Alpha User Study

General Assembly – Data Science: Section 2

Gaurav Sharma, [SharmaG21@gmail.com](mailto:SharmaG21@gmail.com), June 5, 2013

**Abstract**

New users to Saavn.com (an Indian digital music startup, similar to Spotify or Pandora, as well as my employer) can be split into 2 categories: those that are retained, and those that are not. This can be seen as the case for nearly all user-facing businesses. The key in understanding the difference between these groups of users lies in each cohort's engagement. As engagement is driven, retention is driven, and thus the user base grows. I've administered a series of data studies to better understand our retained users, who I'm referring to as our "alpha users". Through these findings we hope to make some data facing product decisions that will hopefully drive our would-be non-alpha users to become highly engaged users in the future.

**Introduction**

Saavn.com is a digital music startup that brings the largest database of streaming Indian music to the fingertips of millions of users every month. Given the large and ever-growing Indian tech market, growth has never been a problem for us. Each month millions of new users try Saavn out, but while India may stand as a great growth opportunity, the majority of our early adopters to Saavn are not completely tech savvy, which makes India a very unique case. While some users are able to navigate our apps and website with ease, many new users are using Saavn's app on their first smartphone, or Saavn.com from an Internet café (both very common cases in any developing country around the world). While Saavn's monthly unique user base is large, it is largely driven by millions of brand new users on Saavn every month. A key focus in leveraging Saavn's growth to further build its user base is Retention. While the percentage of users who return to Saavn.com on a monthly basis may be low, my goal is to better understand the engagement trends of these retained users, and find a way to drive more users into this retained category.

By driving more new users into engaging with Saavn similar to how previously retained users have, we will 1) have a larger stream of returning users on a monthly basis, paired with our already large monthly new user base; 2) increase our user engagement/consumption (track streams, listening hours, playlists created, etc.); 3) drive an overall better user experience.

In terms of datasets, I was given free control in picking the set of data that I felt best suited the studies I chose. This proved very helpful in dealing with usual dataset issues such as sparsity. Nevertheless, I used multiple methods to make sure I was looking at this data in as many ways as possible. Through logistic regression, decision trees, and principle component analysis I was able to come to a few interesting conclusions regarding Saavn.com's "alpha" users, and how to drive more users into this group.

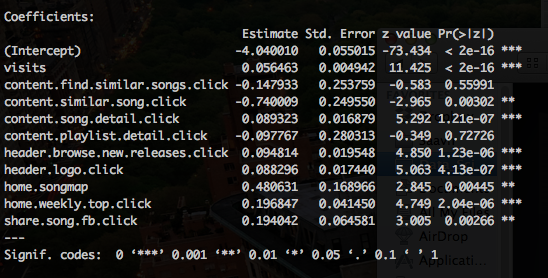
**Studies**

The dataset (alpha\_extracted.csv) being used contains rows of unique, first time users of Saavn.com on an unspecified date. Each column represents an event being tracked on Saavn.com and the amount of times each user triggers this corresponding event during a 30-day period. In addition to events, the first column refers to each user's "alpha status". A user is given an alpha value of "1" if they return to Saavn at least one time after 30 days, a user is given an alpha value of "0" if they do not. This definition of "alpha user" was chosen since our goal is to better understand why certain users are retained after 1 month, and others are not. A sample of the data is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Alpha | Visits | song.detail.click | Browse.featured playlists |
| 1 | 4 | 6 | 3 |
| 0 | 1 | 0 | 7 |

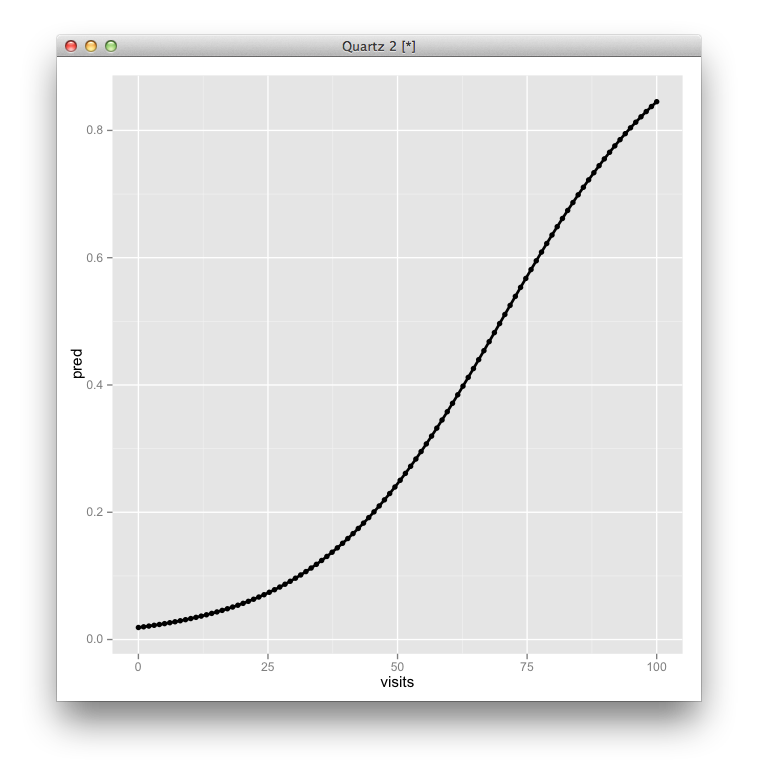
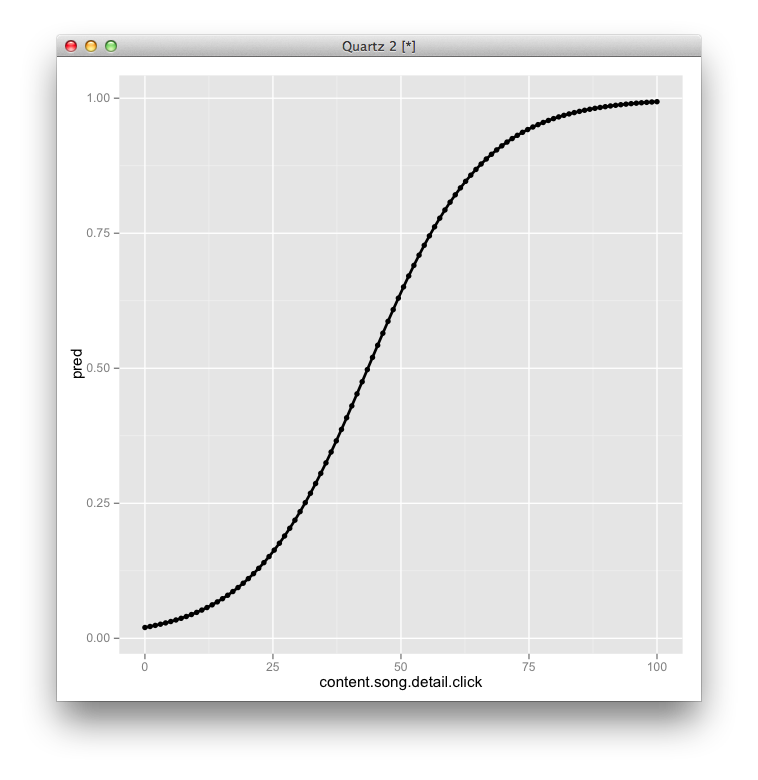
*Logistic Regression*

The first model I chose to build was a logistic regression model with regards to a users alpha status. In order to reduce error I extracted all but 11 events using feature extraction (there were nearly 300 events being tracked to begin with). In the end, I found the following events to be most statistically significant in deciding a user's alpha status:



"Content.song.detail.click", "content.similar.song.click", "Home.Weekly.Top.Click", and "Header.logo.click" were especially intriguing. While the (1) song detail and (2) similar song features are seldom used, and the specific location of the (3) "Weekly Top 15 Playlist" event is only located on the home screen, which cannot be returned to once a user begins streaming content. The clicking of our header logo might mean users are attempting to return to our homepage to browse content within the homepage. While all of the content on our homepage is available in our player page, users might simply be more comfortable with the homepage's layout.

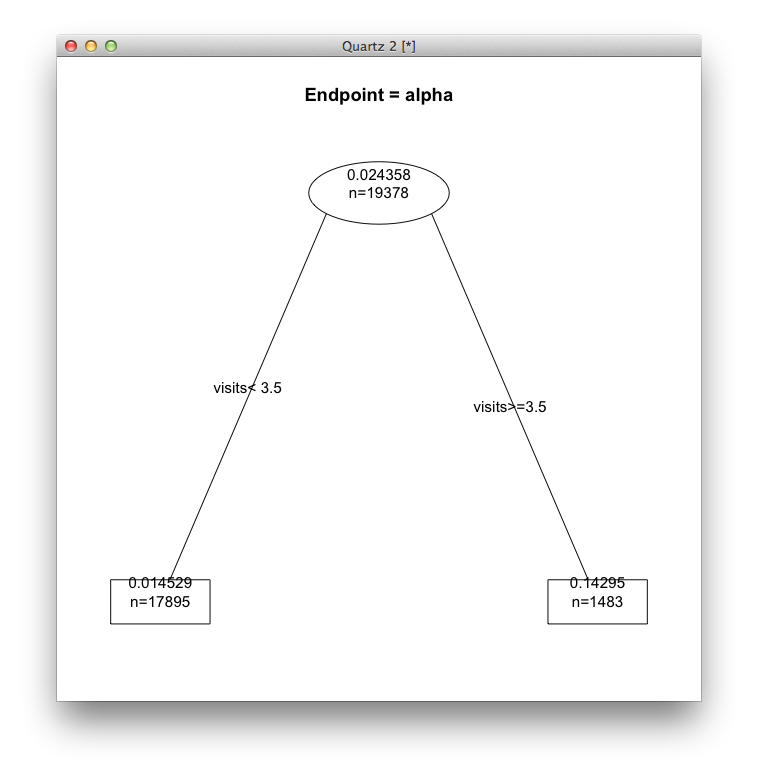
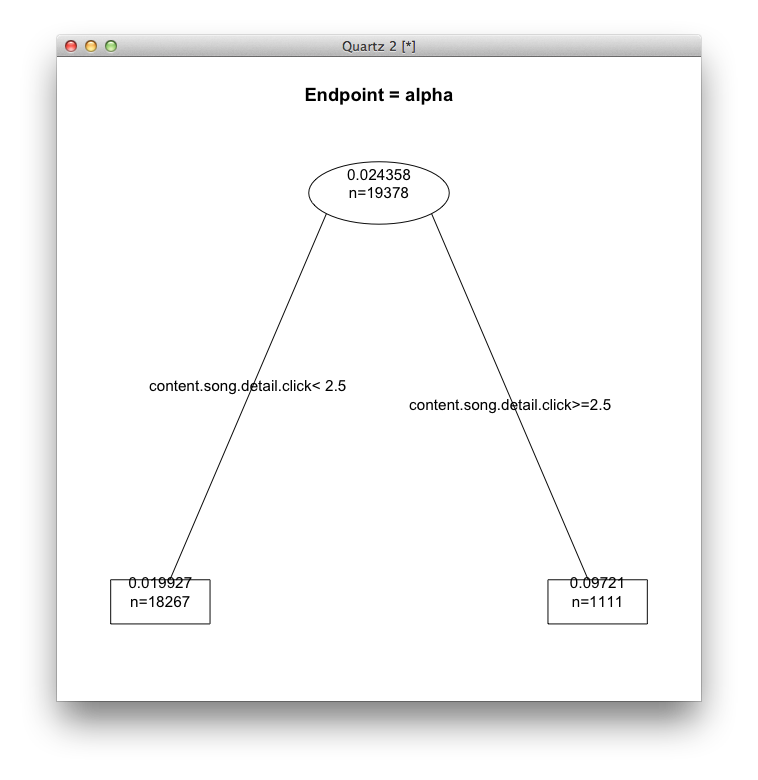
In addition to building a logistic regression model, I also attempted to predict a user's alpha status while varying a user's engagement with some of the aforementioned events. Below are the plots for varying visits and song detail clicks.



*Decision Trees*

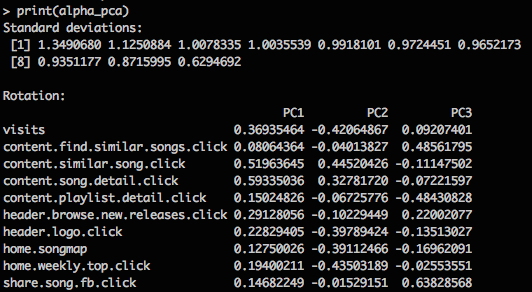
While the logistic regression model was helpful in decreasing the feature space to a group of more statistically significant events, I still wished to look a little deeper into the user engagement with regards to these specific events. Using recursive partitioning I was able to create a pair of decision trees specific to certain events.

Across all users in the dataset, only 2.4% were considered "alpha users". Through recursive partitioning I was able to find a split at 3.5 visits, increasing the mean "alpha usage" to 14.3% when a user is above 3.5 visits / mo., and decreasing to 1% when a user is below it. Similarly, users who clicked "song detail" more than 2.5 times / mo. had an alpha usage 10% of the time, while users with less than the 2.5 song detail clicks were alpha only 1% of the time. According to my dec-tree these were the two more significant variables in creating splits. Below are screenshots of plots from the decision tree.



*Principal Component Analysis*

In addition to logistic regression modeling and decision trees I wanted to conduct principal component analysis as a form of dimensionality reduction. Although I'm not looking to reduce the dimensions in my dataset, understanding how the principal components are built helps give an idea of which features might be more important than others in the extracted feature space. In looking at the first few principal components we can see a large share of "content.song.detail.click", "content.similar.song.click", and visits. This verifies findings from the logistic regression model built earlier.



**Findings**

Saavn.com has had lots of success in attracting new, first time users every day. Our next step is to better retain these users, and thus grow our monthly user base at an even faster rate. My goal was to better understand the engagement of our retained users vs. the engagement of our un-retained users. In building a logistic regression model I found a clear connection between the usage of "song detail" pages, "similar song" links, general engagement on Saavn.com's homepage as a browsing medium, and a user's retention, or alpha status. In extracting the rest of the events being tracked from my feature space, I was able to conduct additional studies on our more significant features using recursive partitioning and principal component analysis.

Being able to identify the key differentiating features utilized by alpha users allows us to better understand how users are attempting to use Saavn, and gives the product team direction in finding a better way to engage our users. A few examples come to mind when looking at my findings:

1. Saavn.com, just like Spotify or Pandora, currently does not have a fully built out "Browse" functionality. The most obvious replacements for browse that we feature are lists of editor-curated playlists, new album releases, and search. Interestingly, the "detail/info" pages for each song contain clickable metadata such as the song's singer, songwriter, movie it's featured in, and year it was released. Logistic Regression and PCA showed "content.song.detail.click" to be one of the more significant features with regards to a user's alpha status. Furthermore, through recursive partitioning we found that a user's chance of becoming an alpha user increases from 2% to nearly 10% if they visit more than 2.5 song detail pages per month. Most first time users do not realize these pages exist, and thus we now have justification in surfacing this feature moving forward.
2. Saavn's home page looks completely different when a user first visits our website vs. once they begin to stream content. Once content is clicked on the homepage, users live in a "player page" until they leave the site. While "player page" contains all the content available on the homepage, it is organized and illustrated in a completely different way. Furthermore, if a user clicks the Saavn logo in the top left of the player page, the natural assumption would be that it would bring a user back to the homepage, which is currently not the case. The statistical significance of content clicked on the homepage, and clicks to the logo in our player page, could mean that users are more comfortable browsing our homepage vs. player pages. Our first step in confronting this issue would be to either allow user to return to the homepage through our header logo, or simply organize our player's homepage similar to the regular homepage.

**Conclusion**

In utilizing logistic regression, recursive partitioning, and PCA to better understand the engagement of our alpha users, I'm able to give my product team direction in how to better engage our total user base. Based on the results of these statistical methods, I'd suggest that we better surface browsing capabilities in general. Specifically, I'd suggest the better surfacing of song detail pages, song metadata, and similar songs. I'd also suggest that content organization be consistent between home and player pages.