Capstone Project: Predicting the Likelihood of Subscribing

Pandora Media has 20 million daily active users and 80 million monthly active users on the service. 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. Features that may help guide Pandora’s advertising strategy may be the listeners’ demographics (gender, age, location), behavior on the platform, engagement with the product, and previous purchase history. By collecting the historical data set for listeners who start a subscription along with those who do not one can create a likelihood of subscribing model to help future advertising efforts to reduce market spend.

Without additional computational power, the available data set was scaled down to a sample and within the sample there was additional filtering to exclude listeners who have not been active within the past 7 days on the service. There could be another model to predict the likelihood of these listeners to return to the service in the subsequent days but that is outside of the scope of this report. This seems to not be very detrimental to the occurrence of subscription starts, and potentially the right move, because of the 3 million observations filtered out there were only 7 subscription starts within it and likely from listeners new to the service. That said, even with the sampling and filtering, the final data set has roughly 550,000 observations across 14 variables of which 7 variables are used along with the dependent variable, whether or not the listener starts a subscription.

Unfortunately, the incidence in which a subscription start occurs out of the 550,000 observations is very low at 33 out of the sample. This means the baseline model which will assume that every listener does not subscribe would be accurate 99% of the time. Given this low incidence, the model that can match the performance of the baseline should be considered for further evaluation and optimization.

The logistic regression algorithm was chosen to move forward with over say a linear regression model because the dependent variable is categorical rather than continuous as in the listener either subscribes or does not. The model was trained on 10% of the data set or roughly 55,000 listeners’ data across total play time on the service in the past week, days on the service in the past week, their age, gender (M/non-M), days since they created their account, whether or not they provided a zip code, and how many coachmarks (ads) they saw in the past week.

While there were no features that were statistically significant or meaningful in predicting the likelihood to subscribe, the model’s true positive rate was in line with the baseline model in that it predicted everyone would stay in the same state (i.e. not subscribe).

Given the current feature set and model’s performance, there can be ways to improve the model in the next version. Some ideas currently being considered are:

1. Expanding the features:

* Hours listened per listener in previous 7 days
* Days active on the service in past 7 days
* Age
* Gender
* Interactions with the service (features used)
* Impressions across all marketing and product touchpoints
* Integration of other data sources (FB profile, Google, etc).
* Subscription tenure
* Previous purchase history

1. Combining clustering techniques with the regression model by identifying clusters of higher concentration of subscribers and removing clusters where there are no subscribers. From there, one can use the listener’s cluster as a feature for the regression model.
2. Expand the date range used for the listener to start a subscription because while the number of subscription starts per day is low relative to listeners, within the week or month the number of unique listeners increases at a slower pace than the number of subscription starts.

Overall, the findings of this study show that the current marketing efforts can already be improved by focusing more on the listeners who are active in the past week. The process and structure for updating the model also does not have to be re-created so the next version can be tested and validated even faster.