What drives users to become active on DubMeNow?

Group project for Data Mining for Business (BUDT733)
Group 8 | Cheng-Tsung Wang, Ji Hee Han, Nicholas Rapagnani, Shahryar Rizvi
Fall, 2009

Executive Summary

The success of a company like DubMeNow, that deals in digital markets, hinges entirely on the value of their network. Not just in the size of it, but also the quality. So for a customer looking to join DubMeNow, what matters is not just the amount of users in the network, but also how active they are within it. In an industry like DubMeNow 's, marginal costs approach zero so adding too many users to their network is not a problem. The goal is to have a larger network than your competitors.

DubMeNow is finding success. This app was first released in October 2008 and the company currently has over 100,000 users. The average growth rate is 133.23% per month. DubMeNow has also been named a featured app on BlackBerry App World this month. Once their user base is broadened some more, DubMeNow plans to commercialize their product to generate revenue.

We were given a large amount of good user data from the company itself. Our main goal was understanding the active users. We were not trying to predict which of the users would become active; instead, we wanted to find out what attributes active users shared and how the information could be used by DubMeNow.

DubMeNow gave us a lot of columns to look at in their dataset and it was difficult figuring out which were necessary and which would just going to slow us down. The data we were given gave us both information on the usage patterns of the customers. We explored many graphs and employed several operations on the data using XLMiner.

While we found that many of the models reinforced the findings of other we also saw that not all methods worked for our project. Classification trees in particular gave us a lot of trouble and even though we spent time tweaking it, did not tell us much about the users. However we had more success using logistical regression to tell us which attributes DubMeNow should pay attention to.

Ultimately, the most important variables found were AcceptYourInvitation and AcceptInvitation. Social Networks were important too. We have sample suggestions to be considered utilizing a rewards program and Facebook to give an idea of the creative solutions that should be sought out.

Technical Summary

Dub's user database contains user data and device information. From here, we extracted ten variables that we felt would be helpful in explaining what separates the active from the inactive users:

User Data			
Description	Variables	Units	Type
A unique user ID generated by system	User ID	Number	Numerical
A flag that shows if the user is defined as an active user or not (**Active user is defined as users who login more than ONCE a week on average**)	Active User	String	Categorical
The number of times when users update their cards	Update Card	Number	Numerical
The number of social networks that user associated with their account	Social Network	Number	Numerical
The number of contacts that user has in DUB network	Contact	Number	Numerical
The number of IM account that user registered in system	IM	Number	Numerical
The number of invitation sent by each user	Invitation Sent	Number	Numerical
The number of invitation received by each user	Invitation Received	Number	Numerical
The number of accepted invitations among all the invitations sent by each user	Accept Your Invitation	Number	Numerical
The number of accepted invitations among all the invitations received by each user	Accept Invitation	Number	Numerical

Table 1 Variables used from DubMeNow database

At the time of the database, DUB had 50,000 users.

Graph & Pivot Table Exploration

We started in Spotfire and then Excel, creating scatter plots and box plots between the different variables. The variables that stood out from our graph exploration were UpdateCardCount, SocialNetwork, and Total Weeks after Registration.

Figure 1 shows that the users who were a part of more social networks were more active than those users with less social networks. **Figures 2 and 3** show two box plots comparing active and inactive users on UpdateCardCount and Total Weeks After Registration (For the UpdateCardCount box plot, we took the Log of it in the y-axis to see it more obviously). From the box plot, there is a difference between the two classes and these two variables could be effective variables to distinguish active from inactive users.

The next graphs examined were the scatter plots involving UpdateCardCount, Total Weeks After Registration, and Social Networks. **Figure 4** shows the amount of weeks a user has been registered and how often they update their card. It shows that the more active users are the ones who update frequently in the beginning weeks after registration (when they are "new" or "young"), while "older" users are inactive and do not update as much. Lastly, **Figure 5** shows that those who registered earlier tend to be less active while users who registered for a short period of time tend to be more active. So it seems users lose their interests after a certain period of time.

Pivot Table

After exploring the graphs, next was the using Excel to generate the pivot table to see what predictors can work better to successfully separate the different types of users.

Data	Active User		Grand Total
	N	Y	
Average of AcceptInvitation	0.10745108	0.17678588	0.12225489
Average of AcceptYourInvitation	0.06751129	0.29562887	0.11621712
Average of UpdateCardCount	0.55245861	2.18233065	0.90045579
Average of Contact	0.05888108	0.28028833	0.10615418
Average of SocialNetwork	0.08356749	0.41382497	0.15408141
Average of IM	0.11954340	0.39377137	0.17809435
Average of Total Weeks After Registration	10.16650778	5.40162647	9.14914860
Average of InvitationReceived	0.14530858	0.21504482	0.16019810
Average of InvitationSent	1.20735073	4.77229461	1.96850891

Table 2 Pivot Table

The pivot table shows all variables selected could separate active user from non-active user to some extents because there are some differences between the average of two classes. To verify if these variables are statistically significant or not, several models are used to explain.

Model Exploration

Classification Tree

The first model is the classification tree because its result could be interpreted and explained easily. Also, the tree is highly automated and could be used as an exploratory tool. The tree requires a large number of records, but our large data set has 50,000 records in it so that is not a problem.

Since the goal is to explain what drives users to become active, no partitioning was done on the data and the full tree was run. However, the results were disappointing. The tree classified all observations as non-active with an overall error rate of 21.20%. This could be because the majority of records are non-active.

Concluding that the tree did not seem to work well to explain, another model was used next.

Discriminant Analysis

Discriminant Analysis is helpful in finding out differentiation by calculating the difference of points in each variable. All variables were used in the Discriminant Analysis to get a general idea of which variables are more important.

Of the variables included in this analysis, AcceptInvitation scored the highest difference (See **Figure 6**). We interpret this to mean that active users are more likely to accept the invitation. However, it is somewhat considered common sense, because the activity is directly related to the usage of DubMeNow.

Among two variables that are not directly related to the activity, IM and SocialNetwork, Social network scored higher in difference.

Although this provides some information that is useful for explaining the characteristic of users, it does have a high percentage error of 42% in classifying the active user (low selectivity). So either the cutoff should be increased, or another classification method should be used.

Logistic Regression Model

Logistic regression is a good explanatory tool and could provide the odds of the variables in addition. Here the data was partitioned to meet the maximum records of 10,000 allowed in XLMiner in order to run the logistic regression model. A default cutoff value 0.5 for the success (active user) class was used.

In the first model that was run (See **Figure 7**), all nine of the variables were put in. Then, the one with the highest p-value was removed and the model was run again. This was repeated again, removing the next variable with the highest p-value. Then, the third model was run with only seven variables that were all statistically significant. The multiple R2 is 25.42%, which is the model deviance to the deviance of the naïve model.

According to the model, "AcceptYourInvitation" is the single predictor that is most useful for separating active users from non-active users. The next most useful predictor is "AcceptInvitation". This result reconfirmed our thoughts from the Discriminant Analysis that "AcceptInvitation" is the most useful predictor with "AcceptYourInvitation" as the second most useful.

Cluster Analysis

Finally, Cluster Analysis was run to understand the characteristics of Active Users (See **Figure 8**). Segmenting the Active Users by their behavior would help DubMeNow with future marketing. Since Hierarchical Clustering limits the number of observation to 4000, K-means clustering was chosen to be run.

The same variables that were used in the Classification Tree were chosen. Being an iterating process, the analysis was run with two means, three means, four means and so on. Also, the K that has best explanatory power and smallest average distance – three – was selected.

Three clusters we found in this analysis had distinct characteristics. Cluster 1 seemed to have average points in each variable, Cluster 2 had high scores in IM and Social network, and Cluster 3 showed extremely high numbers in variables related to 'DubMeNow activity' but scored less in IM and Social Network.

This observation is outstanding because it shows contradictory results. We learned that all the variables are positively correlated to the logit of the active user. In this analysis, we found that there are segments with different character; Cluster 2-Online Social Networker and Cluster 3-Just Dubber.

Conclusions

Like any community the data shows that the more accepting people are of each other the more likely they are to be active. The more invitations that a user sends out that are accepted, the more active that user is within the community. Similarly the more invitations that are sent to a user the more active the user becomes.

One of the problems DubMeNow has is with the users whose activity peaks early and then slowly declines over time. We believe that they are active at first because they are spending time trying to build a network of friends. However once they have added everyone that they wanted, they are no longer motivated to log in. It is important that DubMeNow finds a way to motivate the "older" users into using their service.

An example of an idea for improving this could come from implementing a rewards program. This would get older users to come back. Perhaps users could collect points when they successfully invite a friend to be a DUB user.

We have also found that the users who are active within the DubMeNow community are also active within other communities. There was a strong correlation between the number of social networking sites that a user is part and how active they are within the DubMeNow community. Similarly the more instant messenger accounts the user has the more active they are using DubMeNow.

This knowledge may seem obvious but it also gives us a strategy for DubMeNow. Since the active users are spending a lot of time at social networking sites, we believe that they should use these sites to help attract more active users to their network. By leveraging the existing social network contacts of their active users, DubMeNow should be able to add more valuable users.

One suggestion based on this that we propose is for DubMeNow to create a Facebook app that would give their users access to their DubMeNow profile within Facebook. Users log into Facebook much more often then they log into DubMeNow. By allowing their users to add/edit/update information from within Facebook they would be more likely to be active. The Facebook app would also make it easy for users to send out invitations to their Facebook friends and since these friends are social network users, we believe that they would be more active than the average new user.

Other similar ideas should be considered.

Exhibits

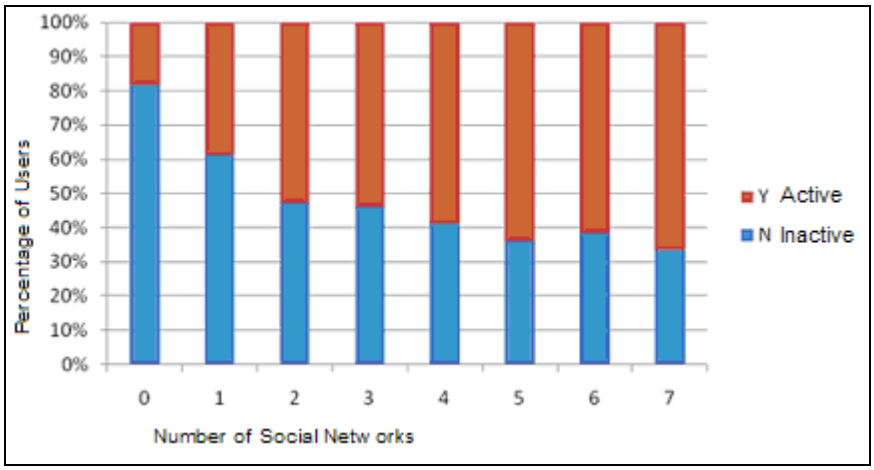

Figure 1 Percentage of active and inactive users per number of social networks

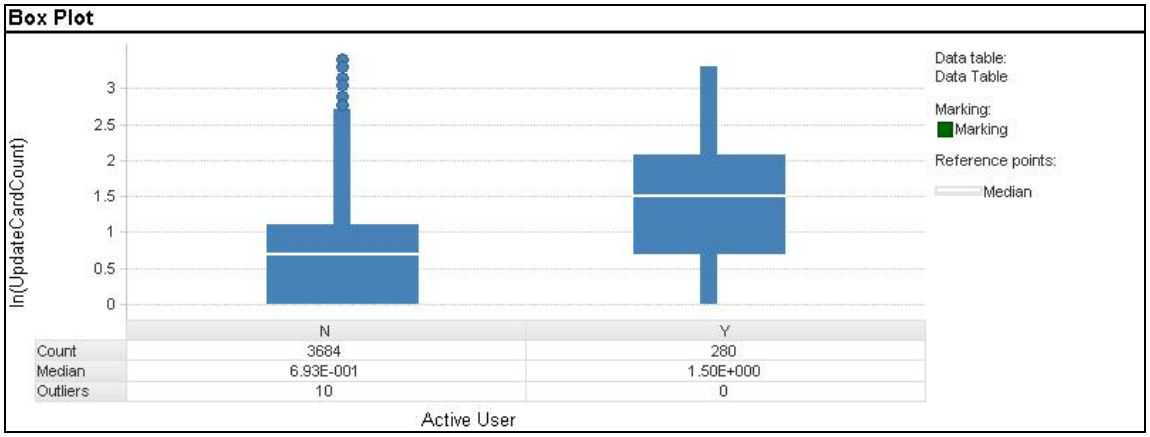

Figure 2 Box Plot comparing active and inactive users and (ln of) UpdateCardCount

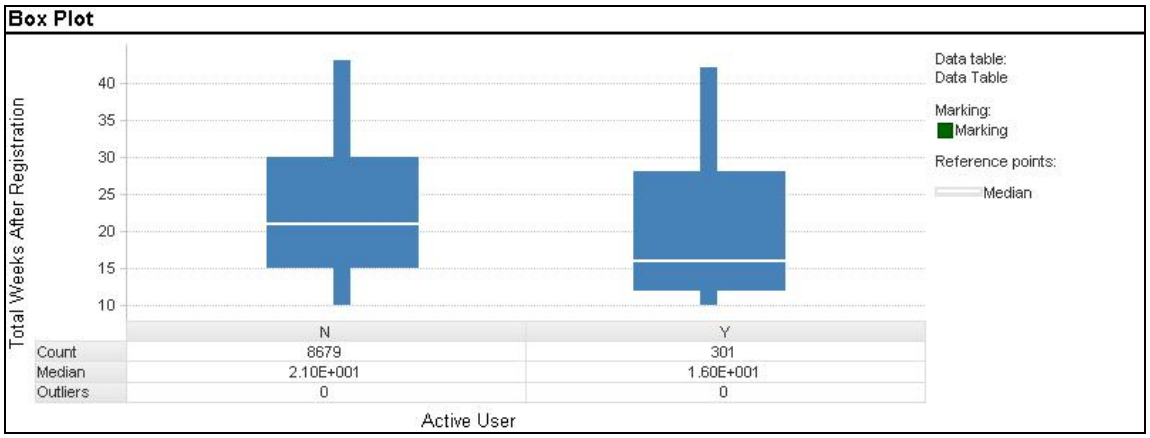

Figure 3 Box Plot comparing active and inactive users on Total Weeks After Registration

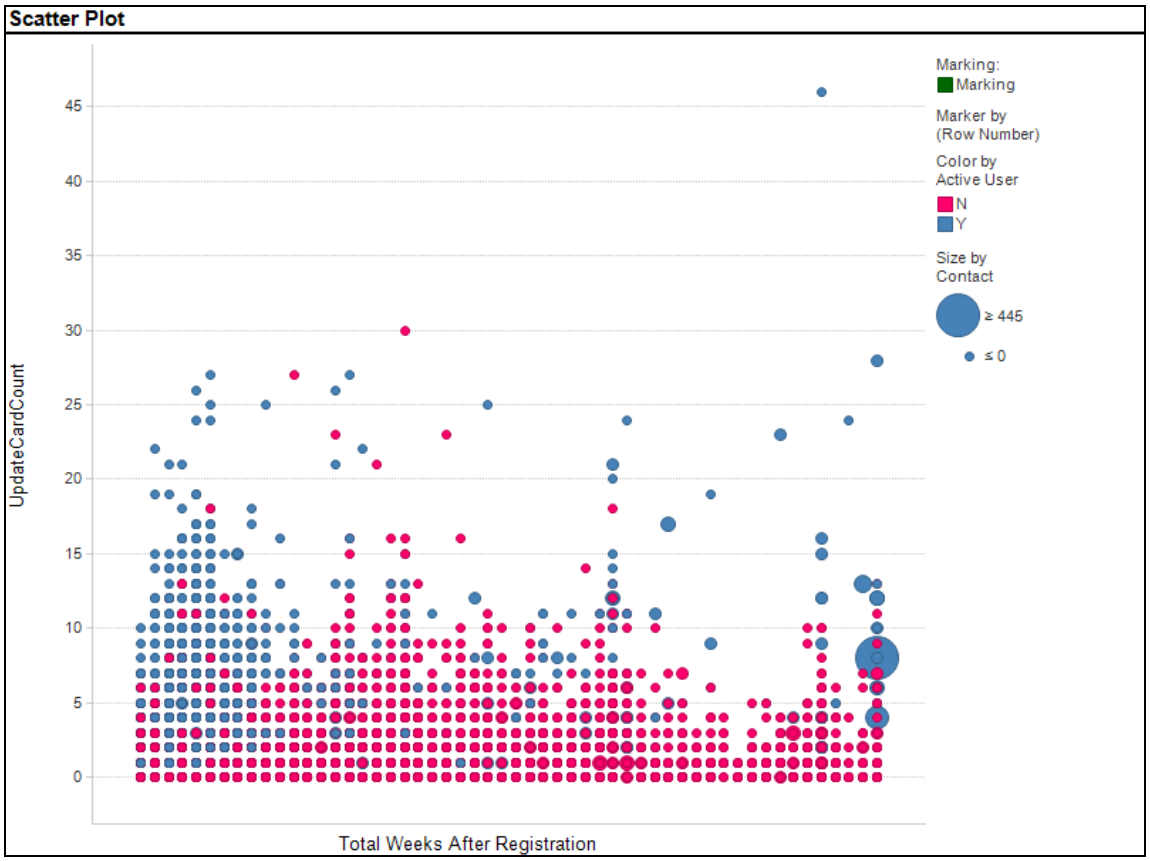

Figure 4 Relationship between UpdateCardCount and Total Weeks After Registration for active and inactive users

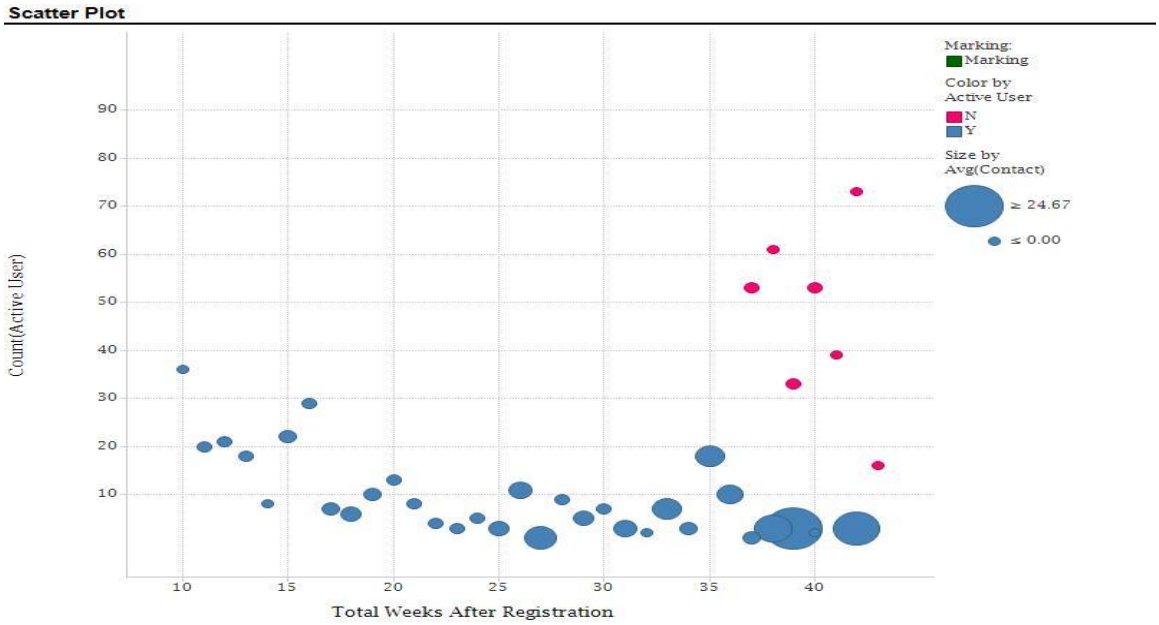

Figure 5 Relationship between Active User count and Total Weeks After Registration for Active and Inactive Users, also considering Contact Size

Prior class probabilities

Equal prior probabilities

Class	Actual Prob.	Misclass. Costs	Altered Prob.	
Y	0.5	1	0.5	<-- Success Class
N	0.5	1	0.5	

Classification Function

Variables	Classification Function		Difference
	Y	N	
Constant	-1.80323923	-1.29257894	
Total Weeks After Registration	0.03906569	0.11315618	0.07409049
UpdateCardCount	0.66981965	0.08803784	0.58178181
IM	0.23786096	0.02571089	0.21215007
SocialNetwork	0.59832996	0.2421916	0.35613836
Contact	-0.64625776	-0.14085963	0.50539813
InvitationSent	0.00038447	0.00021238	0.00017209
InvitationReceived	0.22461542	0.20800151	0.01661391
AcceptYourInvitation	0.62678605	0.03579542	0.59099063
AcceptInvitation	0.3698566	-0.25687221	0.62672881

Training Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	Y	N
Y	1229	891
N	1008	6870

Error Report			
Class	# Cases	# Errors	% Error
Y	2120	891	42.03
N	7878	1008	12.80
Overall	9998	1899	18.99

Figure 6 Discriminant Analysis

Prior class probabilities

According to relative occurrences in training data

Class	Prob.	
Y	0.212	<-- Success Class
N	0.788	

The Regression Model

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-0.6479165	0.0644445	0	*
Total Weeks After Registration	-0.24885428	0.01171437	0	0.7796936
UpdateCardCount	0.65867007	0.02317566	0	1.93222094
IM	0.11965668	0.05472173	0.02876888	1.12710977
SocialNetwork	0.17492075	0.051659	0.00070903	1.19115186
Contact	-0.28207234	0.08846086	0.0014293	0.75421911
AcceptYourInvitation	0.88276178	0.08866508	0	2.41756725
AcceptInvitation	0.71505439	0.0926279	0	2.04429793

Residual df	9992
Residual Dev.	7706.01416
% Success in training data	21.2
# Iterations used	9
Multiple R-squared	0.25415251

Training Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	Y	N
Y	755	1365
N	284	7596

Error Report			
Class	# Cases	# Errors	% Error
Y	2120	1365	64.39
N	7880	284	3.60
Overall	10000	1649	16.49

Figure 7 Logistic Regression

Cluster centers

Cluster	Total Weeks After Registration	UpdateCard Count	IM	SocialNetwork	Contact	AcceptYour Invitation	AcceptInvitation
Cluster-1	5.003371	1.714084	0.14758	0.135074	0.09168	0.159326	0.114845
Cluster-2	5.485564	4.624017	1.852364	2.091862	0.194226	0.290682	0.138451
Cluster-3	40.04901	7.911773	0.794118	0.470588	18.56862	12.656854	6.333332

Distance between cluster centers	Cluster-1	Cluster-2	Cluster-3
Cluster-1	0	3.93241054	42.46638442
Cluster-2	3.93241054	0	41.69060622
Cluster-3	42.46638442	41.69060622	0

Data summary

Cluster	#Obs	Average distance in cluster
Cluster-1	9164	0.915
Cluster-2	1555	2.345
Cluster-3	102	8.307
Overall	10821	1.19

Figure 8 Cluster Analysis