The University of Hong Kong

Faculty of Engineering

Department of Computer Science

ICOM6044 Data Science for Business

Instructor: Professor Alan Montgomery

Group Written Assignment 3

Cell2Cell (Part I) Case

By

CHAN Chui Yi, Christy (UID: 3035752145)

KWAN Ho Tin, Joseph (UID:3035751828)

LEE Tin Chak. Daniel (UID: 3035752183)

NG Wah Hay, Jacky (UID: 3035751725)

Date of submission: 22 July 2021

Cell2Cell

To: Charles R. Morris and Sarah A.

From: Daniel Lee, Jacky Ng, Joseph Kwan, and Christy Chan

Date: 22 July 2021

Subject: Analyse Customer Churn Using Decision Tree and Logistic Regression

---

The purpose of this task is to find out which of our customers are more likely not to continue using our services in the coming months and suggest ways to retain these customers. We outline two ways to predict churn and recommend one of these ways to do so.

Expected Relationships Between Churn and Some Variables

We assume customers call our customer retention team to request for discontinuing our services. We expect that the higher the number of calls previously made to our retention team by a customer, the more probable this customer would unsubscribe from our services. A high number of these calls would result in more churn.

A customer probably looks for another handset when the one he is using becomes old. This customer might consider switching to another cell phone services provider when they look for another handset. We expected that the higher the number of days of the current equipment, the more likely that customer would discontinue our relationship. A higher number of equipment days would result in more churn.

Our customers can refer consumers to our services. We believe that a referral indicates that the customer is satisfied with our services. Hence, we expect the higher the number of referrals, the more likely that customer would continue subscribe for our services. A high number of referrals would result in less churn.

Decision Tree

To predict the churn, a decision tree model has been adopted. In order to further differentiate distinctive groups of people who are expected to be leaving, three conditions have been set in our model [Appendix 2-1]:

1. The cp is 0.00123324 (12 nodes)
2. The probability of end nodes we chose are greater than 50% meaning that these 5 case scenarios are the major causes to the departure of our customers

3. 8 variables have been used. [Appendix 2-2]

A table that summarises the relationships we found is presented in Appendix 2-3. Our key findings are as follows:

- In 4 out of 5 nodes (59% of our sample size) we classified from the decision tree model, expected leaving customers have been using their current handsets equal to or more than 306 days.
- In 3 out of 5 nodes (48% of our sample size), expected leaving customers used our service at least more than a year.
- The strongest reason (the darkest in green colour in the decision tree model) to discontinue our service is that customers barely used our service but signed up a long-term contract with us at the beginning. They may treat our service plan as their secondary usage.

Logistic Regression

Logistic regression model is applied to predict churn. There are 3 types of variables used – Demographic, Behaviour and Marketing. In the result, 21 variables are identified to have relationship in churn rate prediction and a summary table is prepared [Appendix 3-1].

The top 3 reasons contributing to churn rate are number of days of the current equipment, number of unique subscribers in a household, number of calls previously made to our retention team (excluding the interaction variables). All these 3 variables have a positive relationship with churn rate. In the telecom industry, it is quite commonly seen that customers switch handset due to technology advancement and function upgrade of mobile phones; while telecom companies design different marketing programs (e.g. free phone, family plan, limited promotion offer) to attract switching of service provider. The logistic regression model with top 3 identified variables supports our observation [Appendix 3-1].

Some variables are identified to have negative relationship with churn rate. For example, the larger percentage change in minutes of use is observed, the lower is the churn rate of that customer. Similar negative impact is observed for age in the model. The older the customer is, the lower is his churn rate. It is reasonable to say that customers who do not have a stable pattern of minutes usage or customers who are old in age are less likely to make changes including switching to a new telecom company.

Recommended Model

After examining the two prediction churn models generated, we can conclude Decision Tree is a more suitable one for Cell2Cell's use case for several reasons. First, by computing the confusion matrix for the two models obtained [Appendix 4-1], we can observe that Decision Tree manages to give a better precision rate when compared to Logistic Regression (58.9% vs 57.6%). The tree model also gives a better accuracy (59.7% vs 57.6%) and a better lift (1.95 vs 1.73 after adjustment from over-sampling). This implies that our tree model could give a more precise customer churn prediction and indicate if a customer is likely to churn or not more accurately. By using the Decision Tree model, Cell2Cell would have less chance of identifying a churned customers wrongly so that their budgets allocated for churn management can be more effectively utilized with less leakage due to wrong prediction. In addition, based on the plotted ROC curves [Appendix 4-2] for the two models, Decision Tree is clearly having a higher AUC against Logistic Regression (0.63 vs 0.61), which is also a strong indicator that the tree is a better model that can produce more true positives (Predicting Churn correctly) than giving false alarms (Predicting Churn wrongly).

Moreover, we believe the top variables identified in Decision Tree (i.e. Equipment Days, Months in service, MOU) that have the most impact to churn prediction are pretty much aligned with the top customer churn factors that we understand about the telecom industry. For example, it is not surprising to see a customer switches to another provider if he has owned an equipment for long time and been staying with one telecom company for years, given that many providers give out promotional offers with free phone equipment as a gift to attract customer who wants to get a new phone and is willing to sign for longer term contract[1]. From this point of view, it naturally makes sense to select Decision Tree over other models as less efforts will then be required to explain it to the business.

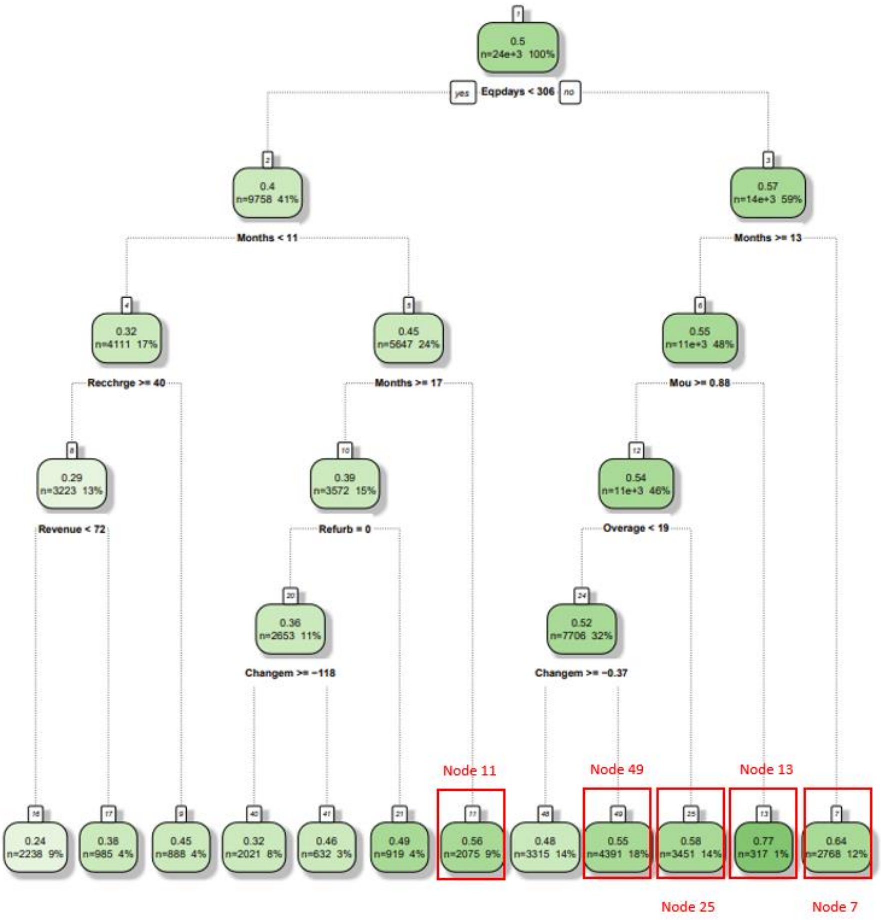
Lastly, from business management perspective, it is important to have a data model that is easy to be interpreted so that misalignment of understanding can be avoided. It is apparent that the tree visualization that Decision Tree model produces is more intuitive than Logistic Regression. Therefore, Decision Tree is preferred where business can directly make use of it to design business strategy in the next stage.

[1] Strategies for Reducing Churn Rate in the Telecom Industry
https://www.omnisci.com/blog/strategies-for-reducing-churn-rate-in-the-telecom-industry

<u>**Appendixes**</u>

Appendix 2-1

Decision Tree



The value of cp is 0.00123324 (12 nodes). The nodes selected represent groups of customers who are more than 50% likely to end our relationship from our data.

Appendix 2-2

Variables used in the decision tree

type	**subtype**	**variable**	**description**
Behavior	Spending	Recchrge	Mean total recurring charge
Behavior	Usage	Mou	Mean monthly minutes of use
Behavior	Usage	Overage	Mean overage minutes of use
Behavior	Usage	Changem	% Change in minutes of use
Behavior	Spending	Revenue	Mean monthly revenue
Behavior	Handset	Refurb	Handset is refurbished
Behavior	Tenure	Months	Months in Service
Behavior	Handset	Eqpdays	Number of days of the current equipment

Appendix 2-3

Summary table to explain relationships of variables in decision tree

	Nodes / Buckets				
	11	**49**	**25**	**13**	**7**
Decision Variables	Number of days using the handset <306 Months in service between 10 months and 17 months	Number of days using the handset >=306 Months in service >=13 months Mean monthly minute of use >=0.88 minutes Mean overage minutes of use < 19 minutes % change in minutes of use < -0.37	Number of days using the handset >=306 Months in service >=13 months Mean overage minutes of use => 19 minutes	Number of days using the handset >=306 Months in service >=13 months Mean monthly minute of use <0.88 minutes	Number of days using the handset >=306 Months in service <13 months

Outcomes (Adjusted for over-sampling)	2.5% churn rate 2075 group size 52 expected leaving customers	2.4% churn rate 4391 group size 105 expected leaving customers	2.7% churn rate 3451 group size 93 expected leaving customers	6.4% churn rate 317 group size 20 expected leaving customers	3.5% churn rate 2768 group size 97 expected leaving customers
Their Story	Love keeping up with the trend of handset along with the telecom service	Used to have a lot of voice calls but has been minimizing	Signed a relatively long and unsuitable plan as the service is not sufficient for their mixed minutes usage	Treated the plan as secondary or backup usage	committed to short term telecom service without changing their handsets frequently

Node 11:

We predict that this group of people love to change their handsets frequently with the flexible telecom service as other telecom services usually offer bundle packages including the services and the new handsets.

Node 49:

We predict that they made a lot of voice calls initially but they started to shrink the usage probably due to many communication methods at the market right now instead of voice calls.

Node 25:

We predict that they signed long term contracts with us, but the actual usage is far more than expected. Probably, they signed the contracts because of the promotional discount at the beginning.

Node 13:

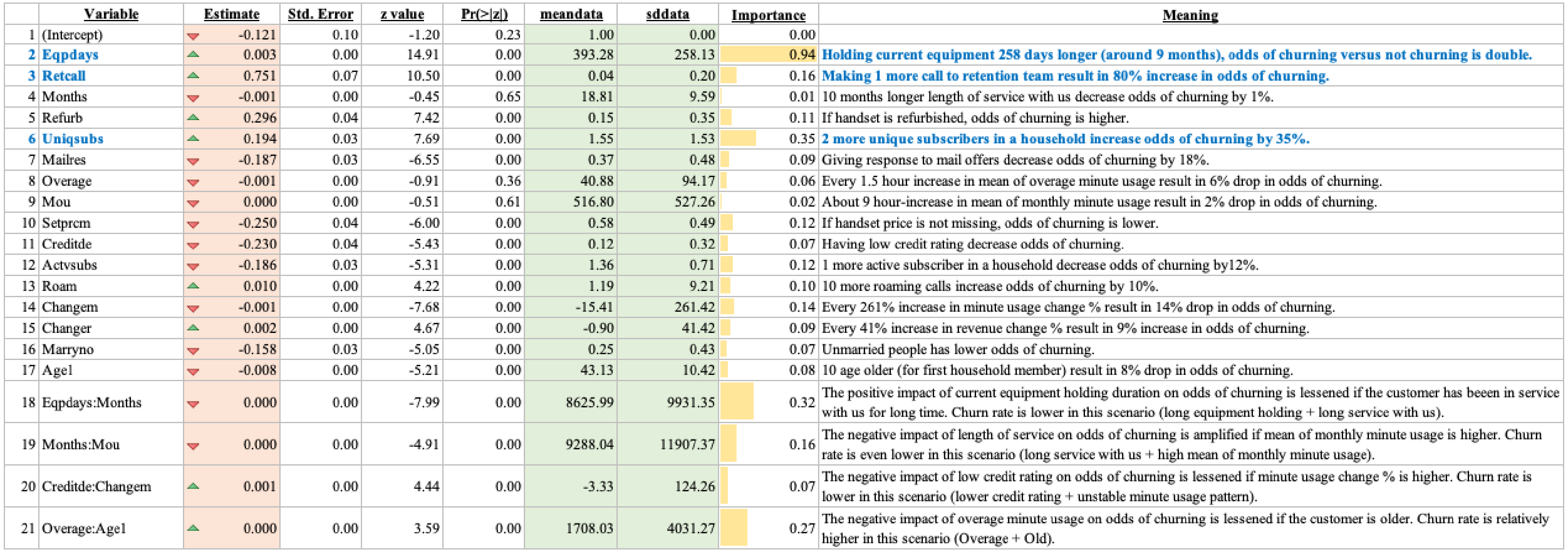
We predict that they used our service as a backup as the usage of the voice calls is close to zero.

Node 7:

We predict that the users usually look for other promotions of the telecom services only and commit to short term contracts.

Appendix 3-1

Summary table to explain relationships of variables in logistic regression model



	Variable	Estimate	Std. Error	z value	Pr(>|z|)	meandata	sddata	Importance	Meaning
1	(Intercept)	▼ -0.121	0.10	-1.20	0.23	1.00	0.00	0.00	
2	**Eqpdays**	▲ 0.003	0.00	14.91	0.00	393.28	258.13	0.94	**Holding current equipment 258 days longer (around 9 months), odds of churning versus not churning is double.**
3	**Retcall**	▲ 0.751	0.07	10.50	0.00	0.04	0.20	0.16	**Making 1 more call to retention team result in 80% increase in odds of churning.**
4	Months	▼ -0.001	0.00	-0.45	0.65	18.81	9.59	0.01	10 months longer length of service with us decrease odds of churning by 1%.
5	Refurb	▲ 0.296	0.04	7.42	0.00	0.15	0.35	0.11	If handset is refurbished, odds of churning is higher.
6	**Uniqsubs**	▲ 0.194	0.03	7.69	0.00	1.55	1.53	0.35	**2 more unique subscribers in a household increase odds of churning by 35%.**
7	Mailres	▼ -0.187	0.03	-6.55	0.00	0.37	0.48	0.09	Giving response to mail offers decrease odds of churning by 18%.
8	Overage	▼ -0.001	0.00	-0.91	0.36	40.88	94.17	0.06	Every 1.5 hour increase in mean of overage minute usage result in 6% drop in odds of churning.
9	Mou	▼ 0.000	0.00	-0.51	0.61	516.80	527.26	0.02	About 9 hour-increase in mean of monthly minute usage result in 2% drop in odds of churning.
10	Setprcm	▼ -0.250	0.04	-6.00	0.00	0.58	0.49	0.12	If handset price is not missing, odds of churning is lower.
11	Creditde	▼ -0.230	0.04	-5.43	0.00	0.12	0.32	0.07	Having low credit rating decrease odds of churning.
12	Actvsubs	▼ -0.186	0.03	-5.31	0.00	1.36	0.71	0.12	1 more active subscriber in a household decrease odds of churning by12%.
13	Roam	▲ 0.010	0.00	4.22	0.00	1.19	9.21	0.10	10 more roaming calls increase odds of churning by 10%.
14	Changem	▼ -0.001	0.00	-7.68	0.00	-15.41	261.42	0.14	Every 261% increase in minute usage change % result in 14% drop in odds of churning.
15	Changer	▲ 0.002	0.00	4.67	0.00	-0.90	41.42	0.09	Every 41% increase in revenue change % result in 9% increase in odds of churning.
16	Marryno	▼ -0.158	0.03	-5.05	0.00	0.25	0.43	0.07	Unmarried people has lower odds of churning.
17	Age1	▼ -0.008	0.00	-5.21	0.00	43.13	10.42	0.08	10 age older (for first household member) result in 8% drop in odds of churning.
18	Eqpdays:Months	▼ 0.000	0.00	-7.99	0.00	8625.99	9931.35	0.32	The positive impact of current equipment holding duration on odds of churning is lessened if the customer has beeen in service with us for long time. Churn rate is lower in this scenario (long equipment holding + long service with us).
19	Months:Mou	▼ 0.000	0.00	-4.91	0.00	9288.04	11907.37	0.16	The negative impact of length of service on odds of churning is amplified if mean of monthly minute usage is higher. Churn rate is even lower in this scenario (long service with us + high mean of monthly minute usage).
20	Creditde:Changem	▲ 0.001	0.00	4.44	0.00	-3.33	124.26	0.07	The negative impact of low credit rating on odds of churning is lessened if minute usage change % is higher. Churn rate is lower in this scenario (lower credit rating + unstable minute usage pattern).
21	Overage:Age1	▲ 0.000	0.00	3.59	0.00	1708.03	4031.27	0.27	The negative impact of overage minute usage on odds of churning is lessened if the customer is older. Churn rate is relatively higher in this scenario (Overage + Old).

Remarks:

1. Column 3 “Estimate” illustrates positive relationship (Green arrow) and negative relationship (Red arrow) of the variable on churn rate. The degree of impact can be calculated by multiplying the “Estimate” coefficient with corresponding variable data of a particular user.
2. To understand the change in odds of churning due to 1 SD movement above/below variable mean (i.e. standardized scale), column 9 “Importance” should be referred.

3. Item 18 to 21 are interaction variables.

4. Interactions exist for some variables in the model. Although the current equipment holding time is the top 1 variable having positive relationship with churn rate, the impact is not as strong if the customer is in long service with Cell2Cell. The reason is that length of service individually has negative relationship with churn rate. Combining two variables into the interactive “Eqpdays:Months” variable results in a negative relationship that require special attention.

Appendix 4-1

Confusion Matrix of Decision Tree for Churn (i.e. Customer Churn = Yes)

		Actual	
		No	Yes
Prediction	No	2,261	1,473
	Yes	1,752	2,514

Accuracy = ((Prediction=No | Actual=No) + (Prediction=Yes | Actual=Yes)) / Total no. of samples

= (2,261+2,514) / (2,261+2,514+1,473+1,752)

= 59.7%

Precision = (Prediction=Yes | Actual=Yes) / ((Prediction=Yes | Actual=Yes) + (Prediction=Yes | Actual=No))

= 2,514 / (2,514+1,752)

= 58.9%

Lift (Before Adjustment* for over-sampling) = Top Decile Prediction / Baseline Prediction

= 66.4% / 49.8%

= 1.33

Lift (After Adjustment* for over-sampling) = Top Decile Prediction adjusted / Baseline Prediction adjusted

= 3.9% / 2.0%

= 1.95

Confusion Matrix of Logistic Regression for Churn (i.e. Customer Churn = Yes)

		Actual	
		No	Yes
Prediction	No	2,343	1,721
	Yes	1,670	2,266

Accuracy = ((Prediction=No | Actual=No) + (Prediction=Yes | Actual=Yes)) / Total samples

= (2,343+2,266) / (2,343 +2,266+1,721+1,670)

= 57.6%

Precision = (Prediction=Yes | Actual=Yes) / ((Prediction=Yes | Actual=Yes) + (Prediction=Yes | Actual=No))

= 2,266 / (2,266+1,670)

= 57.6%

Lift (Before Adjustment* for over-sampling) = Top Decile Prediction / Baseline Prediction

= 63.6% / 49.8%

= 1.28

Lift (After Adjustment* for over-sampling) = Top Decile Prediction adjusted / Baseline Prediction adjusted

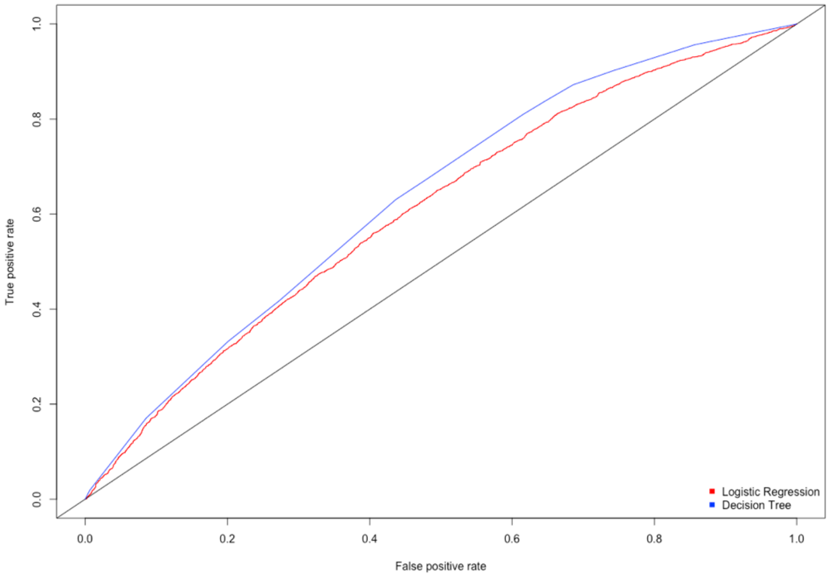
= 3.4% / 2.0%

= 1.73

*Adjustment is required because in training and validation dataset churn rate is over-sampled to 50%, where in original dataset the monthly churn rate is only about 2%.

Appendix 4-2

ROC Curves:



AUC of the two models:

	AUC
Decision	0.63
Logistic Regression	0.61