Embeddings for Data Enrichment
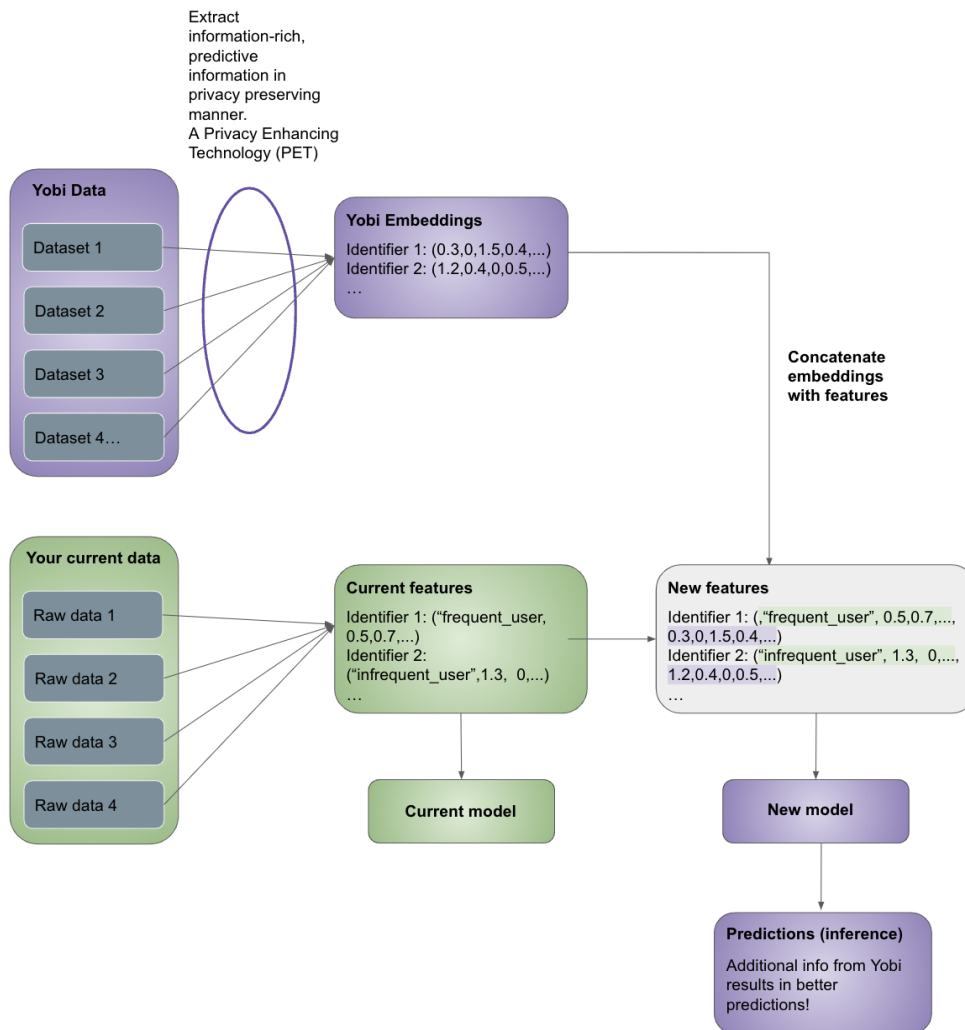
The following information is proprietary and confidential.
Do not share without Yobi consent.

Overview

Yobi's embeddings are generated on our diverse set of opt-in consumer behavioral data, which condenses information into a 200-dimensional vector for each individual user identifier. These identifiers can be matched to 1st party identifiers on the customer's side, providing the customer's company with additional rich behavioral information about their end customers.

Extract information-rich, predictive information in privacy preserving manner.
A Privacy Enhancing Technology (PET)

Yobi Data
Dataset 1
Dataset 2
Dataset 3
Dataset 4…

Yobi Embeddings
Identifier 1: (0.3,0,1.5,0.4,...)
Identifier 2: (1.2,0.4,0,0.5,...)
…

Concatenate embeddings with features

Your current data
Raw data 1
Raw data 2
Raw data 3
Raw data 4

Current features
Identifier 1: ("frequent_user, 0.5,0.7,...)
Identifier 2: ("infrequent_user",1.3, 0,...)
…

New features
Identifier 1: (,"frequent_user", 0.5,0.7,..., 0.3,0,1.5,0.4,...)
Identifier 2: ("infrequent_user", 1.3, 0,..., 1.2,0.4,0,0.5,...)
…

Current model

New model

Predictions (inference)
Additional info from Yobi results in better predictions!

Data Enrichment

By leveraging a large set of behavioral data sets, we generate embeddings, which can enrich Customer's current set of data. The behavioral data includes diverse information about individual consumer behavior including financial transactions, location information, CTV watching, web browsing data, and more. These embeddings are intended to be joined to Customer's features by identifier and concatenated to them, so that the resulting larger set of features per end user can be fed into Customer's models for improved performance.

Predictive modeling

To generate our embedding, we create an enormous graph of the majority of the US populations and their behaviors across our diverse collection of datasets. We then treat the problem as a large graph link prediction task, where we hold out future behavior. We then train deep learning models on this graph to create a user embedding that is predictive of the individual held out behaviors. The embeddings compress the space of behaviors so that users that are close in the latent space are likely to behave similarly. The resulting embeddings are therefore predictive out-of-the-box, adding extra predictive power on the end consumer, beyond what our customers have organically.
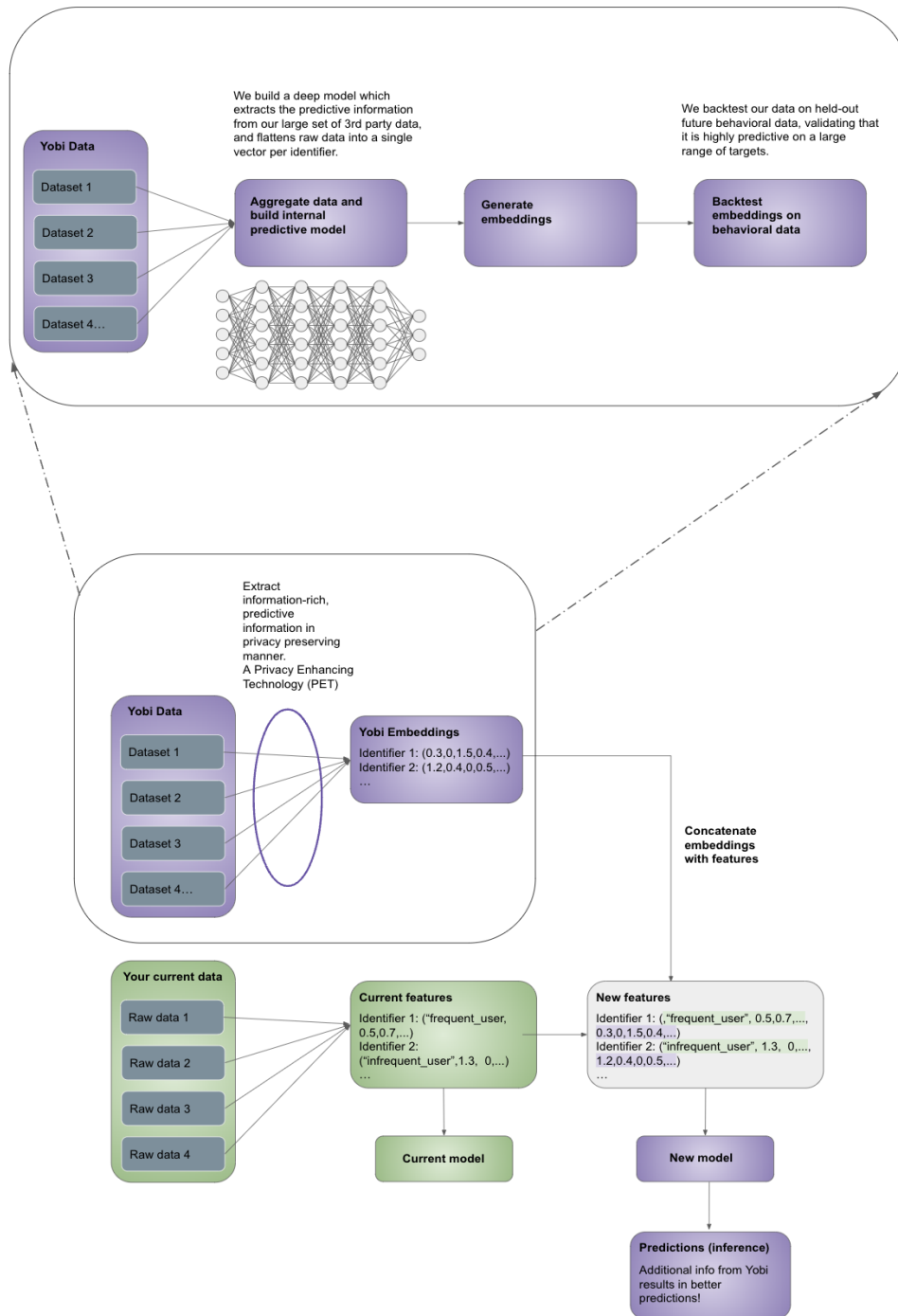
Highly predictive

These models

1. aggregate billions of events across behavioral datasets into one embedding vector per identifier

2. pre-train a representation to make it highly predictive of future behavior without containing sensitive or extraneous signal

The result is that these embeddings are much more predictive then if one were to purchase the 3rd party data and join it to their own feature set with a simpler model.

Privacy preserving

By reducing the dimensionality of these large data sets into a single 200-dimensional vector per identifier, we also preserve the privacy of users, making them less interpretable, and anonymizing the raw 3rd party event-level information. This is often called a Privacy-Enhancing Technology (PET), and protects information of users, as well as de-risks the client's company from having the event-level information about users on their system.

Detailed view of Yobi Embeddings framework

We build a deep model which extracts the predictive information from our large set of 3rd party data, and flattens raw data into a single vector per identifier.

We backtest our data on held-out future behavioral data, validating that it is highly predictive on a large range of targets.

Yobi Data
- Dataset 1
- Dataset 2
- Dataset 3
- Dataset 4…

Aggregate data and build internal predictive model

Generate embeddings

Backtest embeddings on behavioral data

Extract information-rich, predictive information in privacy preserving manner.
A Privacy Enhancing Technology (PET)

Yobi Data
- Dataset 1
- Dataset 2
- Dataset 3
- Dataset 4…

Yobi Embeddings
Identifier 1: (0.3,0,1.5,0.4,…)
Identifier 2: (1.2,0.4,0,0.5,…)
…

Concatenate embeddings with features

Your current data
- Raw data 1
- Raw data 2
- Raw data 3
- Raw data 4

Current features
Identifier 1: ("frequent_user, 0.5,0.7,…)
Identifier 2: ("infrequent_user",1.3, 0,…)
…

New features
Identifier 1: (,"frequent_user", 0.5,0.7,…, 0.3,0,1.5,0.4,…)
Identifier 2: ("infrequent_user", 1.3, 0,…, 1.2,0.4,0,0.5,…)
…

Current model

New model

Predictions (inference)
Additional info from Yobi results in better predictions!

FAQ

- **Q: What is an embedding?**
  - A: An embedding is a low-dimensional vector of features for an object, in this case an identifier. Our vectors are 200 dimensions, meaning that for each identifier there is a vector such as (0.3, 1.5, 0.7, …) containing 200 numbers (more specifically, doubles) which are predictive of that identifier's future behavior.
- **Q: How is using Yobi Embeddings different than just incorporating 3rd party data**
  - **Privacy-preserving:** our embeddings are condensed from event-level data to uninterpretable 200-dimensional vectors per identifier
  - **Pre-trained and highly predictive:** our embeddings are created using deep learning models which distill the information into a set of 200 dimensions which are highly predictive of future user behavior. These models have been extensively iterated upon by our machine learning engineers and back-tested on held-out future behavior, validating their predictive ability.
  - **Easy-to-use:** our embeddings condense the event-level data across a large set of 3rd party datasets into a single embedding per identifier. These embeddings can simply be concatenated to the client's features. There is no need to generate features from the 3rd party data sets or to find a way to aggregate the information into an identifier-level format.
- **Q: What is an identifier?**
  - An identifier is a mostly unique ID for a user which can be used to join our Yobi embeddings to your features. For example, this could be a hashed email address, device ID, or Experian ID.