Who is GroupM?

GroupM is a leading global Media Agency. Advertising agencies make ads, but GroupM is focused on showing the right ads to the right people at the right time, and to this end has become a leader in data-driven solutions.

What are [m]Clusters?

[m]Clusters are segments of the population that are defined by particular online behaviors. GroupM clients can score their website’s visitors against these segments, and can also target these segments for future advertising.

Automated Data Cleaning

Our datasets had bad users, which we can either attribute to online bots or to failures of the semantic engine. Including these bad users hurt our clustering results. To make our entire process replicable, however, we couldn’t leave behind any steps that centered around our ability to find and judge outliers. Our automated data cleaning notebook uses elastic dimensionality reduction with MCA and DBScan clustering to automatically detect outliers.

Dimensionality Reduction

High-dimensional data and binary data are both poorly suited to clustering analysis. The first suffers from the curse of dimensionality, and the latter makes distance calculations difficult. So, we used PCA to compress the user interest data and convert binary interests into continuous features. The curse of dimensionality – that the distance measurements converge in high dimensions – is not just a theoretical problem. Its business consequence is too many users placed in a “leftover cluster.”

Clustering

The next was to perform agglomerative hierarchical clustering with linkage determined by Ward’s Method. We kept the first 30 branches of these trees.

Tree Pruning

“Choosing K” is always a difficult step in clustering, and we included business experts for this step of the process. In this business context, it’s a question of aggregation. Which splits of the tree separated distinguishable audiences, and which split a single audience into two? The importance of this question centers on our ability to sell the segments to GroupM client managers, and on their ability to sell the segments to their clients. So, we asked local experts in each country to help us choose the level of aggregation for their market.

INTEGRATION & RESULTS

Final Segments

One metric of importance to GroupM is the ability of their bidding and insights engine to identify the right users from the population to match a segment. They evaluate this with a statistic called GRP, which measures the efficiency of the lookalike models. It can be thought of as similar to recall, except that smaller GRP numbers mean that the models are more successful. MIT segments outperformed nearly all benchmark models.

Lookalike Results

We activated our segmentation for three different clients, who each selected certain segments to target. These selections were based on segments “over indexed” for their current customers, and the results shown below demonstrate the monetary value of [m]Clusters.

[m]Clusters

AUDIENCES FIRST

Copenhagen, Denmark

GroupM Mentor: Kristian Brødreskift

Will Fein and Gerard Woytash

January - August, 2018

MIT Advisor: Karen Zheng

GroupM’s Proprietary Data

GroupM gave us access to their extraordinary and proprietary user interest data, among the most comprehensive in the world. As GroupM participates in display ad auctions, they record site visits for nearly all online users and nearly all webpages. These webpages are classified by a semantic engine, and then the counts of visits are converted to binary tables specifying whether a given user is “interested in” a given webpage type. It was these behavior tables – and no additional demographic data – that we used to make our behavior-based segmentation.

Process

The next step was to perform agglomerative hierarchical clustering with linkage determined by Ward’s Method. We kept the first 30 branches of these trees.

INTEGRATION & RESULTS

Final Segments

One metric of importance to GroupM is the ability of their bidding and insights engine to identify the right users from the population to match a segment. They evaluate this with a statistic called GRP, which measures the efficiency of the lookalike models. It can be thought of as similar to recall, except that smaller GRP numbers mean that the models are more successful. MIT segments outperformed nearly all benchmark models.

Lookalike Results

We activated our segmentation for three different clients, who each selected certain segments to target. These selections were based on segments “over indexed” for their current customers, and the results shown below demonstrate the monetary value of [m]Clusters.

[m]Clusters

AUDIENCES FIRST

Copenhagen, Denmark

GroupM Mentor: Kristian Brødreskift

Will Fein and Gerard Woytash

January - August, 2018

MIT Advisor: Karen Zheng

GroupM’s Proprietary Data

GroupM gave us access to their extraordinary and proprietary user interest data, among the most comprehensive in the world. As GroupM participates in display ad auctions, they record site visits for nearly all online users and nearly all webpages. These webpages are classified by a semantic engine, and then the counts of visits are converted to binary tables specifying whether a given user is “interested in” a given webpage type. It was these behavior tables – and no additional demographic data – that we used to make our behavior-based segmentation.

Process

The next step was to perform agglomerative hierarchical clustering with linkage determined by Ward’s Method. We kept the first 30 branches of these trees.

INTEGRATION & RESULTS

Final Segments

One metric of importance to GroupM is the ability of their bidding and insights engine to identify the right users from the population to match a segment. They evaluate this with a statistic called GRP, which measures the efficiency of the lookalike models. It can be thought of as similar to recall, except that smaller GRP numbers mean that the models are more successful. MIT segments outperformed nearly all benchmark models.

Lookalike Results

We activated our segmentation for three different clients, who each selected certain segments to target. These selections were based on segments “over indexed” for their current customers, and the results shown below demonstrate the monetary value of [m]Clusters.

[m]Clusters

AUDIENCES FIRST

Copenhagen, Denmark

GroupM Mentor: Kristian Brødreskift

Will Fein and Gerard Woytash

January - August, 2018

MIT Advisor: Karen Zheng

GroupM’s Proprietary Data

GroupM gave us access to their extraordinary and proprietary user interest data, among the most comprehensive in the world. As GroupM participates in display ad auctions, they record site visits for nearly all online users and nearly all webpages. These webpages are classified by a semantic engine, and then the counts of visits are converted to binary tables specifying whether a given user is “interested in” a given webpage type. It was these behavior tables – and no additional demographic data – that we used to make our behavior-based segmentation.

Process

The next step was to perform agglomerative hierarchical clustering with linkage determined by Ward’s Method. We kept the first 30 branches of these trees.

INTEGRATION & RESULTS

Final Segments

One metric of importance to GroupM is the ability of their bidding and insights engine to identify the right users from the population to match a segment. They evaluate this with a statistic called GRP, which measures the efficiency of the lookalike models. It can be thought of as similar to recall, except that smaller GRP numbers mean that the models are more successful. MIT segments outperformed nearly all benchmark models.

Lookalike Results

We activated our segmentation for three different clients, who each selected certain segments to target. These selections were based on segments “over indexed” for their current customers, and the results shown below demonstrate the monetary value of [m]Clusters.