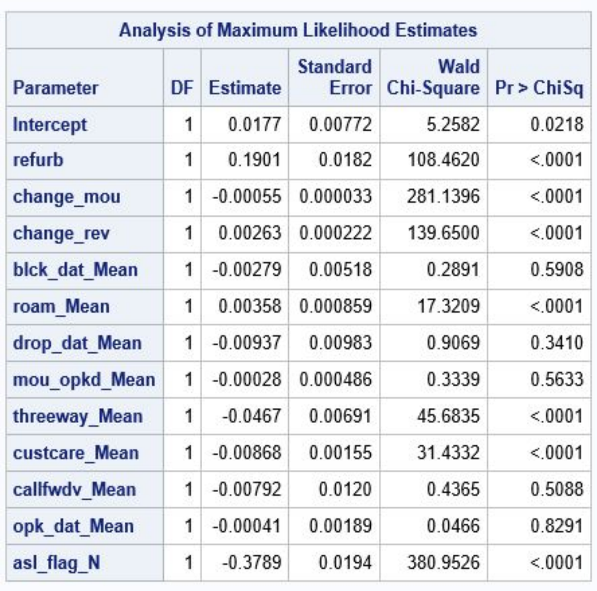
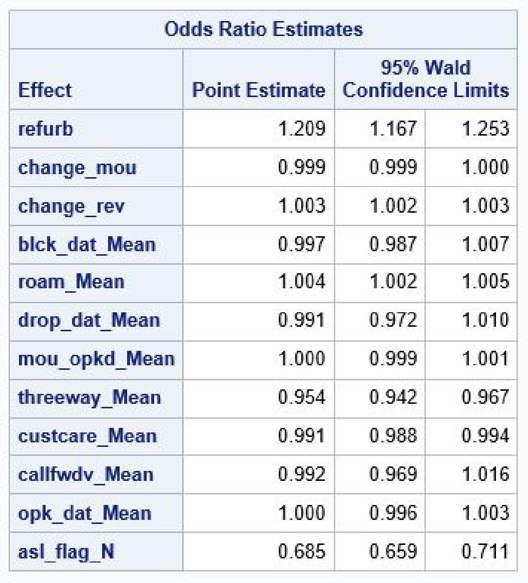
Q1. Include a table of coefficients, t-values, and odds ratio. Interpret the logistic output explaining AIC/BIC, meaning of coefficients, significance, prediction accuracy (percent concordance), odds-ratios etc.

Ans.



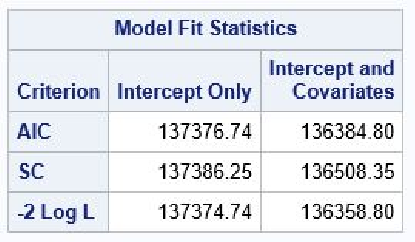
Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.0177	0.00772	5.2582	0.0218
refurb	1	0.1901	0.0182	108.4620	<.0001
change_mou	1	-0.00055	0.000033	281.1396	<.0001
change_rev	1	0.00263	0.000222	139.6500	<.0001
blck_dat_Mean	1	-0.00279	0.00518	0.2891	0.5908
roam_Mean	1	0.00358	0.000859	17.3209	<.0001
drop_dat_Mean	1	-0.00937	0.00983	0.9069	0.3410
mou_opkd_Mean	1	-0.00028	0.000486	0.3339	0.5633
threeway_Mean	1	-0.0467	0.00691	45.6835	<.0001
custcare_Mean	1	-0.00868	0.00155	31.4332	<.0001
callfwdv_Mean	1	-0.00792	0.0120	0.4365	0.5088
opk_dat_Mean	1	-0.00041	0.00189	0.0466	0.8291
asl_flag_N	1	-0.3789	0.0194	380.9526	<.0001

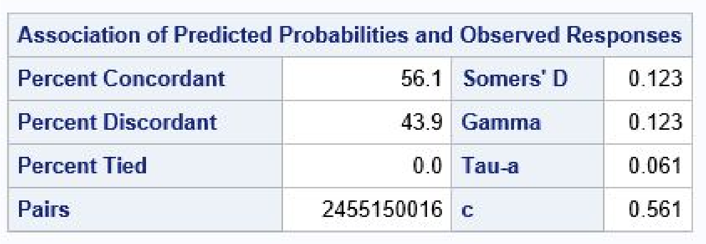


Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
refurb	1.209	1.167	1.253
change_mou	0.999	0.999	1.000
change_rev	1.003	1.002	1.003
blck_dat_Mean	0.997	0.987	1.007
roam_Mean	1.004	1.002	1.005
drop_dat_Mean	0.991	0.972	1.010
mou_opkd_Mean	1.000	0.999	1.001
threeway_Mean	0.954	0.942	0.967
custcare_Mean	0.991	0.988	0.994
callfwdv_Mean	0.992	0.969	1.016
opk_dat_Mean	1.000	0.996	1.003
asl_flag_N	0.685	0.659	0.711



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	137376.74	136384.80
SC	137386.25	136508.35
-2 Log L	137374.74	136358.80



Association of Predicted Probabilities and Observed Responses			
Percent Concordant	56.1	Somers' D	0.123
Percent Discordant	43.9	Gamma	0.123
Percent Tied	0.0	Tau-a	0.061
Pairs	2455150016	c	0.561

The coefficients can be interpreted as follows:

Keeping all other variables constant, if the handset used by the customer is refurbished, odds of the customer churning increases by 20.9% compared to if the handset being used is new. The refurb variable has a very low p value, or a very high significance, at the 99.99% level and definitely at the standard 95% or 90% levels of significance.

For a 1 unit increase in the % change in minutes of use, the odds of the customer churning, decreases by 0.1%, provided all other variables are held constant. This variable also has a high impact, which is confirmed by its p value of less than 0.0001, leading it to be significant at the 99.99% and lower level of significance.

For 1 unit increase in the % change of revenue generated from the customer, the odds of churning go up by 0.3%, if all other variables remain unchanged. This is important as the variable is significant at the 99.99% level of significance and definitely at the 95% level of significance.

If all other variables remain the same and the mean value of the number of blocked data calls goes up by 1 unit, the odds of the customer churning goes down by 0.3%. The p value of .5908 implies it is insignificant at the 95% or even the 90% level of significance, instead being significant at 40.92% level of significance and lower.

For 1 unit increase in the mean value of roaming for the customer, the odds of churning increase by 0.4%, keeping all other variables constant. The p value of less than 0.0001 means it is significant at the 99.99% level of significance and definitely at the 95% level of significance.

Provided that there is no change in any of the other variables, if the mean value of the number of dropped data calls goes up by 1 unit, the odds of the customer churning goes down by 0.9%. The variable is insignificant at the 90% level of significance, with a p value of .3410, meaning it only becomes significant at 65.90% level of significance.

For a unit increase in the mean of the minutes of usage of off peak data calls, the odds of customer churning remains the same, with all other variables held constant. In other words, the change in mean value of minutes of usage of off peak data calls doesn't impact the churn data. This can also be verified by the fact that the p-value of the variable is too high (0.5633), meaning it is significant at 43.67% level of significance, but insignificant at the usual 95% or even the 90% level of significance.

Given other variables remain unchanged, if the mean value of threeway calls increase by 1 unit, the odds of churning reduce by 4.6%. The threeway_Mean variable is significant at the 95% level of significance, with a p value of less than 0.0001.

For 1 unit increase in the mean value of customer care calls, the odds of customer churning go down by 0.9%, if the remaining variables are kept constant. The variable custcare_mean has a p value of less than 0.0001 meaning it is significant at the higher than 99.99% level of significance and definitely significant at the 95% level of significance.

The callfwdv_mean variable has a p value of .5088, means it is only significant at the 49.12% level of significance and lower but not at the 95% level. With all other variables remaining the same, for a unit increase in the mean value of call forward calls, the odds of churning go down by 0.8%.

The odds of customer churning isn't impacted at all by the mean value of number of off peak data calls. For a unit increase in the mean value, provided the other variables are held constant, the odds of churning remain the same (odds ratio of 1). Since the variable has a p value of 0.8291, it is insignificant at the 95% level of significance, being significant only at the 17.09% and lower levels.

With the remaining variables held unchanged, if the account spending limits is flagged as No, i.e. there are no limits set on the account spending, the odds of the customer churning go down by 31.5%, as compared to when account spending limits is flagged as Yes. The variable is very important, which is evident form its p value (<0.0001) meaning a very high level of significance (99.99%). The variable is significant at the 95% level of significance.

From the Maximum Likelihood Estimates table, we can look at the log of odds of a positive response as a linear combination of the predictor variables.

For our model, we have, log[p / (1-p)] = 0.0177 + 0.1901*refurb - 0.00055*change_mou + 0.00263*change_rev - 0.00279*blck_dat_mean + 0.00358*roam_mean – 0.00937*drop_dat_mean – 0.00028*mou_opkd_mean – 0.0467*threeway_mean – 0.00868*custcare_mean – 0.00792*callfwdv_mean – 0.00041*opk_dat_mean – 0.3789*asl_flag_n.

Where p is the probability that churn is 1.

Looking at the Model fit statistics, we can see that the AIC value for the intercept only model is 137376.74 while the corresponding values for the intercept and covariates model is slightly lower at 136384.80. Lower the AIC of the model, better the model.

This means that our model, including the variables we have added, performs better than the model which only contains the intercept. In other words, the variables we have included to predict the churn status are good as they are able to perform better than a model which only has the intercept.

Similarly, the BIC/SC values for the intercept only model is 137386.25 while that for the intercept and covariates model is 136508.35. SC is a model fit criterion which penalizes the addition of more variables in the model. It is analogous to the Adjusted R^2 for a linear regression model. Since the SC value for our model is significantly higher than the corresponding AIC value, when compared to the values for the intercept only model, we can interpret this as the model being penalized for including more variables. In other words, it is possible that we may get a lower value for SC for our model, if some of the variables were dropped from the model.

A pair of observations with different response variables is said to be concordant if the observation with the lower ordered response value (churn=0 in this case) has a lower predicted mean score than the observation with the higher ordered response value (churn=1). A better model is defined by a higher corresponding concordance percentage. We want the model the value to be more than 50% as 50% can be looked at as pure chance and our model should be able to perform better than random chance.

For our model, the % concordant pairs is 56.1% meaning that our model is able to perform better than random assignment/ 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 43.9%.

The pairs represent the total number of possible pairs than can be formed from the dataset given the 2 different output values (churn=0 and churn=1). The dataset has 100,000 observations out of which 99,108 have been used in the model. These have the following split: 50,224 observations have a churn value of 0 while 48,884 have a churn value of 1. The product of these 2 values will give a value of 2,455,150,016, which is the total number of pairs formed.

Somers' D helps determine the strength and direction of relation between pair of variables. A value of -1 would mean that all the pairs disagree while a value of 1 means that all pairs agree. It can take any value between -1 and 1 and is defined as the difference between the concordant 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:

(56.1-43.9)/100 = 0.122. From the output table, we can see that the actual value is 0.123, meaning there was some rounding off done else we would've got the exact match.

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

Ans. Looking at the output, as per our model, the top 3 variables affecting the churn status are

1. asl_flag_N which represents the flag status No for account spending limits,

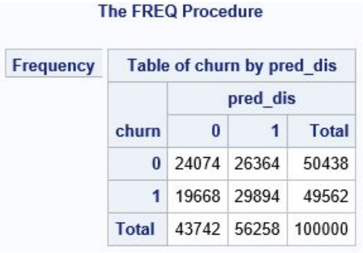
2. refurb, which represents if the customer is using a refurbished handset, and

3. threeway_mean, which represent the mean value of the number of threeway calls made by the customer.

These variables have been chose based on the fact that all of them are significant at the 95% significance level, and their coefficients/ parameters have the biggest absolute values. This can also be verified by the fact that for 1 unit change in their values, they have the biggest % impact on the odds of churning for a customer.

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.

Ans.



The FREQ Procedure

Frequency	Table of churn by pred_dis			
		pred_dis		
	churn	0	1	Total
	0	24074	26364	50438
	1	19668	29894	49562
	Total	43742	56258	100000

The table here gives the frequency distribution for the actual vs predicted values for our model. In other words, we have the confusion matrix to measure the performance of our model.

Hit ratio is defined as the % of observations that have been correctly classified. Hit ratio performance is always compared to the actual class distribution.

In this case, there are actually 49,562 out of 100,000 observations which have a churn value of 1.

Alternatively, there are 50,438 out of 100,000 observations or 50.438% observations with churn value of 0.

The corresponding value for churn value of 0 for our model is 43,742 out of 100,000 or 43.742%.

Compared to the actual data, we can conclude that our model has a decent hit ratio as it is able to predict 43.74% of the churn value 0 observations out of actually 50.44% churn value 0 observations correctly.

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

Ans. Looking at the dataset, we have already captured a lot of information in the various variables (0ver 170). But if we could include a variable that captures the feedback/ review of the service from the customer at regular intervals, it would help us understand how the customers perceive the service to be. We have captured total calls to customer care and minutes of use for customer care calls but I believe that only captures calls which would be made when the customers faced a problem. We need to put in a mechanism to get regular feedback from the customers

Similarly, we have record of retention offers made to the customer and how many of them were accepted. But if we could make other offers to the customer, not only when their term with the company is about to end, and record their feedback about the offer, it might help us design and make more targeted offers to the customers.

Both these approaches might make the customer feel valued which might in turn reduce the chances of churning.