Table of Contents

Introduction

- Our dataset comes from a fictional cell phone company called, “Cell2Cell”
- Some facts from the dataset
    - Over 70,000 rows
    - 78 different variables used
    - Pre-existing column that separates customers into training and testing data
        - Important to note, because of the limitations it provided to us
- We chose to use this dataset because it came clean and structured, and we had to do little work to clean up the data for our analysis
    - We enjoyed the background of the dataset, and knew we would be able to run deep, and great analysis in which we have learned in class
- Of all the datasets that we analyzed for this project, we felt that this one would provide us with the greatest learning experience and also allow us to a decent amount of models in which we learned throughout the past quarter

Business Objective

- The objective of our study is to create a model that can be used to analyze the effect that the variables, individually and collective, have on churn at Cell2Cell.
- The study analyzes the data through multiple models in order to find the best and is able to provide the greatest accuracy in future prediction efforts.
- Once the most appropriate model has been identified and a thorough understanding is established regarding which factors have the greatest impact on churn, a plan is developed with recommended countermeasures to significantly reduce churn.
- In essence, the model will be used to increase profitability by reducing churn. Reducing churn will reduce acquisition costs, which account for a majority of Cell2Cells marketing efforts.

Data Preparation

- Separating Training and Testing Data: The variables *calibrat* and *churndep* in the dataset provided information to set the training and testing data apart. *Calibrat* = 1 for training data and 0 for testing data.

Test Measures

1. AICc: Gives a measure of the goodness of fit of an estimated statistical model that can be used to compare two or more models
2. Confusion matrix: Matrix of actual and predicted values
3. Precision: Fraction of relevant instances among the retrieved instances

$$\text{Precision} = \frac{tp}{tp + fp}$$

4. Recall: Fraction of the total amount of relevant instances that were actually retrieved correctly

$$\text{Recall} = \frac{tp}{tp + fn}$$

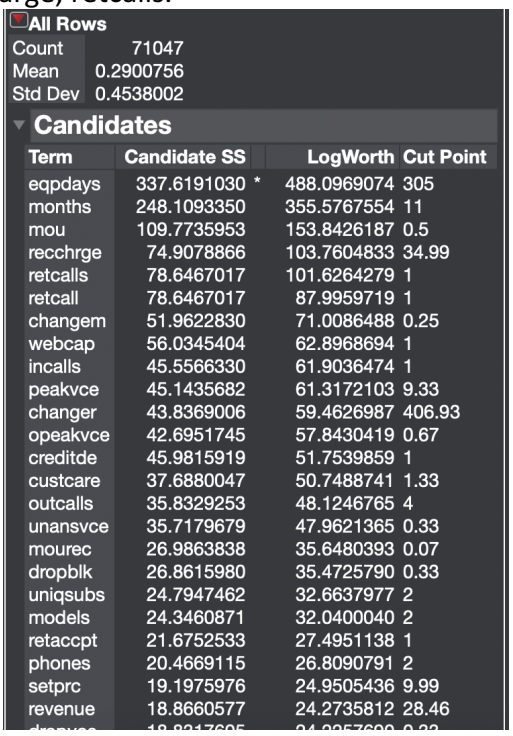
Where tp is true positive, fp is false positive and fn is false negative

Data Modeling

Identifying Significant variables

1. Decision Tree

- The data set is Cell2Cell Original Dataset with 71047 rows and 78 variables
- The dependent variable was churn (0=not churn, 1 = churn). The independent variables are other all variables except calibrate, customer, churndep, and csa.
- The objective is to know the significant variables which drive to customer churn
- We used JMP to do decision tree analysis. JMP make the best spilt based on statistical significance.
- The AICC achieved was 89357
- The results are shown below, the top five drivers to customer churn are eqpdays, months, mou, recharge, retcalls.



All Rows

Count	71047
Mean	0.2900756
Std Dev	0.4538002

Candidates

Term	Candidate SS		LogWorth	Cut Point
eqpdays	337.6191030	*	488.0969074	305
months	248.1093350		355.5767554	11
mou	109.7735953		153.8426187	0.5
recchrge	74.9078866		103.7604833	34.99
retcalls	78.6467017		101.6264279	1
retcall	78.6467017		87.9959719	1
changem	51.9622830		71.0086488	0.25
webcap	56.0345404		62.8968694	1
incalls	45.5566330		61.9036474	1
peakvce	45.1435682		61.3172103	9.33
changer	43.8369006		59.4626987	406.93
opeakvce	42.6951745		57.8430419	0.67
creditde	45.9815919		51.7539859	1
custcare	37.6880047		50.7488741	1.33
outcalls	35.8329253		48.1246765	4
unansvce	35.7179679		47.9621365	0.33
mourec	26.9863838		35.6480393	0.07
dropblk	26.8615980		35.4725790	0.33
uniqsubs	24.7947462		32.6637977	2
models	24.3460871		32.0400040	2
retaccpt	21.6752533		27.4951138	1
phones	20.4669115		26.8090791	2
setprc	19.1975976		24.9505436	9.99
revenue	18.8660577		24.2735812	28.46

2. Logistic Regression

- The entire data set with 71,047 rows was analyzed to get the most significant variables.
- The dependent variable was churn and independent variables were all except calibrat(calibrate customer), churndep and csa(communication service area).
- The method used was R function glm().
- The code used to run the model is:

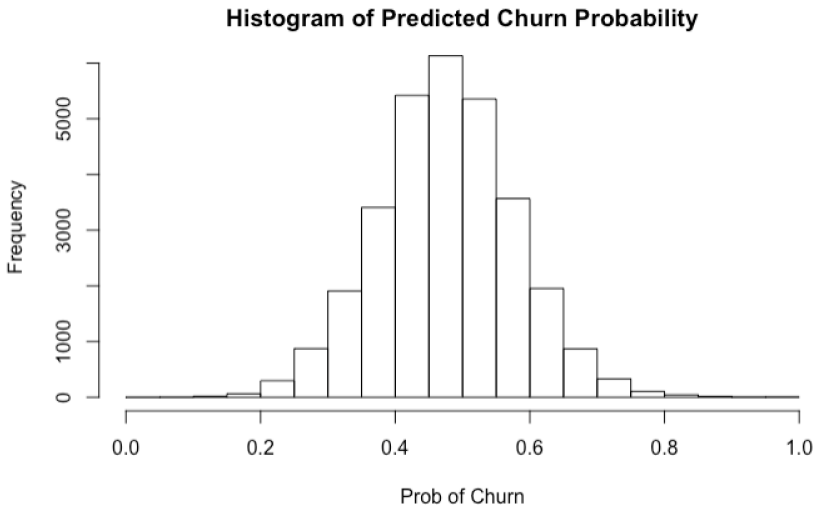
  glm.c2c <- glm(churn ~ .-calibrat-churndep, family=binomial(link='logit'), data=c2c)
- The AICC achieved is 52909.
- It was concluded that as the AICC for Decision Tree model was higher, we will use decision tree derived variables to shortlist the 5 most important ones.
- The results obtained are as shown below:

```
Coefficients: (2 not defined because of
singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.722e+00  8.820e-01   4.220 2.44e-05
revenue       1.720e-03  8.101e-04   2.123 0.033763
mou          -2.703e-04  5.094e-05  -5.306 1.12e-07
recchrge     -2.817e-03  9.048e-04  -3.113 0.001852
directas     -3.410e-03  6.096e-03  -0.559 0.575892
overage       8.332e-04  2.844e-04   2.929 0.003398
roam          7.180e-03  2.110e-03   3.402 0.000668
changem      -5.104e-04  5.451e-05  -9.365  < 2e-16
changer       2.343e-03  3.743e-04   6.260 3.85e-10
dropvce       1.130e-02  7.307e-03   1.546 0.122043
blckvce       6.618e-03  7.213e-03   0.917 0.358939
unansvce      1.083e-03  4.694e-04   2.307 0.021053
custcare     -5.940e-03  2.615e-03  -2.271 0.023135
threeway     -3.205e-02  1.163e-02  -2.755 0.005867
```

Prediction Modeling

1. Logistic Regression with all variables

- The aim is to predict churn for testing data by using logit (glm()) function.
- The data used was Training data (40,000 rows) and Testing data (31047 rows) which contained a few NA values. These values were omitted by setting. `getOption("na.action")` as `na.omit`
- The dependent variable was churn and independent variables were all except calibrat(calibrate customer,) churndep and csa(communication service area).
- Most of the predicted probability lied in the 0.5 range as seen in the histogram below.
- Result: The confusion matrix is shown below.
  - Precision: 2.81% which means that 2.81% of the instances predicted were “relevant” instances, i.e. churn=1.
  - Recall: 58.81% which means that 58.81% of the predicted relevant instances were actually correctly predicted.



Predicted Churn	Actual Churn 0	1	Row Total
0	17881	241	18122
	0.987	0.013	0.597
1	11901	345	12246
	0.972	0.028	0.403
Column Total	29782	586	30368

2. Logistic Regression with 5 variables

- The main goal for this model, was to take to the 5 variables that we found from section 1 and incorporate those into a logistic regression
    - *eqpdays, months, mou, recharge, retcalls*
- We wanted to be more specific in our analysis, and instead of using all 78 variables in our case, we only wanted to use those variables that were important to us
- New results…
    - Precision Rate = 2.62%
    - Recall = 59.96%

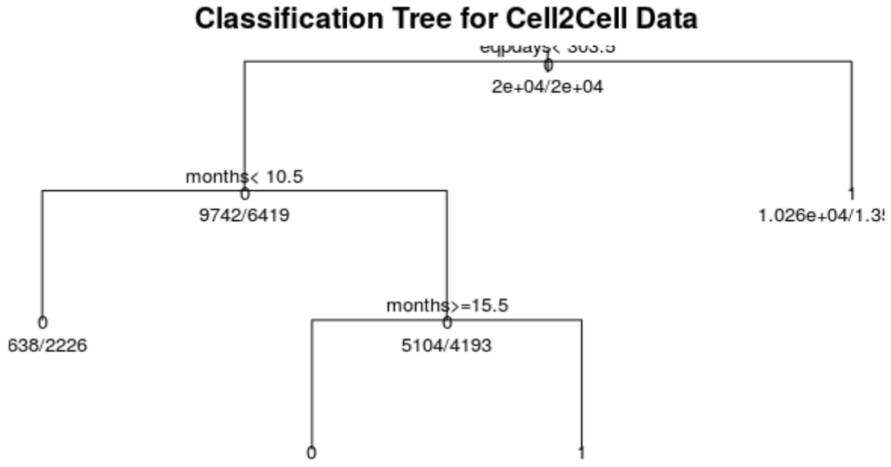
Predicted Response	Real Response 0	1	Row Total
0	18209	277	18486
	0.985	0.015	0.597
1	12158	327	12485
	0.974	0.026	0.403
Column Total	30367	604	30971

- Our New Results...
    - Compared to the logistic regression we ran with all of the variables from the case, this regression has a lower precision rate but a higher recall rate

- We as a group came to a conclusion where having more variables will always come up with a higher precision rate but having the right variables will come up with a more relevant and statistically important rate
- Based on the two logistic regressions that we ran, we would recommend to use the findings from this model in order to have a better outcome

3. Decision Tree

- The objective is to predict churn by using decision tree model.
- The data used are training data (40000 rows) and testing data (31047 rows). They are automatedly divided in original dataset.
- R package used: rpart
- The dependent variable was churn, the independent variables are eqpdays, months, mou, recharge, retcalls
- Results are shown below, R provided tree classification, and predictive value, we made the confusion matrix based on the result.



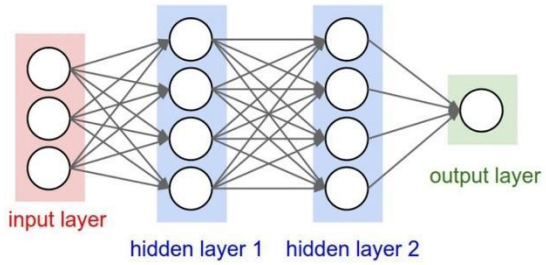
	Actual Churn = 0 (Not Churn)	Actual Churn = 1 (Churn)	Total
Prediction Churn = 0 (Model predicts not churn)	12,836	156	12,992
Prediction Churn = 1 (Model predicts churn)	17,602	453	18,055
Total	30,438	609	31,047

- The precision rate is **2.5%** and recall rate is **42.8%**

4. Neural Network

- Neural Network was chosen in this analysis because it is a powerful machine learning algorithm inspired by the way in which the brain performs a particular learning task.

- Core Concept: Like human brain, a neural net has a large of processing nodes that continuously interact with each other. Each neuron receives a number of inputs that carry different weights, and then each neuron will sum the weighed input signals and applies an activation to determine the output signal.



- To build a neural network model, a package named *"neuralnet"* in R was installed.

- **Data Manipulation**
  - As neural network requires considerable processing power, a few steps were taken in order to transform the data and get it ready for model processing.
  - Step 1: Variable Selection
    - Dependent variables: *churn (1 or 0)*
    - Independent variables: *eqpdays, months, mou, recchrge, retcalls*
  - Step 2: Remove rows with missing values
    - *Training data : 39,859 rows, churn rate is approximately 50%*
    - *Testing data: 30,971 rows, churn rate is approximately 2%*
  - Step 3: Use "Min-Max Normalization" to normalize the variables between 0 and 1 to prevent a particular variable affecting the prediction due to its large numeric value range.

- **Model Processing**
  - Step 1: Among 39,859 rows in the testing data, randomly select 5,000 observations to build the neural network model.
    - Test on a small portion of data and shuffle the order of the rows

```
data.train <- training_data[sample(1:nrow(training_data)),][1:5000,]
```

  - Step 2: Train the neural network by setting different number of hidden layers
    - Model 1: hidden layer = 20

```
nn.model_1<-neuralnet(churn~eqpdays+months+mou+recchrge+retcalls,data=data.train, hidden=20, threshold=0.01,
                      stepmax=1e6, act.fct = "logistic", linear.output = FALSE)
```

    - Model 2: hidden layer = 10* 2

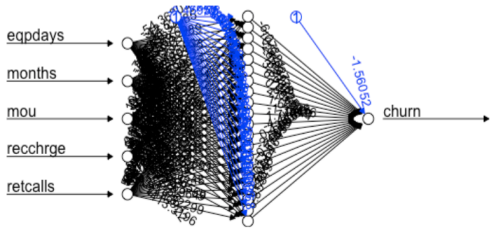
```
nn.model_2<-neuralnet(churn~eqpdays+months+mou+recchrge+retcalls,data=data.train, hidden=c(10,2), threshold=0.01,
                      stepmax=1e6, act.fct = "logistic", linear.output = FALSE)
```

  - Step 3: Predict churn behavior with the neural network
    - Predict the probability for the test data using the compute() function
  - Step 4: Convert probabilities into binary classes

- Use round() to convert the probability score to either 0 to 1

- Step 5: Plot the neural network model and create a cross table to judge the quality of the predictions

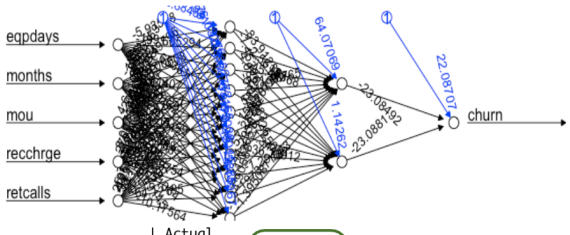
Model 1– Hidden layers: 20



Predict	Actual 0	1	Row Total
0	16753	249	17002
1	13614	355	13969
Column Total	30367	604	30971

Precision: 355/13969 = 2.54%
Recall: 355/604 = 58.8%

Model 2– Hidden layers: 10 * 2



Predict	Actual 0	1	Row Total
0	15719	214	15933
1	14648	390	15038
Column Total	30367	604	30971

★ **Precision: 390/15038 = 2.59%**
Recall: 390/604 = 64.6%

- Result: Model 2 performs better than model 1 in terms of recall rate (64.6% vs. 58.8%), suggesting that by adding one more hidden layer, the accuracy of prediction improved by 5.8%.
- Insights: To improve the precision rate, it is necessary to test different independent variables, try different methods to normalize the data, or adding more hidden layers.

Evaluation of All Models

Summary of Variables

Variables	Description	Min.	Max.	Median	Mean
Eqpdays	Number of days of the current equipment	5	1823	330	380.3
Months	Months in Service	6	61	16	18.75
Mou	Mean monthly minutes of use	0	7667.8	366.0	525.7
Recchrge	Mean total recurring charge	-11.29	399.99	44.99	46.88
Retcalls	Number of calls previously made to retention team	0	4	0	0.037

Comparison of All Models

	Decision Tree (5 Variables)	Logistic Regression (All Variables)	Logistic Regression (5 Variables)	Neural Network (5 Variables)
Precision Rate	2.51%	2.81%	2.62%	2.59%
Recall Rate	42.8%	58.81%	59.96%	64.6%

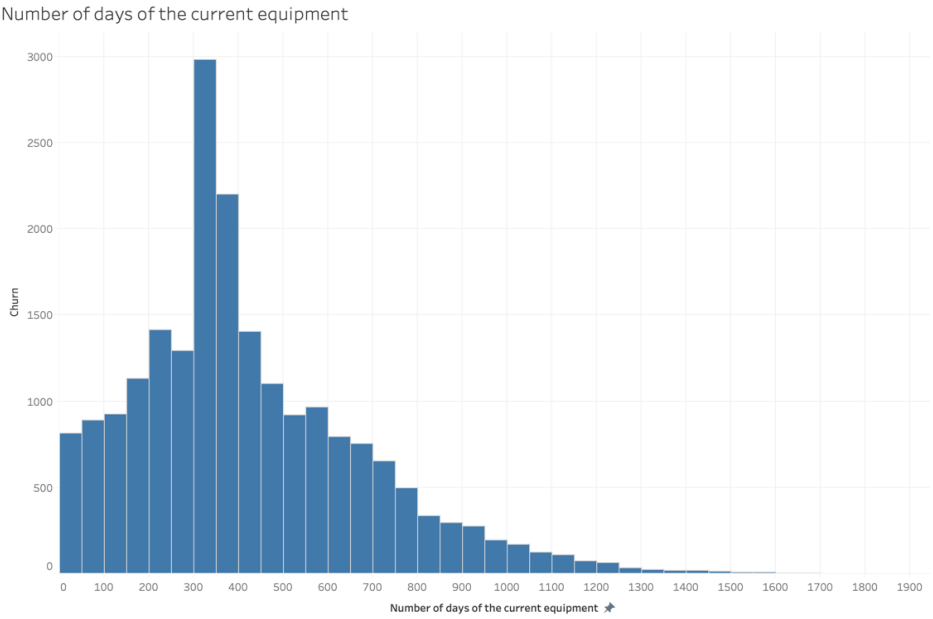
- The precision rate of the four models is low because there are many cases where "predicted as churn, but actually not churn". The reason is that the churn rate in the training data is 50%, but the churn rate in the testing data is only 2%. We infer that the original data has a certain degree of bias.
- The decision tree model with 5 selected variables has the lowest precision rate and recall rate. It predicts customer churn poorly.
- Compared with logistic regression model with all related variables, logistic regression model with 5 selected variables has slightly lower precision rate but higher recall rate.
- Neural network model with 5 selected variables predicts well with highest recall rate and similar precision rate with logistic regression model with 5 selected variables.
- Overall, logistic regression model and neural network are both acceptable models to predict customer churn behavior.

Data Visualization

Graphically seeing the relation between churn and most significant 5 variables helped in understanding relation between the 2 variables.

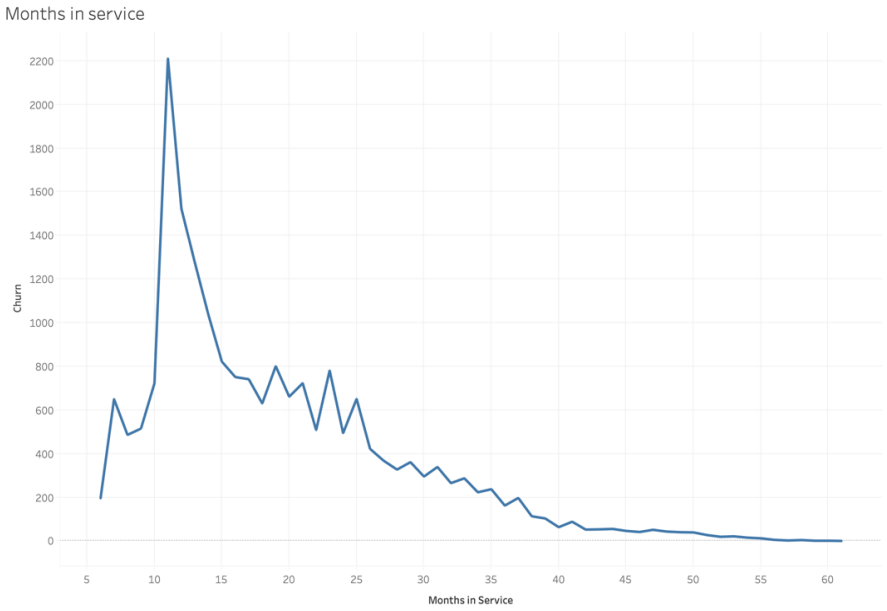
1. Number of days of the current equipment vis-à-vis Number of churns

   It is seen that this is right skewed and users who have used current equipment for 300-350 days' churn more.



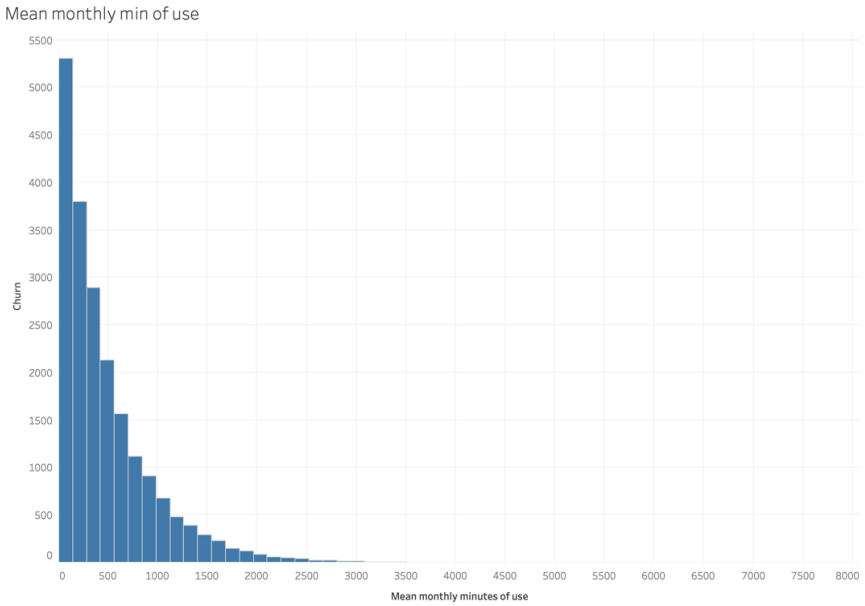
2. Months in service vis-à-vis Number of churns

   There is a spike in no. of churn when the months in service is 11 days.



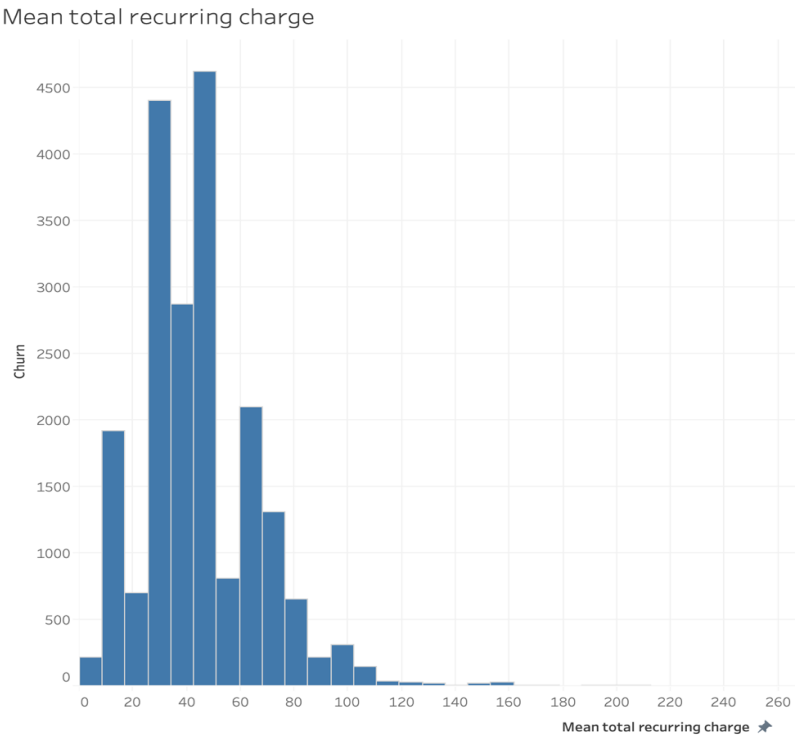
3. Mean monthly minutes of use vis-à-vis Number of churns

   Lower the mean monthly minutes of use, higher is the churn because the customer is not highly involved in using cell2cell service.



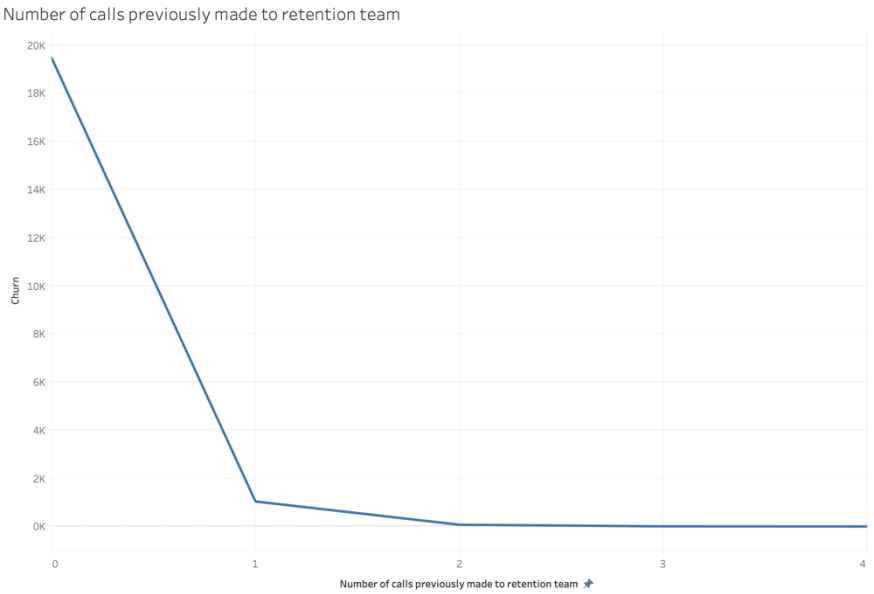
4. Mean total recurring charge vis-à-vis Number of churns

   Very low or very high recurring charge doesn't lead to churn, but a mean total recurring charge between 25-45 leads to more churn.



5. Number of calls previously made to retention team vis-à-vis Number of churns

True to logic, if the no. of calls made to retention team are 0, the churn is higher because this shows that no effort was made to keep the customer from churning



Summary

- Decision tree is useful for knowing significant variables
- Prediction/Scoring Logistic Regression and Neural Network models are the best fit
- Recall and Precision cannot be increased simultaneously because if we try to increase one, the other decreases. Depending on the business case, priority has to be given to either one metric
- In this case, Neural network has the highest recall rate of 64.6%
- Top 5 drivers to Churn: *Number of days of the current equipment*, *Months in Service*, *Mean monthly minutes of use*, *Mean total recurring charge, Number of calls previously made to retention team*

Recommendations

Based on our analysis, there are a number of recommendations that will serve as countermeasures against churn at Cell2Cell.

1. To reduce churn associated with the duration of equipment ownership, further analysis must be performed to investigate the intersection between segments and preferred devices. This can be used to proactively market to customers at a time when they are likely to initiate replacing their phone, and are therefore vulnerable to offers from other service providers. To support this initiative, Cell2Cell must ensure it has selected the most popular devices to offer and that the fulfillment channel is substantially robust.

2. Similar to the previous point, Cell2Cell must fine-tune their segments and reach out to flight-risk clients at the critical point in their months of service, before the client looks for competitive service elsewhere.

3. There Cell2Cell data presents a strong correlation between average monthly minutes of use and churn. Those that use very few minutes per month and those who are power users are both at risk of leaving. Competitive plans must be formulated to best serve these customers. Those that use very few minutes may want a “just-in-case” phone, but do not want to pay much. Those that are on the higher end of use per month are, in general, more conscious of the service provider landscape, and must feel like they are not over paying.

4. To counter churn associated with average recurring charges, Cell2Cell must have competitive plans for each segment and be extremely proactive at ensuring that clients are offered the option to switch to the new lower price plan when available. We must not wait until a customer is calling to cancel their service to offer them a better price. By this point they have made a decision and possibly even a commitment. Cell2Cell must actively investigate whether clients are currently being offered the best price, and reach out to them to make an offer when newer plans are made available. This goodwill should go far in retention efforts.

5. Finally, if a customer has reached out to customer service multiple times, or more than once in a short period, follow-up should occur from a higher level employee to ensure that the customer is satisfied and their needs are being met or exceeded.