Data Science in Service of Performing Arts: Applying Machine Learning to Predicting Audience Preferences

Jake Abernethy
University of Michigan
Ann Arbor, MI
jabernet@umich.edu

Cyrus Anderson
University of Michigan
Ann Arbor, MI
andersct@umich.edu

Alex Chojnacki
University of Michigan
Ann Arbor, MI
thealex@umich.edu

Chengyu Dai
University of Michigan
Ann Arbor, MI
daich@umich.edu

John Dryden
University of Michigan
Ann Arbor, MI
jcdryden@umich.edu

Eric Schwartz
University of Michigan
Ann Arbor, MI
ericmsch@umich.edu

Wenbo Shen
University of Michigan
Ann Arbor, MI
shenwb@umich.edu

Jonathan Stroud
University of Michigan
Ann Arbor, MI
stroud@umich.edu

Laura Wendlandt
University of Michigan
Ann Arbor, MI
wenlaura@umich.edu

Sheng Yang
University of Michigan
Ann Arbor, MI
physheng@umich.edu

Daniel Zhang
University of Michigan
Ann Arbor, MI
dtzhang@umich.edu

ABSTRACT
Performing arts organizations aim to enrich their communities through the arts. To do this, they strive to match their performance offerings the taste of those communities. Success relies on understanding audience preference and predicting their behavior. Similar to most ecommerce or digital entertainment firms, arts presenters need to recommend the right performance to the right customer at the right time. As part of the Michigan Data Science Team (MDST), we partnered with the University Musical Society (UMS), a non-profit performing arts presenter housed in the University of Michigan, Ann Arbor. We are providing UMS with analysis and business intelligence, utilizing historical individual-level sales data. We built a recommendation system based on collaborative filtering, gaining insights into the artistic preferences of customers, along with the similarities between performances. To better understand audience behavior, we used statistical methods from customer-base analysis. We characterized customer heterogeneity via segmentation, and we modeled customer cohorts to understand and predict ticket purchasing patterns. Finally, we combined statistical modeling with natural language processing (NLP) to explore the impact of wording in program descriptions. These ongoing efforts provide a platform to launch targeted marketing campaigns, helping UMS carry out its mission by allocating its resources more efficiently. Celebrating its 138th season, UMS is a 2014 recipient of the National Medal of Arts, and it continues to enrich communities by connecting world-renowned artists with diverse audiences, especially students in their formative years. We aim to con-
tribute to that mission through data science and customer analytics.

Keywords
Machine Learning, Collaborative Filtering, Natural Language Processing

1. INTRODUCTION
The University Musical Society (UMS) is a 501(c)(3) non-profit performing arts organization affiliated with the University of Michigan in Ann Arbor. UMS seeks to engage the community with new and innovative artists. Founded in the winter of 1880, UMS is one of the oldest non-profit performing arts presenters in the country and presents 65 to 75 shows per year in a variety of genres. UMS holds events in the Ann Arbor area across multiple venues. The largest is Hill Auditorium, which has a maximum capacity of 3,536 and accommodates world-class musicians and performers.

On the one hand, UMS is a non-profit organization, depending on donations and grants for funding. On the other, it depends on generating revenue through ticket sales. Much like many other businesses, UMS uses common tactics in multi-channel retail, online marketing, and digital content distribution to maximize its marketing strategy. UMS maintains a large purchase history database, containing information such as what marketing material was sent to customers, how tickets were purchased, and how users' purchasing habits have changed over time.

In this paper, we explore how this rich dataset can be further utilized to gain a better understanding of the UMS audience to eventually guide marketing and event programming decisions. We show how machine learning and other

data analysis techniques can be applied to gain insight into purchasing patterns, as well as suggest improvements in future marketing strategy. Section 2 describes the main UMS dataset, as well as additional data collected, and provides model-free visualizations of the data patterns. Section 3 examines the performances themselves, specifically through the language used in descriptions. Section 4 uses collaborative filtering techniques to analyze purchase history data. Section 5 shows how Markov chains can be used to model the data, linking to a broader stream of statistical modeling in customer-base analysis. Finally, we end in Section 6 with conclusions and future work.

2. DATA

UMS has provided us with five years of anonymized ticket purchasing data from 2011 to 2015. We use the first three years for training, and holdout the most recent two years for test data. This data set includes over 190,000 transactions from 48,000 users, totaling over $13 million in revenue. Each transaction contains the following pieces of information:

- UMS account number
- Date the account was created (or digitalized, for pre-digital accounts)
- Customer type (either a household, individual, or organization)
- Name, date, and venue of the performance
- Price of the tickets and number of seats sold
- Information about whether the ticket was part of a promotion or special offer
- Mode of sale (how the ticket was bought, for example, via the UMS website or over the phone)
- Date of the order
- Postal code of the customer

One particularly important dynamic captured in the dataset is information about subscriptions. At the beginning of each performance season, UMS offers customers the chance to purchase subscriptions, which are packages of tickets for a series of shows. Usually each subscription is thematic and includes related shows. Some current subscription series are Dance and Theater, Jazz, and Choral / Vocal. Subscriptions often need to be treated differently during data analysis because when customers buy a series, they do not individually select each show. This changes their purchasing patterns and behaviors.

2.1 Data Visualizations

Data visualizations reveal patterns in terms of customer activity, revenue composition, and purchase patterns. Figure 1 shows select statistics and visualizations from the UMS dataset. We define the activity duration of each customer as the time span between the first purchase and the last purchase. Interestingly, 66% percent of customers only made only one purchase, never returning to buy another day. This suggests that there is great potential to convert these one-time customers into frequent customers. The importance of frequent customers is highlighted in the pie chart of the revenue composition. This pie chart shows that more than one-third of all revenue is from subscription purchases. This relatively large percentage of revenue is from a relatively small fraction of customers, as only 5.6% customers are subscription buyers. High purchase concentration is to be expected with a large heterogeneous audience. Finally, we find that the non-subscription purchases and subscription purchases have distinct time patterns. Non-subscription (regular ticket) purchases are distributed throughout the year. There are almost no non-subscription purchases in the months of May, June, and July since there are generally no performances scheduled during these months. In addition, at the opening days of various ticket groups, there is always a rush for tickets. Contrary to the non-subscription purchases, subscription purchases are concentrated from April to June when there are very few performances. This is understandable since the subscriptions are for the coming season from September to April of next year. The rushes for tickets in both patterns imply that performances organized by UMS are very popular and well-received.

2.2 Performance Descriptions

To augment the purchasing data provided by UMS, we collect descriptions for each performance. These descriptions are written by UMS and are publicly available on the UMS website. The average description length is 164.4 tokens, where each token is either a word or a symbol. After the descriptions are collected, they are manually categorized into seven categories: Orchestra, Chamber, Jazz, Theater, Dance, Choral, and Other. Figure 2 shows the percentage of performance descriptions in each category, as well as the total number of seats sold in each genre. These charts highlight some discrepancies in the number of shows offered for each genre and the popularity of each genre. For example, 9.1% of performances are in the Dance category, while 21.7% of the total number of seats sold are to Dance shows.

3. WORDING IN PERFORMANCE DESCRIPTIONS

One of the marketing tools that UMS has at their disposal is the written performance descriptions distributed via programs, brochures, posters, and online media. Analyzing the writing style of these descriptions can provide insight into why customers choose to see the shows that they do. They also help to explain similarity of performances and artists.

Several metrics exist to measure the style of a piece of writing. One of these is readability, which assigns a reading grade level to a piece of writing. Readability measures are often used by educators and librarians to measure the reading level of a piece of literature. One standard measure of readability is the Flesch-Kincaid Grade Level [4]. This is calculated according to the following formula:

$$0.39 \frac{\text{total words}}{\text{total sentences}} + 11.8 \frac{\text{total syllables}}{\text{total words}} - 15.59 \quad (1)$$
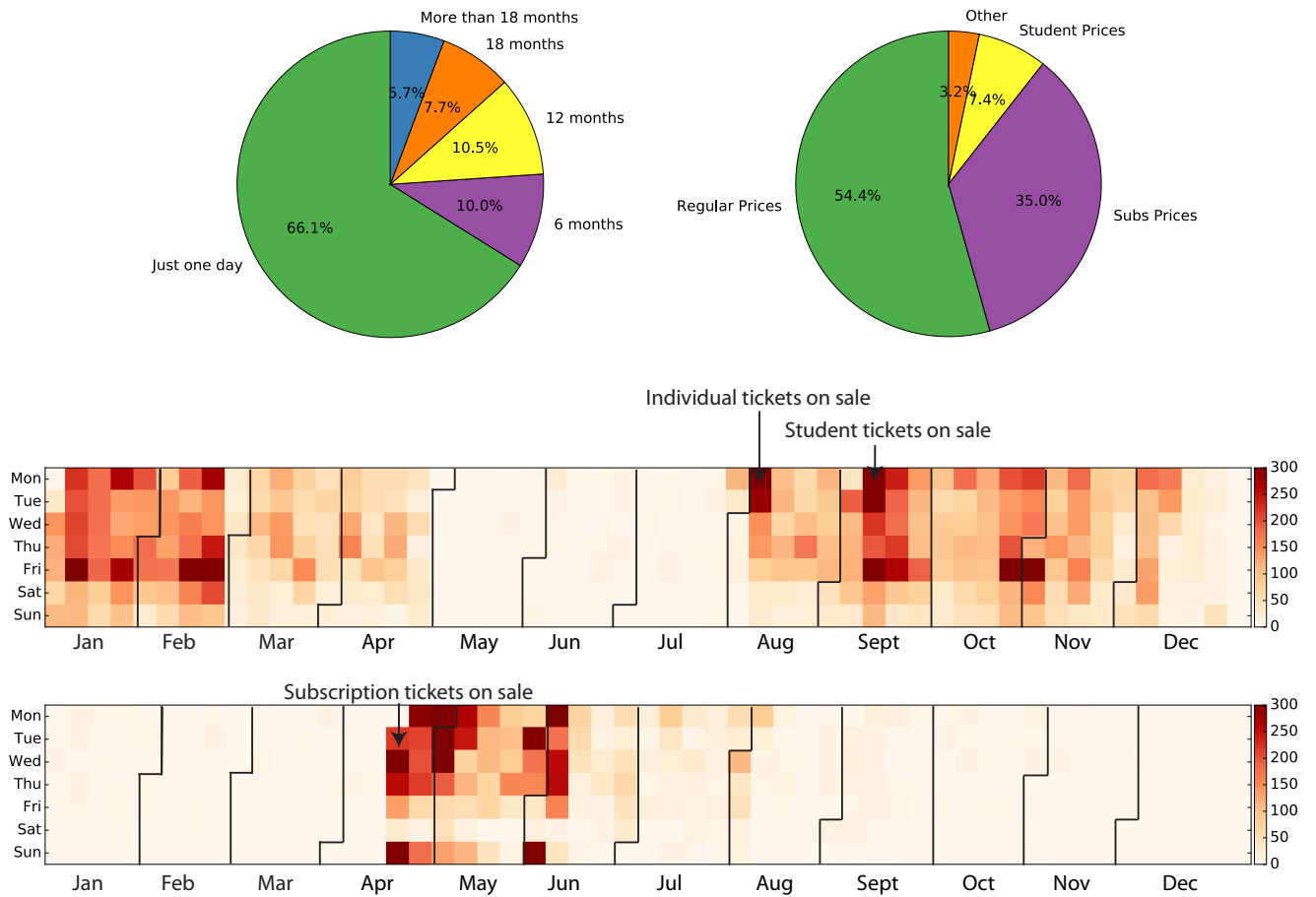
Figure 1: Top left pie chart: the duration of customers' activities. The duration is defined as the time span between the customer's first purchase and his or her last purchase. Top right pie chart: UMS revenue composition by the price groups: regular price, subscription price, student price, or other. Top heat map: the number of non-subscription tickets bought throughout the year. Darker colors indicate that more tickets were bought on that day. Bottom heat map: the number of subscription tickets bought throughout the year. For both heatmaps, purchase data of calendar year 2013 is used.

Another text-based style metric is formality. Formality attempts to quantify the preciseness and informativeness of a statement. The Heylighen and Dewaele measure of formality is calculated according to the following formula [3]:

$$
\begin{aligned}
(\text{noun freq.} + \text{adjective freq.} + \text{preposition freq.} + \\
\text{article freq.} - \text{pronoun freq.} - \text{verb freq.} - \\
\text{adverb freq.} - \text{interjection freq.} + 100)/2
\end{aligned} \tag{2}
$$

A third text-based style metric is the length of a document. Figure 3 shows the relationship between these three metrics and the percentage of seats sold for every show. Only tickets that are not part of a subscription are included for this analysis. A 2-dimensional line of best fit is fitted to the readability plot, while 1-dimensional lines of best fit are fitted to the other plots. These plots show that on average, as formality and description length increase, there is an increase in the number of tickets sold for that performance. There is a similar trend for readability, but this data is better fitted with a polynomial curve, indicating that the optimal

readability for a program description is around grade level 15.

4. COLLABORATIVE FILTERING MODELING FOR PURCHASE HISTORY

In addition to analyzing performance descriptions, another way to understand historical purchasing patterns is through the use of a model based on collaborative filtering. [1, 5]. Popularized by the Netflix Prize in 2006, collaborative filtering is a technique that automatically matches customers to performances they might enjoy based on information such as purchase history and customer or content similarity. Unlike Netflix, where content remains available to all users, a performing arts organization has live performances constraints that come and go. So the collaborative filtering approach applied in practice will have to make recommendations among a small set of possible shows remaining in the given season. We return to this as future work in the final section.

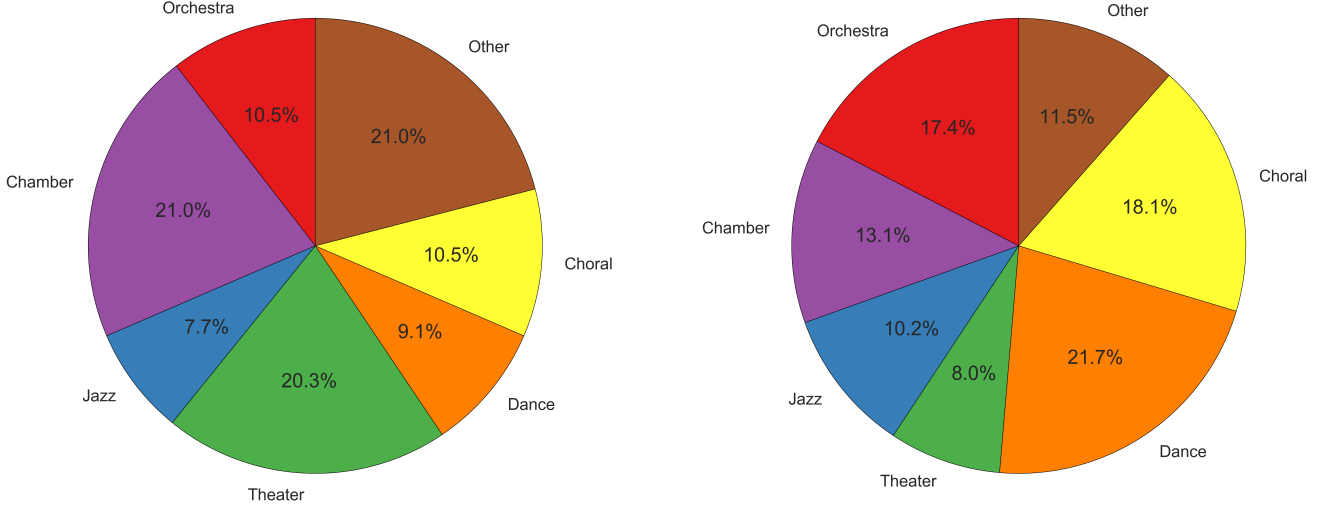Another interesting component of this collaborative filter-

Figure 2: The left pie chart shows performance descriptions broken down by category. The right pie chart compares the total number of seats sold in each genre.
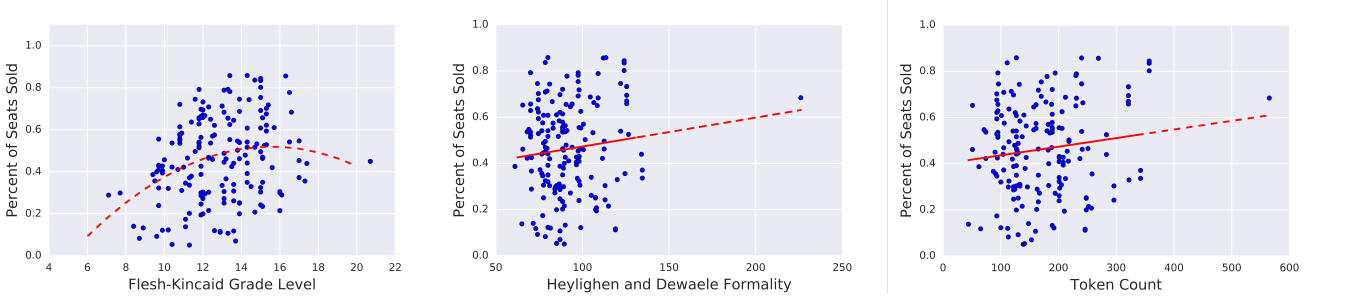

Figure 3: Style metrics applied to the written performance descriptions. These graphs show the relationship between three style metrics (readability, formality, and length) and the percentage of seats sold for that show. These plots do not include tickets sold as part of a subscription. Each graph is fitted with either a 2-dimensional or a 1-dimensional line, to show the trends in the data. The Pearson correlation coefficients for these scatter plots are, from left to right, 0.26, 0.12, and 0.14.

ing modeling problem is the diversity of performances across genres. Due to the great variety of UMS performances, it is difficult to define a reliable similarity metric that can be used to compare performances. Therefore, our system relies primarily on well-documented recent purchase history. Applying matrix factorization (MF) to collaborative filtering achieved great success in both academic research and industrial application. We adopt this approach, which is formally introduced below.

4.1 Methodology

First, we represent purchase history as a binary-valued matrix $\mathbf{X}$ with dimensions N_c x N_p, where N_c and N_p are the numbers of unique customers and performances, respectively, in the dataset. $\mathbf{X}_{ij} = 1$ indicates that customer i purchased a ticket for performance j , and $\mathbf{X}_{ij} = 0$ indicates that no ticket was purchased.

Every customer's willingness to purchase a ticket for a specific performance can be expressed in terms of how well that performance matches each of the artistic styles the customer

is interested in. We also include constant bias terms to model that some customers have more buying power and similarly some performances are more popular than others. This gives us the following equation.

$$\mathbf{X}_{ij} \approx \sum_l \mathbf{L}_{il}\mathbf{R}_{jl} + \mathbf{B}_i^L + \mathbf{B}_j^R \tag{3}$$

Translating into the language of matrices,

$$\mathbf{X} \approx \mathbf{L}\mathbf{R}^T + \mathbf{B}^L + \mathbf{B}^R \tag{4}$$

A more formal way of posing the problem is to express it in terms of a regularized Frobenius norm optimization prob-
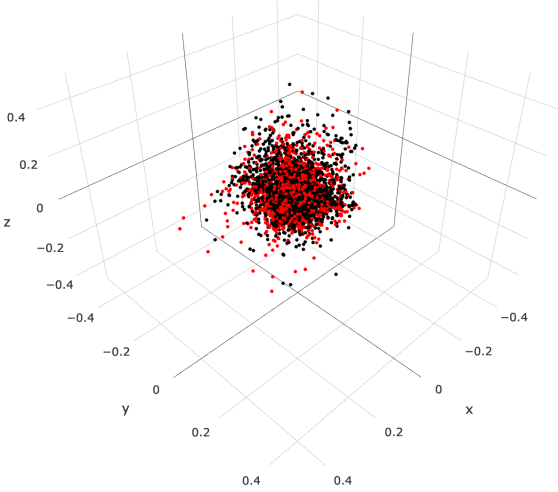
Figure 4: Visualizing artistic preferences of both students (red) and the general public (black). This scatter plot shows the embeddings of different customers in a latent space representing artistic style preferences.
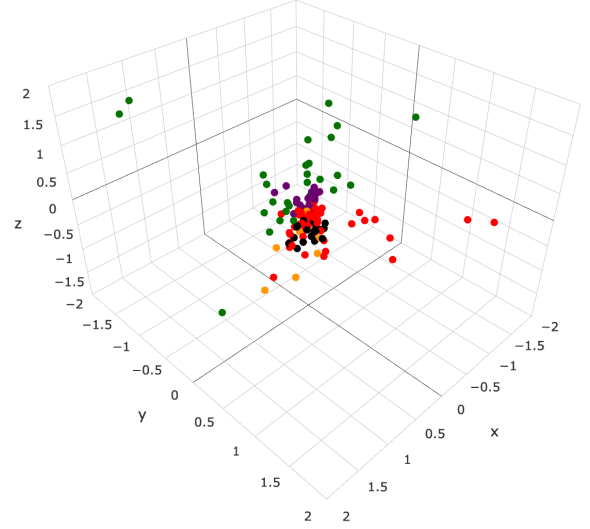

Figure 5: Visualizing artistic styles for performances that belong to different subscription series. Green: Choral Unions, Purple: Chamber Music, Black: Dance, Orange: Jazz, Red: Others

lem, as follows:

$$\underset{\mathbf{L},\mathbf{R},\mathbf{B}^L,\mathbf{B}^R}{\text{minimize}} \quad \|\mathbf{X} - \mathbf{L}\mathbf{R}^T - \mathbf{B}^L - \mathbf{B}^R\|_F^2 + \frac{\lambda}{2}\|\mathbf{L}\|_F + \frac{\lambda}{2}\|\mathbf{R}\|_F$$

$$\text{subject to} \quad \mathbf{L} \in \mathbb{M}_{N_c,L}, \mathbf{R} \in \mathbb{M}_{N_p,L}, \mathbf{B}^L, \mathbf{B}^R, \in \mathbb{M}_{N_c,N_p}$$

In the above equation, $\mathbf{B}^L$ / $\mathbf{B}^R$ are column-wise / row-wise constant matrices respectively.

This optimization problem is closely related to singular value decomposition (SVD). It is generalized to include a constant term in the formula and to make training possible even when some of the data is not present.

Standard approaches to training this model include stochastic gradient descent (SGD) and alternating least square (ALS). We adopt the latter approach, because empirically we found that SGD with all random initialization fails to converge when the matrix is tall-and-skinny (when the number of customers greatly exceeds the number of performances). Note that if we fix either $\mathbf{L}$ or $\mathbf{R}$ and optimize with respect to the other, then the problem becomes a standard quadratic matrix optimization problem which can be solved by least squares. Alternating between fixing either $\mathbf{L}$ or $\mathbf{R}$ and solving for the other, the algorithm converges reasonably well after a few hundred steps for practical use and further study.

4.2 Interpreting the Factorization

This simple matrix factorization model can provide insight into customers and performances, based solely on the purchase history. $\mathbf{L}$ and $\mathbf{R}$ have clear geometric interpretations as collections of vectors of each customer or performance's position in the latent space of artistic style. For each customer vector, the magnitude of the vector is related to the total purchases of the customer, while the direction of the

vector represents the customer's artistic style preferences. For each performance vector, the magnitude of the vector is related to the number of seats sold for the performance, while the direction of the vector represents the artistic style of the performance. Each point in the latent space represents a vector connecting the origin to that point.

4.2.1 Customer Preference Analysis

These customer vectors can be used to explore customer heterogeneity in taste. Since UMS is affiliated with the University of Michigan, we will use those vectors to compare university students' artistic preferences with the preferences of the non-student general public. This analysis furthers one of the core goals of UMS, which is to enrich students' cultural experiences. To investigate this question, we restrict our model to the three most significant latent dimensions, to allow for easy visualization. We separate out students by looking for customers who have purchased tickets from a student promotion or at a student price.

From the visualization of Figure 4, we conclude that both students and the general public have large in-group variation of artistic style preferences. This is reflected in the diverse directions of the latent vectors; they cannot be simply clustered into a few definite groups. The great spectrum of performance types that UMS currently provides serves this diverse community well.

More importantly, our result also suggests that in general students do not have different artistic preferences than regular customers. However, the latent vectors representing student customers have smaller magnitudes, meaning that students in general have less willingness to purchase. This certainly has face validity, as students typically have smaller incomes than working adults. This analysis, while not able

to make strong causal claims, suggests that UMS may be wise to continue their current student discount pricing policy, which is giving general discounts to students while not limiting the discounts to any specific types of performances.

4.2.2 Performance Style Analysis

Another question of interest is if performances in the same subscription series are also similar in terms of artistic style. We investigate this question by specifically analyzing the purchase data excluding subscription tickets. Performance vectors with similar directions indicate similar artistic styles, as shown in Figure 5. In general, Jazz (orange) performances and Choral Union (green) series do show similarity with others of their kind. Chamber (purple) and Dance (black) performances are represented by latent vectors with very small magnitudes due to limited venue sizes relative to other genres. Additionally, a few Other (red) performances approximately form a straight line, showing very related style. Many of these are annually recurring performances of Handel's Messiah, an Ann Arbor traditional holiday show that many people attend. In general, we found individual ticket purchase patterns match the hand-picked subscription series based on genre, as expected.

4.3 Discussion on Missing Data

As most of UMS's customers come from the local college town of Ann Arbor, many of them have only been residents of the town for a few years. This must be taken into account when considering a customer's willingness to purchase a ticket. If a person is not living in Ann Arbor at the time of the performance, he or she should not be considered unwilling to purchase the ticket. Instead, that information should be considered missing. A reasonable approximation of the customer's arrival time is the customer's UMS account creation date.

Our training method ALS supports the use of missing entries, as long as we solve for each L_i vector sequentially, by doing regression only on the performance entries j that are not missing for a particular customer (likewise for training R_j). After this correction, the model contains significantly fewer artifacts and is more interpretable.

5. CUSTOMER LIFECYCLE ANALYSIS

Collaborative filtering provides a way to analyze the artistic preferences of users. Another way to understand transaction history is to analyze dynamics of purchase behavior throughout a customer's lifecycle, with an aim towards calculating customer lifetime value. This information is particularly helpful for identifying which customers will be most valuable in the future (and which should be ignored). Knowing this can improve the efficiency of targeted promotional efforts, from costly as postal mail catalogs to more personalized emails.

We draw on a rich literature of customer-base analysis and customer relationship management in marketing [2]. To begin, we will use lifecycle consists of three stages: 'active,' 'inactive', and 'dead'. While these can be inferred from the data as latent states, we provide an analysis based on observed stages. If a customer purchases any tickets in a year,

he or she is associated with the active state. The inactive state means the customer purchases no tickets in a given year, and the dead state is reached if a customer is inactive for two or more years. These three lifecycle states can be modeled using a Markov chain. A Markov chain models system dynamics as a set of states, where at the end of each season, a customer can transition from one state to any other state according to a set of probabilities. The transition probabilities depend only on the current state. Fitting the model then entails finding the nonzero transition probabilities. Here we define a Markov chain for the customer lifecycle consisting three states.

The fitted Markov chain is shown in Figure 6. Nearly half the customers in a given cohort become inactive, and of the inactive customers, most do not return, i.e., churned. We see that UMS sustains this turnover with a number of completely new customers each year. Nevertheless, understanding customer churn is important in balancing customer acquisition and retention efforts.

6. CONCLUSION & FUTURE WORK

The work presented in this paper analyzes historical purchase history data for UMS. We analyze the correlation between program description wording and ticket sales and utilize collaborative filtering techniques to understand customer artistic preferences and performance styles. We also model the lifecycle of customers as a Markov chain, showing how customer activity falls into regular patterns.

This work is ongoing in all three fronts. The Markov model of customer lifecycles presented here relies on 'observed' states with predefined labels. But current future work will infer 'unobserved' states using hidden Markov models. A special case of these models is known as 'Buy till you die' models [2]. By allowing for customer churn to be inferred, we will aim for deeper understanding of purchase behavior throughout the customer lifecycle.

Moving forward, we are working on building an intelligent recommendation system that will be able to predict which customers will enjoy new shows. While this builds on collaborative filtering techniques, it is made more difficult by the 'cold start problem.' This occurs when there is no available purchasing information for previously unseen shows.

One solution to this problem is to compare new shows with previous shows that are similar in content. This requires having some means of assessing similarity between shows. We have begun working on a way to do this using topic modeling on performance descriptions. Topic models extract a set of topics from a corpus of training documents. Each topic is a distribution of words, and it describes which words are used most frequently in that particular topic. Then, each document is assigned a distribution of topics. Documents that have similar topic distributions should have similar content. Topic modeling can be used to assign topic distributions to each performance description and find similar performances.

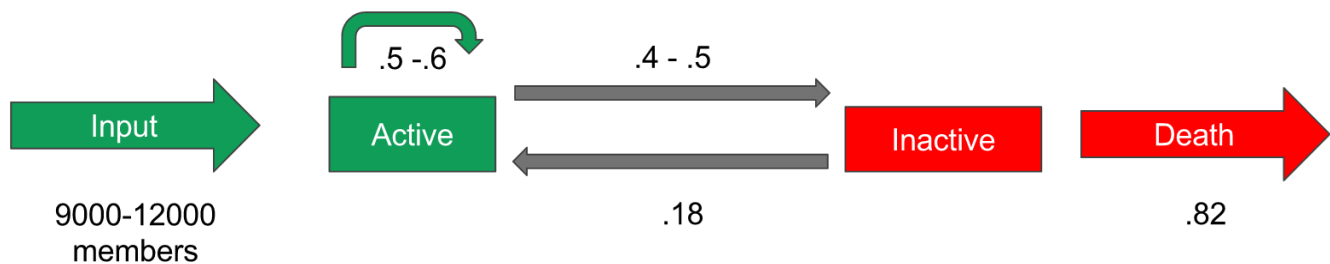In order to have an effective topic model, it must be trained

Figure 6: A Markov chain showing the lifecycles of UMS customers. The three stages in this cycle are "active", "inactive", and "death". The probabilities of moving from one stage to another are shown in the diagram. The transitions occur annually.

over a large corpus of documents. The performance descriptions are short and do not provide enough information to train a good model. We have begun collecting a much larger corpus of performance-related documents to train a model. This larger corpus consists of Wikipedia pages that are relevant to different performance genres, such as theater, dance, and orchestra. The larger corpus will be used to train an effective topic model, which will then be used to find similarity scores between different performance descriptions. This information will be input to the collaborative filtering algorithm to complete a fully functional recommendation system.

From a methodological perspective, there we see a promising direction in combining these three areas: topic modeling, collaborative filtering, and predictive models of repeat customer behavior. Such an integration across data science approaches will not only be useful to UMS and arts organizations, but to a broader methodological and customer analytics audience as well.

Our work demonstrates how data science could help non-profit organizations further achieve their missions. Data visualizations and business analytics help extract information in the underutilized data and identify the areas for improvement. The statistical models and machine learning approaches presented here are only a starting point for ways data science can help inform decision making in the performing arts.

Acknowledgments

References

[1] R. M. Bell and Y. Koren. Lessons from the netflix prize challenge. *ACM SIGKDD Explorations Newsletter*, 9(2):75–79, 2007.

[2] P. S. Fader and B. G. Hardie. Probability models for customer-base analysis. *Journal of Interactive Marketing*, 23:61–69, 2009.

[3] F. Heylighen and J.-M. Dewaele. Formality of language: definition, measurement and behavioral determinants. *Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel*, 1999.

[4] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document, 1975.

[5] Y. Koren, R. Bell, C. Volinsky, et al. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.