Temporal Evolution of Behavioral User Personas
via Latent Variable Mixture Models

Snigdha Panigrahi[1]
psnigdha@umich.edu
Department of Statistics,
University of Michigan, Ann Arbor,
MI, USA.

Nadia Fawaz[1]
Sunnyvale, CA, USA

Ajith Pudhiyaveetil
ajith.pudhiyaveetil@technicolor.com
Technicolor, Palo Alto, CA, USA.

ABSTRACT

This work[1] characterizes the users of a VoD streaming service through user-personas based on a **tenure timeline** and **temporal** behavioral features in the absence of explicit user profiles. A combination of tenure timeline and temporal characteristics caters to business needs of understanding the evolution and phases of user behavior as their accounts age. The personas constructed via latent variable mixture models successfully represent both dominant and niche characterizations while providing insightful maturation of user behavior in the system. With new users entering the system at any time point, the existing user-profiles are updated in our temporally evolving approach.

The two major highlights of our personas are demonstration of *stability* along tenure timelines on a population level, while exhibiting interesting migrations between labels on an individual granularity and clear interpretability of user labels. Finally, we show a trade-off between an indispensable trio of guarantees, *relevance-scalability-interpretability* by using summary information from personas in a CTR (Click Through Rate) predictive model. The proposed method of uncovering latent personas, consequent insights from these and application of information from personas to predictive models are broadly applicable to other streaming based products.

CCS CONCEPTS

• **Computing methodologies** → Modeling and simulation; • **Computing methodologies** → Machine learning.

KEYWORDS

user personas; temporal labels; personalization; CTR prediction; mixture model.

[1]This work was performed while all three authors were with Technicolor Research, CA, USA.

INTRODUCTION

User segmentation, the idea of dividing a market up into homogeneous segments and targeting each group with a distinct product or message is a basic tool to model similar consumers. This is explored in diverse sectors like finance [24], health [16], telecommunications [5] etc. and through focus on different behavioral aspects [1], [4], [10]. The current work adopts a latent parametric *mixture model* approach to construct segments of homogeneous consumers called *user personas* for VoD services from raw transactional logs, using a tenure timeline and temporal behavioral features. Examples of such services in the VoD space include *itune, googleplay, vudu, fandangoNOW*, etc; where users pay per piece of content they watch. This is in contrast with subscription based services, where users pay a monthly subscription, such as netflix, hulu plus, amazon video etc. The work provides explicit user characterizations based on **spending behavior**, **content preference** and **transactional habits** with the main contributions as presented below:

- Align user transaction timelines on a **tenure basis** at a **monthly** granularity, a novel choice for a timeline of comparison, in place of the conventional calendar timeline
- Construct temporal behavioral feature vectors from transaction logs, that are aggregates of transactions over a month along tenure timeline; such features represent the **evolving behavioral** consumer traits
- Capture both **dominant** and **niche** segments of population and provide highly interpretable user labels.
- Capture stable latent structure on a population level, even as individual profiles keep transforming with age. The derived user personas maintain a **consistent** clustering over time while accurately explaining the changes on an individual level.
- Represent insights on inter-relations between behavioral characteristics as layers within user profiles.

Such a construction of temporally evolving personas with new insights into behavioral characteristics is the *first of its kind* in the streaming space, to the best of our knowledge. The extended detailed version of this work is available in [25].

A line of prior works [2], [6], [7], [33] [32] has explored characterization of consumers; another independent set of works has contributed to methods on personalized recommendations [3], [27], [28], [26], [18], [14], [22]. Our work concludes with a unification of these two important goals to demonstrate the utility of *user personas*. In particular, we illustrate an application of persona based features in CTR (Click through Rate) predictions. We show that a model based on the constructed personas achieves a 3 criteria **relevance - scalability - interpretability** tradeoff, when compared

against models that do not include persona information. We show a substantive gain in computational cost through the use of lower dimensional persona features in the form of soft or hard clustering information. This gain occurs with retaining clarity in the interpretation of feature space (as opposed to random projections onto lower dimensional spaces) and does not compromise with predictive ability. The CTR model we describe is interesting in its own right as we use persona features in a logistic model trained per item to capture item specific variability. The use of persona information can also aid in preserving anonymity of individual users as well as of individual transactions. We supplement the CTR model with a discussion on other commonly used collaborative filtering models that can potentially achieve a similar trade-off.

Our methods are by no means limited to the VoD space. They can be extended to lend similar insights and achieve similar benefits for other product based services. Modeling latent structure from raw transactional data can overcome the curse of dimensionality through an efficient reduction in regression size, while maintaining predictive power and interpretability of feature space.

Related works

Consumer segmentation is driven by the intuition that predictive models of customer behavior based on groups of similar customers outperform a single aggregate model, see [29]. A segmented predictive model [2] can be refined further to an individual level, trained per customer. In doing so, we gain a reduced bias in creating increasingly more homogeneous customer groups at the cost of increased variance in estimation as we consider progressively more refined segments containing fewer customers. Thus, there is a classic bias-variance trade-off which is effectively dealt by integrating customer segmentation into such predictive models, termed as *segmented models*, see [19]. In this work, we advocate the use of features based on user personas not only for improvement of predictive power but, as a *meaningful, lower dimensional, summary* space that can be used to achieve scalability in regression models and facilitate storage for future debugging.

Various techniques of segmenting consumers include neural net models [8], latent probabilistic models [12], combinatorial optimization based grouping models [20]. We offer in this work a multinomial latent mixture model analysis with both soft and hard clustering values as outputs, employing the classic Expectation-Maximization (EM) algorithm [11] to estimate the mixing proportions and distribution parameters for building user personas. Most part of the raw data-logs consists of count features for which a multinomial model seems a natural choice; except for the spending amounts which we choose to implement the K-means clustering which gives similar results as as the more commonly used parametric Gaussian mixture model [15]. In comparison to prior art, our goal here goes beyond discovering latent representations. That is, we want labels that can directly render business insights as opposed to non-interpretable clusters.

One of the key features of our personas is that they exhibit stability on a population level even as migrations on an individual level are constantly taking place along the chosen time granularity. [9] explores clusters not shifting dramatically from one time-step

to the next and [23] establishes equilibrium of average network properties, a concept resonating with the stability of clusters.

VoD Dataset

The dataset considered in this work consists of transaction logs of a subset of **730, 000** anonymous users from a large-scale streaming VoD service across a time span of **16** months from January 2014 to April 2015, with over **2** million transactions. Each record in the transaction logs consists of a unique user-id, a unique time-stamp, a unique content-id, the type of transaction–rentals/ purchases, a net price giving the cost of each transaction, and content meta-data such as genres, release year, MPAA ratings corresponding to each transaction. [30] analyzes a processed user-interactional part of this data set, consisting of 3488 users and 26404 viewing sessions, to model binge watching behavior for VoD services; we consider a larger set of users in our analysis and focus on the transactional data instead.

We present summary statistics based on the transactional data; these preliminary statistics and observations lead to the belief that there is a latent structure in the users consumption patterns and guide the pre-processing stage to construct features from raw transaction logs. Note that the characterizations of user behavior discovered as latent structure from raw logs in this work can be viewed as more precise and refined summaries. The transactions break up into two types- rentals and purchases with **88%** rentals and **12%** purchases. The price categories of rentals vary from **0 − 5\$** with higher price categories falling in the **3 − 5\$** range. The purchases range as high as **25\$**, mainly for new movies and tv series. The purchases greater than **10\$** in value are considered as higher end transactions. Most transactions occur in the lower price categories of both types of transactions with only **10%** of consumers transacting in the higher price ranges. A transactional perspective of the content catalogue is observed through segregation of transactions into **15%** TV shows and **85%** movies, with the movie *Frozen* being the most consumed content in the catalogue. The dominant genres in the transactions are *Drama (**18%**), Comedy (**10%**), Action (**10%**), Family (**9%**), Animation (**7%**), Thriller (**6%**), Biography (**5%**), Sci-fi (**4%**), Crime (**4%**)* etc, with the crucial observation that while some users (**20%**) tend to prefer more family-friendly content (Family, Animation, Super-hero). Other segments (**80%**) consume genres such as drama, horror, comedy etc. The time of transactions is seen mostly to range between evenings and nights, evenly split between weekdays (Monday-Friday) and weekends. As part of the pre-processing of raw logs, barely active users (spend less than **1** dollar in a certain month of activity) and one-time deal hunters (transact only once and never return) were filtered out to prevent cluster centers being pulled to **0**. Summarized information is subsequently uncovered from the data as cluster centers and cluster sizes, which preserves anonymity of individual users while not giving information on any particular transaction.

CONSTRUCTION OF PERSONAS

We construct personas on spending traits, content preferences and transactional habits of users, with interest in above characterizations stemming from domain knowledge, product intuition, and business goals. We discuss the timeline, granularity of comparison,

and behavioral features that are aggregated over 1-month windows of transactions; these play a consequential role in excavating meaningful latent structure in raw data.

Timeline of comparison

Transaction logs consist of time-series data. We make a careful choice as to how the timelines of different users are compared with regard to the following 2 aspects:

Temporal alignment of user timelines: User timelines can be aligned on a calendar basis or on a tenure basis. In the calendar basis, transactions of different users happening at the same calendar dates, for instance in January 2014, are compared against each other. Aligning timelines according to a calendar basis allows to detect seasonalities(holidays, end-of-year movie releases), and effects of specific events happening at a particular date (TV-show new episode/season release or end). On the other hand, in the tenure basis, the first transaction of a given user defines the birth of the user timeline, and transactions of different users are compared when they happen at the same age of the user in the system. For instance, if user A made his first transaction on January, 15th 2014 and user B made his first transaction on April, 10th 2014, comparisons would be drawn for their first month of transactions between Jan. 15th-Feb 14th 2014 for user A and April 10th-May 9th for user B. Aligning timelines on a tenure basis allows observations on how users age in the system and helps in understanding behavioral phases and in predicting churn.

Temporal granularity: Timestamps in transaction logs can be specified up to seconds or even milliseconds. When building features based on time-series, the question arises as to the granularity at which events should be grouped to devise the desired features. Transactions can be aggregated at a monthly/ weekly/ daily/hourly granularity. For instance, to compute a count feature at the monthly granularity, transactions happening within the same 30 day period will be aggregated. The granularity level affects the detection of behavioral patterns and cycles.

In this work, user timelines are aligned on a **tenure basis**, and events are considered at a **monthly (30 days) granularity**. The first transaction of a user marks the beginning of its timeline, and user's transaction history is divided into successive periods of 30 days each. Our choice of a monthly granularity is guided by elementary analysis of the transaction logs which show unstable structures with weekly granularity– *Weekly logs were too short a period to capture behavioral patterns*–, and a flat structure at a quarterly granularity–*Quarterly logs were too long to capture the dynamism in user labels due to an over-cumulation effect of data*. Our choice of a tenure basis was motivated by the business need to understand the evolution and phases of user behavior along their transaction histories; this helps model their dynamic behavior, predict loss of interest in system, predict lifetimes etc.

Aggregate feature space

The features used in construction of personas are aggregates of transactions at monthly granularity, binned into categories. The choices of binning, arising from a combination of summary knowledge of data and domain information, lead to the below features.
Monthly Expenditure (ME) characterizes spending behavior: Each feature is the total net amount spent in one month by a user in either a rental/purchase transaction type and a given price category (5 categories for rentals, 8 for purchases).
Transaction frequency (TF) characterizes economic behavior: Features are transaction counts binned into 2 price categories in rentals and 4 categories in purchases.
Dominant genres (DG) indicates content preference: Features are monthly counts of transactions in 15 most popular genres: *Drama, Comedy, Action, Family, Animation, Thriller, Biography, Sci-Fi, Crime, Super Hero, Comedy-Drama, Fantasy, Horror, Romance, Kids, Miscellaneous.*
Content recency (CR) indicates freshness preference: Features are counts binned into ranges of content release year: *Old: < 1990, Nostalgia: 1990 − 2000, Not New: 2000 − 2010, Recent:2010 − 2013 and Latest: 2014 − 2015.*
Time & day of transaction (TDT) gives transacting habits: Timestamps of transactions are processed to generate the day of week and time of transaction as per the geographic region of the user, then counts are binned into weekdays or weekends and 4 time slots: *10 AM-5 PM (Office Hours), 5PM-10 PM (evening and night), 10PM-5AM (late night).*

A mixture model for latent characteristics

To fix notations for this section, we have a $n \times d$ feature matrix $X^T = (x_1, x_2, \cdots, x_n)$, with $x_i \in \mathbb{R}^d$ representing the feature vector of user i in a sample of n users, and d representing the dimension of the feature space. We propose a parametric approach, a mixed multinomial model **MMM** [13],[31],[17], to describe user labels based on count data. The choice of a multinomial distribution is a natural model for count feature vectors. The iterative EM algorithm applied to estimate the mixing proportions and the parameters in mixed multinomial distribution, is in itself a very powerful mechanism, with one of its many merits being the ability to deal with missing features. An MMM assumes that rows of X are independent draws from a multinomial model, that is $x_i \sim MN(d, \theta_{Z_i})$, where Z_i is a latent variable from the categorical distribution taking values $j \in [K]$, where K is the number of clusters; independent of X_i. We have a hierarchically structured model as

- $Z_i \overset{iid}{\sim} MN(1, \pi)$ with $\pi = (\pi_1, \cdots, \pi_K)$ representing mixing probabilities for the K clusters;
- $X_i|(Z_i \overset{ind}{=} j) \sim MN(d, \theta_j)$, where for $j \in [K]$ and the vector $\theta_j = (\theta_{j,1}, \cdots, \theta_{j,d})$ represents parameters in the multinomial density given latent factor $Z = j$.

The mixing probabilities π and the parameters of the mixture model $\theta_j, j \in [K]$ are estimated using an EM algorithm as proposed in [11]. We outline the E and M steps for the (t)-th iteration of the algorithm for the MMM based on iterates $\pi^{(t)}$ and $\theta^{(t)}$–
E-step: computes the posterior probabilities given estimates of parameters π and θ_{Z_j} of the t-th iteration, that is

$$\tau_{i,z}^{(t)} = \mathbb{P}(Z_i = z|x_i; \pi^{(t)}, \theta^{(t)})$$
$$= \mathbb{P}(X = x_i|Z_i = z; \theta^{(t)})\pi_z^{(t)} / \sum_{j=1}^k \mathbb{P}(X = x_i|Z_i = j; \theta^{(t)})\pi_j^{(t)}.$$

where $\mathbb{P}(X = x_i|Z_i = j; \theta^{(t)}) \propto \Pi_{v=1}^d \theta_{j,v}^{x_{i,v}}$.
M-step: maximizes the Expected Complete Log Likelihood (ECLL) to refine estimates of parameters π and θ_j for $j \in [K]$; $\theta^{(t+1)}, \pi^{(t+1)} =$

$\arg\max \mathbb{E}(\mathcal{L}(\theta^{(t)}, \pi^{(t)}; X))$, where ECLL is

$$\mathbb{E}(\mathcal{L}(\theta, \pi; X)) = \sum_{i=1}^{n} \sum_{j=1}^{k} \tau_{i,j}^{(t)} \times \log(\pi_j \cdot \mathbb{P}(X = x_i | Z_i = j; \theta))$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} \tau_{i,j}^{(t)} \log \pi_j + \sum_{i=1}^{n} \sum_{j=1}^{k} \sum_{v=1}^{d} x_{i,v} \tau_{i,j}^{(t)} \log \theta_{j,v} + \text{constant}$$

(where constant does not depend on π, θ) yielding estimates

$$\pi_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_{i,j}^{(t)}, \; \theta_{j,v}^{(t+1)} = \sum_{i=1}^{n} x_{i,v} \tau_{i,j}^{(t)} / d \sum_{i=1}^{n} \tau_{i,j}^{(t)}.$$

Hard cluster assignments are obtained by calculating

$$S_{x_i} = \arg\max_{z} \mathbb{P}(Z_j = z | x_i; \pi, \theta),$$

with $S_{x_i} \in [K]$ for $i \in [n]$.

PERSONAS AND INSIGHTS

Having described our methods of constructing personas, we present the summaries of personas based on preferential and behavioral patterns. The significant highlights of these persona labels are clear characterizations of users in each persona label. We supplement the persona labels with interesting insights that can lead to future business actions to understand evolving patterns of both dominant and niche behavioral traits.

User persona labels for behavioral characterizations

We give interpretable persona labels based on the latent structure excavated from the aggregate features described above. Below, we list the labels with the cluster sizes reported in percentages (beside the label) and for each label, we give a brief explanation of user behavior in that bucket.

Monthly Expenditure: Cluster centers represent monthly expense in each of the 13 price categories (5 rental and 8 purchase price categories). The user persona labels are

- **Economic Renters** (71%) : 10$ spent in a month of activity, including smaller $2 - 3$\$ amounts in higher renting price categories.
- **Heavy Renters** (21%) : 17$ in total, including 13$ spent in the $3 - 5$ rental price category.
- **Movie Buyers** (4.5%): 32$ in total with one purchase on average in the $16 - 20$ price category and 1/4-th of monthly expenses in higher-priced rentals and lower-priced purchases.
- **Movie Buffs** (2.5%): 60$ in total, with around 3 purchases in $10 - 16$ price category and around 7 dollars in $16 - 20$ price category.

Frequency of Transaction: Cluster centers denote transaction counts in 6 price ranges, the persona labels uncovered are

- **Frequent High-End Renters** (61%): over 85% transactions in rentals above 3$.
- **Frequent Low-End Renters** (21%): over 60% and 30% transactions in rentals below and above 3$ respectively.
- **Frequent Movie Buyers & Sporadic Renters** (12%): 45% transactions in purchases in $8 - 16$\$ price category and 35% transactions in rentals as well.

- **Frequent Low End Purchasers** (6%): 80% transactions mostly in the $0 - 8$\$ purchase price category.

Dominant Genre of Content Consumed: The three prime clusters recovered with cluster centers being percentage of monthly transactions in 16 genres are–

- **Happy Family** (23%): content qualifying as family watch with distribution being family genre (28%)–the most consumed genre, followed by animation (20%), comedy (13%); but no or almost no crime, horror, romance, thriller.
- **Drama-Comedy:** (40%) content with dominant genres– drama (28%), followed by biography (10%), comedy (10%), bit of romance but little or almost nothing as compared to other clusters in terms of consuming family, horror, action, crime.
- **Action-Horror-Thrill:** (37%) dominant genre is action (20%), followed by drama (15%), thriller (12%), sci-fi (8%), comedy (6%), horror (5%), but little or almost nothing as compared to other clusters in terms of consuming family, comedy-drama, fantasy content.

Recency of Content consumed: We obtain 3 genre clusters based on the count matrices binned as per release year of content to observe characterizations for recency of content.

- **Latest** (40%): 85% transactions with release year 2014-15.
- **Recent** (30%): 85% transactions with release year 2010-13.
- **Nostalgic** (30%): About 30% with release in 2000-09 followed by recent and latest content in the remaining 65% of transactions.

Time & Day of Transaction: Based on habits or preferences to transact at a certain times and days of the week, the clusters with centers representing counts in each time category of weekday/ weekend are

- **Weekend Evening & Night** (24%): 65% of transactions on weekend nights, followed by 25% in evening.
- **Weekday Evening & Night** (24%): 70% of transactions on weekday nights, followed by 20% in evening.
- **Weekend & Weekday Night** (42%): 45% and 35% of transactions on nights of weekdays and weekends.
- **Weekend Day & Night:** (10%): 25% of transactions on weekend day time and 60% in weekend nights.

Insights into user persona labels

Temporal nature of labels: stability of macro characteristics- A highlight of the derived user personas is that the uncovered clusters stay stable in terms of size and composition on a population level. This attractive property of consistency allows us to use these clusters to model the temporal evolution of tenure timelines at a population level consistently. At the same time, the personas also succeed in explaining individual dynamism. That is, migrations do happen on a user level and individual user labels are not static. Our results show that these migrations between categories are never drastic in nature, but rather migrations between neighboring clusters, although we observe a few interesting migrations into far-off labels. These migrations can be explained as *dominant characterizations* reflecting spending capacity, content preferences and habits staying stable over time while *niche characterizations* being more prone to change. As specific examples, we see the dominant segment of users transacting in lower-priced categories staying stable

in their respective labels over time. However, the niche segment of higher end purchasers keep migrating to lower end categories and migrate back to the niche labels with only availability of new products of their interest. Another niche segment is a proportion of people who buy content in the *happy family* label; over their tenure, they move to other labels of genre consumption to buy content for individual consumption that is different from content consumed in the family context. On the contrary, the other two labels within genre preference together represent the dominant population and show stability along tenure timelines.

Natural hierarchical structure of clusters: We observe that the user personas exhibit a natural, divisive, hierarchical structure (not imposed through algorithm), as we increase the number of clusters. This lends interesting interpretations on the sub-population of users within broad segments. An example of this is upon clustering users based on monthly expenditure into two clusters, cluster centers represent renters and purchasers, the two main segments of users. When increasing the number of clusters, renters break up into economic and heavy renters with 3 clusters, while purchasers mostly decompose into two niche clusters, movie buyers and movie buffs with 4 buckets.

Upon clustering count data representing dominant genres consumed by users into two clusters, we see a segment preferring family content over a segment that consumes content not qualifying as family watch. With three clusters, the non-family content consumers decompose into two buckets- one that consumes drama, comedy etc while other prefers thrill inducing content.

Layered structure of clusters: We explore the inter-relations between the various user persona characterizations by performing a layered clustering using the mixture model technique. An example is the assignment of labels for a characterization such as genre preference within clusters for spending behavior. For instance, we observe that clusters based on genre preference derived within the clusters characterizing economic behavior are similar across all economic clusters. Similarly, the clusters for spending behavior are similar across different genres. This observation statistically validates that genre preferences of consumers are independent of their economic budget. A similar observation goes for recency and economic behavior. On the other hand, we see different clustering results for recency of content when clustered within the genre clusters, with the category preferring family content showing more inclination towards more classic content than the other drama-based or thrill inducing categories that prefer more recent content.

INTEGRATION OF PERSONAS IN PERSONALIZATION

We demonstrate the usefulness of user personas through an application to response prediction, by effectively integrating the information from personas into personalization. Specifically, we focus on a CTR predictive model where the goal is to predict $p_{u,i}$, the probability that user u transacts on item i. The scope of utilizing persona information extends to other popular models in collaborative filtering. We conclude the paper by discussing such possibilities, where one can integrate personas into other commonly used models and expect to attain a relevance-scalability-interpretability tradeoff.

CTR: relevance-scalability-interpretability balance

We model the CTR problem to predict transactional probabilities through an ℓ_1 penalized logistic regression model that is trained per item. Such a fine-grained model at the item level captures the item specific interest in users, leading to more accurate predictions [34]. The challenge in such models, however, is the sparsity of the transactional data, with about **1**% users transacting on any given item. To overcome this imbalance and avoid bias towards the outcome of *not transacting at all*, for every positive sample (users who transacted), we sample 5 negative samples (users who did not transact). The gain with summarized information from personas can be described as a balance between scalability of the training model, interpretability of feature space and relevance of predictions:

Relevance-deliver relevant recommendations to users, quantified by the quality of prediction in transactional probabilities. To fix notations, we denote the evaluation metric to assess the performance of the predictive model as $\mathcal{F}$ on a test set. With the training model M^* giving predicted labels $\widehat{\text{label}}_{M^*}$, the predictive ability is given by $\mathcal{F}(\widehat{\text{label}}_{M^*}, \text{label}_{\text{test}})$. $\mathcal{F}$ here, is the mean AUC over the 100 most popular items in the content catalogue.

Scalability-reduce the size of input feature and sample space (leads to reduction in regression size) by using lower dimensional persona features. Information from personas can be encoded as soft clustering features or incorporated as hard clusters via a model trained per cluster. This brings significant reduction in regression dimensions which in turn, facilitates storage and future use of these feature vectors in the same or other predictive models.

Interpretability-retain the intuitive meaning of the feature space as opposed to random lower dimensional projections which seldom lend business insights. With a meaningful feature set, we can reutilize the same features in a host of predictive tasks and use them in easy debugging of models. While relevance and scalability can be quantified, there is no measure of interpretability.

The trade-off in the above criteria arises as we can use a baseline model with the count features that were used to recover latent user labels as regressors. However, there is a significant computational cost associated with a higher regression size of the baseline based on these aggregate features, without using any knowledge of personas. We see a clear reduction in regression size and the associated complexity with integration of persona information at the cost of losing only a mere 2% predictive ability in Figure 2. Scalability of regression size with comparable predictive power as the baseline model alongside retaining clear meaning of feature space is the trade-off achieved in CTR prediction with persona information. The take away is that persona features can be used to construct interpretable, lower dimensional regressors that preserve predictive power. An added advantage of incorporating these summary features in a model with sub-sampled users is preservation of privacy of individual users and also, of individual transactions in using summaries over a random set of users.

To describe our model and results, we use X_u to denote the feature vector corresponding to user u. This feature vector can be based on 3 characterizations: ME (monthly expenses), DG (dominant genre), CR (content recency). Information from personas can be incorporated into X_u in different ways, yielding different models.

Figure 1: AUC of models based on X_u that uses different characterizations: Recency of a content item is the most predictive of the consumer's interest in transaction.
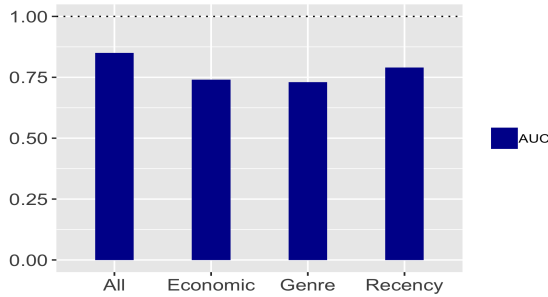
In particular, we construct feature vectors using the personas on ME, DG and CR in the following forms - denoted by **(c)**, **(s)**, **(h)** and **(-)** respectively. **(c)** is used in the baseline model with count features based on a particular characterization, **(s)** and **(h)** integrate soft and hard clustering information based on characterizations. **(-)** uses neither count nor persona information, we call this the null model. These are summarized below:

(c) a feature vector with distribution of ME in price categories and/ or count vectors for DG/CR in feature bins (directly using the constructed features). It uses aggregate count features, but no additional knowledge from latent personas.

(s) a feature vector of soft clustering values in the form of distances of count features from their respective cluster centers.

(h) incorporates hard clustering information for a characterization by training a model cluster-wise.

(-) does not include any information from a characterization at all.

Below, we describe the different CTR models and discuss results on the three criteria trade-off.

We achieve a **gain in relevance** with information from each added characterization, either in the form of soft clustering/ hard clustering/ count feature. Figure 1 highlights the relevance of each characterization in the CTR model. CR (recency) is seen to the most informative characterization adding the most to AUC.

Denote n_i as the samples per item and $n_{i,c}$ as samples per item, per cluster, p the number of predictive features, O as the complexity of regularized logistic with sample size and regression dimension. Table 1 below compares different models illustrating how our proposed integration of user personas into personalized recommendation achieves a tradeoff between relevance and scalability. We note that interpretability comes alongside using summary information from personas. The baseline model is depicted in the first row of the table; representing the model with all count features **(c)**. We see a significant reduction in the predictive power when we do not incorporate any information from the recency feature, this is depicted by the fourth row of the table. When we train a model per recency-cluster using **(h)**, we lose 1% of predictive power, but reduce the sample size for the training model on each cluster as well as the feature space leading to an overall reduction in complexity. We see a similar predictive power when we use soft clustering recency feature **(s)**, but a significant reduction in the size of feature space.

Figure 2: Computational complexity O as sample size per item n_i varies and $n_{i,c} = \lceil n_i/3 \rceil$ varies. The null model can be seen to more scalable than the model based on counts or soft-clustering features, but is seen to lose 13% power in comparison to the baseline model based on counts. The persona based models are comparable to the baseline in terms of predictive ability, but are more scalable. The gain with hard cluster membership based model is clearly considerable.

While we do not incorporate all 64 combinations of **(c)**, **(s)**, **(h)**, **(-)**, we see that using soft clustering features for all the 3 characterizations leads to a loss of only 2% AUC. This is represented in the last row of below table. The computational gain, however, is seen to be significant even in a simple regression model that scales in complexity as p^2 with the size of feature space. Figure 2 shows this as n_i and $n_{i,c} = \lceil n/3 \rceil$ varies per item. Interpretability is inherent in these models due to the clear meaning of soft clustering features that represent distances from cluster centers or hard-coded cluster memberships in training models based on similar users.

Table 1: Tradeoff in AUC versus complexity based on size of sample and feature space

Recency	Genre	Economic	$\mathcal{F}$	n	p	O
c	c	c	0.85	n_i	140	$O\left(n_i \times 140^2\right)$
s	c	c	0.84	n_i	128	$O(n_i \times 128^2)$
h	c	c	0.84	$n_{i,c}$	116	$O(n_{i,c} \times 116^2)$
-	c	c	0.75	n_i	116	$O(n_i \times 116^2)$
s	s	s	0.82	n_i	40	$O(n_i \times 40^2)$

Persona based collaborative filtering models

We finally discuss few models based on popular collaborative filtering techniques that can incorporate information from personas to retain predictive power while gaining in scalability for practical implementations.

User based nearest neighbor similarity: This approach is based on a similarity metric $\text{sim}(u, v)$ (examples include Jaccard, cosine etc.) to predict a weighted average rating based on similarity between users who transacted on the same items. Denoting by $U(i)$ the set of users who transacted on the same item i, the amount of money $r_{u,i}$ that user u is willing to spend on item i can be predicted as

$$r_{u,i} = \sum_{v \in U(i)} \text{sim}(u, v) r_{v,i} / \sum_{v \in U(i)} \text{sim}(u, v),$$

and the probability that user u transacts on item i as

$$p_{u,i} = \sum_{v \in U(i)} \text{sim}(u,v) / \sum_v \text{sim}(u,v).$$

Similarity approaches have scaling issues with high computational cost associated with searching through set of users or even the top K similar users in the set $U(i)$. Persona information can bring in gain in prediction accuracy, also offering better scalability via limiting search of top K neighbors to already formed personas.

We could use clusters from most representative time point of activity for predictions. Alternately, we can use temporal persona information for prediction with the scope of leveraging differently on time points through a weighted similarity prediction along the tenure timeline. Denoting time points of transaction history (months of tenure timeline) as t with weights w_t (that can be tuned) and features-u_t for user u, $U(t,i)$ as the set of users who transacted on the same item i and $C(t,u)$ the set of users present in the same cluster as user u at time t, ratings at a time point T leveraging on temporal history till time T can be predicted as

Table 2: Ratings in CF: Clustering buckets $C(u)$

$$r_{u,i}(T) = \frac{\sum_{t \leq T} \sum_{v \in U(t,i) \cap C(t,i)} w_t \text{sim}(u_t, v_t) r_{v,i}}{\sum_{t \leq T} \sum_{v \in U(t,i) \cap C(t,i)} w_t \text{sim}(u_t, v_t)}$$

$$p_{u,i}(T) = \frac{\sum_{t \leq T} \sum_{v \in U(t,i) \cap C(t,i)} w_t \text{sim}(u_t, v_t) r_{v,i}}{\sum_{t \leq T} \sum_{v \in C(t,i)} w_t \text{sim}(u_t, v_t)}$$

Latent factor model: Without clustering information, the vanilla model with latent factors q_i for item i and p_u for user u is $\hat{r}_{u,i} = \mu + b_i + b_u + q_i^T p_u$, solved either through stochastic gradient descent or alternating least squares [22]. Letting $\mathcal{A}$ to be a set of attributes and a a cluster for $\mathcal{A}$, user persona information can be incorporated into the above model by

(1) adjusting for biases per cluster.
(2) enhancing user representation in the form of latent factors for cluster memberships learnt with $y_a \in \mathcal{A}$–a latent factor for each cluster a in set of characterizations [22].
(3) hard wiring clustering information as features in the form of an enhanced user feature with a latent component p_u concatenated with known added features $\tilde{p}_u$.
(4) training latent factor model per cluster with c_a being clusters corresponding to some attribute a; $\mathbb{I}_{u \in c_a}$ equals 1 if user u is in cluster c_a, 0 otherwise.

Table 3 below describes the enhanced rating models for each case described above.

Table 3: Ratings in CF: Adding Persona Clustering to Vanilla

1.	$\hat{r}_{u,i} = \mu + b_i + b_u + \sum_{a \in \mathcal{A}(u)} b_a + q_i^T p_u$
2.	$\hat{r}_{u,i} = \mu + b_i + b_u + q_i^T (p_u + \sum_{a \in \mathcal{A}(u)} y_a)$
3.	$\hat{r}_{u,i} = \mu + b_i + b_u + \tilde{q}_i^T (p_u : \tilde{p}_u)$
4.	$\hat{r}_{u,i}^{c_a} = \mu_{c_a} + b_i + \mathbb{I}_{u \in c_a} b_{c_a} + \mathbb{I}_{u \in c_a} q_i^T p_{c_a}$

CONCLUDING REMARKS

This work offers temporally evolving personas that lend new perspectives and actionable insights into behavioral patterns of VoD users as they age in the system. As highlighted, our personas do possess the cluster stability on a macro level, while being able to represent dynamic niche characterizations at the same time. Our mixture approach together with the choices of granularity and timeline of comparison and the engineered features give rise to a consistent and robust latent model. That is insights derived from a study of user personas at any time point are also likely to apply to future clusters and models built using these clusters. Information from efficiently built personas can achieve a much practical and vital *relevance-scalability-interpretability* tradeoff in recommendations, highlighted in the work with predictive models that are trained and tested on VoD data. An untapped area of application is churn analysis, see [21], aiming to improve user retention and interest. One can create user buckets based on longevity in system or use existing personas to predict when users slip into a state of inactivity in the system. A potential future direction also includes a possible tradeoff between privacy and predictive power in models based on persona features. Finally, the methods, guarantees and perspectives from this work can be extended to other domains of personalization and can be realized in a host of other predictive tasks.

REFERENCES

[1] Rakesh Agrawal, Manish Mehta, John C Shafer, Ramakrishnan Srikant, Andreas Arning, and Toni Bollinger. 1996. The Quest Data Mining System.. In *KDD*, Vol. 96. 244–249.
[2] Greg M Allenby and Peter E Rossi. 1998. Marketing models of consumer heterogeneity. *Journal of econometrics* 89, 1 (1998), 57–78.
[3] Asim Ansari, Skander Essegaier, and Rajeev Kohli. 2000. Internet recommendation systems. *Journal of Marketing research* 37, 3 (2000), 363–375.
[4] Chidanand Apte, Bing Liu, Edwin PD Pednault, and Padhraic Smyth. 2002. Business applications of data mining. *Commun. ACM* 45, 8 (2002), 49–53.
[5] Judy Bayer. 2010. Customer segmentation in the telecommunications industry. *Journal of Database Marketing & Customer Strategy Management* 17, 3-4 (2010), 247–256.
[6] David Besanko, Jean-Pierre Dubé, and Sachin Gupta. 2003. Competitive price discrimination strategies in a vertical channel using aggregate retail data. *Management Science* 49, 9 (2003), 1121–1138.
[7] Amit Bhatnagar and Sanjoy Ghose. 2004. A latent class segmentation analysis of e-shoppers. *Journal of Business Research* 57, 7 (2004), 758–767.
[8] Derrick S Boone and Michelle Roehm. 2002. Retail segmentation using artificial neural networks. *International journal of research in marketing* 19, 3 (2002), 287–301.
[9] Deepayan Chakrabarti, Ravi Kumar, and Andrew Tomkins. 2006. Evolutionary clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 554–560.
[10] Mu-Chen Chen, Ai-Lun Chiu, and Hsu-Hwa Chang. 2005. Mining changes in customer behavior in retail marketing. *Expert Systems with Applications* 28, 4 (2005), 773–781.
[11] Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)* (1977), 1–38.
[12] José G Dias and Jeroen K Vermunt. 2007. Latent class modeling of website users' search patterns: Implications for online market segmentation. *Journal of Retailing and Consumer Services* 14, 6 (2007), 359–368.
[13] David B Dunson. 2000. Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62, 2 (2000), 355–366.
[14] Michael D Ekstrand, John T Riedl, and Joseph A Konstan. 2011. Collaborative filtering recommender systems. *Foundations and Trends in Human-Computer Interaction* 4, 2 (2011), 81–173.
[15] Nir Friedman and Stuart Russell. 1997. Image segmentation in video sequences: A probabilistic approach. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 175–181.

[16] Johanna Gummerus, Veronica Liljander, Minna Pura, and Allard Van Riel. 2004. Customer loyalty to content-based web sites: the case of an online health-care service. *Journal of services Marketing* 18, 3 (2004), 175–186.
[17] Donald Hedeker. 2003. A mixed-effects multinomial logistic regression model. *Statistics in medicine* 22, 9 (2003), 1433–1446.
[18] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*. Ieee, 263–272.
[19] Tianyi Jiang and Alexander Tuzhilin. 2006. Segmenting customers from population to individuals: Does 1-to-1 keep your customers forever? *IEEE Transactions on Knowledge and Data Engineering* 18, 10 (2006), 1297–1311.
[20] Tianyi Jiang and Alexander Tuzhilin. 2009. Improving personalization solutions through optimal segmentation of customer bases. *IEEE transactions on knowledge and data engineering* 21, 3 (2009), 305–320.
[21] Komal Kapoor, Mingxuan Sun, Jaideep Srivastava, and Tao Ye. 2014. A hazard based approach to user return time prediction. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1719–1728.
[22] Yehuda Koren, Robert Bell, Chris Volinsky, et al. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.
[23] Gueorgi Kossinets and Duncan J Watts. 2006. Empirical analysis of an evolving social network. *science* 311, 5757 (2006), 88–90.
[24] Wei Li, Xuemei Wu, Yayun Sun, and Quanju Zhang. 2010. Credit card customer segmentation and target marketing based on data mining. In *Computational Intelligence and Security (CIS), 2010 International Conference on*. IEEE, 73–76.
[25] Snigdha Panigrahi, Nadia Fawaz, and Ajith Pudhiyaveetil. 2019. Temporal Evolution of Behavioral User Personas via Latent Variable Mixture Models.

https://arxiv.org/abs/1704.07554. *arXiv preprint arXiv:1704.07554* (2019).
[26] Michael J Pazzani and Daniel Billsus. 2007. Content-based recommendation systems. In *The adaptive web*. Springer, 325–341.
[27] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*. ACM, 285–295.
[28] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. 2007. Collaborative filtering recommender systems. In *The adaptive web*. Springer, 291–324.
[29] Wendell R Smith. 1956. Product differentiation and market segmentation as alternative marketing strategies. *Journal of marketing* 21, 1 (1956), 3–8.
[30] William Trouleau, Azin Ashkan, Weicong Ding, and Brian Eriksson. 2016. Just one more: Modeling binge watching behavior. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1215–1224.
[31] Jeroen K Vermunt and Jay Magidson. 2002. Latent class cluster analysis. *Applied latent class analysis* 11 (2002), 89–106.
[32] Michel Wedel and Wagner A Kamakura. 2012. *Market segmentation: Conceptual and methodological foundations*. Vol. 8. Springer Science & Business Media.
[33] Jing Wu and Zheng Lin. 2005. Research on customer segmentation model by clustering. In *Proceedings of the 7th international conference on Electronic commerce*. ACM, 316–318.
[34] XianXing Zhang, Yitong Zhou, Yiming Ma, Bee-Chung Chen, Liang Zhang, and Deepak Agarwal. 2016. GLMix: Generalized Linear Mixed Models For Large-Scale Response Prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 363–372.