**Digital Marketing Analytics Course**

**MIT Sloan School of Management**

Creating a holistic digital Marketing Optimization Plan.

Pablo Aguirre Solana M6 U3 Ongoing project

**I. Executive Summary**

We found evidence that supports the fact that freemium services need to rely on a business model that both increases the likelihood of paying for an upgrade and that can rely on a client base that are socially and actively engaged.

The results of our analysis validate the hypothesis that: user participation is positively associated with the likelihood of subscribing to premium services and that the likelihood of subscribing depends more on homophily rather on influence.

Thus, our recommendation for a Marketing Plan that deals with a freemium service would be the following:

* Segment the client base on two dimensions; activity (engagement) and customer value. (High-engagement, high-value as our target group)
* The client segmentation can help as directing specific efforts to commit in viral marketing campaigns which incentive engaged and highly vale users in spreading awareness through their own content.
* The campaign must be dependent on the content of highly value usures as a way of building homophily communities and as a way of social adverting’s. Expecting that can help build a halo or cascade effect to the less engaged, less value users.

**II. Problem Statement**

High Note is a company that offers a freemium service of Online music to its subscribers, it is a business model in which anyone could sign up for a user account, build a profile and listen to songs for free. This service offers two kinds of subscription, one that is free, that comes with an add every 15 seconds, and the other premium, which for $3 a month fee, you can add extra features and no adds.

This business models had their risks; in the way that it depended on advertising money, and to get advertising money more traffic was needed, until the company hit a plateau at 2011 with new subscribers, that do not necessarily were opting for the premium subscription.

Thus, the main issues that High Note had to solve was:

How can they convert more free users to paid premium subscriptions?

How can they attract more premium subscribers?

How can they keep the users we have engaged so they don’t wander off?

In order to answer these questions, we must first analyze what are the driving factors that make a person become subscriber vs not becoming one, and what is the causal link between them. Once we can know what variables increase the likelihood of become a subscriber vs not becoming one, then we can start to build and suggest strategic avenues to increase the number of paying subscribers and to keep them engaged.

For this present document hence; our first null hypothesis can be the one that Gal Oestereicher-Singer and Lior Zalmanson posed in their paper (Content or Community?) Which states that: *“user participation in websites is positively associated with the likelihood of subscribing to premium services”*

By participation we will refer to those social variables that are available in the data set provided that such as: songs listened, loved tracks, posts, playlist and shouts.

A second null hypothesis that is motivated from the paper of Sinan Aral, Lev Muchnik and Arun Sundararajan (Distinguished influenced-based contagion from homophily-driven diffusion in dynamic networks) is to proof somehow the difference between peer influence process from alternative processes such as homophily. That is; influence driven is self-reinforcing whereas homophily is more determined by its intrinsic characteristics. Thus, our second null hypothesis would state: *“the likelihood of subscribing depends more on homophily (which is understood as similar characteristics) rather than on influence (understood as other factors)”*

Thus; the validation of these hypothesis will help us create a marketing strategy that could be fully supported on the empirical evidence at hand.

**III. Analysis**

From the exploratory data analysis (EDA) of the available data, we can draw the following observations:

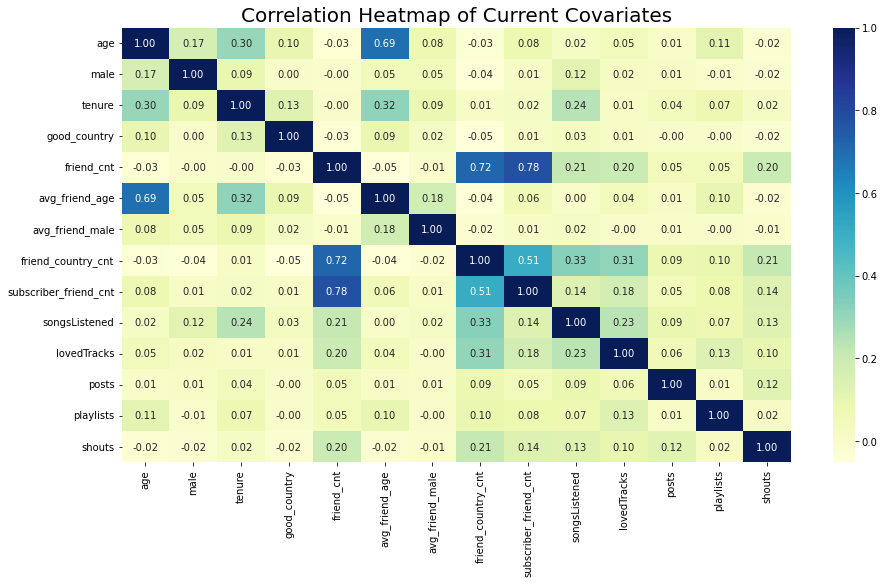
A data set from the company was given to perform the analysis which had 107213 records in three different time stages. Missing values were removed to clean the data so that this could not be susceptible to errors in the models. Hence from having 107213 cases we ended with 43827 cases. The data base is divided in three periods current, pre and post, for this document and analysis we are only using the current period, for simplicity and time’s sake. So, all the conclusions and analysis will refer to the current period of the data set.

The first thing to notice is that; the demographics of the service are: 63% males and 37% women. Also, the average age is 24 years with a standard deviation of 6.43, which means that in the 75% percentile the age is 27. The country of the service is higher in other countries than in the US, UK and Germany 65%, and 35% in the latter.

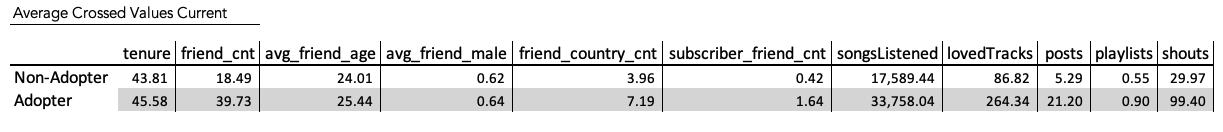
The second thing to notice is that only 8% of the total sample are adopters; demographically they are 73% males and 27% women. With an average age of 26 years and a standard deviation of 6.84, which means that in the 75% percentile the age is 29. Also, 71% of the adopters are from other countries than US, UK and Germany and 29% from the latter.

From a simple correlation analysis, there appears to be some mild positive correlation (.30) between tenure and age. Also, a covariate that is highly correlated (.69) with age, is the average friend age, which seems pretty understandable. Regarding gender our correlation analysis does not yields any important correlation between gender and other variables, which does not mean necessarily that there is no relationship between gender and others, just that it is not easily discernible from the immediate patterns of the data.

In terms of behavior covariates, we found that friend count has a strong correlation with friend country count (.72) and with subscriber friend count (.78). From these results, we see broadly that having a friend that is a subscriber can influence somehow.



From a simple and very basic bivariate analysis crossing the average values of the Behavioral covariates and the adoption variable we found that; tenure, friend count, friend age, friend gender, friend country, subscriber friend count, songs listened, loved tracks, posts, playlists and shouts, are higher for adopters than for no adopters. As the table below shows:



The first implication for this is; that adopters are overall more active with the service; thus, it is crucial for us to find what are the variables that help explain this activeness with further analysis and models.

To test this implication that was found in the EDA (exploratory data analysis): We elaborated a logistic regression model that was done to validate the first hypothesis:

*“User participation in websites is positively associated with the likelihood of subscribing to premium services”*

And therefore, to assess what are the variables that drive the activity or engagement, that is what are the covariates that increase the likelihood of becoming a subscriber vs not.

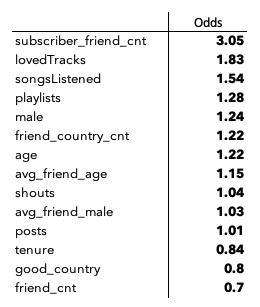
It is important to state that in this case activity and engagement are measured in the available data set as the variable that imply a social interaction with others; such as: loved tracks, posts, playlist, and shouts.

For our model the following steps were taken:

For the pre-processing the data was scaled due to the fact that each covariate is measured in different units and we need to be able to generalize across covariates. The data was split into training and testing sample with size 30% for the test and 70% for the training. Also, a re-sample technique different from what the authors Gal Oestereicher-Singer and Lior Zamalson used in their paper called oversampling was used, to compensate for balancing the classes in the data set, which can be crucial for model interpretation and model outcomes.

The results of the model in the test sample have an overall accuracy of .69. For class 1 which is the one that we are interested in predict (adopters); we have a F1 score of .66 and a recall of .61 and a precision of .72. Which broadly states that we are predicting incorrectly around 30% of cases. That is users who we predict will be adopters, which in reality will be.

What is really interesting is to see the odd ratios, which are essentially the coefficients of the logistic regression exponentiated, they can tell us that a user that has a friend who is subscriber is 3.05 times more likely to become an adopter. Also, we can see that loving tracks, the amounts of songs listened, the creation of playlist, gender and friends’ country can increase the likelihood of becoming a subscriber, as results of table below show, that is social variables or covariates that represent the engagement or activity.



Thus, there appears to be some evidence supporting our HO, still, to proof that there is causality between having friends and being an adopter, some other statistical models need to be run.

For this we ran a Propensity Matching Score model, which tries to account for the causal relationship between having friends and being a subscriber. That is, to proof that being a subscriber or not being a subscriber is independent of being in the treatment group conditional to all their covariates (tenure, age, gender, songs listened, loved tracks, posts, playlists and shouts) The treatment group being; having or no having friends that are subscribers. Matching is created to create an artificial control group and then to estimate the impact of treatment.

The model assumes; unconfoundedness. Selection on treatment (or not) should be solely based on observable characteristics. Assuming there is no selection bias from unobserved characteristics. It is not possible to prove the validity of this unconfoundedness assumption.

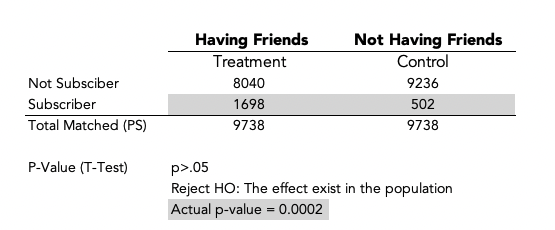
The steps we followed for this model were:

* Estimate the propensity score. This is the propability (logistic regression) that an observation is treated or not.
* Perform matching. For each treated sample, identify an untreated sample with similar logit propensity score. The matching is 1-to-1 with replacement, using Nearest Neighbors technique.
* Once matching is performed, we review the balance of the X variables to assess their balance.
* Estimate the impact of treatment with a T-test of proportions.

The results from the matching variables balanced according two groups is presented in the below table. And we can see that from the T-test between samples (treatment and control) there is a p-value less than .05, that is; those samples are statistically significant independent from one another thus, confirming that there is a casual link between having friends and being a subscriber vs not being a subscriber, also confirming the first hypothesis of Gal Oestereicher-Singer and Lior Zamalson which stated that:

*“User participation in websites is positively associated with the likelihood of subscribing to premium services”*

In our case participation should be understood as a consequence of having friends that are subscriber vs not.

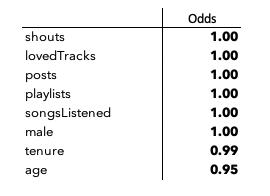


Finally, to test our second hypothesis which states that:

*“The likelihood of subscribing depends more on homophily (which is understood as similar characteristics) rather than on influence (understood as other factors)”*

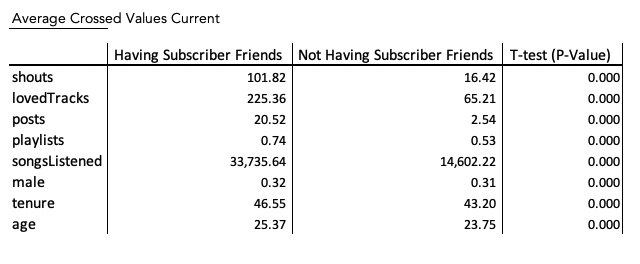
To prove or reject the before mentioned hypothesis, we derived the coefficients and odd ratios of the logistic regression used in the Propensity Matching Score, that had as dependent variable (having friends that are subscriber’s vs not having friends that are subscribers) and the same covariates used in for the Propensity Matching Score: (tenure, age, gender, songs listened, loved tracks, posts, playlists and shouts).

From the results of this model as the below table shows, we can see that the likelihood of having friends that are subscriber’s vs not having friends that are subscribers is increased equally (1.0) by the social variables or the variables that imply activity or engagement (shouts, loved tracks, post, song listened) and gender.



Thus, what we can conclude from this, is that: being engaged increases the likelihood in 1.0 times, of having a friend that is a subscriber vs not having a friend that is a subscriber.

Now to prove if there exists homophily what we did is a simple comparison of groups using a T-test of group means independence, this independence can serve as a proxy of homophily. Hence, the results of the below table can show and prove that people in the group that have subscriber friends are statistically significant dissimilar of those who don’t in all covariates, which can be taken as some evidence validating the hypothesis of homophily as increasing the likelihood of subscribing.



**References:**

Do Your Online Friends Make You Pay? A Randomized Field Experiment

in an Online Music Social Network (Submitted to management science)

Content or Community?

Gal Oestreicher-Singer and Lior Zalmanson. MIS Quarterly Vol. 37 No. 2, pp. 591-616/June 2013

Completing the Virtuous Cycle between Payment and Social Engagement in Freemium Social Communities. Ravi Bapna, Jui Ramaprasad, Akhmed Umyarov, September 2013

Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. Sinan Arala,Lev Muchnika, and Arun Sundararajana. PNAS December 22, 2009