Predicting PicCollage users' first purchase

for targeted promotions

Team 2

Reggie Escobar . Eduardo Salazar
Uni Ang . Lynn Pan

PIC•COLLAGE

Cardinal Blue, Inc.

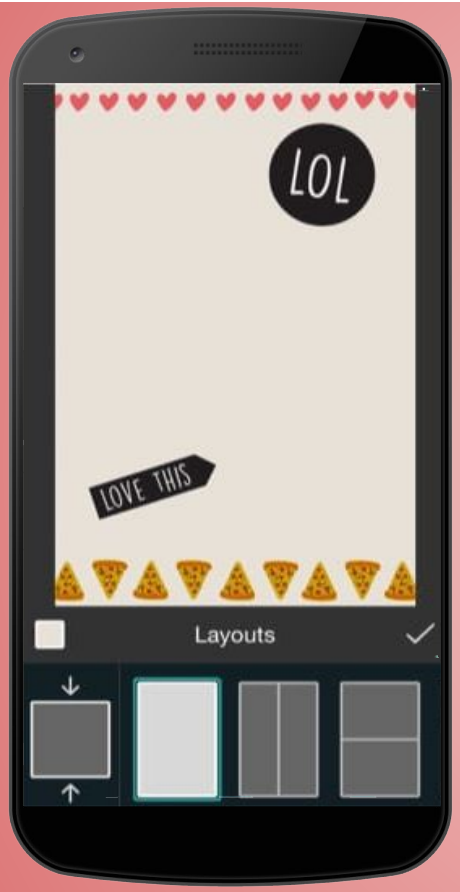
Founded in 2011

100m
installs

$2.3m
Seed funding
(2013)

In-app purchases - backgrounds; stickers & watermark removal

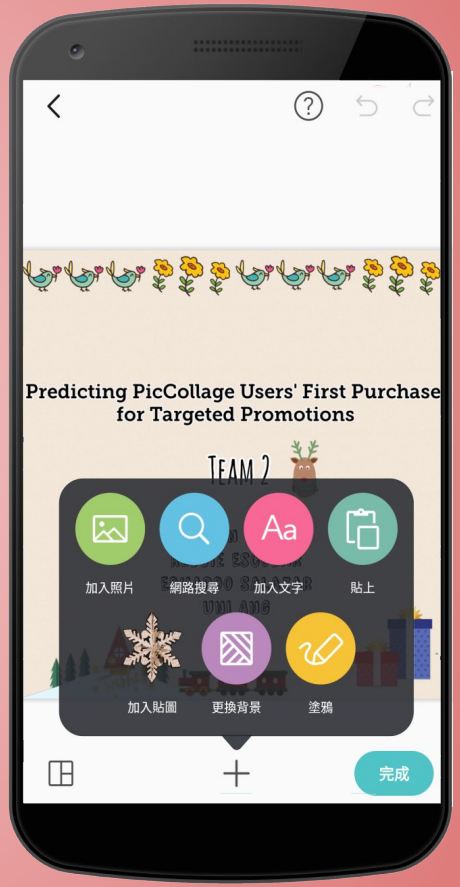
About PicCollage



WATERMARK

STICKER

BACKGROUND



Business Goal & Humanistic Evaluation

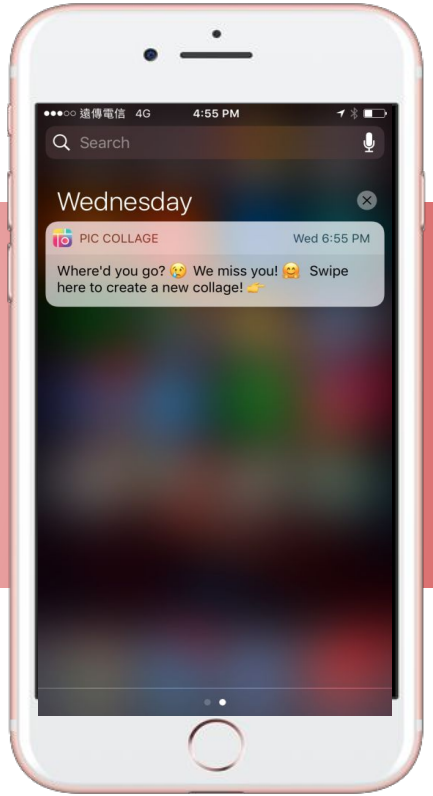
Stakeholder : PicCollage

Business Goal

Target users likely to make a first purchase

↓

Send personalized promotions



Problem : Limited user info data hinders user specific targeted promotions

Data Mining Goal

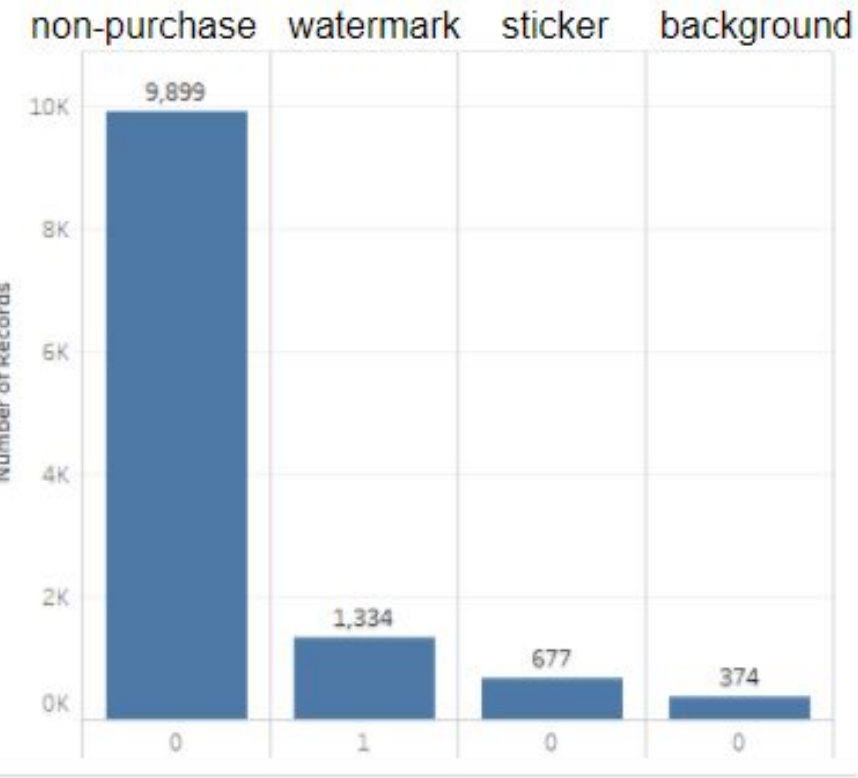
Ranking the user's with high probability of making a first purchase when they create their first collage

Supervised . Forward-looking

Categorical: Binary for first purchase (Y/N)

Data Source

One Month New user (2017/9) data from firebase



of Collage save before first purchase

Watermark

	FALSE	TRUE
0	886	0
1	1303	379
2	1084	287
3	955	169
4	734	105
5	593	81
6	493	62
7	416	49
8	376	29
9	335	26
10	309	17
11	236	16
12	211	16
13	212	16
14	177	8

Sticker

	FALSE	TRUE
0	661	225
1	1548	134
2	1281	90
3	1073	51
4	806	33
5	656	18
6	538	17
7	451	14
8	391	14
9	352	9
10	318	8
11	247	5
12	221	6
13	223	5
14	181	4

Background

	FALSE	TRUE
0	766	120
1	1621	61
2	1325	46
3	1099	25
4	825	14
5	662	12
6	547	8
7	459	6
8	398	7
9	356	5
10	318	8
11	245	7
12	224	3
13	222	6
14	180	5

Share Collage : type + number

Export collage : sticker / background/ …..

28 commits | 1 branch | 0 releases | 2 contributors

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

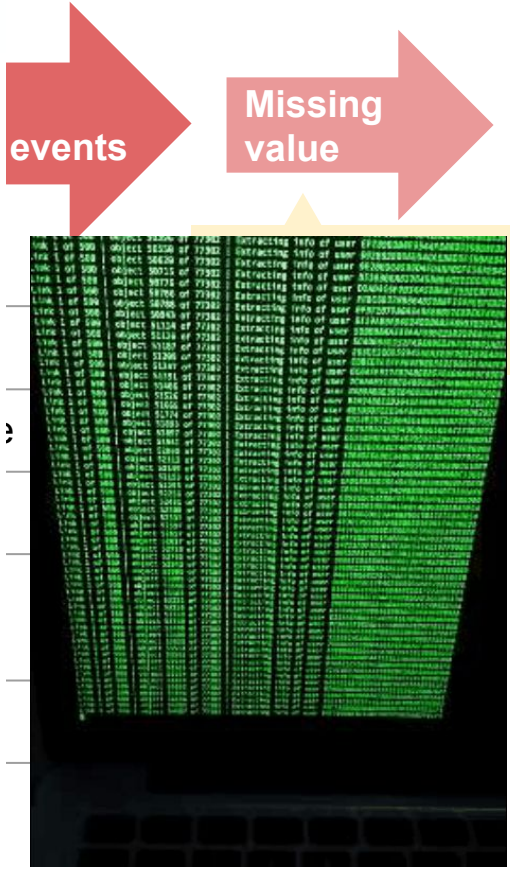
eduardo-salazar updates — Latest commit fa58836 3 days ago

models	Added new colums refering to params in event dim	27 days ago
queries	updates	11 days ago
services	updates	11 days ago
.config.yml.example	Creating schema from sql file	12 days ago
.gitignore	Getting files and sample of files	2 months ago
Gemfile	Creating schema from sql file	12 days ago
Gemfile.lock	Creating schema from sql file	12 days ago
README.md	Finished main_to_mysql. Now it saves directly to mysql tables	12 days ago
Rakefile	Finished data extracting code	2 months ago
correlation.rb	Getting ID from file	7 days ago
download.rb	init structure to dump to mysql database	12 days ago
main.rb	Finished main_to_mysql. Now it saves directly to mysql tables	12 days ago
main_to_mysql.rb	updates	3 days ago

README.md

Pico Extractor

This project is a parser to get the predictors



Methods & Performance Evaluation

- Task : Ranking
- Benchmark : naive (all class “0”)
- Method
  - Naive Bayes (Binned variables)
  - Classification tree (single)
  - Random Forest
  - Boosted Tree
  - Logistic Regression
- Performance measure
  - Lift Chart
  - Decile lift chart
  - Sensitivity
  - Specificity

Method : Random Forest

Non-oversample

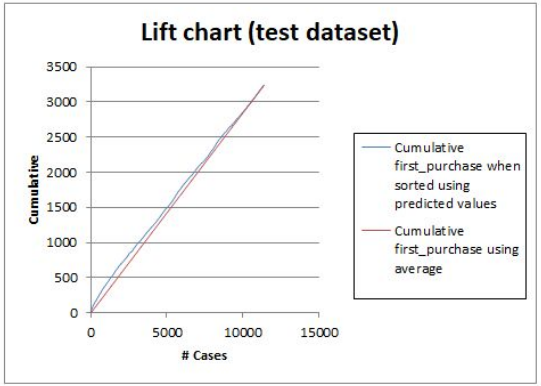
Test Data scoring - Summary Report

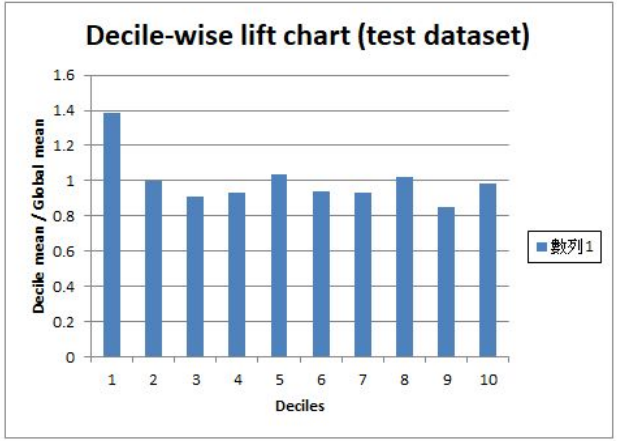
Cutoff probability value for success (UPDATABLE)	0.5

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

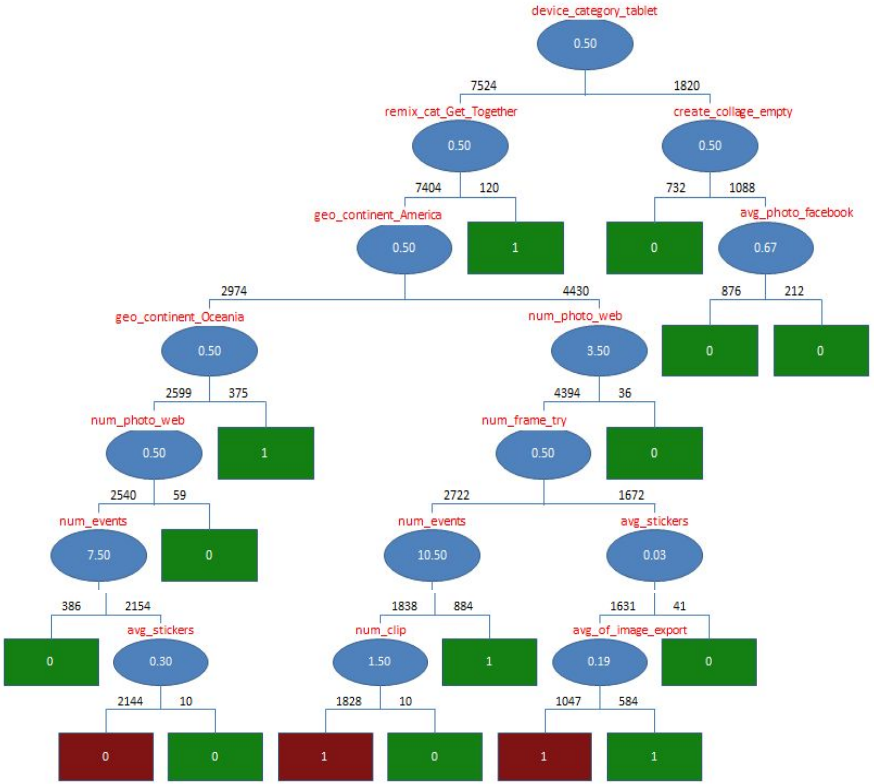
Performance	
Success Class	1
Precision	1
Recall (Sensitivity)	0.020685
Specificity	1
F1-Score	0.040532

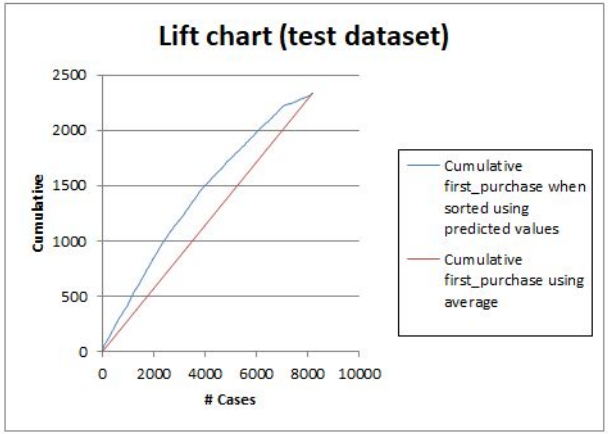


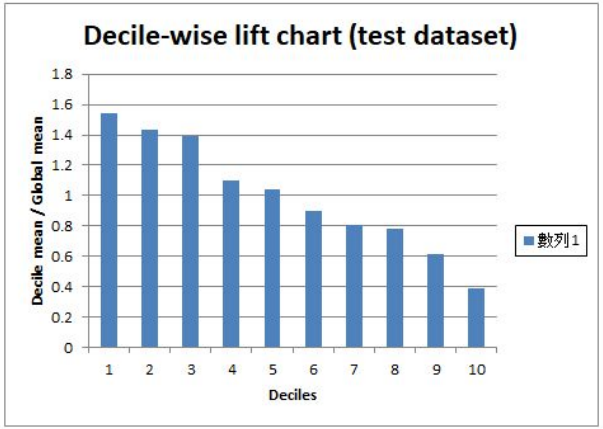


Method : Single Tree

oversample / Full Tree / terminal 934







Method : Random Forest oversampling

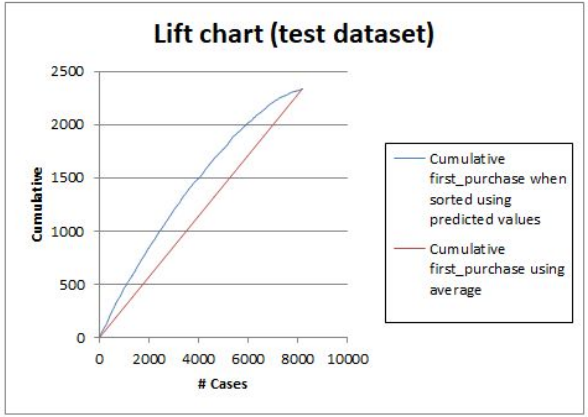
Test Data scoring - Summary Report

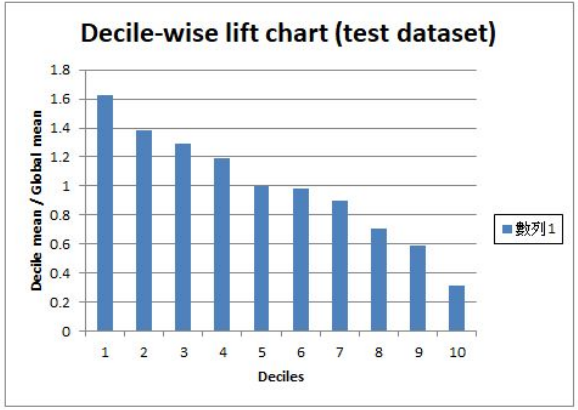
Cutoff probability value for success (UPDATABLE)	0.5

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





Method : Boosted Tree

oversampling

Test Data scoring - Summary Report

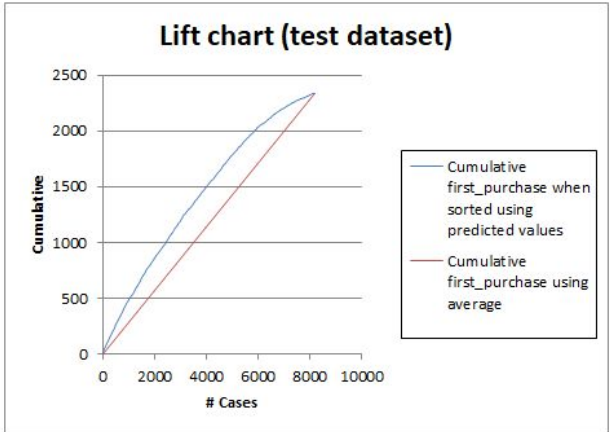
Cutoff probability value for success (UPDATABLE)	0.5

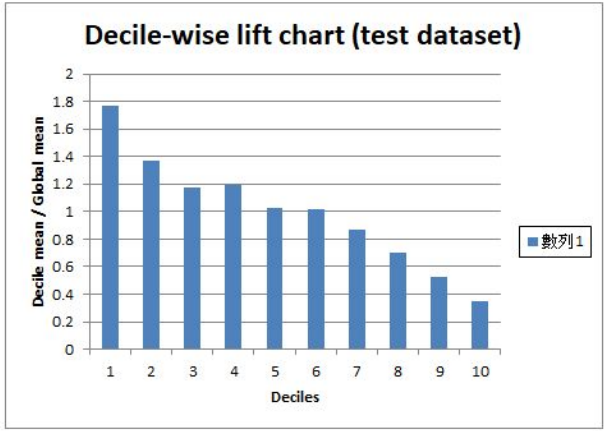
Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	1398	938
0	2273	3593

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

var	rel.inf
avg_photo_library	13.960964164
avg_of_image_export	11.487679575
geo_continent	8.486017858
num_events	7.184206499
avg_photo_facebook	6.801353846



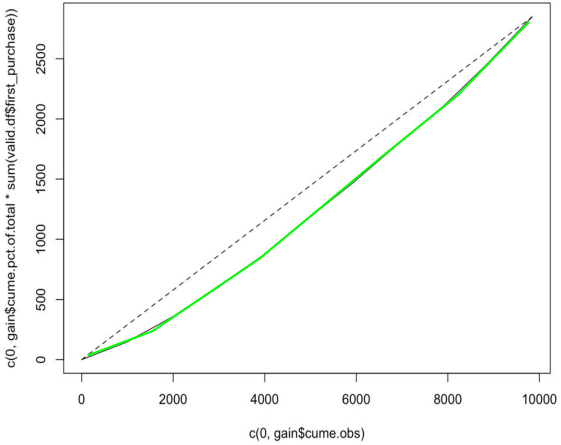


Method : Logistic Regression

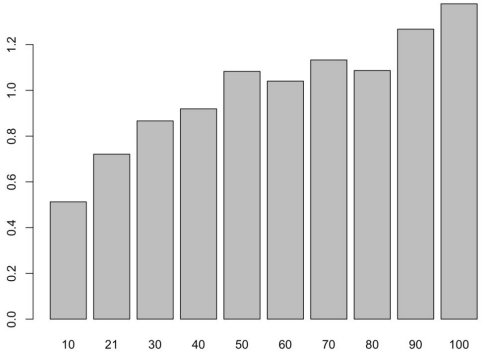
oversampling

Variables selection—Stepwise

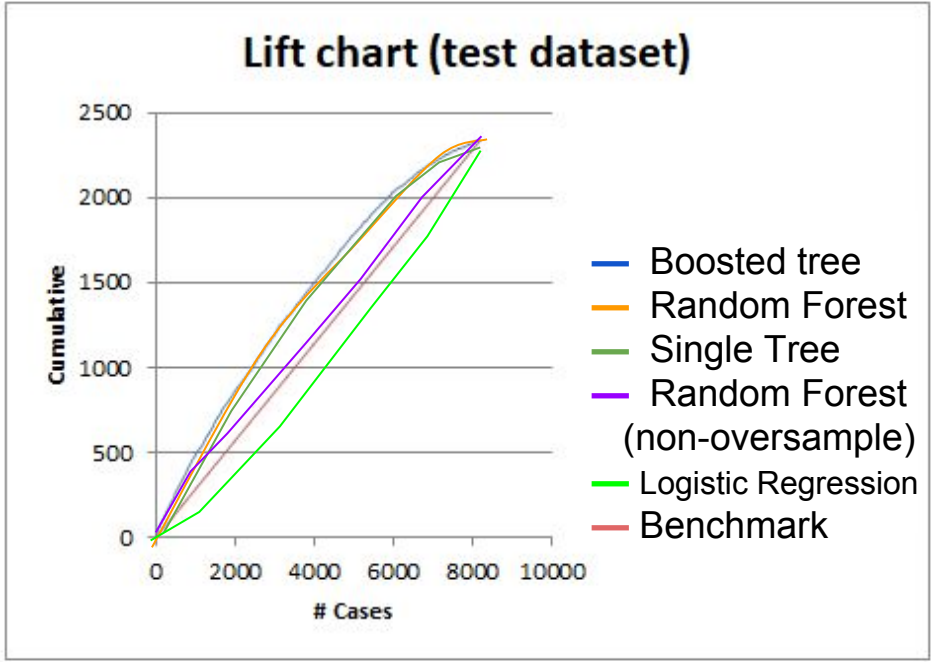
Num_events
Create_collage_empty
Num_background_try
Num_frame_try
Avg_of_image_export
Avg_photo_facebook
remix_cat_Back_to_School
remix_cat_Congrats
remix_cat_Just_for_Fun
remix_cat_Labor_Day_Weekend
font_Roboto_BlackItalic
Create_collage_grid
Login

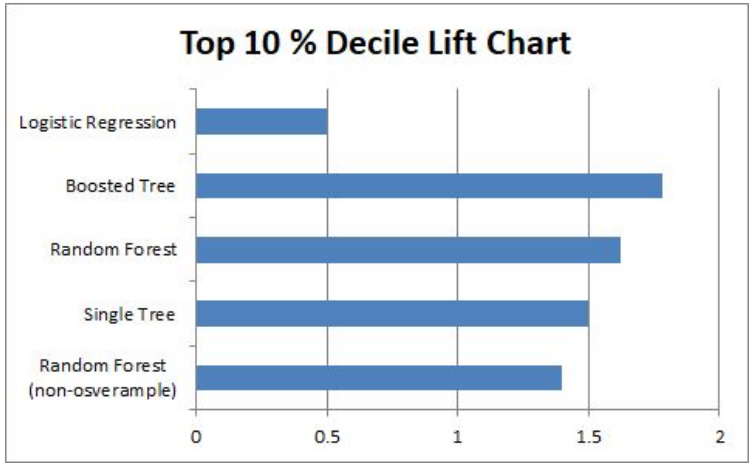


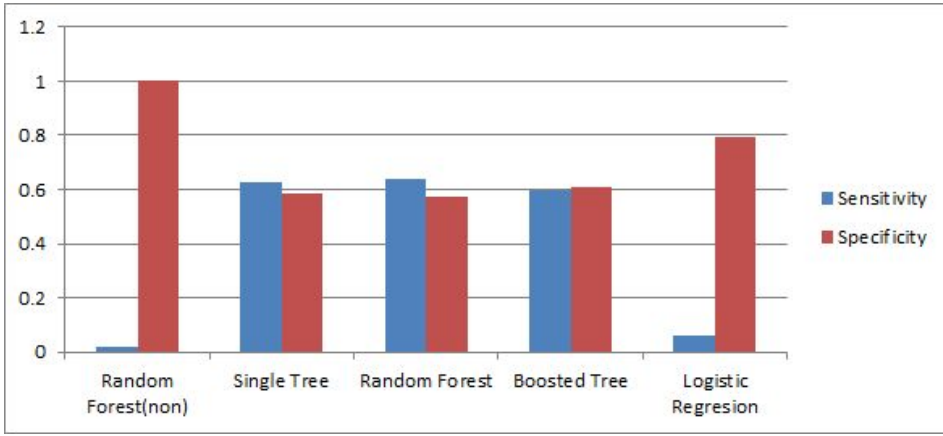

```
Confusion Matrix and Statistics

          Reference
Prediction    1    0
         1  104  635
         0 2745 6359

               Accuracy : 0.6566
              Sensitivity : 0.03650
              Specificity : 0.90921
```



Performance Evaluation







- Boosted Tree and Random Forest are top two best model.

Recommendations

- How to use this model for marketing promotion?

  Offering bundles/discount to users that have a high probability of making a first purchase.

- Model recommendation
  - Due to the unbalanced dataset and ranking goal, we suggest to adopt over-sampling
- Date recommendation
  - The data we are using now is missing the October purchase.
  - Collect events data per user for their 30 days full history.
- Variables recommendation
  - Getting user information might help to predict first purchase earlier.