Data Science for Business
Lecture #9
Pro-active Churn Prevention, Preparing for Cell2Cell Part 2

Prof. Alan L. Montgomery

The University of Hong Kong & Carnegie Mellon University, Tepper School of Business

email: alanmontgomery@cmu.edu

4. Recommendation
Getting Ready for Part 2

How do we make a decision with our model?

How do we communicate our results to convince someone else to use them?

Translating your Model into Potential Promotional Offers

Importance	Variable	Meaning
1	Eqpdays	***New Phone Offer:*** Offer new phones to customers with phones that are one-year old
3	Retcall	***Reach Out Campaign:*** Pro-actively contact those who have not and offer New Phone
4	Months	***Gift Card Campaign:*** Instead of a new phone offer Gift Cards or Rebates for a new phone since loyal customers less likely to churn
2	Overage	***No Overage Plan:*** Reach out to those at risk due to high overage costs with no overage plan (but not for everyone)
5	Mou	***Gift Card Campaign:*** Instead of a new phone offer Gift Cards or Rebates for a new phone since heavy users are less likely to churn
6	Changem	***Reach Out Campaign:*** Pro-actively contact those who use is trending down and offer New Phone

Discussion Exercise
Develop a Proactive Retention Policy

Given everything you know what offer should you make to our four selected customers?

Variable	User #15747	User #29301	User #8695	User #34573
Predicted Adjusted Probability of Churn	4.5%	1.7%	5.7%	1.3%
LTV (with Adjust Churn)	$1,228	$2,508	$83	$2,100
Recommended Offer	?	?	?	?
LTV (with offer)	?	?	?	?
Gain in LTV	?	?	?	?

What will happen to LTV? How to compute these values?

What is the relationship between on offer and the probability of churn?

We can use either a logistic regression or a decision tree to make predictions about whether a customer will churn:

$$\Pr\left(Churn \mid characteristics\right) = p$$

Unfortunately, these predictions do not take into account the actions that we will take. What we want is the conditional probability of churn given all the characteristics we have used in Part 1, but also what is the probability of churn given a new offer:

$$\Pr\left(Churn \mid characteristics, offer\right) = p'$$

Notice our problem

Notice that in our data we have never made promotional offers to proactively retain customers

- *We are using a Data Based approach but we have never seen the target!*
- *We are using predictive models to understand, but we want to use them to make predictions*

Discussion Exercise

Solution: If offer is a characteristic in our model

- Use our model to evaluate what would happen to the user, say if we give them a new phone then Eqpdays=0.
- Need to know the cost of a NewPhone (say $200)
- Notice LTV goes from $1227 to $1552

*Solution: If offer is **not** a characteristic in our model*

- Give them a promotional offer worth $X and recompute churn probability and gain in LTV

	Current	New Phone
	1	1
Eqpdays	391	0
Retcall	1	1
Months	16	16
Overage	0	0
Mou	0	0
Changem	0	0
PromoOffer	$ -	$ -
Revenue	$ 60.33	$ 60.33
Churn	1	1
ChurnProb	69.9%	69.9%
Customer	1039199	1039199
Score	0.8408	0.4292

Churn and LTV calculations (using oversam

Pr(Churn)	70%	61%
MonthProfit	$ 60.33	$ 60.33
NewPhone	$ -	$ 200.00
PromoOffer	$ -	$ -
LTV	$ 86.19	$ (100.67)

Churn and LTV calculations (projected for or

A	0.0279	0.0242
B	0.5907	0.7729
AdjPr(Churn)	4.5%	3.0%
LTV (AdjPr)	$ 1,227.75	$ 1,552.77

Example of how to adjust churn given an offer

Here is an example using Tree Model A (simple tree):

$$\Pr\left(Churn \mid Eqpdays = 361, Months = 61\right) = p = 0.56$$

What if we give a free phone?

$$\Pr\left(Churn \mid Eqpdays = 0, Months = 61\right) = p' = 0.39$$

An exception is if our offer is related to an existing characteristic (e.g., "free phone" → "eqpdays=0")

What if we give a $50 gift card?

$$\Pr\left(Churn \mid Eqpdays = 361, Months = 61, Giftcard = \$50\right) = p' = ?$$

Solution B for predicting effect of promotions: Two methods for adjusting churn probability for Offers

Method A. Assume probability is multiplied by factor.

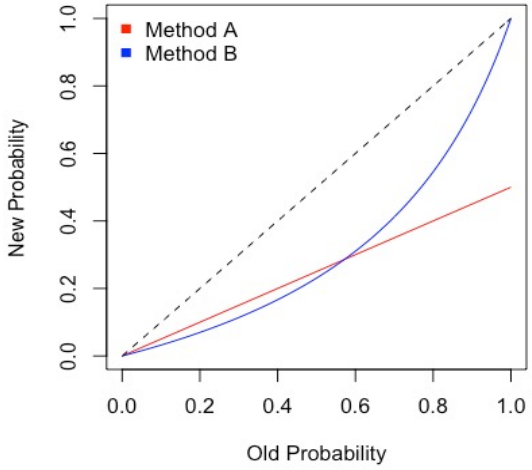
Example: Customer has 8% churn without any offer, but 50%x8%=4% with an offer of $50 gift card.

$$p' = \alpha \cdot p \quad \Rightarrow \quad \frac{p'}{1-p'} = \frac{\alpha p}{1-\alpha p} = \left(\alpha \frac{1-p}{1-\alpha p}\right)\left(\frac{p}{1-p}\right)$$

Method B. Assume that the odds ratio is multiplied by factor.

Example: Customer has odds ratio of 2:1 of churning without any offer, but 50%x2=1:1 with an offer of $50 gift card.

$$\frac{p'}{1-p'} = \beta \cdot \frac{p}{1-p} \quad \Rightarrow \quad p' = \frac{\beta}{1-p+p\beta} p$$



Comments

- Both methods are correct, Work with predictions from Tree or LR
- The challenge: factor is unknown

Solving Problem II of predicting effect of promotions:
Understanding Method B as new variable in logistic regression

You have estimated the following logistic regression model:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

However, we need a model that shows how the probability will change if we make an offer so we can add an offer term to model:

$$\ln\left(\frac{p'}{1-p'}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \gamma \cdot Offer$$

The "new" model is just the same as the first model with a new offer effect:

$$\frac{p'}{1-p'} = \frac{p}{1-p} \cdot \beta \quad \Rightarrow \quad \ln\left(\frac{p'}{1-p'}\right) = \ln\left(\frac{p}{1-p}\right) + \gamma \cdot Offer$$

- Notice we can add "offer term" to the model, but we cannot estimate it since we do not have any data on this variable.

Adjusting Churn Probability given Promotional Offer:
"What if…" simulator using logistic regression for new offers

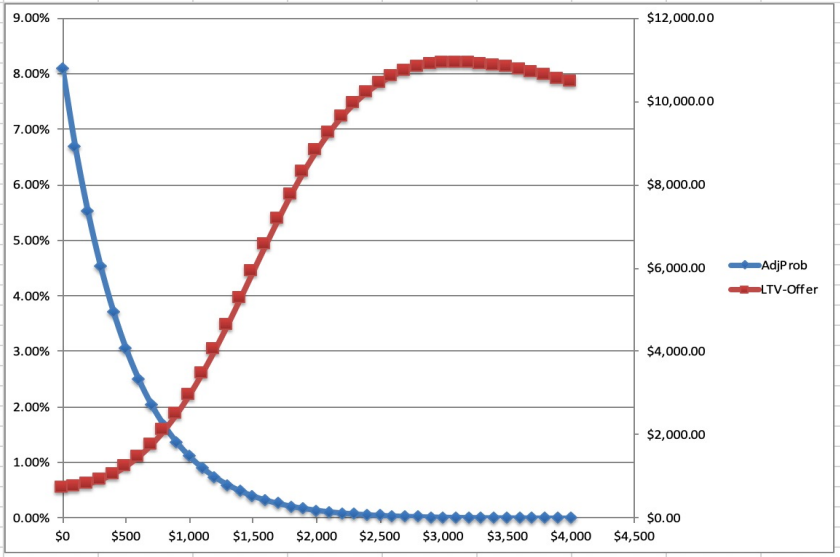
	A	B	C	D	E	F	G	H	I	J	K
1	**Simulator Model for a Specific User**										
2											
3											
4		**Estimate**			**Current**	**New Phone**	**Current + Promo**				
5	*(Intercept)*	-0.1620			1	1	1				
6	*Eqpdays*	0.0011		Eqpdays	391	0	391	Let's assume that the user is given a new equipment			
7	*Retcall*	0.7985		Retcall	1	1	1				
8	*Months*	-0.0130		Months	16	16	16				
9	*Overage*	0.0017		Overage	0	0	0				
10	*Mou*	-0.0002		Mou	0	0	0				
11	*Changem*	-0.0003		Changem	0	0	0				
12	*Promo*	-0.0021		PromoOffer	$ -	$ -	$ 200.00	Use PromoOffer given below			
13				Revenue	$ 60.33	$ 60.33	$ 60.33				
14				Churn	1	1	1				
15				ChurnProb	69.9%	69.9%	69.9%				
16				Customer	1039199	1039199	1039199				
17											
18				Score	0.8408	0.4292	0.4292				
19											
20				Churn and LTV calculations (using oversampled data)							
21				Pr(Churn)	70%	61%	61%				
22				MonthProfit	$ 60.33	$ 60.33	$ 60.33				
23				NewPhone	$ -	$ 200.00	$ -	Assume new phone costs $200			
24				PromoOffer	$ -	$ -	$ 200.00	Use Solver to find the optimal PromoOffer			
25				LTV	$ 86.19	$ (100.67)	$ (100.67)				
26											
27				Churn and LTV calculations (projected for original data)							
28				A	0.0279	0.0242	0.0242				
29				B	0.5907	0.7729	0.7729				
30				AdjPr(Churn)	4.5%	3.0%	3.0%				
31				LTV (AdjPr)	$ 1,227.75	$ 1,552.77	$ 1,552.72				

Alan Montgomery:
This is just a guess. Try different values. This value equates the value of a $200 new phone with a $200 gift card from Goal Seek.

Solving Problem 2 related to predicting effect of promotions: "What if…" simulator using logistic regression for optimal offer

	A	B	C
1	**How does an offer effect response?**		
2			
3	Init Prob	8.1%	
4	Offer ($)	$200.00	
5	Coefficient	-0.0021	
6	Adj Prob	5.5%	
7	InterestRate (Year)	5%	
8	Profit (Month)	$60.33	
9			
10			
11	Offer	AdjProb	LTV-Offer
12	$0	8.10%	$711.33
13	$100	6.69%	$751.95
14	$200	5.52%	$820.83
15	$300	4.54%	$922.68
16	$400	3.73%	$1,062.55
17	$500	3.05%	$1,245.66
18	$600	2.50%	$1,477.11
19	$700	2.04%	$1,761.51
20	$800	1.67%	$2,102.48
21	$900	1.36%	$2,502.02
22	$1,000	1.11%	$2,959.93
23	$1,100	0.91%	$3,473.25
24	$1,200	0.74%	$4,035.85
25	$1,300	0.60%	$4,638.35
26	$1,400	0.49%	$5,268.47
27	$1,500	0.40%	$5,911.77
28	$1,600	0.33%	$6,552.73
29	$1,700	0.27%	$7,175.99
30	$1,800	0.22%	$7,767.68
31	$1,900	0.18%	$8,316.31
32	$2,000	0.14%	$8,813.49
33	$2,100	0.12%	$9,254.05
34	$2,200	0.10%	$9,635.91
35	$2,300	0.08%	$9,959.59
36	$2,400	0.06%	$10,227.66
37	$2,500	0.05%	$10,444.08
38	$2,600	0.04%	$10,613.65
39	$2,700	0.03%	$10,741.56

Alan Montgomery:
This is just a guess. Try different values. This value equates the value of a $200 new phone with a $200 gift card from Goal Seek.



Adjusting Churn Probability given Promotional Offer: Using Method A

Variable	User #15747	User #8695	User #29301	User #34573
Predicted Probability of Churn	4.5%	5.7%	1.7%	1.3%
Recommended Offer	Large $100	Small $50	None	None
Expected Profit without offer	$ 60.33	$ 53.72	$ 5.00	$ 34.99
LTV	$1,228	$2,508	$83	$2,100
Predicted Probability of Churn with offer	25%x4.5%= 1.1%	50%x5.7%= 2.8%	0.7%	0.7%
LTV	$3,918.31	$154.53	$13,697.93	$7,314.91
Gain	$3,918.31-$1,227.75= +$2,950.56	$154.53-$82.56= +$21.97	0	0

Focus of Part 2

Method A: Multiply churn probability by "factor"

Effect of offers on Churn and LTV (projected for original data)				
Offer	$ 100	$ 50		$ -
SuccessRate	1.1%	2.8%	1.7%	1.3%
LTV (AdjSuccess)	$ 3,918.31	$ 154.53	$ 2,507.88	$ 2,100.17
LTV Change	$ 2,590.56	$ 21.97	$ -	$ -

Part 2 of the Assignment

1. Translate your model into a proactive retention campaign. Use your predictive model to design a strategy using your best model that you believe would increase retention (decrease churn rate) amongst mobile phone customers. Tell a story in plain English about why customers are leaving.
2. Recommend actions that can be taken proactively to keep customers that are likely to leave. Explain how your strategy relates to your model. Be specific in your recommendations about who to target, what to offer and how you will communicate the offer to the customers (phone, email, text, mail, or other).
3. What gain in profits or LTV would you expect from your proactive retention campaign? (Hint: you will need to make assumptions about the cost of your offers and how consumers will respond to potential promotional offers.)

Hints

Which consumers to target?

- All 40,000?
- Probably not all 40,000 since we want a *Pro-active campaign*, try to focus on those with high probabilities

What offer to make?

- Think about a small number of offers (say free phone, $10 off, no overage minutes), and evaluate effects on churn and LTV. Just select the best plan
- Remember that LTV measures long-term gains in profits, do you really care whether you think customer will stay 20 years instead of 15 years? Or do you care more about profits in next one year?

Logistic Regression Spreadsheet Simulator

Details on the Logistic Spreadsheets for Part 2

Understanding the Logistic "Model" Sheet

The "Model" sheet has the estimated parameters, and defines the named region "coefvec" (e.g. B2:B8) (see Blue)

All customers have been grouped into 21 clusters (see Green)

- The customers are clustered according why they are likely to churn (multiply variable by corresponding coefficient) – not clusters on raw data
- The first 20 clusters are the 25% of the customer base that is most likely to churn, the 21st cluster has the remaining 75% of the customer base.
- The fraction of customers in each cluster is given (both for unadjusted as well as adjusted).

	A	B	C	D	E	F	G	H	I
1		Estimate	Std. Error	z value	Pr(>|z|)		Cluster	PredFrac.ui	PredFrac
2	Intercept	-0.168	0.049	-3.449	0.001		1	0.0%	0.0%
3	Eqpdays	0.001	0.000	18.684	0.000		2	1.7%	1.9%
4	Retcall	0.766	0.071	10.796	0.000		3	0.3%	0.4%
5	Months	-0.016	0.002	-9.525	0.000		4	0.7%	1.0%
6	Refurb	0.344	0.038	8.975	0.000		5	0.9%	1.1%
7	Uniqsubs	0.195	0.025	7.775	0.000		6	1.8%	2.2%
8	Mailres	-0.210	0.028	-7.573	0.000		7	3.4%	4.0%
9	Overage	0.002	0.000	9.241	0.000		8	2.0%	2.5%
10	Mou	0.000	0.000	-7.348	0.000		9	1.4%	1.8%
11	Creditde	-0.238	0.042	-5.678	0.000		10	0.4%	0.5%
12	Actvsubs	-0.182	0.035	-5.227	0.000		11	1.0%	1.3%
13							12	2.5%	3.0%
14							13	1.5%	1.9%
15							14	0.3%	0.3%
16							15	1.0%	1.3%
17							16	0.0%	0.0%
18							17	2.8%	3.3%
19							18	1.0%	1.2%
20							19	0.7%	0.8%
21							20	1.6%	2.2%
22							21	75.0%	69.4%

Model parameters (Do not change)

Cluster analysis is done on data scaled by the corresponding estimate (e.g., cluster .001*eqpdays not eqpdays). This gives 20 clusters of customers that churn for similar reasons.

Logistic “Simulator” Sheet provides “What if…” alculations for your model (change an input and Pr(Churn) and LTV is updated)

The “Simulator” sheet has one prototypical customer for each cluster.

- The cluster analysis weights the variables by their impact on score (not clustering raw data, but the variable times its coefficient)
- There are 20 clusters + one for all other observations
- A prototype is chosen by choosing the customer that is closest to the Centroid of the cluster.

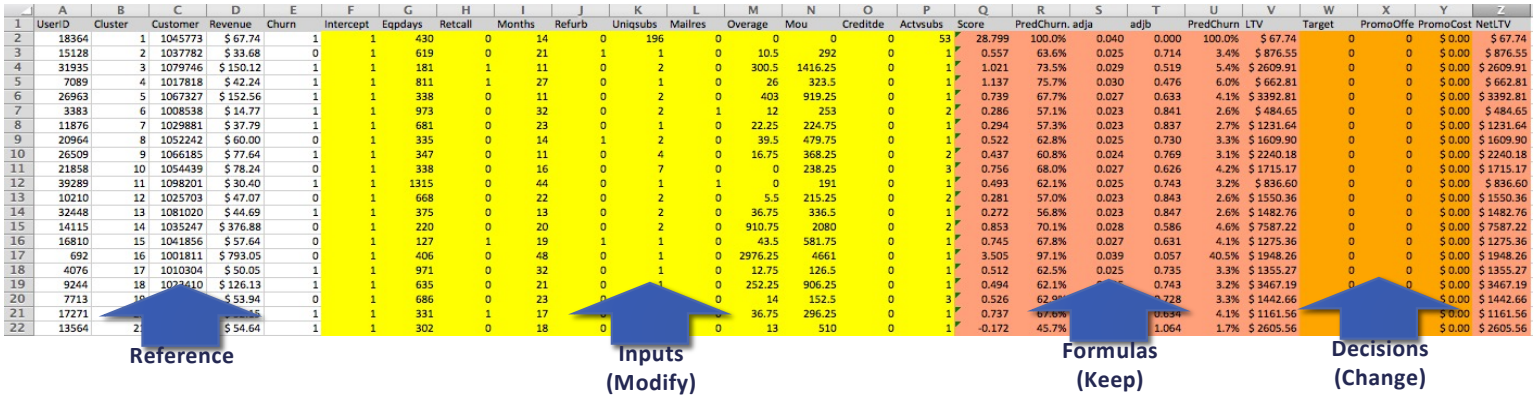
Yellow area gives input data needed for model (must be in same order as coefvec)

White area gives extra information (like ID, Cluster, Customer, Revenue, Churn)

Red area gives computed values from the model or profit calculations. You can copy and paste these formulas (each predictions uses the values on the row, keep inputs in same order)

Orange area is for promotional campaign decisions (Target, PromoOffer, PromoCost)

Notice that the Logistic Regression model will automatically recompute the probability of each user if you change an input



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	UserID	Cluster	Customer	Revenue	Churn	Intercept	Eqpdays	Retcall	Months	Refurb	Uniqsubs	Mailres	Overage	Mou	Creditde	Actvsubs	Score	PredChurn	adja	adjb	PredChurn	LTV	Target	PromoOffe	PromoCost	NetLTV
2	18364	1	1045773	$ 67.74	1	1	430	0	14	0	196	0	0	0	0	53	28.799	100.0%	0.040	0.000	100.0%	$ 67.74	0	0	$ 0.00	$ 67.74
3	15128	2	1037782	$ 33.68	0	1	619	0	21	1	1	0	10.5	292	0	1	0.557	63.6%	0.025	0.714	3.4%	$ 876.55	0	0	$ 0.00	$ 876.55
4	31935	3	1079746	$ 150.12	1	1	181	1	11	0	2	0	300.5	1416.25	0	1	1.021	73.5%	0.029	0.519	5.4%	$ 2609.91	0	0	$ 0.00	$ 2609.91
5	7089	4	1017818	$ 42.24	1	1	811	1	27	0	1	0	26	323.5	0	1	1.137	75.7%	0.030	0.476	6.0%	$ 662.81	0	0	$ 0.00	$ 662.81
6	26963	5	1067327	$ 152.56	1	1	338	0	11	0	2	0	403	919.25	0	1	0.739	67.7%	0.027	0.633	4.1%	$ 3392.81	0	0	$ 0.00	$ 3392.81
7	3383	6	1008538	$ 14.77	1	1	973	0	32	0	2	1	12	253	0	2	0.286	57.1%	0.023	0.841	2.6%	$ 484.65	0	0	$ 0.00	$ 484.65
8	11876	7	1029881	$ 37.79	1	1	681	0	23	0	1	0	22.25	224.75	0	1	0.294	57.3%	0.023	0.837	2.7%	$ 1231.64	0	0	$ 0.00	$ 1231.64
9	20964	8	1052242	$ 60.00	0	1	335	0	14	1	2	0	39.5	479.75	0	1	0.522	62.8%	0.025	0.730	3.3%	$ 1609.90	0	0	$ 0.00	$ 1609.90
10	26509	9	1066185	$ 77.64	1	1	347	0	11	0	4	0	16.75	368.25	0	2	0.437	60.8%	0.024	0.769	3.1%	$ 2240.18	0	0	$ 0.00	$ 2240.18
11	21858	10	1054439	$ 78.24	0	1	338	0	16	0	7	0	0	238.25	0	3	0.756	68.0%	0.027	0.626	4.2%	$ 1715.17	0	0	$ 0.00	$ 1715.17
12	39289	11	1098201	$ 30.40	1	1	1315	0	44	0	1	1	0	191	0	1	0.493	62.1%	0.025	0.743	3.2%	$ 836.60	0	0	$ 0.00	$ 836.60
13	10210	12	1025703	$ 47.07	0	1	668	0	22	0	2	0	5.5	215.25	0	2	0.281	57.0%	0.023	0.843	2.6%	$ 1550.36	0	0	$ 0.00	$ 1550.36
14	32448	13	1081020	$ 44.69	1	1	375	0	13	0	2	0	36.75	336.5	0	1	0.272	56.8%	0.023	0.847	2.6%	$ 1482.76	0	0	$ 0.00	$ 1482.76
15	14115	14	1035247	$ 376.88	0	1	220	0	20	0	2	0	910.75	2080	0	2	0.853	70.1%	0.028	0.586	4.6%	$ 7587.22	0	0	$ 0.00	$ 7587.22
16	16810	15	1041856	$ 57.64	0	1	127	1	19	1	1	0	43.5	581.75	0	1	0.745	67.8%	0.027	0.631	4.1%	$ 1275.36	0	0	$ 0.00	$ 1275.36
17	692	16	1001811	$ 793.05	0	1	406	0	48	0	1	0	2976.25	4661	0	1	3.505	97.1%	0.039	0.057	40.5%	$ 1948.26	0	0	$ 0.00	$ 1948.26
18	4076	17	1010304	$ 50.05	1	1	971	0	32	0	1	0	12.75	126.5	0	1	0.512	62.5%	0.025	0.735	3.3%	$ 1355.27	0	0	$ 0.00	$ 1355.27
19	9244	18	1023410	$ 126.13	1	1	635	0	21	0	1	0	252.25	906.25	0	1	0.494	62.1%	[illegible]	0.743	3.2%	$ 3467.19	0	0	$ 0.00	$ 3467.19
20	7713	19	[illegible]	$ 53.94	0	1	686	0	23	0	[illegible]	0	14	152.5	0	3	0.526	62.9%	[illegible]	0.728	3.3%	$ 1442.66	0	0	$ 0.00	$ 1442.66
21	17271	[illegible]	[illegible]	[illegible]	1	1	331	1	17	0	[illegible]	0	36.75	296.25	0	1	0.737	67.6%	[illegible]	0.634	4.1%	$ 1161.56	0	0	$ 0.00	$ 1161.56
22	13564	2[illegible]	[illegible]	$ 54.64	1	1	302	0	18	0	[illegible]	0	13	510	0	1	-0.172	45.7%	[illegible]	1.064	1.7%	$ 2605.56	0	0	$ 0.00	$ 2605.56

Data Columns have same values as Part 1 in R, the additional columns provide calculations or decisions

Column	Description
UserID	The row number from the original dataset
Cluster	Cluster that customer has been assigned to
Score	Sum of the product of the coefficients and data
PredChurn.unadj	Predicted Churn (in oversampled data)
PredFrac.unadj	Fraction of Users similar to this user (cluster size)
adja	Adjustment A for computing adjusted probability
adjb	Adjustment B for computing adjusted probability
PredChurn	Predicted Churn (in population)
PredFrac	Fraction of Users similar to this user (population)
LTV	Customer Lifetime Value (assumes 5% discount rate, Revenue = Profits, and depends on PredChurn)
Target	Set to either =1 if you want to target or =0 if you do not want to target
PromoOffer	Set to cost in $ of monetary offer/rebate/gift card (if any)
PromoCost	Set to cost in $ of your promotional offer (if any)
NetLTV	=LTV-PromoCost

Set these values

Understanding the Logistic "Data" Sheet

The entire dataset of 40,000 records and the data used in the model is given on this Sheet. Gives Cluster for every customer, and initial Churn prediction from model (static)

Use this dataset if you want to:

- Understand more about the customers (Pivot tables)
- Use Pivot tale to compute Centroids (average by Cluster)
- Look at other customers in the clusters
- Copy customers into the Simulator sheet

Warning: This Sheet makes the file large

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	UserID	Cluster	Customer	Revenue	Churn	Intercept	Eqpdays	Retcall	Months	Refurb	Uniqsubs	Mailres	Overage	Mou	Creditde	Actvsubs	PredChurn	Target	PromoOffe	PromoCost
2	1	21	1000001	$ 24.00	1	1	361	1	61	0	2	1	0	219.25	0	1	2.1%	0	$ 0.00	$ 0.00
3	2	11	1000003	$ 16.99	1	1	1504	0	58	0	1	1	0	10.25	0	1	3.4%	0	$ 0.00	$ 0.00
4	3	11	1000004	$ 38.00	0	1	1812	0	60	0	1	0	0	7.5	0	1	5.8%	0	$ 0.00	$ 0.00
5	4	21	1000005	$ 55.23	0	1	434	0	57	0	1	0	0	570.5	0	1	1.0%	0	$ 0.00	$ 0.00
6	5	21	1000007	$ 17.15	1	1	852	0	53	0	2	1	0	0	0	2	1.7%	0	$ 0.00	$ 0.00
7	6	21	1000008	$ 38.05	0	1	231	0	53	0	1	0	0	682.5	0	1	0.8%	0	$ 0.00	$ 0.00
8	7	18	1000009	$ 97.34	0	1	700	0	55	0	1	0	419.5	1039	0	1	2.7%	0	$ 0.00	$ 0.00
9	8	21	1000012	$ 35.30	0	1	544	0	53	0	3	0	0	24.25	0	3	1.5%	0	$ 0.00	$ 0.00
10	9	21	1000013	$ 81.00	0	1	388	0	55	0	1	0	0	1056.25	0	1	0.9%	0	$ 0.00	$ 0.00
11	10	21	1000017	$ 63.02	0	1	530	0	57	0	1	1	5.75	440.25	0	1	1.0%	0	$ 0.00	$ 0.00
12	11	21	1000021	$ 172.44	1	1	143	1	58	0	2	1	362.25	1978	0	1	2.0%	0	$ 0.00	$ 0.00
13	12	21	1000022	$ 60.19	1	1	162	1	54	0	1	1	0	1424.5	1	1	0.9%	0	$ 0.00	$ 0.00
14	13	21	1000026	$ 29.99	0	1	573	0	54	0	3	1	0	47	0	3	1.2%	0	$ 0.00	$ 0.00
15	14	21	1000031	$ 24.49	1	1	776	0	58	0	2	0	9.5	41.75	0	1	2.2%	0	$ 0.00	$ 0.00
16	15	21	1000034	$ 21.15	0	1	138	0	58	0	1	1	14	45.5	0	1	0.7%	0	$ 0.00	$ 0.00
17	16	21	1000037	$ 30.00	0	1	630	0	56	0	2	1	0	155.75	0	2	1.2%	0	$ 0.00	$ 0.00
18	17	21	1000040	$ 33.48	1	1	179	0	54	1	2	1	0.25	195.5	0	1	1.2%	0	$ 0.00	$ 0.00
19	18	11	1000044	$ 16.14	1	1	1661	0	55	0	1	1	0	3.75	0	1	4.3%	0	$ 0.00	$ 0.00
20	19	21	1000046	$ 57.98	1	1	864	0	55	0	2	0	3	683.75	0	1	2.1%	0	$ 0.00	$ 0.00
21	20	21	1000048	$ 78.29	1	1	151	0	53	1	5	1	0	851.5	0	4	1.1%	0	$ 0.00	$ 0.00

Inputs (Modify)

Static Model Prediction

Decisions (Change)

Understanding the Logistic “Pilot” Sheet

A test dataset of 31,047 records for all the reference and input variables used in your model. Same format as “Data”, but this sample is not oversampled as the Data set is. Fill this sheet with the appropriate values of your campaign.

Use this dataset to provide the decisions that you will make for your proactive retention campaign.

- Set Target=1 if you want to target this customer, =0 If not
- Set PromoOffer=$ if you are providing cash (e.g., PromoOffer=50 would be a $50 cash rebate, and PromoOffer=0 is nothing)
- Set PromoCost=$ to tally the total promotional cost (e.g., new phone that sets Eqpdays=0 implies PromoCost=50, notice PromoCost>=PromoOffer)

Copy customers into the Simulator sheet

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	UserID	Cluster	Customer	Revenue	Churn	Intercept	Eqpdays	Retcall	Months	Refurb	Uniqsubs	Mailres	Overage	Mou	Creditde	Actvsubs	PredChurn	Target	PromoOffe	PromoCost
2	40001	21	1000002	$ 57.49	#N/A	1	240	0	56	0	1	1	22.75	482.75	1	1	0.6%	0	$ 0.00	$ 0.00
3	40002	21	1000006	$ 82.28	#N/A	1	458	0	59	0	2	1	0	1312.25	0	2	0.7%	0	$ 0.00	$ 0.00
4	40003	21	1000010	$ 31.66	#N/A	1	601	0	57	0	2	1	0	25.5	0	2	1.2%	0	$ 0.00	$ 0.00
5	40004	21	1000011	$ 62.13	#N/A	1	464	0	59	0	2	1	0	97.5	0	2	1.0%	0	$ 0.00	$ 0.00
6	40005	21	1000014	$ 25.23	#N/A	1	354	0	53	0	2	0	0	2.5	0	2	1.2%	0	$ 0.00	$ 0.00
7	40006	21	1000015	$ 212.52	#N/A	1	199	0	59	0	5	1	249.5	1971.5	0	1	1.5%	0	$ 0.00	$ 0.00
8	40007	21	1000016	$ 42.57	#N/A	1	697	0	55	0	2	0	6	270.5	0	2	1.6%	0	$ 0.00	$ 0.00
9	40008	21	1000018	$ 35.59	#N/A	1	48	0	59	0	3	1	16	153	0	2	0.7%	0	$ 0.00	$ 0.00
10	40009	21	1000019	$ 55.27	#N/A	1	408	0	52	0	1	1	0	1212.75	0	1	0.8%	0	$ 0.00	$ 0.00
11	40010	21	1000020	$ 50.97	#N/A	1	253	0	56	0	1	1	2.5	162	0	1	0.8%	0	$ 0.00	$ 0.00
12	40011	21	1000025	$ 25.49	#N/A	1	752	0	52	0	2	0	0.75	299.5	0	2	1.8%	0	$ 0.00	$ 0.00
13	40012	21	1000028	$ 37.66	#N/A	1	452	1	58	0	1	0	4.75	971.75	0	1	2.0%	0	$ 0.00	$ 0.00
14	40013	21	1000030	$ 30.26	#N/A	1	319	0	52	0	1	0	0	34.5	0	1	1.1%	0	$ 0.00	$ 0.00
15	40014	11	1000032	$ 30.00	#N/A	1	1395	0	54	1	1	0	0	94	0	1	5.3%	0	$ 0.00	$ 0.00
16	40015	21	1000033	$ 35.55	#N/A	1	628	0	52	0	1	1	0	138.75	0	1	1.3%	0	$ 0.00	$ 0.00
17	40016	6	1000035	$ 28.50	#N/A	1	917	0	56	0	2	0	0	65.75	0	1	2.6%	0	$ 0.00	$ 0.00
18	40017	21	1000036	$ 99.91	#N/A	1	258	0	54	0	1	0	97.25	1193.75	0	1	0.9%	0	$ 0.00	$ 0.00
19	40018	21	1000041	$ 82.16	#N/A	1	107	0	57	0	3	1	100.75	660	0	3	0.7%		$ 0.00	$ 0.00
20	40019	21	1000042	$ 30.00	#N/A	1	616			0	2	1	0	56	0					$ 0.00
21	40020	11	1000058	$ 20.18	#N/A	1	1536	0		0	1	0	0	1.25	0	1	.9%		0.00	$ 0.00

Inputs (Modify)

Static Model Prediction

Decisions (Change)

More Comments about the Logistic Spreadsheet

There are several named regions on the “Model” sheet to make it easier to reference values to construct the model

These regions are used for computing the Score=“MMULT(F2:P2,coefvec)”

Create a “Scenario” column and change the covariates and costs appropriately

Or create new “Scenario” tabs with different covariates and costs

Use the Data to learn more about the customers within a Cluster

Column	Description
CoefVec	Vector of Logistic Regression Coefficients
ClusterIndex	Cluster Index (e.g., 1 to 21)
PredFrac.unadj	Unadjusted Fraction of Observations for each Leaf
PredFrac	Adjusted Fraction of Observations for each Leaf

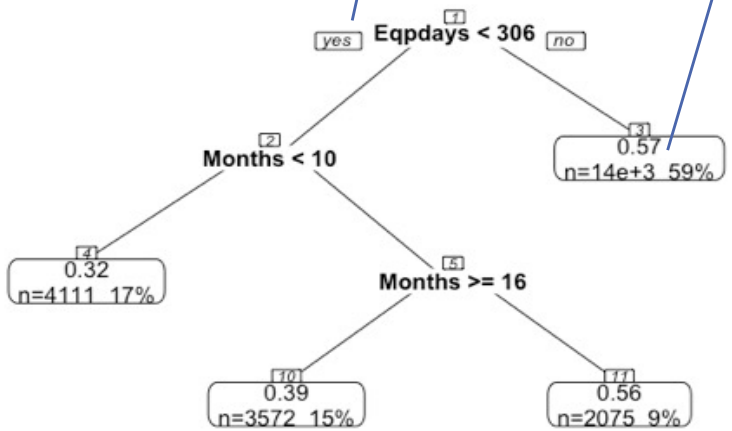
Decision Tree Spreadsheet Simulator

Details on the Tree Spreadsheets for Part 2

Understanding the Tree "Model"

The "Model" sheet shows how the tree works for one user

	A	B	C	D	E	F	G	H	I	J	K	L
1		Values		Subrule	formula		LeafRule	PredChurn.	PredFrac.ur	PredChurn	PredFrac	formula
2	Eqpdays	361		CONDL1	FALSE		3	56.5%	59.3%	2.6%	51.6%	TRUE
3	Months	61		CONDL2	FALSE		4	32.1%	17.1%	1.0%	22.9%	FALSE
4				CONDL3	TRUE		10	38.9%	14.9%	1.3%	17.9%	FALSE
5				CONDR1	TRUE		11	56.1%	8.6%	2.5%	7.6%	FALSE
6				CONDR2	TRUE							
7				CONDR3	FALSE			formula				
8							LeafIndex	1				
9							LeafRule	3				
10							PredChurn.	0.5653				
							PredChurn	0.025853				



- The data is in red and must be named region according to variable
- The branches in the tree are encoded as Conditions (e.g., CONDL1 is branch left on level 1?). Do not change blue rules.
- The green area is the lookup table for the probability and fraction (must be named)
- Yellow has the predictions. In this example the User would have a 56.5% unadjusted probability of churn and 2.6% adjusted probability of churn

Tree “Simulator” Sheet provides “What if…” calculations for your model (change an input and Pr(Churn) and LTV is updated)

The “Simulator” spreadsheet has one prototypical customer from each branch (e.g., the one closest to the average customer within that leaf). The Data for each user is in Yellow (change it if you want to ask “what if…”), the White area gives extra information about the customer

	A	B	C	D	E	F
1	UserID	Customer	Revenue	Churn	Eqpdays	Months
2	14515	1036284	$ 14.81	0	545	21
3	28834	1072124	$ 78.49	0	211	8
4	6093	1015282	$ 40.59	0	146	28
5	23397	1058322	$ 89.22	1	136	13

The Rules are in Blue (do not change)

G	H	I	J	K	L	M	N	O	P
CONDL1	CONDL2	CONDL3	CONDR1	CONDR2	CONDR3	RULE3	RULE4	RULE10	RULE11
FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE
TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE

The Red has the predictions (do not change). These values are computed only using the data on this spreadsheet, and can be copy/pasted with formulas intact. The calculations are in each row. Orange (decisions to make) is for the settings of your promotional campaign.

Q	R	S	T	U	V	W	X	Y	Z	AA
LeafIndex	LeafRule	PredChurn.	adja	adjb	PredChurn	LTV	Target	PromoOffe	PromoCost	NetLTV
1	3	56.5%	0.023	0.852	2.6%	$ 495.31	0	$ 0.00	$ 0.00	$ 495.31
2	4	32.1%	0.013	1.330	1.0%	$ 5737.22	0	$ 0.00	$ 0.00	$ 5737.22
3	10	38.9%	0.016	1.197	1.3%	$ 2395.28	0	$ 0.00	$ 0.00	$ 2395.28
4	11	56.1%	0.022	0.860	2.5%	$ 3023.85	0	$ 0.00	$ 0.00	$ 3023.85

Understanding the Tree "Data" Sheet

The entire dataset of 40,000 records and the data used in the model is given on this Sheet.

Gives LeafIndex for every customer, and initial Churn prediction from model (static)

Use this dataset if you want to:

- Understand more about the customers (Pivot tables)
- Use Pivot tale to compute Centroids (average by Cluster)
- Look at other customers in the clusters
- Copy customers into the Simulator sheet

Warning: This Sheet makes the file large

	A	B	C	D	E	F	G	H	I	J	K	L
1	UserID	LeafIndex	LeafRule	Customer	Revenue	Churn	Eqpdays	Months	PredChurn	Target	PromoOffe	PromoCost
2	1	1	3	1000001	$ 24.00	1	361	61	2.6%	0	$ 0.00	$ 0.00
3	2	1	3	1000003	$ 16.99	1	1504	58	2.6%	0	$ 0.00	$ 0.00
4	3	1	3	1000004	$ 38.00	0	1812	60	2.6%	0	$ 0.00	$ 0.00
5	4	1	3	1000005	$ 55.23	0	434	57	2.6%	0	$ 0.00	$ 0.00
6	5	1	3	1000007	$ 17.15	1	852	53	2.6%	0	$ 0.00	$ 0.00
7	6	3	10	1000008	$ 38.05	0	231	53	1.3%	0	$ 0.00	$ 0.00
8	7	1	3	1000009	$ 97.34	0	700	55	2.6%	0	$ 0.00	$ 0.00
9	8	1	3	1000012	$ 35.30	0	544	53	2.6%	0	$ 0.00	$ 0.00
10	9	1	3	1000013	$ 81.00	0	388	55	2.6%	0	$ 0.00	$ 0.00
11	10	1	3	1000017	$ 63.02	0	530	57	2.6%	0	$ 0.00	$ 0.00
12	11	3	10	1000021	$ 172.44	1	143	58	1.3%	0	$ 0.00	$ 0.00
13	12	3	10	1000022	$ 60.19	1	162	54	1.3%	0	$ 0.00	$ 0.00
14	13	1	3	1000026	$ 29.99	0	573	54	2.6%	0	$ 0.00	$ 0.00
15	14	1	3	1000031	$ 24.49	1	776	58	2.6%	0	$ 0.00	$ 0.00
16	15	3	10	1000034	$ 21.15	0	138	58	1.3%	0	$ 0.00	$ 0.00
17	16	1	3	1000037	$ 30.00	0	630	56	2.6%	0	$ 0.00	$ 0.00
18	17	3	10	1000040	$ 33.48	1	179	54	1.3%	0	$ 0.00	$ 0.00
19	18	1	3	1000044	$ 16.14	1	1661	55	2.6%	0	$ 0.00	$ 0.00
20	19	1	3	1000046	$ 57.98	1		55				$ 0.00

Inputs (Modify) | Static Model Prediction | Decisions (Change)

Understanding the Tree "Pilot" Sheet

A test dataset of 31,047 records for all the reference and input variables used in your model. Same format as "Data", but this sample is not oversampled as the Data set is. Fill this sheet with the appropriate values of your campaign.

Use this dataset to provide the decisions that you will make for your proactive retention campaign.

- Set Target=1 if you want to target this customer, =0 If not
- Set PromoOffer=$ if you are providing cash (e.g., PromoOffer=50 would be a $50 cash rebate, and PromoOffer=0 is nothing)
- Set PromoCost=$ to tally the total promotional cost (e.g., new phone that sets Eqpdays=0 implies PromoCost=50, notice PromoCost>=PromoOffer)

Copy customers into the Simulator sheet

	A	B	C	D	E	F	G	H	I	J	K	L
1	UserID	LeafIndex	LeafRule	Customer	Revenue	Churn	Eqpdays	Months	PredChurn	Target	PromoOffe	PromoCost
2	1	1	3	1000001	$ 24.00	1	361	61	2.6%	0	$ 0.00	$ 0.00
3	2	1	3	1000003	$ 16.99	1	1504	58	2.6%	0	$ 0.00	$ 0.00
4	3	1	3	1000004	$ 38.00	0	1812	60	2.6%	0	$ 0.00	$ 0.00
5	4	1	3	1000005	$ 55.23	0	434	57	2.6%	0	$ 0.00	$ 0.00
6	5	1	3	1000007	$ 17.15	1	852	53	2.6%	0	$ 0.00	$ 0.00
7	6	3	10	1000008	$ 38.05	0	231	53	1.3%	0	$ 0.00	$ 0.00
8	7	1	3	1000009	$ 97.34	0	700	55	2.6%	0	$ 0.00	$ 0.00
9	8	1	3	1000012	$ 35.30	0	544	53	2.6%	0	$ 0.00	$ 0.00
10	9	1	3	1000013	$ 81.00	0	388	55	2.6%	0	$ 0.00	$ 0.00
11	10	1	3	1000017	$ 63.02	0	530	57	2.6%	0	$ 0.00	$ 0.00
12	11	3	10	1000021	$ 172.44	1	143	58	1.3%	0	$ 0.00	$ 0.00
13	12	3	10	1000022	$ 60.19	1	162	54	1.3%	0	$ 0.00	$ 0.00
14	13	1	3	1000026	$ 29.99	0	573	54	2.6%	0	$ 0.00	$ 0.00
15	14	1	3	1000031	$ 24.49	1	776	58	2.6%	0	$ 0.00	$ 0.00
16	15	3	10	1000034	$ 21.15	0	138	58	1.3%	0	$ 0.00	$ 0.00
17	16	1	3	1000037	$ 30.00	0	630	56	2.6%	0	$ 0.00	$ 0.00
18	17	3	10	1000040	$ 33.48	1	17	54	1.3%	0	$ 0.00	$ 0.00
19	18	1	3	1000044	$ 16.14	1		55			00	$ 0.00
20	19	1	3	1000046	$ 57.98	1		55	%		0.00	$ 0.00

Inputs (Modify) **Static Model Prediction** **Decisions (Change)**

Data Columns have same values as Part 1 in R, the additional columns provide calculations or decisions

Column	Description
UserID	The row number from the original dataset
LeafIndex	Row number of the Rule (like a cluster)
LeafRule	Leaf number from rpart of the prediction
PredChurn.unadj	Predicted Churn (in oversampled data)
PredFrac.unadj	Fraction of Users similar to this user (leaf size)
adja	Adjustment A for computing adjusted probability
adjb	Adjustment B for computing adjusted probability
PredChurn	Predicted Churn (in population)
PredFrac	Fraction of Users similar to this user (population)
LTV	Customer Lifetime Value (assumes 5% discount rate, Revenue = Profits, and depends on PredChurn)
Target	Set to either =1 if you want to target or =0 if you do not want to target
PromoOffer	Set to cost in $ of monetary offer/rebate/gift card (if any)
PromoCost	Set to cost in $ of your promotional offer (if any)
NetLTV	=LTV-PromoCost

Set these values

More comments about the Tree Spreadsheet

There are several named regions on the “Model” sheet to make it easier to reference values to construct the model

The only named regions that are used in “Simulator” are LeafVec and UnadjProbVec, all the other regions only pertain to the simulation on the “Model” sheet

Create a “Scenario” column and change the covariates and costs appropriately

Or create new “Scenario” tabs with different covariates and costs

Use the Data to learn more about the customers within a Leaf

Column	Description
RulesVec	List of Conditions used in Tree
LeafVec	**Named region of the Leaf Node numbers**
UnadjProbVec	**Unadjusted Probability for each Leaf**
UnadjFracVec	Unadjusted Fraction of Observations for each Leaf
AdjProbVec	Adjusted Probability for each Leaf
AdjFracVec	Adjusted Fraction of Observations for each Leaf

Appendix

More technical details

Solving Problem 2 related to predicting effect of promotions: Alternative Form for Method B

We can use either a logistic regression or a decision tree to make predictions about whether a customer will churn:

$$\Pr\left(Churn \mid characteristics\right) = p$$

Unfortunately, these predictions do not take into account the actions that we will take. What we want is the conditional probability of churn given all the characteristics we have used in Part 1, but also what is the probability of churn given a new offer:

$$\Pr\left(Churn \mid characteristics, offer\right) = p'$$

The idea is to build a new "logistic regression" that takes the original churn probabilities and adjusts for our offer:

$$\Pr\left(Churn \mid characteristics, offer\right) = \frac{1}{1+\exp\left\{-\left(\ln\left(\frac{p}{1-p}\right)+\gamma\cdot Offer\right)\right\}}$$

How do we estimate the response (ϒ)? And recommend offer (Offer)?

$$= \frac{1}{1+\left(\frac{p}{1-p}\cdot Offer^{\gamma}\right)^{-1}} = \frac{p}{p(1-Offer^{-\gamma})+Offer^{-\gamma}}$$

Solving Problem 2 of predicting promotional effect
Alternative form of Method B

You have estimated the following logistic regression model:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

However, we need a model that shows how the probability will change if we make an offer:

$$\ln\left(\frac{p'}{1-p'}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \gamma \cdot Offer$$

The "new" model is just the same as the first model with a new offer effect:

$$\ln\left(\frac{p'}{1-p'}\right) = \ln\left(\frac{p}{1-p}\right) + \gamma \cdot Offer$$

Alternatively we can write the model in terms of the probability instead of the log of the odds ratio:

$$p' = \frac{1}{1+\exp\left\{-\left(\ln\left(\frac{p}{1-p}\right) + \gamma \cdot Offer\right)\right\}} = \frac{1}{1+\left(\frac{p}{1-p}\cdot Offer^{\gamma}\right)^{-1}} = \frac{p}{p(1-Offer^{-\gamma}) + Offer^{-\gamma}}$$

Understanding the Part2 Script

(Optional, just use the Excel spreadsheet that corresponds with your model)

The goal of the script is to enable you to create an Excel simulator spreadsheet from the morning so that you can evaluate selected users and ask “what if” questions

The users selected are meant to prototypical users and give a cross-section. Every user has a weight (or predicted fraction of the total population that they represent), this enables you to take your predictions to the population level.

- Tree model: Computes the leaf nodes for each observation, and then finds the observation that is closest to the mean of all other observations in that group
- Logistic model: Computes the partial score for each variable (e.g., what is the contribution of the variable as weighted by its coefficient to the score) and clusters the users. The prototype is the observation closest to the centroid.

Creates other useful predictions like adjusted probability of churn and LTV

Outputs the selected users along with Excel formulas to compute the probability of churn and LTV. You can evaluate alternative scenarios (e.g., what if the user had a new phone) using the spreadsheet.

- There is CSV output which requires you to require named regions, but if possible use the already generated Excel spreadsheets
- If you want to generate the Excel spreadsheets directly you need to install openxlsx. This package also requires a zip application which you need to install by installing Rtools. After installing Rtools make sure shell(“zip”) works. See https://cran.r-project.org/web/packages/openxlsx/README.html