Use the Churn data from a telecom company to understand what factors are good predictors of churn. Use a logistic regression model to build the best model (that is **best** in terms of model fit criteria).

Churn is the dependent variable that takes the value 1 if a customer has churned and 0 otherwise.

Make sure that there are no two explanatory variables that are highly correlated. Use correlation analysis to determine the correlation between the variables.

1. 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.

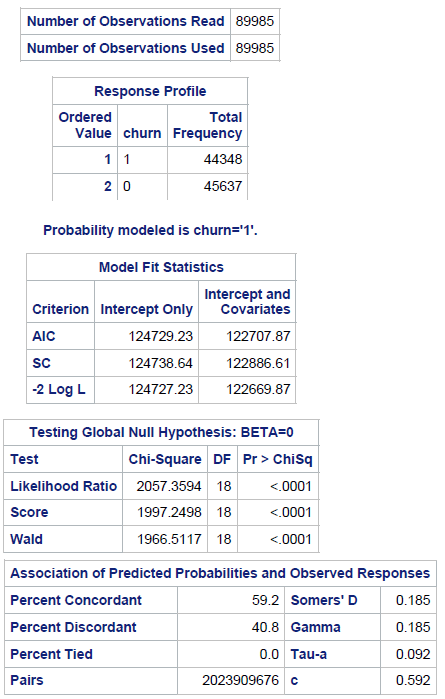
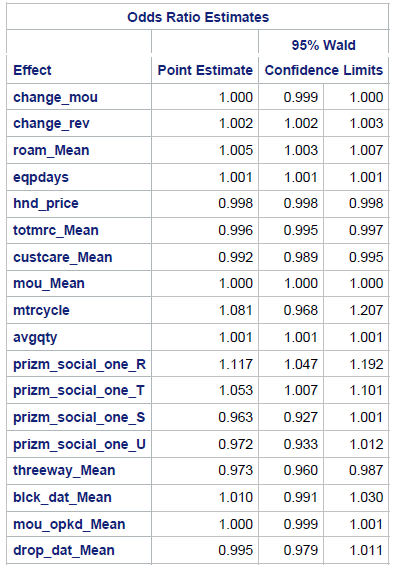
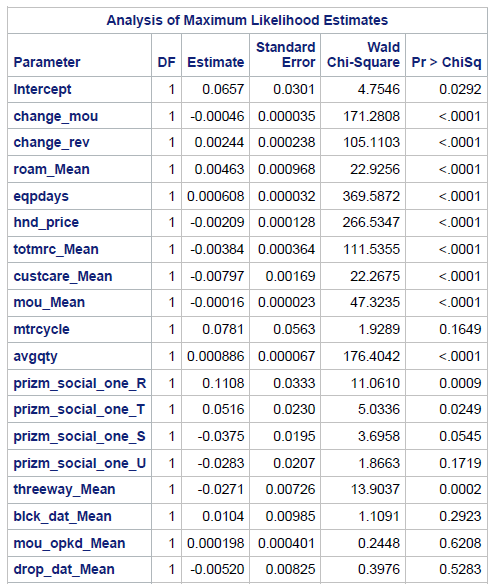
The given dataset has 100000 records with 173 variables.

**Variable selection procedure used**:

1. Calculated percentage difference in mean values corresponding to the category distribution of target variable.
2. Used Decision Tree as variable selection method to check for the feature importance matrices for categorical and continuous variables.
3. Checked the distribution of target categories with categorical independent variables to find the most diverse categories for independent variables selection.
4. After selection of top 30 variables based on maximum percentage difference and feature importance, checked for amount of NULL/missing values.
5. Selected top 15 variables after checking for correlation among them.

**Results of Logistic regression:**

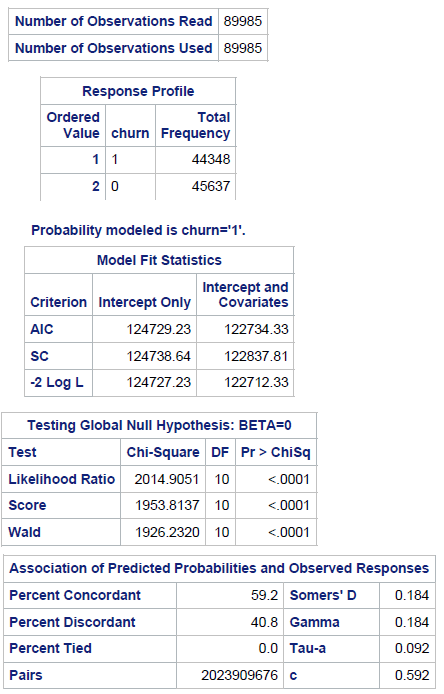
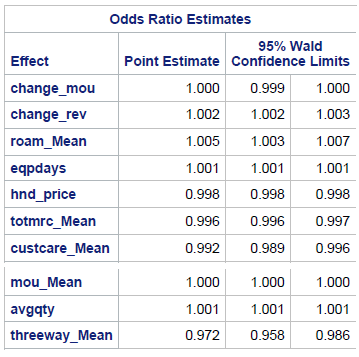
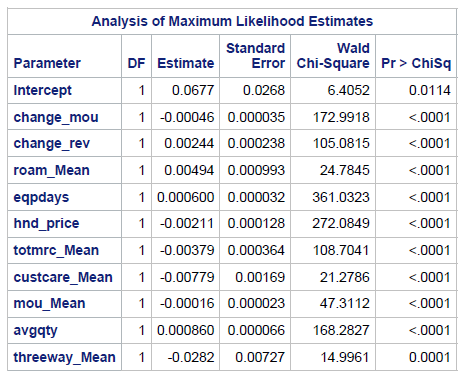
The response variable is “churn” with 2 response levels: 1 (Yes) and 0 (No), which corresponds to whether the customer will churn or not. The data is quite evenly distributed in the 2 categories.

The number of observations read and used is 89985 which is 89% of given observations after removing the null/missing value records.

**MODEL INTERPRETATION:**

**After removing the insignificant variables:**



**Model Fit Statistics:**

AIC – This is the Akaike Information Criterion. AIC, like Adjusted R-square in linear regression, penalize the log-likelihood for the number of predictors in the model. Ultimately, the model with the smallest AIC is considered good.

SC – This is the Schwarz Criterion. Like AIC, SC penalizes for the number of predictors in the model and the smallest SC is most desirable.

As AIC and SC value for Intercept and covariates (122734.33 and 122837.81 respectively) is better (lower) than Intercept only (124729.23 and 124738.64 respectively) model, we can say that our model does better.

**Meaning of coefficients:**

change\_mou (% change of minutes of use): For every 1 unit change in variable change\_mou, the log odds of churn (versus non-churn) decreases by 0.00046 units, keeping all other variables constant.

change\_rev (% change of revenue): For every 1 unit change in variable change\_rev, the log odds of churn (versus non-churn) increases by 0.0024 units, keeping all other variables constant.

roam\_Mean (mean value of roaming): For every 1 unit change in variable roam\_Mean, the log odds of churn (versus non-churn) increases by 0.00496 units, keeping all other variables constant.

eqpdays (Number of days of the current equipment): For every 1 unit change in variable eqpdays, the log odds of churn (versus non-churn) increases by 0.0006 units, keeping all other variables constant.

hnd\_price (Handset price): For every 1 unit change in variable hnd\_price, the log odds of churn (versus non-churn) decreases by 0.002 units, keeping all other variables constant.

totmrc\_Mean (total MRC, mean (monthly recurring charge)): For every 1 unit change in variable totmrc\_Mean, the log odds of churn (versus non-churn) decreases by 0.00379 units, keeping all other variables constant.

custcare\_Mean (Customer Care Calls): For every 1 unit change in variable custcare\_Mean, the log odds of churn (versus non-churn) decreases by 0.0077 units, keeping all other variables constant.

mou\_Mean (MINUTE\_QTY): For every 1 unit change in variable mou\_Mean, the log odds of churn (versus non-churn) decreases by 0.00016 units, keeping all other variables constant.

mtrcycle (Motorcycle Indicator): Log odds of churn (versus non-churn) increases by 0.0802 units, when the motorcycle indicator is 1 as compared to 0, keeping all other variables constant.

avgqty (average number of calls): For every 1 unit change in variable avgqty, the log odds of churn (versus non-churn) increases by 0.00086 units, keeping all other variables constant.

threeway\_Mean (Three Way Calls): For every 1 unit change in variable threeway\_Mean, the log odds of churn (versus non-churn) decreases by 0.0251 units, keeping all other variables constant.

**Significance**:

**Pr > ChiSq** : Out of the initially chosen 15 variables, all variables came out significant based on P-value except for mtrcycle, prizm\_social\_one, blck\_dat\_Mean, mou\_opkd\_Mean and drop\_dat\_Mean independent variables. So we later ran the model after dropping these variables. We are testing the null hypothesis that all of the regression coefficients in the model are equal to zero and alternate that at least one of them is not equal to zero.

The likelihood ratio chi-square of 2014.9051 with a p-value < 0.0001 tells us that our model fits significantly better than an empty model.

**Prediction accuracy (percent concordance)**: Percentage concordance tells us how good our model is. The higher the concordance ratio, the better is the model. For the calculation of Concordance, we will make all the possible pairs of (1) and (0). The total number of Concordant pairs are counted and divided by the total number of pairs to calculate concordance ratio. The above model gives us the concordance percentage of 59.2 which is good.

**Odds-ratios**:

change\_mou (% change of minutes of use): For a one unit increase in change\_mou variable, the odds of churn (versus non-churn) increase by a factor of 1, keeping all other variables constant.

change\_rev (% change of revenue): For a one unit increase in change\_rev variable, the odds of churn (versus non-churn) increase by a factor of 1.002, or approximately 0.002 percent increase, keeping all other variables constant.

roam\_Mean (mean value of roaming): For a one unit increase in roam\_Mean variable, the odds of churn (versus non-churn) increases by a factor of 1.005, or approximately 0.005 percent increase, keeping all other variables constant.

eqpdays (Number of days of the current equipment): For a one unit increase in eqpdays variable, the odds of churn (versus non-churn) increases by a factor of 1.001, keeping all other variables constant.

hnd\_price (Handset price): For a one unit increase in hnd\_price variable, the odds of churn (versus non-churn) decreases by a factor of 0.998, or approximately 0.002 percent decrease, keeping all other variables constant.

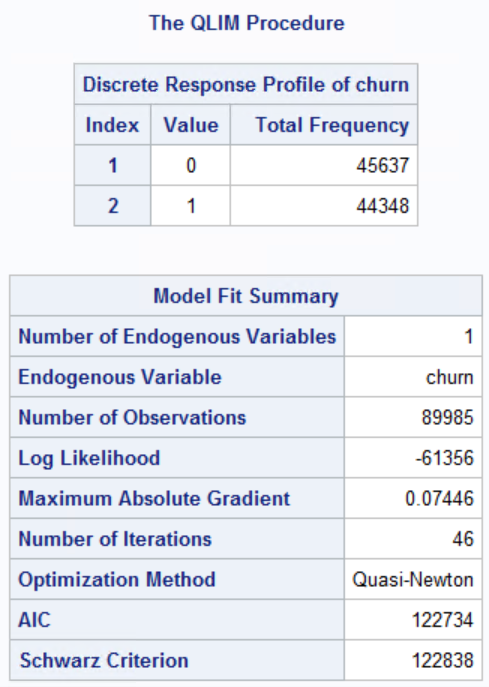
totmrc\_Mean (total MRC, mean (monthly recurring charge)): For a one unit increase in totmrc\_Mean variable, the odds of churn (versus non-churn) decreases by a factor of 0.996, or approximately 0.004 percent decrease, keeping all other variables constant.

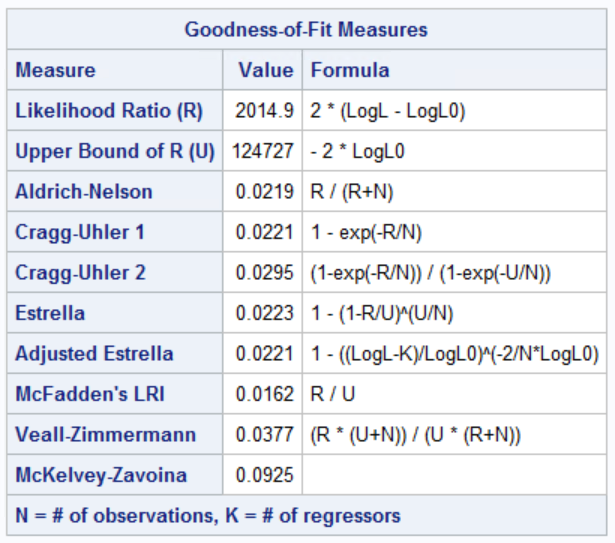
custcare\_Mean (Customer Care Calls): For a one unit increase in totmrc\_Mean variable, the odds of churn (versus non-churn) decreases by a factor of 0.992, or approximately 0.008 percent decrease, keeping all other variables constant.

mou\_Mean (MINUTE\_QTY): For a one unit increase in mou\_Mean variable, the odds of churn (versus non-churn) increase by a factor of 1, keeping all other variables constant.

avgqty (average number of calls): For a one unit increase in avgqty variable, the odds of churn (versus non-churn) increase by a factor of 1.001, keeping all other variables constant.

threeway\_Mean (Three Way Calls): For a one unit increase in threeway\_Mean variable, the odds of churn (versus non-churn) decreases by a factor of 0.972, or approximately 0.03 percent decrease, keeping all other variables constant.

**Goodness of fit - McFadden’s R2**:



McFadden suggested a likelihood ratio index that is analogous to the http://support.sas.com/documentation/cdl/en/etsug/63348/HTML/default/images/etsug_qlim0090.pngin the linear regression model to measure the Goodness of fit. Used GLIM procedure to calculate the McFadden’s R-square for the churn dataset of value 0.0162. The model with McFadden's values of 0.2 to 0.4 for rho-squared represent Good fit.



Logistic regression generates **Cox-Snell R2**, which can also be used to measure the predictive power of the model. For the given dataset, both Cox-Snell (0.02) and McFadden R-square (0.016) can be calculated but usually Cox-Snell measures for binary logistic regression and McFadden’s measure for multinomial and ordered logit.

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

From the results we can see that the coefficient estimates of all the 10 variables are not very different (wide) from each other. That means a unit increase in the value of an independent variable will not cause a significant increase or decrease in the value of target variable as compared to other variable.

But on running the logistic model for different combination of the 10 variabes, change\_mou, eqpdays and hnd\_price gives us a concordance percentage of 58.1, which is highest for 3 variable pair. Hence, the top 3 factors that affect churn in our model are:

1. change\_mou (% change of minutes of use)
2. eqpdays (Number of days of the current equipment)
3. hnd\_price (Handset price)
4. What other variables (that if collected) would help to improve the fit of the model.

a. **Customer Satisfaction Rating**

The survey data on the customer satisfaction can be helpful in improving the model fit. If the customer is rating the experience as poor after a customer care call, then the customer is more likely to churn as compared to a customer who is rating the customer service as good. This variable shall enhance the model fit and improve model accuracy.

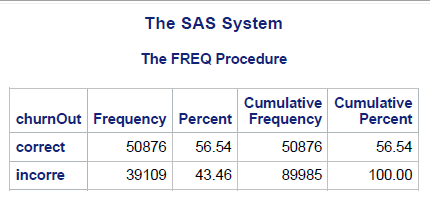
b. **Competitor Promotional Offer**

If there are any attractive schemes or discounts being offered by the competitor companies, then customers are more likely to switch to those careers. So, a variable or a combination of multiple variables capturing the promotion offers by one or more competitors shall improve the model fit and accuracy.

c. **International Plan Subscription**

If a customer gets involved in making calls to international numbers due to any reason, then depending upon the rates of international plans as compared to rates of other service providers, the customer churn may happen if the call rates are not suitable for the customer. Introducing a variable which captures that if customer has international calling plan may be helpful in improving the model fit.

1. 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.



**Hit Rate**: Percentage of observations that are correctly predicted by the model.

Hit Rate = 50876/(50876+39109) = 56.54

