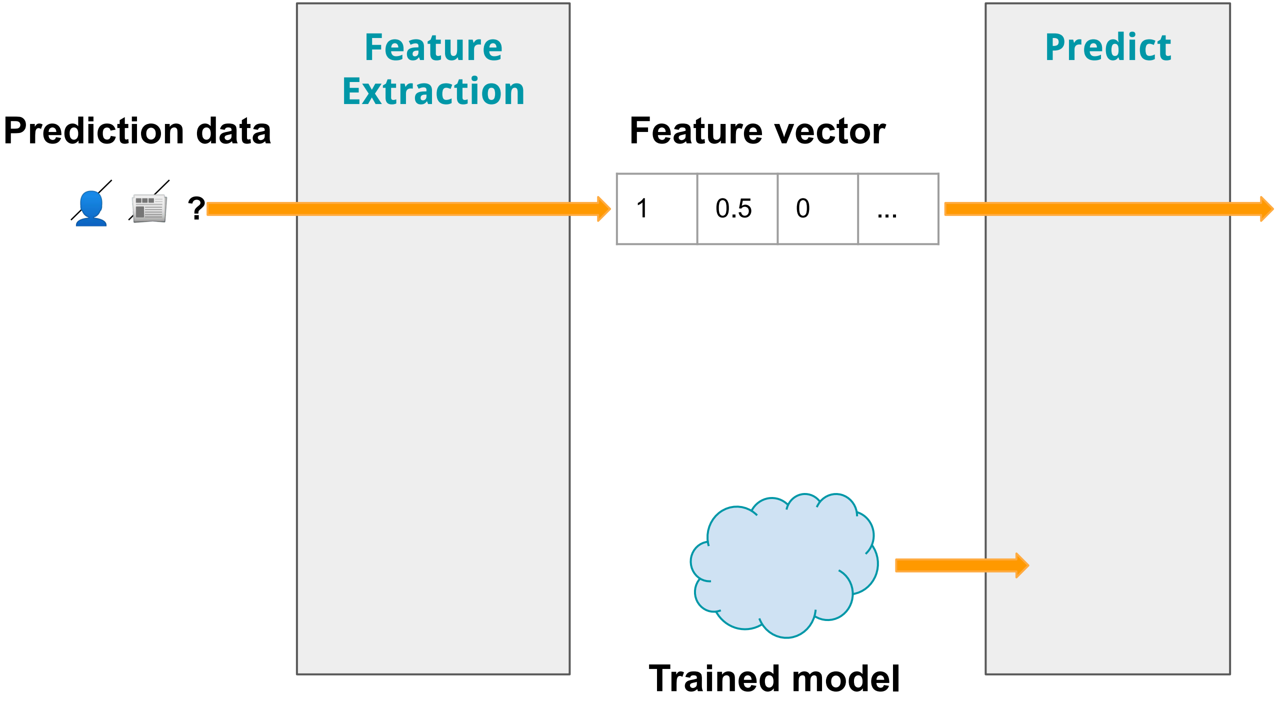
Anca Dumitrache
@anca_dmtrch

Prediction process



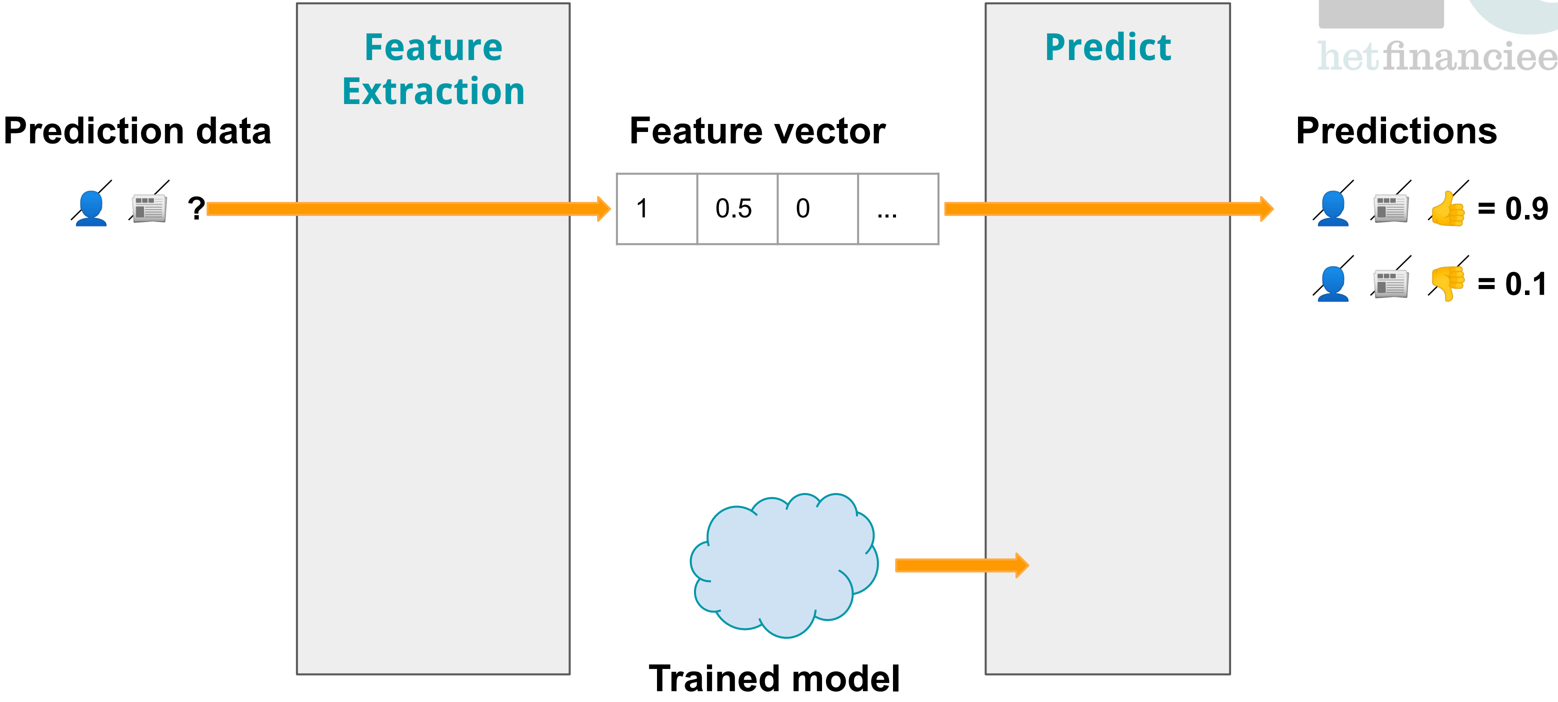
Anca Dumitrache
@anca_dmtrch

Prediction process



Anca Dumitrache
@anca_dmtrch

Prediction process



Anca Dumitrache
@anca_dmtrch

How it looks on FD.nl

articles from the past 24h

articles from the past 7 days

Gemist de afgelopen 7 dagen?

MIJN NIEUWS

AANBEVOLEN VOOR U

Dit kan er beter in de zorgverzekering

Deze week maakten de zorgverzekeraars de nieuwe premies voor de basisverzekering bekend. Waarom verschilt de prijs voor dezelfde zorg? En werkt het systeem eigenlijk wel goed genoeg?

Bewaren

AANBEVOLEN VOOR U

VEB kritisch over transactie tussen Mountainshield en DGB

Vereniging voor Effectenbezitters zet vraagtekens bij aandelentransactie van Mountainshieldfonds, dat in een half jaar tijd €3,7 mln fondsvermogen verloor.

Bewaren

AANBEVOLEN VOOR U

Kapitalisme redt, met wat hulp, de aarde

MIT's Andrew McAfee schrijft deze keer een optimistisch boek: Meer uit minder

Bewaren

AANBEVOLEN VOOR U

Bouwers nieuwe windmolen gooien modder bij Ondernemingskamer

De mogelijkheden leken ongekend. Een windmolen die met minder overlast net zoveel vermogen levert als de standaardmodellen. De initiatiefnemers van windmolenbouwer Mega Windforce zagen gouden bergen. Nog geen drie jaar later dreigt de ondergang. Deze week diende de zaak in de Ondernemingskamer.

Bewaren

AANBEVOLEN VOOR U | NET BINNEN

PRIVACY EN CYBERSECURITY

Justitie dwingt Microsoft tot aanpassing wereldwijde cloudcontracten

TECH EN MEDIA

SoftBank probeert met Japanse internetkampioen op te boksen tegen Google

FINANCIËLE SECTOR

VEB kritisch over transactie tussen Mountainshield en DGB

MARKTEN

Franse toezichthouder eist €5mln van persbureau Bloomberg

OPINIE

Mkb'ers kunnen wél meer investeren, ook in mensen

MIJN NIEUWS

Future work

Online testing - currently ongoing

Measuring usefulness (dynamicness, serendipity, diversity) aspects of recommendations & seeing how readers respond to them

Closing the loop: tweaking the model based on how the recommendations are presented to readers

Conclusions

Simpler model (GBDT) is sometimes better + easier to understand.

Less data (days, user interactions) does not mean worse performance.

Combined user-article features are the most meaningful, but all features contribute a little.

Anca Dumitrache
@anca_dmtrch