Lookalike Modeling in E-commerce Context

Xiao Lu[1] and Juong-sik Lee[2]

Abstract— **One major challenge in e-commerce is to acquire high-value customers. We consider the problem of expanding a limited set of high-value customers a.k.a. source audience to a larger set of potentially high-value customers a.k.a. lookalike audience. The Lookalike Model examines the behavioral patterns, shopping interests, on-line engagement, price sensitivity and spending power across all product categories, and identify lookalike audience using the top most important features weighted by Jensen-Shannon divergence. Our results show that the Lookalike Model is capable of identifying the main criteria used to generate the seed audience, as well as correlated features that are not directly used to generate the seed audience. As a result, the lookalike audience is built on a richer, more comprehensive set of features, weighted by their significance in defining the source audience. Also, by selecting and weighting the top features, the Lookalike Model remains fully interpretable, and circumvent the curse of dimensionality. Unlike traditional dimension-reduction approach such as Principle Component Analysis, the Lookalike Model produces interpretable results, and runs much faster by directly ranking customers based on top most important features.**

I. INTRODUCTION

Traditional segmentation approach applies hard filters to user population, such as age range, account monetary value and purchase frequency, to create customer segments. This methodology has two disadvantages.

On one hand, hard filters may introduce human bias. For example, when promoting a baby product, requiring seed audience to be all females will leave out male customers who tend to their babies on their wives' behalf, or female customers who did not declare their gender on the e-commerce platform. If such bias is held to be absolute, the lookalike audience will all be females too.

On the other hand, hard filters reduce the population so fast that only a few hundred customers remain after applying just a few filters. As the number of filters increases, the size of segment shrinks exponentially. When a segment becomes too small, it represents a too specialized set of customers who may not generalize to a meaningful persona to build lookalike audience.

We propose a Lookalike Model that ”softens” the filters, and accounts for the underlying statistical structure of different metrics a.k.a. features. The assumption is that all features are not independent, as commonsense affirms. In other words, if all features are independent, applying hard filters would generate the most accurate segment.

Still, the lookalike mmodel does not deny that human bias is often helpful. Since the initial segments must be specifically defined for a business goal, human bias often aligns (albeit imperfectly) with the business goal. The role of the Lookalike Model is to take the source audience as input, inspect its underlying structure, assign weights to features, and build a larger segment using unsupervised machine learning. For example, knowing that most diaper buyers are women is a meaningful piece of information, but the Lookalike Model will override the human bias that ”diaper buyers must be female” if there are male customers who conform to most of other selected features of the initial segments.

II. METHODOLOGY

A. Feature Extraction

The goal is to represent every customer as a feature vector. The features we used can be divided into two groups: product features and account features. Product features represent how a customer interacts with each product category, e.g. views count, session count, placing items in cart, and placing an order. Account features are macro features relevant to customer demographic and account history, which are independent of product categories.

The product features are summarized in the table below. We apply the 17 product features ($f01$-$f17$) for every product category once. With a total of 105 product categories, this results in 1,785 features. Then we apply those features to all products regardless of categories, resulting in 17 more features. With three additional account features ($f18$ gender, $f19$ age, $f20$ account age), the complete Lookalike Model has 1,805 features.

The 17 product features are designed to capture the following five aspects of an e-commerce customer:

- Need: annualized aggregate spending (aas), annualized order count (cto), annualized quantity count (ctq), days per order (dpo), day per quantity (dpq) quantify customer's need for each product category.
- Habit: annualized view count (ctv) measures indirectly how much interests a customer places in each product category.
- Engagement: view per order (vpo), view per quantity (vpq) measure how much attention a customer spends on looking for the best offer in each product category.
- Spending power: GMV per order (gpo), per quantity (gpq), per day (gpd) measure how much a customer is willing to spend (per unit order / quantity) in each product category.
- Churn: days since last order (dal), views since last order (vsl) measure how likely a user is going to place the next order in each product category.

*This work was supported by Coupang Inc.

[1] X. Lu, MS Management Science & Engineering, Stanford University

[2] J. Lee, Audience Platform, Growth Engineering, Coupang Inc.

TABLE I

FEATURE ENGINEERING

	Symbol	Name
$f01$	aas	Annualized aggregate spending
$f02$	cto	Annualized order count
$f03$	ctq	Annualized quantity count
$f04$	ctv	Annualized view count
$f05$	cts	Annualized session count
$f06$	vpo	View per order
$f07$	vpq	View per quantity
$f08$	spo	Session per order
$f09$	spq	Session per quantity
$f10$	dpo	Days per order
$f11$	dpq	Days per quantity
$f12$	gpo	GMV per order
$f13$	gpq	GMV per quantity
$f14$	gpd	GMV per day
$f15$	dsl	Day since last order
$f16$	vsl	View since last order
$f17$	ssl	Session since last order
$f18$	age	Customer age
$f19$	sex	Customer age
$f20$	aay	Account age

B. Feature Weighting

Feeding all 1,805 features to the Lookalike Model is not scalable, as high-dimensional near-neighbor queries are substantial open problem in computer science. Furthermore, because of the curse of dimensionality, every pair of customers will be equally similar/dissimilar in high dimension. Hence we must perform feature selection and feature weighting before applying near neighbor method.

$$\begin{aligned} M &= \frac{1}{2}(Q+P) \\ D_{JS}(Q||P) &= \frac{1}{2}D_{KL}(Q||M) + \frac{1}{2}D_{KL}(P||M) \\ &= \sum_i Q(i) ln \frac{Q(i)}{M(i)} + \sum_i Q(i) ln \frac{P(i)}{M(i)} \end{aligned}$$

We use Jensen-Shannon divergence (JS divergence) to calculate weights for every feature. The JS divergence measures the similarity between two probability distributions. The more different the two distributions are, the higher the JS divergence.

In the Lookalike Model, we compare the distribution of the seed audience (Q) with that of a randomly drawn population sample (P). For each feature, we approximate the probability density function of the two samples using histogram, and directly compute JS divergence based on the histogram's bins. We selected 0.05 as the minimum threshold, above which the JS divergence can be attributed to meaningful signal instead of noise. After selecting top features with highest JS divergence, we compute their weights by normalizing JS divergence.

$$w_j = \frac{D^j_{JS}(Q||P)}{\sum_k D^k_{JS}(Q||P)}$$

C. Near Neighbor Ranking

Having selected a subset of features and assigned weights, the problem of finding lookalike audience reduces to classic near neighbor query. We normalize each selected feature to zero means and unit standard deviation, and multiplied weights obtained from earlier steps.

The same transformation was applied to the seed audience, which is averaged out to provide a Persona. Every customer in the population is compared against the Persona, and similarity scored is computed using Euclidean distance. The entire ranking process takes $O(n)$ time.

D. Lookalike Audience

Having ranked all customers by similarity to the Persona, the model outputs top m customers with lowest Euclidean distance as the lookalike audience. Because every customer must be compared against the Persona, the lookalike audience size can be selected or adjusted after the ranking process has completed, without having to rerun the model.

III. RESULTS

Because the lookalike modeling problem belongs to unsupervised learning, here we present three sample use cases to demonstrate its power and limitation.

A. Meal Essential

Suppose we would like to promote Meal Essential category product, and we apply only one criterion to segment seed audience: the customers must spend at least 100,000 KRW, or 86.36 USD, per average order placed in Meal Essential category.

From the features ranking table, the Lookalike Model successfully identified the product category used to generate seed audience – Meal Essential. The relevant features for this product category are ranked among the top: gpq, dsl, gpd, ass, and dpo. This is reasonable, as a large divergence in gpo surely causes divergence in other metrics of the same product category.

Interestingly, the model picks up relevant food categories: snack, coffee, tea, beverages etc. This is reasonable, as you may expect those who often order food also tend to order drinks and snacks. Furthermore, kitchen disposable (gpo, gpq), detergent (gpo) are listed among the top 20 features. It logically makes sense: people who order food frequently online may also order kitchen supplies together.

B. Baby Core

Suppose we would like to promote certain baby products, and we apply the following criteria in selecting the seed audience:

- Customers must be females
- Spend over 70,000 KRW per order on diaper product
- Have not ordered diaper for more than 30 days but fewer than 180 days
- Age between 25 and 35

The model accurately returned the defining features (top three) except age. This can be explained by the fact that

TABLE II

TOP FEATURES - MEAL ESSENTIAL USE CASE

sym	category	D_{JS}	weight
gpo	Meal & Healthy - Meal Essentials	0.972	0.333
gpq	Meal & Healthy - Meal Essentials	0.328	0.112
dsl	Meal & Healthy - Meal Essentials	0.176	0.060
aas	Meal & Healthy - Meal Essentials	0.168	0.058
gpd	Meal & Healthy - Meal Essentials	0.168	0.058
dpo	Meal & Healthy - Meal Essentials	0.131	0.045
gpo	Snacks, Coffee & Tea - Snacks	0.102	0.035
dpq	Meal & Healthy - Meal Essentials	0.101	0.035
gpo	Beverages - Drinks	0.099	0.034
gpo	Fresh, Cold & Frozen - Fresh	0.082	0.028
gpo	Kitchen - Kitchen Disposable	0.070	0.024
gpo	Cold & Frozen Food	0.069	0.024
gpo	Snacks, Coffee & Tea - Coffee & Tea	0.062	0.021
aay	null	0.061	0.021
gpo	Household - Detergent	0.060	0.020
gpo	Household - Tissue	0.060	0.020
cto	Meal & Healthy - Meal Essentials	0.057	0.019
gpq	Snacks, Coffee & Tea - Coffee & Tea	0.055	0.019
gpo	Meal & Healthy - Healthy food	0.053	0.018
gpq	Kitchen - Kitchen Disposable	0.049	0.017

33.6% of female customers fall within this age range, and so the age feature is not a significant in defining the source audience.

Other features, relevant to pet, beauty, and bag categories are returned. It should not be surprising for female mom to be interested in those items. However, it should be noted that all the model receives is a set of user IDs, and that it is capable of finding those association among interests, behavioral patterns, and spending power in relevant product categories.

To check that the lookalike audience accurately captures the distinguishing characteristics of the source audience, we may examine the histograms for the top features. If the distribution of lookalike audience is closer to the source audience than to the population, then the lookalike audience is similar to the source audience. For example, the in the gpo015 feature, both source audience and lookalike audience are truncated at 70,000, meaning that the differentiating criteria is upheld. The dsl feature doesn't observe a clear cuts on upper and lower ends, but observes a fatter tail, shifting toward the source audience direction. For the sex feature, lookalike audience is mostly females, compared to the 100% female source audience.

Generally, when the population distribution is normal, lookalike audience's shift to source audience is exhibited as a shift of mean (see gpq015 - Baby Core - Diapering below). When the population distribution is exponential, the lookalike audience's shift is exhibited as a flattened exponential distribution, with a fatter tail.

Note that the shape of histogram depends on the lookalike audience size. The more lookalike audience, the more its shape converges to the population (if the lookalike audience is chosen to be the same size of the population, then the two become exactly identical).

TABLE III

TOP FEATURES - BABY CORE USE CASE

sym	category	D_{JS}	weight
gpo	Baby Core - Diapering	0.765	0.337
dsl	Baby Core - Diapering	0.307	0.135
sex	null	0.257	0.113
dpo	Baby Core - Diapering	0.152	0.067
cto	Baby Core - Diapering	0.147	0.065
gpq	Baby Core - Diapering	0.140	0.062
dpq	Baby Core - Diapering	0.104	0.046
dsl	Accessory - Bag	0.078	0.035
cto	Baby Core - Formula	0.070	0.031
dpo	Baby Core - Formula	0.066	0.029
dpo	Pet - Cat Food	0.062	0.027
gpq	Baby Core - Wipes	0.061	0.027
dpq	Pet - Cat Supply	0.059	0.026

IV. CONCLUSION

The Lookalike Model can reliably identify the criteria used to select the source audience, and it can also identify correlated features that are not directly used to produce source audience. Hence, the ranking of lookalike customers takes into account a more comprehensive set of features than the source audience alone.

As shown in the gender and age features in the baby core use case above, the Lookalike Model can override human bias if necessary. That is, when male customers are buying baby product just as female customers do, those male customers will not be excluded by the model because of their gender. In contrast, when a criterion (age) proves to be unhelpful in segmenting source audience, the Lookalike Model may simply discard that criterion.

The Lookalike Model is superior to Principle Component Analysis, which transforms high-dimensional feature space to lower dimension, before running near-neighbor queries. On one hand, PCA is a black box model, meaning that it is impossible to logically interpret the rule of transformation. On the other hand, PCA treats all variations indiscriminately, and tries to preserve as much variation in lower space as possible. If signals are concentrated among a few features, those variations in irrelevant features, when projected onto lower dimensional space, may easily overwhelm the signals.

The disadvantage of the Lookalike Model is that all divergences between source audience and population are valued equally (note that this shall not be confused with PCA, which values variation among every single feature equally). This is a fundamental limitation of unsupervised learning. Because no label of any kind is available, if two features are assigned equal JS divergence score, it is impossible to tell which one is more helpful in advancing the business goal.

APPENDIX

Though the proprietary data are not available to the public, the figures, table, presentation and discussion could be found at GitHub repository: `https://github.com/shawlu95/Lookalike_Model`

ACKNOWLEDGMENT

The proprietary data, project funding, and mentorship were generously supported by Coupang Inc.

REFERENCES

[1] Jain A.K. (2008) Data Clustering: 50 Years Beyond K-means. In: Daelemans W., Goethals B., Morik K. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2008. Lecture Notes in Computer Science, vol 5211. Springer, Berlin, Heidelberg

[2] Dullaghan, Cormac, and Eleni Rozaki. ”Integration of Machine Learning Techniques to Evaluate Dynamic Customer Segmentation Analysis for Mobile Customers.” arXiv preprint arXiv:1702.02215 (2017).

[3] Qiu, Jiangtao. ”A predictive Model for Customer Purchase Behavior in E-Commerce Context.” PACIS. 2014.

[4] Pandey, Sandeep, et al. ”Learning to target: what works for behavioral targeting.” Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011.

[5] Jacobs, Bruno JD, Bas Donkers, and Dennis Fok. ”Model-based purchase predictions for large assortments.” Marketing Science 35.3 (2016): 389-404

[6] Kooti, Farshad, et al. ”Portrait of an online shopper: Understanding and predicting consumer behavior.” Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. ACM, 2016.

[7] Jagabathula, Srikanth, Lakshminarayanan Subramanian, and Ashwin Venkataraman. ”A Model-based Projection Technique for Segmenting Customers.” arXiv preprint arXiv:1701.07483 (2017).