Predicting Conversion of Free Trial Users to Paying Customers to Increase Sales by Developing an Effective Free Trial Program

Team 3
107077424 | Bui Thuc Han
107077425 | Chawisa Mahajindapla
107077433 | Jariya Tienmongkol
107077443 | Ryjill Roa

Executive Summary

Migo is a digital company that utilizes satellite technology to serve digital contents to consumers who live in urban areas that have limited data connection due to high internet prices and find it too expensive to subscribe to the existing streaming services in the market.

As a recently launched services application, we are looking to find a way to increase sales by recruiting new customers. But engaging new customers is very challenging, especially for a new brand where people have no previous brand awareness. Free trial promotions are a good way to engage new customers but its effectivity should be evaluated to ensure that return on investments and efforts are maximized.

For starters, we are looking for an increase in the conversion rate of free trial customers to a paying customer within one month from their first subscription to a free trial by 10%. To support the business goal, we will focus on predicting whether free trial customers will purchase a subscription for their next transaction within one month after their first free trial ends.

We used four tables as inputs to the prediction: (1) transactions, (2) download, (3) engagement and (4) title. The tables were cleaned, filtered then merged to reflect one record (user) per row while the columns from the four tables were used as columns in the merged table. A new column was created called ‘Customer Spending’ with a ‘Yes’ when the user converts to a paid user for their next transaction within one month from their first free trial while ‘No’ if otherwise. Three predictive algorithms were applied to the merged data - Naive Bayes, Random Forest and Lasso Regression - and the results were compared using a lift and gains chart.

Comparing the accuracy and lift and gains charts from the three algorithms, we found that the Random Forest showed the best performance with a top decile of 4.089 which implies that the Random Forest model can perform approximately 4 times better than random selection in the highest 10%. Based on the variables that had the most predictive power, the unique and absolute engagement of the users, followed by the two most effective free-trial promotions such as the 2-Day Pass and PisoMigo_trial, and download space had the most significant impact on the conversion of users to a paying one. Based on this, the Marketing team can consider providing short-periods of free trial promotions and creating a rewards program to encourage users to be more engaged with the application which will therefore, increase the conversion rate of free-trial users to paying customers.

Problem Description

Business Goal

Migo is a direct-to-customer company that designed a platform to serve digital products and services to emerging markets. Customers can use Migo's mobile application to download content at Migo hotspots which uses satellite technology to enable faster download speeds that are otherwise not available in Migo's target countries. Migo has recently launched its services in the Philippines and is currently looking at making them available in Jakarta. They are targeting customers who live in urban areas that have limited data connection due to high internet prices and find it too expensive to subscribe to Netflix or other streaming services. Also, Filipinos like to watch movies and series and where better to do that than during their long commute to and from word due to traffic congestion.

As a recently launched brand in the Philippines, the Migo brand and its products and services suffer from low brand awareness which makes it challenging to increase the market share by recruiting new customers. Free-trial promotions are a good way to engage new customers to the brand. However, its effectivity should be evaluated to ensure that it is maximized and that it results in an increase in paying customers. Given this situation, the first priority for the company should be to increase sales by offering an effective free-trial to customers whose effectivity would be measured by a 10% increase in the conversion rate of free-trial customers to a paying customer for their next transaction within one month from subscription.

As an ethical company, we know that Migo always has the best interest of their customers in mind which is why we identified ethical issues such as concerns regarding the collection of user data that some customers may not feel comfortable sharing like watching habits, genre preferences, duration and time of watching activity, usual locations (tracked by hotspot ID), etc.. These ethical issues should be taken into consideration throughout the duration of the project.

Data Mining Goal

To support the business goal, we will focus on predicting whether free trial customers will purchase a subscription on their next transaction within one month after their first free trial ends.This is a supervised and predictive task, and an outcome variable called "Customer Spending" will be created to reflect the conversion of a customer to a paying subscriber in their next transaction within one month after the first free trial ends.

Data Description and Preparation

There are four tables provided by Migo which we can use in developing our algorithm: (1) transactions - reflects all subscriptions of every user whether free trial or paid subscription, (2) download - contains all the downloaded titles of every user, (3) engagement - duration of watched content of users per title and (4) title - list of titles available in the Migo library. These raw tables were filtered to reflect only users who had a free trial in their first transaction. Missing data were filled with the average for that variable while some were replaced with 0. We then created new columns that we believe will be good predictors of the algorithm. We also created the "Customer Spending" variable which is our outcome column that has a value of 'Yes' if the customer converts to a paying subscription in their next transaction within 30 days and a value of ‘No’ if otherwise. The screenshot of the raw tables and detailed preparation done for each table are in Appendix I.

After cleaning the data, we merged the tables using the user ID column. We were left with 46 predictors, 1 outcome column and 1,083 rows of first free-trial users. The merged table can be seen in Appendix II. We partitioned this data into three samples with data from September and October as training data, November as validation data and December as test data. We partitioned it this way to reflect the reality that previous months, the data available, are used to train the algorithm that will predict for the upcoming months. Since we are only applying the model to new users (given that a free trial should be their first transaction), there would be no duplication of users and we can evaluate the generalizability of the model.

Data Mining Solution

We applied Lasso Regression, Naive Bayes and Random Forest, and used the Lift and Gains Chart to evaluate the predictive performance of each method.

The Lift & Gains Chart (which can be seen in Appendix III) shows that Random Forest performed better than other models with a top decile equal to 4.089. This implies that the Random Forest model can perform approximately 4 times better than random selection in the highest 10%, while top deciles for both Naive Bayes and Lasso Regression were 3.578. We therefore decided to analyze the data using the Random Forest model.

Looking at the variable importance plot of Random Forest in Appendix IV, the variables which had the most predictive power are the unique and absolute engagement of the users. We also found the two most effective free-trial promotions which are the 2-Day Pass (prod_9) and PisoMigo_trial (prod_2). Download space seem to be also an important variable in predicting the conversion of users.

Recommendation

Based on the data mining results, the Marketing team can consider providing short-periods of free trial promotions since there is a strong correlation between the short-period free-trial program and the conversion of free-trial subscribers to a paying subscription. There is also a strong correlation between conversion to a paying subscriber and duration of watched content during the trial period. Given this data, we can only assume that the limited time given to a user to explore the Migo application gives them the motivation to make the most out of their free-trial program. These customers, therefore, tend to be more engaged with the application than customers that have long-trial period. Aside from this, Migo’s Marketing team can launch encouragement campaigns, such as a rewards program, where users can earn points when the continuously watch for more than one hour in the Migo application. The users can then use these points to redeem rewards or discounts in the Migo application. The company could also offer more subscription promotion to the most prospect customers before or when their free-trial ends to encourage them to subscribe to Migo’s services.
Despite the relatively low significance of other factors, the data mining result also shows that users who choose to watch action or romance genres, movies or series, or content provided by KBS tend to convert to a paying subscriber. Hence, Migo’s purchasing team can use this

information to improve the value of its services to the customers by providing their preferred content.

Conclusion

The most effective free-trial promotion that we previously identified was the 2-Day Pass. Corresponding to this, the dataset and the model could be run 3 times a week and should be analyzed on an on-going basis. When the accuracy of the model becomes lower than 75%, the data and algorithm used should be updated to improve the model's accuracy and predictive power.

Though we provided some recommendations based on the results from our data mining algorithm, it is important to know that our data and analysis had some limitations. First is the presence of missing values in the title and engagement datasets which could have tampered with the overall accuracy of the data. Moreover, the model was built based on the conversion of users within 30 days after their first-free trial and only covers their immediate succeeding transaction after the first free trial. Therefore, there might have been some users who have converted after 30 days or after several free trial programs which were not taken into consideration in our model. Lastly, there is no data regarding customer feedback which helps to understand the reason behind a customer's action that could have

Nevertheless, the model and results along with the recommendations presented are a good foundation to develop Migo's free-trial programs especially for new customers who have never interacted with Migo's mobile application before. As the business problem and goals of Migo change, they should update the parameters and measures used in the algorithm so that it properly reflects these new goals.

APPENDIX

I. Data

Table 1 - User's Purchase Transactions

unique_id_sk	spend	Customer Spending	hotspot_id	product_id	product_name	top_up	transaction_visit	paid_transaction_visit	spend_date_sk
1.4461E+13	1	Y	C00300	1	PisoMigo	1	1	1	20170927
1.4461E+13	0	N	C00300	2	PisoMigo_trial	20	1	1	20170927
1.4461E+13	0	N	C00300	2	PisoMigo_trial	50	1	1	20170927
1.4461E+13	79	Y	C00300	16	Red MC Starter P	79	1	1	20170927
1.4461E+13	25	Y	C00300	28	Buy 1 Take 1 Pro	25	1	1	20170927
1.4461E+13	0	N	C00300	2	PisoMigo_trial	50	1	1	20170927
1.4461E+13	0	N	C00324	2	PisoMigo_trial	50	1	1	20170915
1.4461E+13	79	Y	C00324	16	Red MC Starter P	79	1	1	20170915
3.57952E+14	0	N	C00225	2	PisoMigo_trial	50	1	1	20170929
3.57952E+14	49	Y	C00225	30	BYOC Starter Pac	49	1	1	20170929

For the transaction dataset, we first filtered the free-trial users who have transaction_visit = 1 and paid_transaction_visit = 0. The product ID was checked to make sure they all belong to free trial programs. There were users with multiple records so we compressed the table so users only take up one row in the table. The product ID was converted to a binary so we would not lose this information during the compression.

Afterwards, we filtered the original data to reflect users with transaction_visit = 2 and paid_transaction_visit = 1 and compared it with our previously filtered data. We created that outcome column "Customers Spending" based on the results of this comparison.

As a reference to the next tables, we again filtered the original data to reflect users with transaction_visit = 2 regardless of the value in paid_transaction_visit. We got the data for each user which will be used to compare and limit the data in the following tables.

Table 2 - User's Download Transactions

unique_id_sk	title_sn	expected_space_taken_mb	dl_takt_time_sec	interact_at	hotspot_id	mbps
8.69966E+14	5102	800.947	781.43	8/31/2017 0:08	C00235	4425.97
3.55613E+14	5462	745.391	356.95	8/31/2017 0:08	C00066	7316.81
3.59884E+14	5833	722.924	364.84	8/31/2017 0:09	C00066	7348.69
3.55613E+14	4387	623.747	420.78	8/31/2017 0:15	C00066	7316.81
3.59884E+14	5148	851.861	616.69	8/31/2017 0:20	C00066	7348.69
3.55613E+14	4390	612.703	395.08	8/31/2017 0:22	C00066	5373.34
3.57213E+14	4439	630.18	707.62	8/31/2017 0:23	C00323	3424.94
3.55613E+14	5384	664.239	24.65	8/31/2017 0:30	C00066	5373.34
3.57213E+14	4341	699.958	47.92	8/31/2017 0:36	C00323	2346.72
3.55613E+14	5833	722.924	348.28	8/31/2017 0:36	C00066	4664.09

For the download dataset, we filtered the downloads only for those reflected in our free-trial user database and made sure to cap the record filtered to before their next transaction visit. We kept the columns expected_space_taken_mb and dl_takt_time_sec and created a new variable named "date_difference" which equals the free-trial date minus the download date. Afterwards, we stored the titles download by the users in a separate table to use as comparison in the List of Title Films. Finally, the users with the same ID are compressed in one row where the predictor columns reflect the sum of the total space taken, the average of the download time, and the average of date_difference.

Table 3 - User's Video Engagement

unique_id_sk	engage_at_sk	title_sn	absolute_engage_minute	unique_engage_minute
8.66E+14	20171219	6482	0	0
8.66E+14	20171220	6482	12	0
8.66E+14	20171221	4178	2	2
8.66E+14	20171221	4179	0	0
8.66E+14	20171221	4413	12	3
8.66E+14	20171221	4748	1	0
8.66E+14	20171221	5374	4	0
8.66E+14	20171222	2973	1	0

For the engagement dataset, we filtered the engagement only for those reflected in our free-trial user database and made sure to cap the record filtered to before their next transaction visit. We then compress the users with the same ID to one row and summed the absolute_engagement_minute and unique_engage_minute columns.

Table 4 - List of Film Titles

title_sn	title	content_type	episode_number	running_length	major_migo_genre	migo_genre	season	content_provider	country	country_group
1838	Modern Farmer	SERIAL	1	60	Romance	Comedy, Romance	(Series)	SBS	KR	Regional
1839	Modern Farmer	SERIAL	2	60	Romance	Comedy, Romance	(Series)	SBS	KR	Regional
1840	Modern Farmer	SERIAL	3	60	Romance	Comedy, Romance	(Series)	SBS	KR	Regional
1841	Modern Farmer	SERIAL	4	60	Romance	Comedy, Romance	(Series)	SBS	KR	Regional
3731	Art of the Devil 2	MOVIE	0	103	Horror	Horror, Suspense	Art of the Devil 2	Five Star	TH	Regional
3736	Spongebob Squarepants - DELETE (10/17/18)	EPISODIC	1	24	Comedy	Comedy, Kids	(Series)	Viacom	US	Western
3737	Spongebob Squarepants - DELETE (10/17/18)	EPISODIC	2	24	Comedy	Comedy, Kids	(Series)	Viacom	US	Western
3738	Spongebob Squarepants - DELETE (10/17/18)	EPISODIC	3	24	Comedy	Comedy, Kids	(Series)	Viacom	US	Western
3880	Back to the 90s - DELETE (10/18/18)	MOVIE	0	112	Romance	Comedy, Romance	Back to the 90s - DELETE (10/18/18)	Mono Film	TH	Regional
3901	The Dog - DELETE	MOVIE	0	95	Action	Action	The Dog - DELETE	Mono Film	TH	Regional
858	Oh My Venus - DELETE	SERIAL	2	70	Romance	Romance	(Series)	KBS	KR	Regional
859	Oh My Venus - DELETE	SERIAL	3	70	Romance	Romance	(Series)	KBS	KR	Regional
860	Oh My Venus - DELETE	SERIAL	4	70	Romance	Romance	(Series)	KBS	KR	Regional
861	Oh My Venus - DELETE	SERIAL	5	70	Romance	Romance	(Series)	KBS	KR	Regional

For the title description dataset, we matched title_sn column from the download dataset then kept necessary columns such as content_type, running_length, migo_genre, content_provider, and country. We then created a binary table for each of the title's content_type, migo_genre, content_provider and country. Afterwards, we compressed the data so that users with the same ID are reflected in only one row. We summed the binary table we previously created to reflect the characteristics of all downloaded content and the running_length.

We merged the cleaned and filtered tables from the four separate tables to get one encompassing data set. We then created a new column "unwatched" which is the difference of the total running_length and the total unique_engage_minutes.

II. Merged Data

Table 5 - Merged Data

unique_id	hotspot_ic	spend_date_sk	prod_10	prod_2	prod_8	prod_9	prod_25	prod_1	prod_3	prod_16	prod_15	prod_18	prod_30	prod_28	prod_33	prod_32	prod_34
1.18E+15	C00335	12/31/2017	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1.18E+15	C00021	12/21/2017	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
1.20E+15	C00184	11/2/2017	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1.33E+15	C00021	12/16/2017	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1.35E+13	C00346	10/11/2017	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

unique_id	customer_	absolute_	unique_er	dl_takt_tir	avg_diff_c	expected_	SERIAL	MOVIE	EPISODIC	running_l	genre_Cor	genre_Ror	genre_Hor	genre_Dra	genre_Adv	genre_Far	genre_Act	genre_Crii
1.18E+15	No	0	0	153.21	0	668.291	0	0	0	0	0	0	0	0	0	0	0	0
1.18E+15	No	108	122	76.59036	0	6857.614	0	0	0	0	0	0	0	0	0	0	0	0
1.20E+15	No	959	791	81.87778	1.6	23735.98	56	10	8	4888	14	49	4	8	2	2	1	1
1.33E+15	No	0	0	481.585	0	1957.575	0	2	0	220	0	0	0	0	0	0	2	0
1.35E+13	No	0	0	139.815	0	783.404	0	0	0	0	0	0	0	0	0	0	0	0

unique_id	genre_Kid	SBS	KBS	Viva.Films	JTBC	Lionsgate	Regal.Ente	BBC	Turner.Br	KR	PH	US	GB	unwatched
1.18E+15	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.18E+15	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.20E+15	8	20	32	5	4	3	2	8	0	56	7	3	8	4097
1.33E+15	0	0	0	2	0	0	0	0	0	0	2	0	0	220
1.35E+13	0	0	0	0	0	0	0	0	0	0	0	0	0	0

III. Lift and Gains Chart

Figure 1 - Lift Chart

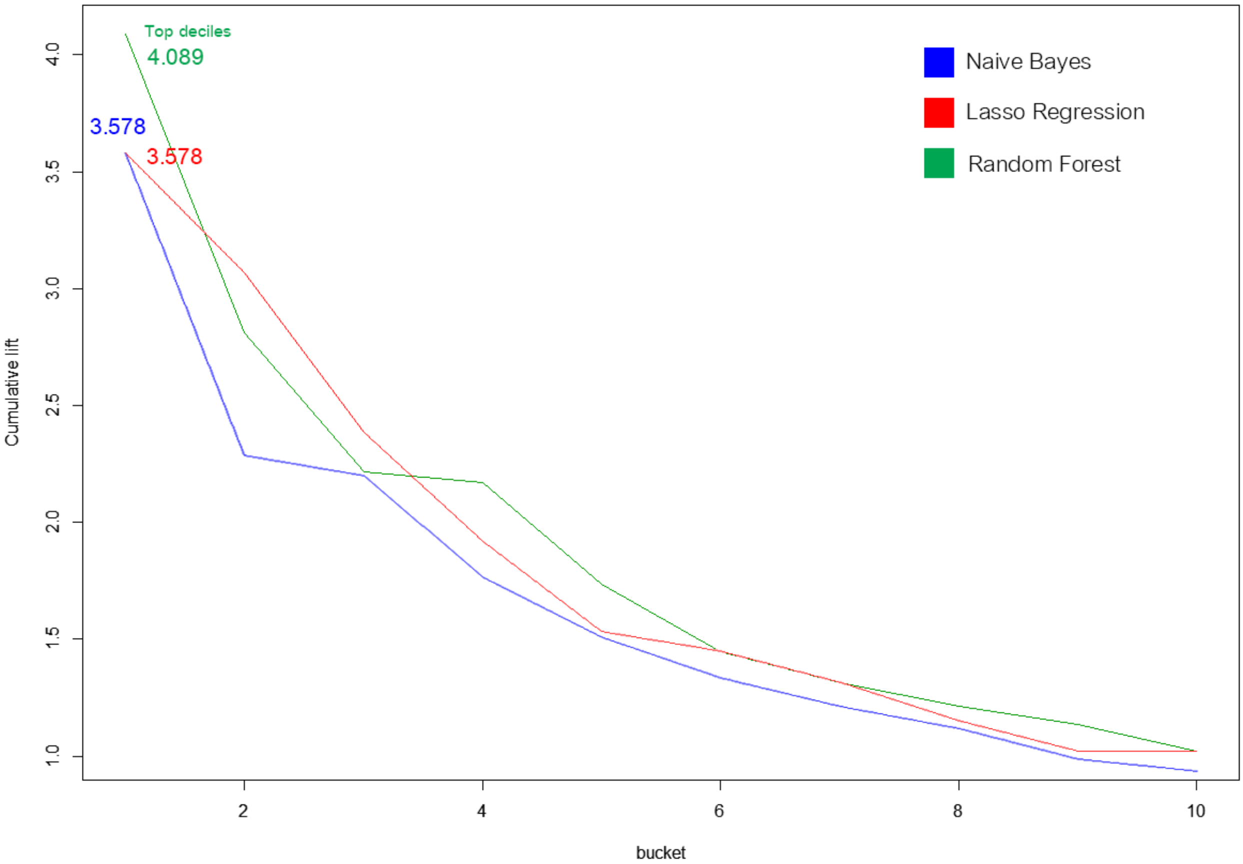
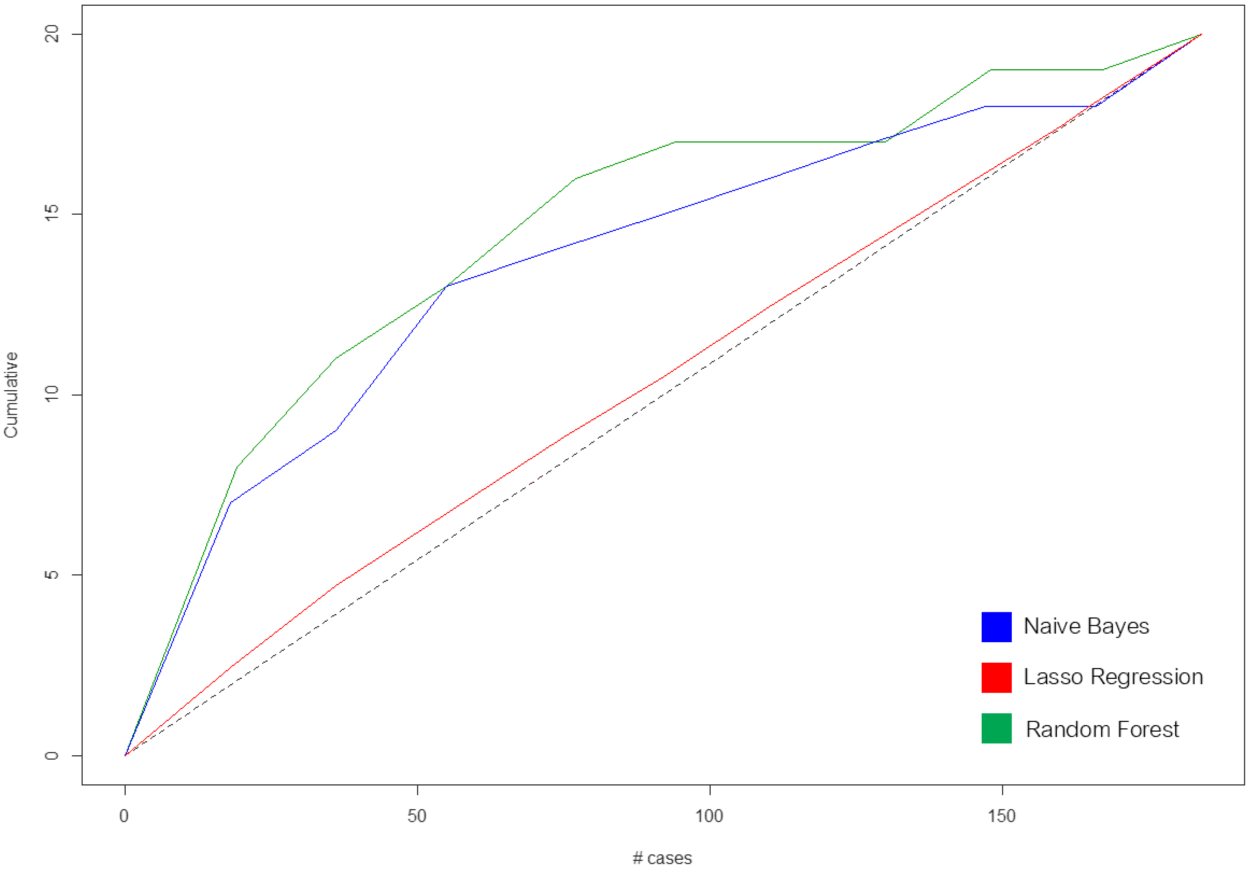


Figure 2 - Gains Chart



IV. Variable Importance in Random Forest

Figure 3 - Variable Importance Plot