Expanding Spanish Language Audiences

Sergey Fogelson
TelevisaUnivision
Date

Edouardo Vitale
TelevisaUnivision
Date

Expanding Spanish Language Audiences Through a Custom Lookalike Modeling Architecture

Sergey Fogelson
Edouardo Vitale

04.25.2023

TelevisaUnivision

Motivation

Misidentification

4 in 10 Hispanics are excluded from 3p datasets

Waste

70% of impressions targeted at Hispanics are wasted

Scale

The true scale of the Hispanic population within a given brand’s 1p dataset is hard to identify without extensive validation

Data Sources

1P VIEWERSHIP

- ViX
- UNow
- Digital Apps
- Univision.com

3P VIEWERSHIP

- AVOD platform video impressions
- SVOD/VMVPD Programmatic video impressions

DEMOGRAPHICS

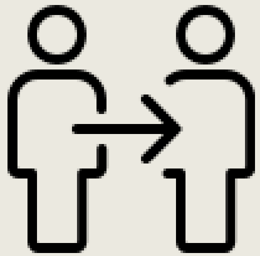
Age/gender/income/education taken from our spine partner

HOUSEHOLD GRAPH

Combine these 3 sources to create robust household-level representation of ~17MM Hispanic households in the US

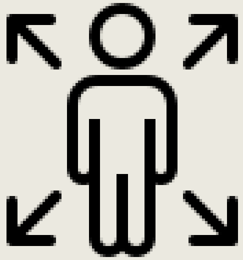
Expanding An Audience With Look-alike Modeling

WHAT IS LAM?



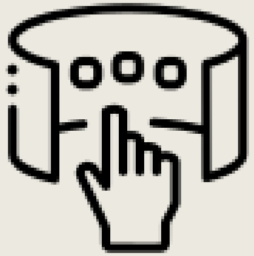
Look-alike modeling (LAM) is a process that helps identify individuals who look and act just like a given target audience

WHAT IS IT USED FOR?



Look-alike models are used to build larger audiences from smaller segments to create reach for marketers and advertisers

HOW DOES OUR LAM WORK?



Embeddings are used to create a latent space that allows for comparing user similarities, very similar users have (mathematically) very similar embeddings

Embeddings

What is an embedding?

Embeddings are low dimensional, learned continuous vector representations of discrete variables

Technical Level

Embeddings are representations of content or users that help describe viewing similarities numerically

High Level

How are embeddings created?

Embeddings are extracted from the parameters of a neural network which are adjusted to minimize loss on a content viewing prediction task.

Technical Level

We find similarities between individuals based on:

1. What content they viewed
2. Where the content was viewed (zip code)
3. Demographics of the viewer

High Level

TelevisaUnivision

Example Data and Embeddings

INPUT

Title Embedding Input

DEVICE ID	TITLE INPUT
ae5c55a7-c7db-e8b3-490f	[‘Lamentos’, ‘Liga MX’,’La Rosa’]

Zip Code Embedding Input

ZCTA	TITLE INPUT
43220	[‘El Privilegio de Amar ‘Las Amazonas’, ’La Rosa’]

Demographic Embedding Input

DEMOGRAPHIC GROUP	TITLE INPUT
[‘Male’, ’18-24’, ‘< 100K income’, ‘Highschool Graduate’]	[‘Tom & Jerry’, ‘ESPN’, ‘Law and Order’,’Godfather’]

OUTPUT

Title Embedding Output

Title	Vec_0	Vec_1	Vec_2	Vec_3	…	Vec_19
Liga MX	-1.67986	-0.82079	2.30146	0.06019		-1.45798

Zip Code Embedding Output

Title	Vec_0	Vec_1	Vec_2	Vec_3	…	Vec_19
33136	-1.59740	-0.67312	0.97731	1.09124		-1.86311

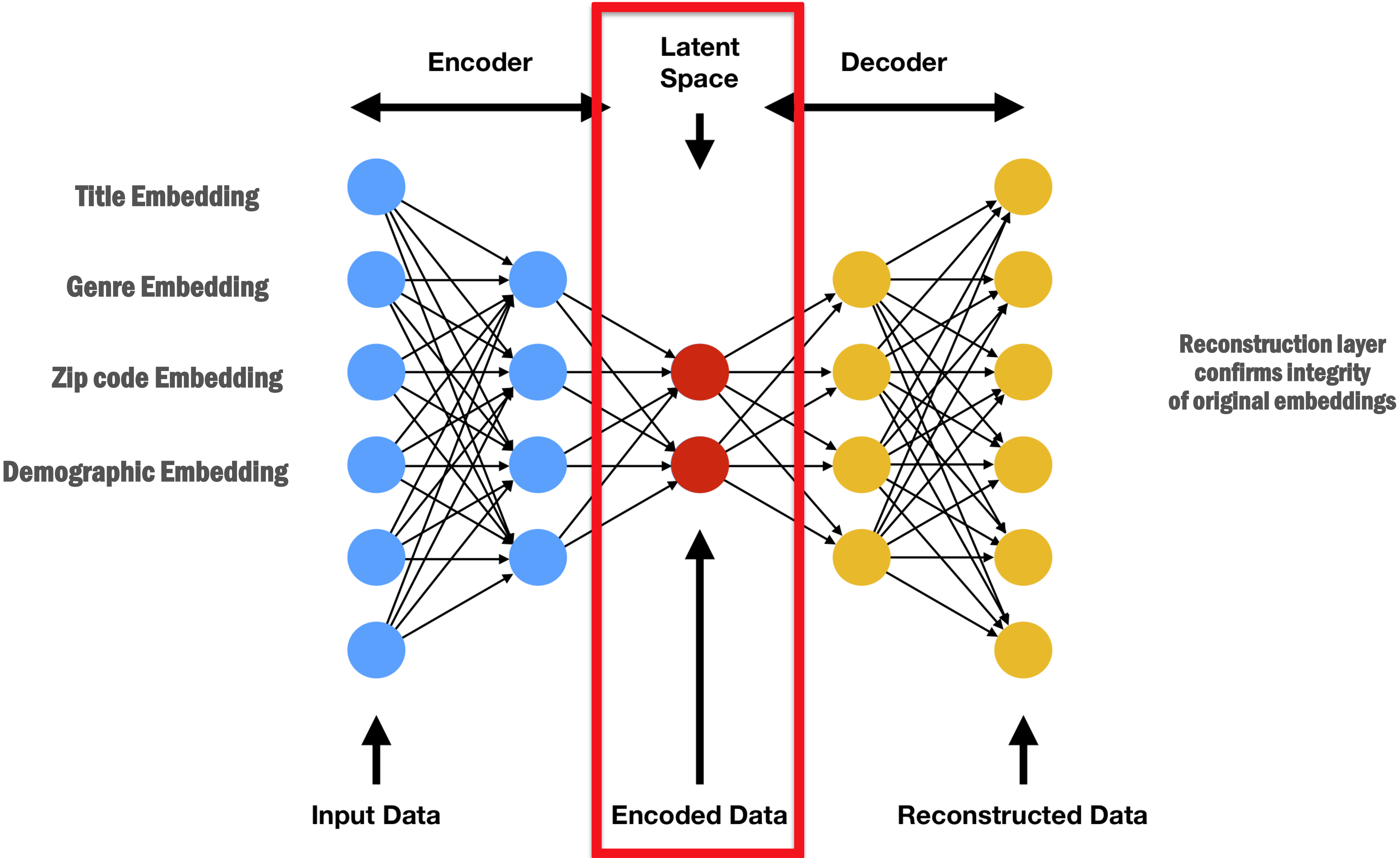
Demographic Embedding Input

Title	Vec_0	Vec_1	Vec_2	Vec_3	…	Vec_19
[‘Male’,’18-24’,‘< 100K income’,‘Highschool Graduate’]	-1.59740	-0.67312	0.97731	1.09124		-1.86311

Embeddings Allow Us to Perform Title Math

TITLE 1		TITLE 2		OUTPUT
A Dog Named Christmas	+	A Summer Romance	›	• A Very Merry Brides maid • Jingle Bell Bride
El zorro, La Espada y La Rosa	+	Silwood School of Cookery: How to Debone a Chicken	›	• Los Chefs • Top Chef Mexico

Autoencoder Architecture



Expanding Our Target Audience with LAM: Process

Create content/geo/demo embeddings,
Use autoencoder

Generate embeddings of titles/zipcodes/demographic attributes, combine with autoencoder

Generate aggregate seed audience

Average the auto-encoded embeddings for all identifiers in a seed list

Calculate Expansion Leveraging household graph

Find the most similar other identifiers in our graph to the aggregated audience and find all other identifiers in those households within our graph

How do we Validate?

01

Generate an audience identifier list
- First Party video audience
- Third Party video audience

02

Select a random seed list of ~1K ids from the audience and average their embeddings into an aggregate seed vector

03

Calculate N percent most similar identifiers to aggregate seed vector over our household graph

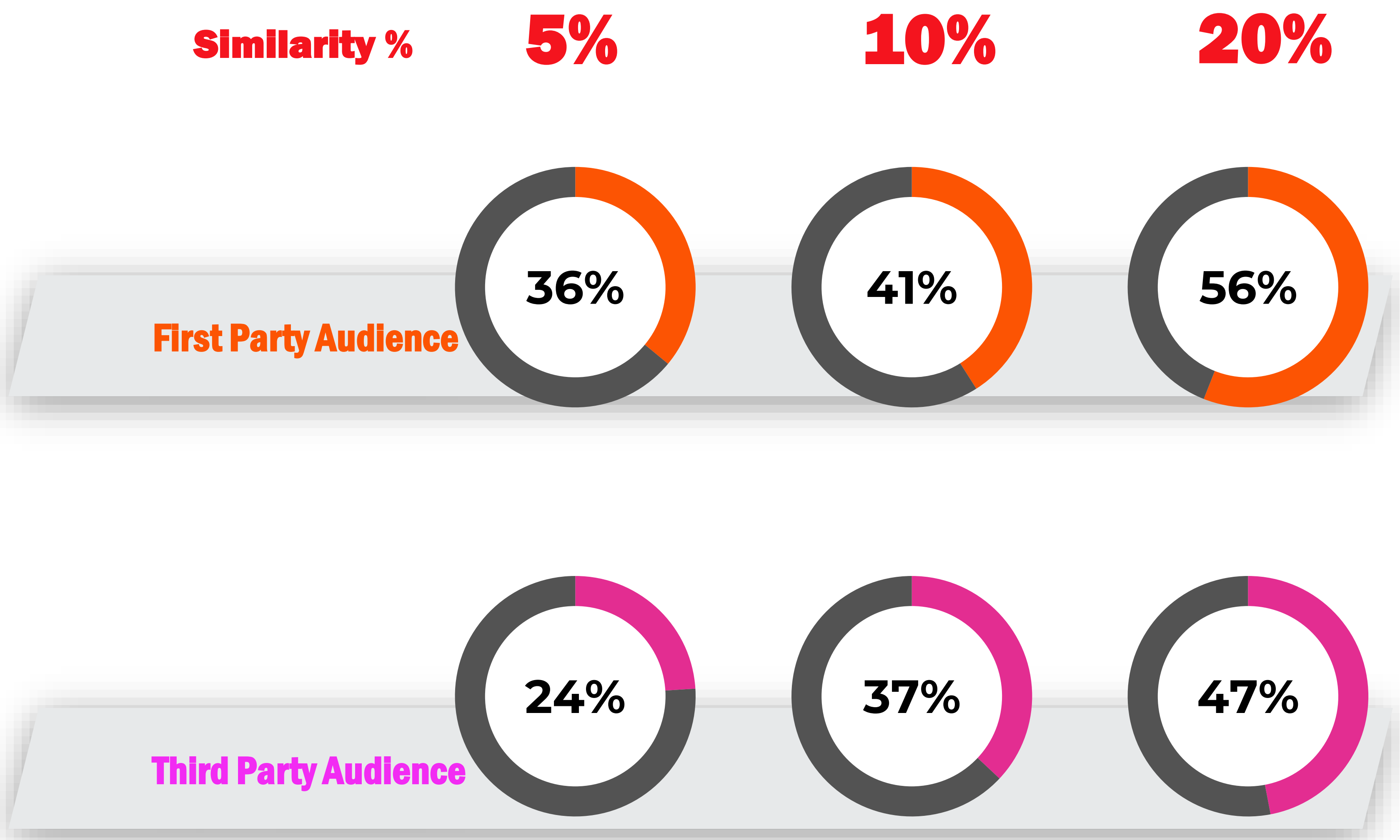
04

Calculate % of original audience list found in top N percent similar identifiers list

Reasoning: Selecting ids at random should give a proportional % of ids in the top percentile in relation to list size.

E.G. A random list of ~10% of our most similar devices should contain ~10% of the original identifier list. Lift is shown by having a larger proportion of the original audience in the Lookalike audience list.

How do we Validate?: Results



Televisa Univision

Expansion – Resulting Audience Sizes

DETERMINISTIC DEFINITION

Propagating IDs through graph to household level and finding all household-level identifiers, based on having a given number of matched ids

Deterministic	8K	203K	1.7MM
Seed Audience Size	1k	20k	150k
LAM	411K	1.5MM	2.7MM
Similarity, 1k seed	1%	5%	10%

TelevisaUnivision

Conclusions

Misidentification

Leveraging our graph, we can reliably identify what identifiers in a 1p audience are likely to belong to Hispanic media consumers

Waste

LAM reliably finds other identifiers in our graph that are very similar to the 1p audience based on zip/demo/geo viewing similarities

Scale

LAM + our graph achieves significant increases in overall audience scale, allowing brands to maximize in-target reach within the Hispanic population.