Capstone Project: Predicting the Likelihood of Starting a Subscription for Pandora Listeners

Background: Pandora Media has 20MM active listeners every day and of 80 million monthly active users on the ad-supported tier. Pandora also offers a subscription tier that allows a listener to access more product features such as offline mode and more skips and replays in an ad free environment. The subscription service grew rapidly at the takeoff point but given recent years of competition and increased appetite for an on-demand experience, subscriber growth has not been as lively as previous years. Due to this, Pandora would like to optimize its current marketing spend and product touch points in order to convert the appropriate ad supported listeners to the subscription state.

The objective of this project is to analyze the data of listeners who started a subscription with the service to find any patterns or features that may help identify other listeners likely to subscribe.

Current feature set:

1. Listener ID
2. Subscription\_Start\_Flag
3. Hours listened per listener in previous 7 days
4. Days active on the service in past 7 days
5. Age
6. Gender (is\_male\_flag)
7. Days since account creation
8. Zip code flag

Process:

1. Establish ground truth for subscription starts
2. Choose appropriate subset (% of total listeners) of daily data
3. Brainstorm and identify feature set that will help identify listeners likely to subscribe
   1. It helps to be creative here
   2. Be expansive with your list
   3. Ensure there are no product features that may completely throw off your assessment
   4. Is there enough sample of that feature to make an impact? Regressions are heavily impacted by sample sizes (e.g. business page views). Random Forests are better at correcting for this.
4. Create data tables to feed the master data table export for R
5. Data cleaning, wrangling, and munging
6. Choosing ML model for data set
   1. Clustering?
   2. Regression? (Linear, Logit, Generalized)
   3. Decision Tree -> Random Forest
7. Create train and test data sets
8. Try the model(s) out
9. Score the model
   1. RF.score, logistic.score
   2. TPR/FPR – precision/recall
   3. Score it against each other (AUC curve)
10. Tweak it as needed – there may be many iterations to improve the model
11. Look at who is being identified
    1. Does it make sense?

**Complications:**

There are currently 4.5 million subscribers out of 80 million monthly active users meaning the incidence of subscribers are fairly low relative to the overall population. Additionally, subscription starts are a much smaller proportion of daily active users meaning an even lower incidence rate. Some potential options for this is to create a ratio within the training data set to have more subscription starts than reality but that may lead to elevated false positives. Another option is to use the accurate sample size and incidence of true positives and train the model to gain any learnings we can.

The baseline model for comparison will perform incredibly well if it just assumes that every listener continues in their current state (i.e. not start a subscription) and have > 90% accuracy. Given this, the metric to compare the model to the baseline may be simply matching the assumption of everyone staying the same and any gains above this will be a great breakthrough.

**Limitations:**

This project will not allow one to say with certainty whether or not a listener will subscribe if they display certain feature usage or sees x impressions but the probability of subscribing can be used to say identify listeners that should get more or less subscription ads allowing for the sale of these ads to other companies.

There is no unsupervised learning right now to identify clusters or for segmentation purposes. After initial analysis, it looks like the majority of the data set has listeners who are not active within the past week and excluding them will allow the model to run much faster (from ~4MM rows to 500k rows).

Does not understand all of the current business rules such as the need for trials prior to subscription or availability of subscription on different devices or platforms.

**Further steps for consideration:**

Future feature set:

1. Hours listened per listener in previous 7 days
2. Days active on the service in past 7 days
3. Age
4. Gender
5. Interactions with the service (features used)
6. Impressions across all marketing and product touchpoints
7. Integration of other data sources (FB profile, Google, etc).
8. Subscription tenure

Optimizing data table creation and export for faster run time.

Combination of supervised and unsupervised learning to increase model’s performance:

* Perform unsupervised learning to identify clusters which may show higher concentrations of subscribers. This could also help reduce the noise of the model by excluding clusters that are mostly inactive or extremely low incidence of subscription starts.
* Leverage the cluster identify as a feature within a supervised model like regressions.

Currently, the data is using the listener’s previous state on day 0 with the aggregated and grouped features and then the listener’s current state on day 1. The future state will have this happening on a daily cadence and simulate movements across multiple listener states.

Creation of a multi-class identification model allowing listeners to move from more than just ad supported to subscriber but also, for example, subscriber back to ad supported (churn event) or active listener to inactive.