***Background:***

Telcos are in the middle of a dramatic change in their average revenues per user; increased competition along with similarity of offerings and diversity of cheap phones have created a lot of customer churn in their user base. Cell2Cell is a telecommunication firm that seeks your help to manage its churn. Your task is to (1) develop a statistical model for predicting customer churn, (2) use the model to identify the most important drivers of churn, and (3) with these new insights, recommend a customer churn management program to the CBM (Customer Base Management) Group.

Currently, Cell2Cell uses mass marketing techniques for building brand awareness and acquire customers. While the explosive growth of cell subscribers helped this strategy, as the growth slowed, the net additions minus exiting customers has fallen into the red. To counter this, the CBM group is trying to increase revenue per user, up-sell new products and decrease churn.

When churn was first measured, the rate of customers leaving the company was close to 5% per month, refreshing about half the subscriber base every year. In response, Cell2Cell implemented a reactive retention program when customers called with the intention to leave. While this was partially successful, a truly data-driven churn management strategy that is proactive is still missing.

The client recommends a three-stage process for implementing the proactive churn management strategy. First is to develop an accurate predictive churn model. Second is to identify the most important factors that drive subscribers churning. Third, use the insights from the second stage to identify offers that should be targeted to customers with a high (at least 75% higher than average) risk of churning. Offers might include cash incentives, price discounts, or any other enticement that could be expected to work best. There is no need to target the same incentive to each of the high risk customers, although the offer should be financially sound.

To assist your prototype analysis, your contact Sarah has assembled a “calibration” database consisting of 40,000 customers and a “test” database consisting of 31,047 customers. Each database contained (1) a “churn” variable signifying whether the customer had left the company two months after observation, and (2) a set of 75 potential predictor variables that could be used in a predictive churn model. Following usual model development procedures, the model would be estimated on the calibration data and tested on the validation data. At the time, Cell2Cell’s churn rate was about 2% per month. However, Sarah created the calibration database so that it contained roughly 50% churners. This was to make it easier for whatever data mining tool she used to identify the factors influencing customer churn. The validation data contained 2% churners.

***Instructions:***

The questions for this exercise are based upon the assigned reading for this class: "Cell2Cell: The Churn Game".  Please complete this reading before answering these questions.  In this exercise you want to design a model that will predict which customers are most likely to churn (e.g., not renew their mobile phone contracts) and then intervene with these customers.

Links for the dataset (as comma delimited file), documentation about variables, and R script to perform a suggested analysis are provided here:

[cell2cell\_data.csvPreview the document](https://canvas.cmu.edu/courses/9526/files/3644732/download?wrap=1)

[cell2cell\_doc.csvPreview the document](https://canvas.cmu.edu/courses/9526/files/3644733/download?wrap=1)

[cell2cell\_doc.xlsxPreview the document](https://canvas.cmu.edu/courses/9526/files/3644735/download?wrap=1)

[cell2cell\_Part1.R](https://canvas.cmu.edu/courses/9526/files/3644741/download?wrap=1)

You are welcome to consider alternative techniques (see [cell2cell\_Extra.R](https://canvas.cmu.edu/courses/9526/files/3644737/download?wrap=1) for optional code to estimate random forests and boosted trees), but this is not necessary to complete the assignment.  The main challenge is to interpret your model and explain why you have chosen it.

**This exercise is to be completed in your group as listed on Canvas.**

Please provide a clear, concise, and well organized essay that addresses at least the following questions.  You are free to address other issues in the case as well.  The intent of the assignment is to have you think critically about the business problem faced in the case and how it can be solved through data mining.  Analyze the quantitative material in the case to support your answers.  Spend most of your time in defining and defending your recommendation for what should be done.

Good answers may require assumptions of facts that may not be presented in the case.  You are welcome to make these assumptions, but please state these assumptions and briefly justify why that are reasonable.  Also, you may use whatever resources you can locate to provide further information about this industry or the web in general.  Please reference your sources.

***Required:***

Prepare a powerpoint presentation that communicates to the Chief Marketing Officer (CMO) an analysis of customer churn using Classification & Regression Tree (or Decision Tree) and a Logistic Regression using the information provided to you.  You should assume that the CMO is not familiar with analytic models -- but is very knowledgeable about the business.  Consider the following points in your presentation:

1. Purpose – what is the marketing purpose of your task? How do you think the company could use your results to target customers who are likely to churn?  Before you being the analysis name three relationships that you expect to see between churn and the predictive variables.  (Hint: focus on the direction of their influence, do you think a high/low value of this variable will result in more or less churn?) [1 point]
2. Estimate at decision tree that predicts CHURN. Prepare a graphic, visualization or table that summarizes the relationships that you have found. (Hint: use the provided script, you may want to change the settings used in the script to change the complexity of the decision tree.) [3 points]
3. Estimate a logistic regression that predicts CHURN. Prepare a table that summarizes the relationships that you have found. At a minimum include the following columns in your table: Variable, Parameter Estimate, Importance, and Meaning. The meaning should provide a description of the effect in plain English (e.g., if the parameter for “Eqpdays” is “0.0010527”, you could make a statement like “For every extra month (30 days) that a customers odds of churning versus not churning goes up by 3%.”). [3 points]
4. Which model do you find is best?  Be sure to justify your assessment based upon the performance of each of these models in the test sample using a confusion matrix and/or lift in the top decile. [2 points]
5. Complete predictions for the four users given at the end of the script (rows #8695, 15747, 29301, 34573).  Specifically compute the probability of churn.  Discuss a pro-active offer that you could give each.  Then give your best guess about how the probability of churn will change based upon your pro-active offer? This is meant to be a qualitative exercise to get you to think about how you will use this output to score all the users in the test dataset.  [1 point]

Note: The rubric for grading this assignment is a little different then the others.  The points for each question are given above and your answer will be evaluated per each question.

Your powerpoint presentation should be self-contained.  I would encourage you to keep the number of slides short -- perhaps only 5 to 10.  Write your slides so that they are self explanatory.  Use the Presentor Notes to make detailed comments about your slide -- if you think it is not clear.  Clearly title your slides.  Your grade is largely determined by how effective you are at presenting your results and convincing the audience/reader that they should follow your recommendations.  Finally, designate one of your slides (and only one) as a "Killer Slide" -- this is the page that you think is the most impactful and convincing your audience/reader that you have a good model.

The following variables are available for your analysis:

|  |  |  |  |
| --- | --- | --- | --- |
| **Original Position** | **Type** | **Variable** | **Variable Description** |
| 1 | Behavior | revenue | Mean monthly revenue |
| 2 | Behavior | mou | Mean monthly minutes of use |
| 3 | Behavior | recchrge | Mean total recurring charge |
| 4 | Behavior | directas | Mean number of director assisted calls |
| 5 | Behavior | overage | Mean overage minutes of use |
| 6 | Behavior | roam | Mean number of roaming calls |
| 7 | Behavior | changem | % Change in minutes of use |
| 8 | Behavior | changer | % Change in revenues |
| 9 | Behavior | dropvce | Mean number of dropped voice calls |
| 10 | Behavior | blckvce | Mean number of blocked voice calls |
| 11 | Behavior | unansvce | Mean number of unanswered voice calls |
| 13 | Behavior | threeway | Mean number of threeway calls |
| 14 | Behavior | mourec | Mean unrounded mou received voice calls |
| 15 | Behavior | outcalls | Mean number of outbound voice calls |
| 16 | Behavior | incalls | Mean number of inbound voice calls |
| 17 | Behavior | peakvce | Mean number of in and out peak voice calls |
| 18 | Behavior | opeakvce | Mean number of in and out off-peak voice calls |
| 19 | Behavior | dropblk | Mean number of dropped or blocked calls |
| 20 | Behavior | callfwdv | Mean number of call forwarding calls |
| 21 | Behavior | callwait | Mean number of call waiting calls |
| 23 | Behavior | months | Months in Service |
| 24 | Behavior | uniqsubs | Number of Uniq Subs |
| 25 | Behavior | actvsubs | Number of Active Subs |
| 27 | Behavior | phones | # Handsets Issued |
| 28 | Behavior | models | # Models Issued |
| 29 | Behavior | eqpdays | Number of days of the current equipment |
| 44 | Behavior | refurb | Handset is refurbished |
| 45 | Behavior | webcap | Hanset is web capable |
| 74 | Behavior | setprcm | Missing data on handset price |
| 75 | Behavior | setprc | Handset price (0=>missing) |
| 26 | Demo | csa | Communications Service Area |
| 30 | Demo | customer | Customer ID |
| 31 | Demo | age1 | Age of first HH member |
| 32 | Demo | age2 | Age of second HH member |
| 33 | Demo | children | Presence of children in HH |
| 34 | Demo | credita | Highest credit rating - a |
| 35 | Demo | creditaa | High credit rating - aa |
| 36 | Demo | creditb | Good credit rating - b |
| 37 | Demo | creditc | Medium credit rating - c |
| 38 | Demo | creditde | Low credit rating - de |
| 39 | Demo | creditgy | Very low credit rating - gy |
| 40 | Demo | creditz | Lowest credit rating - z |
| 41 | Demo | prizmrur | Prizm code is rural |
| 42 | Demo | prizmub | Prizm code is suburban |
| 43 | Demo | prizmtwn | Prizm code is town |
| 46 | Demo | truck | Subscriber owns a truck |
| 47 | Demo | rv | Subscriber owns a recreational vehicle |
| 48 | Demo | occprof | Occupation - professional |
| 49 | Demo | occcler | Occupation - clerical |
| 50 | Demo | occcrft | Occupation - crafts |
| 51 | Demo | occstud | Occupation - student |
| 52 | Demo | occhmkr | Occupation - homemaker |
| 53 | Demo | occret | Occupation - retired |
| 54 | Demo | occself | Occupation - self-employed |
| 55 | Demo | ownrent | Home ownership is missing |
| 56 | Demo | marryun | Marital status unknown |
| 57 | Demo | marryyes | Married |
| 58 | Demo | marryno | Not Married |
| 62 | Demo | travel | Has traveled to non-US country |
| 63 | Demo | pcown | Owns a personal computer |
| 64 | Demo | creditcd | Possesses a credit card |
| 67 | Demo | newcelly | Known to be a new cell phone user |
| 68 | Demo | newcelln | Known not to be a new cell phone user |
| 70 | Demo | incmiss | Income data is missing |
| 71 | Demo | income | Income (0=>missing) |
| 72 | Demo | mcycle | Owns a motorcycle |
| 73 | Demo | creditad | Number of adjustments made to customer credit rating |
| 12 | Marketing | custcare | Mean number of customer care calls |
| 59 | Marketing | mailord | Buys via mail order |
| 60 | Marketing | mailres | Responds to mail offers |
| 61 | Marketing | mailflag | Has chosen not to be solicited by mail |
| 65 | Marketing | retcalls | Number of calls previously made to retention team |
| 66 | Marketing | retaccpt | Number of previous retention offers accepted |
| 69 | Marketing | refer | Number of referrals made by subscriber |
| 76 | Marketing | retcall | Customer has made call to retention team |
| 77 | Marketing | churn | Dependent variable (1=churn, 0=retained) |
| 78 | Marketing | sample | Sample (1=training, 2=test, 3=prediction) |