**Identifying Likely to Churn Users for a Television Everywhere Platform**

*Jennifer Prasol*

*3/6/2017*

(specifics of the data herein have been obfuscated to conceal the identity of the source)

**Executive Summary**

Television Everywhere (TVE) is a platform used across the television industry that allows content providers such as ABC, NBC, Turner Broadcasting & Viacom to offer access to digital content to qualified cable subscribers. Users can utilize their cable subscription to view programming directly from the content provider through a variety of platforms such as mobile apps, connected devices (AppleTV, Roku, FireTV etc.) and a standard web browser. Monetization occurs through direct programmer ad sales.

Like many online services and mobile apps, TVE offerings have a large number of users who try the experience, including successfully authenticating, but then do not return. For the TVE provider in question, 30% of the authenticated users only view content on a single day and never return, another 30% view only 2-4 days. The objective of this analysis is to predict which users are likely to become loyal users vs. which are likely to abandon (aka churn). This capability would allow the TVE provider to 1) predict usage growth 2) cut marketing costs by targeting only likely loyal users or 3) improve the user experience to develop additional loyal users.

Data access across the platform is complicated by the shared business model. The cable company owns the connection to the subscriber including their identity and demographics; the network programmer owns the content and the historical viewing data on the platforms. Therefore, this analysis will utilize viewing behavior on the TVE platform only.

In this analysis a sample of ~100K authenticated subscriber records and their viewing behavior was used to predict which users were likely to abandon after their initial visit and which would be loyal returning users. Most models predicted within 60-65% success range overall, but some accurately predict one classification at 80%. Key learnings actually came from the variables that were identified as predictive, such as whether the user participated in a free trial prior to authenticating.

**Analysis**

**Discovery Phase**

Data for the TVE platform is captured through an industry standard digital analytics application that is transferred to and stored in Amazon Web Services Redshift environment. The data include information about whether a user has authenticated, a unique account identifier, content viewed, platform viewing occurred on, and other viewing details such as time spent.

Previous analysis work for the TVE provider has identified that 30%+ of the subscribers who authenticate and stream a video on the platform do no return after their first use. Using six months of data, user loyalty (calculated as # days streamed) was determined and users were binned into low (1 day), medium (2-4 days) and high (5+) days were calculated.

Because about half of the initial variables available were categorical variables and because the outcome to be predicted could be stated clearly as a classification but may be inaccurate as a numerical output, a classification decision tree was selected as the target model to be used.

Since this is the first time this model has been attempted, the classification decision tree will also help identify which variable are useful in the model and therefore inform future attempts. Future iterations should also compare the decision tree output with a logistic regression in order to see the overlap between the two and assess significance of the various variables.

**Data Preparation**

Using the data stored in the AWS environment and custom SQL, the following variables were prepared in a data file, were each row was an authenticated user who had begun viewing a content episode online between August 2016-October 2016.

* Churned Bin – categorical. Val= 1, 2-4, 5+ based on # days viewed. Also refined as churned = 1-2 days and stayed = 3+ days.
* Platform – categorical. Val = web, app, cnd (connected devices).
* Device Type – categorical. Val = desktop, android, ios, tvos, roku.
* Network – categorical. Val= net1, net2, net3.
* Day of Week – categorical. The name of the week day on the first viewing day

Because some series show on specific days, this may be a consideration.

* Daypart – categorical. An hour range within the day.

Dayparting is standard in TV and some ad pricing is based on airing daypart.

* Number of Episodes Viewed (first day) – numerical. Count of unique episodes viewed on the first day.
* Number of Series Viewed (first day) – numerical. Count of unique series viewed on the first day.
* Number of Top 5 Series Viewed (first day) – numerical. Number of top 5 series on each network viewed on the first day.
* Total Time Spent Viewing (min) – numerical. Estimated total viewing minutes on the first day.
* Trial – categorical. Val = sd (same day), pd (previous day) no (no trial). Based on whether or not the user viewed an episode prior to signing in, and when the episode viewing occurred.

**Data Exploration & Sample**

Variable distributions were assessed in Tableau and for any numerical values outlier and cutoff points were identified, with the objective of identifying users which viewed an excessively large amount of content and excluding them from the analysis. See the appendix for output from the data exploration.

The following observations and decisions were made:

* + Although 6 months of data were available only the first three months would be used in order to allow at least three months for the user to return. Remaining n= 500K+
  + Users with 90+ days of viewing would be excluded in order to remove outlying accounts. This happens due to users sharing accounts, testing and other reasons. N excluded < 3K
  + Users with more than 10 episodes on their first day would be excluded. N excluded < 250
  + Users with more than 480 minutes (8 hrs) of viewing on their first day would be excluded. 8 hrs is a typical industry cut off point for continuous viewing. The concept being represented is whether the user is actually still ‘watching’ vs. the content just playing out on the device. N excluded < 5K
  + Variables with influence on the dependent variable appeared to be: platform, the consumption variable (number of episodes, time tuned, number of series, top 5 series) and trial participation.
  + Variables that appeared to have no influence on the dependent variable were: network, daypart, week part, day of week.
  + A random sample of 100K records was selected for modeling purposes. A random id was used to generate a row number and records were sorted.

**Model**

* The 100k record sample was split into a training and a test data set, with 60% of the records in the training set.

Classification decision trees were run using rpart in R.

* Several trees were run and compared including 2 bin churn, 3 bin churn and a tree for each platform.

**Results**

The 3 bin method did increase the accuracy of the classification for the 5+ bins and reduced accuracy for the 1 bin by 4% points in either direction. However, the tree itself was very similar to the 2 bin method (where churned was defined as 1-2 days streaming and stayed was 3+), therefore the 2 bin confusion matrix and tree are outlined below.

Using the two bin method, overall accuracy was 63% with true negatives for churned at 50% and true positives for stayed at 75%. The model was more effective at predicting likelihood to be a loyal user and less effective at identifying users who would be likely to abandon. This could be just as useful from a business perspective but would be deployed differently.

Variables that were used in the tree were: number of episodes viewed, platform, time spent viewing and trial participation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Predicted** |  |  |  |
| **Actual** | **churned** | **stayed** | **Sum** | **True Neg/Pos** |
| **churned** | 9,281 | 9,175 | 18,456 | 50% |
| **stayed** | 5,694 | 16,713 | 22,407 | 75% |
| **Sum** | 14,975 | 25,888 | 40,863 |  |

|  |  |  |
| --- | --- | --- |
|  | **Predicted** |  |
| **Actual** | **churned** | **stayed** |
| churned | **23%** | 22% |
| stayed | 14% | **41%** |

**Interpretation & Recommendations**

The model demonstrates that how much content the user engages with and views prior to or during their initial authenticated session influences the user’s future loyalty. This makes intuitive sense, since consuming more content is reflective of having a positive experience and therefore suggests a user is more likely to return.

To increase the effectiveness of the model, the following should be attempted:

* Balance the sample at 50/50 loyal vs. churn, since the tree is known to perform better.

Also repeat excluding the 2-4 day bin with a 50/50 sample.

* Since time to return is a factor, use the month of acquisition as a potential variable in the model.
* Additional experience variables should be identified, such as occurrences of an error or interacting with other features. This may also help to identify what features have a impact on consumer experience and provide direction for product development.

**Appendix:**

**N and Bins**

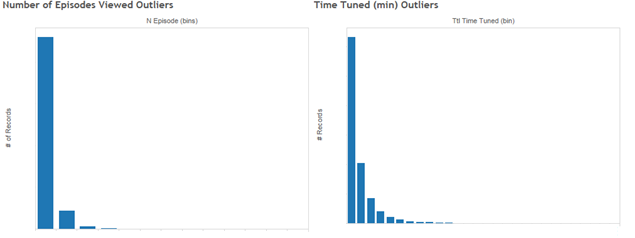
Shows the total number of records available. From the histogram, there is a long tail on the # of days a visitor returns, and 37% are only 1 day. The values past 90 days were discarded as potentially shared or test accounts. The amount of time available to the user to return has an influence on number of days viewed, as does the content available during a particular time period, therefore just the first 3 months were used.

3 bins were used initially to try to create more differentiation between the 1 and the 5+.



**Data cleansing**

Both the number of episodes viewed and the total spent viewing had a long right tail. Since these figures are supposed to represent the total number of episodes viewed in a day, and the amount of viewing time, the larger numbers may represent users who are not really viewing. This is a standard problem in TV viewing. Number if episodes was capped at 10 and Time spent viewing was capped at 8 hrs (480 min).

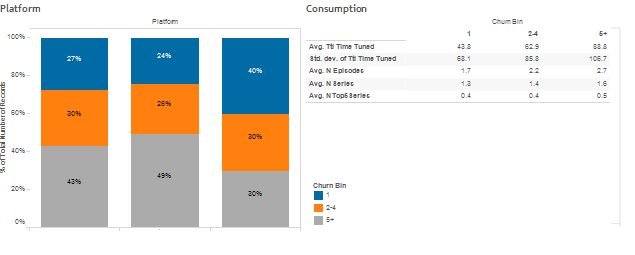


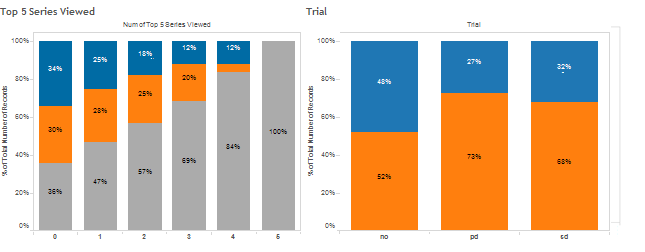
*Labels removed for confidentiality.*

**Dependent Variables with Potential Impact**

Platforms, the various consumption metrics (number of episodes, number of series, top 5 series and time spent viewing) and participation in a trial on or before the first authentication day appear to have an influence on the churned bin. Consumption variables tend to move together. No scatterplot was created because the volume of data was difficult for tableau to process. The top 5 series are the most popular on the network are drive a large volume of activity, but are sometimes seen as source of non-loyal viewing.

*Some labels removed for confidentiality.*

****

****

*Some labels removed for confidentiality. Churned bin for trial just ‘churned’ or ‘stayed’ as pictured.*

**Dependent Variables without impact in initial exploration**

Network and the variables regarding when the viewing occurred within the week or time of day initially appeared to have little influence on churn bin.

****

**Decision Tree**

