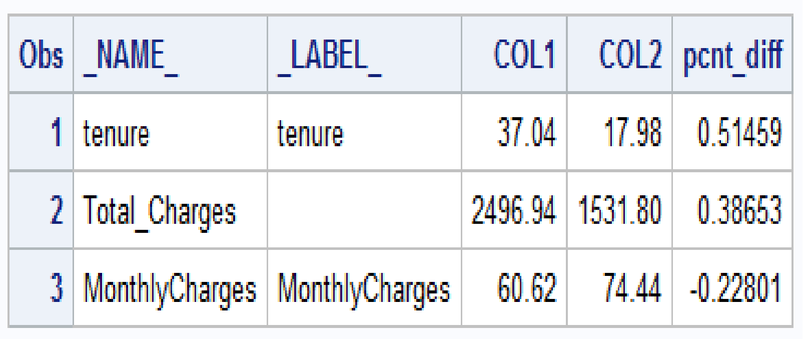
CUSTOMER CHURN DATA ANALYSIS FOR A LEADING TELECOMMUNICATION COMPANY

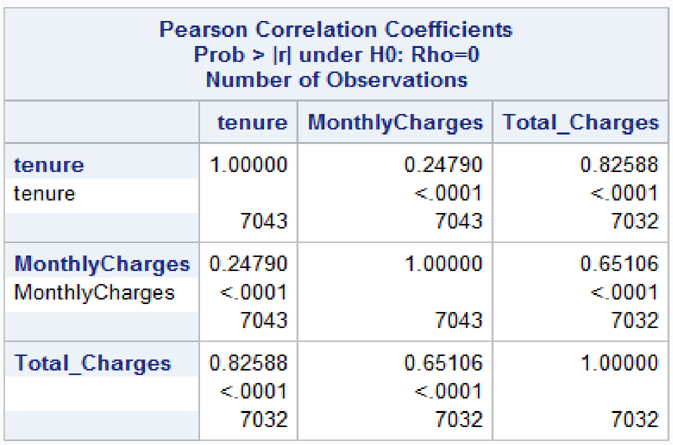
Model Selection:

- We have 3 continuous and 18 categorical variables in the given dataset.
- Firstly, we computed percentage mean difference for the continuous variables based on churn and no churn and found that the maximum % difference is given by tenure.
- Then, we also tested for the correlation of these 3 variables and found that tenure and total charges are highly corelated, we kept tenure for our model.
- Then we developed a logistic regression model including tenure and monthly charges along with all other categorical variables (after converting then to dummies).
- Finally, we came up with below best model using economic theory, significance of the variables and the model which gives best fit criteria and percentage concordant

```
churn=b0+b1*Dep+b2*PS+b3*ML_Y+b4*IS_FO+b5*IS_D+b6*OS_Y+b7*TS_Y+b8*ST_Y+b9
*SM_Y+b10*Con_1+b11*Con_2+b12*PB+b13*PM_Elec+b14*tnure+b15*monthlyCharges
```



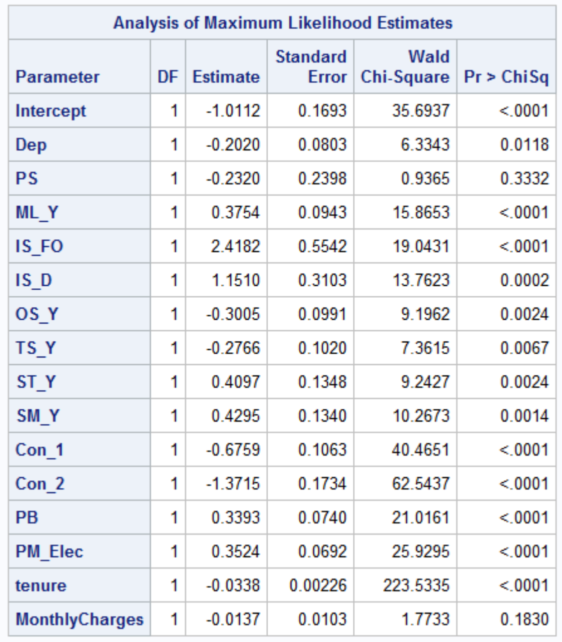
Obs	_NAME_	_LABEL_	COL1	COL2	pcnt_diff
1	tenure	tenure	37.04	17.98	0.51459
2	Total_Charges		2496.94	1531.80	0.38653
3	MonthlyCharges	MonthlyCharges	60.62	74.44	-0.22801



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

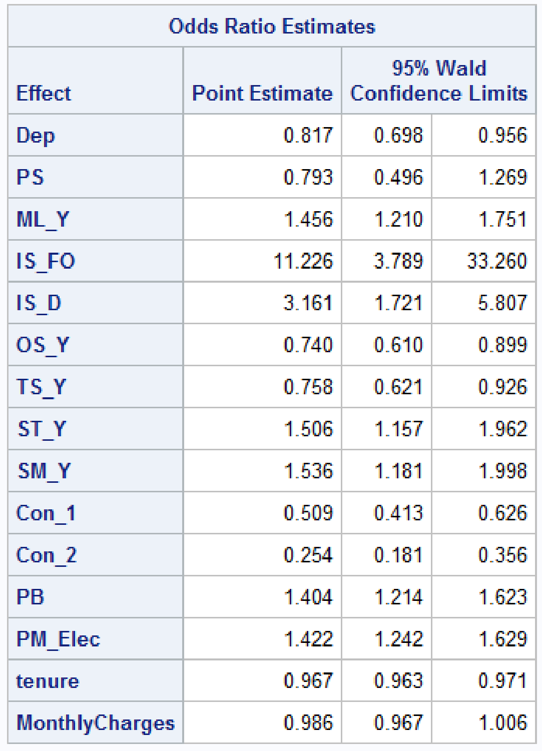
	tenure	MonthlyCharges	Total_Charges
tenure tenure	1.00000 7043	0.24790 <.0001 7043	0.82588 <.0001 7032
MonthlyCharges MonthlyCharges	0.24790 <.0001 7043	1.00000 7043	0.65106 <.0001 7032
Total_Charges	0.82588 <.0001 7032	0.65106 <.0001 7032	1.00000 7032

Interpreting the logistic output explaining AIC/BIC, meaning of coefficients, significance, prediction accuracy (percent concordance), odds-ratios etc.



Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.0112	0.1693	35.6937	<.0001
Dep	1	-0.2020	0.0803	6.3343	0.0118
PS	1	-0.2320	0.2398	0.9365	0.3332
ML_Y	1	0.3754	0.0943	15.8653	<.0001
IS_FO	1	2.4182	0.5542	19.0431	<.0001
IS_D	1	1.1510	0.3103	13.7623	0.0002
OS_Y	1	-0.3005	0.0991	9.1962	0.0024
TS_Y	1	-0.2766	0.1020	7.3615	0.0067
ST_Y	1	0.4097	0.1348	9.2427	0.0024
SM_Y	1	0.4295	0.1340	10.2673	0.0014
Con_1	1	-0.6759	0.1063	40.4651	<.0001
Con_2	1	-1.3715	0.1734	62.5437	<.0001
PB	1	0.3393	0.0740	21.0161	<.0001
PM_Elec	1	0.3524	0.0692	25.9295	<.0001
tenure	1	-0.0338	0.00226	223.5335	<.0001
MonthlyCharges	1	-0.0137	0.0103	1.7733	0.1830



Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Dep	0.817	0.698	0.956
PS	0.793	0.496	1.269
ML_Y	1.456	1.210	1.751
IS_FO	11.226	3.789	33.260
IS_D	3.161	1.721	5.807
OS_Y	0.740	0.610	0.899
TS_Y	0.758	0.621	0.926
ST_Y	1.506	1.157	1.962
SM_Y	1.536	1.181	1.998
Con_1	0.509	0.413	0.626
Con_2	0.254	0.181	0.356
PB	1.404	1.214	1.623
PM_Elec	1.422	1.242	1.629
tenure	0.967	0.963	0.971
MonthlyCharges	0.986	0.967	1.006

The p values of the variables indicate that all variables are statistically significant except Phone Service and monthly charges.

Dep: Controlling for other factors and as compared to a customer without dependents, the odds of churn of a customer with dependents is 18.3% less.

ML_Y: Controlling for other factors and as compared to a customer without Multiple Lines, the odds of churn of a customer with multiple lines is 45.6% more.

IS_FO: Controlling for other factors and as compared to a customer without Internet service, the odds of churn of a customer with Fiber Optic Cable is 1022.6% more.

IS_D: Controlling for other factors and as compared to a customer without internet service, the odds of churn of a customer with DSL is 216.1% more.

OS_Y: Controlling for other factors and as compared to a customer without internet service, the odds of churn of a customer with Online Security is 26% less.

TS_Y: Controlling for other factors and as compared to a customer without Internet Service, the odds of churn of a customer with Tech Support is 42.2% less.

ST_T: Controlling for other factors and as compared to a customer without Internet Service, the odds of churn of a customer with Streaming TV is 50.6% more.

SM_Y: Controlling for other factors and as compared to a customer without Internet Service, the odds of churn of a customer with Streaming Movies is 53.6% more.

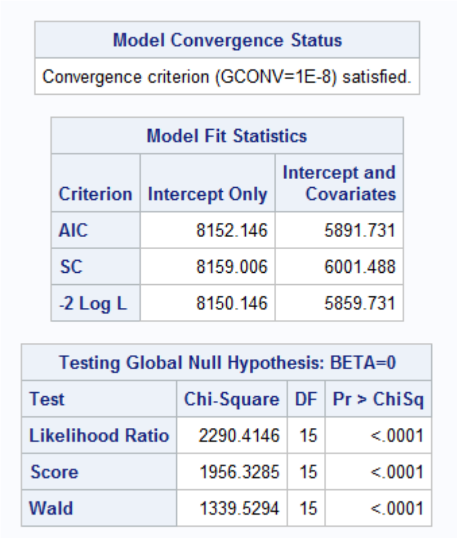
Con_1: Controlling for other factors and as compared to a customer with monthly contract, the odds of churn of a customer with one-year contract is 49.1% less.

Con_2: Controlling for other factors and as compared to a customer with monthly contract, the odds of churn of a customer with two-year contract is 74.6% less.

PB: Controlling for other factors and as compared to a customer without paperless billing, the odds of churn of a customer with paperless billing is 40.4% more.

PM: Controlling for other factors and as compared to a customer having credit card payment method, the odds of churn of a customer with payment methods as electronic cheque is 42.2% more.

Tenure: Controlling for other factors for every month increase in tenure, the odds of churn of a customer decreases by 3.3% less.



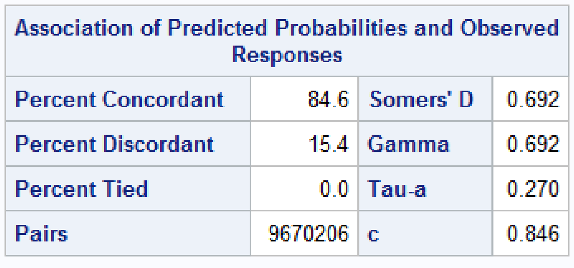
Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	8152.146	5891.731
SC	8159.006	6001.488
-2 Log L	8150.146	5859.731

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2290.4146	15	<.0001
Score	1956.3285	15	<.0001
Wald	1339.5294	15	<.0001

As all our model fit statistics for our chosen model are less than that for an intercept only model, we can conclude that we have picked good explanatory variables.

The McFadden's R-square is 28.1%, implying our model can explain 28.1% of the variation in our data.



Association of Predicted Probabilities and Observed Responses			
Percent Concordant	84.6	Somers' D	0.692
Percent Discordant	15.4	Gamma	0.692
Percent Tied	0.0	Tau-a	0.270
Pairs	9670206	c	0.846

Concordance signifies the how much our data agrees with our predictions. Our model shows 84.6% concordance implying a good model fit.

Top three factors that affect churn :-

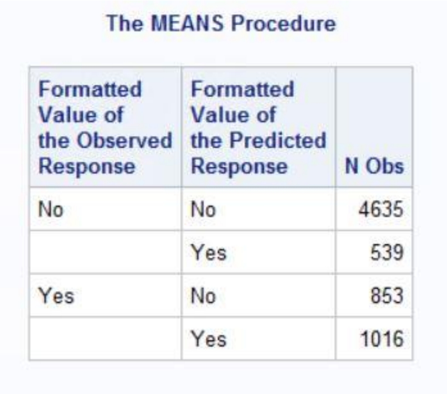
Internet Service, Contract Period, Streaming Movies are the top three factors that affect the churn in our model. These parameters also make intuitive sense.

Other variables (that if collected) would help to improve the fit of the model.

Competitors pricing: This is a significant contributor of churn rate as the customer who observes cheaper plans for similar service he uses is most likely to switch to a cheaper alternative.
Cost paid/Usage cost ratio: If the customer usage is way too less than what he pays eventually he will churn out to a cheaper provider. It will be good to know this ratio and recommend them cheaper plans.
Customer service: Aspects such as promptness (waiting time) either at the service outlet or on a phone call is an important factor of how the provider treats the customer. It can certainly instigate a frustrated customer to churn out of the system.
Frequency of Offers: Discounts coupled with right advertising will certainly help retain the customer for longer time.
Quality of service index: Irrespective of all the other factors mentioned above, if the quality of service is bad will not help retain the customer.
Minutes of Use: If a customer uses more minutes in recent months, probably they are less likely to churn
Age of device: If the device a customer is using is old, there is good chance of churn.

HIT RATIO:-

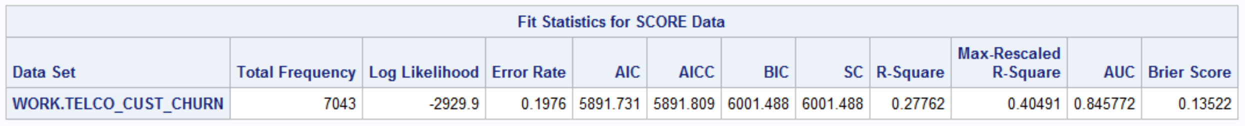
Hit ratio is defined as the percentage of correct predictions using the logit model. Use the model to predict 1 or 0 using the same data



The MEANS Procedure

Formatted Value of the Observed Response	Formatted Value of the Predicted Response	N Obs
No	No	4635
	Yes	539
Yes	No	853
	Yes	1016

The hit ratio is measured as the percentage of accurate predictions. Thus, our hit ratio is (4635+1016)/7043=80.2%. Below are the fit statistics for our prediction model:



Fit Statistics for SCORE Data

Data Set	Total Frequency	Log Likelihood	Error Rate	AIC	AICC	BIC	SC	R-Square	Max-Rescaled R-Square	AUC	Brier Score
WORK.TELCO_CUST_CHURN	7043	-2929.9	0.1976	5891.731	5891.809	6001.488	6001.488	0.27762	0.40491	0.845772	0.13522

Since our ROC is greater than 0.5, we can conclude that we have a reliable model.