Assignment-5 (Churn Prediction)

Group11

Initially we checked for the missing values and in cases where missing values were more than 60% we dropped those variables. Also, checked the missing frequency of categorical variables and dropped them.

Finally the variables we dropped are – rtcount, rmcalls, rmmou, rmrev, REF_QTY, tot_ret, tot_acpt, pre_hnd,_price last_swap, occu1, educ1, numbcars, retdays.

We then checked for the percentage mean difference between the churn and non churn variables and selected top 20 variables with high mean difference as high diff could be a significant reason for churn.

After selecting the top 20 we performed correlation analysis to make sure no two explanatory variables are highly correlated.

The final model consists of the following variables:

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > Chi Sq
Intercept		1	-0.3824	0.0268	203.5922	<.0001
refurb_new	N	1	-0.1746	0.00984	314.8635	<.0001
prizm_social_one	C	1	-0.0270	0.0147	3.3853	0.0658
prizm_social_one	R	1	0.0999	0.0241	17.1821	<.0001
prizm_social_one	S	1	-0.0529	0.0121	18.9868	<.0001
prizm_social_one	T	1	0.0212	0.0153	1.9142	0.1665
new_cell	N	1	-0.00476	0.0136	0.1222	0.7266
new_cell	U	1	0.0199	0.00999	3.9877	0.0458
months		1	-0.0112	0.000844	174.8840	<.0001
drop_blk_Range		1	0.00114	0.000455	6.2280	0.0126
actvsubs		1	-0.1636	0.0179	83.1548	<.0001
change_mou		1	-0.00009	0.000044	3.7947	0.0514
roam_Mean		1	0.00809	0.00258	9.8482	0.0017
roam_Range		1	-0.00199	0.000771	6.6517	0.0099
eqpdays		1	0.00130	0.000032	1594.5555	<.0001
mtrcycle		1	0.1100	0.0562	3.8272	0.0504
uniqsubs		1	0.2177	0.0131	276.1998	<.0001
mou_Range		1	0.000038	0.000023	2.5819	0.1081
ovrmou_Range		1	0.000533	0.000101	28.0434	<.0001
ovrmou_Mean		1	-0.00028	0.000144	3.8994	0.0483
rev_Range		1	0.000106	0.000195	0.2959	0.5865
change_mou*eqpdays		1	-6.26E-7	1.37E-7	20.8624	<.0001

Q1. Include a table of coefficients, t-values, and odds ratio. Interpret the logistic output explaining AIC/BIC, meaning of coefficients, significance, prediction accuracy (percent concordance), odds-ratios etc.

ODDS RATIO :

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
refurb_new N vs R	0.705	0.679	0.733
prizm_social_one C vs U	1.014	0.974	1.056
prizm_social_one R vs U	1.151	1.081	1.227
prizm_social_one S vs U	0.988	0.955	1.023
prizm_social_one T vs U	1.064	1.020	1.110
new_cell N vs Y	1.010	0.964	1.059
new_cell U vs Y	1.036	1.000	1.072
months	0.989	0.987	0.991
drop_blk_Range	1.001	1.000	1.002
actvsubs	0.849	0.820	0.879
roam_Mean	1.008	1.003	1.013
roam_Range	0.998	0.997	1.000
mtrcycle	1.116	1.000	1.246
uniqsubs	1.243	1.212	1.276
mou_Range	1.000	1.000	1.000
ovrmou_Range	1.001	1.000	1.001
ovrmou_Mean	1.000	0.999	1.000
rev_Range	1.000	1.000	1.000

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	59.5	Somers' D	0.190
Percent Discordant	40.5	Gamma	0.190
Percent Tied	0.0	Tau-a	0.095

-Months : As months of service increases by one unit (one month) the odds of churning will decrease by 1.1%.
- Drop_blk_range: As brop/block calls increases by one unit the odds of churning will increase by 0.1%.
-Actvsubs- As the number of active subs increases by one the odds of churning will decrease by 15.1%.
-roam_mean: As the roaming mean increases by one unit the odds of churning will increase by 0.8%.
-Roam_range : As the roaming range increases by one unit the odds of churning will decrease by 0.2%.
-mtrcycle: An increase of one unit of motorcycle increases the odds of churning 11.6%.
-uniqsubs: As number of unique subs increase by one unit the odds of churning by 24.3%.
- mou_range: As the minute range increases by one unit the odds of churning does not change, which does not make sense.

- ovrmou_range: As the Overage minutes of use increases by one unit the odds of churning will increase by 0.1%.
-ovrmou_mean : As Overage revenue does not change which does not make sense.
-rev_range: As charge amount increases by one dollar the odds of churning does not get affected.
-Refurb_new: As handset refurbished/new increase by one unit the odds of churning decreases by 30%.
-prizm_social_one : As social group letter increase by one unit the odds of churning increases by 1.4%.

The model with maximum concordance ratio is considered for churn prediction. The final model has a concordance ratio of 59.5% that means model is doing a good job in showing the higher probability of churners than non-churners because the distribution of the data shows only 50% are churners.

AIC/BIC:

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	125844.89	123515.62
SC	125854.30	123722.78
-2 Log L	125842.89	123471.62

So AIC/BIC is a measure of goodness of fit of a model. Here both AIC and BIC prefers the intercept and covariates which gives the lowest values as compared to the intercept only.

Q2. Which are the top three factors that affect churn in your model.

As per our model, the top three factors affecting churn in our model is change_mou, roam_Mean and roam_Range.

The change_mou is the change in the minutes of use which is a very good predictor of churn as based on how the customer utilizes the service, it can be a good measure of his/her satisfaction.

The roam_Mean is the mean roaming value, which can give the company better insight into how well their product is being used while on roaming and can indicate on how to change the plan to give better benefits.

The roam_Range is the roaming range value, which can help the company gauge how far to expand their roaming range so that the customer is better satisfied. If the customer travels a lot and does not get good connectivity, that could lead to a high probability of churn.

Q3. What other variables (that if collected) would help to improve the fit of the model.

From a very customer centric perspective, one variable that would drastically help fit the model would be whether there are any offers such a free 2-month unlimited data service and so on. It could be to lure more customers and overall gain a much higher market share. It could be a binary variable 1(if the customer had the free package) and 0(otherwise).

Another variable that could help would be whether the person is part of a family plan or a single person plan as that person would end up spending lesser when on a family plan and the chance of churning would be much lesser.

Another variable could be whether the carrier charges the customer for hotspot service because in many cases, customers are charged for hotspot and that could in a major way contribute to churn.

Q4. Compute the hit ratio for your model. Hit ratio is defined as the percentage of correct predictions using the logit model. Use the model to predict 1 or 0 using the same data.

The SAS System

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of churn by P_final		
	P_final		
churn	no	yes	Total
0	33562 33.56 66.54 54.82	16876 16.88 33.46 43.52	50438 50.44
1	27660 27.66 55.81 45.18	21902 21.90 44.19 56.48	49562 49.56
Total	61222 61.22	38778 38.78	100000 100.00

The overall hit ratio is 55.4% for the dataset.