1. Short-listing a few important variables that could affect churn:

We considered 82% of the given data after removing the missing values and standardized the rest of the data. A snippet of the standardized data is shown below:

churn												
1	N	42460	42460	42460	42460	42460	42460	42460	42460	42460	42460	42460
	MIN	-1.835547925	-1.400995046	-1.006394014	-2.359220465	-0.410179511	-0.42541993	-0.449798815	-0.447325672	-0.089030284	-0.082714455	-0.501226557
	MAX	1.7032469832	24.66294626	13.390762022	13.214491491	70.714456084	43.038102399	28.423489141	28.873350435	75.608788246	70.036393969	49.481311178
	MEAN	0.0307446055	0.0036919916	0.0463013184	0.0592973739	0.0099206283	-0.023365429	-0.027758244	-0.028137338	-0.000518654	-0.009223883	-0.024623189
	STD	1.0356613916	0.9501429854	1.0253137859	1.0231238122	1.0413403242	1.0053323777	0.9661638053	0.9688357959	0.838139633	0.5638838623	0.8126153344
0	N	39668	39668	39668	39668	39668	39668	39668	39668	39668	39668	39668
	MIN	-1.835583314	-1.349324934	-1.006394014	-3.12831613	-0.410179511	-0.42541993	-0.449798815	-0.447325672	-0.089030284	-0.082714455	-0.501226557
	MAX	1.7032823722	80.199201435	21.913300086	15.329346148	29.850923211	27.297884482	35.071188412	28.699112316	131.91872712	232.20055137	149.27372711
	MEAN	-0.03290854	-0.003951849	-0.049560199	-0.063470971	-0.010618884	0.0250099858	0.0297119858	0.0301177618	0.0005551594	0.0098730988	0.0263562719
	STD	0.9592842498	1.0507458385	0.9697408117	0.9706480477	0.9536653389	0.9936646672	1.0341797308	1.031477842	1.1482543005	1.3152480139	1.1671524763

After taking the difference of the mean values of churn 1 and 0, the absolute value of the difference is sorted in the descending order to find out the important variables. The shortlisted variables in the order of importance are shown below:

eqpdays	0.219
hnd_price	0.178
asl_flag	0.138
hnd_webcap	0.124
totmrc_Mean	0.123
mou_Mean	0.096
mou_cvce_Mean	0.088
age1	0.087
complete_Mean	0.085
comp_vce_Mean	0.084
mou_opkv_Mean	0.082
avg3mou	0.081
uniqsubs	0.080
mou_peav_Mean	0.080
peak_vce_Mean	0.079
mou_rvce_Mean	0.078
attempt_Mean	0.078
lor	0.078

As many of these variables are likely to be correlated to each other, we carried out a correlation procedure to avoid multi-collinearity in our model.

Pearson Correlation Coefficients, N = 82128
Prob > |r| under H0: Rho=0

	months	hnd_price	eqpdays	avg3mou	change_rev	asl_flag	change_mou	mou_Range	totmrc_Mean	ovrmou_Range	adults	avgrev	phones	roam_Mean
months	1.00000	-0.18737 <.0001	0.37062 <.0001	-0.03922 <.0001	-0.00774 0.0265	-0.27358 <.0001	-0.00079 0.8206	-0.07592 <.0001	-0.01210 0.0005	-0.00362 0.3001	0.00679 0.0518	0.00900 0.0099	0.43161 <.0001	-0.00274 0.4329
hnd_price	-0.18737 <.0001	1.00000	-0.41275 <.0001	0.18842 <.0001	-0.00378 0.2786	0.10078 <.0001	0.00065 0.8528	0.13848 <.0001	0.19030 <.0001	0.07744 <.0001	-0.00749 0.0318	0.15994 <.0001	0.11777 <.0001	0.01432 <.0001
eqpdays	0.37062 <.0001	-0.41275 <.0001	1.00000	-0.28657 <.0001	-0.00157 0.6536	-0.21087 <.0001	-0.01486 <.0001	-0.22475 <.0001	-0.21721 <.0001	-0.13561 <.0001	0.00497 0.1545	-0.21189 <.0001	-0.36538 <.0001	-0.01794 <.0001
avg3mou	-0.03922 <.0001	0.18842 <.0001	-0.28657 <.0001	1.00000	-0.09084 <.0001	0.15591 <.0001	-0.18050 <.0001	0.61413 <.0001	0.55557 <.0001	0.54730 <.0001	-0.01701 <.0001	0.67743 <.0001	0.28443 <.0001	0.07588 <.0001
change_rev	-0.00774 0.0265	-0.00378 0.2786	-0.00157 0.6536	-0.09084 <.0001	1.00000	-0.01888 <.0001	0.68020 <.0001	0.16309 <.0001	-0.02188 <.0001	0.01817 <.0001	-0.00062 0.8580	-0.10332 <.0001	-0.01170 0.0008	0.48647 <.0001
asl_flag	-0.27358 <.0001	0.10078 <.0001	-0.21087 <.0001	0.15591 <.0001	-0.01888 <.0001	1.00000	-0.03162 <.0001	0.18047 <.0001	0.09763 <.0001	0.04670 <.0001	-0.03056 <.0001	0.12140 <.0001	-0.03817 <.0001	0.00366 0.2942
change_mou	-0.00079 0.8206	0.00065 0.8528	-0.01486 <.0001	-0.18050 <.0001	0.68020 <.0001	-0.03162 <.0001	1.00000	-0.01359 <.0001	-0.02068 <.0001	-0.01914 <.0001	-0.00127 0.7165	-0.12612 <.0001	-0.00647 0.0636	0.27990 <.0001
mou_Range	-0.07592 <.0001	0.13848 <.0001	-0.22475 <.0001	0.61413 <.0001	0.16309 <.0001	0.18047 <.0001	-0.01359 <.0001	1.00000	0.31067 <.0001	0.60623 <.0001	-0.01639 <.0001	0.50146 <.0001	0.17303 <.0001	0.36075 <.0001
totmrc_Mean	-0.01210 0.0005	0.19030 <.0001	-0.21721 <.0001	0.55557 <.0001	-0.02188 <.0001	0.09763 <.0001	-0.02068 <.0001	0.31067 <.0001	1.00000	0.20641 <.0001	-0.02912 <.0001	0.66131 <.0001	0.21624 <.0001	0.04171 <.0001
ovrmou_Range	-0.00362 0.3001	0.07744 <.0001	-0.13561 <.0001	0.54730 <.0001	0.01817 <.0001	0.04670 <.0001	-0.01914 <.0001	0.60623 <.0001	0.20641 <.0001	1.00000	-0.02067 <.0001	0.49047 <.0001	0.13931 <.0001	0.03948 <.0001
adults	0.00679 0.0518	-0.00749 0.0318	0.00497 0.1545	-0.01701 <.0001	-0.00062 0.8580	-0.03056 <.0001	-0.00127 0.7165	-0.01639 <.0001	-0.02912 <.0001	-0.02067 <.0001	1.00000	-0.03742 <.0001	0.00214 0.5392	-0.00106 0.7604
avgrev	0.00900	0.15994	-0.21189	0.67743	-0.10332	0.12140	-0.12612	0.50146	0.66131	0.49047	-0.03742	1.00000	0.26694	0.11753

age1	uniqsubs	lor
0.10994 <.0001	0.01992 <.0001	0.01986 <.0001
-0.08530 <.0001	-0.01409 <.0001	-0.02999 <.0001
0.10001 <.0001	-0.01719 <.0001	0.02515 <.0001
-0.15101 <.0001	-0.02615 <.0001	-0.04870 <.0001
0.00631 0.0707	0.00022 0.9493	-0.00470 0.1776
-0.15633 <.0001	-0.16307 <.0001	-0.01873 <.0001
0.01134 0.0011	0.00086 0.8045	-0.00201 0.5641
-0.12714 <.0001	-0.02240 <.0001	-0.03480 <.0001
-0.09639 <.0001	-0.02702 <.0001	-0.04149 <.0001
-0.07170 <.0001	-0.00319 0.3603	-0.03262 <.0001
0.29183 <.0001	0.04373 <.0001	0.34976 <.0001
-0.13474	-0.02047	-0.05482

We removed variables that showed correlation above 0.75 and considered the next variables in the order. We then proceeded to run the logistic regression. The results of the logistic regression are given below:

The SAS System

The LOGISTIC Procedure

Model Information	
Data Set	WORK.CHURNDATA
Response Variable	churn
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	82128
Number of Observations Used	82128

Response Profile		
Ordered Value	churn	Total Frequency
1	0	42460
2	1	39668

Probability modeled is churn='1'.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	113760.65	110569.36
SC	113769.96	110737.05
-2 Log L	113758.65	110533.36

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	3225.2897	17	<.0001
Score	3079.0228	17	<.0001
Wald	3000.0901	17	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.0743	0.0397	3.4981	0.0614
months	1	-0.0194	0.00115	285.9432	<.0001
hnd_price	1	-0.00183	0.000137	179.5980	<.0001
eqpdays	1	0.00118	0.000047	621.1427	<.0001
avg3mou	1	-0.00036	0.000023	262.2666	<.0001
change_rev	1	0.00269	0.000254	112.1084	<.0001
asl_flag	1	-0.3691	0.0224	271.2067	<.0001
change_mou	1	-0.00054	0.000037	217.1182	<.0001
mou_Range	1	0.000274	0.000025	116.3238	<.0001
totmrc_Mean	1	-0.00533	0.000443	144.5847	<.0001
ovrmou_Range	1	0.000395	0.000056	50.2559	<.0001
adults	1	0.0296	0.00602	24.2388	<.0001
avgrev	1	0.00429	0.000341	158.7388	<.0001
phones	1	0.0860	0.00787	119.2429	<.0001
roam_Mean	1	0.00158	0.000969	2.6587	0.1030
age1	1	-0.00550	0.000355	240.7386	<.0001
uniqsubs	1	0.0862	0.00844	104.2547	<.0001
lor	1	-0.0181	0.00192	88.7207	<.0001

T-values are calculated separately: it is the square root of Wald Chi-Square or Estimate/Standard Error.

Table of coefficients and t-values:

Analysis of Maximum Likelihood Estimates						
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**	**t-value**
Intercept	1	0.0743	0.0397	3.4981	0.0614	1.87
eqpdays	1	0.00118	0.000047	621.1427	<.0001	25.11
hnd_price	1	-0.00183	0.000137	179.598	<.0001	-13.36
asl_flag	1	-0.3691	0.0224	271.2067	<.0001	-16.48
totmrc_Mean	1	-0.00533	0.000443	144.5847	<.0001	-12.03
age1	1	-0.0055	0.000355	240.7386	<.0001	-15.49
avg3mou	1	-0.00036	0.000023	262.2666	<.0001	-15.65
uniqsubs	1	0.0862	0.00844	104.2547	<.0001	10.21
months	1	-0.0194	0.00115	285.9432	<.0001	-16.87
change_rev	1	0.00269	0.000254	112.1084	<.0001	10.59
change_mou	1	-0.00054	0.000037	217.1182	<.0001	-14.59
mou_Range	1	0.000274	0.000025	116.3238	<.0001	10.96
phones	1	0.086	0.00787	119.2429	<.0001	10.93
adults	1	0.0296	0.00602	24.2388	<.0001	4.92
ovrmou_Range	1	0.000395	0.000056	50.2559	<.0001	7.05
avgrev	1	0.00429	0.000341	158.7388	<.0001	12.58
roam_Mean	1	0.00158	0.000969	2.6587	0.103	1.63
lor	1	-0.0181	0.00192	88.7207	<.0001	-9.43

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
months	0.981	0.979	0.983
hnd_price	0.998	0.998	0.998
eqpdays	1.001	1.001	1.001
avg3mou	1.000	1.000	1.000
change_rev	1.003	1.002	1.003
asl_flag	0.691	0.662	0.722
change_mou	0.999	0.999	1.000
mou_Range	1.000	1.000	1.000
totmrc_Mean	0.995	0.994	0.996
ovrmou_Range	1.000	1.000	1.001
adults	1.030	1.018	1.042
avgrev	1.004	1.004	1.005
phones	1.090	1.073	1.107
roam_Mean	1.002	1.000	1.003
age1	0.995	0.994	0.995
uniqsubs	1.090	1.072	1.108
lor	0.982	0.978	0.986

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	61.2	Somers' D	0.230
Percent Discordant	38.2	Gamma	0.232
Percent Tied	0.6	Tau-a	0.115
Pairs	1684303280	c	0.615

Akaike's Information Criterion, AIC of the model = [-2logL + 2p] = 110569.36

Bayesian Information Criterion, BIC of the model = [-2logL + plog(n)] = 110737.05

The lower these values, the better the model.

McFadden's R-square = difference in (-2LogL)/Null model's (-2LogL) = 0.03

Meaning of coefficients, significance and odds-ratios:

Coefficient of *eqpdays* = 0.00118
tvalue = 25.11. This value is greater than the critical value = 1.96 at 95% confidence. Hence, this coefficient is statistically different from zero with more than 95% confidence level.
The logistic regression model can be written as follows:
Ln(odds ratio) = β X
e^{β} gives the odds ratio for a feature, keeping all other features constant.
From the odds ratio estimate table, the odds ratio estimate of *eqpdays* is 1.001. This means that when the number of days of the current equipment (measured as *eqpdays*) increases by one unit, the odds of the customer churning over odds of customer not churning increases 1.001 times keeping other factors the same.
In other words, we can say for a one unit increase in the number of days of the current equipment, we expect to see about 0.1% increase in the odds of customer churn.

Coefficient of *hnd_price* = -0.00183
tvalue = -13.36. This value is less than the critical value = -1.96 at 95% confidence. Hence, this coefficient is statistically different from zero with more than 95% confidence level.
The odds ratio estimate of *hnd_price* is 0.998. This means that when the handset price increases by one unit, the odds of the customer churning over odds of customer not churning changes 0.998 times keeping other factors the same.
In other words, for a one unit increase in the handset price, we expect to see about 0.2% decrease in the odds of customer churn.

Coefficient of *asl_flag* = -0.3691
The odds ratio estimate of *asl_flag* is 0.691. This means that when there is an Account Spending Limit, the odds of the customer churning over odds of customer not churning changes 0.691 times compared to when there are no Account Spending Limits, keeping other factors the same.
In other words, when there is an Account Spending Limit, we expect to see about 31% decrease in the odds of customer churn compared to when there are no Account Spending Limits. Probably, customers who have account spending limits are less like to leave the plan/business than customers who have no account spending limits.

Prediction accuracy (percent concordance):

Concordance is one of the measures for telling how well the model is predicting. For any random pairing of 1's and 0's from the actual data, Percent Concordance is the percentage of number of cases where the model has thrown a probability for a 1 > probability for a 0. Higher the concordance percent, better the model.

For our model, the percent concordance = 61.2%

2. The top three factors that affect churn in our model are:

eqpdays - Number of days of the current equipment

hnd_price - Handset Price

asl_flag - Account Spending Limits

3. Variables (that if collected) would help to improve the fit of the model:

Apart from the variables that are present in the dataset, there can be several other key factors which if included could enhance the model and help in more accurate prediction of churn. Few of these are:

a) Customer Plan: Whether the customer has an individual plan or a family plan. A customer might find a better individual plan with the competing telecom firms and hence churn out, but if it's a family plan, the customers are more likely to stay.

b) Customer Priority: We are not told if a customer is a Premium customer or Standard customer. A Premium customer is less likely to get churned compared to a Standard customer.

c) Market competition: Competition against other telecom firms are not captured in the data. Market competition is one key factor to decide why a customer churned out from a telecom firm.

4) The hit ratio (% events correctly classified) of our model with the top 17 variables is 0.58. It implies that our model successfully predicts churn with an accuracy of 58%. It is meaningful as several features are not captured in the top selected variables. Several important factors are not available in the data. With those factors, the hit ratio is likely to improve with respect to the current performance. The confusion matrix of prediction is shown below.

The SAS System

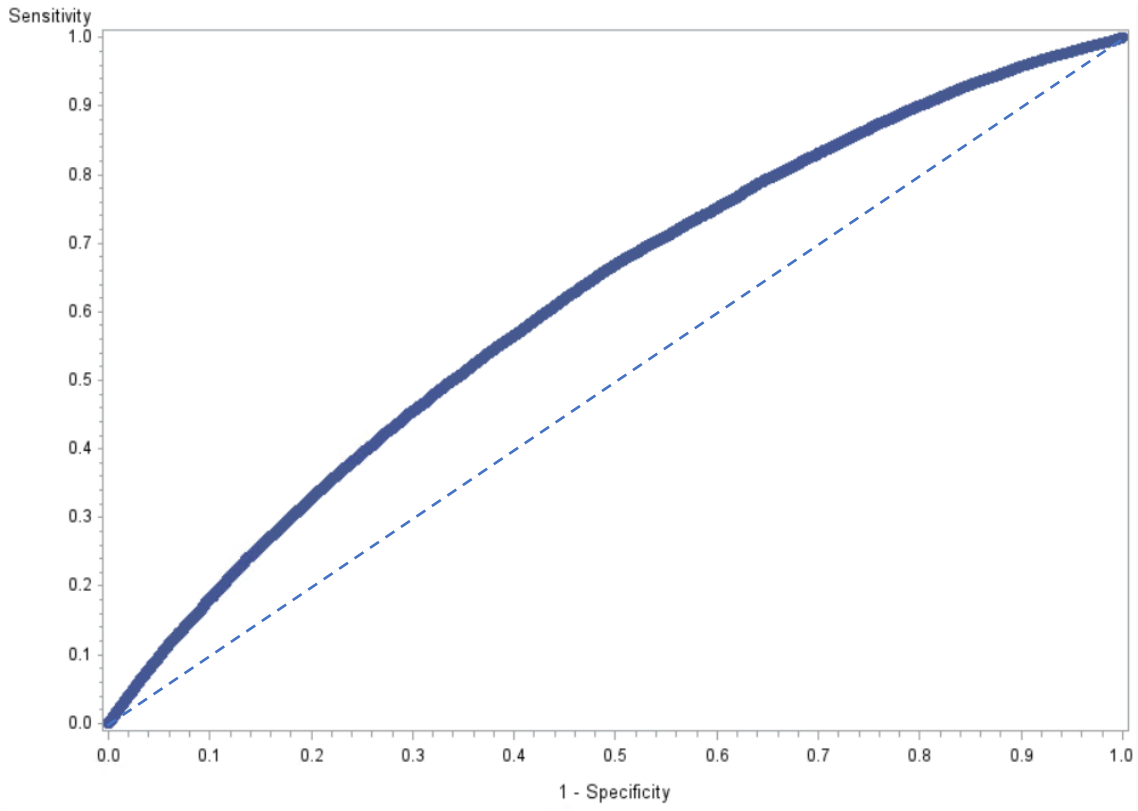
The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of churn by _INTO_		
	INTO(Formatted Value of the Predicted Response)		
churn	0	1	Total
0	27040 32.92 63.68 58.99	15420 18.78 36.32 42.49	42460 51.70
1	18795 22.89 47.38 41.01	20873 25.42 52.62 57.51	39668 48.30
Total	45835 55.81	36293 44.19	82128 100.00

The calculated hit ratio is shown below:

Analysis Variable : Match
Mean
0.5833942

ROC Curve:



Lift Curve: