Report on Company Y's Research Challenge

From: Eugene Yankovsky
To: Company Y's
Date: July 17th, 2017

A marketing manager has proposed a new television campaign to drive awareness of "The All New Company Y," with the enhancements intended to make it easier to upload, view and organize your photos and videos as well as make photo books, cards and photo gifts.

1. What would you choose as the primary success metric of the campaign and why?

2. What would you choose as the secondary metric(s) of the campaign and why?

3. Design an experiment to test whether the television campaign has an effect on the primary success metric. Describe the design of your experiment, and please be explicit about assumptions.

Experiment and metrics design

The choice of primary and secondary success metrics depends on selection of the experiment approaches.
Two alternative approaches are considered:
1) classic survey to estimate awareness rates,
2) **advertising attribution approach using number of the customer visits to the web-site**. This web-site traffic can be used to estimate awareness since there should be relationship between general awareness of the enhancement and its' realization in a number of customers visiting the web-site.

Below I will describe two approaches while giving my preference to web-site customer traffic as most feasible option.

Classic survey approach: an ideal that is hard and expensive to achieve

+ directly addressing awareness question: "Are you aware of the web-site enhancements?"
by estimating ratio of survey participants who positively responded on this question to the total number of the survey participants
- takes more time and resources on preparing and conducting the survey that requires
    a) survey sample should reflect the target audience distribution;
    b) participants of sufficient size;
    c) the participants should not have prior knowledge of the enhancement;
    d) the participants should be randomly selected (not volunteers) for a survey results to be generalizable to the target audience.
- more expensive because it is most often involved in significant expenditures related to
    a) paying survey participants
    b) overhead expenses on survey's staff
    c) paying other marketing companies who does surveys for the company.

The approach can be used to test if the ratio during the campaign is different from ratio prior to the marketing campaign by using multinomial distribution goodness of fit test[1]:

Ho: The target audience (population) follows a binomial distribution with specified probabilities for each of 2 categories (yes, no awareness)
vs
Ha: The target audience (population) does not follows a binomial distribution with specified probabilities for each of 2 categories

The probabilities above are estimated by the ratios before the campaign and compared to the ratios during campaign.

[1]See details on multinomial goodness of fit test in Anderson, Sweeney, Williams “Statistics for business and economics”, 9th edition, p.459-461.

Advertising attribution approach (a viable proxy in dynamically changing environment)

+ A number of customers' visits to the web-site per hour or day seemed to be a reasonable proxy for the awareness by making a reasonable assumption that there should be relationship (some ratio) between awareness and its' realization in a web-site visit[2].
+ The web-site traffic per unit of time (hour or day) as an estimate of the awareness realized into into a web-site visit reflects the on of key goals of any marketing campaign.
+ TV marketing approach that focused on answering customer acquisition question directly by comparing the number of the web-site visits before the TV campaign and during the campaign[3];
+ Faster and more frequent to get the internal data on the customer traffic through the web-site
+ Significantly cheaper since there is no need in doing surveys involved paying

In the settings of more feasible attribution approach, the design of experiment will be as follows:

2.1. Primary success metric

The web-site customer traffic estimated by a number of distinct customers per unit of time (hour or day)

2.2. Secondary success metrics

The following **6** secondary metrics will be estimated by the numbers of distinct customers using the corresponding web-site capabilities per unit of time (hour, day or week):

1) uploading their photos and videos;
2) viewing their photos and videos;
3) organizing their photos and videos;
4) making photo books;
5) making cards;
6) making photo gifts.

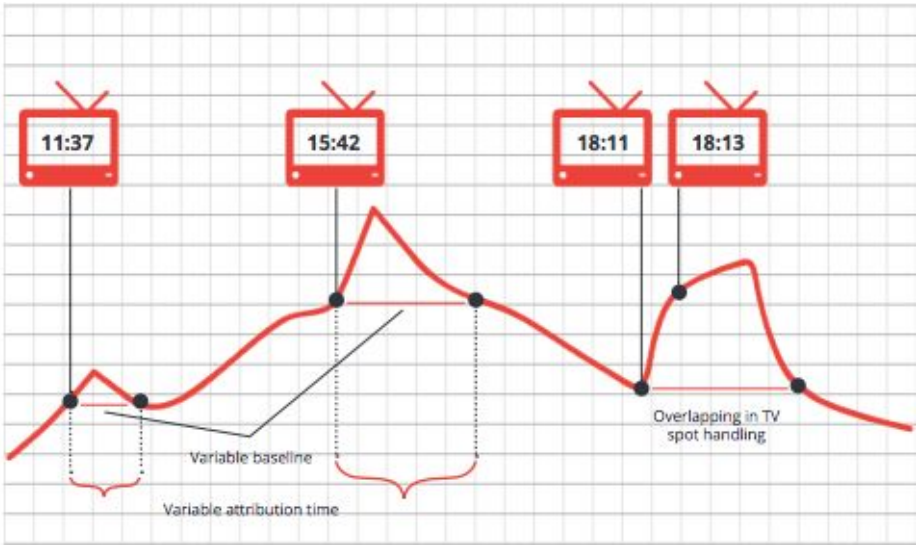
[2] In social sciences, proxy measurements are often required to stand in for variables that cannot be directly measured. This process of standing in is also known as operationalization. E.g., per-capita GDP is often used as a proxy for measures of standard of living or quality of life. (https://en.wikipedia.org/wiki/Proxy_(statistics))

[3] "TV Attribution: The 5 things you need to know", http://wywy.com/market-view/tv-attribution-5-things-you-need-to-know/

2.3. Design of experiment

1. The number of the distinct customers is collected for the metrics of interest
   a) Before campaign web-site traffic data (Y1, …, YT) which start from a baseline set by the previous ad campaign and end before the new campaign started;
   b) Campaign web-site traffic data (YT+1,..., YT+k) which start from the moment the first ad of the campaign aired up to about ½ of the attribution window. The attribution window can be estimated by average attribution window from the prior ad campaigns.

Chart. Expected pattern in the web-site traffic in response to the TV advertising campaign[4]



2. Apply Chow F-test[5] (if no evidence of different variances) or Wald test (if there is evidence of significant difference in variances) for a structural break web-site traffic volume before and after the campaign.

The core ideas of the test evaluate the difference between the squared residuals calculated from the alternative hypotheses:

Ho: The coefficient vectors (ß) of the time series models are the same

$$\begin{bmatrix} Y_{t1} \\ Y_{t2} \end{bmatrix} = \begin{bmatrix} X_{t1} \\ X_{t2} \end{bmatrix} \beta + \begin{bmatrix} e_{t1} \\ e_{t2} \end{bmatrix}$$

vs

Ha: The coefficient vectors (ß1, ß2) of the time series model are the different

$$\begin{bmatrix} Y_{t1} \\ Y_{t2} \end{bmatrix} = \begin{bmatrix} X_{t1} & 0 \\ 0 & X_{t2} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} e_{t1} \\ e_{t2} \end{bmatrix}$$

where

t1 = 1, …, T (time from a baseline to the start of the ad campaign) ;

t2 = T+1, …, T+k (time from the beginning of the campaign to ½ of the attribution window;

---

[4] "Measuring TV's Impact for Mobile Advertisers" http://wywy.com/wp-content/uploads/2015/10/Whitepaper-SM-wywy.pdf

[5] More details on Chow F-test and Wald structural tests can be found on p.130-141 of at W. Greene "Econometric Analysis", 5th edition or at https://en.wikipedia.org/wiki/Chow_test

X_{t1}, X_{t2} are predictors' metrics that may include autoregressive (AR) and moving average (MA) terms. The design settings above may result in small samples to work with. In this case, the Bayesian approach using previous campaign results for prior distributions can be a solution to the small sample problem[6].

Task 3. Data analysis

3.1. Details of exploratory data analysis are provided in Appendix 3.1.

The exploratory data analysis (see details in Appendix 3.1) resulted in the following observations and data transformations.

1. Premium Content is reported with 0 units and non-zero revenues. It is assumed to be a unitless service that is entering prediction model with Premium Content revenue as a predictor.

2. Revenues<0 are assumed to be returns associated with the customers' purchases in the prior periods. They are dropped off from the sample. To note, this change resulted in a loss of only 2 observations in a final sample.

3. After data transformations, the final sample of 22,034 customers is almost balanced with
10710 (48.6%) customers retained;
11324 (51.4%) customers churned.

The data set was randomly split in training (80%) and test (20%) samples:
Data Size : 22034
Training Data Size : 17627
Test Data Size : 4407

4. The final sample included about 260 features:
    1) Features for revenue and units aggregated by category for each customer for the 1st purchase;
    2) Features for revenue and units aggregated by product for each customer for the 1st purchase;
    3) Binary dummy variables associated with timing of the first purchase: US official holidays, and weekend.

To note, month variable was excluded from consideration because 100% of the first purchases were done in January.

5. Dependent variable is
Customer retained = 1 if the maximum sequence order was more than 1, 0 otherwise.

Key Assumption

The setting of the dependent variable implies that the retention model is focused only on factors affecting the customer's decision to buy the second Company Y's service after their first purchase.

[6] An example of application Bayesian approach to evaluating marketing campaigns can be found in https://www.datascience.com/blog/introduction-to-bayesian-inference-learn-data-science-tutorials

3.2. Two-stage predictive modelling approach

Stage 1. Apply random forest (100 trees, up to 5 levels of tree depth) including all features as candidate predictors

Stage 2. Develop several competing models that use top predictors with the importance at least 0.001 as a candidate predictors:

2.1. Random Forest Model (200 trees)

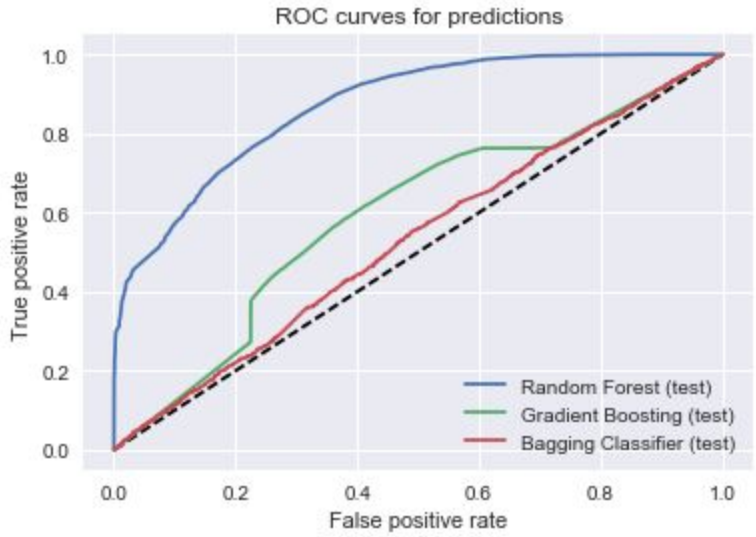
2.2. Gradient Boosting Classifier (200 trees)

2.3. Bagging Classifier (200 trees)

To note, top 100 predictors were selected to minimize potential data over-fitting by the model.

Champion model and goodness-of-fit

Three models were considered: Random Forest, Gradient Boosting Classifier, and Bagging Classifier. See details of their estimation output in Appendix 3.2.The champion model is the Random forest model with the highest Area Under the ROC (AUC) of 0.543.



However, the predictive power of the champion model in a test sample is mediocre:

Accuracy Score = 0.525

Recall Score = 0.482

Precision Score = 0.504.

Potential model alternatives

Potential improvements in the data goodness-of-fit might be achieved by:

1) obtain more data on the customers (gender, age, place of living, and etc.);
2) use sample of bigger size;
3) estimate neutral network model with top 20 predictors discovered by the Random Forest model.

Due to time constraints, the following feature engineering options that might improve the data goodness-of-fit were not tried:

1) Using dummy variables taking 1 if the customer bought the category 0 otherwise;
2) Using dummy variables taking 1 if the customer bought the product 0 otherwise;
3) Using dummy variables taking 1 if the customer falls in certain quartiles by revenue the customer generated.

3.3. Interpretation of modelling results

Although the model has mediocre goodness-of-fit indicators, it found evidence between the customer retention and the following top 20 predictors for marketing managers to focus on.

Table 1. Top 20 predictors ranked by the importance

Rank	Feature	Importance
1	total_revenue	0.239
2	total_units	0.102
3	('revenue', 'Photo Books')_category	0.096
4	('revenue', 'Prints')_category	0.079
5	('revenue', '4x6')_product	0.069
6	('units', 'Prints')_category	0.059
7	('units', '4x6')_product	0.054
8	('revenue', '8x11 Classic Book')_product	0.052
9	('revenue', 'Gifts')_category	0.04
10	('revenue', '5x7')_product	0.039
11	('revenue', 'Premium Content')_product	0.03
12	('revenue', 'Calendars')_category	0.027
13	('revenue', '8x10')_product	0.026
14	('units', '5x7')_product	0.023
15	('revenue', 'Home Decor')_category	0.021
16	('revenue', 'Wall Calendars')_product	0.016
17	('revenue', 'Magnets')_product	0.014
18	('units', 'Calendars')_category	0.007
19	('units', 'Wall Calendars')_product	0.005
20	('units', '4x4')_product	0.003

2. Primary effect signs estimation by logistic regression

The ensemble tree models tend to give better data goodness-of-fit by capturing often complex relationship (including primary and multi level interaction effects) between dependent variable and its' predictors. However, It is extremely challenging to deduce a direction (i.e., sign) of the effects from the random forest model where the same predictor may be used in multiple splitting rules in many trees. To address potential business question of the signs of the effect, I ran logistic regression.

The logistic regression detected evidence of probability of retaining a customer is

- positively affected by amounts spent on "Photo Books" category, total units bought and units of "Wall Calendar" product and, amounts spent on "Magnets" products and

- negatively affected by amount spent in total as well as amounts spent on "Wall Calendar" product and "8x10" product.

To note, the champion predictive Random Forest model suggest complex interaction between more than 20 factors that affect customer's' decision to buy Y products again.

A few examples of the interpretations are:

1) More the customer spends, the less probability of his/her retention. I.e, the estimated odds of retaining a customer who spends $100 is 0.78 times smaller than the odds of retaining a customer who spends only $10.

2) More units customer buys, the higher probability of his retention. I.e., the estimated odds of retaining a customer who buys 100 units is 1.21 times higher than the odds of retaining a customer who buys only 10 units.

3) The "Wall Calendar" product seemed to be very popular and successive in retaining customers, although some of them believe it is overpriced. I.e., the estimated odds of retaining a customer who buys 10 units is 1.10 times higher than the odds of retaining a customer who buys only 1 unit.

A complete list of significant primary effects provided in Table 3.1 below, while interaction effects are reported by the Random Forest model.

Table 3.1. Odds calculation from the logistic regression output (statistically significant only)

Predictor	Coefficient	Estimated Odds ratio for increase from 1 to 10	Estimated Odds ratio for increase from 10 to 100
Total revenue	-0.0027***	0.997	0.78
Revenue from “Photo Books” category	0.0029**	1.003	1.30
Revenue from “Wall Calendar” product	-0.0162**	0.984	0.23
Total units	0.0021*	1.002	1.21
Units of “Wall Calendar” product	0.3373**	1.401	Excessively extreme value is resulted from extreme unit value
Revenue from “8x10” product	-0.023**	0.977	0.13
Revenue from “Magnets” product	0.0145**	1.015	3.69

Note:
***/**/* denotes that statistical significance of the predictor’s effect with p-value at 1%, 5%, 10%

Appendix 3.1. Details of Exploratory data analysis

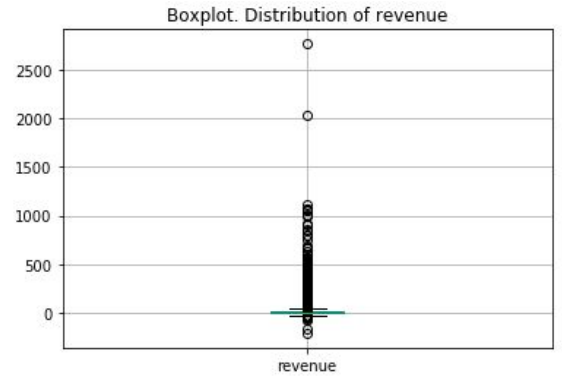
Before deleting negative revenue values

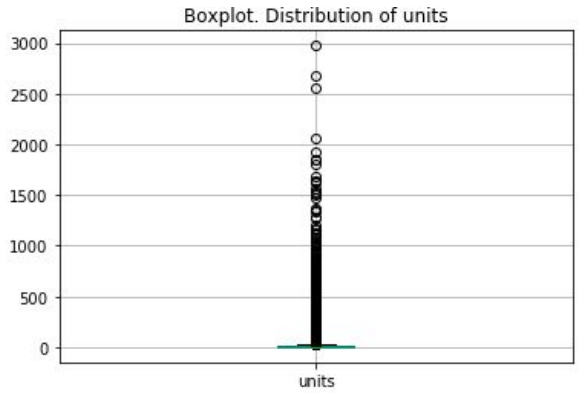
	order_sequence	revenue	units
count	89636.000	89636.000	89636.000
mean	4.456	19.524	22.710
std	6.332	35.721	62.664
min	1.000	-206.150	0.000
25%	1.000	3.936	1.000
50%	2.000	7.990	1.000
75%	5.000	23.380	13.000
max	93.000	2761.990	2974.000

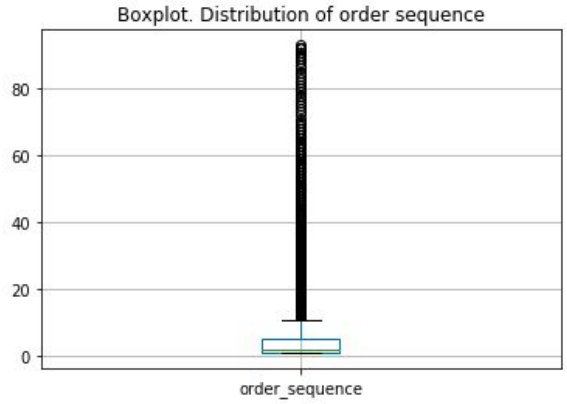
After deleting negative revenue values

	order_sequence	revenue	units
count	31367.000	31367.000	31367.000
mean	1.000	22.290	24.356
std	0.000	39.996	66.942
min	1.000	0.000	0.000
25%	1.000	4.459	1.000
50%	1.000	8.490	1.000
75%	1.000	27.990	19.000
max	1.000	2761.990	2676.000

Summary statistics on continuous variables







Selection of lowest revenue observations

	order_date	category_name	product_name	order_sequence	revenue	units	\
1046	2015-12-26	Other	Unassigned	5	-58.470	0	
3569	2014-11-19	Other	Unassigned	2	-206.150	0	
19292	2015-07-23	Other	Unassigned	26	-68.100	0	
70234	2015-04-11	Other	Unassigned	5	-164.720	0	

	customer_id	order_id
1046	363293	363293
3569	82783278	82783278
19292	48865200	48865200
70234	24783955	24783955

Category names

15

['Gifts' 'Large Formats' 'Photo Books' 'Prints' 'Stationery' 'Cards' 'Calendars' 'Unassigned' 'Home Decor' 'Card Upsell' 'Other' 'Services' 'Shipping' 'Yearbooks' 'OTHER']

Product name

165

['Archive DVDs' '11x14' '12x12 Memory Book' 'Memorabilia Pocket' 'Premium Content' '8x11 Classic Book' '4x6' '5x7' '8x10' 'Address Labels' 'iPhone Case' '8x8 Story Book' '11x14 Collage' 'Wallet' '5x7s' 'Wall Calendars' 'Unassigned' 'Desk Art' 'Fleece Photo Blanket (60x80)' '16x20' 'Magnets' 'SS Travel Mug' 'Greeting Cards' 'Ornaments' 'Liners' '5x7 Flat Card (Premium)' '5x7 Flat Foil Card' '20x20' 'Card Trims' '5x7 Casual Book' 'Key Ring' '4x8s' 'Mugs' 'Canvas Prints' 'Acrylic Prints' 'Photo Cubes' '10x10 Photo Book' 'Pillow' '4x4' '4x8 Stationery Card' 'Pearl Paper' 'Large Calendars' 'Mix and Match' 'Tablet Cover' '3x5 Stationery Card' 'Mousepads' 'Gift Box' 'Trifold Greeting Cards' '6x8 Stationery Card' 'Stationery - Notepads' 'Videograms' 'Desk Calendars' '4X4' 'Reusable Shopping Bag' 'Cube Ornaments' 'Luggage Tag' '5x5 Stationery Card' 'Water Bottle' 'Pet Tag' 'Gift Box PB' '20x30' 'Playing Cards' '6x8 Flat Foil Card' 'Wood Art' 'Fleece Photo Blanket (50x60)' 'Art Prints' '4.25x5.5 Stationery Cards' '20x30 Collage' 'Coasters' 'Serving Tray' 'Plates' 'Envelopes' 'Candle' '11x14 Photo Book' '8x10 Collage' 'Electronic Accessories' 'Pre-Add Envelopes From' 'Cups' 'Metal Prints' 'Framed Canvas' 'Glass Plate' 'Rubber Stamps' 'Ghiradelli Mug' 'Puzzles' 'Shipping' '12x18' 'Shams and Pillowcases' 'Hallmark Inserts' 'Glass Prints' 'Placemat' 'Desk Set' '8x8 Square Print' 'Acrylic Photo Blocks - 5X7' '12x12 Square Print' 'Stickers' 'Charm' '7x9 Photo Book' 'Notebook' 'Wall Art' 'Gift Wrap' 'Ornament Cards' '16x20 Collage' 'Calling Cards' 'Framed Art Prints' 'Ceramic Tiles' 'Collage Frame' 'Travel Mug (16oz)' 'Pre-Add Envelopes From & To' 'Photo Stockings' 'Decorative Wall Decals' 'Duvets' 'Table Runners' 'Smartphone Case' 'Lunch Bags' 'Adventure Books' 'Bracelet' 'Travel Mug (20oz)' 'Keepsake Box' 'Notepad' 'Wedding Invites' 'Dry Erase Decals' 'Easel' 'Framed Mounted Wall Art' 'Stamps' 'Necklace' 'Woven Photo Blanket (54x70)' 'Stein' 'Denim Wallet' 'Personalized Desk Frame' 'Outdoor Pillows' 'Photo Quilt' 'Growth Chart' 'Woven Photo Blanket (60x80)' 'Acrylic Photo Blocks - 5X5'
'Three Quarter Fold Card' 'Gift Tags' 'Dimensional Wall Art' 'Pet Placemat' 'Calendar Gift Box' 'iPad Sleeve' '5x7 Pearl' 'Mason Jars' 'Framed Prints' 'PhotoShow DVDs' 'Totebags' 'Wall Decals' 'Make My Book' '12X12' '11x14 Pearl' '8x32 Panoramic' '8X8' 'Stick It Notes' 'Tshirts' 'Dog Tag' 'Stationery Magnets' '3x5 Response Cards' 'Pet Bowl' '8x10 Pearl' '8x8 Pearl' 'Aprons' '16x20 Pearl' 'Wine Glasses'
'Pre-Add Envelopes To' 'OTHER' '11x14 MIAM']

Appendix 3.2. **Details on estimation results by using two-stage modelling approach**
Stage 1: Apply random forest including all predictors (no_of_estimators = 100, tree_depth = 5)
Top variables by their importance

Rank	Feature	Importance
1	total_revenue	0.045
2	('revenue','Calendars')_category	0.045
3	('units','Calendars')_category	0.045
4	('revenue','PhotoBooks')_category	0.043
5	('revenue','WallCalendars')_product	0.042
6	total_units	0.036
7	('units','Prints')_category	0.029
8	('units','WallCalendars')_product	0.029
9	('revenue','Prints')_category	0.027
10	('revenue','4x6')_product	0.026
11	('units','4x6')_product	0.02
12	('revenue','8x11ClassicBook')_product	0.019
13	('revenue','PremiumContent')_product	0.019
14	('revenue','8x10')_product	0.018
15	('revenue','HomeDecor')_category	0.017
16	('revenue','5x7')_product	0.016
17	('units','4x4')_product	0.016
18	('revenue','Magnets')_product	0.016
19	('units','5x7')_product	0.015
20	('revenue','Gifts')_category	0.015
21	('units','8x10')_product	0.014
22	('units','PhotoBooks')_category	0.014

23	('revenue','4x4')_product	0.013
24	('revenue','MemorabiliaPocket')_product	0.013
25	('units','Gifts')_category	0.013
26	weekend	0.012
27	('revenue','Wallet')_product	0.012
28	('revenue','Cards')_category	0.012
29	('revenue','8x8StoryBook')_product	0.011
30	('revenue','LargeFormats')_category	0.011
31	('units','Magnets')_product	0.01
32	('units','6x8StationeryCard')_product	0.009
33	('units','Cards')_category	0.008
34	('units','5x7FlatCard(Premium)')_product	0.008
35	('revenue','DeskCalendars')_product	0.008
36	('units','8x8StoryBook')_product	0.008
37	('revenue','12x12MemoryBook')_product	0.008
38	('revenue','CardUpsell')_category	0.007
39	('revenue','Stationery')_category	0.007
40	('units','8x11ClassicBook')_product	0.007
41	('revenue','3x5StationeryCard')_product	0.007
42	('units','Stationery-Notepads')_product	0.007
43	('units','LargeCalendars')_product	0.006
44	('revenue','PearlPaper')_product	0.006
45	('revenue','LargeCalendars')_product	0.006
46	('revenue','CardTrims')_product	0.006
47	('units','11x14')_product	0.006

48	('units','7x9PhotoBook')_product	0.006
49	('revenue','7x9PhotoBook')_product	0.006
50	('revenue','PhotoCubes')_product	0.006
51	('units','3x5StationeryCard')_product	0.005
52	('units','Wallet')_product	0.005
53	('revenue','5x7FlatCard(Premium)')_product	0.005
54	('revenue','5x7s')_product	0.005
55	('revenue','Stationery-Notepads')_product	0.005
56	('revenue','SmartphoneCase')_product	0.005
57	('revenue','6x8StationeryCard')_product	0.005
58	('revenue','Mousepads')_product	0.004
59	('revenue','Shipping')_category	0.004
60	('revenue','8x10Collage')_product	0.004
61	('revenue','11x14')_product	0.004
62	('revenue','AddressLabels')_product	0.004
63	('revenue','10x10PhotoBook')_product	0.004
64	('revenue','Mugs')_product	0.004
65	('units','AddressLabels')_product	0.004
66	('revenue','11x14PhotoBook')_product	0.003
67	('revenue','16x20')_product	0.003
68	('units','4.25x5.5StationeryCards')_product	0.003
69	('units','Mugs')_product	0.003
70	holiday	0.003
71	('units','GreetingCards')_product	0.003
72	('revenue','KeepsakeBox')_product	0.003

73	('units','KeepsakeBox')_product	0.003
74	('units','LargeFormats')_category	0.003
75	('revenue','20x30')_product	0.003
76	('revenue','GreetingCards')_product	0.003
77	('units','20x30')_product	0.003
78	('units','5x7CasualBook')_product	0.003
79	('revenue','Puzzles')_product	0.003
80	('revenue','iPhoneCase')_product	0.003
81	('revenue','5x5StationeryCard')_product	0.003
82	('revenue','4x8StationeryCard')_product	0.003
83	('revenue','WaterBottle')_product	0.003
84	('units','GhiradelliMug')_product	0.003
85	('units','HomeDecor')_category	0.003
86	('units','HallmarkInserts')_product	0.002
87	('units','DeskCalendars')_product	0.002
88	('revenue','4x8s')_product	0.002
89	('units','12x12MemoryBook')_product	0.002
90	('revenue','20x30Collage')_product	0.002
91	('revenue','KeyRing')_product	0.002
92	('revenue','DeskArt')_product	0.002
93	('units','4x8s')_product	0.002
94	('units','16x20')_product	0.002
95	('units','11x14PhotoBook')_product	0.002
96	('units','SmartphoneCase')_product	0.002
97	('units','8x10Collage')_product	0.002

98	('revenue','Pillow')_product	0.002
99	('units','CanvasPrints')_product	0.002
100	('units','TrifoldGreetingCards')_product	0.002

RMSE : 0.680200720519

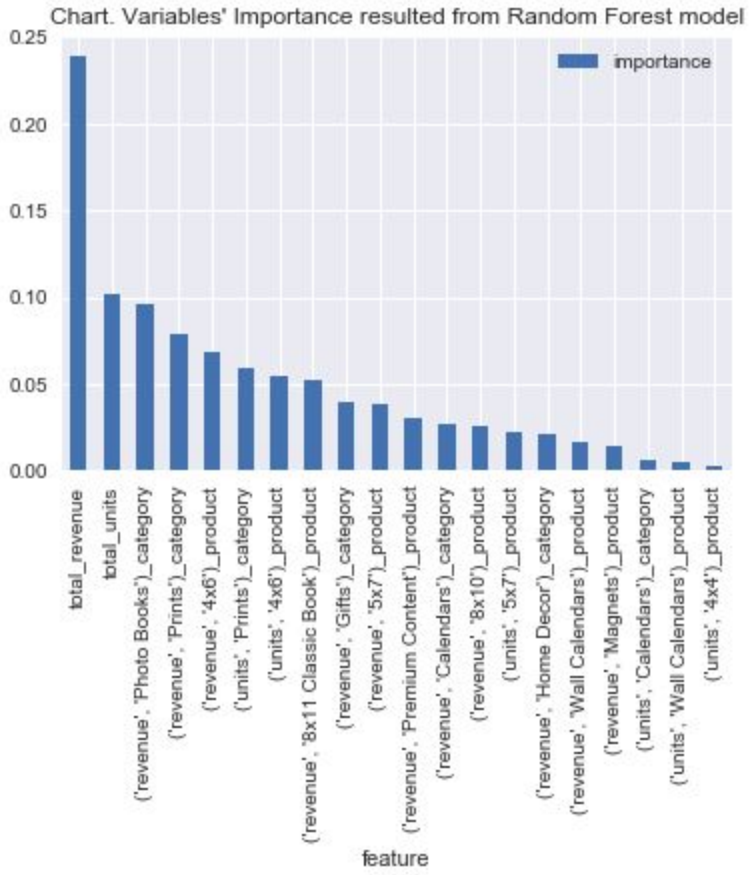
Confusion Matrix :
[[True_positive, False_negative],
[False_positive, True_negative]]

[[1920 378]
[1661 448]]

Accuracy Score : 0.537326979805
Recall Score : 0.212422949265
Precision Score : 0.542372881356

Stage 2: Develop competing models using top 20 importance predictors from stage 1

Random forest (no_of_estimators = 200, tree_depth = 30)



Rank	Feature	Importance
1	total_revenue	0.239
2	total_units	0.102
3	('revenue', 'Photo Books')_category	0.096
4	('revenue', 'Prints')_category	0.079
5	('revenue', '4x6')_product	0.069
6	('units', 'Prints')_category	0.059
7	('units', '4x6')_product	0.054
8	('revenue', '8x11 Classic Book')_product	0.052

9	('revenue', 'Gifts')_category	0.04
10	('revenue', '5x7')_product	0.039
11	('revenue', 'Premium Content')_product	0.03
12	('revenue', 'Calendars')_category	0.027
13	('revenue', '8x10')_product	0.026
14	('units', '5x7')_product	0.023
15	('revenue', 'Home Decor')_category	0.021
16	('revenue', 'Wall Calendars')_product	0.016
17	('revenue', 'Magnets')_product	0.014
18	('units', 'Calendars')_category	0.007
19	('units', 'Wall Calendars')_product	0.005
20	('units', '4x4')_product	0.003

RMSE : 0.678865028076

Confusion Matrix :
[[True_positive, False_negative],
[False_positive, True_negative]]

[[1449 849]
[1182 927]]

Accuracy Score : 0.539142273656
Recall Score : 0.439544807966
Precision Score : 0.521959459459

Gradient Boosting Classifier
'max_depth': 30, 'n_estimators': 200, 'subsample': 0.5, 'random_state': 2,
learning_rate =0.5

RMSE : 0.695050412571

Confusion Matrix :

[[True_positive, False_negative],
  [False_positive, True_negative]]

[[1239 1059]
 [1070 1039]]

Accuracy Score : 0.516904923985

Recall Score : 0.492650545282

Precision Score : 0.495233555767

Bagging Classifier
'max_depth': 30, 'n_estimators': 200
RMSE : 0.689313546522

Confusion Matrix :

[[True_positive, False_negative],
  [False_positive, True_negative]]

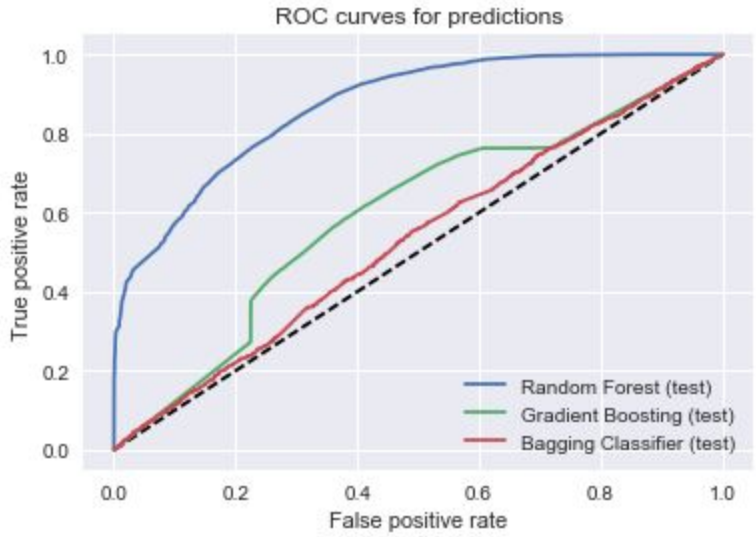
[[1296 1002]
 [1092 1017]]

Accuracy Score : 0.524846834581

Recall Score : 0.482219061166

Precision Score : 0.503714710253

Selecting a champion model by highest AUC



AUC for Random Forest 0.543243428945
AUC for Gradient Boosting 0.505665656037
AUC for Bagging 0.529850621544

Estimation results of the logistic model on main effects as predictors

```
Optimization terminated successfully.
         Current function value: 0.691493
         Iterations 6
                          Logit Regression Results
==============================================================================
Dep. Variable:      customer_retained   No. Observations:                17627
Model:                          Logit   Df Residuals:                    17607
Method:                           MLE   Df Model:                           19
Date:                Sun, 16 Jul 2017   Pseudo R-squ.:               0.001968
Time:                        00:07:10   Log-Likelihood:                -12189.
converged:                       True   LL-Null:                       -12213.
                                        LLR p-value:                 0.0002514
================================================================================================================
                                                 coef    std err          z      P>|z|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------------------------------------
total_revenue                                 -0.0027      0.001     -2.497      0.013        -0.005     -0.001
('revenue', 'Calendars')_category              0.0064      0.006      1.062      0.288        -0.005      0.018
('units', 'Calendars')_category               -0.0412      0.120     -0.343      0.732        -0.277      0.194
('revenue', 'Photo Books')_category            0.0029      0.001      2.285      0.022         0.000      0.005
('revenue', 'Wall Calendars')_product         -0.0162      0.007     -2.229      0.026        -0.031     -0.002
total_units                                    0.0021      0.001      1.637      0.102        -0.000      0.005
('units', 'Prints')_category                  -0.0061      0.021     -0.290      0.772        -0.047      0.035
('units', 'Wall Calendars')_product            0.3373      0.140      2.407      0.016         0.063      0.612
('revenue', 'Prints')_category                 0.0173      0.013      1.384      0.166        -0.007      0.042
('revenue', '4x6')_product                    -0.0147      0.013     -1.090      0.276        -0.041      0.012
('units', '4x6')_product                       0.0041      0.021      0.194      0.846        -0.037      0.045
('revenue', '8x11 Classic Book')_product      -0.0004      0.001     -0.398      0.691        -0.002      0.001
('revenue', 'Premium Content')_product         0.0082      0.011      0.778      0.437        -0.012      0.029
('revenue', '8x10')_product                   -0.0230      0.010     -2.365      0.018        -0.042     -0.004
('revenue', 'Home Decor')_category             0.0011      0.001      0.728      0.467        -0.002      0.004
('revenue', '5x7')_product                     0.0039      0.023      0.171      0.864        -0.040      0.048
('units', '4x4')_product                       0.0122      0.020      0.611      0.541        -0.027      0.051
('revenue', 'Magnets')_product                 0.0145      0.007      2.041      0.041         0.001      0.028
('units', '5x7')_product                      -0.0064      0.028     -0.231      0.818        -0.060      0.048
('revenue', 'Gifts')_category                  0.0005      0.002      0.246      0.806        -0.004      0.005
================================================================================================================
```