Content-based recommendations for financial news

PyData Eindhoven

30 November 2019

Anca Dumitrache @anca_dmtrch

Feng Lu @dsflu

Team:

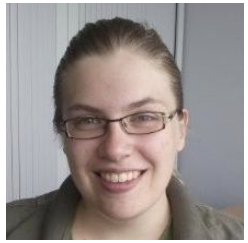
Feng

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 & Feng Lu @dsflu

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

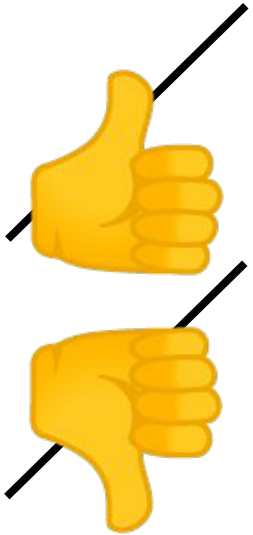
13:48 Mollie mikt met vers groeigeld op Europese klanten

13:47 ASML nieuwe partner Van Gogh Museum

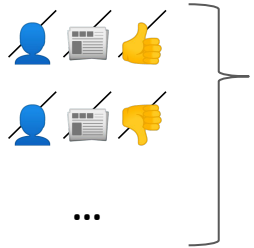
Articles seen but not clicked

Anca Dumitrache @anca_dmtrch & Feng Lu @dsflu

What data do we have?



Train data: clicks on articles from the past



Training Process

Trained model

Predict data: clicks on articles from tomorrow

Prediction Process

Predictions

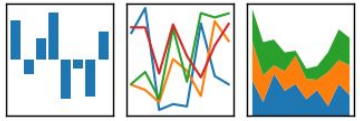
0.9

new articles every day ➡️ cold start problem

Anca Dumitrache @anca_dmtrch & Feng Lu @dsflu

Python stack

`pytrec_eval`



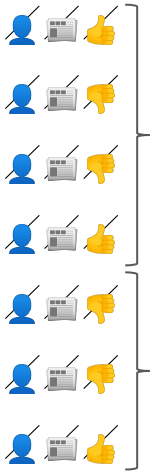
Training process

Train data

...

Training process

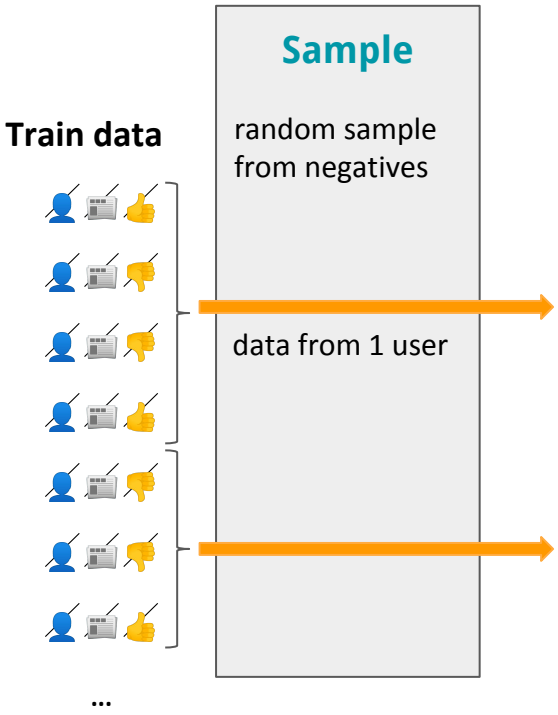
Train data



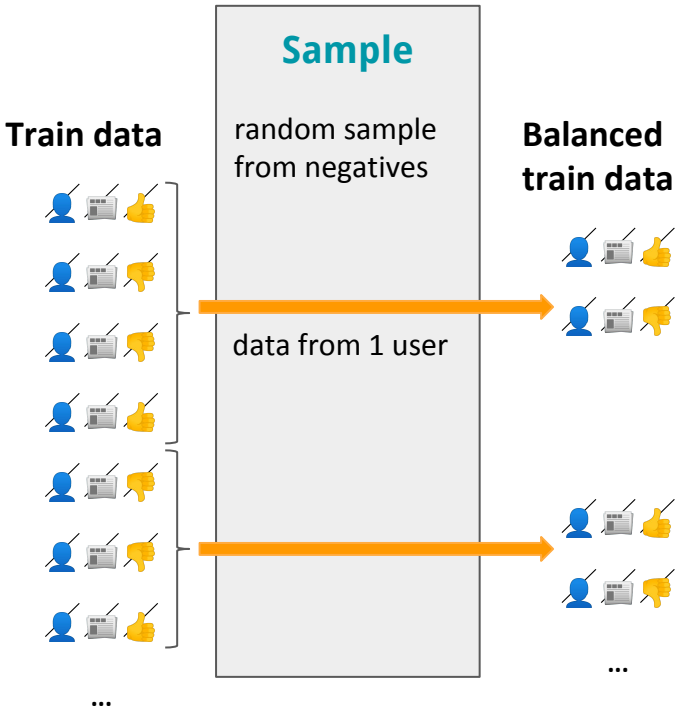
data from 1 user

...

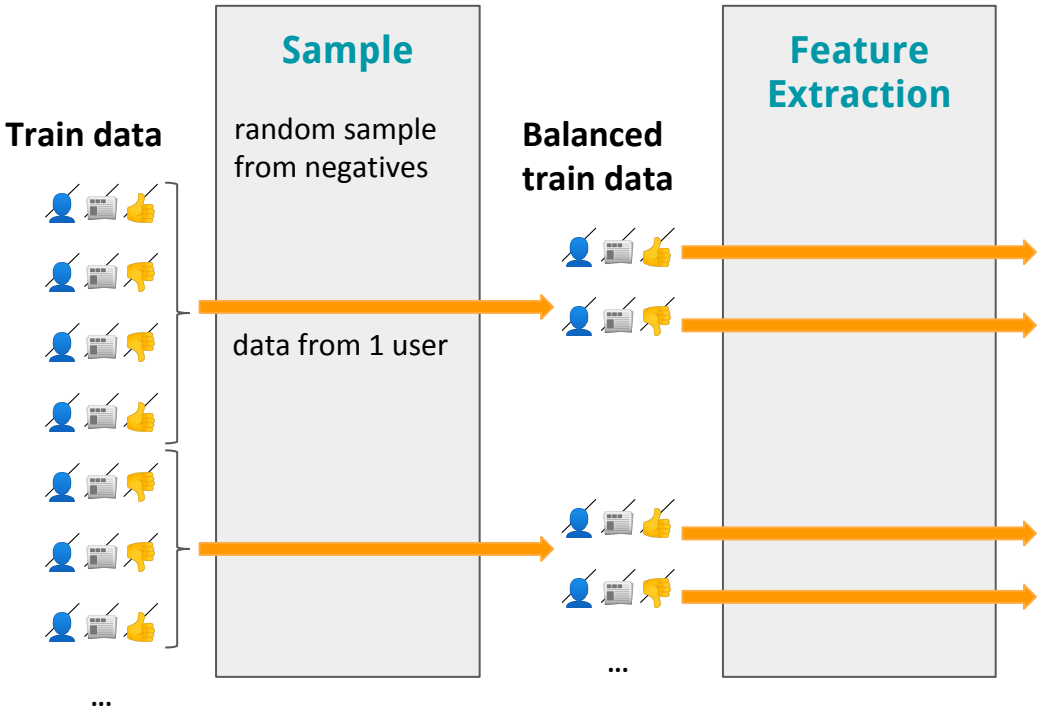
Training process



Training process



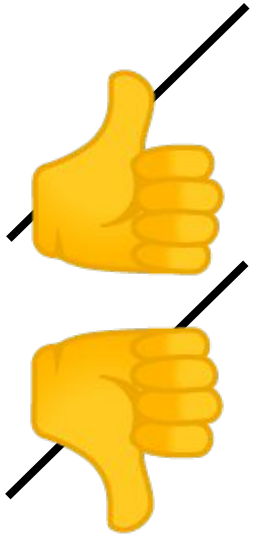
Training process



Features

User Features

Article Features



User-Article Features

Anne Schuth / Data Science at Blendle / Sanoma TechTalks23 (YouTube)

Blendle

Article Features

Metadata ✔
- ➔ Authors
- ➔ Content
- ➔ Section
- ➔ Publication date
- ➔ Long story?

Enrichments ✔
- ➔ #paragraphs, #sentences, #words
- ➔ Tags
- ➔ Article length
- ➔ Article complexity
- ➔ Hapax legomenon
- ➔ Sentiment
- ➔ Word Embeddings

Anca Dumitrache @anca_dmtrch & Feng Lu @dsflu

Article Features

Metadata ✔

- ➔ Authors
- ➔ Content
- ➔ Section
- ➔ Publication date
- ➔ Long story?

Enrichments ✔

- ➔ #paragraphs, #sentences, #words
- ➔ Tags
- ➔ Article length
- ➔ Article complexity
- ➔ Hapax legomenon
- ➔ Sentiment
- ➔ Word Embeddings

Anca Dumitrache @anca_dmtrch & Feng Lu @dsflu

Article Features

text pipe

SCAN ME

Metadata ✔
- ➔ Authors
- ➔ Content
- ➔ Section
- ➔ Publication date
- ➔ Long story?

Enrichments ✔
- ➔ #paragraphs, #sentences, #words
- ➔ Tags
- ➔ Article length
- ➔ Article complexity
- ➔ Hapax legomenon
- ➔ Sentiment
- ➔ Word Embeddings

spaCy

fastText

NLTK

Anca Dumitrache @anca_dmtrch & Feng Lu @dsflu

Article Features

BANKEN

Rabobank terechtgewezen om gebrekkige witwasscreening

Grootbank moet duizenden klantdossiers opnieuw onderzoeken na ingrijpen door toezichthouder DNB

40.000
dossiers van klanten moet Rabobank onder de loep nemen wegens onduidelijk witwasrisico

Marcel de Boer en Johan Leupen
Amsterdam

De Nederlandsche Bank heeft ingegrepen bij de Rabobank, omdat de grootbank de screening van nieuwe Nederlandse klanten heeft verwaarloosd, en witwasrisico's daardoor niet goed in beeld heeft. De toezichthouder heeft Rabo daarom in 2018 een last onder dwangsom opgelegd, en de bank is sindsdien bezig om tienduizenden klantdossiers opnieuw door te lichten op mogelijke witwasrisico's. Dat blijkt uit interne Rabo-stukken in handen van het FD.

Een belangrijke zorg van DNB is dat Rabobank de afgelopen jaren mogelijk ten onrechte veel particuliere Nederlandse klanten heeft ingedeeld in de laagste risicoklasse, die niet actief wordt gecontroleerd. Zo'n 40.000 klanten hebben dat stempel 'potentieel ten onrechte' gekregen, zo bevestigt de coöperatieve bank uit Utrecht desgevraagd. Rabo heeft op last van de toezichthouder enkele honderden werknemers ingeschakeld voor een reeks verbeterplannen, die in april 2020 moeten zijn afgerond.

Ook andere banken liggen onder een vergrootglas van de toezichthouder in verband met witwaspraktijken. Gisteren meldde NRC Handelsblad dat ABN Amro zonder enig onderzoek vele particuliere klanten als betrouwbaar heeft bestempeld, en dat die nalatigheid onderdeel uitmaakt van het strafrechtelijk onderzoek van het Openbaar Ministerie naar de bank.

De last onder dwangsom is een van de zwaardere maatregelen die DNB in zijn arsenaal heeft om instellingen te dwingen reparaties te doen na een overtreding. Vorig jaar is zo'n last vier keer uitgedeeld, waarvan één keer aan een bank. Begin dit jaar bleek al dat Rabobank van de toezichthouder een boete van €1 mln heeft gekregen, vanwege onvoldoende controle van klanten. In februari zei een woordvoerder daarover nog dat de problemen speelden 'tot en met 2016' en dat ze 'inmiddels zijn aangepakt'.

Naar nu blijkt heeft Rabobank nog altijd veel werk te verzetten. 'Uiteraard willen we deze dossiers up-to-date hebben, in het belang van onze klanten en voor de bank zelf. Hierin waren wij naar de mening van DNB destijds tekortgeschoten', zegt een woordvoerder tegenover het FD. Rabo benadrukt dat de kwestie geen verband houdt met specifieke witwaspraktijken die zijn doorgelaten.

Ingewijden zeggen dat controleurs regelmatig klanten tegenkomen die ten onrechte laag zijn ingeschaald of zelfs helemaal niet hadden mogen worden geaccepteerd. Het gaat dan bijvoorbeeld over ondernemers in verhoogde risicosectoren zoals de autohandel of het vastgoed, maar ook over particulieren die veel contante stortingen doen met grote coupures, of die rekeningen en betrekkingen hebben in landen als Malta, Rusland en Angola.

Als een klant in de laagste risicoklasse komt, betekent dit dat een bank er op zich niet meer naar om hoeft te kijken. Een klant in de middencategorie moet elke drie jaar door de molen, in de hoogste risicoklasse ieder jaar.

Vervolg op pagina 3

FOTO ISTOCK


```
"132xxxx": {
  "content": {
    "NSentences": 39,
    "POSTags": {...},
    "NImages": 1,
    "NWords": 902,
    "AvgParagraphLength": 394.46153846153845,
    "HapaxDislegomena": 0.06208425720620843,
    "AvgWordLength": 4.919068736141907,
    "Sentiment": [-0.006161616161616151, 0.48404040404040394],
    "ArticleLength": 4437,
    "NParagraphs": 13,
    "WordEmbeddings": {
      "word2vec_average": [...],
      "word2vec_l2norm": 0.6940994407851842,
      "word2vec_sum": [...],
      "word2vec_variation": [...]
    },
    "HapaxLegomena": 0.3492239467849224,
    "Entities": [
      ["Utrecht", "LOC"],
      ["Rabobank", "ORG"],
      ...],
    "Complexity": 46.68405076730045,
    "CleanFDMGContent": "De Nederlandse Bank heeft...",
    "AvgSentenceLength": 132.94871794871796
  },
  "profile": {
    "tags": ["Rabobank", "Toezichthouder", "Banken"],
    "free": false,
    "title": "Rabo moet tienduizenden dossiers opnieuw doorlichten op witwasrisico's ",
    "short_article": false,
    "section_name": "ondernemen",
    "authors": ["Marcel de Boer", "Johan Leupen"],
    "original_article_id": "1325348",
    "publication_date": "2019-11-21T23:30:00Z"
  },
  "entity-linker": {...},
  "tagger": {...}
```

User Features

Demographic ❌

- ➔ Gender
- ➔ Age range
- ➔ ...

User Reading Behavior ✓

- ➔ Tags followed
- ➔ Most read tags/authors
- ➔ Average #words / #sentences / #paragraphs
- ➔ Average article length
- ➔ Average sentiment
- ➔ Average Word Embeddings

User Features

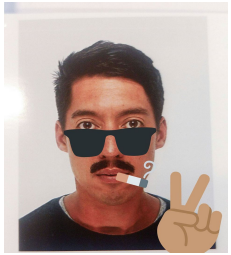

```
"profile": {
  "tags": ["algoritmen", "kunstmatige intelligentie", "data science", "machine learning", "da
  Zeemeijer", "Panama Papers", "Privacy", "Sandra Olsthoorn", "Vastgoed", "Woningmarkt", "Ams
},
"representation": {
  "user_word_embeddings_avg": {
    "average": [...],
    "sum": [...],
    "variation": [...],
    "l2norm": 0.7525256299231794
  },
  "user_editor_tags_distribution": {
    "Amazon": 3,
    "Randstad": 4,
    "Software": 8,
    "Media": 5,
    "Sport": 1,,
    "Tech en media": 22,
    "Technologie": 28,
    "Opinie": 12,
    ...
  },
  "user_author_distribution": {
    "Van onze redacteur": 44,
    "Hella Hueck": 8,
    "Rik Winkel": 2,
    ...
  },
  "user_word_length_avg": 4.7314486756453675,
  "user_sentiment_avg": [0.07734743408416966],
  "user_article_length_avg": 3116.9691358024693,
  "user_nparagraph_avg": 12.074074074074074,
  "user_nwords_avg": 667.4938271604939
},
"_ts": "2019-11-27T14:31:15.634Z"
```

User-Article Features

Article & Avg. User Set Overlap ✓
- ➔ Tags
- ➔ Authors

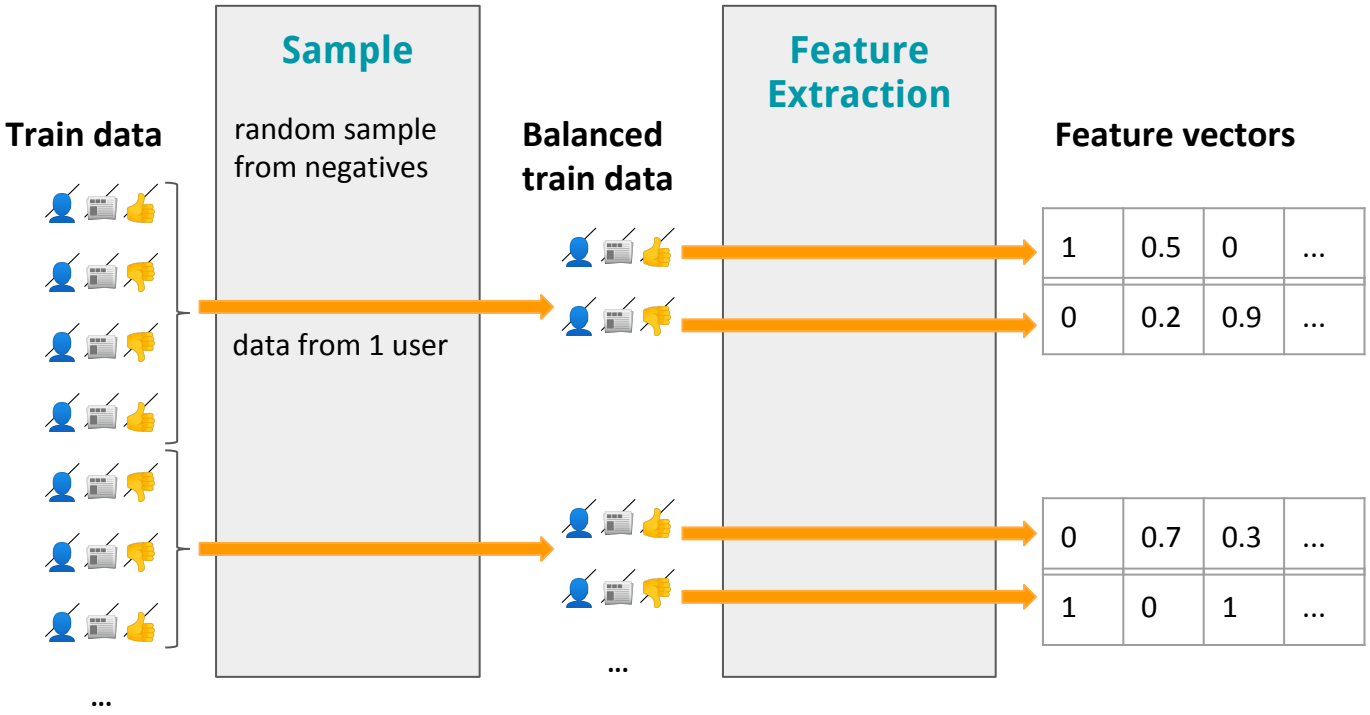
Article & Avg. User Reading Habits Comparison ✓
- ➔ Article length
- ➔ #words / #sentences / #paragraphs
- ➔ Word Embeddings similarity

Feature Representations (~14k)

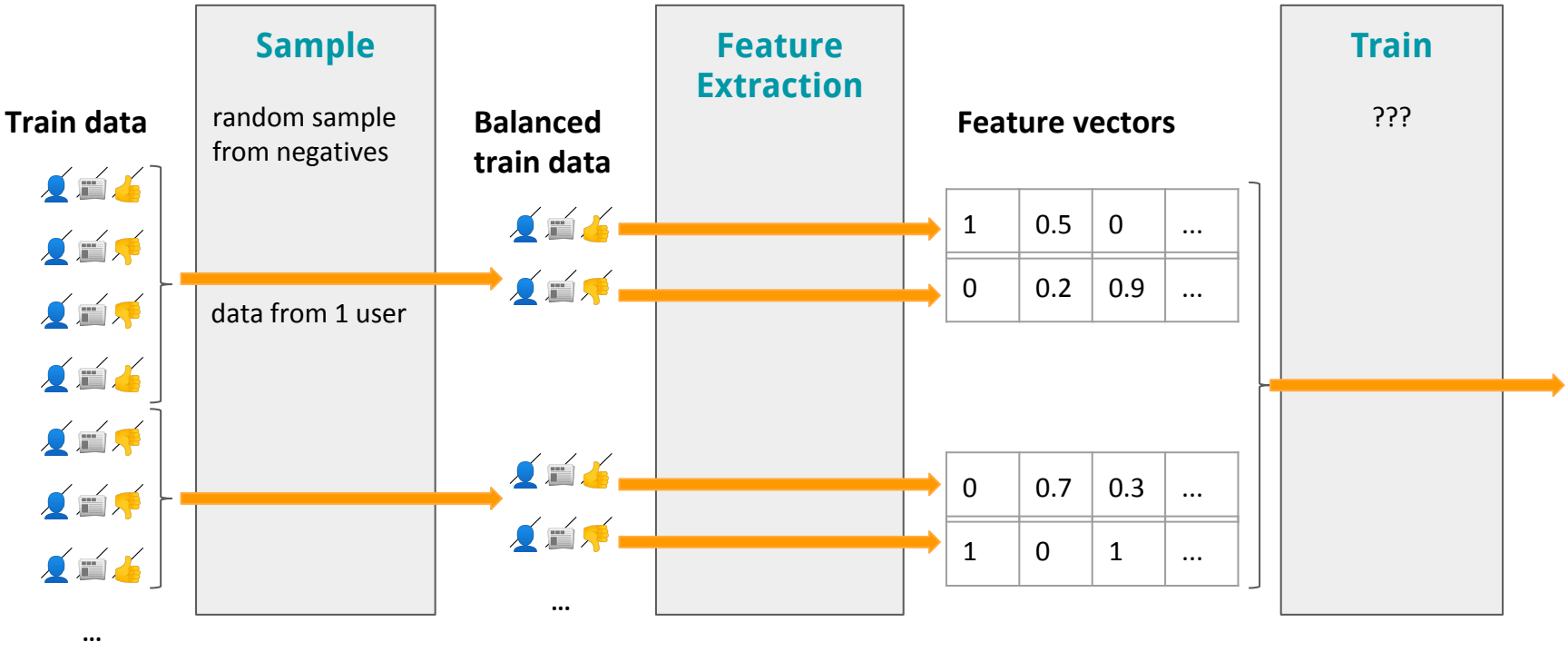


Anca Dumitrache @anca_dmtrch & Feng Lu @dsflu

Training process



Training process



Research Questions

1. What **model**?

2. What **data**?

3. What **features**?

`pytrec_eval`

Practical Lessons from Developing a Large Scale Recommender System at Zalando. Freno. RecSys 2017.

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.

dmlc XGBoost

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

Anca Dumitrache @anca_dmtrch & Feng Lu @dsflu

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?

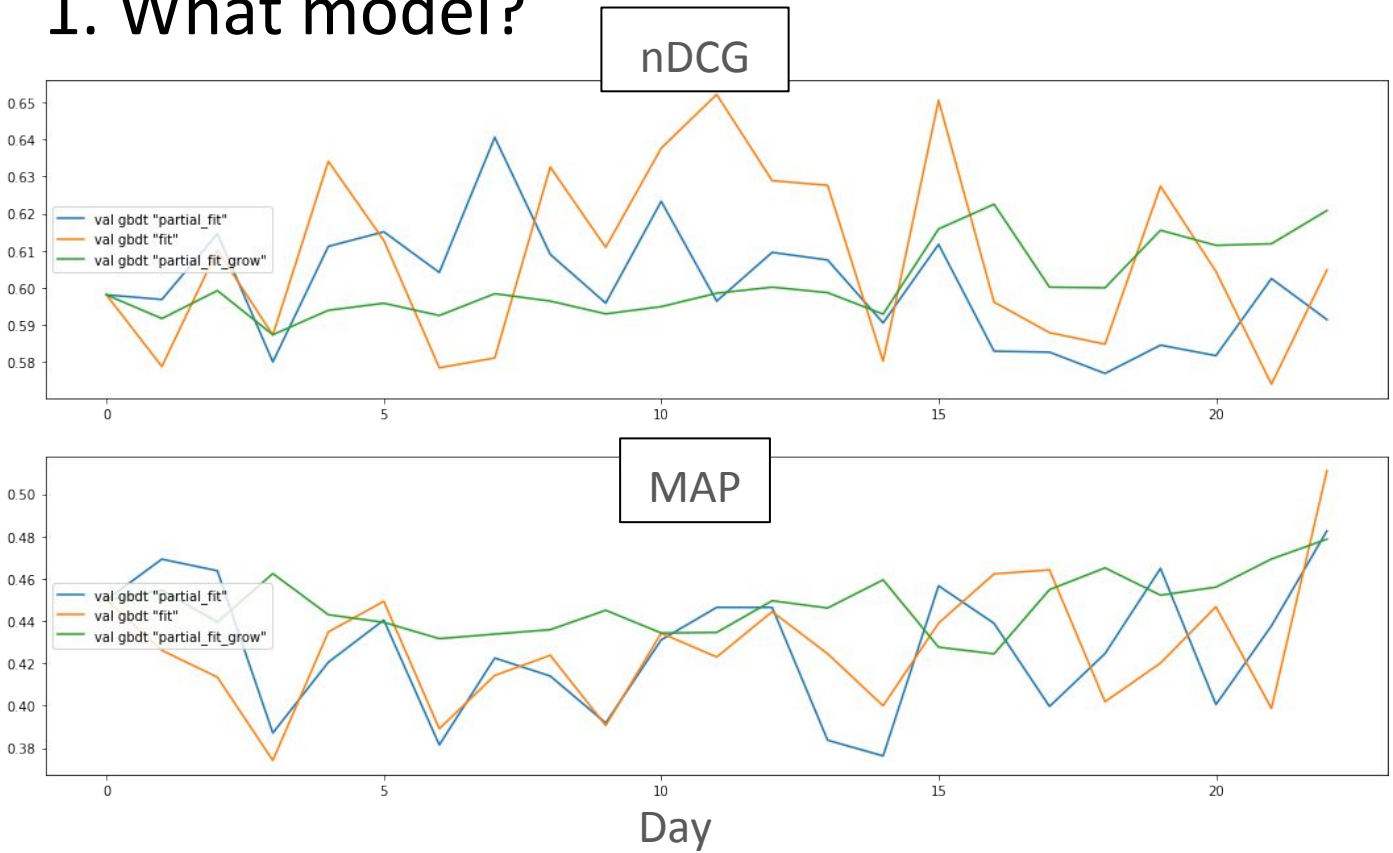
Deals with a heterogeneous mix of continuous, discrete, categorical features

Feature normalization is not required

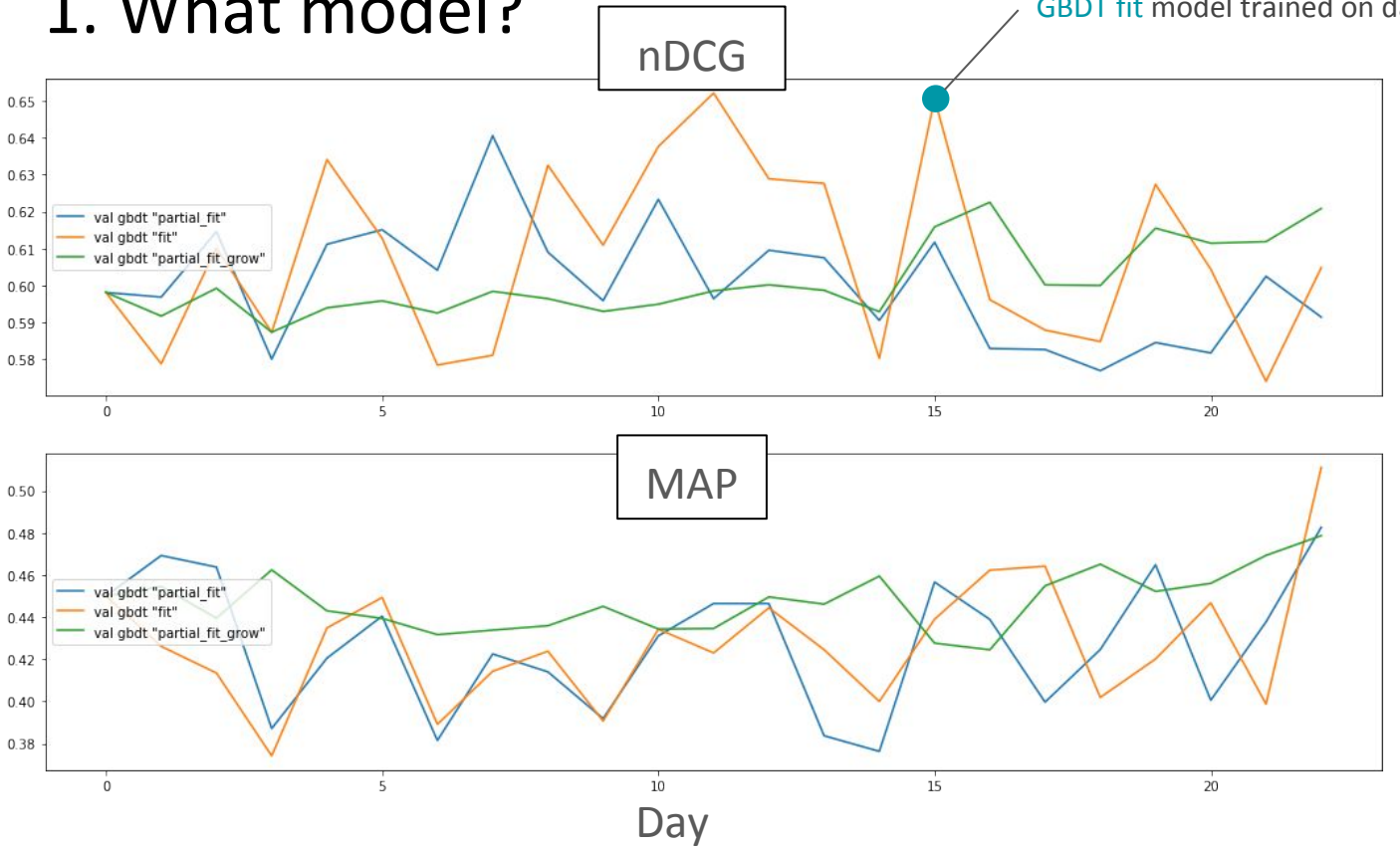
Feature selection is inherently performed during the learning process

Can easily capture non-linear, non-additive relations

1. What model?


nDCG
val gbdt "partial_fit"
val gbdt "fit"
val gbdt "partial_fit_grow"
0.65
0.64
0.63
0.62
0.61
0.60
0.59
0.58
0
5
10
15
20
MAP
val gbdt "partial_fit"
val gbdt "fit"
val gbdt "partial_fit_grow"
0.50
0.48
0.46
0.44
0.42
0.40
0.38
0
5
10
15
20
Day

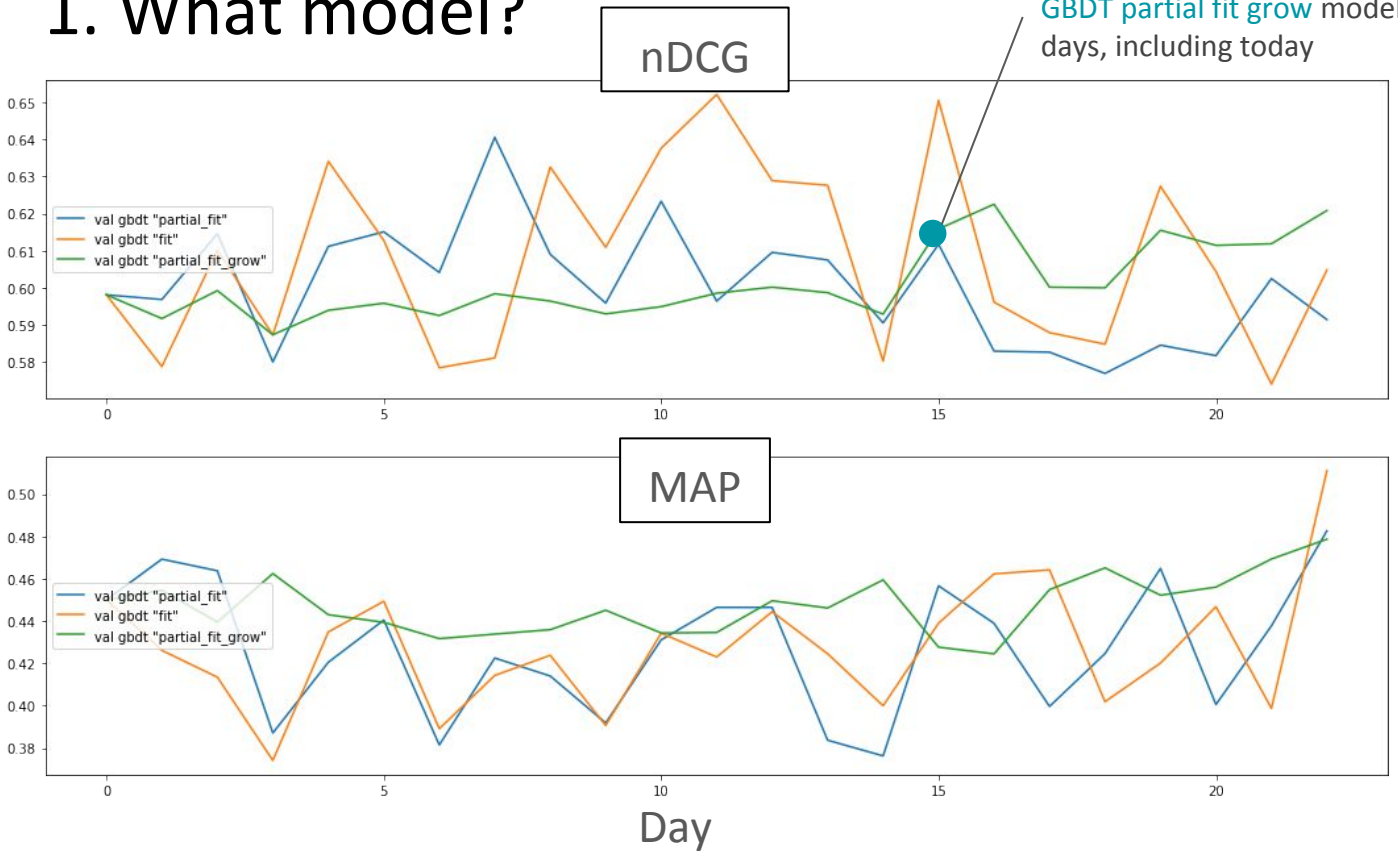
1. What model?



Anca Dumitrache @anca_dmtrch & Feng Lu @dsflu

1. What model?

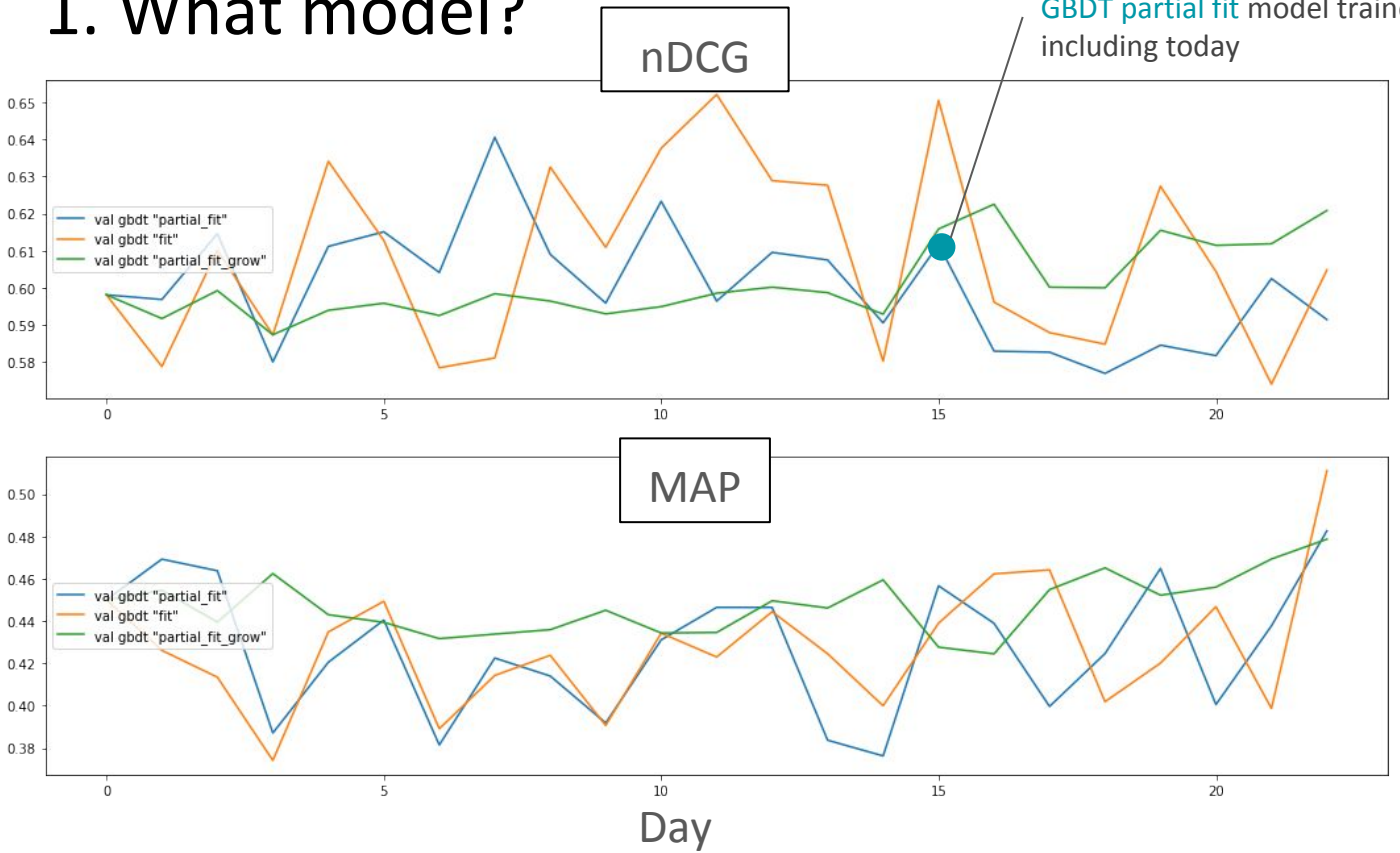
GBDT partial fit grow model trained on data from previous 7 days, including today



Anca Dumitrache @anca_dmtrch & Feng Lu @dsflu

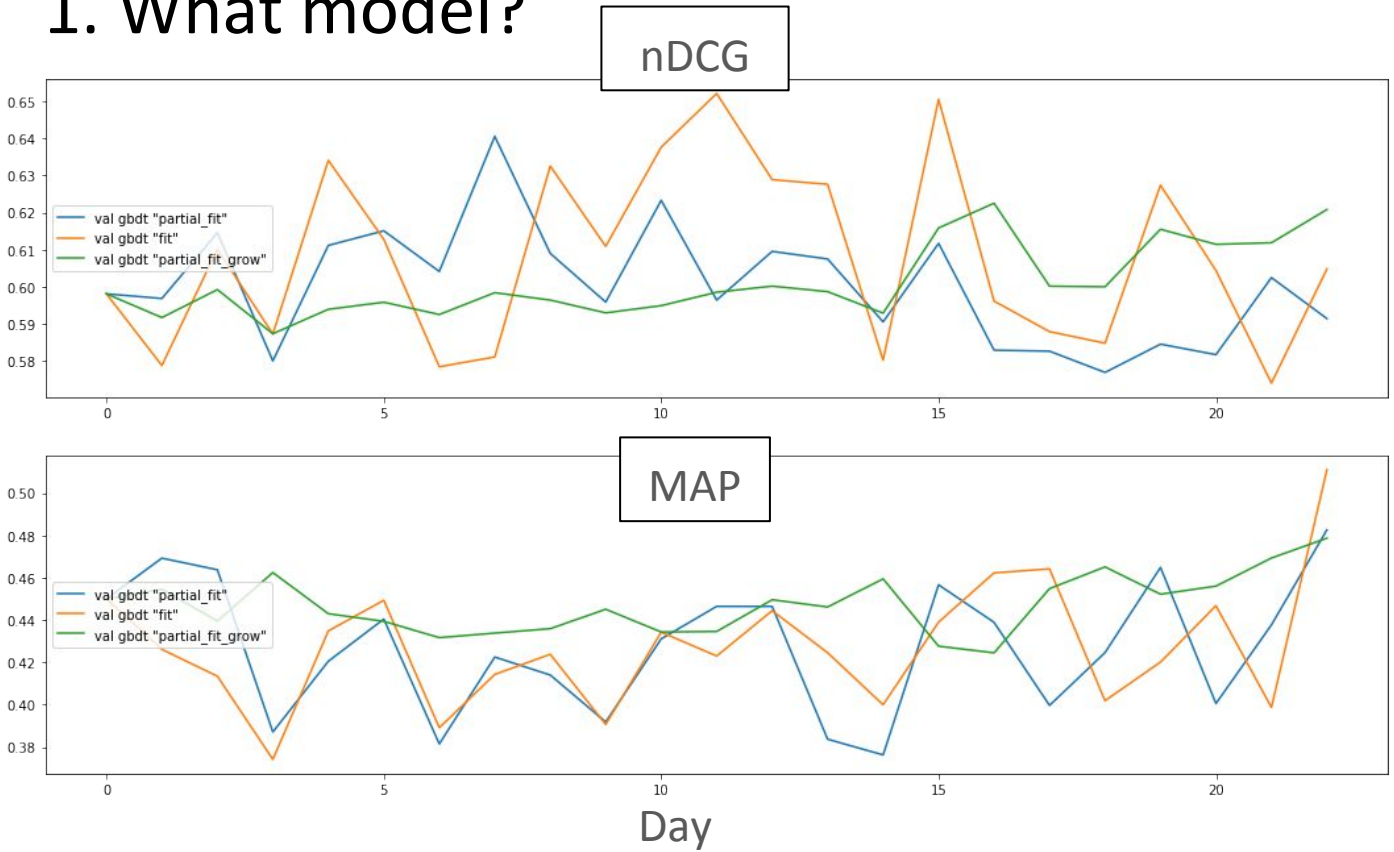
1. What model?

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



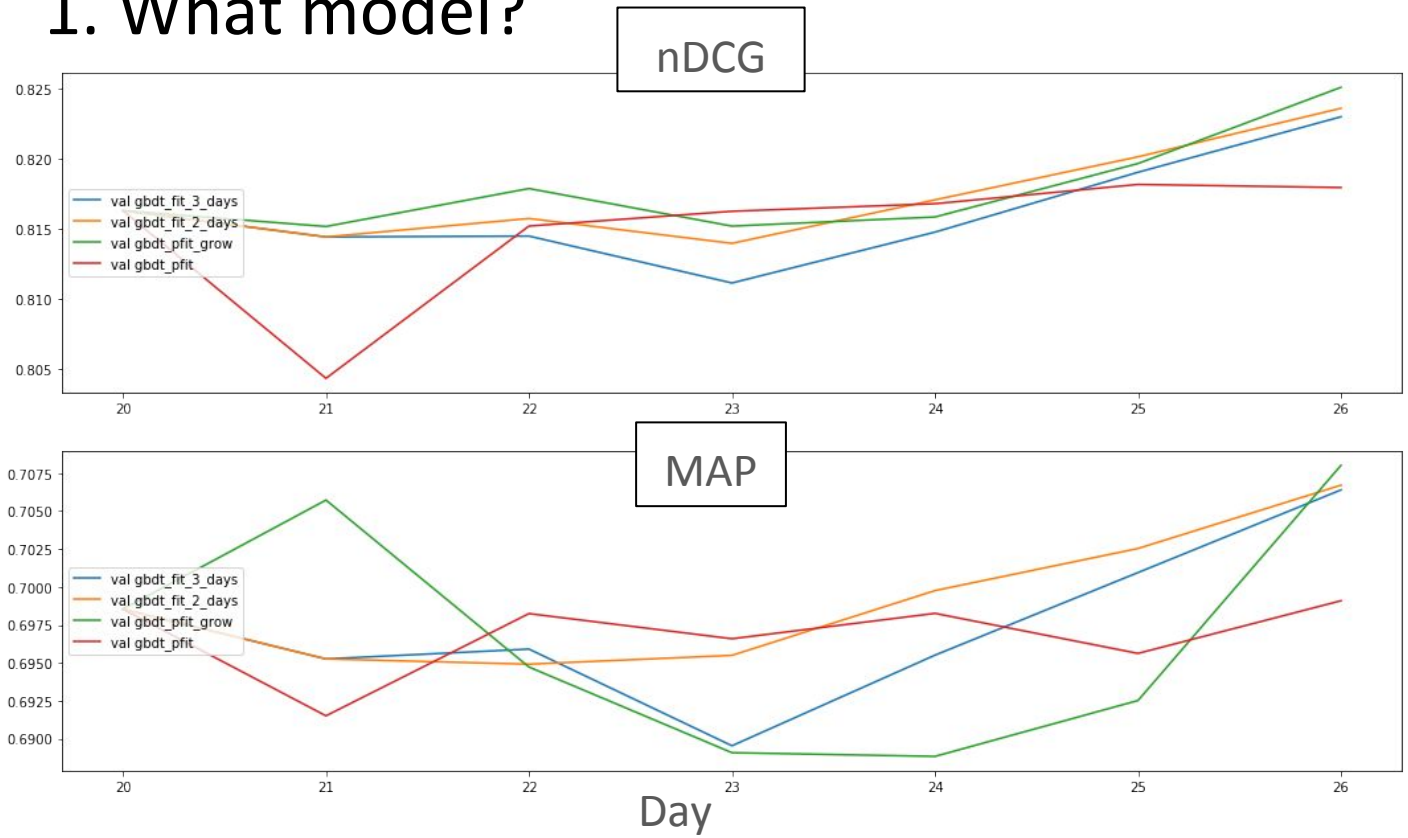
Anca Dumitrache @anca_dmtrch & Feng Lu @dsflu

1. What model?



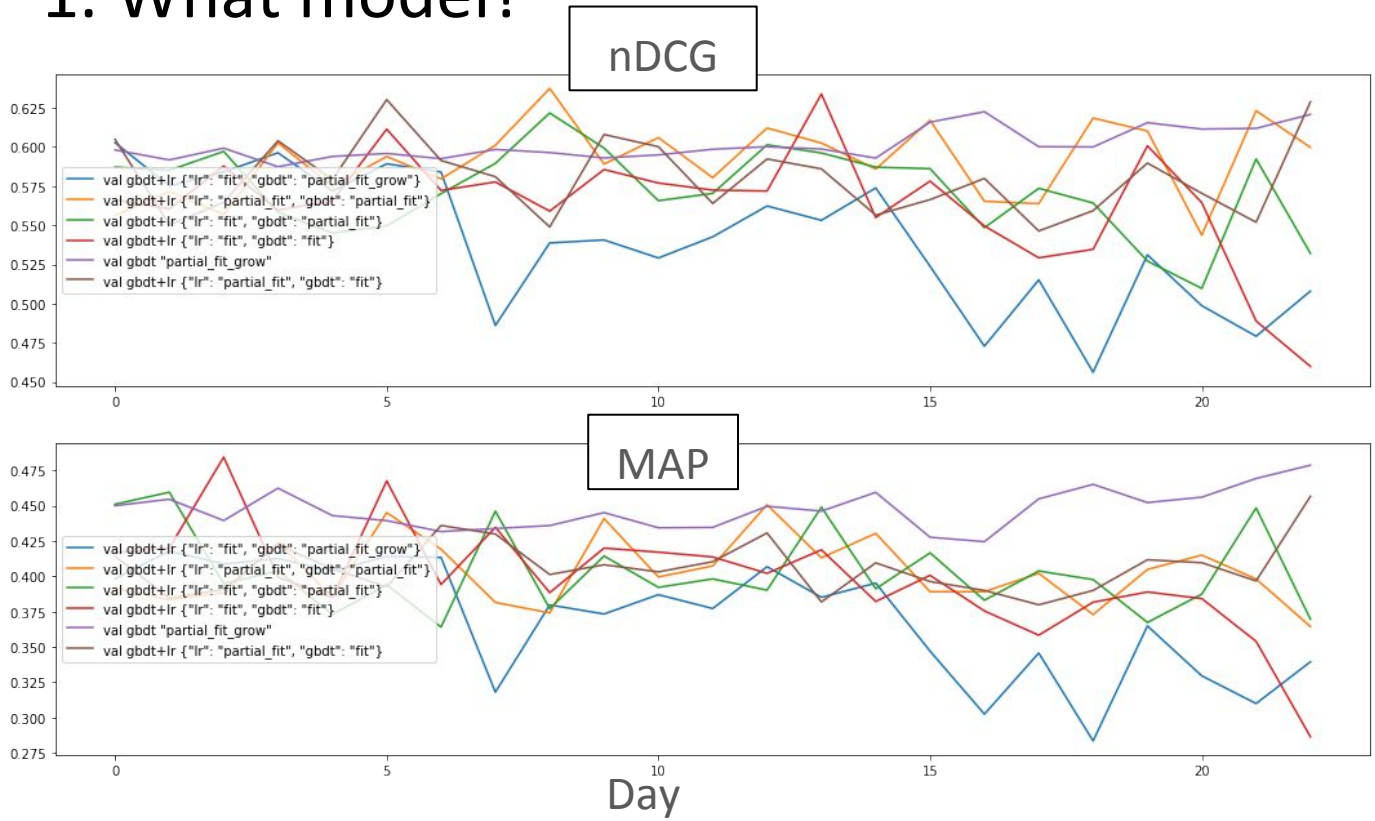
GBDT partial fit grow has the most stable performance, overall better evaluation results.

1. What model?



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

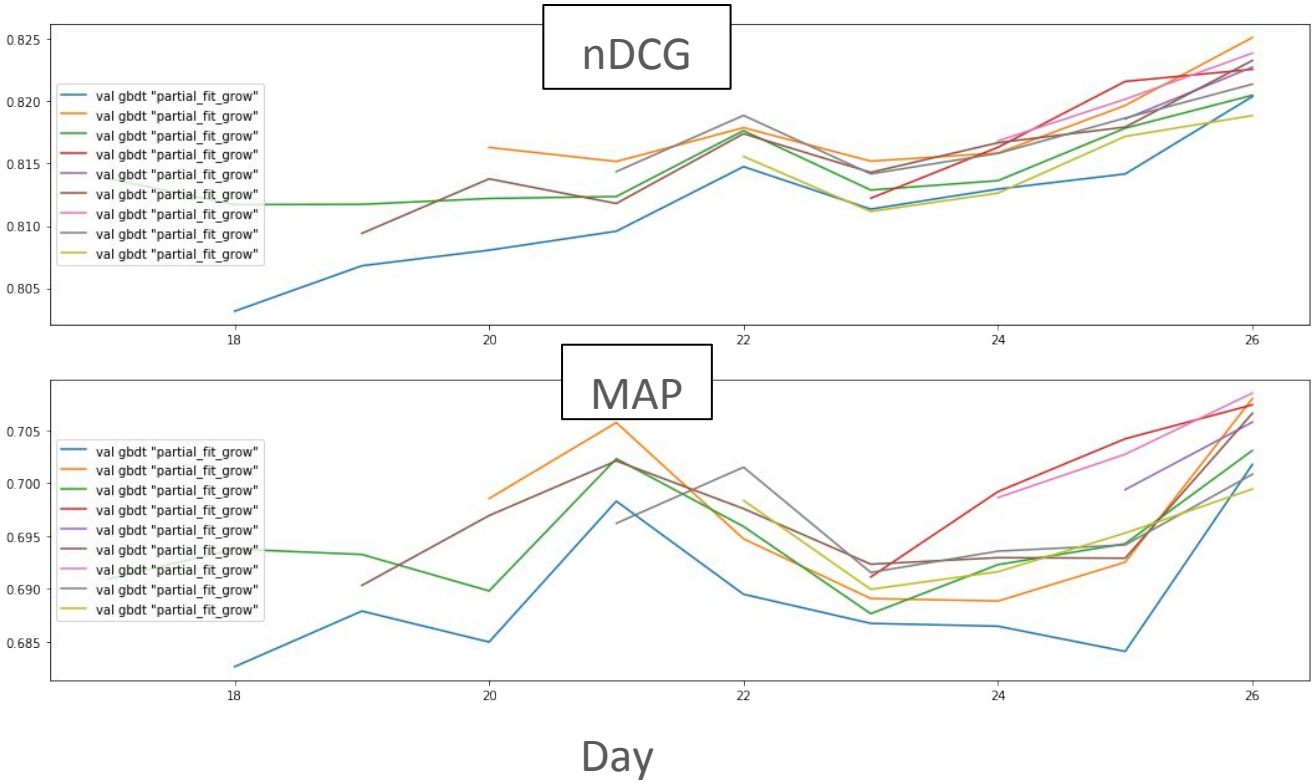
1. What model?



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?



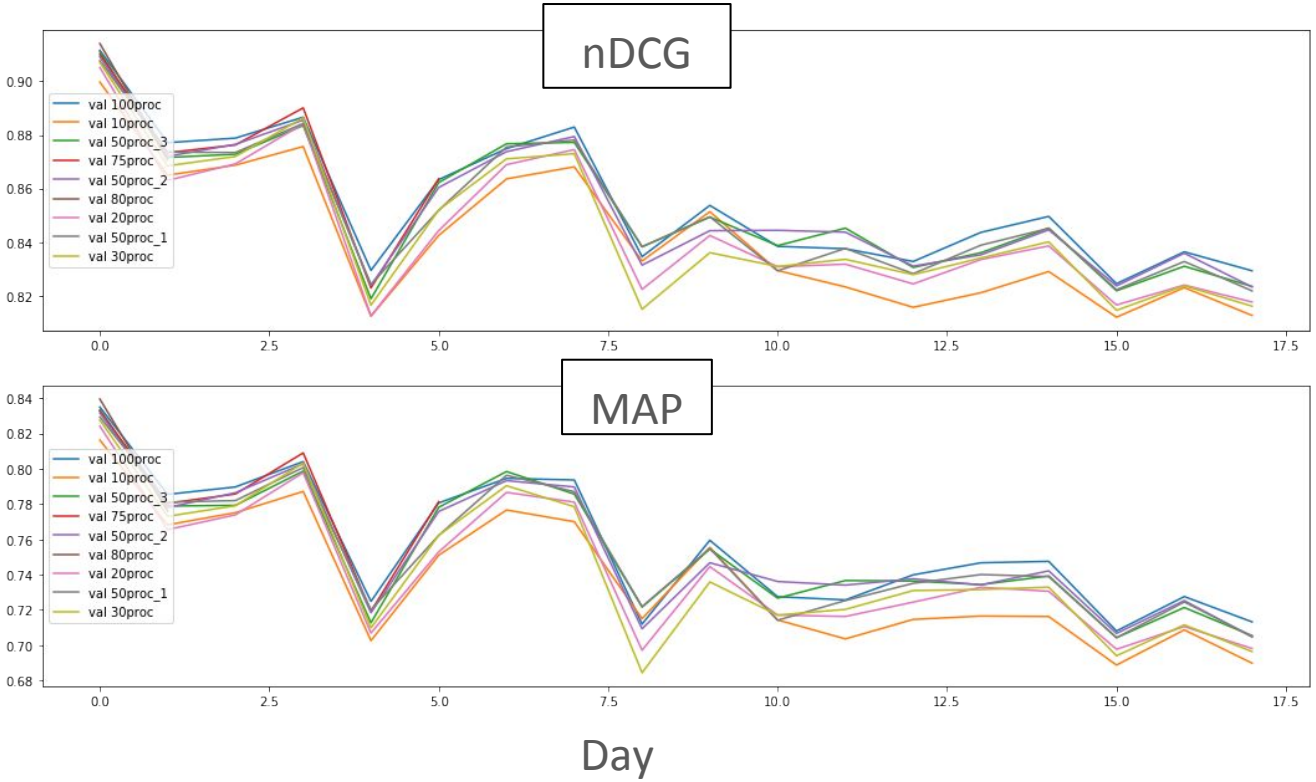
More data is not always better, because data recency is important.

Relation between number of days & performance is not linear.

3 & 7 days in train is best.

Anca Dumitrache @anca_dmtrch & Feng Lu @dsflu

2. Do we need all interactions?



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

Anca Dumitrache @anca_dmtrch & Feng Lu @dsflu

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

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, all features contribute a little.

Experiment conclusions

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

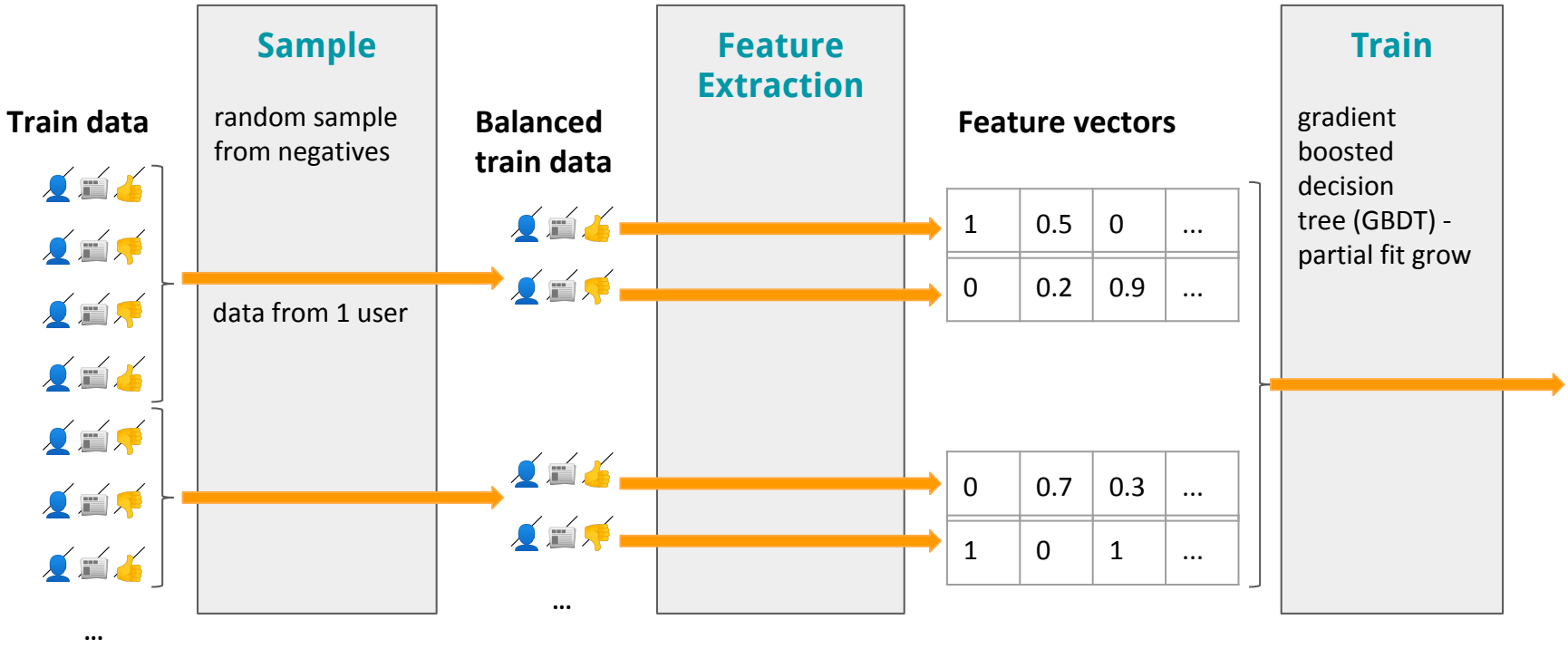
Less data (days, user interactions) does not mean worse performance.

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

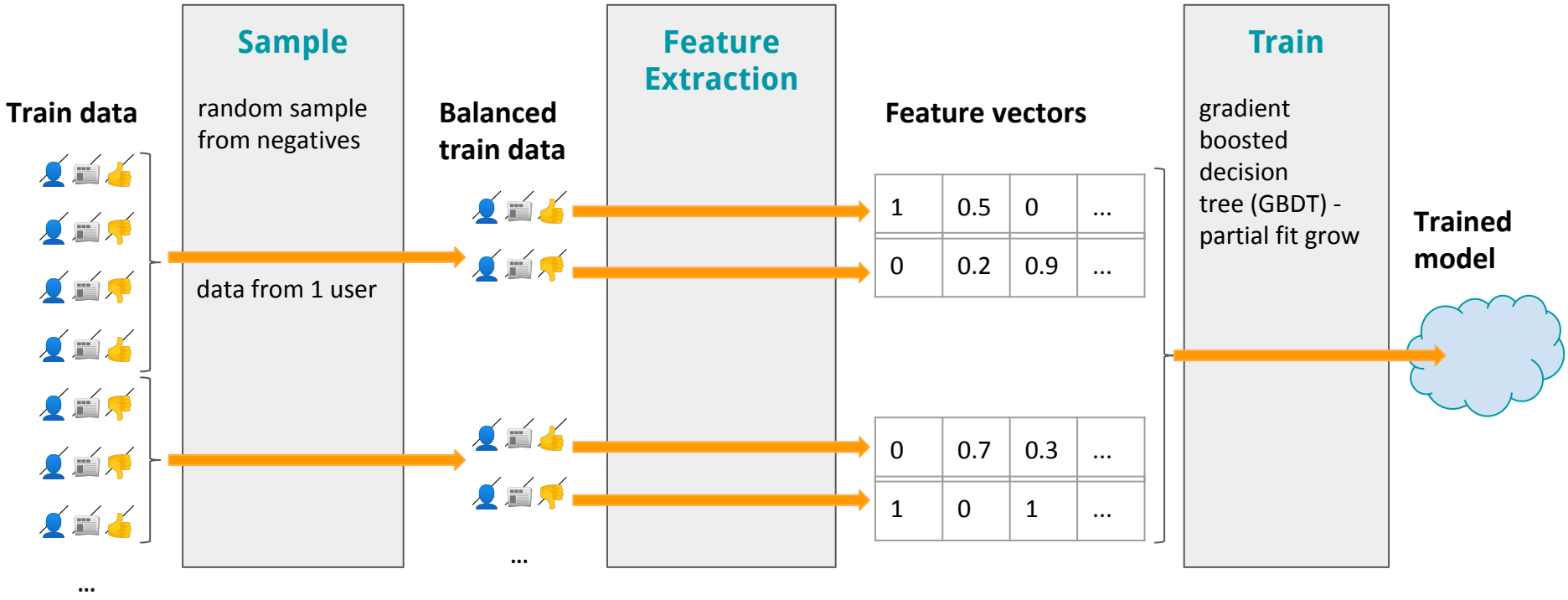
Practical tip

XGBoost does NOT support multi-threading.

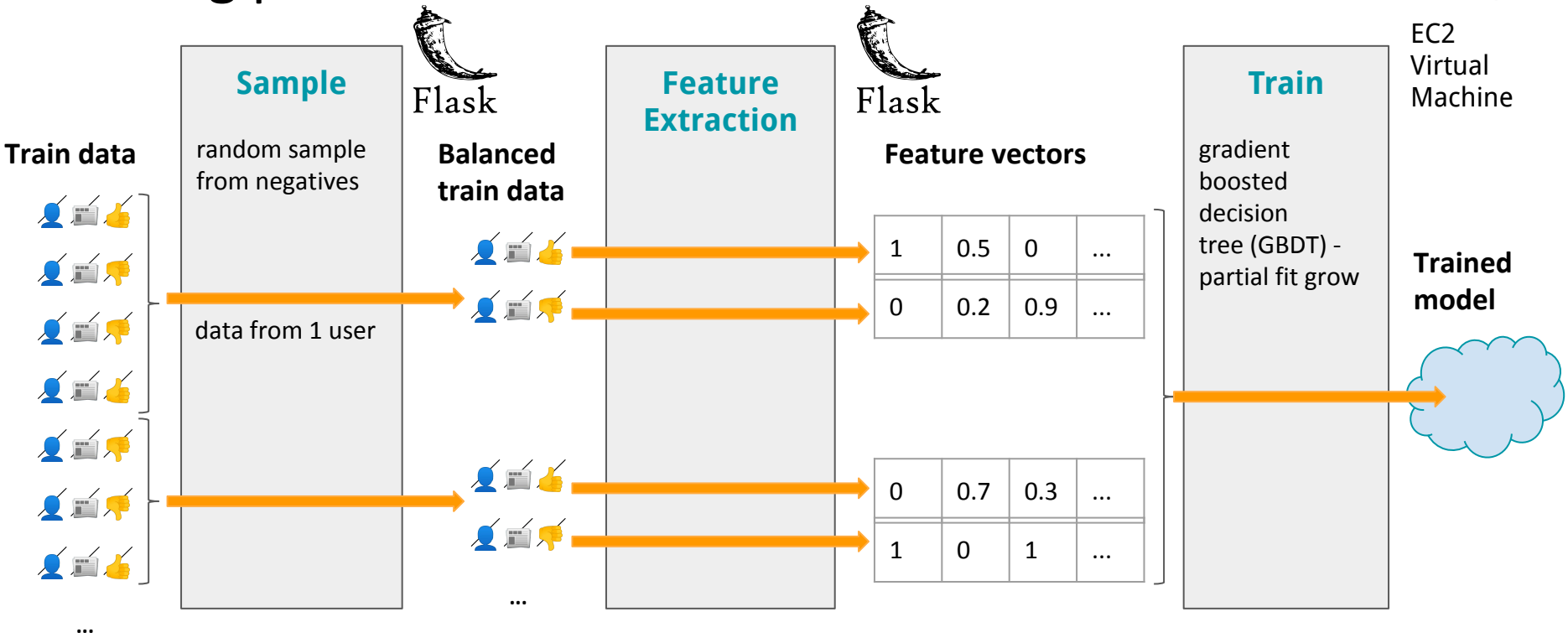
Training process



Training process



Training process

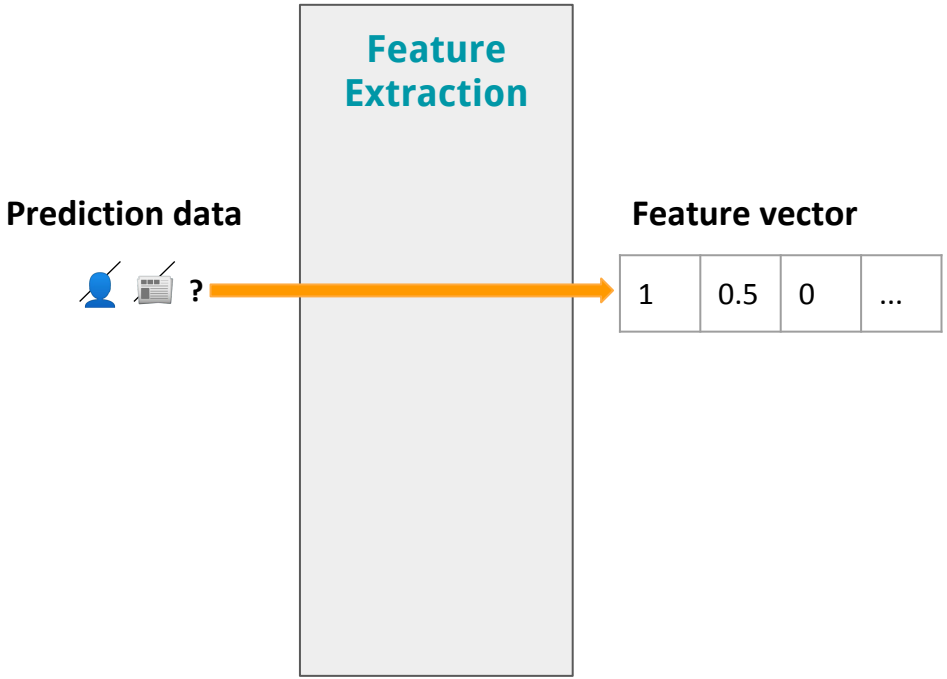


Prediction process

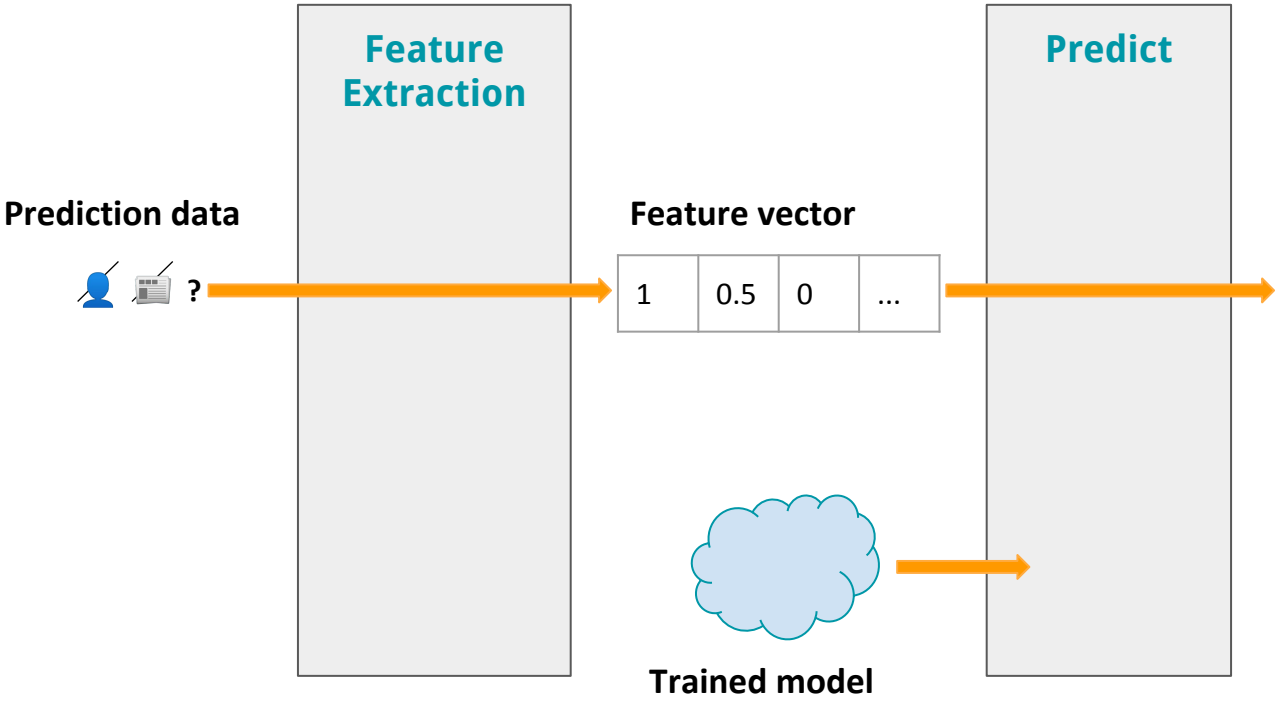
Prediction data

?

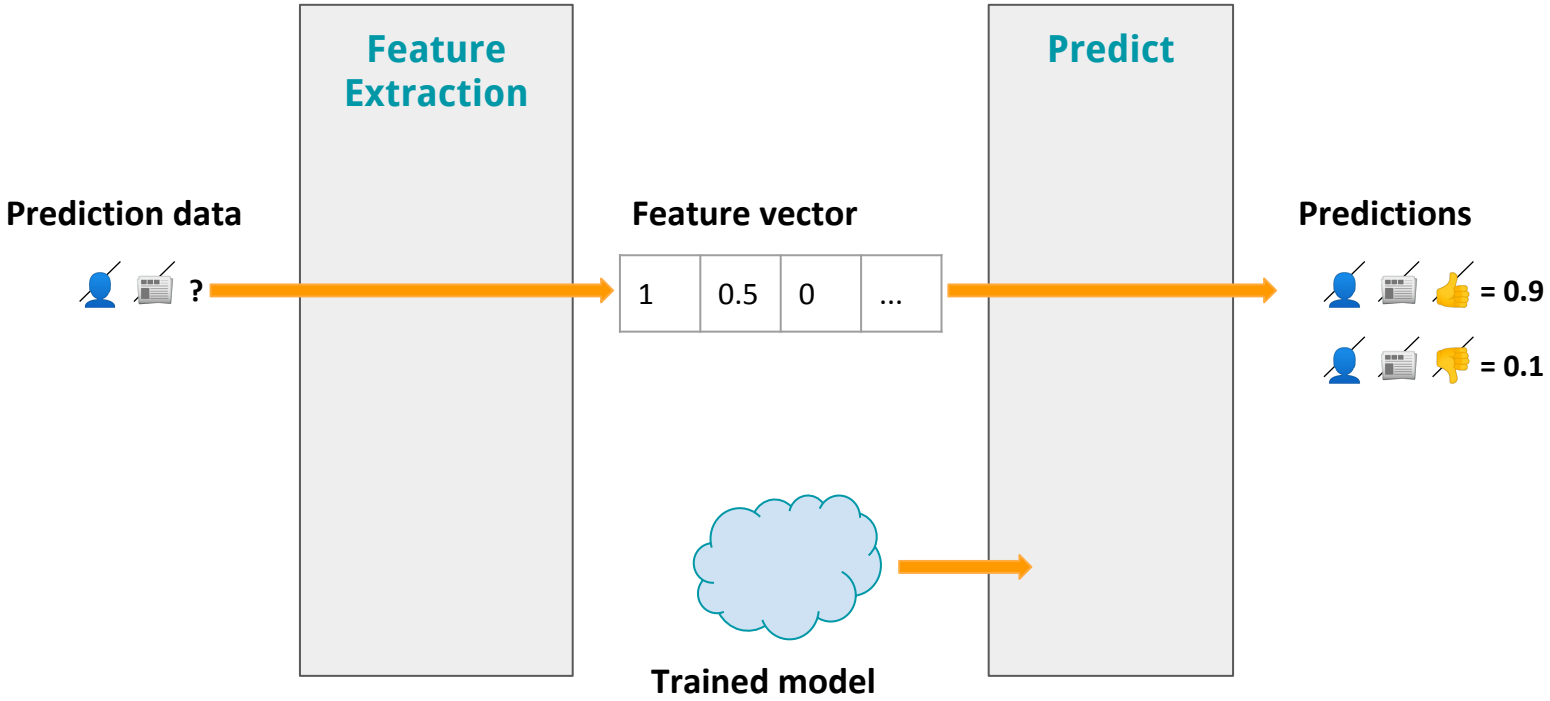
Prediction process



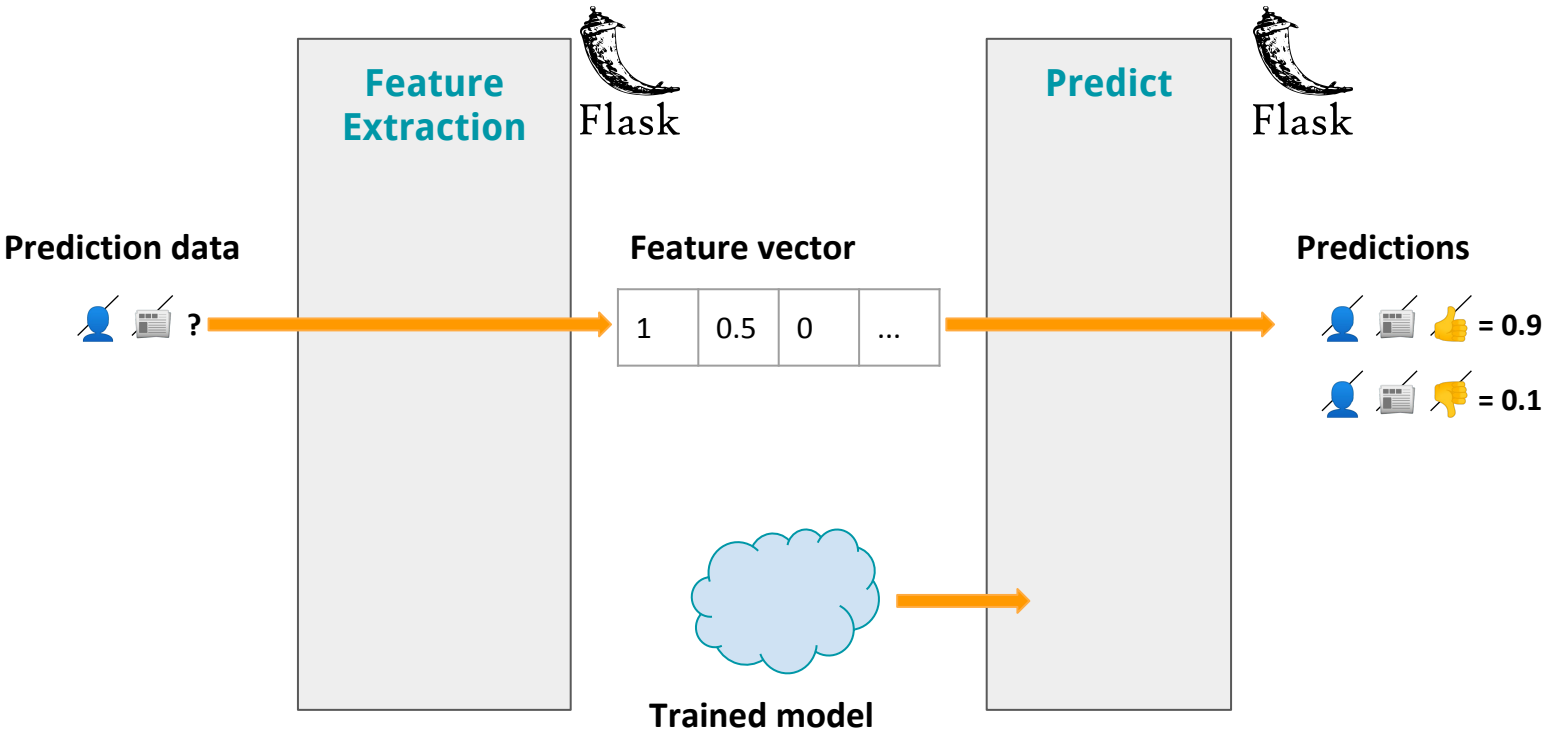
Prediction process



Prediction process



Prediction process



How it looks on FD.nl

articles from the past 7 days

articles from the past 24h

Anca Dumitrache @anca_dmtrch & Feng Lu @dsflu

Training schedule:

New model trained every day, starting at 11PM

➡️ stored in AWS S3 bucket

Predict schedule:

Once in the morning for all users

For each user, 1h after their last visit

➡️ cached in AWS DynamoDB

SENTRY

Building BNR Smart Radio & FD.nl Recommender system using Clojure - Bahadir Cambel (YouTube)

Ongoing work

Online testing - currently collecting data

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

Usefulness Preliminary Results

Coverage:

- Exposure of daily publications
- Highlighted articles v.s. Top-5 recommendation

Conclusion:

1. **At the individual user level:** personalized, less coverage
2. **Over the whole user set:** more coverage, broader articles

Diversity:

- How diverse a list of items in terms of `Section`, `Tags`, `Authors`, and `Content`
- Highlighted articles v.s. Top-5 recommendation

Conclusion:

1. Top-5 recommendations are more diverse on `Section` and `Content`
2. Highlighted articles are more diverse on `Authors`
3. `Tags` are hard to compare due to sparsity

Lessons learned

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, all features contribute a little.

Further reading

Building BNR Smart Radio & FD.nl Recommender system using Clojure - Bahadir Cambel. YouTube.

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

Practical Lessons from Developing a Large Scale Recommender System at Zalando. RecSys 2017.

Anne Schuth / Data Science at Blendle / Sanoma TechTalks23. YouTube.