# Data set Properties

* There are no NA's or Missing values in the data, ref NA Count vector
* All the columns are numeric.
* The data set is unbalanced ~ 85% skewed to non-churn customers

We have divided the given dataset into 7:3 ratios for training and testing respectively, this is a random selection of records, done using in R.

# Model Properties (training data)

Log Likelihood Ratio Test:

|  |
| --- |
| Likelihood ratio test  Model 1: Churn ~ AccountWeeks + ContractRenewal + DataPlan + DataUsage +  CustServCalls + DayMins + DayCalls + MonthlyCharge + OverageFee +  RoamMins  Model 2: Churn ~ 1  #Df LogLik Df Chisq Pr(>Chisq)  1 11 -771.91  2 1 -981.21 -10 418.59 < 2.2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |

* Overall test of the model is significant. Indicating that Churn response depends upon all the other measures in the data set.
* This implies that the null hypothesis of all Betas are zero is rejected and we conclude that at least one Beta is nonzero.

Pseudo R Square:

|  |
| --- |
| llh llhNull G2 McFadden r2ML r2CU  -771.914589 -981.209562 418.589946 0.213303 0.164182 0.288731 |

* From McFadden R Square, we can say that 21.33% of the uncertainty of Intercept model has been explained by Full model.
* This is a moderate fit.

Individual Coefficients:

|  |
| --- |
| Estimate Std. Error z value Pr(>|z|)  (Intercept) -5.6407726 0.6508575 -8.667 < 2e-16 \*\*\*  AccountWeeks 0.0002124 0.0016589 0.128 0.89812  ContractRenewal -1.9961398 0.1720420 -11.603 < 2e-16 \*\*\*  DataPlan -0.9020348 0.6420953 -1.405 0.16007  DataUsage 0.5588354 2.3022013 0.243 0.80821  CustServCalls 0.5269013 0.0464005 11.356 < 2e-16 \*\*\*  DayMins 0.0228111 0.0388610 0.587 0.55721  DayCalls 0.0014684 0.0032850 0.447 0.65486  MonthlyCharge -0.0572357 0.2283152 -0.251 0.80206  OverageFee 0.2305793 0.3896235 0.592 0.55398  RoamMins 0.0768815 0.0266086 2.889 0.00386 \*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |

* From P-values we can see ContractRenewal, CustServCalls and RoamMins are highly significant in explaining the Churn effect.
* We can't conclude any practical business insights at this stage, let's see the individual odds.

ODDS:

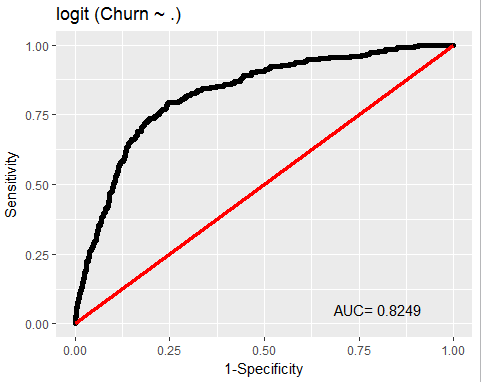
|  |
| --- |
| odds  (Intercept) AccountWeeks ContractRenewal DataPlan DataUsage CustServCalls  0.003550124 1.000212412 0.135858721 0.405743228 1.748634928 1.693675999  DayMins DayCalls MonthlyCharge OverageFee RoamMins  1.023073216 1.001469515 0.944371461 1.259329288 1.079914060  > prob  (Intercept) AccountWeeks ContractRenewal DataPlan DataUsage CustServCalls  0.003537566 0.500053097 0.119608820 0.288632533 0.636183041 0.628760103  DayMins DayCalls MonthlyCharge OverageFee RoamMins  0.505702516 0.500367109 0.485694982 0.557390768 0.519210904 |

* After seeing OODs we can say that the highly significant ContractRenewal is having low odds i.e. by every 1-unit increase of ContractRenewal (If a customer renews his/her contract) then there is a 11% probability in turning that customer from churn to being non-churn, which is every low % when compared to other attributes.
* CustServCalls and DataUsage are still highly helpful in changing a customer non-churn, I.e. 1-unit increase in these are having 62.8% and 63.6% probability respectively in turning a churn customer to non-churn.
* Similarly, AccountWeeks, DayMins, DayCalls, MonthlyCharge, OverageFee, RoamMins have significant probability of 50.0%, 50.5%, 50.0%, 48.5%, 55.7%, 51.9% probability respectively in turning a churn customer to non-churn.

Classification Table:

|  |
| --- |
| Predicted  Actual 0 1 Sum  0 1972 15 1987  1 320 27 347  Sum 2292 42 2334  **Percentage representation**  Predicted  Actual 0 1  0 0.844901457 0.006426735  1 0.137103685 0.011568123 |

* 1999 out of 2234 observations are correctly classified, I.e. overall accuracy of 85.6%.
* 320 Actual churn observations are miss classified as non-churn and 15 vice versa.
* In this model we have unbalanced Type I and Type II error and this is due to the data being skewed towards non-churn observations.

ROC-Plot:

* This tell us that 82.5% of the points are explained by this model, but this % is highly subjected to the skewness in the data.

# Results on Test data set

Classification Table: (Original data – Logistic Regression)

|  |
| --- |
| Predicted  Actual 0 1 Sum  0 857 6 863  1 127 9 136  Sum 984 15 999  **Percentage representation**  Predicted  Actual 0 1  0 0.857857858 0.006006006  1 0.127127127 0.009009009 |

* 866 out of 999 observations are correctly classified, I.e. overall accuracy of 86.6%. Almost similar to the training set results, thus we conclude that there is no overfitting in this model.
* If we observe the type I error (Incorrect prediction of a churn customer) is of 12.7 % while the type II error (Incorrect prediction of non-churn customer) is only 0.6% this is due to the data being skewed towards more non churn customers.

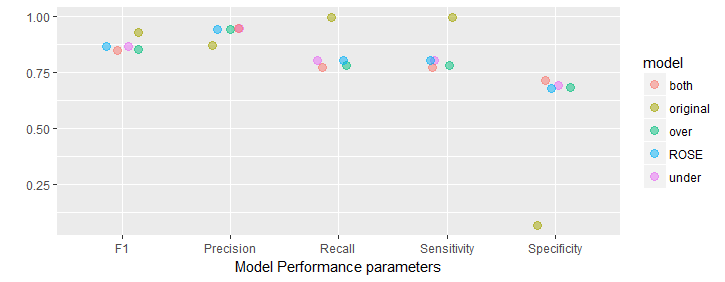
*To improve the type I error we can apply sampling techniques on the data.*

# Results after Sampling

We have applied 3 different kinds on sampling techniques and compared the results in the following table. The main reason for sampling the data even though we have 86.6% accuracy overall is that we need to have less incorrect predictions of churn customers (focus/target customers for all campaigns) - Type I error rate.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Original Data** | **ROSE (Random Over Sampling Examples)** | **Oversampling** | **Undersampling** | **Both Sampling** |
| **No of 0's** | 1987 | 1218 | 1987 | 403 | 1195 |
| **No of 1's** | 347 | 1116 | 1513 | 347 | 1139 |
| **Train count** | 2334 | 2334 | 3500 | 750 | 2334 |
| **LogLiklyhood** | 2.2e(-16) | 2.2e(-16) | 2.2e(-16) | 2.2e(-16) | 2.2e(-16) |
| **McFadden R2** | 21.30% | 16.90% | 24.93% | 23.00% | 24.12% |
| **Type1** | 320-13.7% | 370-15.8% | 387-11.0% | 97-12.9% | 270-11.5% |
| **Type2** | 15-0.6% | 298-12.7% | 421-12.0% | 81-10.8% | 256-10.9% |
| **Overall%** | 85.60% | 71.3%% | 76.91% | 76.26% | 77.46% |
| **ROC Plot** | 82.49% | 77.18% | 82.74% | 81.87% | 82.07% |
| **Test-Type1** | 127-12.7% | 44-4.4% | 43-4.3% | 42-4.2% | 39- 3.9% |
| **Test-Type2** | 6-0.6% | 170-17.01% | 189-18.9% | 171-17.1% | 198-19.81% |
| **Overall-Fault%** | 133-13.3% | 214-21.4% | 232-23.2% | 213-21.3% | 237-23.7% |
| **Overall%** | 86.60% | 78.50% | 76.70% | 78.70% | 76.27% |

* Out of all the samplings, ROSE sampling yielded better results. This has the lowest type II error 17.01% while keeping the type I error to 4.4%.

The following is the graphical representation of various modeling performance indicators.

# Conclusion

ROSE sampled logistic regression model is very much useful if the company needed to perform any campaigns to lower the churning ratio (as here correct prediction of churn customers is important) but in case if they want to reward any loyalty points or such kind of campaigns they may consider logistic regression model on original dataset (as here correct prediction of non-churn customers is important).