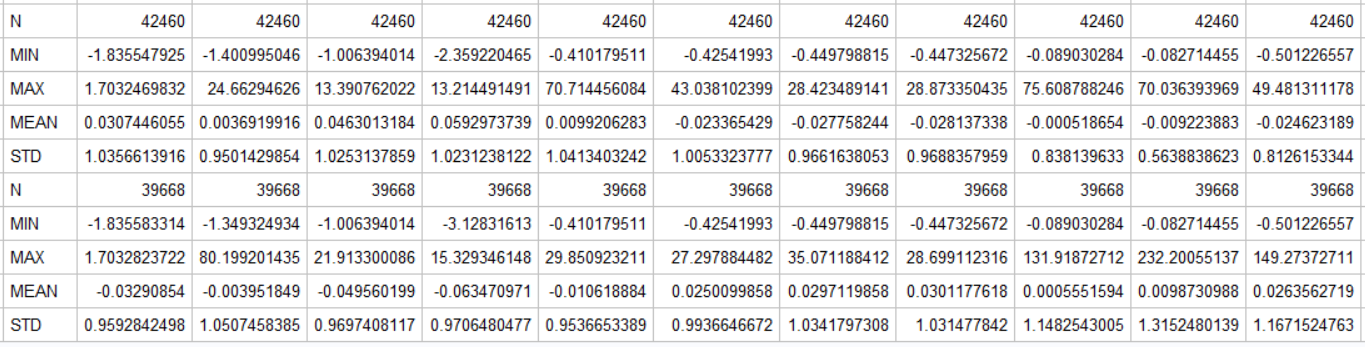
1. Short-listing a few important variables that could affect churn:

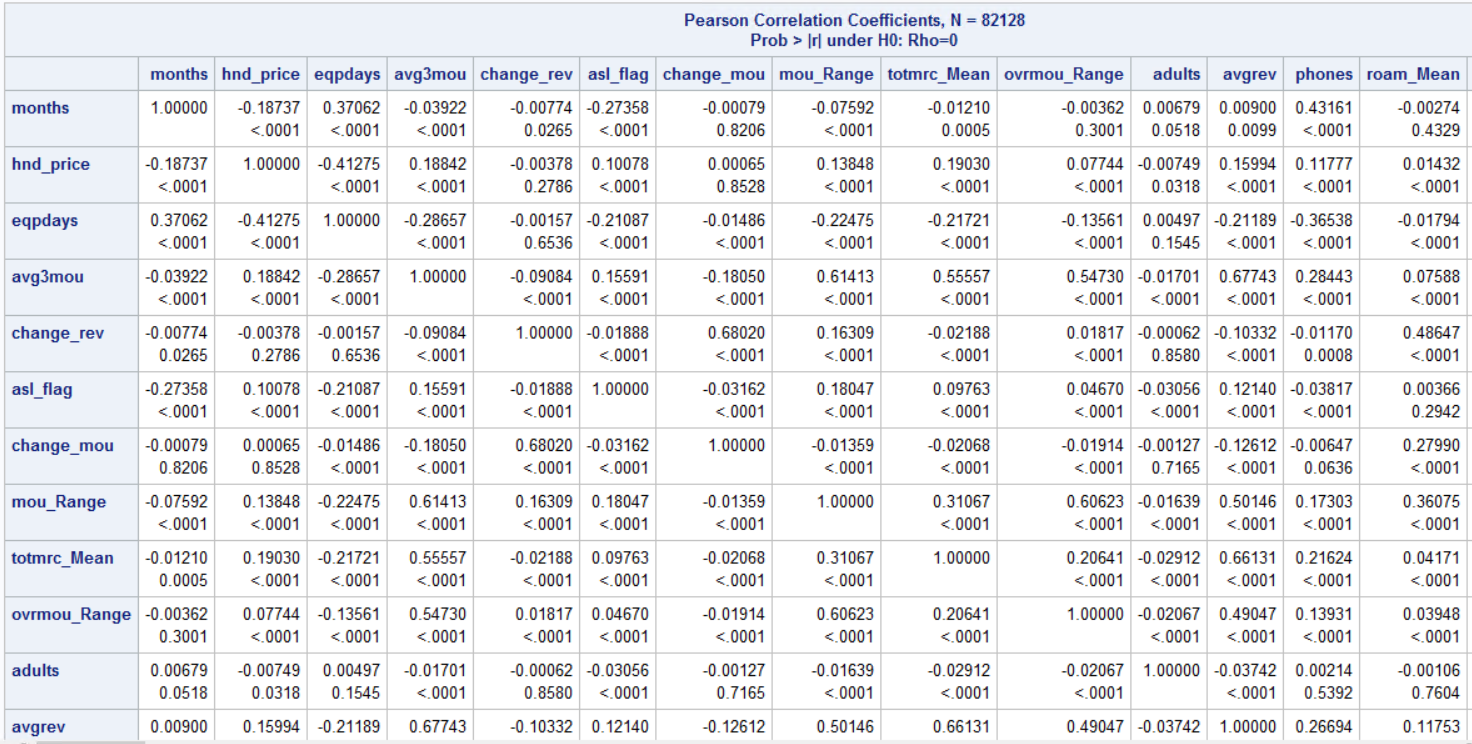
We considered 82% of the given data after removing the missing values and standardized the rest of the data. A snippet of the standardized data is shown below:

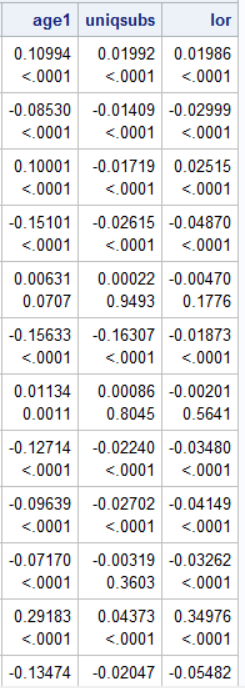


After taking the difference of the mean values of churn 1 and 0, the absolute value of the difference is sorted in the descending order to find out the important variables. The shortlisted variables in the order of importance are shown below:

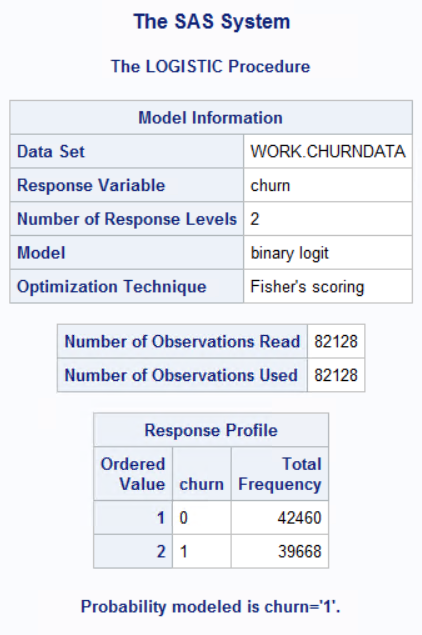


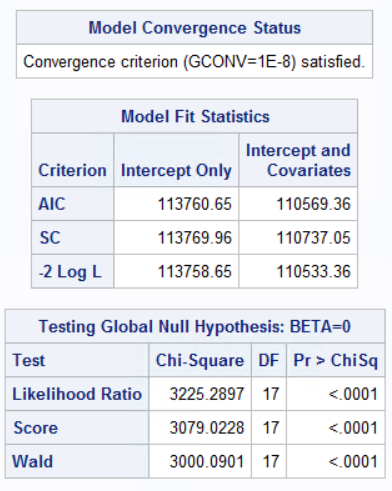
As many of these variables are likely to be correlated to each other, we carried out a correlation procedure to avoid multi-collinearity in our model.





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:





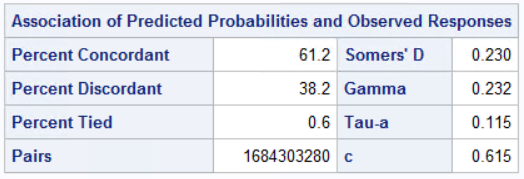


T-values are calculated separately: it is the square root of Wald Chi-Square or Estimate/Standard Error.

Table of coefficients and t-values:







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

eβ 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 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 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:

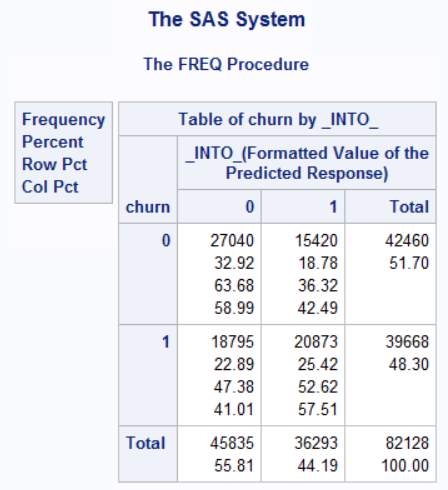
Apart from the variables that are present in the dataset, there can be several other key factors which if included could enhance the model and help in more accurate prediction of churn. Few of these are:

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.

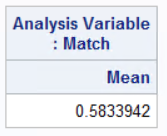
b) Customer Priority: We are not told if a customer is a Premium customer or Standard customer. A Premium customer is less likely to get churned compared to a Standard customer.

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.

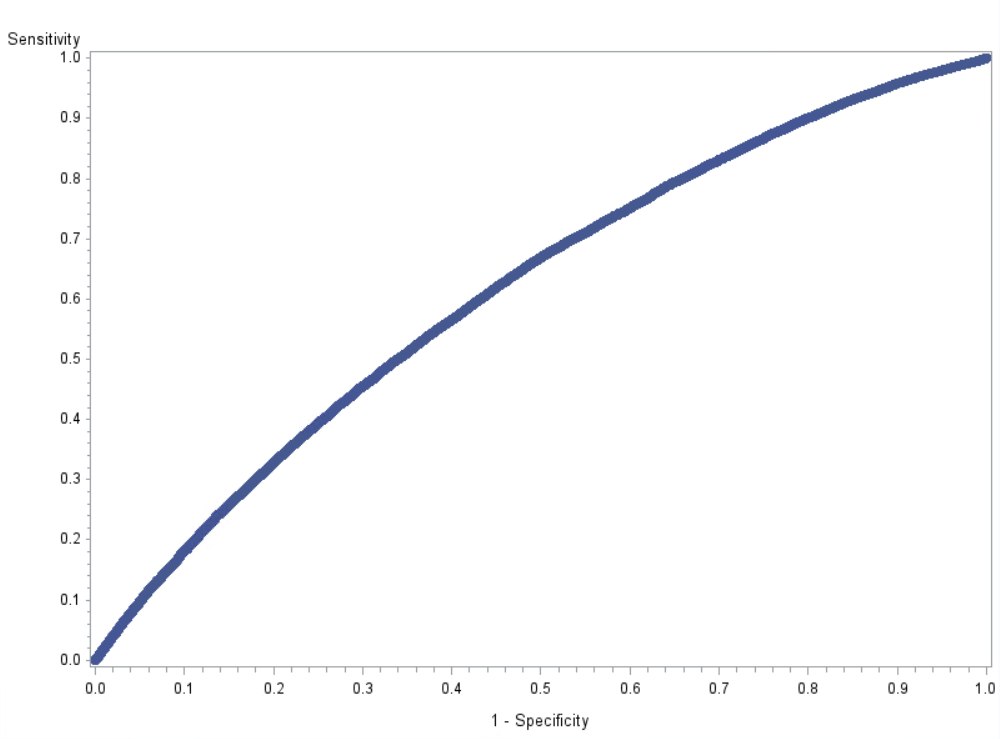
4) The hit ratio (% events correctly classified) of our model with the top 17 variables is 0.58. It implies that our model successfully predicts churn with an accuracy of 58%. It is meaningful as several features are not captured in the top selected variables. Several important factors are not available 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 calculated hit ratio is shown below:



ROC Curve:



Lift Curve:

