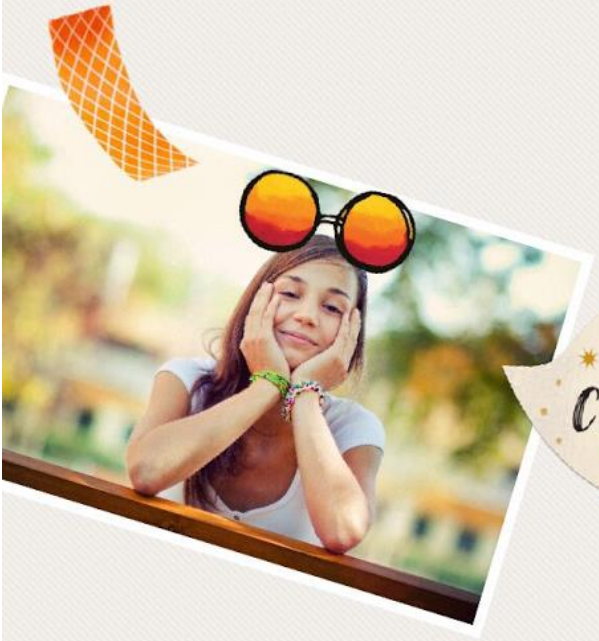
PIC•COLLAGE

Predicting PicCollage users' first purchase for targeted promotions

Team 2

Lynn Pan
Reggie Escobar
Eduardo Salazar
Uni Ang



Instructor：Galit Shmueli
Soumya Ray

National Tsing Hua University

Executive summary

PicCollage is an app to create amazing photo collages with custom stickers, fonts, background that makes creating collages a creative experience. It's biggest revenue streams comes from In-app purchase where users could pay to remove watermark, add custom stickers, and backgrounds. Also the have popup ads; but because ads strategy is implemented different for Android and IOS users in this project the model is not using features extracted from ads behavior and the scope of the model proposed is directly extracted from the behavior when users create their first collage.

PicCollage strategy to increase revenue from in-app purchase is targeting users that are more likely to make a purchase or offering promotions to users that are not likely to make a purchase and considering that most of the users that make a purchase they do it in the first collages. Therefore, in this project, "Predicting PicCollage user's first purchase probability for targeted promotions" the goal is to rank users that are more likely to make a first purchase when they create their first collage.

The dataset came from PicCollage, it contains the events(click, open_page, create_collage,etc) of September, 2017 and only new users. In total the data has 38,748,087 session of new users and 44 events triggered. The sample dataset we used for building our model is columns from user dimension and 79 derived variables from event data. Using users' behavior from first open the app to first collage save to predict if that user will make a first purchase after that. Partitioning the data into training data, validation data, and test data, and using oversampling to deal with the imbalanced data.

Logistic regression, decision tree, random forest, boosted tree were included as our predictive models. Decision trees are easy to interpret and are capable of giving insights about the important features; random forest and boosted tree are improved version of decision tree, which can produce really good and robust predictions. Models mentioned above were implemented and their performances were compared based on top 10% decile lift chart, which is created to test the model's ability to predict the top 10% of first purchase user. From the performance evaluation, boosted tree produced the best results. By using boosted tree model, it will get 1.78 times of first purchase users than randomly send promotion message to all users.

The recommendations for PicCollage is to use the model of boosted tree and oversampling for offering bundles/ discount to users that have a high probability of making a first purchase. For the date, in this predicting the data we are using is missing the October purchase; therefore, it should be more accurate if we add more data for purchase. In addition, to collect events Data per user for their days full history. Moreover, for variables, getting user information might help to predict first purchase earlier.

I. Problem Description

Business Goal

Our dataset came from our shareholder PicCollage. PicCollage wants to know more about what trigger their users to make the first purchase. And might apply different strategy to different group, for example sending different promotion to encourage users to purchase. However, the shortcoming now for PicCollage targeted promotions is hindered by the limited users' demographic data. So our business goal is to target users that are likely to make a first purchase for the purpose of sending personalize promotion message. The benefit of this is to create an effective marketing campaign for targeted segment. Nevertheless, depending on the type of marketing strategy PicCollage uses with the prediction, can possibly antagonize a user whose intention is not to many a purchase.

Data Mining Goal

Our data mining goal is ranking the user's with high probability of making a first purchase when they create their first collage. This is a predictive, forward-looking, and supervised task, since the model is trained on the previous records and predict the probability of making the first purchase when new users that had at least create a first purchase come in. The main outcome variable is the "binary for first purchase (1/0)."

The method we used are logistic regression, decision tree, random forest, boosted tree, and we use oversampling to deal with the imbalanced data. For performance evaluation, we use Lift chart, Decile lift chart, Sensitivity, and Specificity to evaluate our results, in order to find the best-performing model. But our main performance evaluation is the top 10% of decile lift chart.

II. Data Description

The raw dataset is new user' app events for the month of September of 2017 from PicCollage's database, containing 38,748,087 rows and 44 columns. The app event data is comprised of user dimensions and user events. The user dimensions include their device information, geographical information, app information and the time they first opened the app. The event dimensions include all the event names and data about when a user triggered such event. The sample dataset we used for building our model is columns from user dimension (app_instance_id, device_category, geo_conienet, first_open_time in DayofWeek) and 79 derived variables from event data. (Completed variables description please see **Table A1**.) And the total rows is 32,810. The figure 1 below is a screenshot of the dataset for only a couple of samples.

In our dataset, we only use the events before first collage save as our derived variables. As the timeline above, we want to use users' behavior from first open the app to first collage save to predict if that user will make a first purchase after that. **Figure A2** shows the trade-off incurred, by using the first collage a user saves, as predicting factor.

app_instan	device_cat	geo_contin	first_open_	num_event	purchase_v	purchase_s	purchase_t	num_share	num_share	create_coll	num_share	num_share	num_share	num_share	num_share	num_share
000991CC	mobile	Europe	2	19	0	0	0	0	0	0	0	0	0	0	0	0
000A55B4	mobile	Americas	2	11	0	0	0	0	0	1	0	0	0	0	0	0
00160CDE	mobile	Americas	5	12	0	0	0	0	0	1	0	0	0	0	0	0
0028F508	tablet	Europe	3	75	0	0	0	0	0	4	0	0	0	0	0	0
002A1C46	mobile	Americas	4	10	0	0	0	0	0	1	0	0	0	0	0	0
002C5AA4	mobile	Americas	5	8	0	1	0	0	0	0	0	0	0	0	0	0
002EF54E	mobile	Europe	1	10	0	0	0	0	0	1	0	0	0	0	0	0

remix_cat_	remix_cat_	font_SFUIT	font_Baske	font_Dawn	font_Huds	font_Impac	font_King	font_Muse	font_Pacifi	font_Robot	font_Thirst	create_coll	create_coll	Login	num_sticke	first_purcha
0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0
0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0

Figure 1. Screenshot of final used dataset

III. Brief Data Preparation Details

The dataset first is in json structure and each row is by session (open the app to close the app), so we first extract the sample dataset from database and make it into two tables — user_info and events_info[1]. Then, we build a mysql database to filter data by user, and filter those users' events before first purchase and first collage save. After that, we create 79 derived variables from event_info table, and add them to user_info to make our sample dataset.

Last, we deal with the missing value. The only missing value is geo_continent. We use column **user_default_language** to trace back the continent (using Mode). We partition the data into training dde _ ata, validation data, and test data, and using oversampling to deal with the imbalanced data, which illustrated in **Figure A3.**

IV. Data Mining Solutions

- **Benchmark**： naive , using the most popular class "0" .
- **Logistic Regression**

  We use stepwise for our variables selection. It comes out 12 predictors, which are num_events, create_collage_empty, num_background_try, num_frame_try, avg_of_image_export, avg_photo_facebook, remix_cat_Back_to_School, remix_cat_Congrats, remix_cat_Just_for_Fun, remix_cat_Labor_Day_Weekend, font_Roboto_BlackItalic, create_collage_grid, Login.

  The summary report and performance are illustrated in **Figure A4**.
- **Decision Tree**

  Decision tree can automatically select the best predictors for us. Also, can see which predictor is significant to the outcome. **Figure A5** shows the decision tree diagram. **Figure A6** shows the summary report & performance evaluation.

[1] "Pico Extractor" project is a parser to get the predictors https://github.com/eduardo-salazar/pico_extractor

- **Random Forest**

  Since our sample size is not big, different seed set would come out different single tree results. Random Forest can help solve this problem. From **Figure A7** can see that it has a better performance than decision tree. We also try random forest without oversampling, which is illustrated in **Figure A8**. Can see random forest with oversampling performs better than one without oversampling.

- **Boosted Tree**

  When random forest has a well performance, we expected boosted tree would have a better performance than random forest. This time, we also look at variables' relative influence, which is illustrated in **Figure A9-1** and **Figure A9-2**. Boosted tree summary report and performance evaluation show in **Figure A10**.

- **Performance evaluation**

  Our main performance standard is top 10% decile lift chart, which shows below. From the lift chart Random forest and boosted tree's curve are very close, so we restored the model to the top 10% decile lift chart. In **Figure 2** below, Boosted tree has the highest around 1.78 times better than using random forest . In **Figure A11**, can see their sensitivity and specificity are similar.

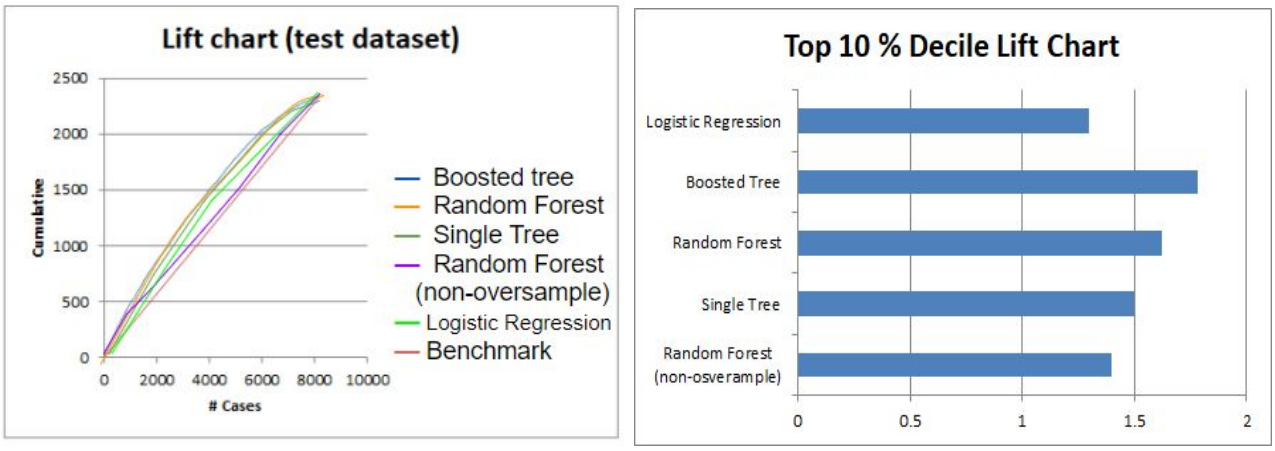


Figure 2. Lift chart and top 10% Decile lift chart

V. Conclusion

In this study, we applied different data mining methods to filter out the most likely users that will make the first purchase. After evaluating our solutions with top 10% decile lift curves, we suggested that boosted tree showed the best results. The followings are some operational and data collecting recommendations:

1. Due to the unbalanced dataset and ranking goal, we suggest to adopt over-sampling.
2. Limitation: The dataset we are using now is missing the purchase in October. Better to collect events data per user for their 30 days full history.
3. Variables recommendation: Getting more demographic user data can help predict first purchase behavior earlier. From our boosted tree variables relative influence results, avg_photo_library and avg_image_export have very high influence on model. If can get more relative params value, such as number of photos in library, may bring about improvements in predictions.
4. Offering bundles/discount to users that have a high probability of making a first purchase.

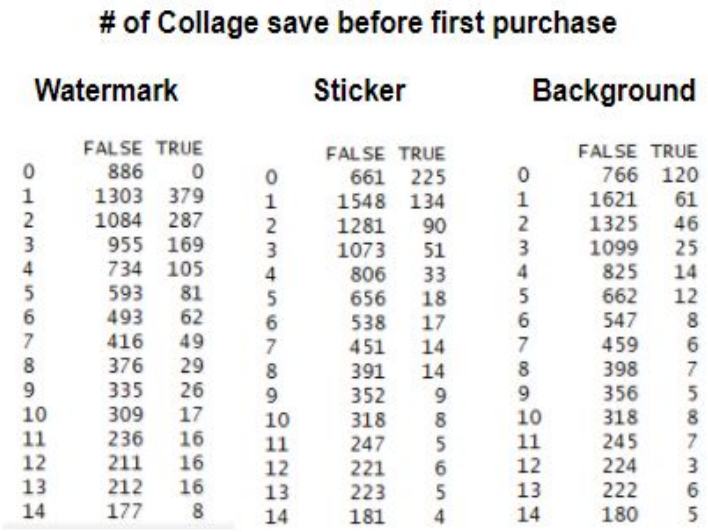
Appendix

Table A1 variables Description

Column	Data Type	How to get it
UserId	number	User App instance ID
country	categorical	country column from user file
first_open_DOW	categorical	1: Sunday ~ 7: Saturday
Login	binary(login:1/Notlogin: 0)	(Not login or login) from all collages Heartbeat – CBAuth logged in
purchase_watermark	number	count event Purchased_Product AND type=watermark
purchase_sticker	number	count event Purchased_Product AND type=sticker
purchase_background	number	count event Purchased_Product AND type=background
first_purchase	binary(purchase:1/non-purchase:0)	use column purchase_watermark,purchase_sticker,ppurcjase_background to create binary outcome
create_collage_empty	number	Count event : Tap_create_collage_empty_collage
create_collage_grid	number	Count event : Create screen - tap grid icon
create_collage_remix	number	Count event : Create screen - tap remix button
remix_cat_Autumn	number	Count : Remix collage AND category=Autumn
remix_cat_Baby	number	Count : Remix collage AND category=Baby
remix_cat_Back to School	number	Count : Remix collage AND category=Back to School
remix_cat_Causes	number	Count : Remix collage AND category=Causes
remix_cat_Congrats	number	Count : Remix collage AND category=Congrats
remix_cat_Everyday Life	number	Count : Remix collage AND category=Everyday Life
remix_cat_Family	number	Count: Remix collage AND category=Family
remix_cat_Fashion	number	Count : Remix collage AND category=Fashion
remix_cat_Featured	number	Count : Remix collage AND category=Featured
remix_cat_For Dad	number	Count : Remix collage AND category=For Dad
remix_cat_For Mom	number	Count : Remix collage AND category=For Mom
remix_cat_Get Together	number	Count : Remix collage AND category=Get Together
remix_cat_Get Well	number	Count : Remix collage AND category=Get Well
remix_cat_Happy Birthday	number	Count : Remix collage AND category=Happy Birthday
remix_cat_Just for Fun	number	Count : Remix collage AND category=Just for Fun
remix_cat_Labor Day Weekend	number	Count : Remix collage AND category=Labor Day Weekend
remix_cat_Love	number	Count : Remix collage AND category=Love
remix_cat_phone case	number	Count : Remix collage AND category=phone case
remix_cat_Summer	number	Count : Remix collage AND category=Summer
remix_cat_Thank You	number	Count : Remix collage AND category=Thank You
remix_cat_Wedding	number	Count : Remix collage AND category=Wedding
avg_photo_library	number	Average event Add photo AND from=Photo Library AND num_of_image=2}

avg_photo_facebook	number	Average event Add photo AND from=Facebook Photo Picker AND num_of_image=2}
avg_photo_instagram	number	Average event Add photo AND from=Instagram Picker AND num_of_image=2}
num_photo_web	number	count event Add_Photos___Image_from_Web
avg_stickers	number	Avg sticker grouped by collage Count: Add_sticker_from_bundle avergae by (Save_collage) to find Avg sticker grouped by collage
font_.SFUIDisplay	number	count event Text editor-finish font AND font_name=Andale (top 8 used font type)
font_Baskervile		
font_DawningofaNewDay		
font_HudsonNY		
font_Impact		
font_KingBasil-Regular		
font_MuseoSlabW01-700		
font_Pacifico-Regular		
font_Roboto-BlackItalic		
font_ThirstyRoughReg		
num_share_Facebook	number	Count Event : Share menu options- FB
num_share_Instagram	number	Count Event : Share menu options- Instagram
num_share_Library	number	Count Event : Share menu options- save to library
num_share_PicCollage	number	Count Event: Share menu options- PicCollage
num_share_message	number	Count Event : Share menu options- message
num_share_Twitter	number	Count Event : Share menu options- Twitter
num_share_copy_link	number	Count Event : Share menu options- copy_link
num_share_Others	number	Count Event : Share menu options- Others
num_share_FBmessage	number	Count Event : Share_menu_optoins___FB_Messenge
num_share_email	number	Count Event : Share menu options- email
num_share_kiteHP	number	Count Event : Share menu options- kiteHP
num_back_photo_icon_click	number	Count Event : Background_picker__tap_photo_icon
num_back_search_icon_click	number	Count Event : Background_picker__tap_search_icon
num_background_try	number	Count Event : Select URL background
doodle_stroke_count	number	AVG Event : Doodle editor-finish doodle AND stroke_count=2
num_frame_try	number	Count Event : Picked grid
num_clip	number	Count Event : Clip_picker___done_button_presse
avg_scrap_collage_save	number	AVG event : Scrap_number_in_collage AND number=7
total_num_collage_save	number	Count Event : Save_collage
num_of_sticker_export	number	Count event and AVG by collage Collage editor-export button AND num_of_stickers=2

type_of_background_export	categorical	Most frequent type from all collages Collage editor-export button AND background_type=bundled
avg_of_image_export	number	Count event and AVG by collage Collage editor-export button AND num_of_image_scraps=2
avg_of_text_export	number	Count event and AVG by collage Collage editor-export button AND num_of_texts=2
avg_of_doodle_export	number	Count event and AVG by collage Collage editor-export button AND num_of_doodle=2
type_of_collage_export_freedom	categorical	Most frequent type from all collages Collage_editor___export_buttonAND collage_style=freeform
type_of_collage_export_freedom	categorical	Most frequent type from all collages Collage_editor___export_button AND collage_style=grid
type_of_collage_export_freedom	categorical	Most frequent type from all collages Collage_editor___export_button AND collage_style=template
Sticker_store__buy_sticker_free	number	Count event :Sticker_store___buy_sticker AND type=Sticker - free
num_sticker_try	number	Count event : Sticker_picker___preview_sticker
num_event	number	Count event before one collage



	Watermark FALSE	Watermark TRUE	Sticker FALSE	Sticker TRUE	Background FALSE	Background TRUE
0	886	0	661	225	766	120
1	1303	379	1548	134	1621	61
2	1084	287	1281	90	1325	46
3	955	169	1073	51	1099	25
4	734	105	806	33	825	14
5	593	81	656	18	662	12
6	493	62	538	17	547	8
7	416	49	451	14	459	6
8	376	29	391	14	398	7
9	335	26	352	9	356	5
10	309	17	318	8	318	8
11	236	16	247	5	245	7
12	211	16	221	6	224	3
13	212	16	223	5	222	6
14	177	8	181	4	180	5

Figure A2 This is a small sample 12,294 users, including 9899 non-purchased users and 2395 purchased users. This graph shows the full history before first purchase. Can see some users didn't save any collage before they make the first purchase. The trade-off of waiting users for at least creating one collage is losing those users don't save any collage before purchase.

	Sample		Over-sampling	
	# record	% purchase	# record	% purchase
Training data	10,000	28%	9344	50%
Validation data	11,405	28%	8202	28%
Test data	11,405	28%	8202	28%

Figure A3 Partition

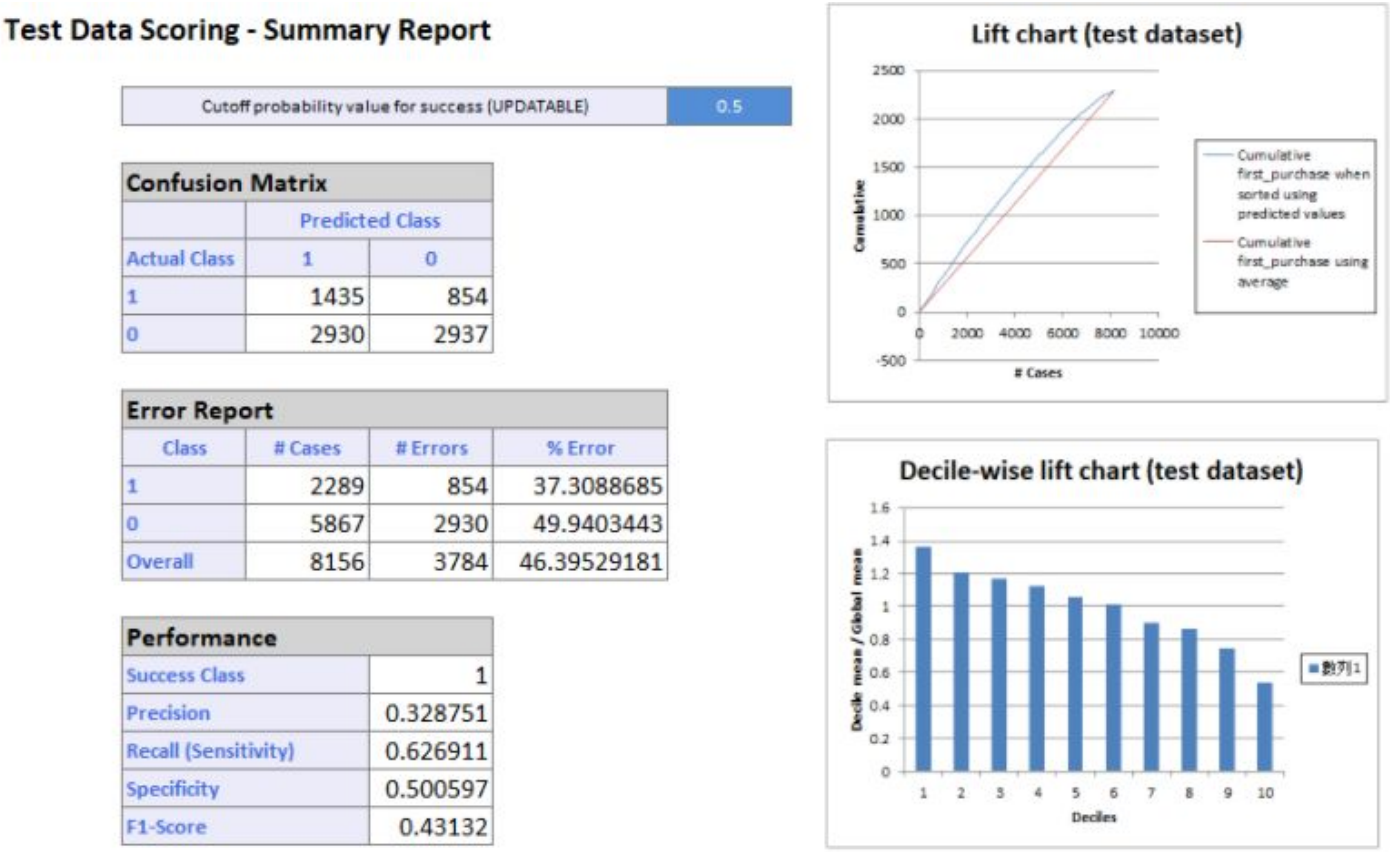


Figure A4 Logistic Regression summary report & performance evaluation

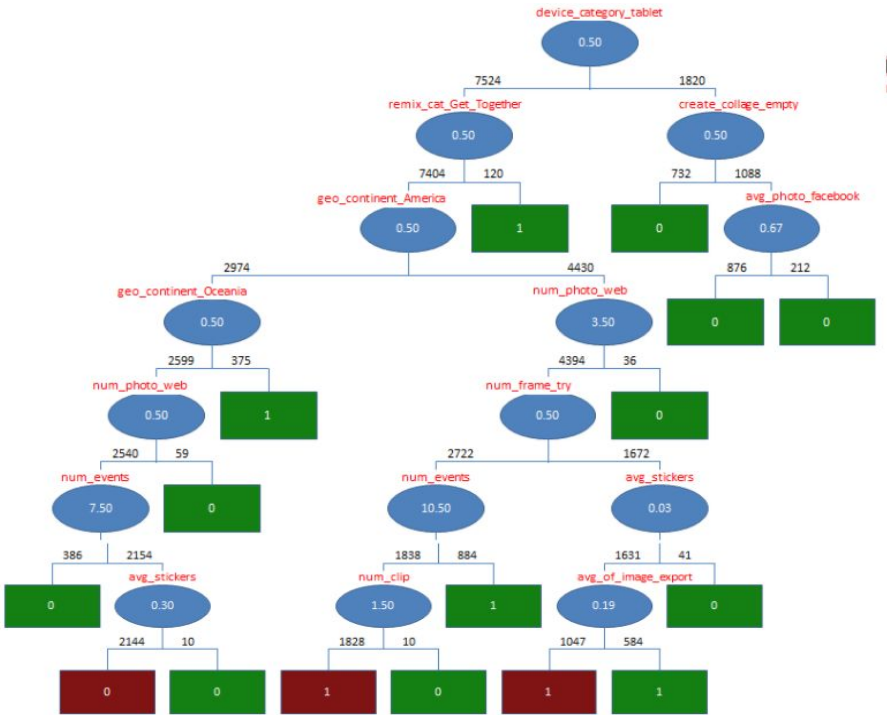


Figure A5 Decision Tree diagram

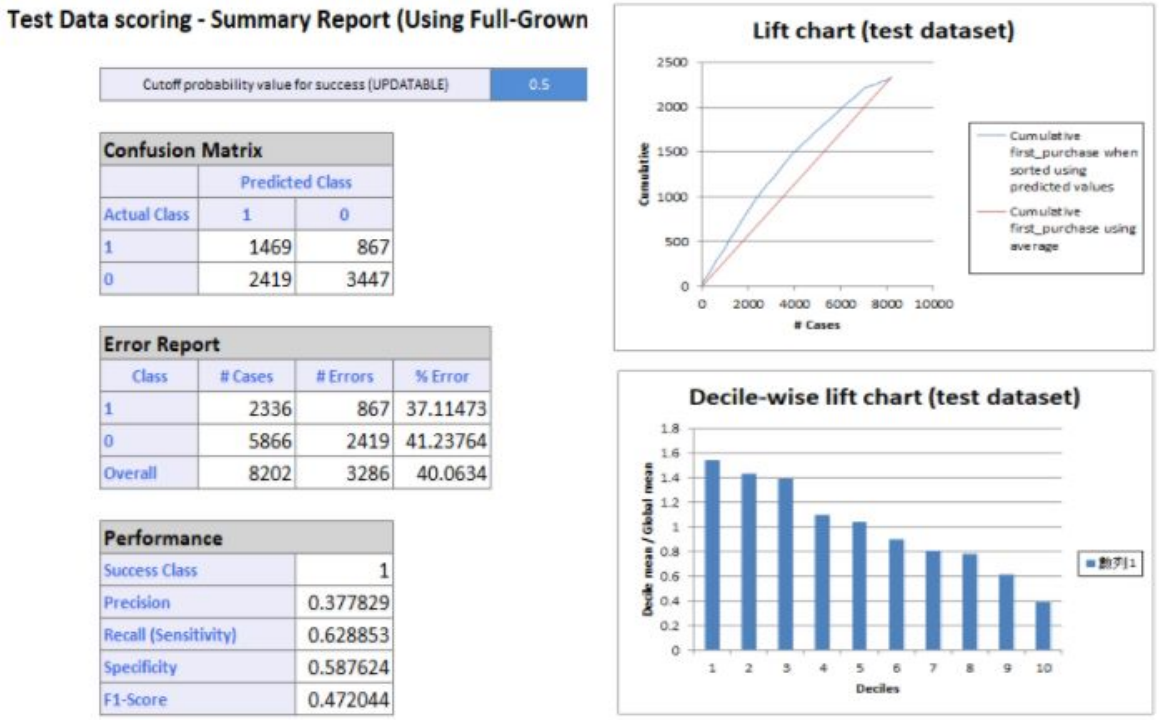


Figure A6 Decision Tree summary report & performance evaluation

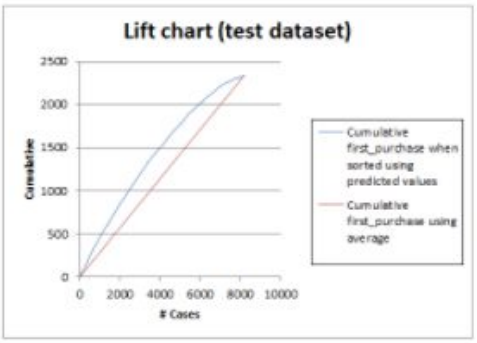
Test Data scoring - Summary Report

Cutoff probability value for success (UPDATABLE)	0.5

Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	1494	842
0	2501	3365

Error Report			
Class	# Cases	# Errors	% Error
1	2336	842	36.04452
0	5866	2501	42.63553
Overall	8202	3343	40.75835

Performance	
Success Class	1
Precision	0.373967
Recall (Sensitivity)	0.639555
Specificity	0.573645
F1-Score	0.471963



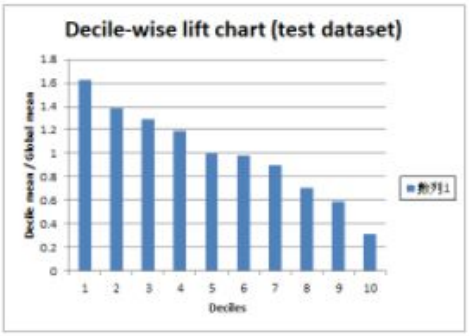


Figure A7 Random forest (with over-sampling) summary report & performance evaluation

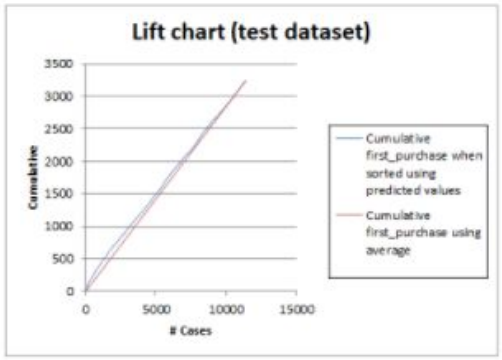
Test Data scoring - Summary Report

Cutoff probability value for success (UPDATABLE)	0.5

Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	67	3172
0	0	8166

Error Report			
Class	# Cases	# Errors	% Error
1	3239	3172	97.93146
0	8166	0	0
Overall	11405	3172	27.81236

Performance	
Success Class	1
Precision	1
Recall (Sensitivity)	0.020685
Specificity	1
F1-Score	0.040532



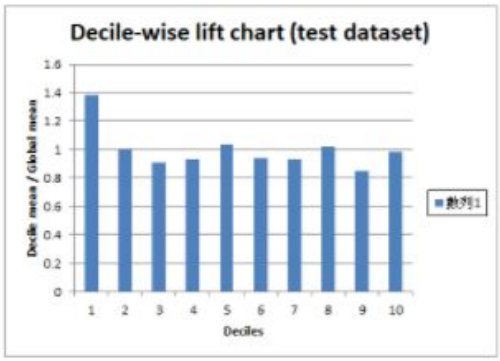


Figure A8 Random forest (without over-sampling) summary report & performance evaluation

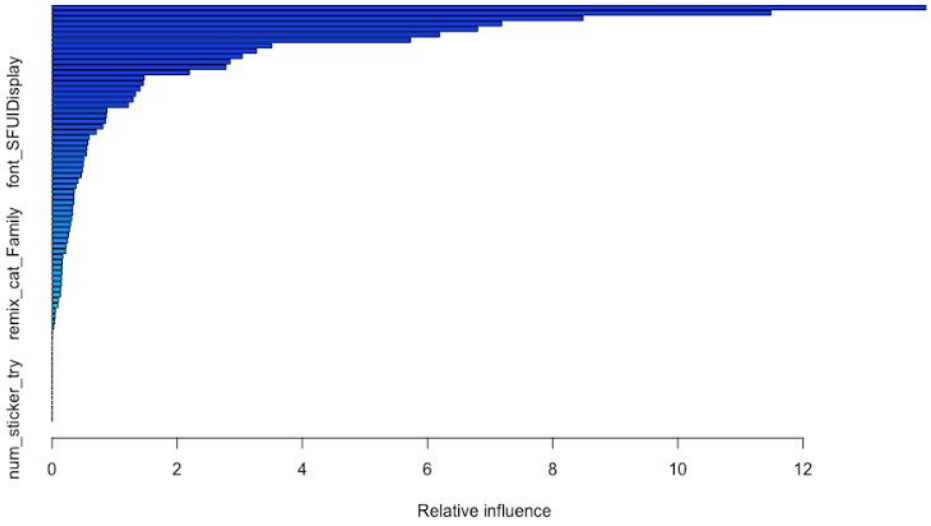


Figure A9-1 Variables' relative influence

```
> summary(objModel)
```

	var	rel.inf
avg_photo_library	avg_photo_library	13.960964164
avg_of_image_export	avg_of_image_export	11.487679575
geo_continent	geo_continent	8.486017858
num_events	num_events	7.184206499
avg_photo_facebook	avg_photo_facebook	6.801353846
device_category	device_category	6.193313815
num_frame_try	num_frame_try	5.727832967
first_open_DOW	first_open_DOW	3.512511398
num_background_try	num_background_try	3.267728578
create_collage_grid	create_collage_grid	3.040554492
avg_of_text_export	avg_of_text_export	2.848137418
num_of_sticker_export	num_of_sticker_export	2.780245657
create_collage_empty	create_collage_empty	2.192813909
remix_cat_Wedding	remix_cat_Wedding	1.474219084
Login	Login	1.461028291
font_KingBasil_Regular	font_KingBasil_Regular	1.407394761
num_clip	num_clip	1.335647305
create_collage_remix	create_collage_remix	1.293903692
avg_stickers	avg_stickers	1.220519738
num_back_photo_icon_click	num_back_photo_icon_click	0.879232486
num_photo_web	num_photo_web	0.867505566
font_SFUIDisplay	font_SFUIDisplay	0.853069123
font_Pacifico_Regular	font_Pacifico_Regular	0.813860072
type_of_collage_export_grid	type_of_collage_export_grid	0.706611377
avg_doodle_stroke_count	avg_doodle_stroke_count	0.593417727
type_of_collage_export_freeform	type_of_collage_export_freeform	0.573701546
remix_cat_Happy_Birthday	remix_cat_Happy_Birthday	0.556877651
font_ThirstyRoughReg	font_ThirstyRoughReg	0.548637610
type_of_background_export_bundled	type_of_background_export_bundled	0.509975480
type_of_background_export_photo_framework	type_of_background_export_photo_framework	0.499573043
avg_of_doodle_export	avg_of_doodle_export	0.486197353
remix_cat_Featured	remix_cat_Featured	0.465465077
remix_cat_Just_for_Fun	remix_cat_Just_for_Fun	0.416651691
font_Impact	font_Impact	0.384183001
remix_cat_For_Dad	remix_cat_For_Dad	0.355920576
Sticker_store__buy_sticker_free	Sticker_store__buy_sticker_free	0.350078667
remix_cat_Causes	remix_cat_Causes	0.348690540
type_of_collage_export_template	type_of_collage_export_template	0.329030936
remix_cat_Labor_Day_Weekend	remix_cat_Labor_Day_Weekend	0.328967962
remix_cat_Back_to_School	remix_cat_Back_to_School	0.315533707
num_back_search_icon_click	num_back_search_icon_click	0.297123556
remix_cat_Thank_You	remix_cat_Thank_You	0.278755478
remix_cat_Love	remix_cat_Love	0.260214332
font_Baskerville	font_Baskerville	0.249196544
font_MuseoSlabW01_700	font_MuseoSlabW01_700	0.224569447
remix_cat_Congrats	remix_cat_Congrats	0.219510415
remix_cat_Summer	remix_cat_Summer	0.171615940
remix_cat_Baby	remix_cat_Baby	0.161929785
remix_cat_Get_Together	remix_cat_Get_Together	0.158402277
remix_cat_Family	remix_cat_Family	0.156834246
font_Roboto_BlackItalic	font_Roboto_BlackItalic	0.152058751
remix_cat_Fashion	remix_cat_Fashion	0.150594645
font_DawningofaNewDay	font_DawningofaNewDay	0.138749587
remix_cat_Everyday_Life	remix_cat_Everyday_Life	0.136987797
type_of_background_export_assets_library	type_of_background_export_assets_library	0.103511472

Figure A9-2 Variables' relative influence

Can see avg_photo_library and avg_image_export has very high influence in boosted tree.

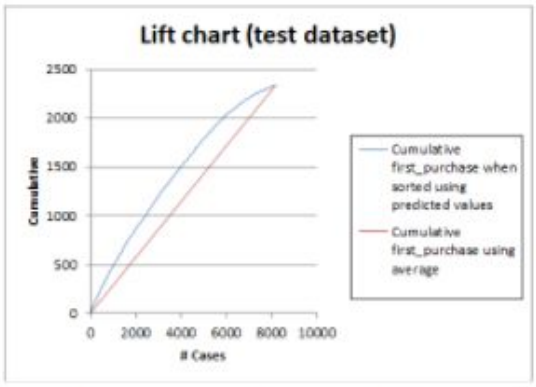
Test Data scoring - Summary Report

Cutoff probability value for success (UPDATABLE)	0.5

Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	1398	938
0	2273	3593

Error Report			
Class	# Cases	# Errors	% Error
1	2336	938	40.15411
0	5866	2273	38.74872
Overall	8202	3211	39.14899

Performance	
Success Class	1
Precision	0.380823
Recall (Sensitivity)	0.598459
Specificity	0.612513
F1-Score	0.465457



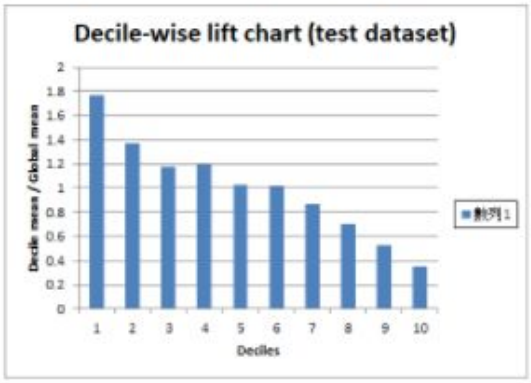


Figure A10 Boosted tree summary report & performance evaluation

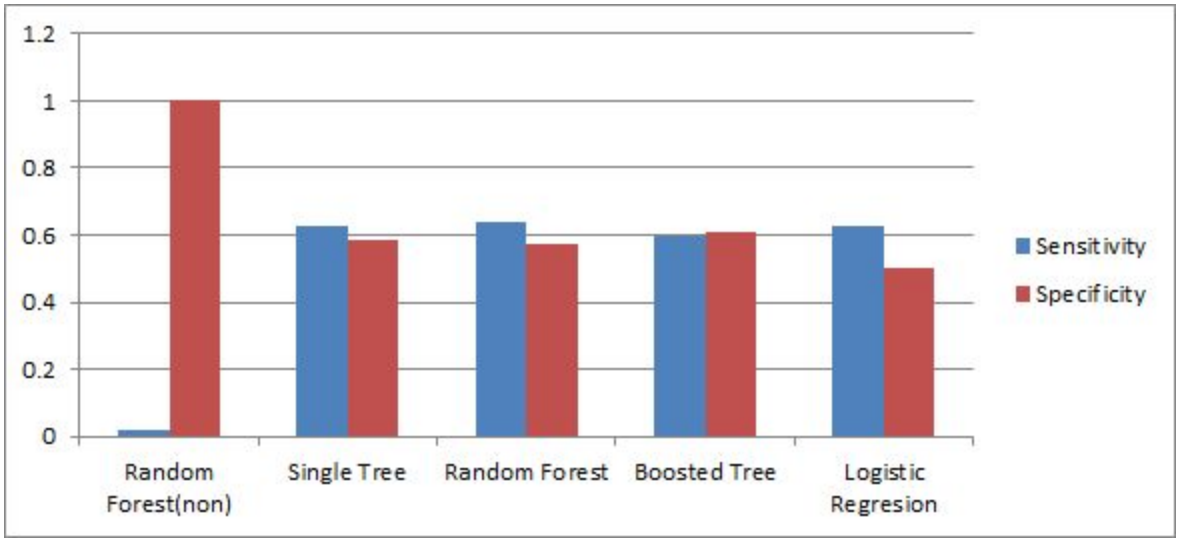


Figure A11 Sensitivity and Specificity performance