Lookalike Modeling

A Cxense DMP Feature

Lookalike modeling

- Can be enabled for segments in a DMP
- Finds similar users to segment members
- Used to extend reach, for example to
  - Target users similar to those who already bought a subscription
  - Target a small set of known users based on 1st party data, and a bigger set of anonymous but similar users
- Percentage of audience to select is configurable

Lookalike modeling in Cxense DMP

- Powered by machine learning / AI
- Input data:
  - Pageview events
  - 1st party data
  - Content profiles
- Selects most similar users among all unique users
- No overlap between original segment and lookalikes

Guidelines for best results

- The best segments to enable lookalike modeling on are
  - Narrowly defined (matching a small subset of users)
  - Based on 1st party data or directly observable event characteristics
- Set an appropriate percentage, not too high (1 - 20% is fine)
  - Quantity vs quality tradeoff, higher percentage means less similar users
- Specify “negative segments” whenever it makes sense
  - Lookalike for women: Use men as negative segment
  - Lookalike for age: Use other age segments as negative segments

ML model #1: Cosine similarity

- Represent each user as word vector, based on consumed content
- Use logistic regression to find the set of most significant words
- Compute centroid (average vector) for all segment members
- Compute similarity between non-members and the centroid
- Output a ranked list of non-members, from most to least similar
- Special cases: demographic properties (gender, age, etc)

ML model #2: kNN-based classification

- Based on the metadata fields in pageview events
- Filtering to find the most relevant fields to use for model training
  - Few missing values, an appropriate value distribution, etc.
- Events are labeled as member/nonmember and used for training
- k Nearest Neighbors algorithm used to classify all events
- Nonmembers with high ratio of “member events” are lookalikes

Example of features selected by model #2

userId	site	sessionStart	connectionSpeed
sessionStop	sessionBounce	activeTime	isoRegion
userAgent	browser	browserVersion	exitLinkUrl
browserLanguage	browserTimezone	os	exitLinkHost
mobileBrand	deviceType	url	exitLinkQuery
host	query	referrerUrl	capabilities
referrerHost	referrerHostClass	referrerSearchEngine	adspaces
referrerSocialNetwork	referrerQuery	resolution	customParameters
colorDepth-	country	region	userParameters
city	metrocode	company	retargetingParameters

Darker yellow means more important feature. Feature selection is dynamic, i.e. the chosen set of features varies from segment to segment based on the input event dataset for that segment.

Quality

- We regularly evaluate model precision in an automated fashion
- Evaluation starts by splitting the set of segment members in two parts
  - The first for training the models
  - The other as a test set, together with a sample of non-members
- The models predict lookalikes from the users in the test set
- Precision is the ratio of segment members among predicted lookalikes
- All precision scores are compared to a baseline of random sampling
- The precision scores are used to ensure we deliver good quality results