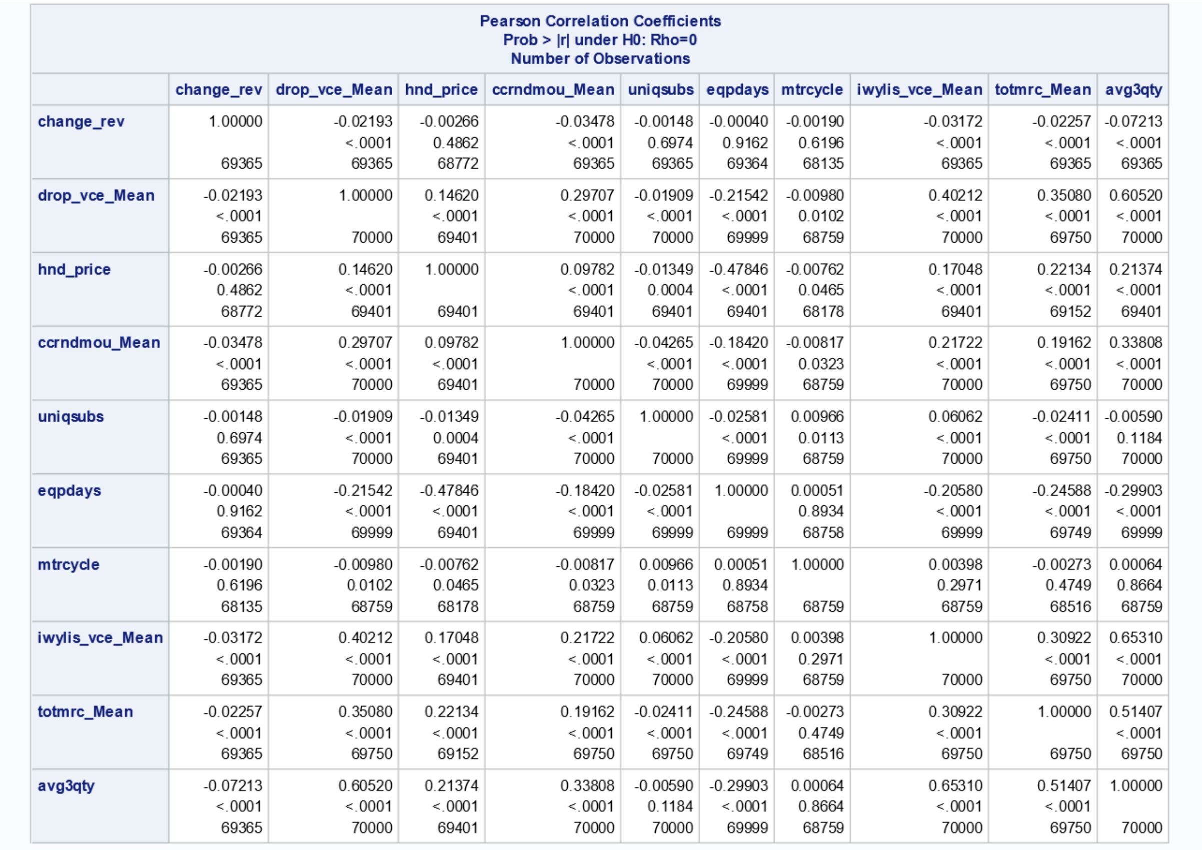
CHURN DATA SET

Steps that we followed for selecting the variables are:

- Computed means of all x variables for churners and non-churners separately and then computed the percentage difference for each variable.
- Sorted the variables from high to low based on percentage difference in means.
- After that, we selected top 20 variables and then checked the correlation and eliminated those variables that have correlation>0.70.
- Then, we started dropping one of two variables that are highly correlated and kept on dropping and adding the variables based on next high percentage difference in means till we got top 10 variables for good candidates' selection.
- We then proceeded to run the logistic regression. The results of the logistic regression are given below:

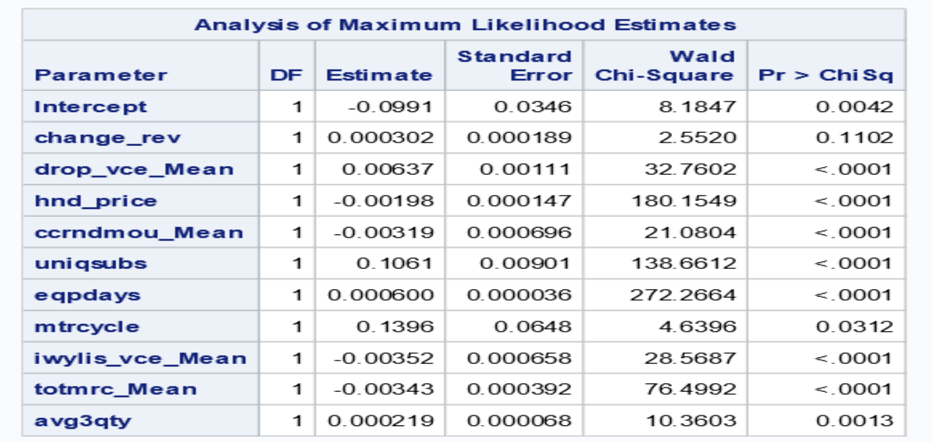
These variables are selected that are mentioned in correlation table



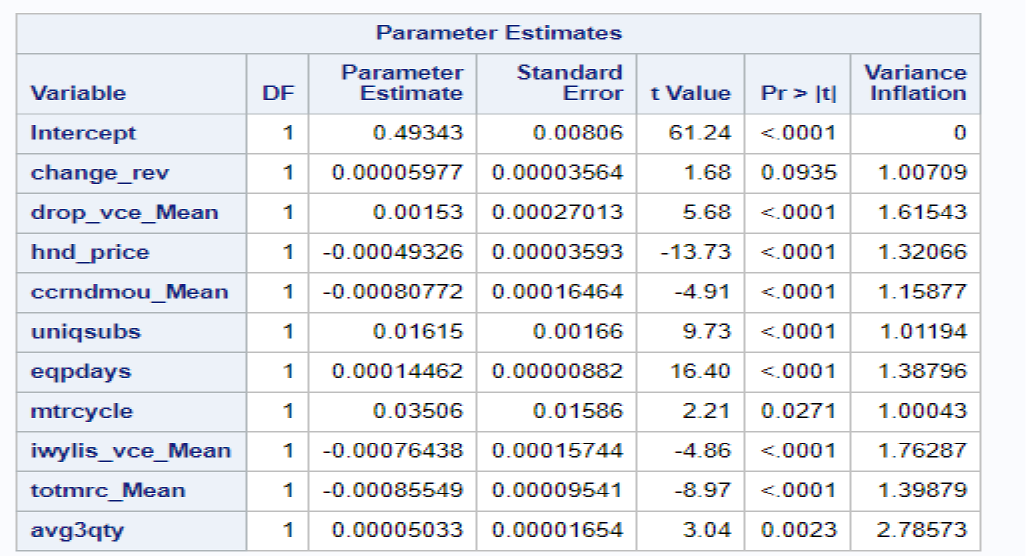
Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	change_rev	drop_vce_Mean	hnd_price	ccrndmou_Mean	uniqsubs	eqpdays	mtrcycle	iwylis_vce_Mean	totmrc_Mean	avg3qty
change_rev	1.00000 69365	-0.02193 <.0001 69365	-0.00266 0.4862 68772	-0.03478 <.0001 69365	-0.00148 0.6974 69365	-0.00040 0.9162 69364	-0.00190 0.6196 68135	-0.03172 <.0001 69365	-0.02257 <.0001 69365	-0.07213 <.0001 69365
drop_vce_Mean	-0.02193 <.0001 69365	1.00000 70000	0.14620 <.0001 69401	0.29707 <.0001 70000	-0.01909 <.0001 70000	-0.21542 <.0001 69999	-0.00980 0.0102 68759	0.40212 <.0001 70000	0.35080 <.0001 69750	0.60520 <.0001 70000
hnd_price	-0.00266 0.4862 68772	0.14620 <.0001 69401	1.00000 69401	0.09782 <.0001 69401	-0.01349 0.0004 69401	-0.47846 <.0001 69401	-0.00762 0.0465 68178	0.17048 <.0001 69401	0.22134 <.0001 69152	0.21374 <.0001 69401
ccrndmou_Mean	-0.03478 <.0001 69365	0.29707 <.0001 70000	0.09782 <.0001 69401	1.00000 70000	-0.04265 <.0001 70000	-0.18420 <.0001 69999	-0.00817 0.0323 68759	0.21722 <.0001 70000	0.19162 <.0001 69750	0.33808 <.0001 70000
uniqsubs	-0.00148 0.6974 69365	-0.01909 <.0001 70000	-0.01349 0.0004 69401	-0.04265 <.0001 70000	1.00000 70000	-0.02581 <.0001 69999	0.00966 0.0113 68759	0.06062 <.0001 70000	-0.02411 <.0001 69750	-0.00590 0.1184 70000
eqpdays	-0.00040 0.9162 69364	-0.21542 <.0001 69999	-0.47846 <.0001 69401	-0.18420 <.0001 69999	-0.02581 <.0001 69999	1.00000 69999	0.00051 0.8934 68758	-0.20580 <.0001 69999	-0.24588 <.0001 69749	-0.29903 <.0001 69999
mtrcycle	-0.00190 0.6196 68135	-0.00980 0.0102 68759	-0.00762 0.0465 68178	-0.00817 0.0323 68759	0.00966 0.0113 68759	0.00051 0.8934 68758	1.00000 68759	0.00398 0.2971 68759	-0.00273 0.4749 68516	0.00064 0.8664 68759
iwylis_vce_Mean	-0.03172 <.0001 69365	0.40212 <.0001 70000	0.17048 <.0001 69401	0.21722 <.0001 70000	0.06062 <.0001 70000	-0.20580 <.0001 69999	0.00398 0.2971 68759	1.00000 70000	0.30922 <.0001 69750	0.65310 <.0001 70000
totmrc_Mean	-0.02257 <.0001 69365	0.35080 <.0001 69750	0.22134 <.0001 69152	0.19162 <.0001 69750	-0.02411 <.0001 69750	-0.24588 <.0001 69749	-0.00273 0.4749 68516	0.30922 <.0001 69750	1.00000 69750	0.51407 <.0001 69750
avg3qty	-0.07213 <.0001 69365	0.60520 <.0001 70000	0.21374 <.0001 69401	0.33808 <.0001 70000	-0.00590 0.1184 70000	-0.29903 <.0001 69999	0.00064 0.8664 68759	0.65310 <.0001 70000	0.51407 <.0001 69750	1.00000 70000

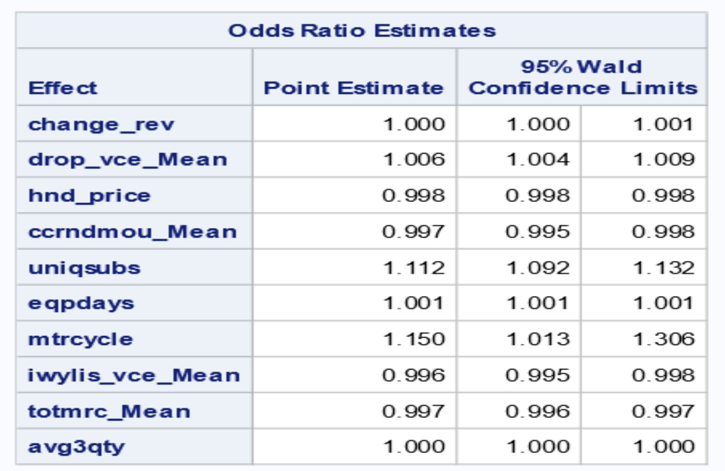
1. **Include a clean table of coefficients, t-values, and odds ratio only. I do not want the SAS output as it is. Interpret the logistic output explaining AIC/BIC, meaning of coefficients, significance of betas, prediction accuracy (percent concordance), odds-ratios etc.**



Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.0991	0.0346	8.1847	0.0042
change_rev	1	0.000302	0.000189	2.5520	0.1102
drop_vce_Mean	1	0.00637	0.00111	32.7602	<.0001
hnd_price	1	-0.00198	0.000147	180.1549	<.0001
ccrndmou_Mean	1	-0.00319	0.000696	21.0804	<.0001
uniqsubs	1	0.1061	0.00901	138.6612	<.0001
eqpdays	1	0.000600	0.000036	272.2664	<.0001
mtrcycle	1	0.1396	0.0648	4.6396	0.0312
iwylis_vce_Mean	1	-0.00352	0.000658	28.5687	<.0001
totmrc_Mean	1	-0.00343	0.000392	76.4992	<.0001
avg3qty	1	0.000219	0.000068	10.3603	0.0013



Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.49343	0.00806	61.24	<.0001	0
change_rev	1	0.00005977	0.00003564	1.68	0.0935	1.00709
drop_vce_Mean	1	0.00153	0.00027013	5.68	<.0001	1.61543
hnd_price	1	-0.00049326	0.00003593	-13.73	<.0001	1.32066
ccrndmou_Mean	1	-0.00080772	0.00016464	-4.91	<.0001	1.15877
uniqsubs	1	0.01615	0.00166	9.73	<.0001	1.01194
eqpdays	1	0.00014462	0.00000882	16.40	<.0001	1.38796
mtrcycle	1	0.03506	0.01586	2.21	0.0271	1.00043
iwylis_vce_Mean	1	-0.00076438	0.00015744	-4.86	<.0001	1.76287
totmrc_Mean	1	-0.00085549	0.00009541	-8.97	<.0001	1.39879
avg3qty	1	0.00005033	0.00001654	3.04	0.0023	2.78573



Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
change_rev	1.000	1.000	1.001
drop_vce_Mean	1.006	1.004	1.009
hnd_price	0.998	0.998	0.998
ccrndmou_Mean	0.997	0.995	0.998
uniqsubs	1.112	1.092	1.132
eqpdays	1.001	1.001	1.001
mtrcycle	1.150	1.013	1.306
iwylis_vce_Mean	0.996	0.995	0.998
totmrc_Mean	0.997	0.996	0.997
avg3qty	1.000	1.000	1.000

Interpretations of all the variables with odd ratios and beta coefficients:

- **Change_rev**

Coefficient of change_rev = 0.0003

tvalue= 1.68. This value is smaller than the critical value = 1.96 at 95% confidence. Hence, this coefficient is not statistically different from zero with more than 95% confidence level.

Interpreting the percent change in revenue, a 1% change in revenue will result in an 0.0003 increase probability that a person will have churn. From an odds ratio estimate of 1.00, this is interpreted as increasing the percent change in revenue by 1% changes the odds of a customer churning is 1 times as likely for having someone churn.

- **Drop_vce_mean**

Coefficient of drop_vce_mean = 0.00637

tvalue= 5.68. 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.

Interpreting the number of dropped voice calls, for each additional dropped call increases the probability of someone churning by 0.00637. From an odds ratio estimate of 1.06, an increase in dropped calls by 1 makes the odds 1x more likely for someone to churn.

- **hnd_price**

Coefficient of *hnd_price* = -0.00198

tvalue = -13.73. 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 estimates 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 the 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.

- **Ccrndmou_mean**

Coefficient of ccrndmou_mean = -0.000807

tvalue= |-4.91| = 4.91. 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.

Interpreting the rounded estimate index of customer care, for each increase in the average customer care index, results in a 0.000807 probability that a person will not churn (since the coefficient is negative). From an odds ratio estimate of 0.997, an increase in the index mean of customer care results in the odd of 0.997 x more likely to churn.

- **eqpdays**

Coefficient of *eqpdays* = 0.0006

tvalue = 16.4. 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.

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 the 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 0.06% increase in the odds of customer churn.

- **uniqsubs**

Coefficient of uniqsubs: = 1.061

From the odds ratio estimate table, the odds ratio estimate of *uniqsubs* is 1.112. This means that when uniqsubs increases by one unit, the odds of the customer churning over odds of the customer not churning increases 1.112 times keeping other factors the same.

In other words, we can say for a one unit increase in uniqsubs we expect to see about 1.112% increase in the odds of customer churn.

- **mtrcycle**

Coefficients of mtrcycle: 0.1396

The logistic regression model can be written as follows:
From the odds ratio estimate table, the odds ratio estimate of *mtrcycle* is 1.112. This means that when mtrcycle increases by one unit, the odds of the customer churning over odds of the customer not churning increases by 0.1396 times keeping other factors the same.

In other words, we can say for a one unit increase in *mtrcycle*. we expect to see about 0.1396% increase in the odds of customer churn.

- **iwylis_vce_Mean**

Coefficients of iwylis_vce_Mean: -0.00352

From the odds ratio estimate table, the odds ratio estimates of iwylis_vce_Mean is -0.00352. This means that when iwylis_vce_Mean increases by one unit, the odds of the customer churning over odds of the customer not churning decreases by 0. 00352 times keeping other factors the same.

In other words, we can say for a one unit increase in iwylis_vce_Mean we expect to see about 0.00352% decrease in the odds of customer churn.

- **totmrc_Mean**

Coefficients of totmrc_Mean: -0.00343
From the odds ratio estimate table, the odds ratio estimates of totmrc_Mean is -0. 00343. This means that when totmrc_Mean increases by one unit, the odds of the customer churning over odds of the customer not churning decreases by 0. 00343 times keeping other factors the same. In other words, we can say for a one unit increase in totmrc_Mean we expect to see about 0.00343% decrease in the odds of customer churn.

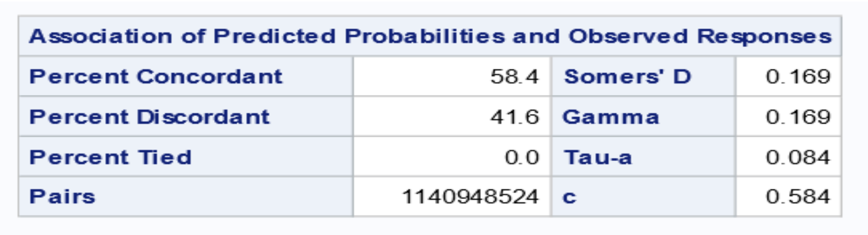
- **Avg3qty**

Coefficients of avg3qty: 0.000219

From the odds ratio estimate table, the odds ratio estimate of avg3qty is 0.000219. This means that when avg3qty increases by one unit, the odds of the customer churning over odds of the customer not churning increases by 0. 000219times keeping other factors the same.

In other words, we can say for a one unit increase in avg3qty we expect to see about 0.000219% increase in the odds of customer churn.

Percentage Concordance and Somers' D Interpretation:



Association of Predicted Probabilities and Observed Responses			
Percent Concordant	58.4	Somers' D	0.169
Percent Discordant	41.6	Gamma	0.169
Percent Tied	0.0	Tau-a	0.084
Pairs	1140948524	c	0.584

A pair of observations with different response variables is said to be concordant if the observation with the lower ordered response value has a lower predicted mean score than the observation with the higher ordered response value. A better model is defined by a higher corresponding concordance percentage. We want the model with the value to be more than 50% as 50% can be looked at as a pure chance and our model should be able to perform better than random chance.

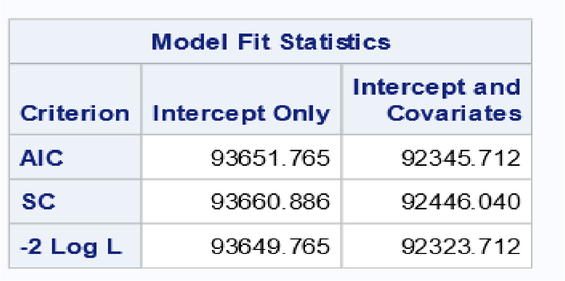
For our model, the percentage concordant is 58.4% meaning that our model is able to perform better than the pure chance model.

Out of 100% possible pairs, whatever pairs are not concordant are said to be discordant pairs. The percentage of discordant pairs for our model is 41.6%.

Somers’ D helps to determine the strength and the direction of relation between pair of variables. It can take any value between -1 and 1 and is defined as the difference between concordance and discordant pairs divided by the total number of pairs with different responses. For our model, since we already have % of concordance and discordance, we can substitute these into the formula and take the total pairs to be 100. The corresponding Somers’ D value will be:

(58.4 - 41.6)/100 = 0.168. From the output table, we can see that the actual value is 0.169, meaning there was some rounding off done else we would’ve got the exact match.

AIC /BIC Interpretation:



Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	93651.765	92345.712
SC	93660.886	92446.040
-2 Log L	93649.765	92323.712

Looking at the model fit statistics, we can see that the AIC for the intercept only model is:93651.765 while the corresponding values for the intercept and covariates model is slightly lower that is: 92345.712. Lower the AIC, the better is the model. This means the variables that we have included to predict the churn status are good as they are able to perform better than a model which has only the intercept that is, the null model.

Similarly, BIC / SC for the intercept only model is:93660.886 while the corresponding values for the intercept and covariates model is slightly lower that is: 92446.040. SC is a model fit criterion which penalizes the addition of more variables in the model. Since the SC is slightly higher than the corresponding AIC value, when compared to the values for the intercept only model, we can interpret this as a model being penalized for including more variables.

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

We have decided the three top factors that affect churn in your model depending upon the significance and coefficients of variables used in the model. So, the three variables that are significant with high coefficient values are:

Mtrcycle - Coefficient of mtrcycle = 0.03506. Also, tvalue= 2.21. 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.

Interpretation - Interpreting whether a person owns a motorcycle, for each person who ownes a motorcycle, they have a 0.03506 increase in probability that a person will churn. From an odds ratio estimate of 1.150, if a person ownes a motorcycle, their odds are 1.15x more likely to churn.

Uniqsubs - Coefficient of uniqsubs = 0.01615. Also, the tvalue= 9.73. 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.

Interpretation - Interpreting the number of unique subs, for each increase in unique subs, results in a 0.01615 increase in probability that a person will churn. From an odds ratio estimate of 1.112, an increase in the unique subs will result in the odds of 1.112 x more likely to churn.

Drop_vce_mean - Coefficient of drop_vce_mean = 0.00637. Also, the tvalue= 5.68. 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.

Interpretation - Interpreting the number of dropped voice calls, for each additional dropped call increases the probability of someone churning by 0.00637. From an odds ratio estimate of 1.06, an increase in dropped calls by 1 makes the odds 1x more likely for someone to churn.

3. What other 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.

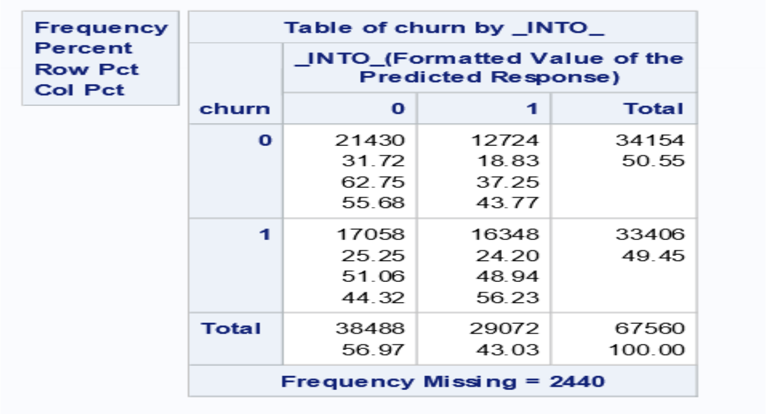
c) **Market competition**: Competition against other telecom firms are not captured in the data. Market competition is one key factor to decide why a customer churned out from a telecom firm.

4. 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 table here gives the frequency distribution for the predicted vs actual values for our model.

The hit ratio (% events correctly classified) of our model for training data set with the top 10 variables is 0.55.9. It implies that our model successfully predicts churn with an accuracy of 55.9%. 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.

There are 21430 that are correctly classified out of 67560 which has a value of 0.
Also, there are 16348 that are correctly classified out of 67560 which has a value of 1.

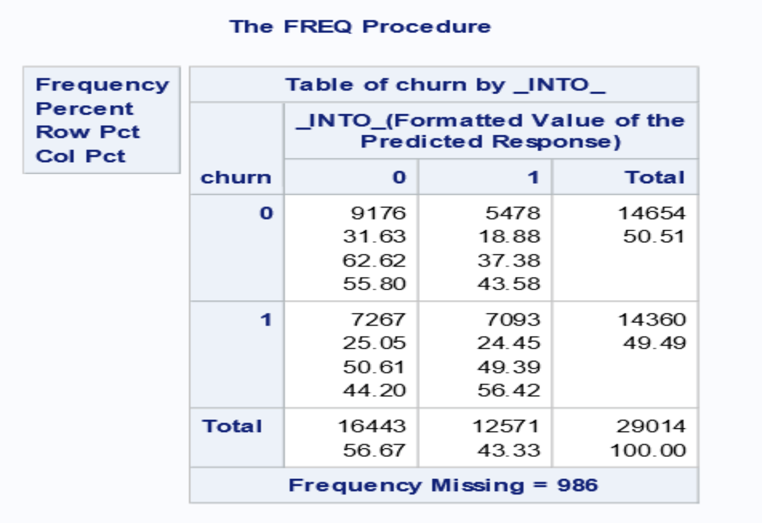


Frequency Percent Row Pct Col Pct	Table of churn by _INTO_		
	INTO(Formatted Value of the Predicted Response)		
churn	0	1	Total
0	21430 31.72 62.75 55.68	12724 18.83 37.25 43.77	34154 50.55
1	17058 25.25 51.06 44.32	16348 24.20 48.94 56.23	33406 49.45
Total	38488 56.97	29072 43.03	67560 100.00
Frequency Missing = 2440			

5. **Using the model parameters predict the churn for the holdout sample as well and compute the hit ratio.**

The hit ratio (% events correctly classified) of our model for testing sample with the top 10 variables is 0.56. It implies that our model successfully predicts churn with an accuracy of 56%. 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.

There are 9176 that are correctly classified out of 29014 which has a value of 0.
Also, there are 7093 that are correctly classified out of 29014 which has a value of 1.



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	9176 31.63 62.62 55.80	5478 18.88 37.38 43.58	14654 50.51
1	7267 25.05 50.61 44.20	7093 24.45 49.39 56.42	14360 49.49
Total	16443 56.67	12571 43.33	29014 100.00
Frequency Missing = 986			