Customer churn prediction

At home fitness brand

Developed a predictive model to analyze historical customer characteristics & behavior to identify ‘at risk’ customers, enabling the firm to target them through effective incentives and campaigns

Customer churn prediction model for a fitness brand

Situation

- An at-home fitness brand faced an increase in the customer attrition rate due to cancellations of subscriptions
- Partnered with the company to identify customers who are likely to cancel their subscription in near future so that they can proactively target them with appropriate preventive actions

Accordion Value Add

- Built a predictive model that estimates a customer's propensity to churn in the near future based on historical characteristics and behavior
- The model also helped in identifying factors that drive 'churn' (cancellation of subscription)
- Used advanced classification algorithms such as Random Forest, Adaboost and Logistic Regression to classify customer subscriptions
- Used SMOTE sampling technique as the data set had 'churns' as a very small minority class and the cost of misclassification of 'churns' was high

Impact

- The churn prediction model helped the client to identify top-100 'at risk' customers with 66% precision
- The predictive model enabled the company to "treat" these customers through suitable incentives and campaigns to prevent them from cancelling their subscriptions

More than 600 iterations were performed based on combinations of variables, data sets and algorithms

Independent Variables

a. Iteration 1: 16 independent variables
b. Iteration 2: All independent variables except **A and B**
c. Iteration 3: All independent variables except **A, B and C**
d. Iteration 4: All independent variables except **A, B, C and D**
e. Iteration 5: All independent variables in iteration 1 and **7 additional variables**

Dependent Variables

a. # of workouts in "N+1", "N+2" and "N+3" month
b. Workout duration in in "N+1", "N+2" and "N+3" month
c. # days worked out in in "N+1", "N+2" and "N+3" month
d. Classification of usage behavior (method 1)
e. Classification of usage behavior (method 2)
f. Classification of usage behavior (method 3)
g. Classification of usage behavior (method 4)

Dataset Used

a. 80% (randomly sampled) of 22 months of account snapshots[1]
b. 3 months of account snapshots
c. 1 month of account snapshot
d. 3 months data sampled using SMOTE
e. 3 months snapshot of only package A subscriptions
f. 3 months snapshot of only package A subscriptions with data sampled using SMOTE

Algorithm Used

a. Logistic Regression
b. Classification using Random Forest (RF) algorithm
c. Classification using Adaboost methodology

05 X 07 X 06 X 03 = 630

Model recall and precision values were evaluated for each iteration

Confusion Matrix definition

	Predicted: No	Predicted: Yes
Actual: No	True Negative	False Positive
Actual: Yes	False Negative	True Positive

Precision = TP / (TP+FP)
Recall = TP / (TP+FN)
False Positive Rate = FP / (FP+TN)

Confusion Matrix at 0.5 cut-off probability[1]

	Predicted: No	Predicted: Yes
Actual: No	505	110
Actual: Yes	48	265

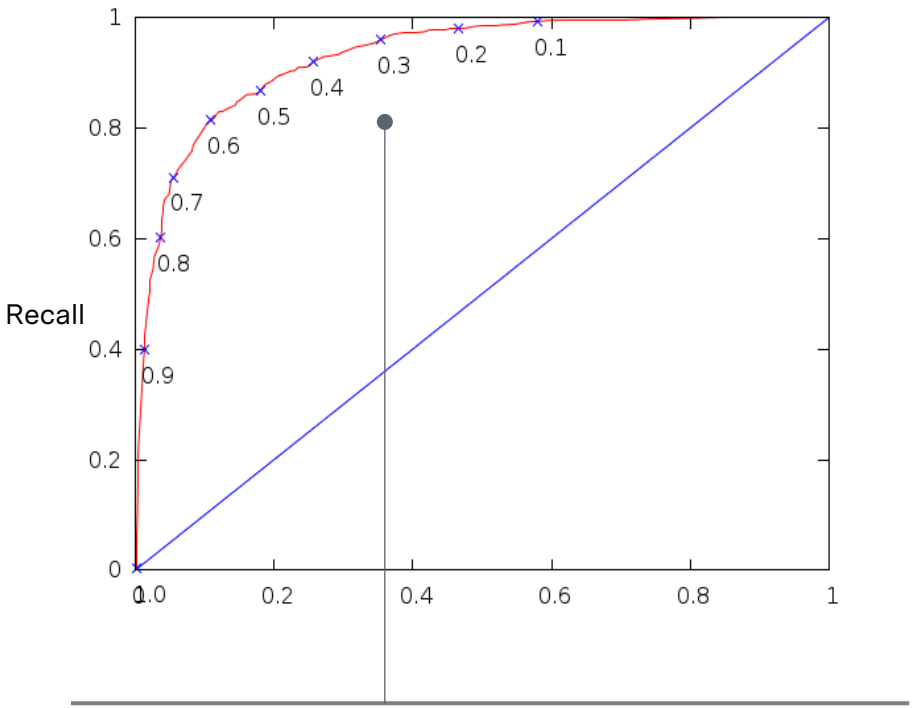
Precision = 265 / (265+110) = 73%
Recall = 265 / (265+48) = 85%
False Positive Rate = 110/(505+110) = 18%

Confusion Matrix at 0.1 cut-off probability

	Predicted: No	Predicted: Yes
Actual: No	261	354
Actual: Yes	4	309

Precision = 309 / (309+354) = 47%
Recall = 309 / (309+4) = 99%
False Positive Rate = 354(261+354) = 58%

ROC (Receiver Operating Characteristic) curve



Reducing cut-off probability leads to increase in 'Recall' at the expense of increase in False Positive rate

Similarly, importance of each variables was evaluated for each iteration

Independent Variable	Mean decrease in accuracy[1]	Mean decrease in GINI[2]	Variable importance in predicting Churn[3]	Variable importance in predicting Non-Churns[4]
A				
B				
C				
D				
E				
F				
G				
H				
I				
J				
K				
L				
M				
N				
O				

Key independent variables identified for each iteration

Rank order of importance of independent variables, by iteration #

Independent Variable	1	2	3	4	5	6	7	8	9	10
A	1	1	2	1	1	4	1	1	8	8
B	2	3	1	2	2	1	2	4	10	2
C	3	2	3	3	5	3	3	9	9	9
D	4	4	4	4	4	2	4	8	4	5
E	5	6	5	5	3	6	13	14	2	13
F	6	5	6	6	6	7	7	13	5	10
G	7	7	7	7	9	5	5	6	7	1
H	8	11	8	9	7	8	6	5	1	6
I	9	14	11	13	14	19	9	12	3	3
J	10	10	9	11	11	10	10	2	6	4
K	11	9	10	12	8	20	8	15	15	15
L	12	13	13	14	12	21	13	11		14
M	13	8	12	13	12	18	13	10	11	12
N	14	12	14	15		11	13	16	12	
O	15		15	8	15	17		7	14	11
P			16		10	13	11	3	13	7
Q						9				
R						12				
S						14				
T						15				
U						16				
V						22				
W						23				