Assignment 2: Tree-Based Methods

BUAD 5082 All Sections – Spring 2018

---

1. Objectives

The purpose of this assignment is to exercise recently developed skills in Tree-Based Methods and to connect these modeling skills to the economic analysis of a common business situation.

2. What You Will Need

- Access to a Windows computer with a recent version of R installed.
- A file named Assignment2TrainingData.csv, downloadable from the Class Schedule page of the course web site.

3. What You Will Hand In

A zip file containing various files, as described in the section below named "Requirements for This Assignment."

4. Due Date

Monday March 19th, just before midnight.

5. Note on Collaboration

This is a Category D team assignment.

6. The case:

Your team has been engaged to develop a Customer Churn model. Your client is a large telecommunications company facing intense competition from industry rivals due to a number of structural factors in the industry. The most important of these factors are:

- A market approaching saturation
- Low switching costs
- Little differentiation in the product and service
- Credible threats to forward integration from suppliers
- Low barriers to entry and high barriers to exit

The client is therefore forced to defend its existing market share aggressively, and has just completed an internal project to design a retention offer for at-risk customers. Some of the findings from this project, which your team must take as given, are as follows:

- The percentage of at-risk customers in the current customer base is 26.448% (this number is supported by the data you will be provided).
- The retention offer will cost the company $1,600 per customer (one-time)
- A lost customer costs the company $11,500 (one time)
- The project has concluded through detailed customer surveys and historical data analysis that the retention offer will be 45% effective. That is, 45% of the customers who have been identified as at-risk and who receive the offer will change their minds and remain with the company. The remaining 55% will leave anyway.

The objective of your engagement is to develop a predictive model to identify the customers who should receive the retention offer, and to evaluate the resulting economic consequences. This model will be used on an ongoing basis to make this decision about future customers. Accordingly, you will be provided with a dataset containing usage information for a random sample of their current and past customers, and an indication of whether or not they have left the company.

Your model must do the following:

- Predict which customers are at-risk given their usage history and should therefore receive the retention offer.
- Identify the associated "expected cost per customer," taking into account the probabilities of both Type I and Type II errors, and the probability that the retention offer will be effective in preventing only 45% of at-risk customer from leaving.
    - Note that this expected cost per customer should be less than
        - The "do-nothing" option, which would cost $0.00 per customer in retention offer cost but incur a lost-customer cost of $3,041.52 (26.448% * $11,500)
        - The "offer-to-all" option, which would cost $1,600 per customer in retention offer costs plus 26.448% * 55% * $11,500 = $1,672.84 per customer in lost-customer costs, totaling $3,272.84 per customer

In order to evaluate your work (and determine if you will get paid), the client will run your model using a portion of the randomly sampled data (this withheld portion is approximately one-sixth the size of the sample provided to your team) and construct a confusion matrix, from which an expected cost will be computed. The details of this analysis are illustrated in

Figure 1 using the results of a very simple model constructed by the client's internal modeling group (needless to say, your team is expected to do better).

Your work will be judged on two criteria:

1. The % difference between the expected cost per customer that you compute using your model (with a specified cutoff) and the expected cost per customer calculated using your model and specified cutoff, but with the withheld portion of the sample
    a. This will measure the variance of your model, so you should take care to avoid over-training).
2. The size of the expected cost per customer calculated by using your model (with a specified cutoff) to compute a confusion matrix from the withheld portion of the sample and thereby compute an expected cost per customer
    a. This will measure your model's ability to produce an effective tradeoff between Type I and Type II errors – you may want to replicate the analysis outlined in Figure 1 in order to evaluate candidate models during your own modeling activities.

7. Requirements for This Assignment:

1. The only models allowed are tree-based models. At a minimum, you must build and evaluate the following:
    a. a decision tree using rpart() or tree()
    b. a random forest using randomForest()
    c. a boosting model using ada() or gbm()
2. For each of these, find the cutoff that minimizes the expected cost per customer as outlined in Figure 1.
3. Submit the following in a zip file named **Assignment2Teams-*nn*.zip** where *s* is your section number and *nn* is your team number (one submission per team please):
    a. One script per team member showing all of that person's modeling work. Each team member should build their "best" versions of each of the three tree-based methods mentioned earlier, where "best" is defined as the model that is expected to minimize the economic cost per customer. These scripts should be named **Assignment2Full*FirstnameLastname*.R**.
    b. A script named **Assignment2Final.R** that does the following:
        i. rm(list = ls())
        ii. "requires" the necessary packages
        iii. Reads in the data
        iv. Performs whatever data management steps you require
        v. Executes your team's "best-of-the-best" model and produces a confusion matrix using your optimal cutoff
        vi. From the confusion matrix, creates a csv file named **Assignment2Results.csv** structured as indicated in Figure 2.
    c. Your economic analysis to compute your model's expected cost per customer
        i. This can be done in R or in an Excel workbook.
            1. If done in R, it should be done in Assignment2Final.R as the last step following the creation of the .csv file. It should be based on the confusion matrix produced by your final model's code and it must perform the calculations done in Figure 1 in a way that's easy to follow (chart optional)

2. If done in an Excel workbook, name the Workbook **Assignment2.xlsx** and include it in your submission. It should look a lot like Figure 1 (chart optional), where the yellow cells are from your Assignment2Results.csv file.

Percent of at-risk customers	26.448%
Retention Offer Effectiveness	45%
Cost of making one retention offer	$1,600.00
Cost of losing one customer	$11,500.00

Minimum Cost	$2,644.36
Optimal Cutoff	28%
False Positive Rate	23%
False Negative Rate	29%
True Positive Rate	71%
True Negative Rate	77%

Probabilities computed from the gold cells in these two tables

Cutoff	False Positive Rate	False Negative Rate	True Positive Rate	True Negative Rate	Overall Error Rate	Cost to do Nothing	Expected Cost Per Customer
0	1	0	1	0	0.7391065	$3,041.52	$3,272.84
0.01	0.8261194	0.00845666	0.99154334	0.1738806	0.6127965	$3,041.52	$3,076.20
0.02	0.75522388	0.01479915	0.985200846	0.2447761	0.5620518	$3,041.52	$2,998.77
0.03	0.7141791	0.03171247	0.968287526	0.2858209	0.536128	$3,041.52	$2,966.46
0.04	0.67761194	0.03805497	0.961945032	0.3223881	0.5107557	$3,041.52	$2,929.42
0.05	0.64179104	0.05073996	0.949260042	0.358209	0.4875896	$3,041.52	$2,899.26
0.06	0.60597015	0.05708245	0.942917548	0.3940299	0.4627689	$3,041.52	$2,863.10
0.07	0.56716418	0.06342495	0.936575053	0.4328358	0.4357419	$3,041.52	$2,823.43
0.08	0.54626866	0.06976744	0.930232558	0.4537313	0.4219526	$3,041.52	$2,804.84
0.09	0.51865672	0.08668076	0.913319239	0.4813433	0.405957	$3,041.52	$2,788.33
0.1	0.48283582	0.10570825	0.894291755	0.5171642	0.3844457	$3,041.52	$2,764.17
0.11	0.45746269	0.12473573	0.875264271	0.5425373	0.3706564	$3,041.52	$2,752.30
0.12	0.4380597	0.13319239	0.866807611	0.5619403	0.3585218	$3,041.52	$2,737.46
0.13	0.42462687	0.14587738	0.854122622	0.5753731	0.3519029	$3,041.52	$2,733.65
0.14	0.40373134	0.15856237	0.841437632	0.5962687	0.3397683	$3,041.52	$2,721.05
0.15	0.38432836	0.17547569	0.824524313	0.6156716	0.32984	$3,041.52	$2,714.21
0.16	0.37164179	0.18181818	0.818181818	0.6283582	0.322118	$3,041.52	$2,705.28
0.17	0.35522388	0.18604651	0.813953488	0.6447761	0.3110866	$3,041.52	$2,689.95
0.18	0.33955224	0.19661734	0.803382664	0.6604478	0.3022614	$3,041.52	$2,681.50
0.19	0.32835821	0.20930233	0.790697674	0.6716418	0.2972973	$3,041.52	$2,680.33
0.2	0.31492537	0.21987315	0.78012685	0.6850746	0.2901269	$3,041.52	$2,674.51
0.21	0.30746269	0.22198732	0.778012685	0.6925373	0.2851627	$3,041.52	$2,667.73
0.22	0.29850746	0.23044397	0.769556025	0.7014925	0.2807501	$3,041.52	$2,665.19
0.23	0.2858209	0.24524313	0.754756871	0.7141791	0.2752344	$3,041.52	$2,664.25
0.24	0.27089552	0.25792812	0.742071882	0.7291045	0.2675124	$3,041.52	$2,658.68
0.25	0.26119403	0.25792812	0.742071882	0.738806	0.260342	$3,041.52	$2,647.26
0.26	0.25223881	0.26849894	0.731501057	0.7477612	0.256481	$3,041.52	$2,646.72
0.27	0.23955224	0.28964059	0.710359408	0.7604478	0.25262	$3,041.52	$2,651.78
0.28	0.22985075	0.29386892	0.706131078	0.7701493	0.2465527	$3,041.52	$2,644.36
0.29	0.2238806	0.30232558	0.697674419	0.7761194	0.2443464	$3,041.52	$2,645.33
0.3	0.21716418	0.32135307	0.678646934	0.7828358	0.2443464	$3,041.52	$2,655.41

At-Risk	Retention Offer?	Leave?	Probability	Cost	Total Expected Cost
Y	Y	Y	10.27%	$13,100	$2,644.36
Y	Y	N	8.40%	$1,600	
Y	N	Y	7.77%	$11,500	
Y	N	N	0.00%	$0	
N	Y	Y	0.00%	$13,100	
N	Y	N	16.91%	$1,600	
N	N	Y	0.00%	$11,500	
N	N	N	0.00%	$0	
			100.00%		

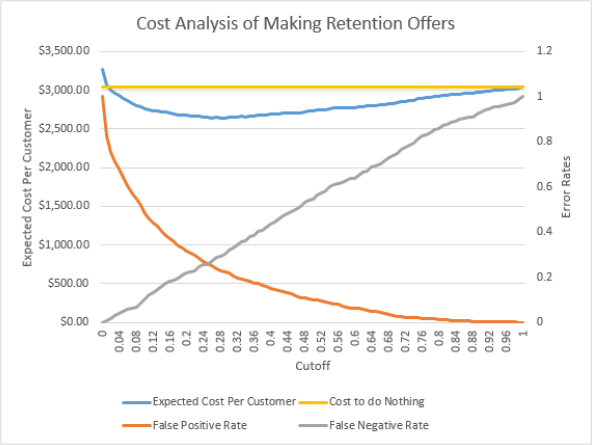

Figure 1

cutoff	FP.rate	FN.rate	TP.rate	TN.rate	OE.rate
0	1	0	1	0	0.739106
0.01	0.826119	0.008457	0.991543	0.173881	0.612796
0.02	0.755224	0.014799	0.985201	0.244776	0.562052
0.03	0.714179	0.031712	0.968288	0.285821	0.536128
0.04	0.677612	0.038055	0.961945	0.322388	0.510756
0.05	0.641791	0.05074	0.94926	0.358209	0.48759
0.06	0.60597	0.057082	0.942918	0.39403	0.462769
0.07	0.567164	0.063425	0.936575	0.432836	0.435742
0.08	0.546269	0.069767	0.930233	0.453731	0.421953
0.09	0.518657	0.086681	0.913319	0.481343	0.405957
0.1	0.482836	0.105708	0.894292	0.517164	0.384446
0.11	0.457463	0.124736	0.875264	0.542537	0.370656
0.12	0.43806	0.133192	0.866808	0.56194	0.358522
0.13	0.424627	0.145877	0.854123	0.575373	0.351903
0.14	0.403731	0.158562	0.841438	0.596269	0.339768
0.15	0.384328	0.175476	0.824524	0.615672	0.32984
0.16	0.371642	0.181818	0.818182	0.628358	0.322118
0.17	0.355224	0.186047	0.813953	0.644776	0.311087
0.18	0.339552	0.196617	0.803383	0.660448	0.302261

Figure 2